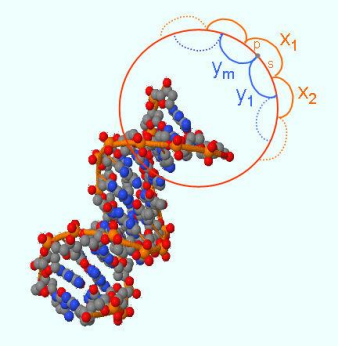


UFR Mathématique Informatique Strasbourg

Professeur Christian Michel

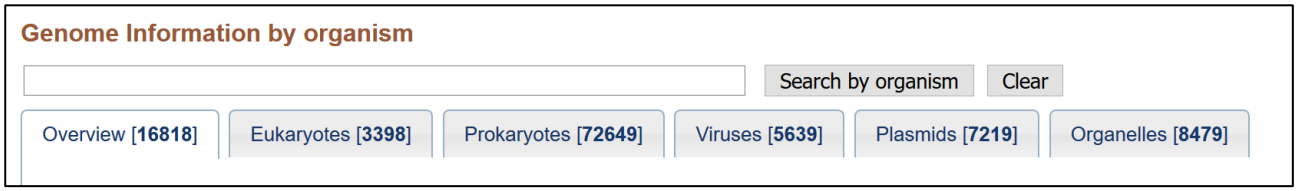
Projet Bioinformatique Master SDSC 2019 - 2020

BIG DATA EN BIOINFORMATIQUE

Statistiques sur les trinucléotides dans les gènes

1. Recherche des gènes et génomes des eucaryotes, procaryotes (bacteria, archaea), virus, plasmides et organelles (mitochondries, chloroplastes) dans la base de données de gènes GenBank

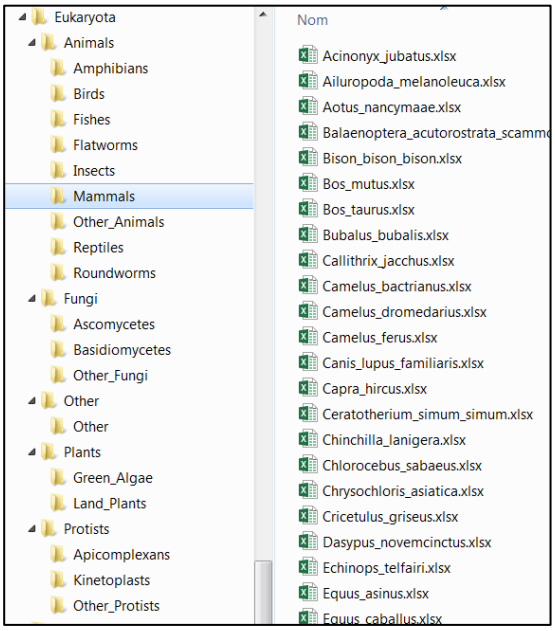
http://www.ncbi.nlm.nih.gov/genome/browse/



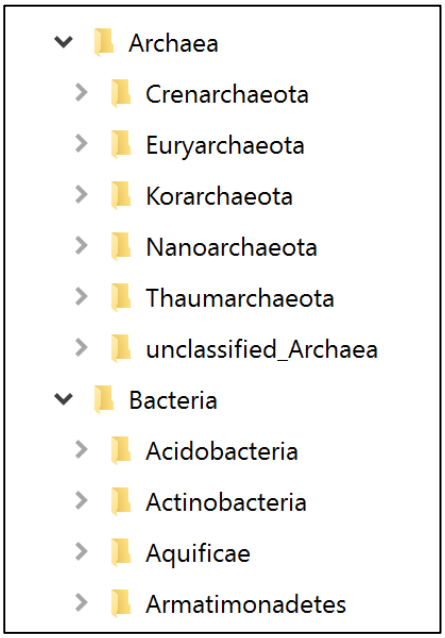
1.1. Stockage des résultats dans des fichiers Excel (extension xlsx)
Impératif: Respectez l'arborescence du site.

Results\Kingdom\Group\SubGroup\Organism.xlsx

Pour les eucaryotes

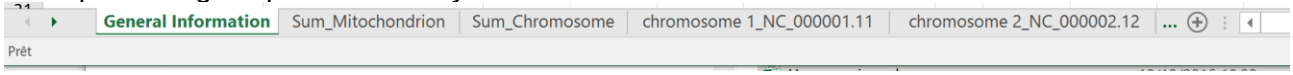


Pour les procaryotes

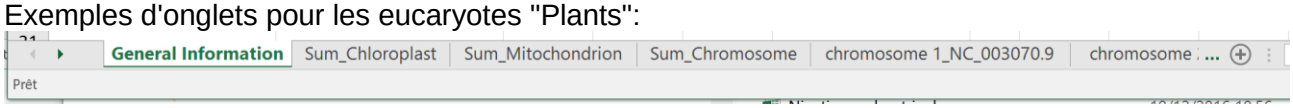


Dans chaque fichier Excel "Organism", plusieurs onglets sont associés à chaque mot-clé.

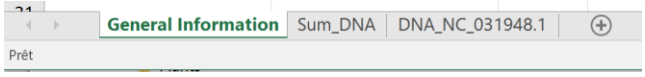
Exemples d'onglets pour les eucaryotes "Mammals":



Exemples d'onglets pour les eucaryotes "Plants":



Exemples d'onglets pour les virus:



IMPORTANT:
- Ne considérer que les identifiants NC (donc rejeter les identifiants NW, AC, etc.).
- Prendre tous les génomes.
- Utiliser les API de GenBank.

IMPORTANT: Dans chaque fichier Excel, créer un onglet "Sum_Chromosome" (en général plusieurs chromosomes), "Sum_Mitochondrion" (en général un mitochondrion), "Sum_Chloroplast" (en général plusieurs chloroplasts), "Sum_DNA" (en général plusieurs DNA), etc.
Il est inutile de créer un onglet faisant la somme générale (somme des "Sum") puisque l'information génétique est différente.

IMPORTANT: Pour chaque "SubGroup", chaque "Group" et chaque "Kingdom", on créera un fichier Total_SubGroup.xlsx, Total_Group.xlsx et Total_Kingdom.xlsx qui fera la somme de tous les onglets "Sum" des fichiers Excel associés au répertoire donné.
Exemple: Total_Mammals.xlsx, Total_Animals.xlsx et Total_Eukaryota.xlsx.

Pour des raisons non détaillées ici, gardez les 2 termes "Sum" et "Total" pour avoir une différence dans les mots-clés, et utiliser le symbole "_".
IMPORTANT: Mettre à jour les fichiers Excel dès que possible.

1.2. Stockage des fichiers de gènes et génomes dans des fichiers texte (extension txt)

On créera, en option utilisateur en raison de la taille des fichiers, 2 répertoires:

(i) Genome\Kingdom\Group\SubGroup\Organism\Genome_Organism.txt pour stocker la séquence entière du génome.

(ii) Gene\Kingdom\Group\SubGroup\Organism\Gene_1_Organism.txt, Gene_2_Organism.txt, etc. pour stocker les gènes valides (utilisés pour le traitement statistique des trinucléotides).
Les fichiers de gènes sont nommés selon leur identificateur GenBank et devraient être compressés (extension zip).

IMPORTANT: Tous les résultats sont remontés aux nœuds parents de l'arborescence en sommant les nombres et en recalculant les fréquences.

1.3. Présentation des fichiers résultats Excel

	A	B	C	D	E	F	G
1	**Information**						
2							
3	**Name**	Procaryotes				**Genome**	
4						**Chromosome**	2152
5	**Modification Date**	2016/111/19				**Plasmid**	1875
6						**DNA**	4
7	**Number of CDS sequences**	6339722					
8							
9	**Number of invalids CDS**	2899					
10							
11	**Number of Organisms**	2076					
12							
13							

	A	B	C	D	E	F	G	H	I	J
1		Phase 0	Freq Phase 0	Phase 1	Freq Phase 1	Phase 2	Freq Phase 2	Pref Phase 0	Pref Phase 1	Pref Phase 2
2	AAA	55736232	2.85	48359279	2.47	43743339	2.23	3631565	2535287	1704318
3	AAC	35198120	1.80	28108179	1.44	22362191	1.14	3050568	2243505	1545131
4	AAG	39498112	2.02	53192166	2.72	14991830	0.77	2525464	3568207	400053
5	AAT	37328243	1.91	24378688	1.24	31815665	1.62	3203130	1955184	2495527
6	ACA	19702968	1.01	33029858	1.69	19684531	1.01	1646500	3586178	1615804
7	ACC	41648883	2.13	32925758	1.68	14393415	0.74	3379165	2197866	1364041
8	ACG	27182580	1.39	49334342	2.52	12678191	0.65	1737618	4184414	958571
9	ACT	17713807	0.90	20374111	1.04	21171637	1.08	1929986	2548229	2673734
10	AGA	11178853	0.57	31188471	1.59	44398344	2.27	384532	2772702	3390592
11	AGC	26878865	1.37	38440007	1.96	29749366	1.52	1441054	2863388	2374104
12	AGG	7122730	0.36	46967299	2.40	24894552	1.27	320869	3962000	2179114
13	AGT	15842010	0.81	18393640	0.94	22081707	1.13	1668882	2742803	2805339
14	ATA	18804157	0.96	36322604	1.85	11429037	0.58	1904164	4590328	1509540
15	ATC	53183965	2.72	30890278	1.58	14204964	0.73	3924681	1932434	670886
16	ATG	45383956	2.32	49111710	2.51	7505932	0.38	3180823	3103730	139454
17	ATT	47071076	2.40	26976284	1.38	28901267	1.48	3963127	1801267	1707737
18	CAA	30580929	1.56	23883125	1.22	49660588	2.54	1687244	1233006	3611387
19	CAC	20794860	1.06	16686876	0.85	34633234	1.77	1662408	1944501	3514085

General Information | Sum_Chromosome | Sum_Plasmid | Sum_DNA

K	L	M	N	O	P	Q	R	S
	STATISTIQUES DINUCLEOTIDES		Phase 0	Freq Phase 0	Phase 1	Freq Phase 1	Pref Phase 0	Pref Phase 1
		AA	217017332	7.39	217694712	7.42	3122993	3216902
		AC	154725937	5.27	155114144	5.28	3091264	3132555
		AG	158209304	5.39	158926540	5.41	3064267	3150589
		AT	186569143	6.36	182339580	6.21	3359329	2842813
		CA	173670924	5.92	173510526	5.91	3108432	3094001
		CC	180199915	6.14	180334052	6.14	3159552	3185408
		CT	156614319	5.33	157319977	5.36	3065086	3155030
		CG	238024554	8.11	238082302	8.11	3116084	3116020
		GA	199970489	6.81	197420094	6.72	3228752	2944140
		GC	251201406	8.56	250321505	8.53	3125371	3040905
		GG	208163470	7.09	206857602	7.05	3237868	3040058
		GT	145702091	4.96	144792343	4.93	3173707	3056002
		TA	123416231	4.20	123621795	4.21	3168380	3199242
		TC	163119599	5.56	163571150	5.57	3084906	3130968
		TG	194994216	6.64	201192509	6.85	2728120	3453974
		TT	184127004	6.27	184627103	6.29	3147874	3220952
		Total	2935725934	100	2935725934	100		
	Informations							
	Number of cds sequences	6139429						
	Number of invalid cds	2708						

Plasmid | Sum_DNA

L'information pour chaque "Organism" (nombre de génomes, trinucléotides, CDS valides, etc.) est sommé dans l'onglet "Sum".
Remarque: le nombre de génome est égal à 1 pour un "Organism" mais cette information deviendra intéressante dans les nœuds parents.

2. Acquisition des gènes
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
//
```

Il faut impérativement réaliser le join avant de tester la validité du gène.

```
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 gènes

Les gènes doivent commencer par un trinucléotide initiation

$$t_{init} = \{\mathrm{ATG}, \mathrm{CTG}, \mathrm{TTG}, \mathrm{GTG}, \mathrm{ATA}, \mathrm{ATC}, \mathrm{ATT}, \mathrm{TTA}\}$$

et se terminer par un trinucléotide stop:

$$t_{stop} = \{\mathrm{TAA}, \mathrm{TAG}, \mathrm{TGA}, \mathrm{TTA}\}.$$

Les gènes ne doivent comporter que des lettres sur $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

La longueur des gènes doit être un multiple de 3. Dans le cas d'un join() (ou complement(join())), la longueur des gènes doit être un multiple de 3 après la concaténation des séquences.

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 un gène ne vérifie pas un des tests précédents, il est éliminé de l'analyse statistique.

IMPORTANT: On garde la séquence des gènes de ATG…TAA.

3. Analyse statistique 1: Nombres et fréquences des trinucléotides dans les 3 phases des gènes

3.1. Nombres de trinucléotides dans les 3 phases des gènes

Il existe 64 trinucléotides $\{\mathrm{AAA}, \mathrm{AAC}, \ldots, \mathrm{TTT}\}$.

Comptage des trinucléotides:
- en phase 0 des gènes: [ATG,NNN,…,NNN]TAA (sauf les trinucléotides stop)
- en phase 1 des gènes: A[TGN,NNN,…,NNT]AA
- en phase 2 des gènes: AT[GNN,NNN,…,NTA]A

IMPORTANT: Les mêmes nombres de trinucléotides sont comptés en phases 0, 1 et 2.

Remarque: En phase 0, les trinucléotides stop $\{\mathrm{TAA}, \mathrm{TAG}, \mathrm{TGA}\}$ ont une valeur proche de 0, contrairement aux phases 1 et 2. Il existe quelques trinucléotides stop aux milieux des gènes qui dans ce cas doivent être considérés dans l'analyse statistique.

3.2. Fréquences des trinucléotides dans les 3 phases des gènes

Pour une phase donnée, la fréquence d'un trinucléotide t est égale au nombre d'occurrence de t divisé par le nombre total de trinucléotides.

3.3. Présentation des résultats

IMPORTANT: Les nombres et fréquences d'occurrence des trinucléotides dans les fichiers Excel doivent être des nombres et non des chaînes de caractères.

L'affichage des fréquences des trinucléotides se fera sur 2 décimales mais gardera les valeurs avec toutes les décimales.

Le nombre de génomes, trinucléotides, le nombre de CDS valides (vérifiant les tests), le nombres de CDS "invalides" et diverses informations sur le génome doivent être donnés.

4. Analyse statistique 2: Nombres de trinucléotides en phase préférentielle dans les 3 phases des gènes

Let G be a genome. Let $Pr_f(t, g)$ be the occurrence frequency of a trinucleotide $t \in A_4^3$ in a frame $f \in \{0,1,2\}$ of a gene g belonging to G. Thus, there are $3 \times 64 = 192$ trinucleotide occurrence frequencies $Pr_f(t, g)$ in the three frames f of a gene g. Then, the preferential frame $F(t, g) \in \{0,1,2\}$ of a trinucleotide $t \in A_4^3$ in a gene g is defined by the frame having the maximal occurrence frequency $Pr_f(t, g)$ among the three frames $f \in \{0,1,2\}$ of g

$$F(t,g) = \arg\max_{f\in\{0,1,2\}} Pr_f(t,g). \quad (1)$$

At the gene level, particularly for genes g of small lengths, a trinucleotide t may have an identical occurrence frequency $Pr_f(t,g)$ in two or three frames f. In this case, two or three preferential frames are assigned to the trinucleotide. There is no preferential frame if there is no trinucleotide t in g.

Let the indicator function $\delta_f(F(t,g)) \in \{0,1\}$ be equal to 1 if the preferential frame $F(t,g) \in \{0,1,2\}$ of a trinucleotide $t \in A_4^3$ is equal to a given frame $f \in \{0,1,2\}$ of a gene g, 0 otherwise

$$\delta_f\big(F(t,g)\big) = \begin{cases} 1 & \text{if } F(t,g) = f \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $F(t,g)$ is defined in Equation (2).
The number $Nb_f(t,G) \in \mathbb{N}$ of preferential frames of a trinucleotide $t \in A_4^3$ for each frame $f \in \{0,1,2\}$ in a genome G is simply defined by

$$Nb_f(t,G) = \sum_{g\in G} \delta_f(F(t,g)) \quad (3)$$

where $\delta_f(F(t,g))$ is defined in Equation (3).

IMPORTANT: Si un trinucléotide t n'existe pas dans g (cas de gènes g de très petite longueur), il n'existe pas de phase préférentielle: $\delta_f\big(F(t,g)\big) = 0$ pour les 3 phases f.

Dans chaque fichier Excel, trois colonnes sont ajoutées pour les nombres $Nb_0(t,G)$ en phase 0, $Nb_1(t,G)$ en phase 1 et $Nb_2(t,G)$ en phase 2 pour chaque trinucléotide t.

5. Analyse statistique 3: Nombres et fréquences des dinucléotides dans les 2 phases des gènes

Il existe 16 dinucléotides $\{\mathrm{AA}, \mathrm{AC}, \ldots, \mathrm{TT}\}$.
Si le gène se terminant par un trinucléotide stop, e.g. TAA, a pour longueur un nombre modulo 3 impair alors:
- la lecture de la phase 0 se termine avant TAA, donc avant le TAA: [AT,GN,…,NN]TAA
- la lecture de la phase 1 se termine avant AA.: A[TG,NN,…,NT]AA

Si le gène se terminant par un trinucléotide stop, e.g. TAA, a pour longueur un nombre modulo 3 pair alors:
- la lecture de la phase 0 se termine avant NTAA: [AT,GN,…,NN]NTAA
- la lecture de la phase 1 se termine avant TAA, donc avant le TAA: A[TG,NN,…,NN]TAA

IMPORTANT: Les mêmes nombres de dinucléotides sont comptés en phases 0 et 1.

6. Analyse statistique 4: Nombres de dinucléotides en phase préférentielle dans les 2 phases des gènes

Principe similaire à l'analyse statistique 2.

7. Programmation en Java

Structurer votre programme de façon qu'aux exécutions suivantes le calcul statistique des trinucléotides (fichiers Excel) reprend l'arborescence locale et ne porte que sur les données génomiques modifiées (nouvelles ou modifiés).
IMPORTANT: Le calcul des fichiers Excel associés aux nœuds parents doit être recalculé en totalité, même si des nœuds parents ne sont pas modifiés, par simplification et pour éviter des erreurs de programmation.
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.

Gérer les statistiques générales (nombre de génomes, nombre de gènes, nombre de trinucléotides) et les statistiques d'erreur associées à chaque fichier Excel.
Interface graphique:
- montrer l'arborescence locale.
- donner une barre d'avancement du programme et un calcul estimé du temps (IMPERATIF).

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).

Les points importants du programme doivent être commentés, en particulier les balises web doivent être mentionnées dans les accès aux différentes pages web.

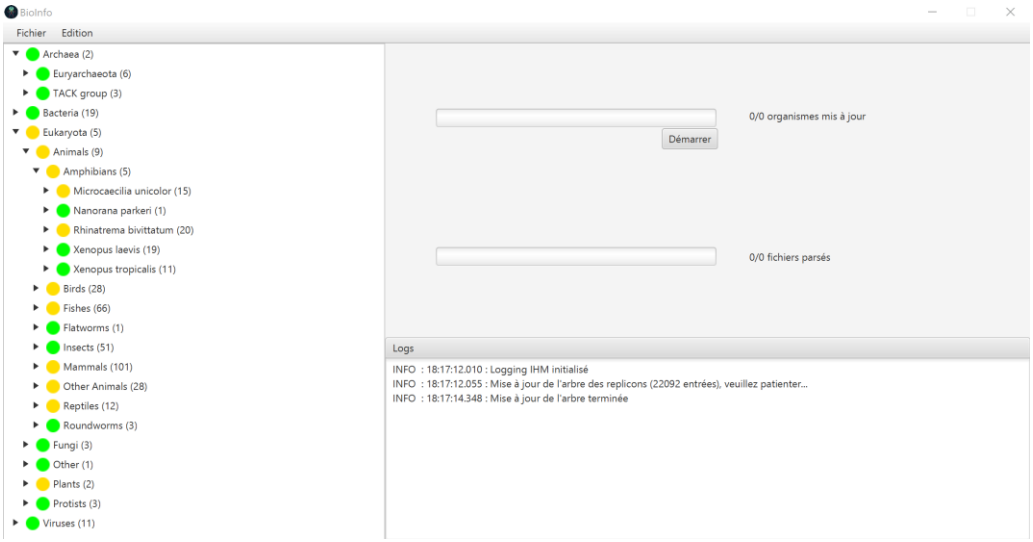
8. Modalités du projet

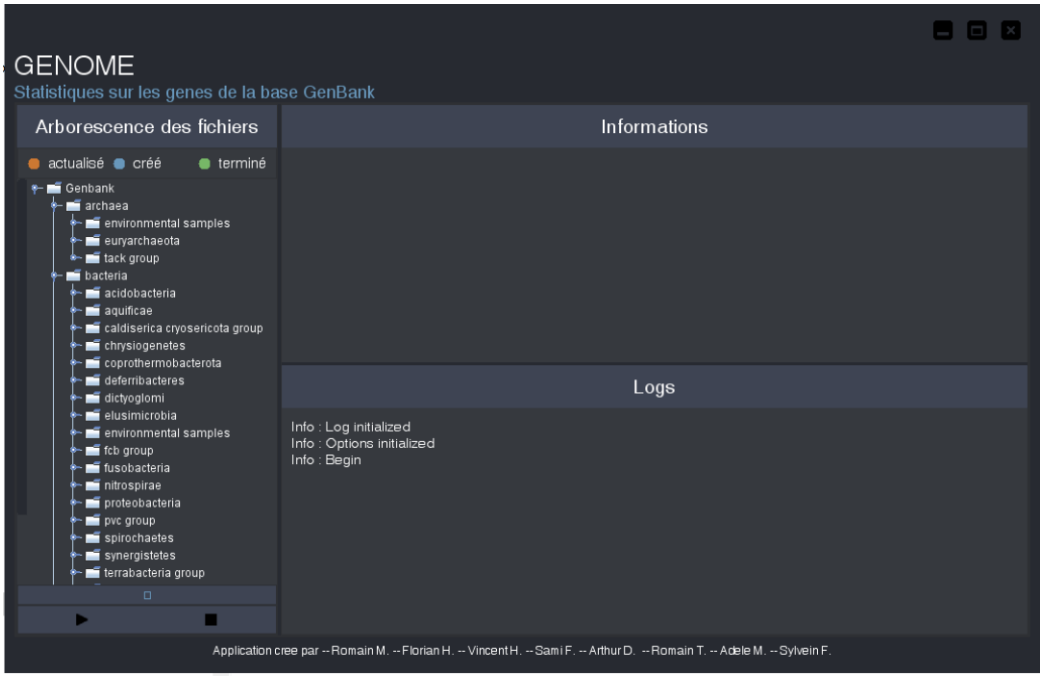
(i) Renvoyer par email à l'adresse c.michel@unistra.fr ou mettre sur un site de téléchargement:
- un dossier "Eclipse" pour les programmes sources pour Eclipse;
- un dossier "Jar" un fichier jar exécutable en double cliquant;
- un dossier "Results" (cf. l'arborescence Results\Kingdom\Group\SubGroup\Organism.xlsx).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

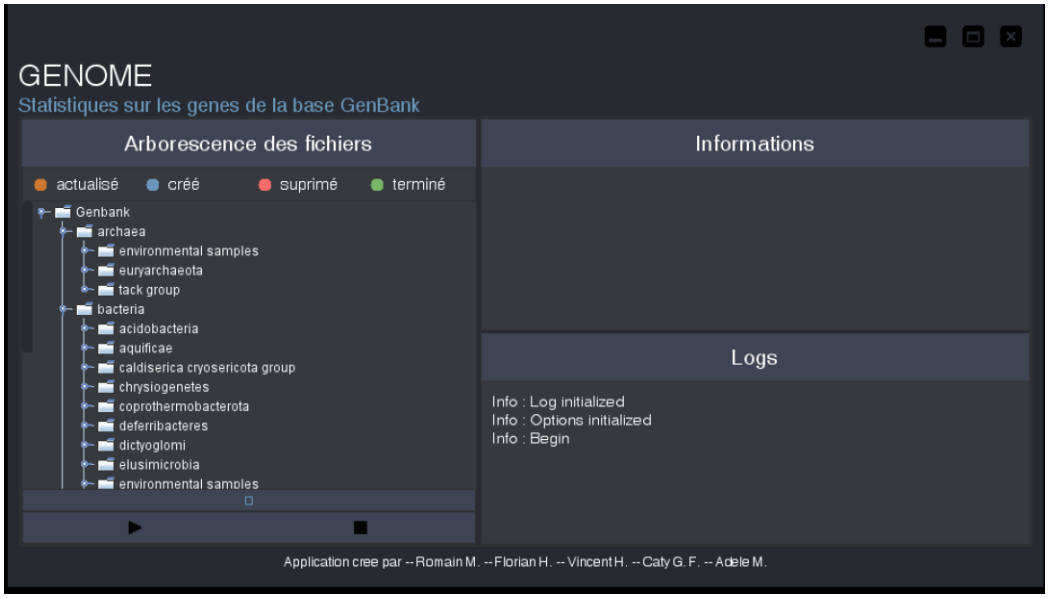
IMPORTANT ET IMPERATIF:
Le dossier "Results" doit également être un sous-répertoire du dossier "Eclipse"
Le dossier "Results" doit être également un sous-répertoire du dossier "Jar"

(ii) 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.

9. Exemples de logiciels







10. 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.