Projet HMSN204

BLAISON, LAIDLAW

Jeudi, 11 avril 2019

Introduction

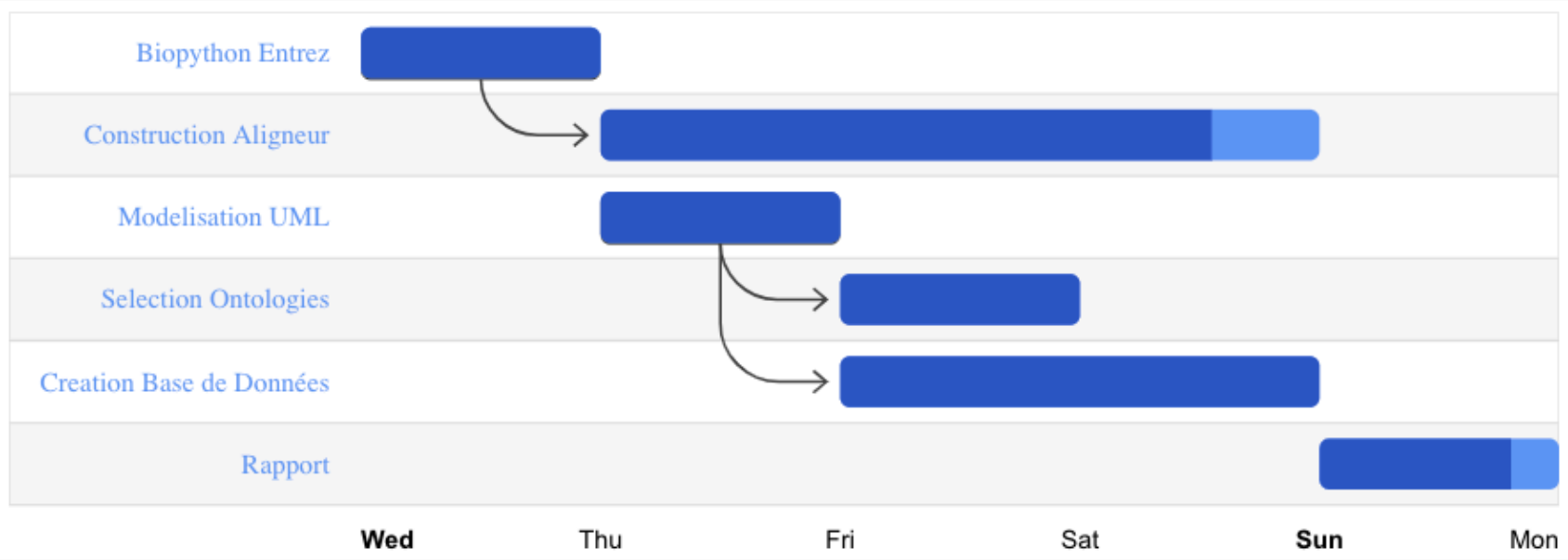
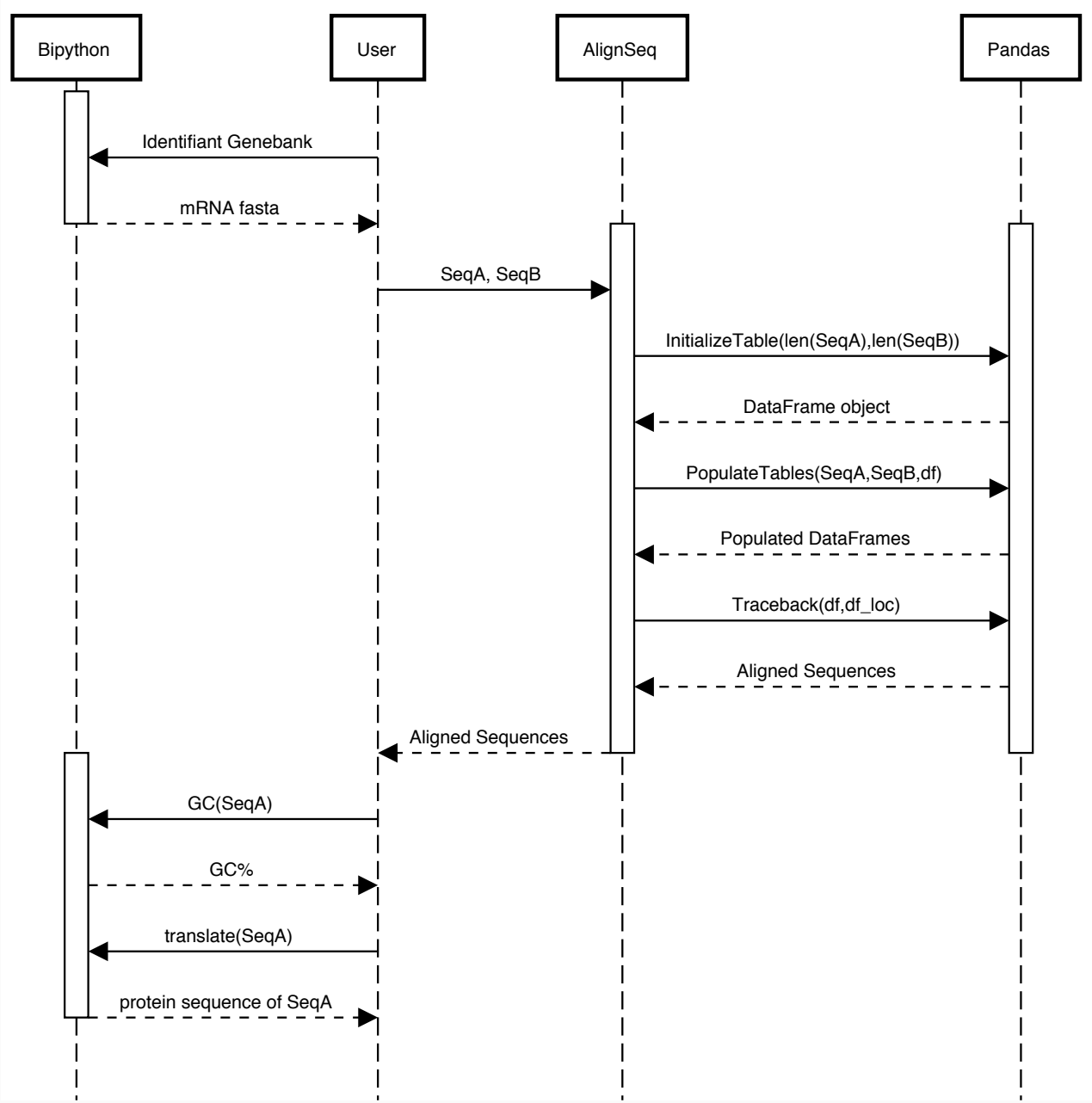


Fig. 1: Gantt Diagram

Vue d'Ensemble



Recuperation des sequences

- NCBI detient 3 sequences RNAm pour SEX1, et 1 sequence cDNA genomique
  - Les 3 ne sont pas des mutants, mais des variants d'epissage alternative variants

Recuperation des sequences

- les *SNP variants* décrit dans dbSNP servait pour creer le fasta de SEX1 mutant.

Region	Chr. position	mRNA pos	dbSNP rs# cluster id	Hetero-zygosity	Validation	MAF	Allele origin	3D	Clinically Associated	Clinical Significance	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos	PubMed
	3581249	4621	rs1106309714	N.D.							missense	-	Glu [E]	3	1387	
											contig reference	C	Asp [D]	3	1387	
	3581250	4620	rs1095915538	N.D.							missense	-	Val [V]	2	1387	
											contig reference	A	Asp [D]	2	1387	
	3581273	4597	rs1102763773	N.D.							synonymous	-	Gln [Q]	3	1379	
											contig reference	G	Gln [Q]	3	1379	
	3581278	4592	rs1104980834	N.D.							missense	-	Thr [T]	1	1378	
											contig reference	G	Ala [A]	1	1378	
	3581332	4538	rs346781335	N.D.							missense	-	Ile [I]	1	1360	
											contig reference	G	Val [V]	1	1360	
	3581350	4520	rs347439271	N.D.							missense	-	Val [V]	1	1354	
											contig reference	T	Leu [L]	1	1354	
	3581370	4500	rs1104490444	N.D.							missense	-	Lys [K]	2	1347	
											contig reference	C	Thr [T]	2	1347	

Fig. 3: dbSNP page for geneID :837619

Recuperation des sequences

Used dbSNP IDs			
rs1105066589	rs1103971843	rs1095089377	rs1100808719
rs1095780989	rs1095659046	rs1097407346	rs1097236347
rs1105152302	rs1097124183	rs347038182	rs346885812
rs1101762250	rs1100942745	rs1106840358	rs346897346
rs1099291378	rs1102995172	rs1096948278	rs1099609436
rs1104510198			

Instantiation du Tableau

Comment faire le tableau ?

- Pour creer un tableau 2D en code, on a utilisé les `DataFrame` de la librarie pandas.

Pourquoi Pandas ?

- Permet l'acces et l'enregistrement de données par coordonées de cellule (permettant ecriture automatique)
- permet visualisation de tout le tableau

Remplissage du Tableau

- Tableau des scores d'alignements selon le standard du Needleman–Wunsch

		G	C	A	T	G	A	C	T	A	A	C	G	T	A	A
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15
G	-1	1	0	-1	-2	-1	-2	-3	-4	-5	-6	-7	-6	-7	-8	-9
C	-2	0	2	1	0	-1	-2	-1	-2	-3	-4	-3	-4	-5	-6	-7
C	-3	-1	3	2	1	0	-1	0	-1	-2	-3	-2	-3	-4	-5	-6
T	-4	-2	2	2	3	2	1	0	1	0	-1	-2	-3	-2	-3	-4
T	-5	-3	1	1	4	3	2	1	2	1	0	-1	-2	-1	-2	-3
A	-6	-4	0	2	3	3	4	3	2	3	4	3	2	1	2	3
C	-7	-5	1	1	2	2	3	5	4	3	3	5	4	3	2	2
T	-8	-6	0	0	3	2	2	4	6	5	4	4	4	5	4	3
A	-9	-7	-1	1	2	2	3	3	5	7	8	7	6	5	6	7
A	-10	-8	-2	2	1	1	4	3	4	8	9	8	7	6	7	8
A	-11	-9	-3	3	2	1	5	4	3	9	10	9	8	7	8	9
T	-12	-10	-4	2	4	3	4	4	5	8	9	9	8	9	8	8
C	-13	-11	-3	1	3	3	3	5	4	7	8	10	9	8	8	7
G	-14	-10	-4	0	2	4	3	4	4	6	7	9	11	10	9	8
T	-15	-11	-5	-1	3	3	3	3	5	5	6	8	10	12	11	10
A	-16	-12	-6	0	2	2	4	3	4	6	7	7	9	11	13	14
A	-17	-13	-7	1	1	1	5	4	3	7	8	7	8	10	14	15

Remplissage du Tableau

- Dans un autre tableau le direction qui a amené au score a été enregistré
- permet à l'algorithme d'ensuite effectuer le backtracking

		G	C	A	T	G	A	C	T	A	A	A	A
	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
G	NaN	\	<	<	<	<	<	<	<	<	<	<	<
C	NaN	|	\	<	<	<	\<	\<	<	<	<	<	<
C	NaN	|	|	<	<	<	<	|<	<	<	<	<	<
T	NaN	|	|	\	\<	<	<	<	\<	<	<	<	<
T	NaN	|	|	|\	|	<	<	<	|<	<	<	<	<
A	NaN	|	|	|\	|	\	\<	<	<	\<	<	<	<
C	NaN	|	|	|	|	|\	|	\	<	<	|	|	|
T	NaN	|	|	|\	|	<	|	|	\	<	<	|<	|
A	NaN	|	|	|\	|	\	|\<	|	|	\	<	<	<
A	NaN	|	|	|	|<	|\	|	<	|	|	|<	|<	|<
G	NaN	|	|	|	\	|\<	|	\	|	|	|	|	|
T	NaN	|	|	|	|\	|<	|	|\	|\	|	|	|	|
A	NaN	|	|	|	|	\	|	<	|	|	|<	|<	|<
A	NaN	|	|	|	<	|\<	|	<	|<	|	|<	|<	|<

Backtracking du Tableau

- la sortie du fonction est deux listes contenant le version aligné des deux sequences.

SeqA = [‘G’, ‘-’, ‘A’, ‘T’, ‘T’, ‘A’, ‘C’, ‘A’, ‘A’]

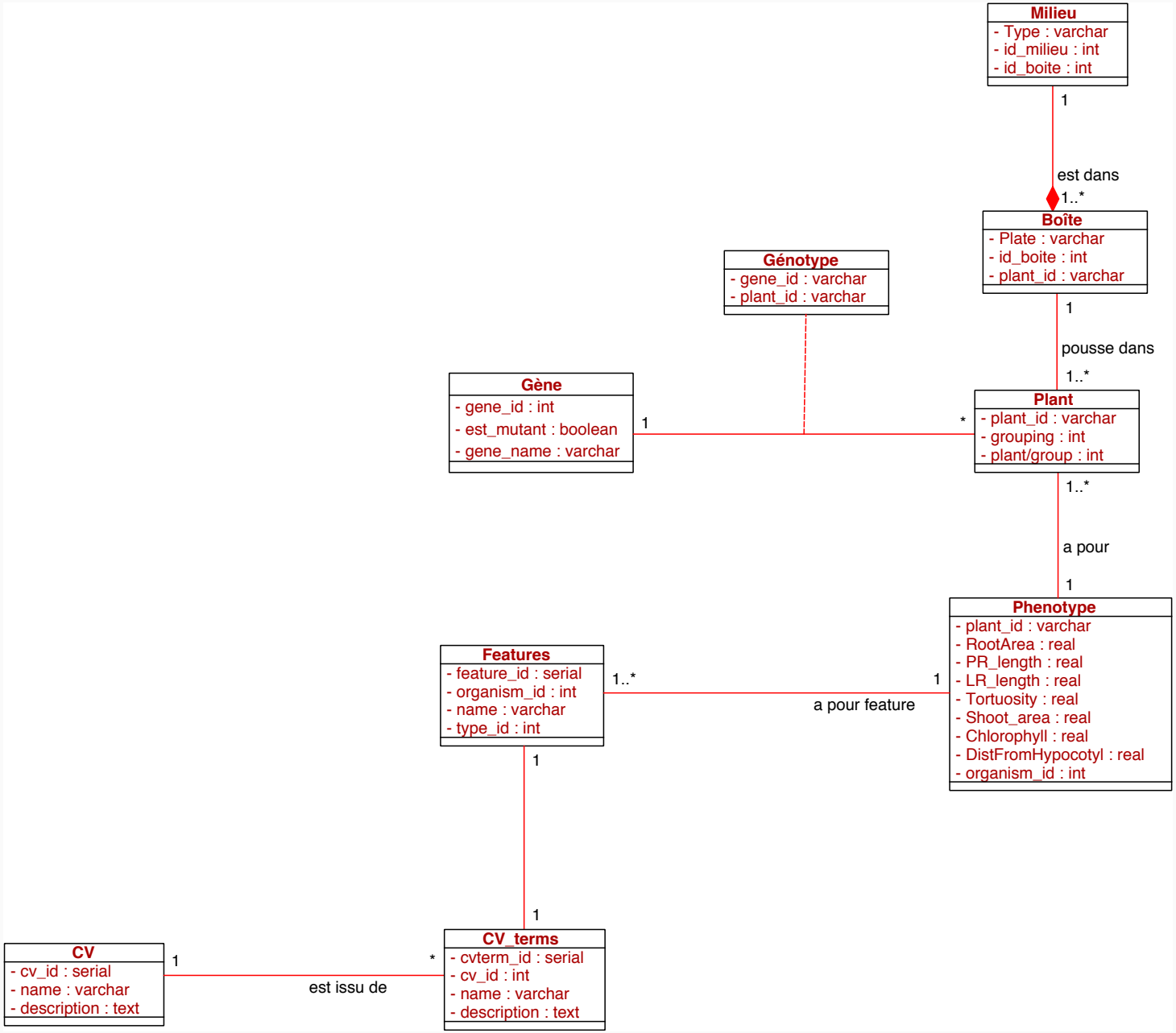
SeqB = [‘G’, ‘C’, ‘A’, ‘T’, ‘-’, ‘G’, ‘C’, ‘-’, ‘U’]

Variations = $\{1, 4, 5, 7, 8\}$

Problèmes d'Alignement

Optimisation	Alignment Time (s)
None	18.6
Function Caching	18.16
Cython	18.02
JIT Compilation	18.22
Combined	18.55

Schema de la base de données



Tab. 3: Table features

feature_id	organism_id	name	uniquename	type_id
1	1	PR_lenght	APO_000001	1
2	1	LR_lenght	APO_000002	2
3	1	ShootArea	APO_000003	3
4	1	RootArea	APO_000004	4
5	1	DistanceFromHypocotyl	APO_000005	5
6	1	Chlorophylle	TO_0000495	6
7	1	Tortuosity	APO_000006	7

Tab. 4: Table cvterm

cvterm_id	cv_id	name	definition
1	5	Root Lenght	a root lenght (FLOPO_0009325) which is part of a primary root (PO_0020127
2	5	Root Lenght	a root lenght (FLOPO_0009325) which is part of a lateral root (PO_0020121
3	5	area	an area (PATO_0001323 which is part of the shoot axis (PO_0025029)
4	5	area	an area (PATO_0001323) which is part of the root (PO_0009005)
5	5	distance	a distance(PATO_000040) between roots (PO_0009005) and hypocotyl (PO_0020100)
6	4	chlorophyll contents	Measures the chlorophyll content in a green tissue. Includes both chlorophyll-a and chlorophyll-b. Chlorophyll is the green pigment found in plants.
7	5	curvature	a curvature (PATO_0001591) which is part of roots (PO_0009005)

Requêtes

```
select m.type, avg(ph.Root_area)
from milieu m, boite b, plant p, phenotype ph
where m.id_boite = b.id_boite and b.plant_id = \
p.plant_id and p.plant_id = ph.plant_id
group by m.type ;
```

Tab. 5: Résultat de la requête :

type	avg
Milieu_5	0.127853366556921
Milieu_1	0.179657717053341
Milieu_2	0.309144144189787
Milieu_3	0.151981368513419
Milieu_4	0.302773334148029