Linking heterogeneous data from model plant species in a graph database

Confais J, Francillonne N

Agde, 2023-10-11

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

Présentateurs

Johann Confais & Nicolas Francillonne

Unité qui porte le projet : URGI

2 thématiques centrales : système d'information & analyse des génomes

⇒ Opportunité 1 : valoriser les données produites dans l'unité (annotation des génomes) et les liant avec d'autres données d'intérêt.

⇒ Opportunité 2 : visualisation des interactions

⇒ Opportunité 3 : Explorer les approches graphe de connaissance

Domaine d'application du jeu de données

Recherche : régulation des gènes/réseaux de gènes

Application : amélioration variétale

Application : l'amélioration variétale qui doit répondre à de nouveaux enjeux comme le réchauffement climatique et la transition agro-écologique

Question : Quels sont les éléments régulateurs en amont de gènes impliqués dans la réponse à un stress ?

Potentiel de réutilisation

Biologie translationnelle

Le nœud gène est un pivot qui permet des liaisons avec d'autres ensembles de données agrégés autour du gène.

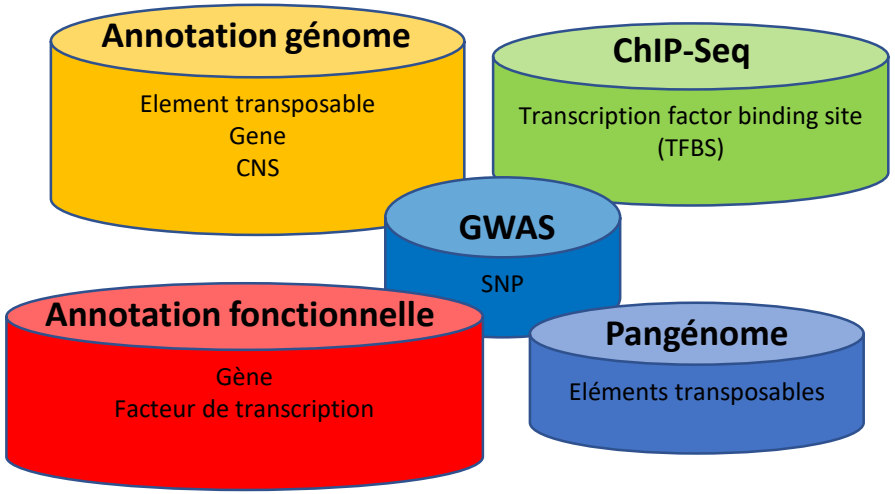
A travers les relation orthologues ont peut créer des ponts inter espèces

Exemple :

Import dans notre base de données issues d'AgroLD (ou autre base de données en rdf, ttl)

Description générale du contenu du jeu de données

Des données hétérogènes, des bases indépendantes et pas de liens



ReMap2022

tair

JASPAR2022

PlantRegMap

JGI Phytozome 12
THE PLANT GENOMICS RESOURCE

TFBS & TF : PlantRegMap Db (http://planttfdb.cbi.pku.edu.cn/download.php) & Heyndrix et al 2014 (Plant Cell):
TAIR V10 repository : https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_gff3
GWAS: Nordborg study (Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines; Atwell et al. - Nature 2010)
REPETDB https://urgi.versailles.inrae.fr/repetdb
CNS : Van de Velde et al 2014 (Plant Cell)

=> Comment articuler ces données pour répondre à notre question ?

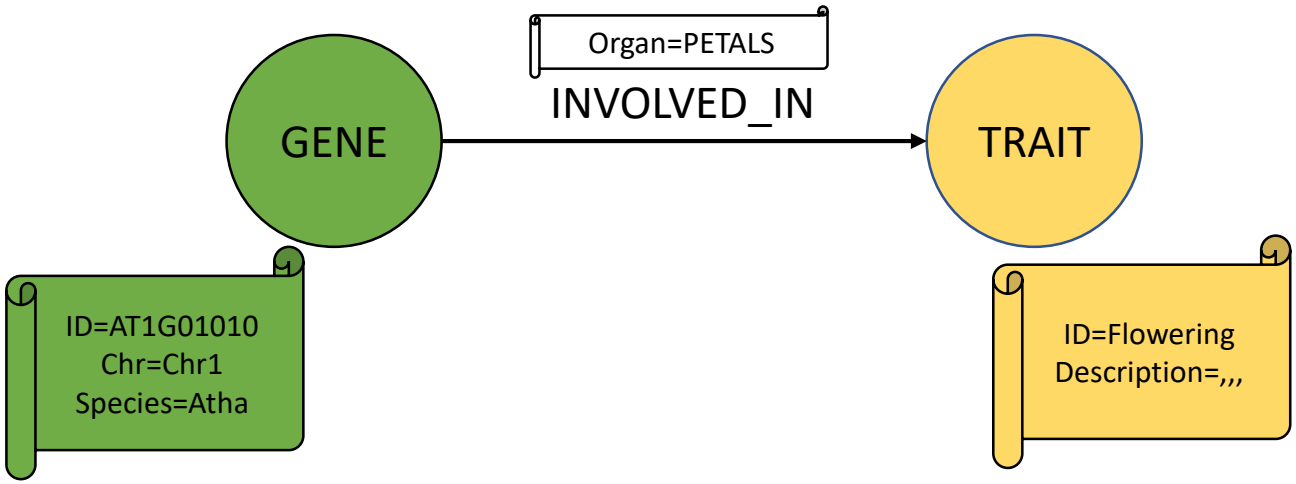
Méthodologie de construction du jeu de données

(Entité)-[relation]-(Entité)

Nœud

Relation

Propriété



Processus d'intégration

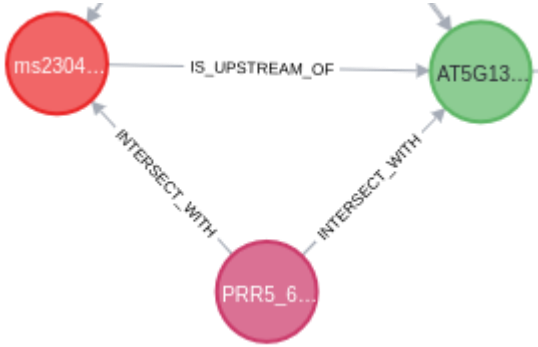
Des données et des formats très hétérogènes

données d'entrée

scripts *ad hoc*.

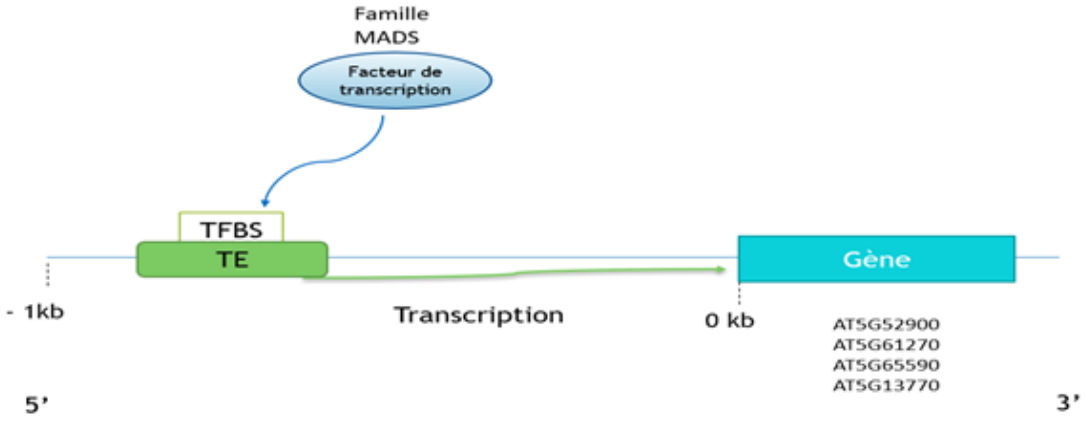
fichiers prêts à être intégrés

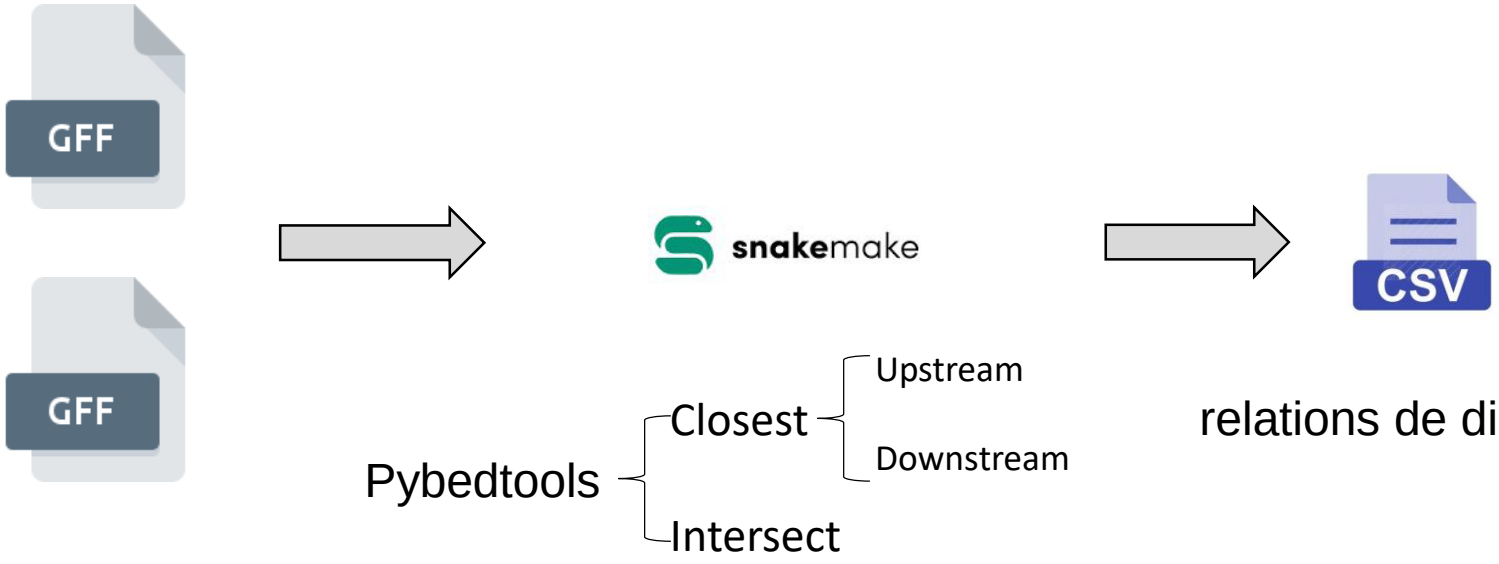
Intégration via Neo4J



Cas de relations de distance

Données d'annotation fichiers .gff





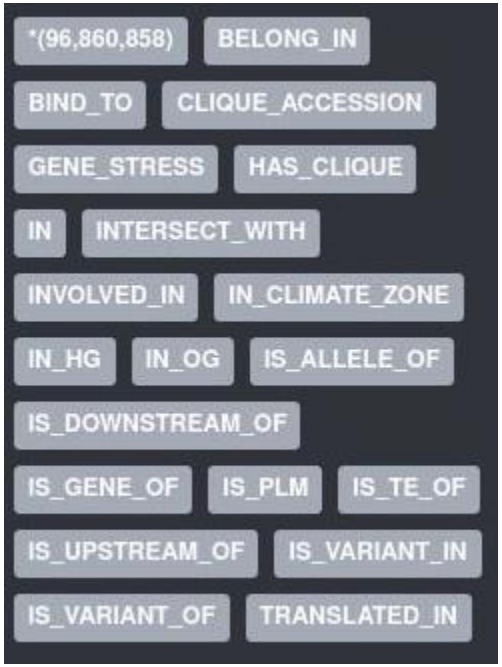
Volumétrie

40 millions

Nœuds

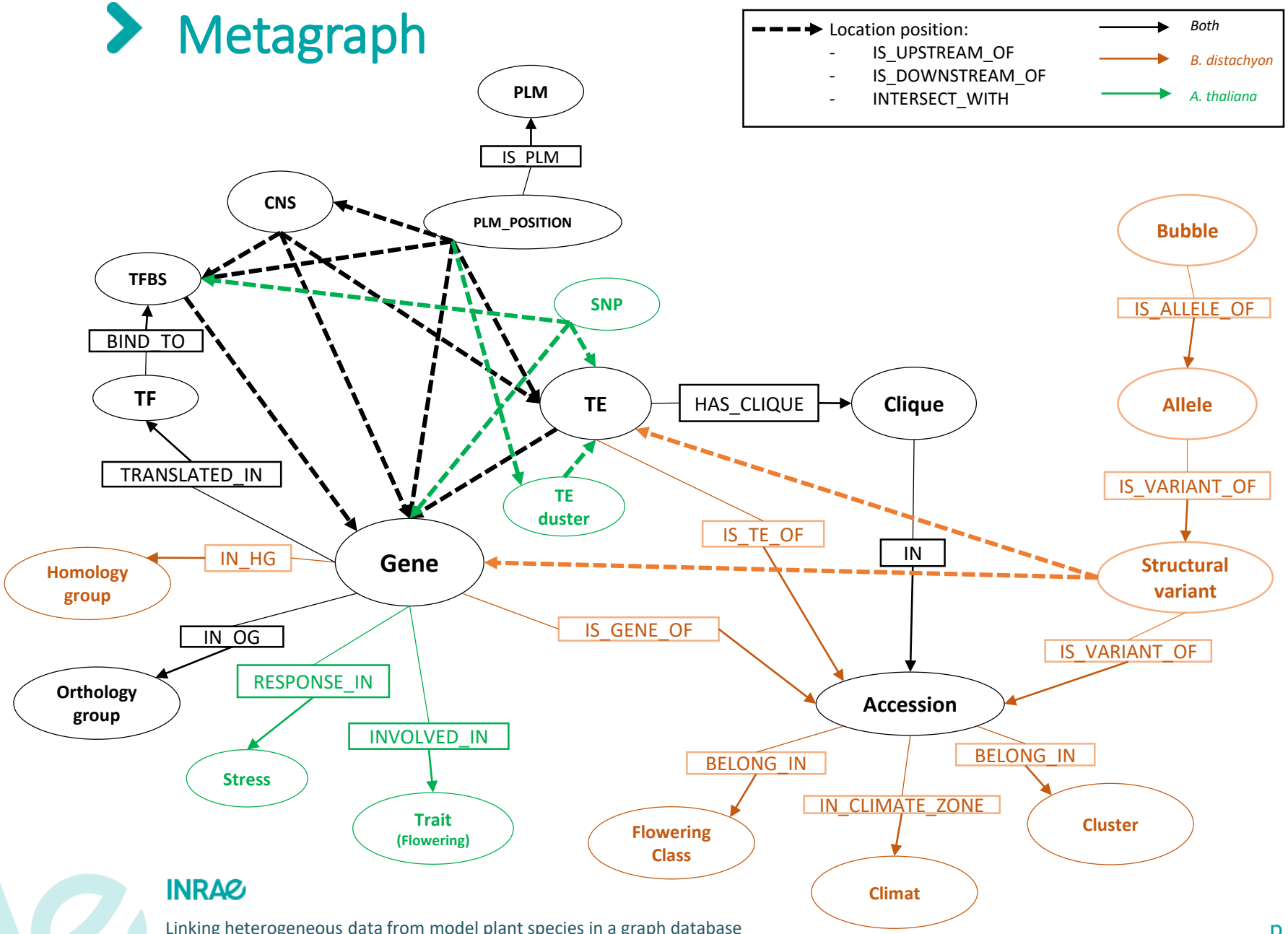
100 millions

Relation



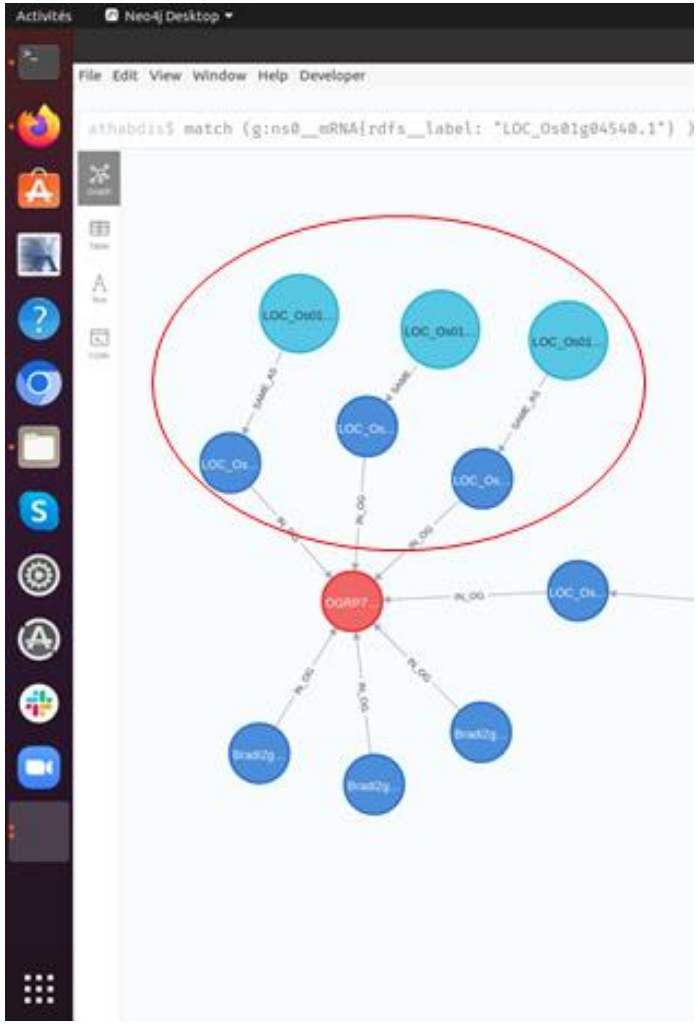
Quelques remarques !
- Plus d'une centaine de propriétés décrivant les nœuds ou relations
- 54 accessions de B.distachyon avec 54 annotations en TE et en gènes

Metagraph


Location position:
- IS_UPSTREAM_OF
- IS_DOWNSTREAM_OF
- INTERSECT_WITH
Both
B. distachyon
A. thaliana
PLM
IS_PLM
CNS
PLM_POSITION
TFBS
SNP
BIND_TO
TF
TE
HAS_CLIQUE
Clique
TRANSLATED_IN
TE duster
Bubble
IS_ALLELE_OF
Allele
IS_VARIANT_OF
Structural variant
IS_TE_OF
IN
IN_HG
Homology group
Gene
IS_GENE_OF
IS_VARIANT_OF
IN_OG
Orthology group
RESPONSE_IN
Accession
INVOLVED_IN
Stress
Trait (Flowering)
BELONG_IN
BELONG_IN
IN_CLIMATE_ZONE
Flowering Class
Cluster
Climat

Interopérabilité de la base

Exemple de l'import de données issues d'AgroLD



- Existant :

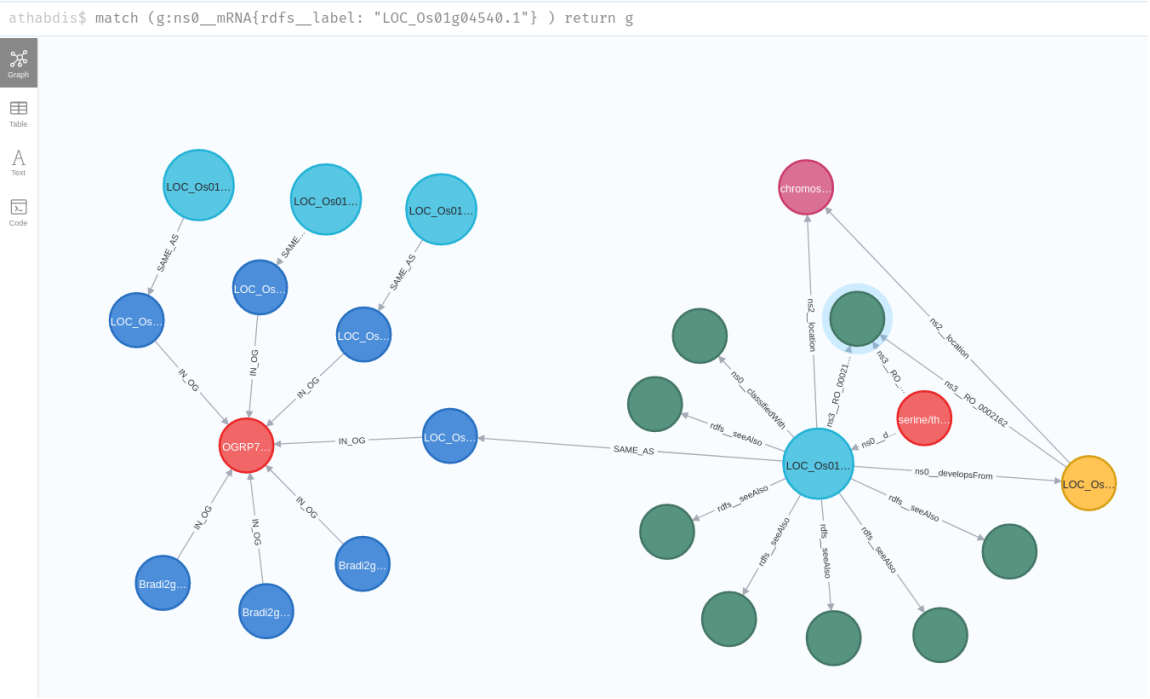
Nœud « orthogroup » (rouge) qui lient les gènes orthologues des différentes espèces en base (Bdis, Atha, Osat)

- À créer :

Sur la base des ID gènes du riz créer un lien entre notre base et celle d'AgroLD (relation nœud bleu foncé-bleu clair)

Interopérabilité de la base

Exemple de l'import de données issues d'AgroLD



Points positifs de l'import

- Cout d'import très faible
- Masse de données facilement accessible

Limite de l'import :

- Modélisation des données différentes.
- Requêtage à adapter ou faire des opérations pour harmoniser les modèles

Import d'ontologie

Cas de la GO (gene ontology)

On peut importer des ontologies et les lier à nos jeux de données à partir des identifiant.

Par exemple les ID gène peuvent avoir les les ID GO en propriété et on peut faire l'import comme pour AgroLD

Moyen d'accès au jeu de données

- Projet pas encore ouvert au grand public
- Publication en cours (mise à disposition des données dans un dépôt public)
- Partenariat ouvert à toutes suggestions

Au sein du groupe de travail « Graph » du CATI GREP : 3 projets co-existent

⇒ Modélisation compatible

⇒ Même technologie de visualisation : import/export aisé (dump, création de sous graph, import des fichiers sources)

Contraintes liées à l'utilisation

- On a fait le choix d'utiliser que des données publiques.
- Nécessite de connaitre le modèle de données

=> Guide d'utilisation dans l'interface :

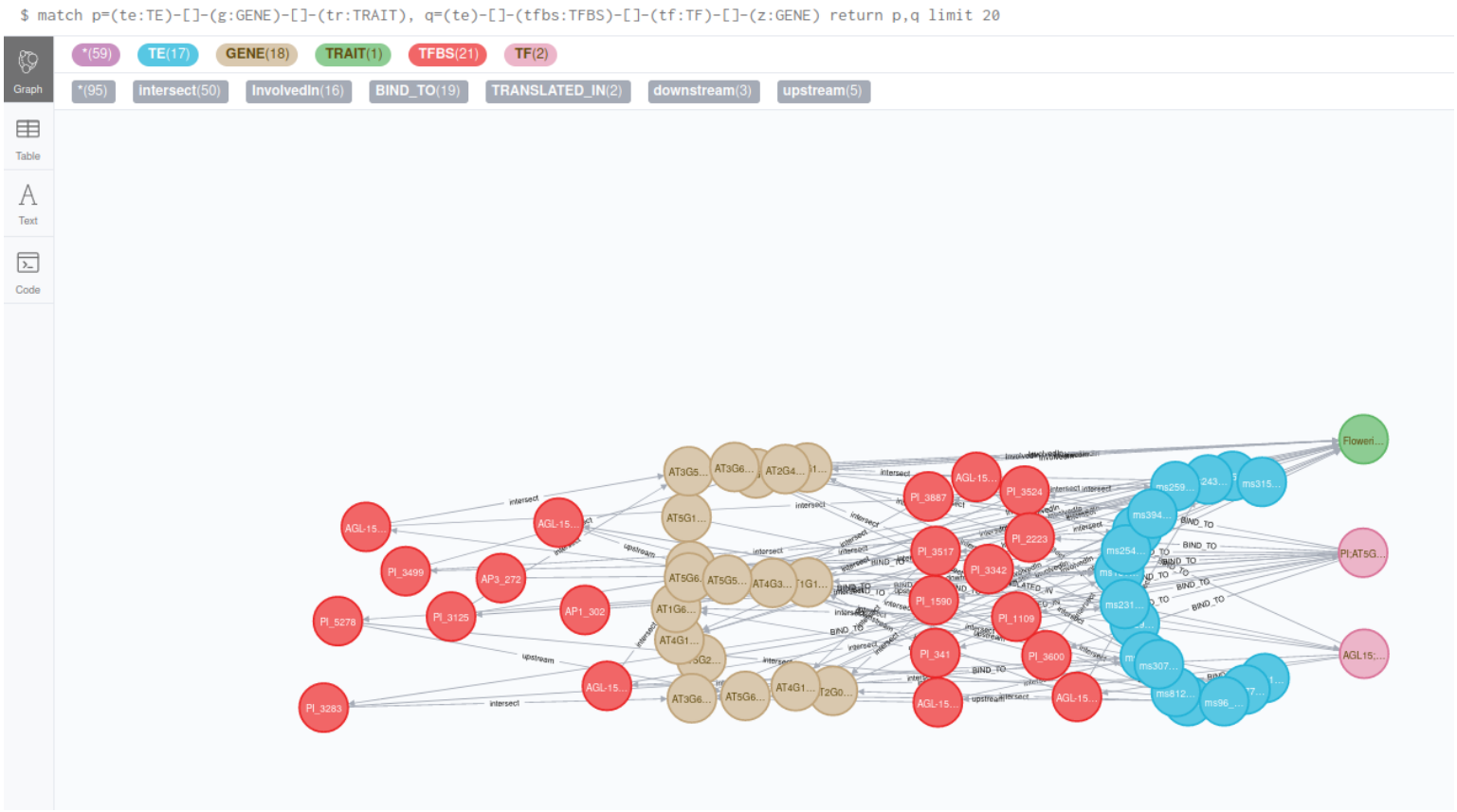
- Requêtes à trou
- Requêtes exemple
- Descriptif des données

Contraintes liées à la ré-utilisation

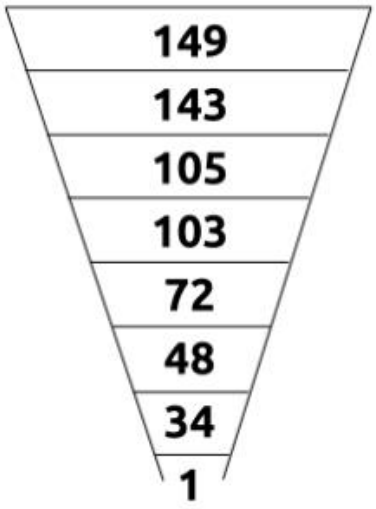
System rdf / ttl compliant

Export en rdf qui ne prend pas en compte les propriétés sur les relations.

Questions ?


$ match p=(te:TE)-[]-(g:GENE)-[]-(tr:TRAIT), q=(te)-[]-(tfbs:TFBS)-[]-(tf:TF)-[]-(z:GENE) return p,q limit 20
Graph
Table
Text
Code
*(59)
TE(17)
GENE(18)
TRAIT(1)
TFBS(21)
TF(2)
*(95)
intersect(50)
involvedIn(16)
BIND_TO(19)
TRANSLATED_IN(2)
downstream(3)
upstream(5)

1er cas: Quelles seraient les séquences régulatrices ancestrales impliquées dans la floraison



Conditions cumulées

Gènes d'*A.thaliana* impliqués dans la floraison

\+ CNS en amont des gènes d'*A.thaliana*

\+ CNS qui chevauche un TFBS validé par analyse ChIP-Seq

\+ ChIP-Seq TFBS se liant avec un TF

\+ Gène d'*A.thaliana* orthologue à un gène de *B.distachyon*

\+ CNS en amont des gènes de *B.distachyon*

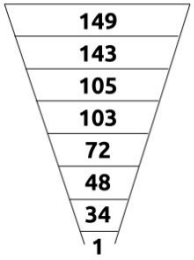
+CNS de *B.distachyon* chevauche un TFBS prédit se liant à un TF

+TF d'*A.thaliana* et de *B.distachyon* de la la même famille

Puis affichage des TFBS prédits d'*A.thaliana* et des groupes d'homologie et d'orthologie de *B.distachyon*

Use case

Quelles seraient les séquences régulatrices ancestrales impliquées dans la floraison ?



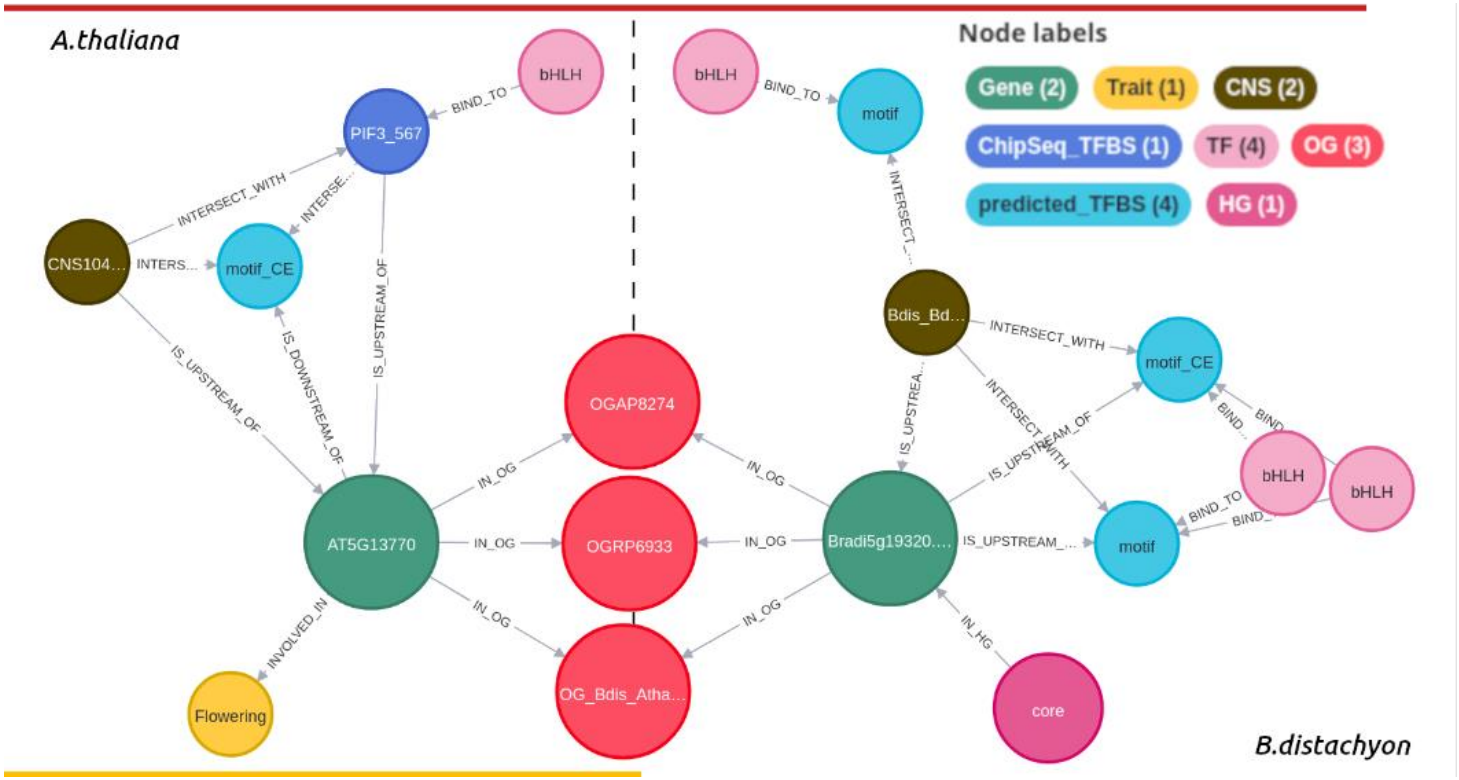
Conditions cumulées

Gènes d'*A.thaliana* impliqués dans la floraison
+ CNS en amont des gènes d'*A.thaliana*
+ CNS qui chevauche un TFBS validé par analyse ChIP-Seq
+ ChIP-Seq TFBS se liant avec un TF
+ Gène d'*A.thaliana* orthologue à un gène de *B.distachyon*
+ CNS en amont des gènes de *B.distachyon*
+CNS de *B.distachyon* chevauche un TFBS prédit se liant à un TF
+TF d'*A.thaliana* et de *B.distachyon* de la la même famille

Puis affichage des TFBS prédits d'*A.thaliana* et des groupes d'homologie et d'orthologie de *B.distachyon*

```
MATCH a=(g1:Gene{specie:"Atha"})--(:Trait{TRAIT_NAME:"Flowering"}),
b=(g1)-[:IS_UPSTREAM_OF]-(cns1:CNS),
c=(cns1)-[:INTERSECT_WITH]-(tfbs1:ChipSeq_TFBS),
d=(tfbs1)--(tf1:TF),
e=(tfbs1)-[:INTERSECT_WITH]-(pTFBS1:predicted_TFBS)-[:INTERSECT_WITH]-(cns1),
f=(pTFBS1)-[:IS_DOWNSTREAM_OF]-(g1),
g=(g1)--(:OG{name_orthogroup:"OG_Bdis_Atha"})--(g2:Gene{specie:"Bdis"}),
h=(g2)-[:IS_UPSTREAM_OF]-(cns2:CNS),
i=(cns2)-[:INTERSECT_WITH]-(tfbs2:predicted_TFBS),
j=(tfbs2)-[:BIND_TO]-(tf2:TF), k=(g2)--(:HG), l=(g1)--(:OG)--(g2)
WHERE tf1.family=tf2.family
RETURN a,b,c,d,e,f,g,h,i,j,k,l
```

Use case


A.thaliana
B.distachyon
Node labels
Gene (2)
Trait (1)
CNS (2)
ChipSeq_TFBS (1)
TF (4)
OG (3)
predicted_TFBS (4)
HG (1)
bHLH
PIF3_567
BIND_TO
INTERSECT_WITH
CNS104...
motif_CE
IS_UPSTREAM_OF
IS_DOWNSTREAM_OF
AT5G13770
OGAP8274
OGRP6933
OG_Bdis_Atha...
IN_OG
INVOLVED_IN
Flowering
motif
Bdis_Bd...
Bradi5g19320....
IS_UPSTREAM_...
IS_UPSTREAM_WITH
IN_HG
core