Bioinformatique

Emploi du temps, groupes de TP, supports de cours et de TP disponibles sur http://silico.biotoul.fr sur la page Enseignement

Contrôle avec report : 30%
1er mars 13h45 (durée 1h30)

Contrôle terminal : 70%
12 avril 13h45 (durée 1h30)

Objectifs pédagogiques :

- Apperçu de quelques domaines d'application de la bioinformatique
  - traitement d'image : mesures phénotypiques
  - gestion des données
  - statistiques : corrélation phénotype - génotype
  - banques de données et sites Web publiques
  - (modélisation de systèmes biologiques et simulations)

Intervenant·e·s

- Roland Barriot (intro, bases de données, ...), Maxime Bonhomme et Raphaël Mourad (génétique, statistiques), Franck Delavoie et Silvia Kocanova (imagerie), Gwennaele Fichant (biologie des systèmes), Elodie Gaulin (bioanalyse), Matthieu Genais (doctorant)
- contact : barriot@univ-tlse3.fr mais de préférence de vive voix après un cours ou un TD/TP

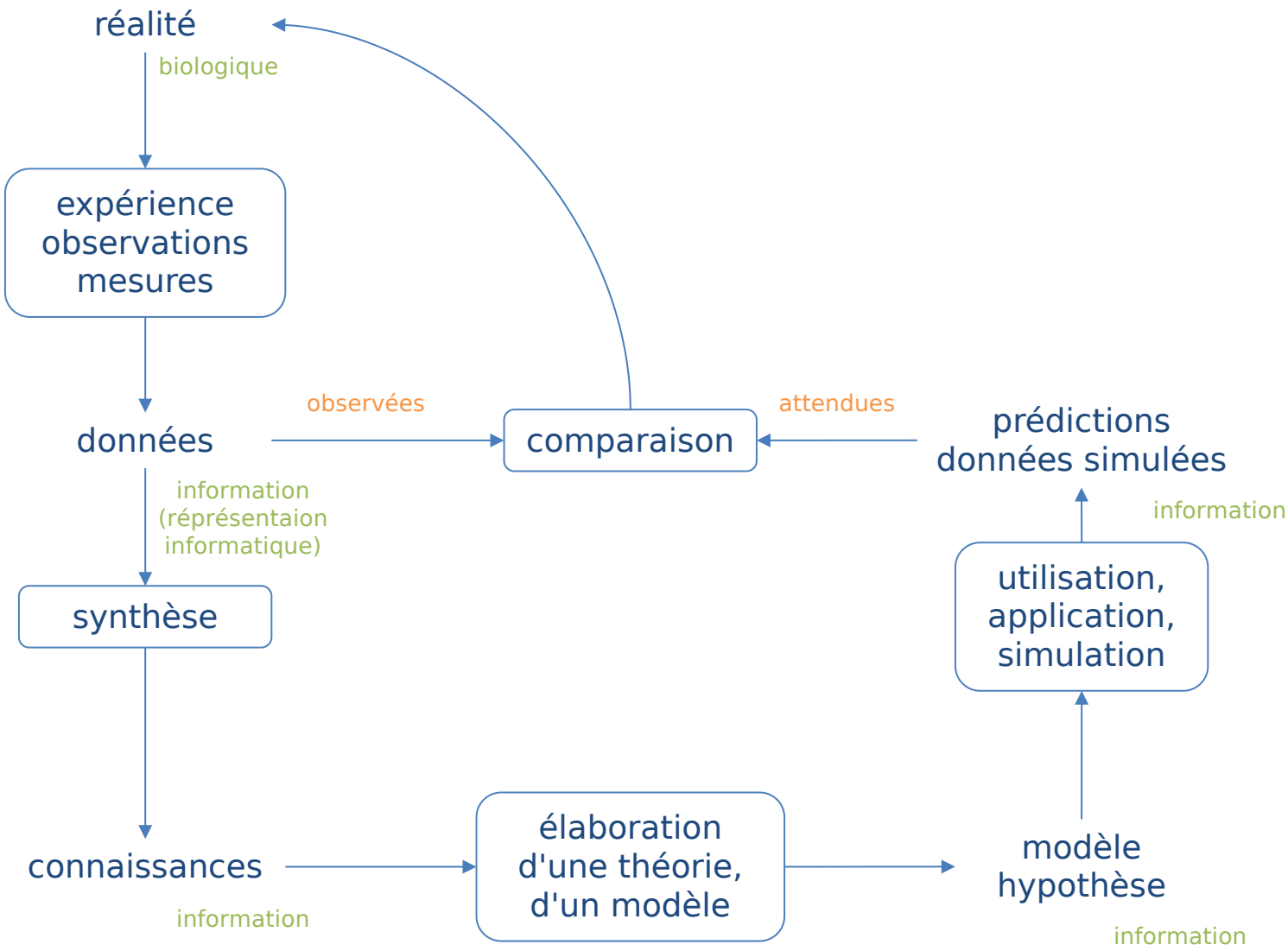
Définition

- Qu'est-ce que la bioinformatique ?
- Plusieurs réponses :
  - pas de l'informatique *bio*
  - traitement de l'information biologique
  - domaine spécialisé de la biologie
  - domaine de recherche multidisciplinaire : biologie, santé, mathématique, statistique, informatique, physique, éthique

- Une définition : modélisation et traitement des informations biologiques par des méthodes informatiques et/ou mathématiques et/ou statistiques.

- Les données et connaissances biologiques posent de nouveaux défis spécifiques pour leur gestion, représentation, et leur analyse/traitement.

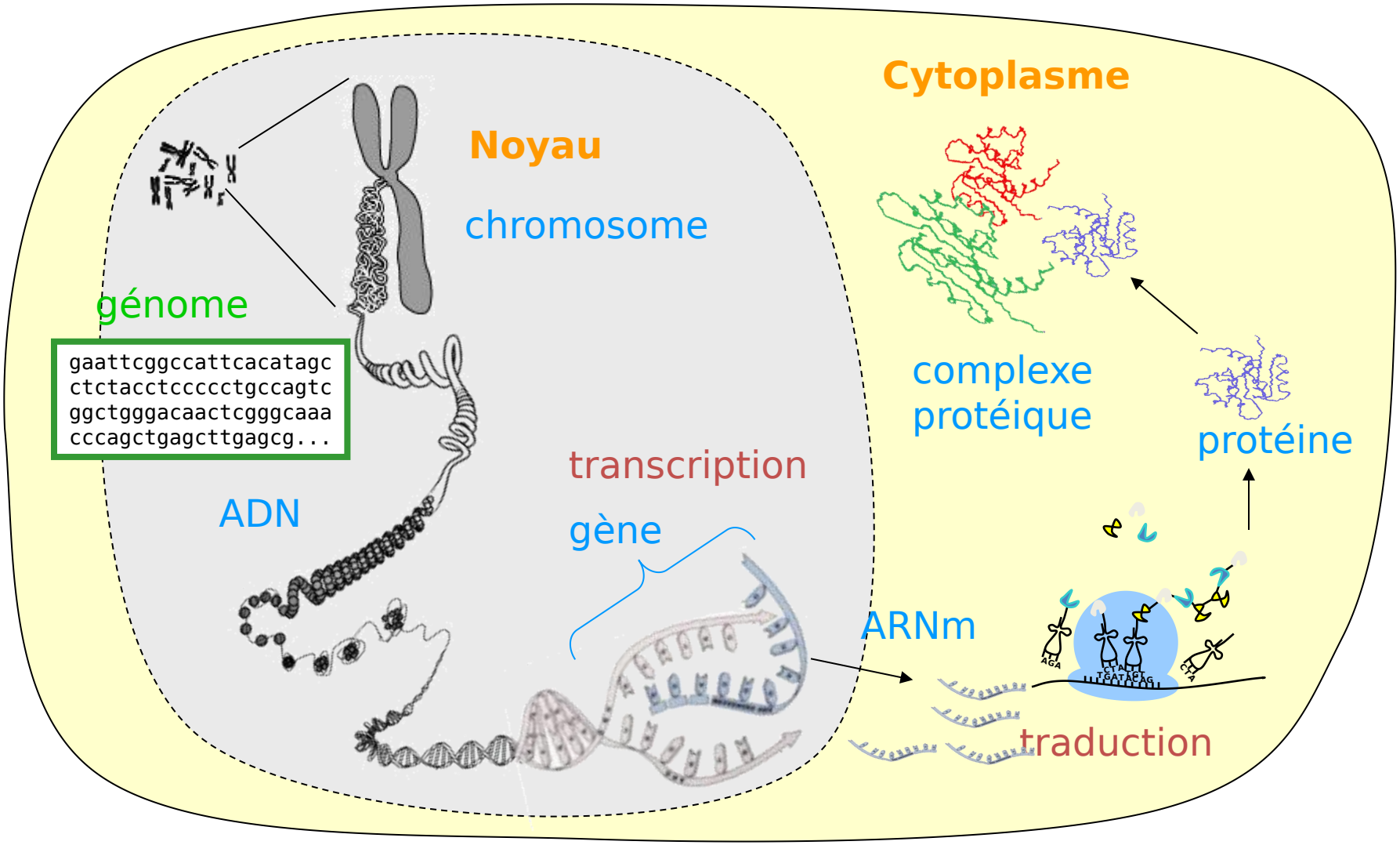
Méthode scientifique



Historique et domaines d'application liés à la bioinformatique

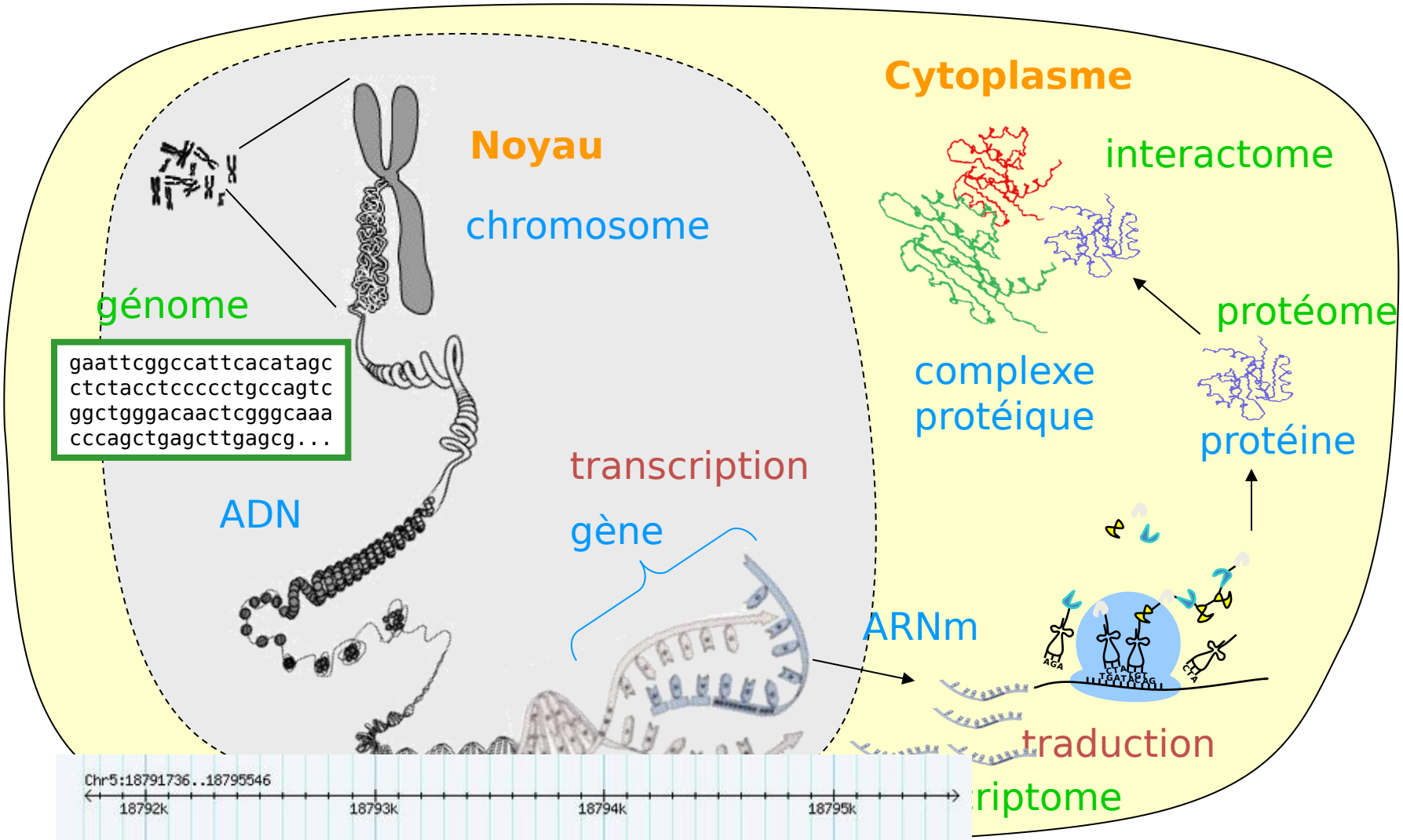
1970	Apparition du terme bioinformatique Disponibilité de séquences de gènes (et du 1er génome complet en 1977, bactériophage phi X174) Comparaison de séquences
1980	Alignement de séquences et recherche de séquences similaires Analyse phylogénétique (basée sur l'alignement de séquences) Prédiction de fonction (par similarité de séquence) Banques de données (de séquences)
1990	Débuts de la génomique et génomique comparative ; analyse du contenu d'un génome (1er génome complet d'organisme vivant à réplication autonome *Haemophilus influenzae*, 1995) Débuts des analyses globales de l'expression des gènes et des protéines Famille de gènes, de protéines, de domaines Prédiction de la structure 3D des protéines
2000	Généralisation des approches globales Traitement de graphes : interactions protéine-protéine, réseau de régulation de l'expression des gènes, réseau métabolique Apprentissage automatique (*data mining*) Fouille de texte (*text mining*) Biologie des systèmes : modélisation de systèmes biologiques Intégration de données hétérogènes, visualisation Médecine personnalisée NGS/Séquençage très haut débit Traitement d'images liées à la microscopie
2010	Ethique, confidentialité des données, manipulation du génome Modélisation d'une cellule, d'un organisme, d'une population, d'un écosystème Biologie synthétique

(Quelques) données (mesures et prédictions) et connaissances disponibles

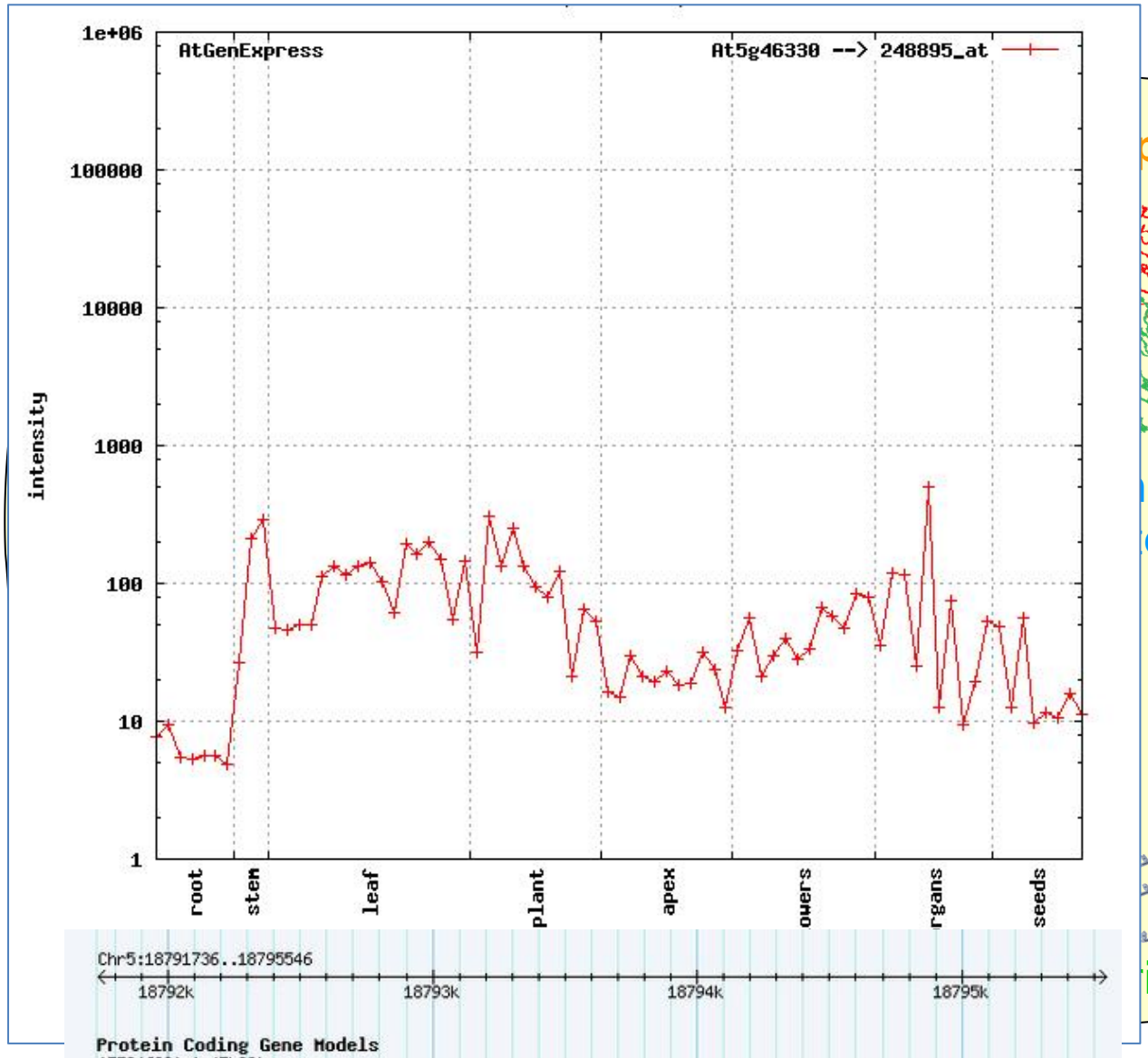


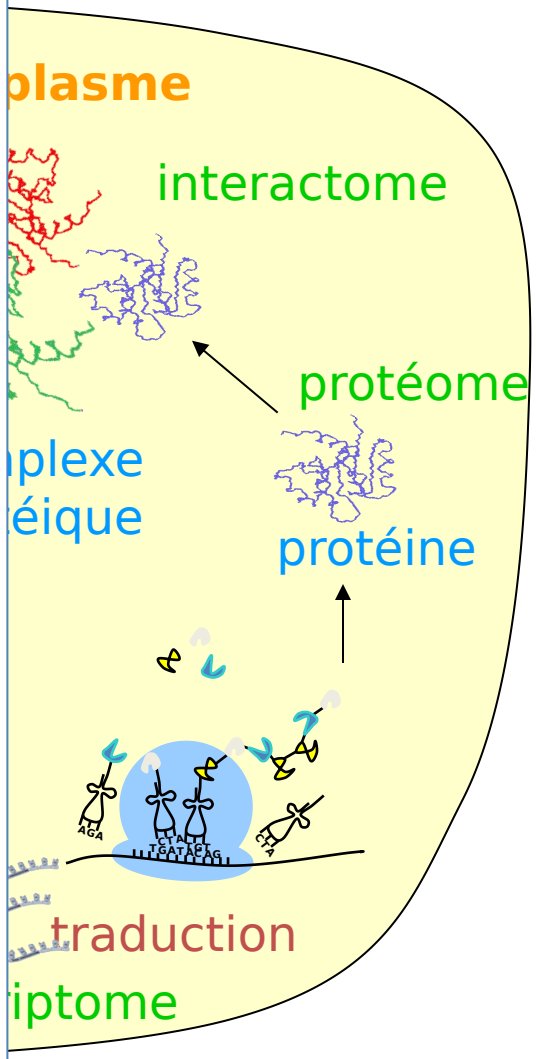
Cellule eucaryote

(Quelques) données (mesures et prédictions) et connaissances disponibles



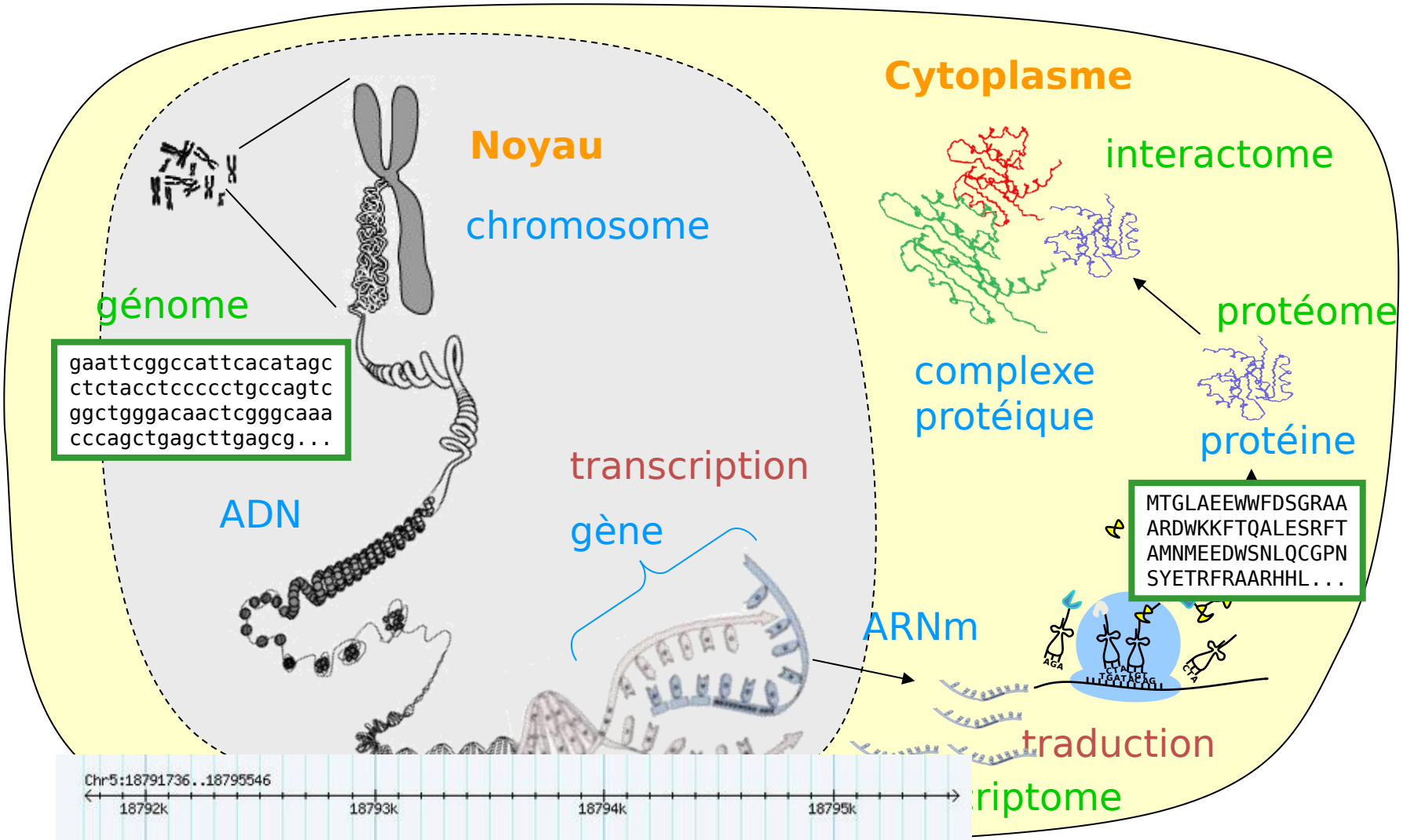
(Quelques) données (mesures et prédictions) et connaissances disponibles



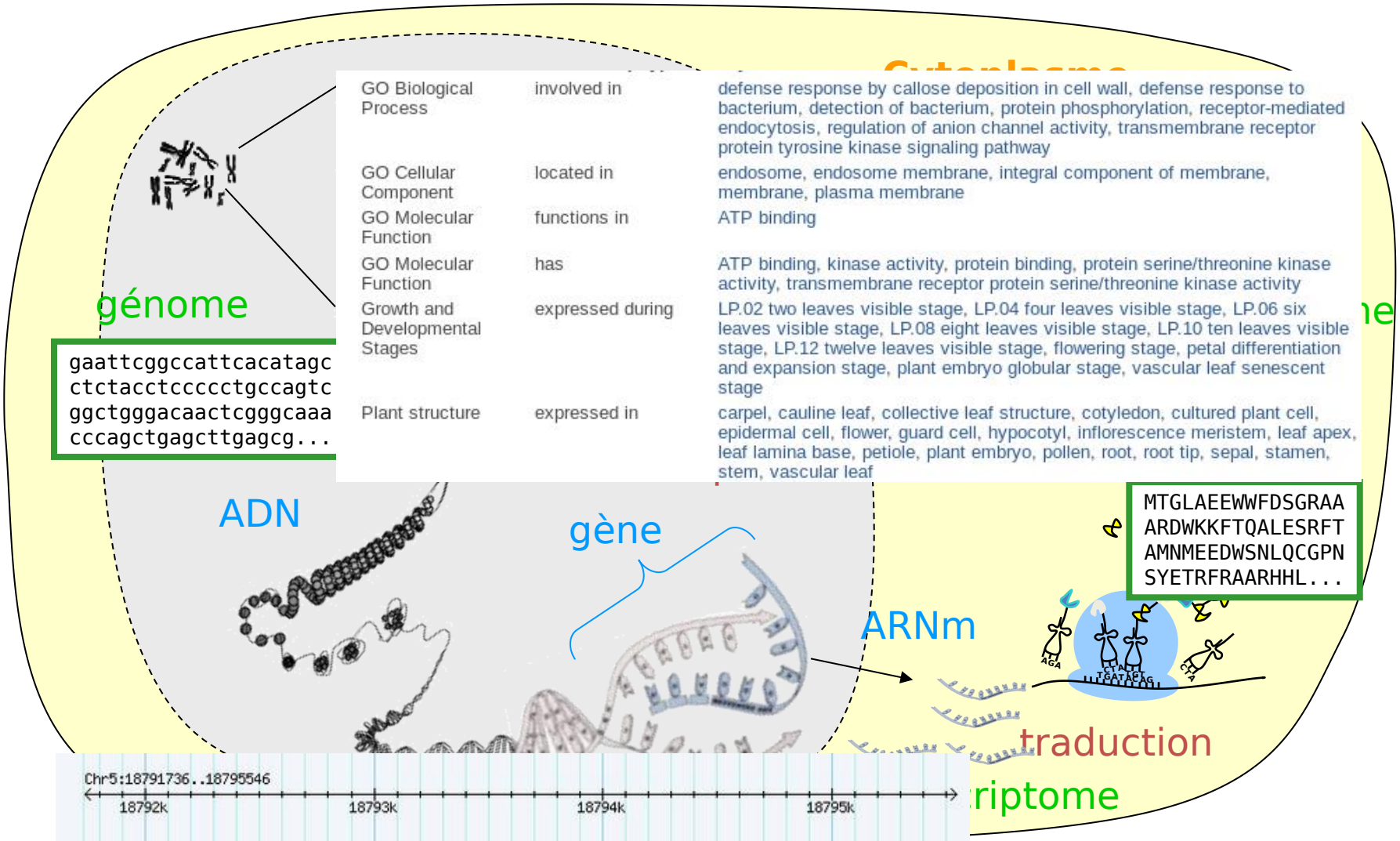


Cellu

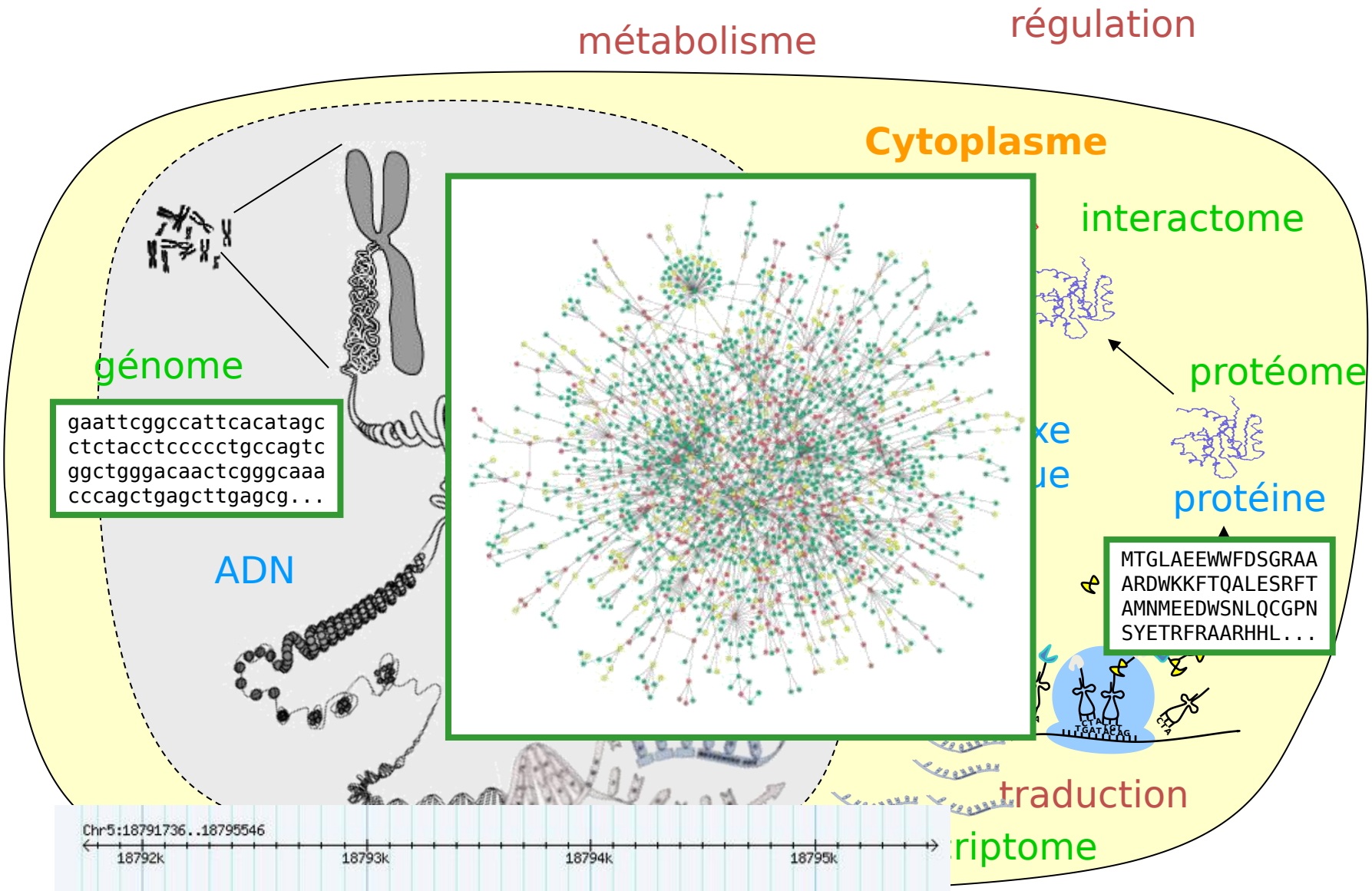
(Quelques) données (mesures et prédictions) et connaissances disponibles



(Quelques) données (mesures et prédictions) et connaissances disponibles



(Quelques) données (mesures et prédictions) et connaissances disponibles



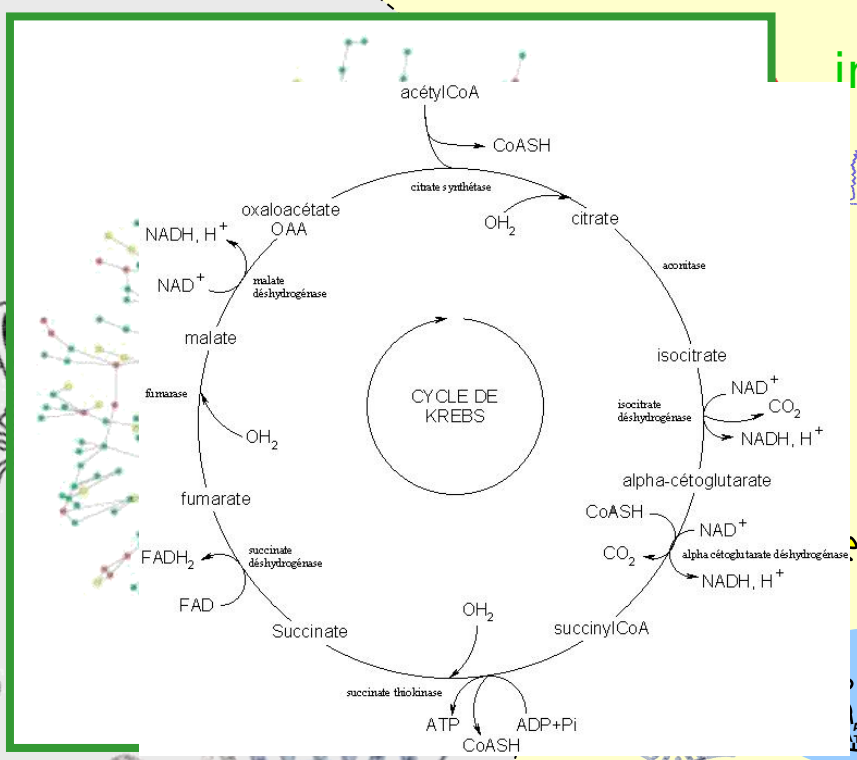
(Quelques) données (mesures et prédictions) et connaissances disponibles

Cytoplasme

génome

```
gaattcggccattcacatagc
ctctacctccccctgccagtc
ggctgggacaactcgggcaaa
cccagctgagcttgagcg...
```

ADN



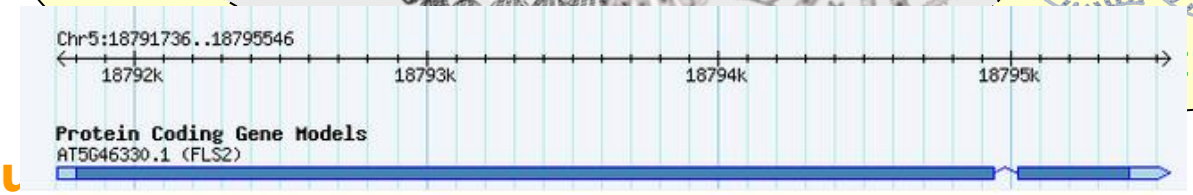
interactome

protéome

protéine

```
MTGLAEEWWFDSGRAA
ARDWKKFTQALESRFT
AMNMEEDWSNLQCGPN
SYETRFRAARHHL...
```

traduction



riptome

Cellu

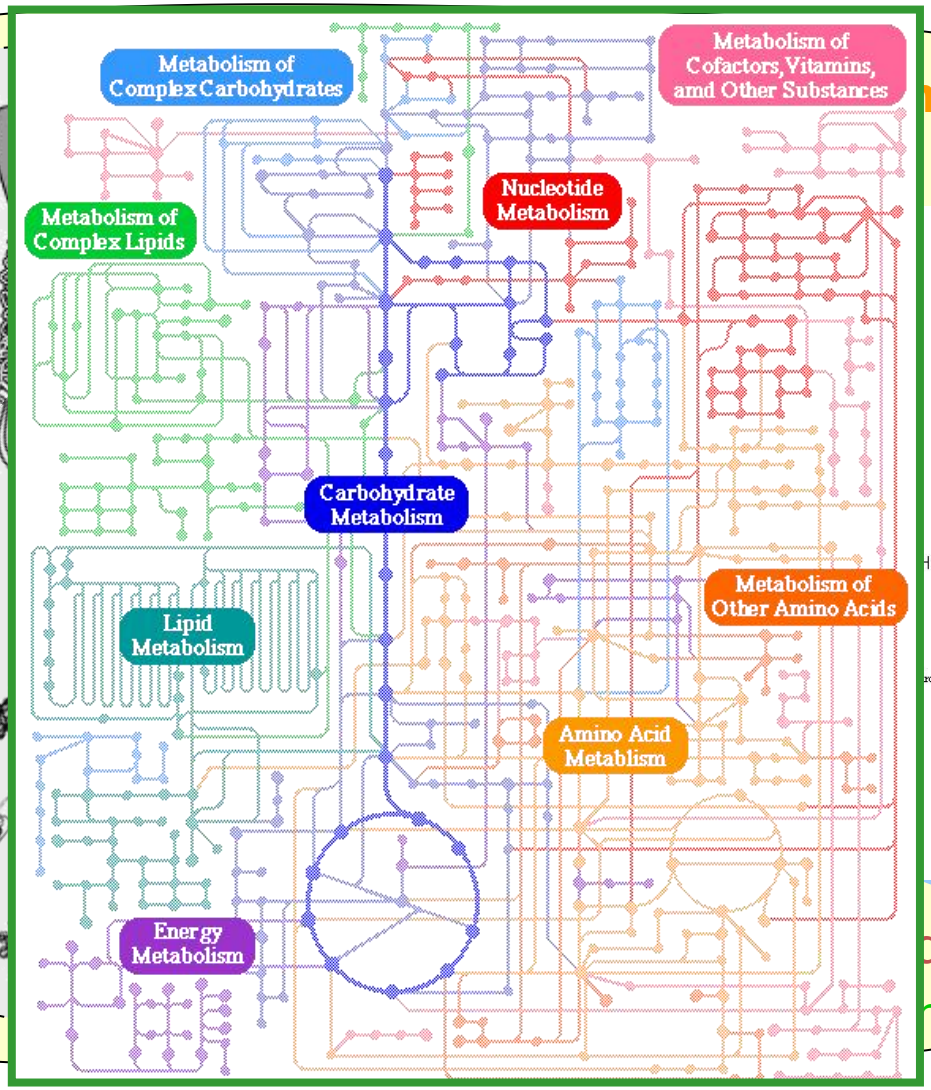
(Quelques) données (mesures et prédictions) et connaissances disponibles

métabolisme

génome

gaattcggccattcacatagc
ctctacctccccctgccagtc
ggctgggacaactcgggcaaa
cccagctgagcttgagcg...

ADN



métabolome

interactome

protéome

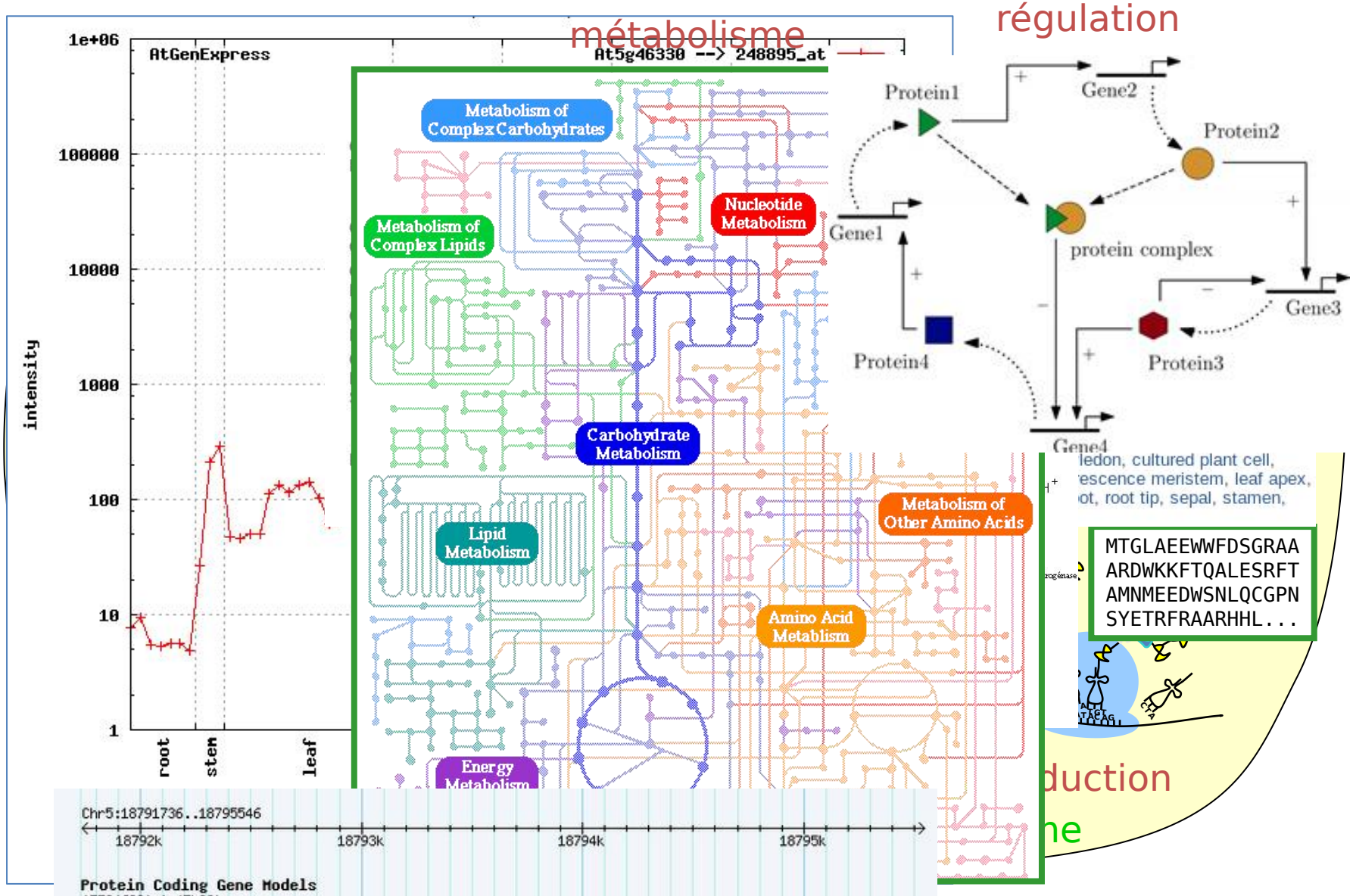
protéine

MTGLAEEWWFDSGRAA
ARDWKKFTQALESRFT
AMNMEEDWSNLQCGPN
SYETRFRAARHHL...

duction

Cellule eucaryote

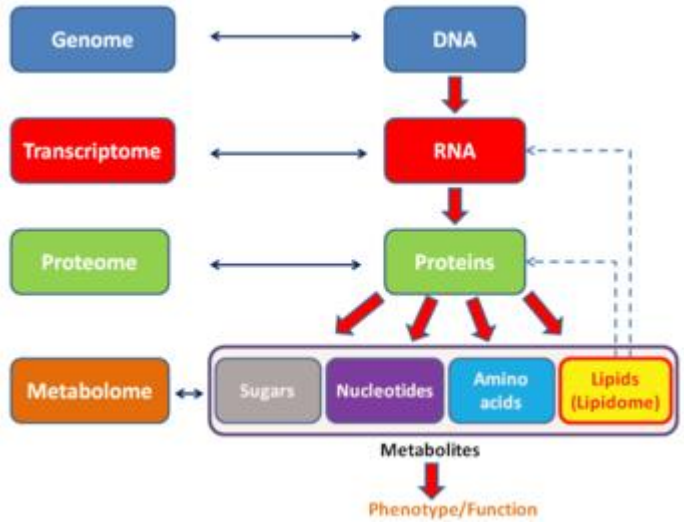
(Quelques) données (mesures et prédictions) et connaissances disponibles



Cellu

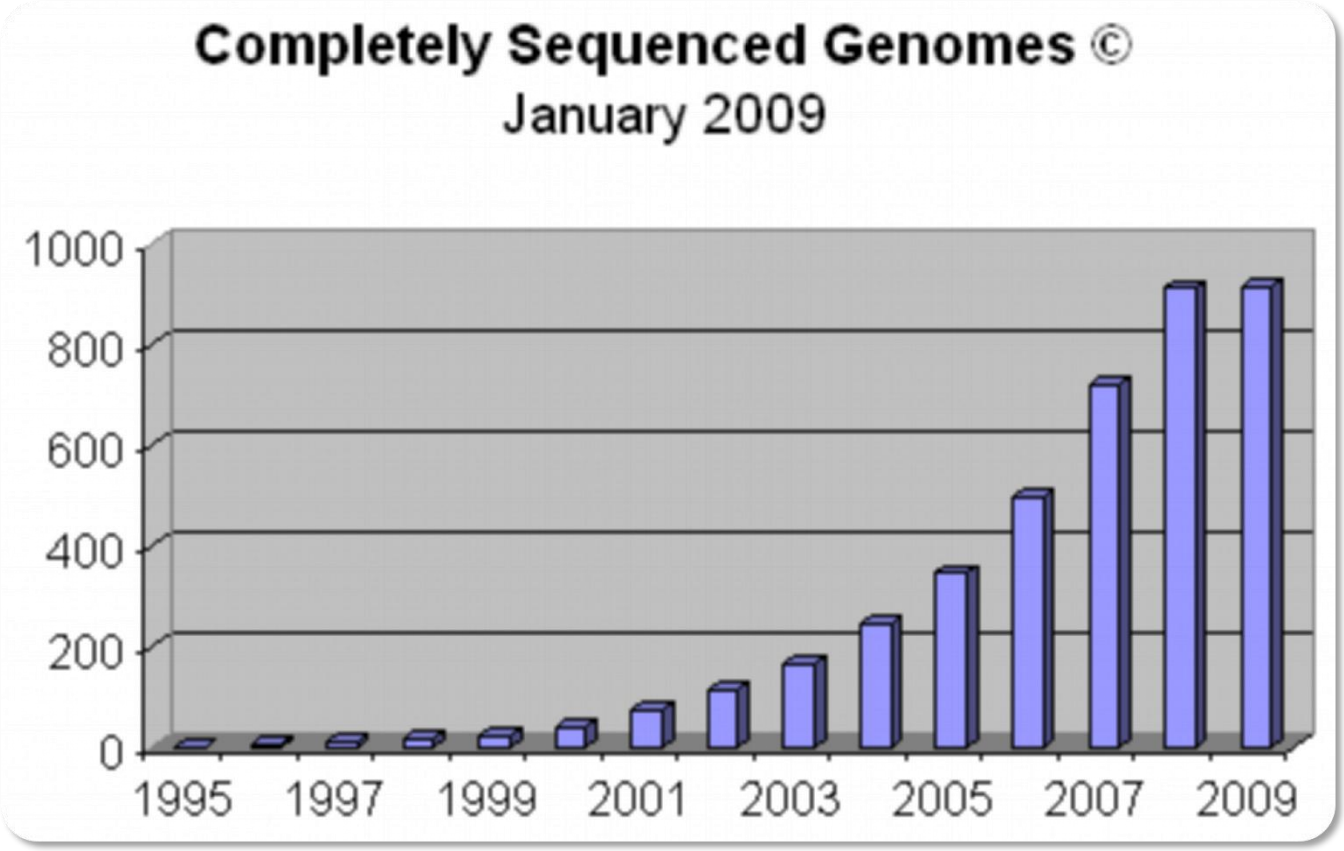
Données omics

- **Génome**
  - séquence(s) nucléique(s) de l'ensemble des chromosomes d'un organisme
  - ensemble des gènes d'un organisme
- **Transcriptome**
  - ensemble des ARNm ou transcrits présents dans une cellule ou une population de cellules dans des conditions données
- **Protéome**
  - ensemble des protéines présentes dans une cellule ou une population de cellules dans des conditions données
- **Interactome**
  - ensemble des intéractions moléculaires pouvant survenir *in vivo*
  - ensemble des intéractions moléculaires dans des conditions données
  - ensemble des intéractions au sein d'un organisme : moléculaires, physiques, génétiques, fonctionnelles, …
- **Métabolome**
  - ensemble des métabolites présents dans une cellule ou une population de cellules dans des conditions données
- Exome, lipidome, phénome, régulome, sécrétome, épigénome, méthylome, …



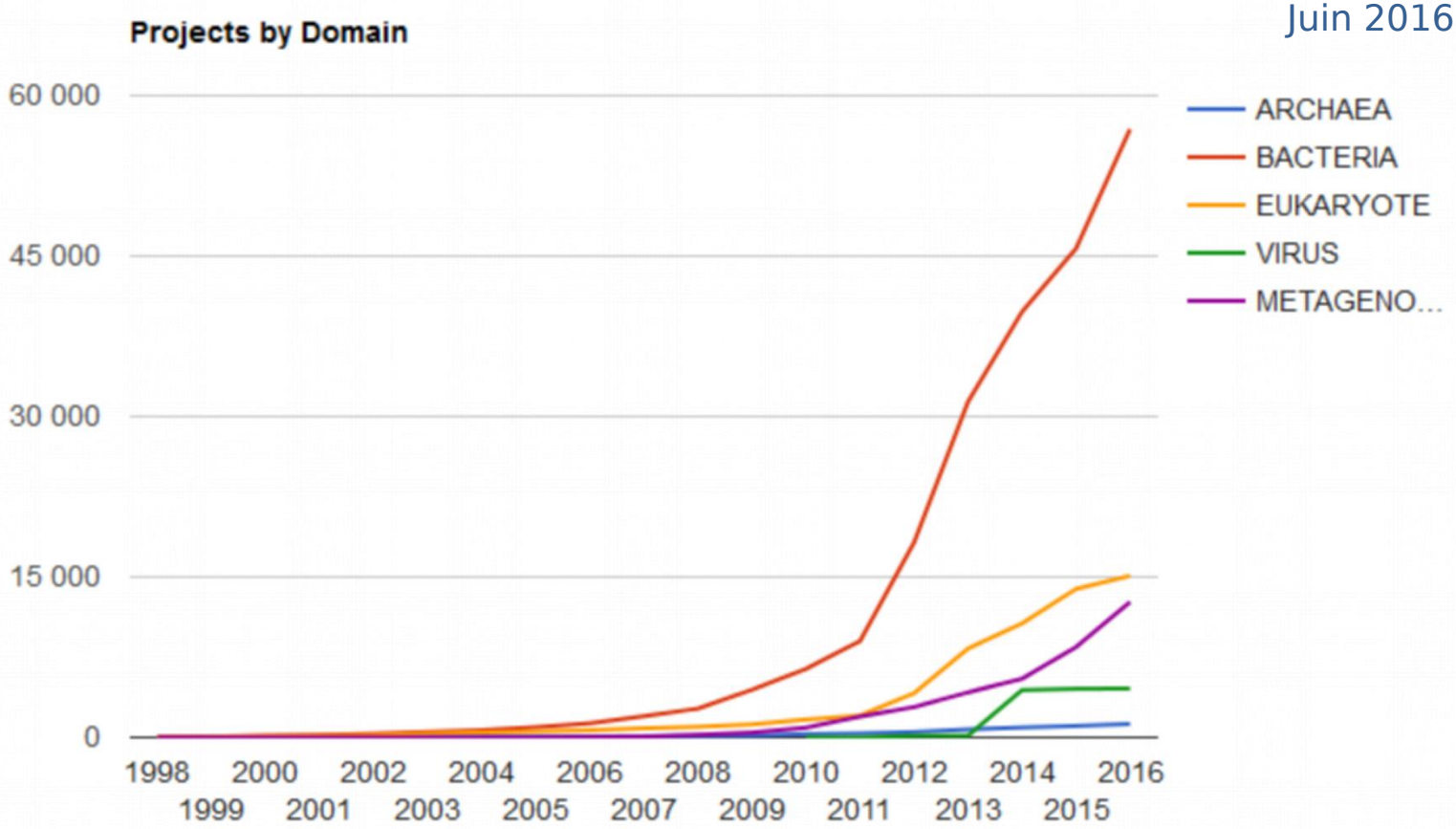
source : Wikipedia

Séquences disponibles : quelques chiffres

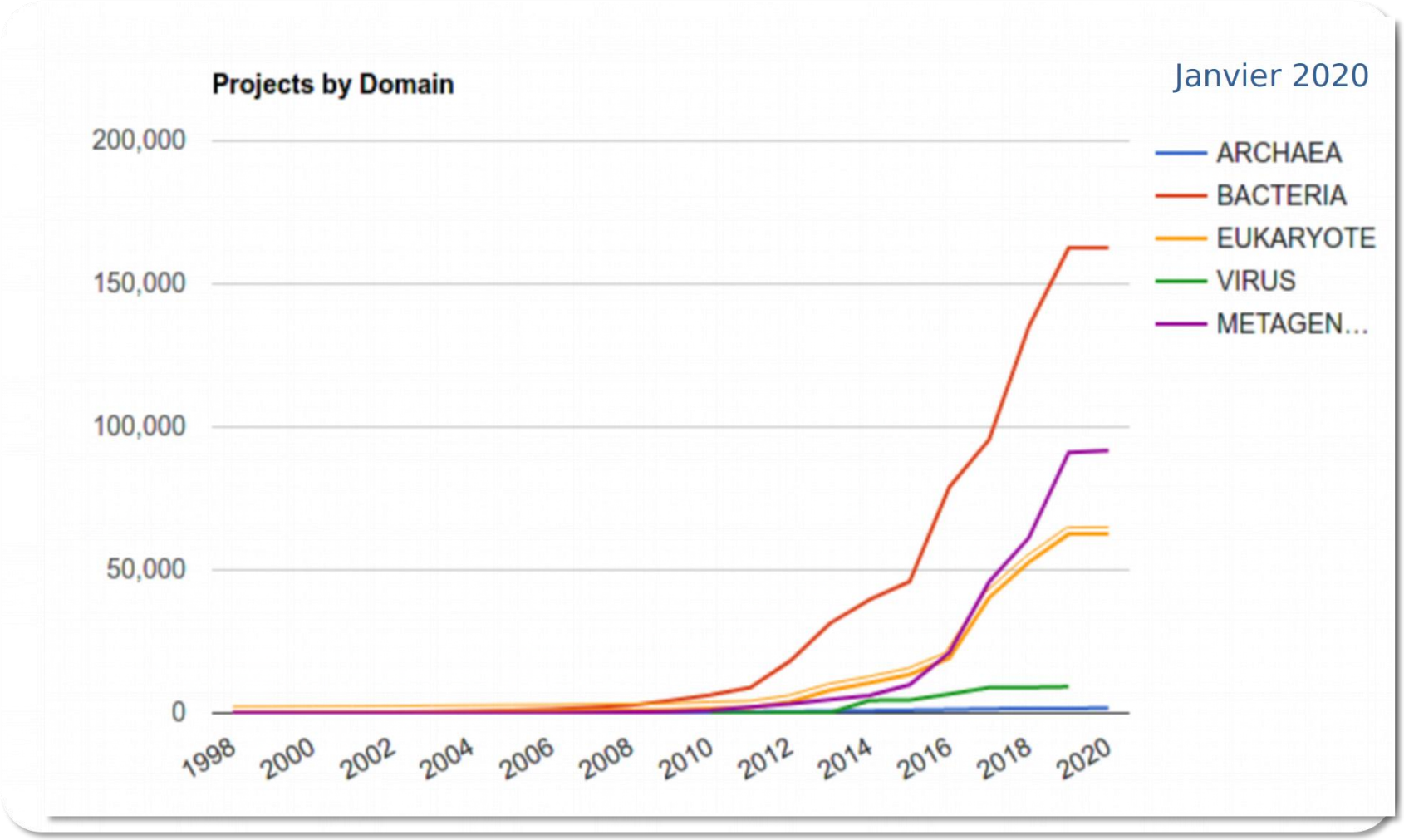


Séquences disponibles : quelques chiffres

Juin 2016

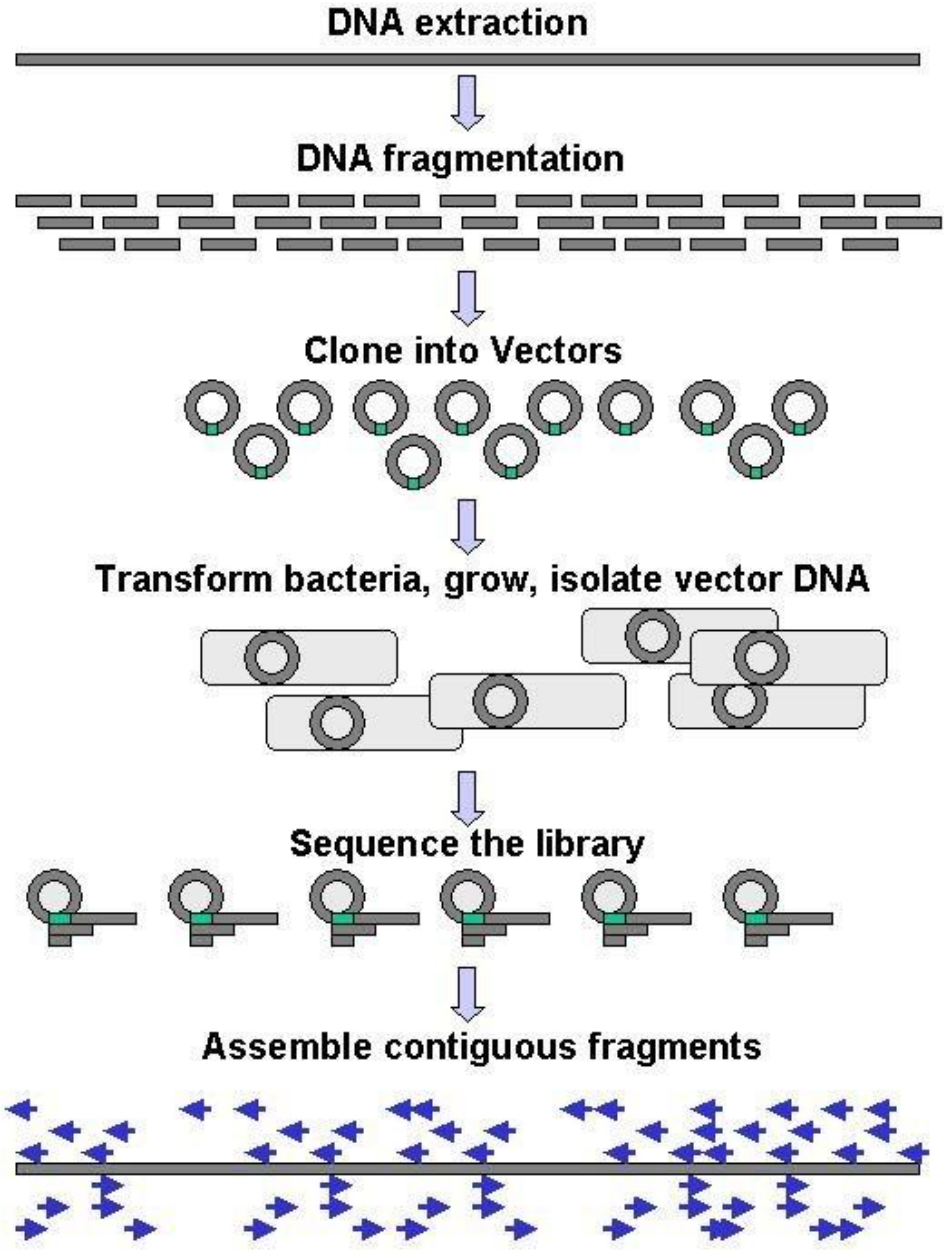


Séquences disponibles : quelques chiffres



source : GOLD

Séquençage d'un génome



source : Wikipedia

L’analyse manuelle d’une séquence peut s’avérer laborieuse

TCCTGGCCTACATGTTCTTTGGCAAAGGATCTTCAAAATCAACGGCTCCCGGTGCGGCGATCATCCATTTCTTCGGAGGGATTCACGAGATTT
ACTTCCCGTACATTCTGATGAAACCTGGCCCTGATTCTCGCAGCCATTGCCGGCGGAGCAAGCGGACTCTTAACATTACGATCTTTAATGCCG
GACTTGTCGCGGCAGCGTCACCGGGAAGCATTATCGCATTGATGGCAATGACGCCAAGAGGAGGCTATTTCGGCGTATTGGCGGGTGTATTGG
TCGCTGCAGCTGTATCGTTCATCGTTTCAGCAGTGATCCTGAAATCCTCTAAAGCTAGTGAAGAAGACCTGGCTGCCGCAACAGAAAAAATGC
AGTCCATGAAGGGGAAGAAAAGCCAAGCAGCAGCTGCTTTAGAGGCGGAACAAGCCAAAGCAGAGAAGCGTCTGAGCTGTCTCCTGAAAGCGC
GAACAAAATTATCTTTTCGTGTGATCCGGGATGGGATCAAGTGCCATGGGGGCATCCATCTTAAGAAACAAAGTGAAAAAGCGGAGCTTGACA
TCAGTGTGACCAACACGGCCATTAACAATCTGCCAAGCGATGCGGATATTGTCATCACCCACAAAGATTTAACAGACCGCGCGAAAGCAAAGC
TGCCGAACGCGACGCACATATCAGTGGATAACTTCTTAAACAGCCCGAAATACGACGAGCTGATTGAAAAGCTGAAAAGTAATCTTATAGAAA
GAGAGTATTGTCATGCAAGTACTCGCAAAGGAAACATTAAACTCAATCAAACGGTATCATCAAAAGAAGAGGCTATCAAATTGGCAGGCCAGA
CGCTGATTGACAACGGCTACGTGACAGAGGATTACATTAGCAAAATGTTTGACCGTGAAGAAACGTCTTCTACGTTTATGGGGAATTTCATTG
CCATTCCACACGGCACAGAAGAAGCGAAAAGCGAGGTGCTTCACTCAGGAATTTCAATCATACAGATTCCAGAGGGCGTTGAGTACGGAGAAG
GCAACACGGCAAAAGTGGTATTCGGCATTGCGGGTAAAAATAATGAGCATTTAGACATTTTGTCTAACATCGCCATTATCTGTTCAGAAGAAG
AAACATTGAACGCCTGATCTCCGCTAAAGCGAAGAAGATTTGATCGCCATTTCAACGAGGTGAACTGACATGATCGCCTTACATTTCGGTGCG
GGAAATATCGGGAGAGGATTTATCGGCGCGCTGCTTCACCACTCCGGCTATGATGTGGTGTTTGCGGATGTGAACGAAACGATGGTCAGCCTC
CTCAATGAAAAAAAAGAATACACAGTGGAACTGGCGGAAGAGGGACGTTCATCGGAGATCATTGGCCCGGTGAGCGCTATTAACAGCGGCAGT
CAGACCGAGGAGCTGTACCGGCTGATGAATGAGGCGGCGCTCATCACAACAGCTGTCGGCCCGAATGTCCTGAAGCTGATTGCCCCGTCTATC
GCAGAAGGTTTAAGACGAAGAAATACTGCAAACACACTGAATATCATTGCCTGCGAAAATATGATTGGCGGAAGCAGCTTCTTAAAGAAAGAA
ATATACAGCCATTTAACGGAAGCAGAGCAGAAATCCGTCAGTGAAACGTTAGGTTTTCCGAATTCTGCCGTTGACCGGATCGTCCCGATTCAG
CATCATGAAGACCCGCTGAAAGTATCGGTTGAACCATTTTTCGAATGGGTCATTGATGAATCAGGCTTTAAAGGGAAAACACCAGTCATAAAC
GGCGCACTGTTTGTTGATGATTTAACGCCGTACATCGAACGGAAGCTGTTTACGGTCAATACCGGACACGCGGTCACAGCGTATGTCGGCTAT
CAGCGCGGACTCAAAACGGTCAAAGAAGCAATTGATCATCCGGAAATCCGCCGTGTTGTTCATTCGGCGCTGCTTGAAACTGGTGACTATCTC
GTCAAATCGTATGGCTTTAAGCAAACTGAACACGAACAATATATTAAAAATCAGCGGTCGCTTTTAAAATCCTTTCATTTCGGACGATGTGAC
CCGCGTAGCGAGGTCACCTCTCAGAAAACTGGGAGAAAATGTAGACTTGTAGGCCCGGCAAAGAAAATAAAAGAACCGAATGCACTGGCTGAA
GGAATTGCCGCAGCACTGCGCTTCGATTTCACCGGTGACCCTGAAGCGGTTGAACTGCAAGCGCTGATCGAAGAAAAGGATACAGCGGCGTAC
TTCAAGAGGTGTGCGGCATTCAGTCCCATGAACCGTTGCACGCCATCATTTTAAAGAAACTTAATCAATAACCGACCACCCGTGACACAATGT
CACGGGCTTTTTACTATCTCGCAATCTAGTATAATAGAAAGCGCTTACGATAACAGGGGAAGGAGAATGACGATGAAACAATTTGAGATTGCG
GCAATACCGGGAGACGGAGTAGGAAAGAGGTTGTAGCGGCTGCTGAGAAAGTGCTTCATACAGCGGCTGAGGTACACGGAGGTTTGTCATTCT
CATTCACAGCTTTTCCATGGAGCTGTGATTATTACTTGGAGCACGGCAAAAATGATGCCCGAAGATGGAATACATACGCTTACTCAATTTGAA
GCAGTTTTTGGGAGCTGTCGGAAATCCGAAGCTGGTTCCCGATCATATATCGTTATGGGGCTGCTGCTGAAATCCGGAGGGAGCTTGAGCTTT
CCATTAATATGAGACCCGCCAAACAAATGGCAGGCATTACGTCGCCGCTTCTGCATCCAAATGATTTTTGACTTCGTGGTGATTCGCGAGAAC
AGTGAAGGTGAATACAGTGAAGTTGTCGGGCGCATTCACAGAGGCGATGATGAAATCGCCATCCAGAATGCCGTGTTTACGAGAAAAGCGACA
GAACGTGTCATGCGCTTTGCCTTCGAATT

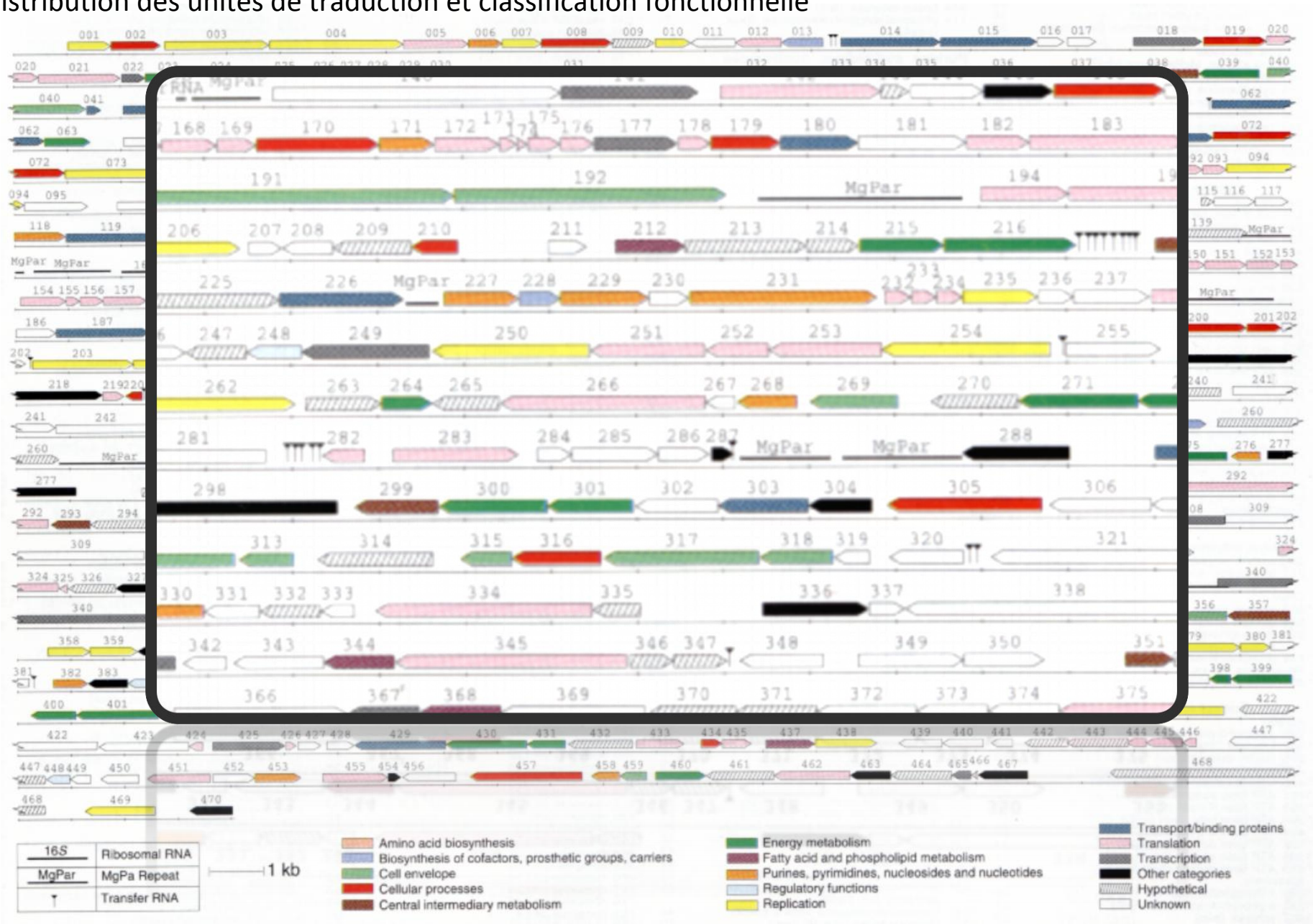
Des séquences... et après ?

- Annotation
  - régions codantes, régions régulatrice, ...
  - prédiction fonctionnelle
- Identification de protéine/Prédiction de structure
- Analyse des relations génotype/phénotype
- Analyses évolutives
- Analyses d'expression des gènes/protéines
- Réseau d'interaction protéine-protéine
- Reconstruction du réseau métabolique
- Reconstruction du réseau de régulation de l'expression des gènes

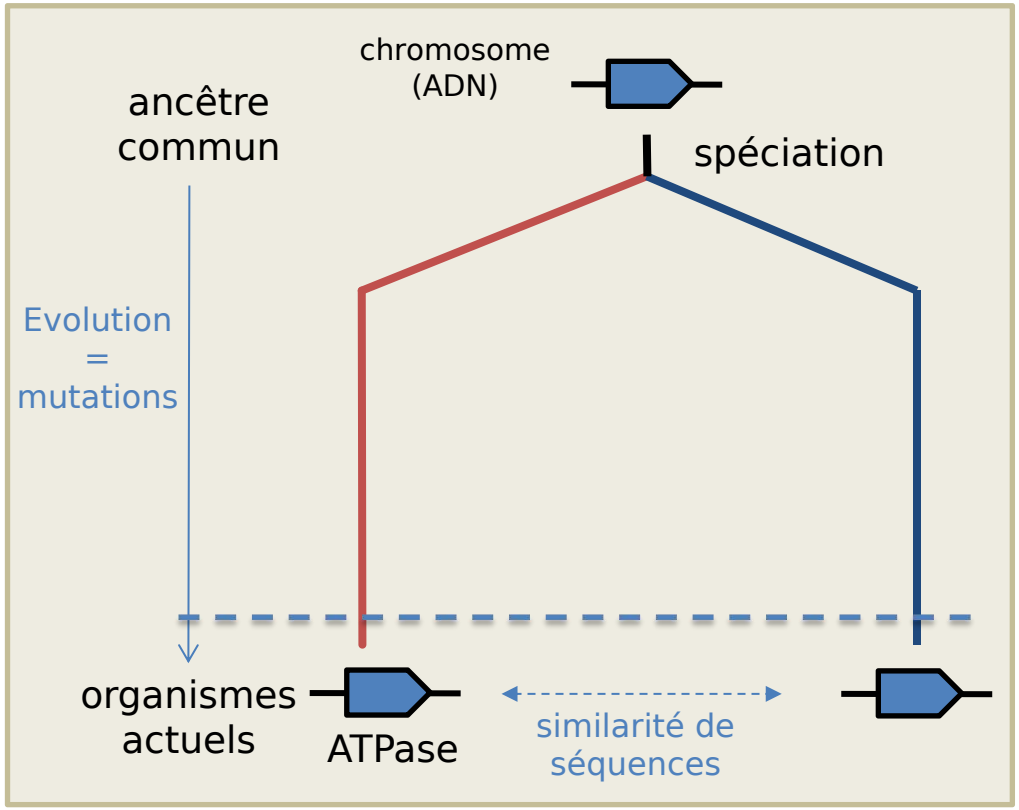
- Identification des gènes codant pour :
  - les ARNr
  - les ARNt
  - les protéines
- Identification des unités de transcription (promoteur et terminateur)
- Identification des unités de traduction
- Pour les gènes codant pour les protéines, prédiction fonctionnelle par recherche de similarité de séquences (BLAST) et prédiction de fonction ou classification en grandes classes fonctionnelles (ex: biosynthèse des acides aminés, métabolisme énergétique, …)

Génome de *Mycoplasma genitalium*

Distribution des unités de traduction et classification fonctionnelle



Homologie et transfert d'annotation



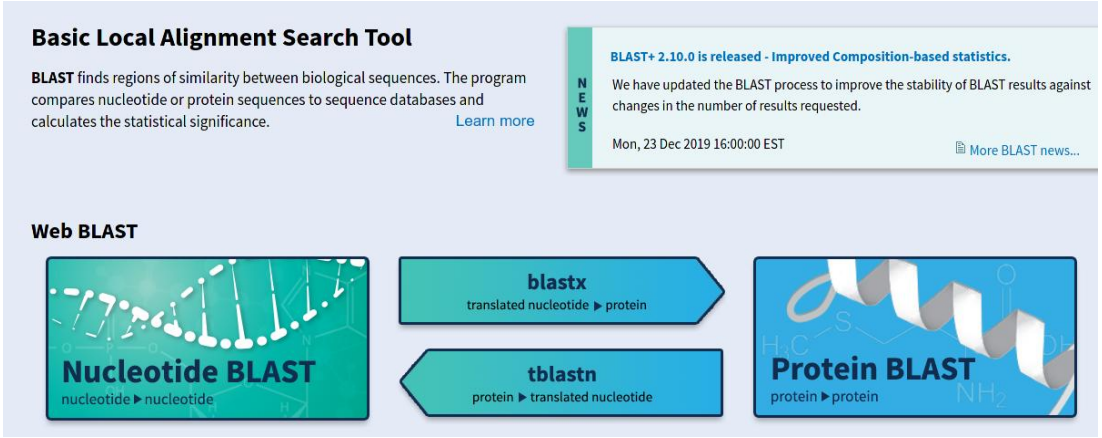
Du temps du séquençage des premiers gènes et génomes

- la conservation/similarité des séquences impliquait des fonctions similaires
- les annotations des gènes caractérisés expérimentalement étaient transférées aux nouveaux gènes/génomes séquencés

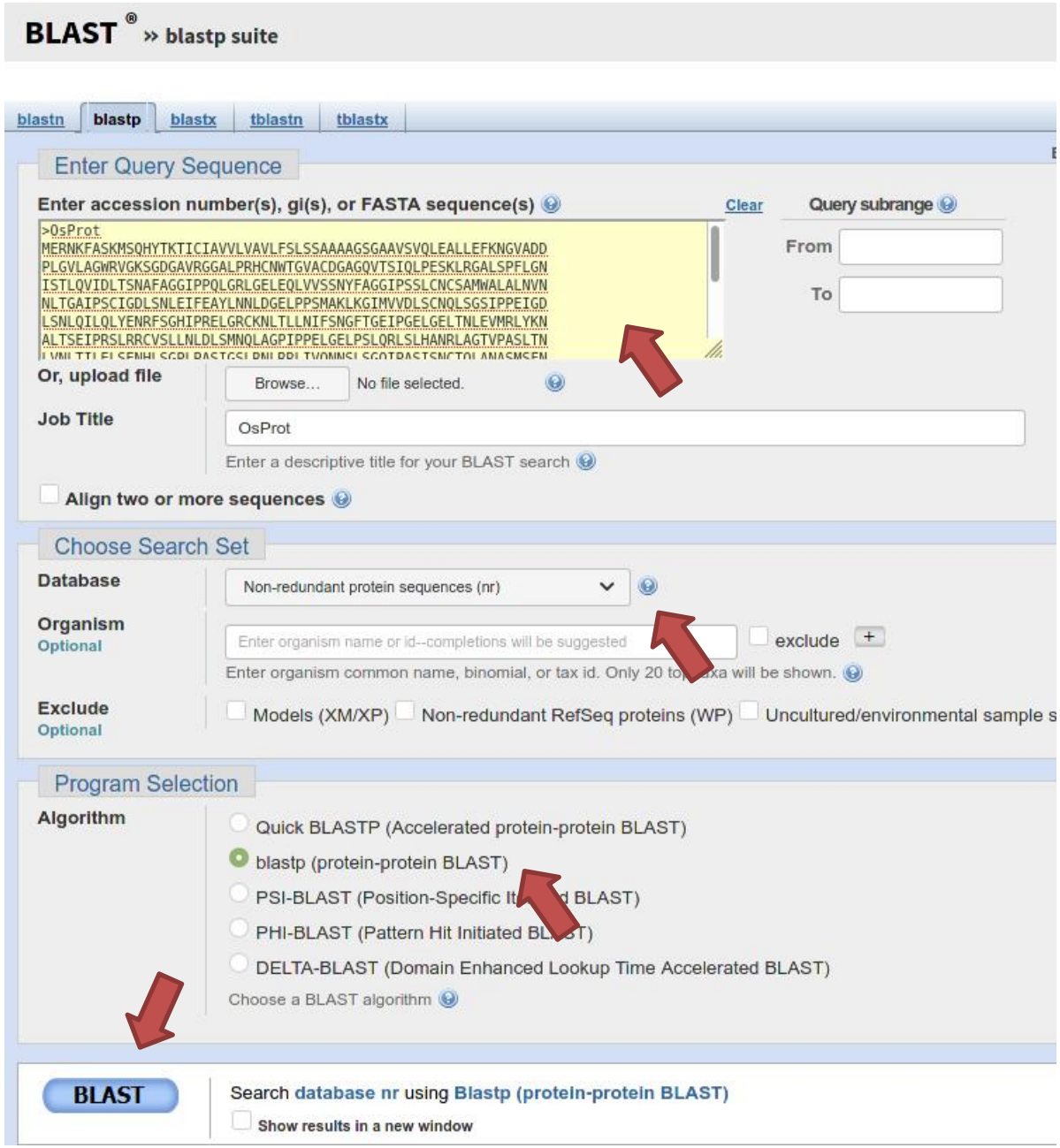
Séquence d'intérêt = Query

Quelles séquences sont proches (suffisamment similaires) et nous indiqueraient la fonction du gène ou de la protéine ?

BLAST : recherche de séquences similaires (Hits) par alignement local de la séquence query avec les séquences d'une banque



BLAST formulaire de requête



BLAST résultats

Job Title OsProt
RID 2ANVU5N101R
Search expires on 01-21 14:32 pm
Download All
Program BLASTP Citation
Database nr See details
Query ID lcl|Query_91880
Description OsProt
Molecule type amino acid
Query Length 1183
Other reports Distance tree of results Multiple alignment
MSA viewer

Filter Results

Organism *only top 20 will appear* exclude
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity to
E value to
Query Coverage to
Filter Reset

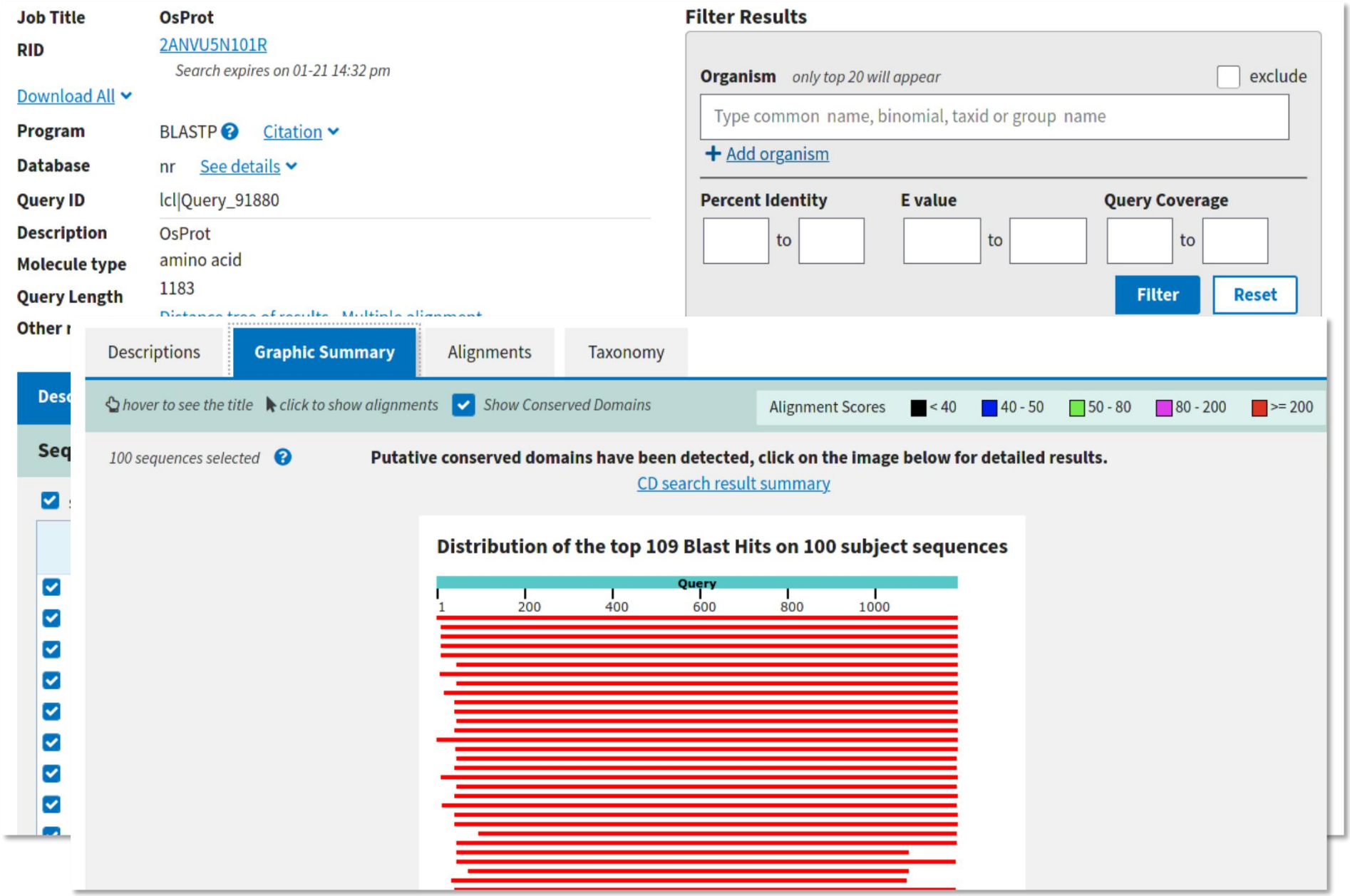
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

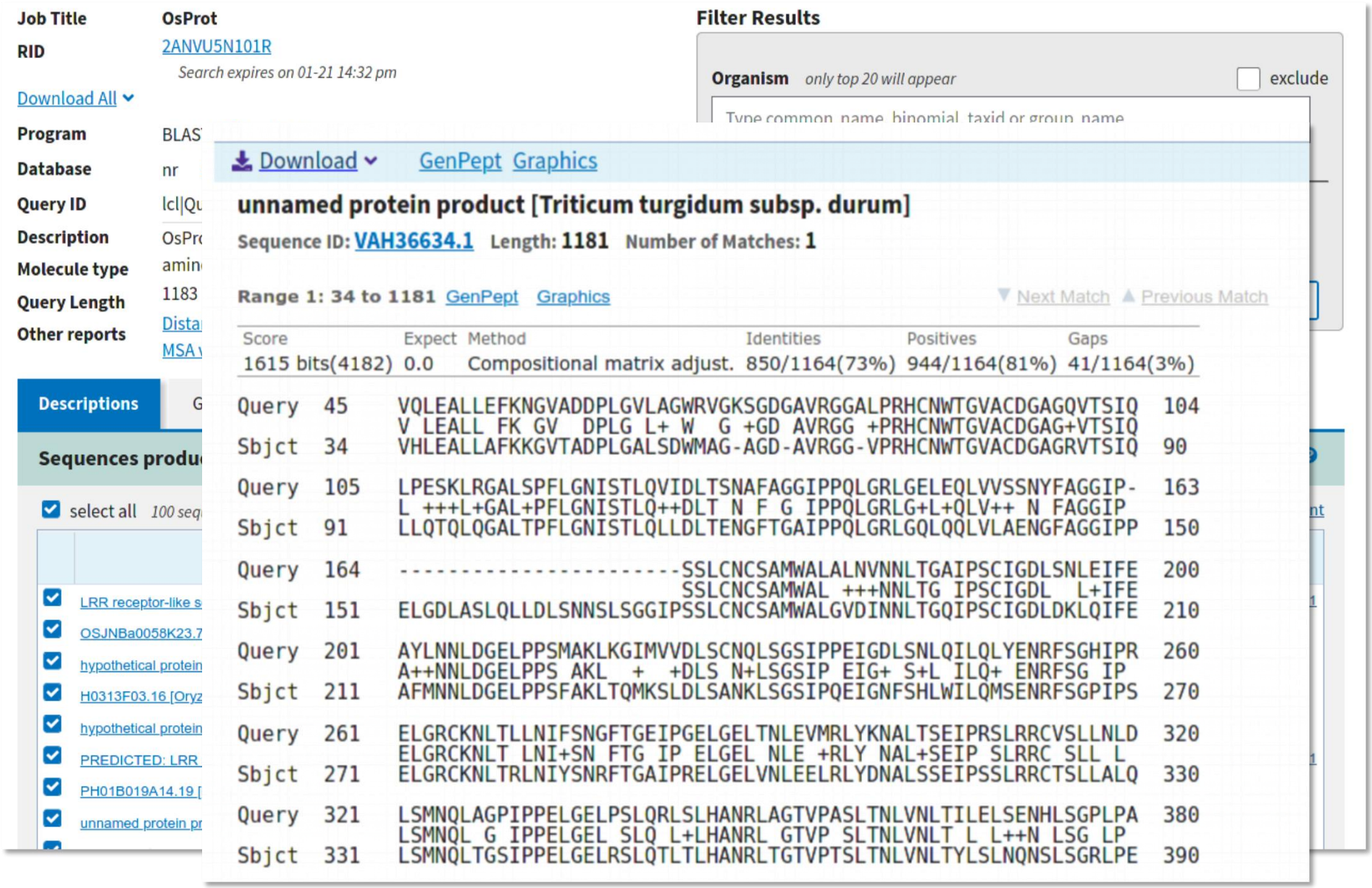
select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
LRR receptor-like serine/threonine-protein kinase FLS2 [Oryza sativa Japonica Group]	2372	2372	100%	0.0	100.00%	XP_015634951.1
OSJNBa0058K23.7 [Oryza sativa Japonica Group]	2357	2357	99%	0.0	100.00%	CAE02151.2
hypothetical protein OsJ_16186 [Oryza sativa Japonica Group]	2352	2352	99%	0.0	99.91%	EAZ32006.1
H0313F03.16 [Oryza sativa]	2326	2326	99%	0.0	98.89%	CAH68341.1
hypothetical protein OsI_17436 [Oryza sativa Indica Group]	2055	2055	99%	0.0	89.59%	EEC78020.1
PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Oryza brachyantha]	1855	1855	96%	0.0	85.69%	XP_015691635.1
PH01B019A14.19 [Phyllostachys edulis]	1638	1638	99%	0.0	73.86%	CCI55350.1
unnamed protein product [Triticum turgidum subsp. durum]	1615	1615	96%	0.0	73.02%	VAH36634.1

BLAST résultats



BLAST résultats



Banque de séquences protéiques : UniProt

UniProt

UniProtKB | Advanced | Search

BLAST | Align | Retrieve/ID mapping | Peptide search | Help | Contact

UniProtKB - Q0JA29 (FLS2_ORYSJ)

Basket

Display

Entry
Publications
Feature viewer
Feature table

None

Function
Names & Taxonomy
Subcellular location
Pathology & Biotech
PTM / Processing
Expression
Interaction
Structure
Family & Domains
Sequence
Similar proteins
Cross-references
Entry information
Miscellaneous

Top

BLAST | Align | Format | Add to basket | History | Help video | Add a publication | Feedback

Protein **LRR receptor-like serine/threonine-protein kinase FLS2**

Gene **FLS2**

Organism *Oryza sativa subsp. japonica (Rice)*

Status Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level

Function

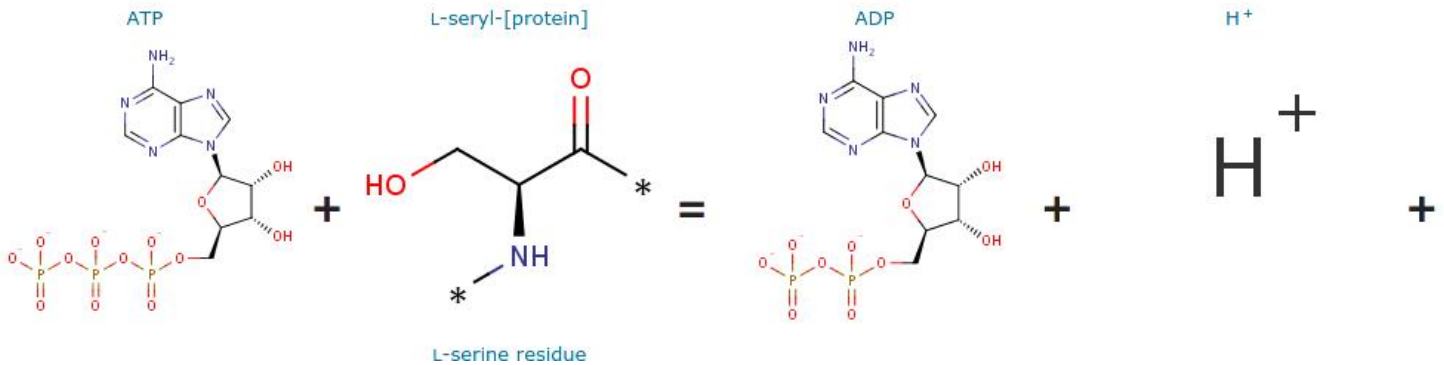
Constitutes the pattern-recognition receptor (PPR) that determines the specific perception of flagellin (flg22), a potent elicitor of the defense response to pathogen-associated molecular patterns (PAMPs). Recognizes flg22 from Pseudomonas aeruginosa and Acidovorax avenae. flg22 is a peptide derived from the bacterial flagellin N-terminus sequence (PubMed:18986259, PubMed:25617720). Does not recognize flg22 from Xanthomonas oryzae pv. oryzae (Xoo) or Xanthomonas oryzae pv. oryzicola (Xoc) (PubMed:25617720). 2 Publications

Catalytic activity

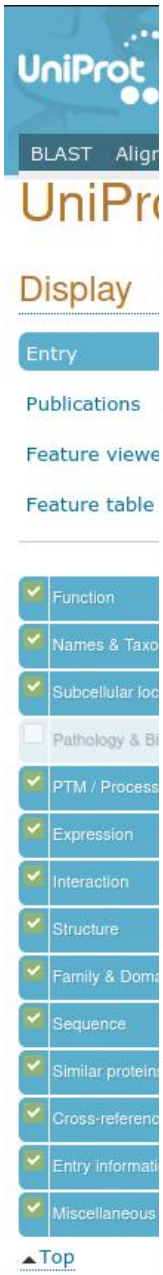
- ATP + L-seryl-[protein] = ADP + H^+ + *O*-phospho-L-seryl-[protein]

  EC:2.7.11.1

  Source: Rhea. « Hide



Banque de séquences protéiques : UniProt

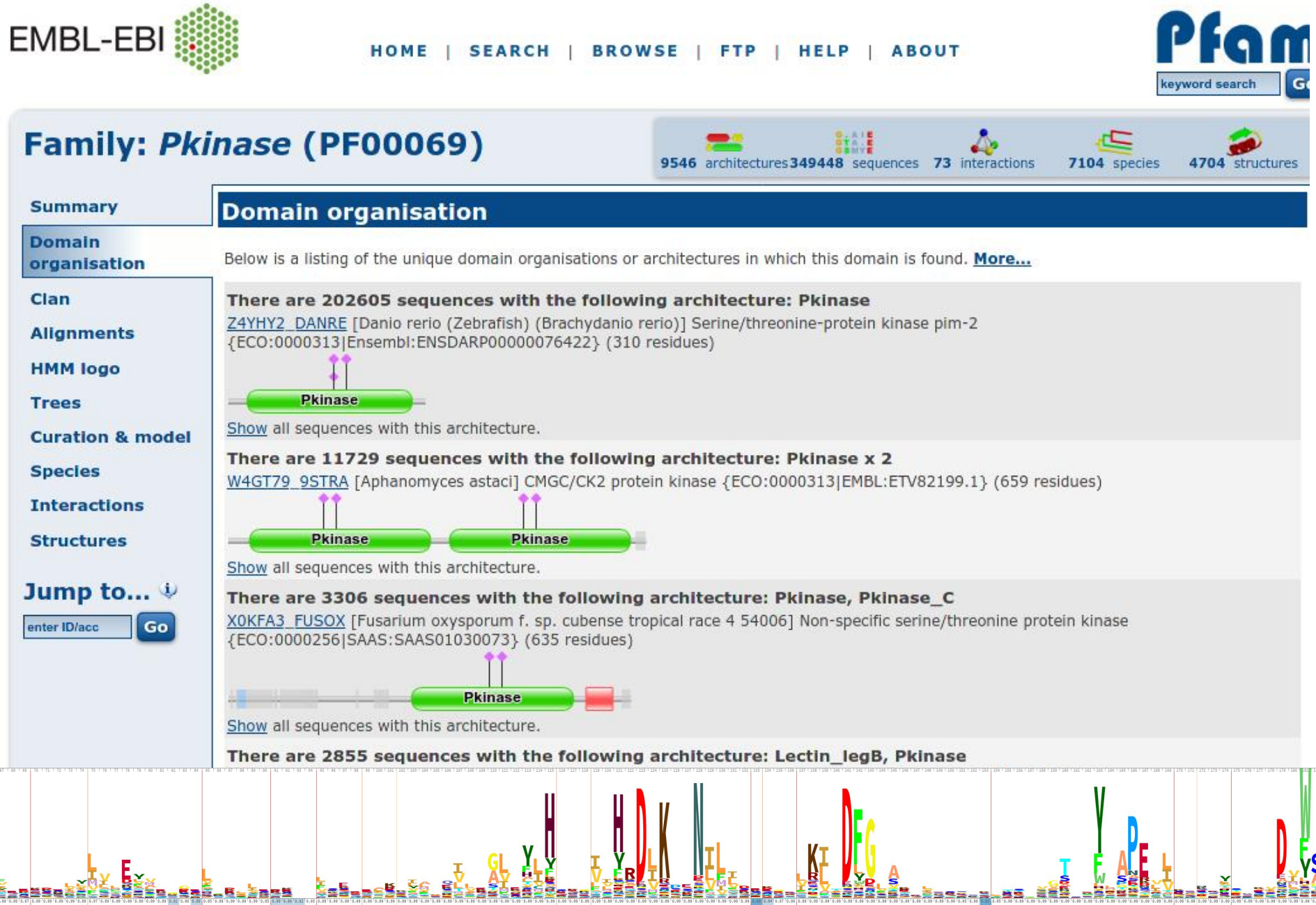



```
ID   FLS2_ORYSJ              Reviewed;        1183 AA.
AC   Q0JA29; A3AXG7; Q7XS37;
DT   30-AUG-2017, integrated into UniProtKB/Swiss-Prot.
DT   03-OCT-2006, sequence version 1.
DT   11-DEC-2019, entry version 107.
DE   RecName: Full=LRR receptor-like serine/threonine-protein kinase FLS2 {ECO:0000305};
DE            EC=2.7.11.1 {ECO:0000305};
DE   AltName: Full=Protein FLAGELLIN-SENSING 2 homolog {ECO:0000305};
DE            Short=OsFLS2 {ECO:0000303|PubMed:18986259};
DE   AltName: Full=Protein FLAGELLIN-SENSITIVE 2 homolog {ECO:0000305};
DE   Flags: Precursor;
GN   Name=FLS2 {ECO:0000303|PubMed:18986259};
GN   OrderedLocusNames=Os04g0618700 {ECO:0000312|EMBL:BAF15808.1},
GN   LOC_Os04g52780 {ECO:0000305};
GN   ORFNames=OsJ_16186 {ECO:0000312|EMBL:EAZ32006.1};
OS   Oryza sativa subsp. japonica (Rice).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliopsida; Liliopsida; Poales; Poaceae; BOP clade;
OC   Oryzoideae; Oryzeae; Oryzinae; Oryza; Oryza sativa.
OX   NCBI_TaxID=39947;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=cv. Nipponbare;
RX   PubMed=12447439; DOI=10.1038/nature01183;
RA   Feng Q., Zhang Y., Hao P., Wang S., Fu G., Huang Y., Li Y., Zhu J., Liu Y.,
RA   Hu X., Jia P., Zhang Y., Zhao Q., Ying K., Yu S., Tang Y., Weng Q.,
RA   Zhang L., Lu Y., Mu J., Lu Y., Zhang L.S., Yu Z., Fan D., Liu X., Lu T.,
RA   Li C., Wu Y., Sun T., Lei H., Li T., Hu H., Guan J., Wu M., Zhang R.,
RA   Zhou B., Chen Z., Chen L., Jin Z., Wang R., Yin H., Cai Z., Ren S., Lv G.,
RA   Gu W., Zhu G., Tu Y., Jia J., Zhang Y., Chen J., Kang H., Chen X., Shao C.,
RA   Sun Y., Hu Q., Zhang X., Zhang W., Wang L., Ding C., Sheng H., Gu J.,
RA   Chen S., Ni L., Zhu F., Chen W., Lan L., Lai Y., Cheng Z., Gu M., Jiang J.,
RA   Li J., Hong G., Xue Y., Han B.;
RT   "Sequence and analysis of rice chromosome 4.";
RL   Nature 420:316-320(2002).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=cv. Nipponbare;
```

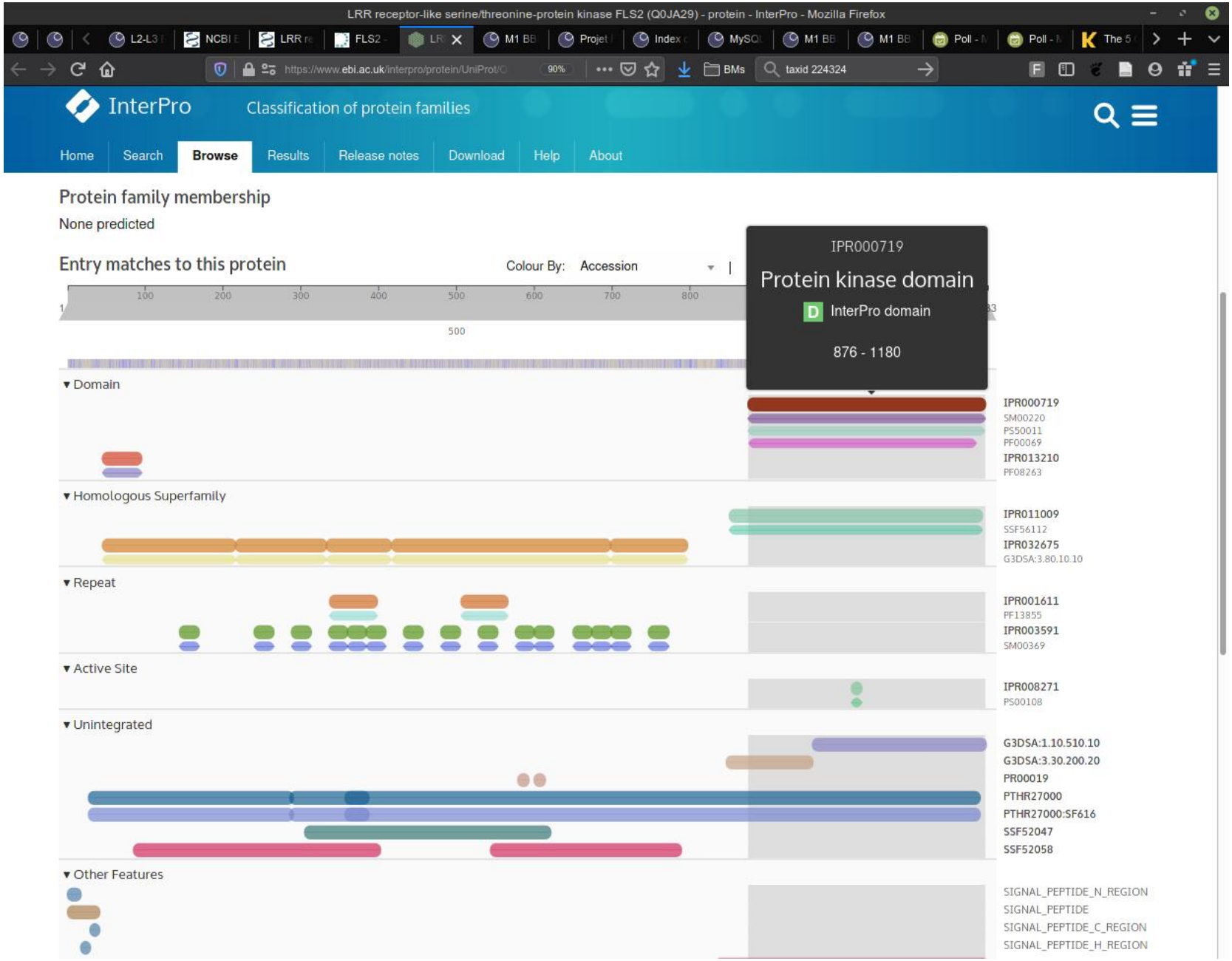
Familles de protéines et de domaines protéiques

- Pression de sélection → conservation de séquence
- Totalité de la séquence mais plus souvent certaines régions → domaines proteiques
  - ex : domaine de liaison à l'ADN, domaine d'interaction avec une autre protéine
- Banques de domaines protéiques
- Méthode de détection des domaines présents sur une séquence donnée

Banques de domaines



Détection de domaines



Transcriptome

Transcriptome : ensemble des ARNm ou transcrits présents dans une population de cellules dans des conditions données.

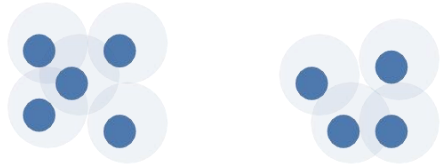
➡ Accès au niveau d'expression de milliers de gènes simultanément (potentiellement l'ensemble des gènes d'un organisme)
= *instantané* de l'état d'une cellule ou d'une population de cellules

Données d'expression des gènes obtenues par :

- qPCR
- Puces à ADN
- Séquençage ultra-haut débit

Transcriptome : acquisition des données (RNAseq)

Échantillon 1 Échantillon 2

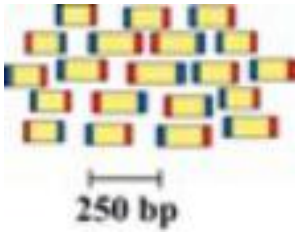


extraction

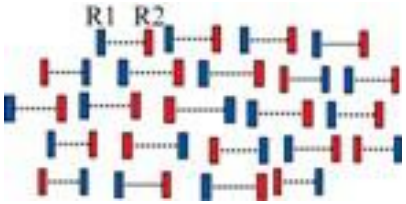
ARNm

Préparation de la librairie
(sélection des fragments,
ajout des linkers, …)

ADNc

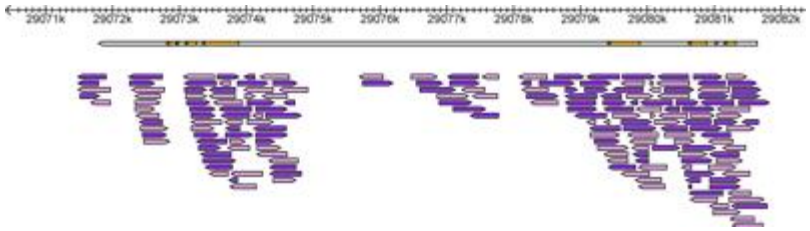


séquençage



(centaines de)
millions de séquences
(reads)

Traitement des données
(assemblage, alignement
sur le génome,
normalisation…)

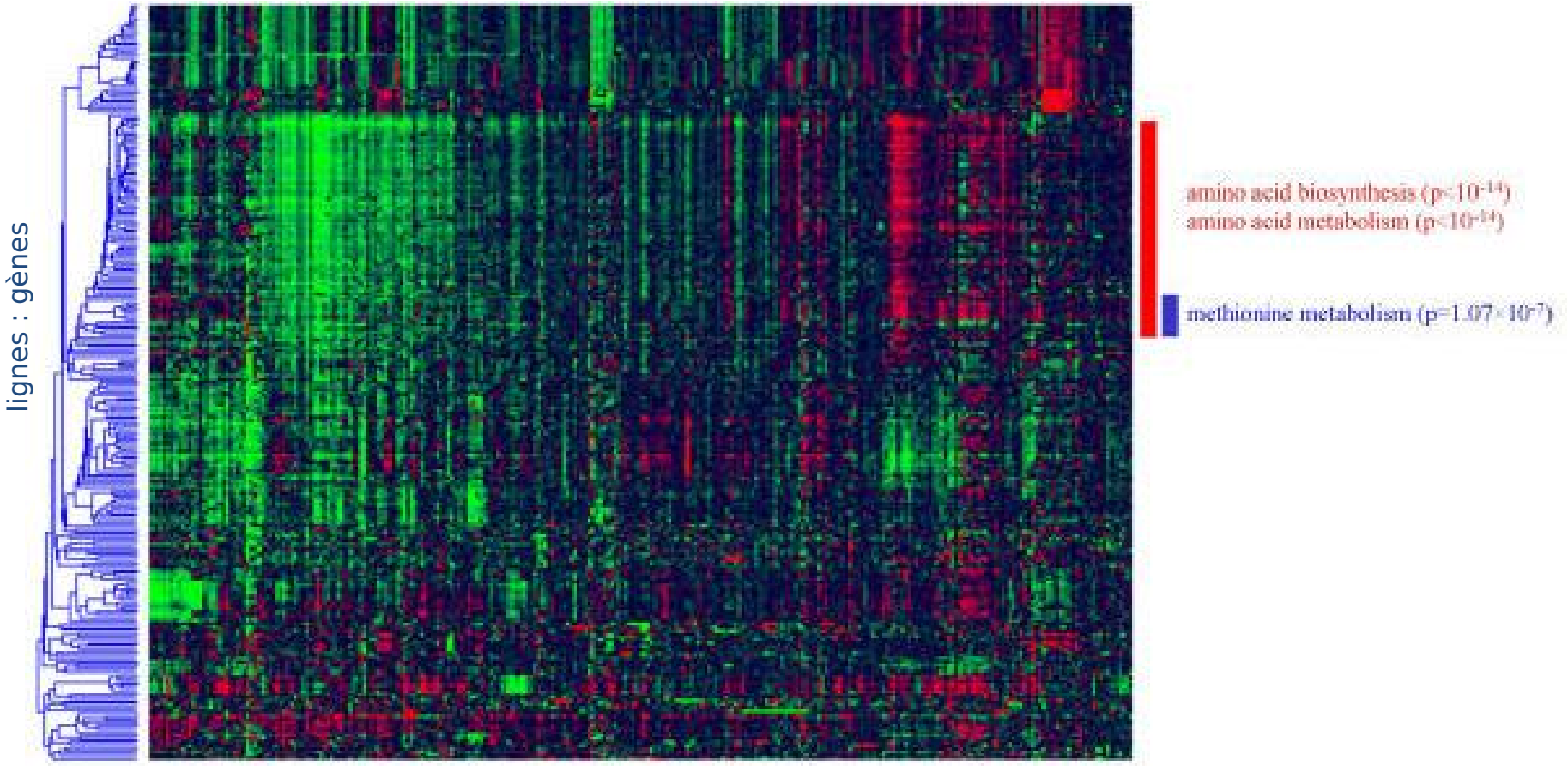


Transcriptome : gènes co-exprimés

- Motivation : les gènes ayant des profils d'expression similaires sont potentiellement co-régulés et participeraient donc à un même processus biologique

- But : regrouper les gènes impliqués dans un même processus biologique

colonnes : condition expétimentales (ex: mutants)

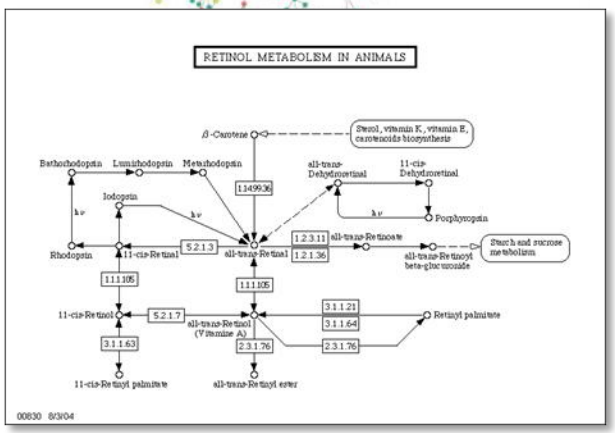


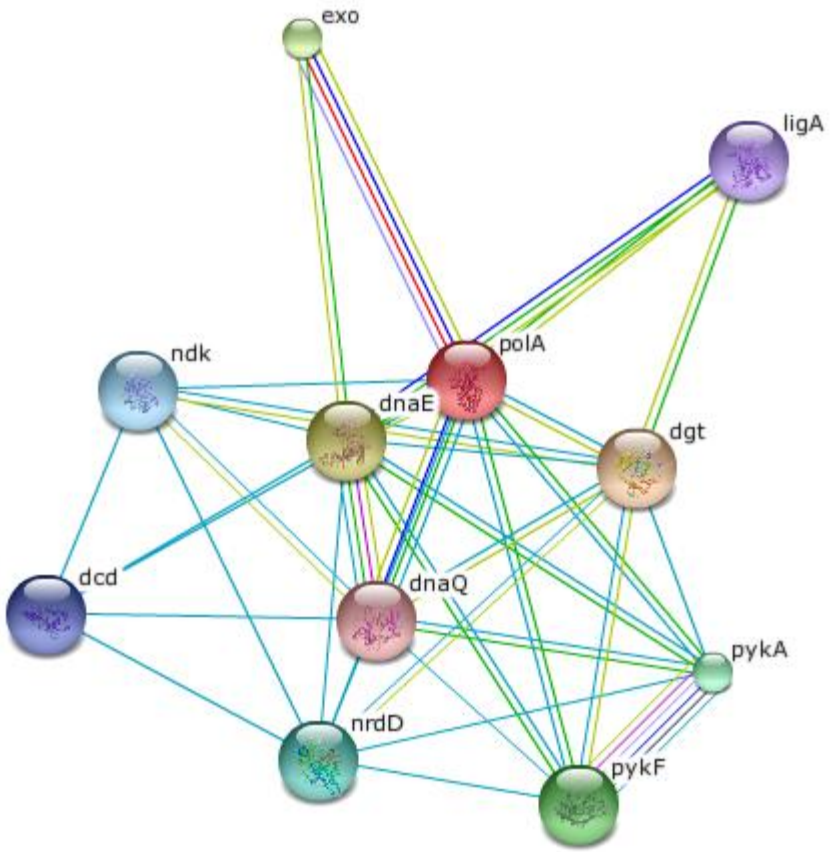
Robinson *et al.* 2002

- Métagénomique
  - Echantillon provenant de l'environnement
- ADN ancien · Paléogénomique
- Police scientifique
- Variant de virus
- Biologie intégrative
  - Genome complet
  - Expression
  - Epigénétique (méthylation, modification des histones)
  - Conformation 3D des chromosomes *in vivo*

Réseaux de gènes et de protéines

Réseaux :
- d'interactions protéine - protéine, génétiques, fonctionnelles, …
- de régulation des gènes
- métabolisme (enzymes – substrats)
- transduction du signal



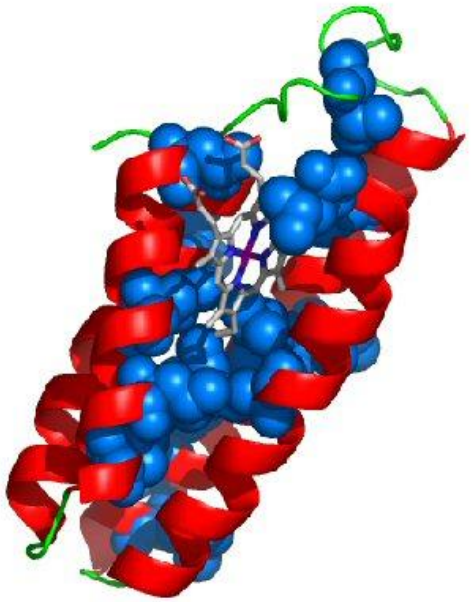


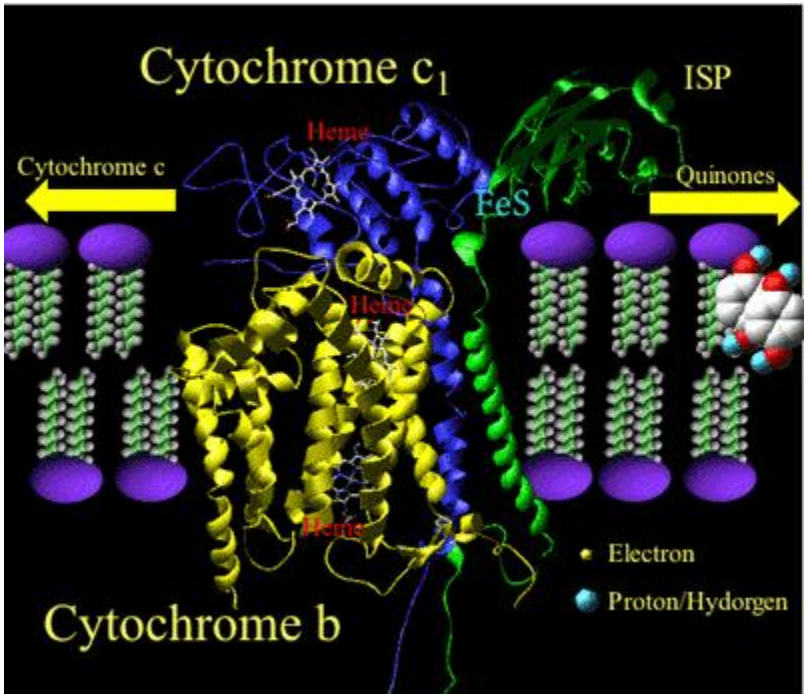
Prédiction de structure

Séquence protéique

```
>gi|5524211|gb|AAD44166.1| cytochrome b
 LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

Prédiction ou résolution de la structure tridimensionnelle

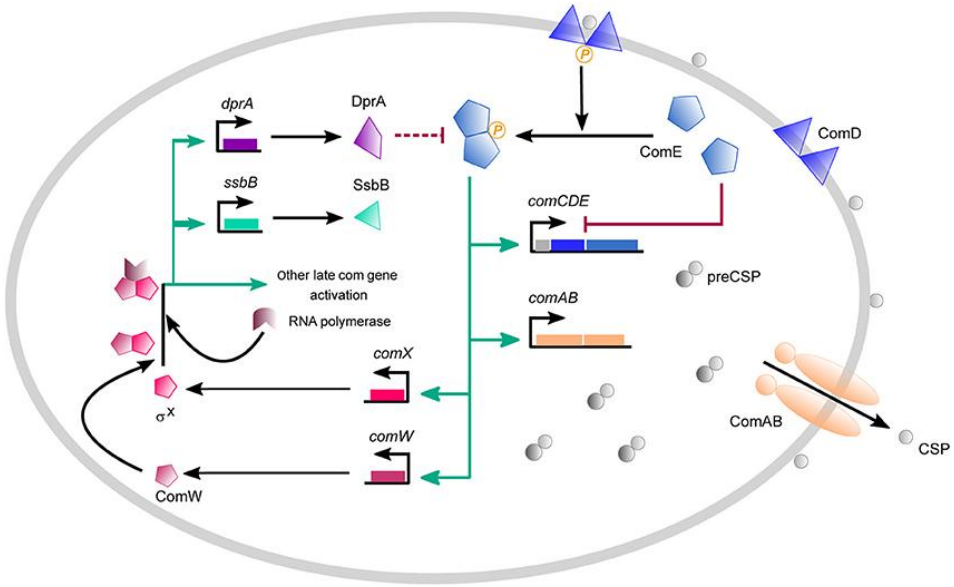


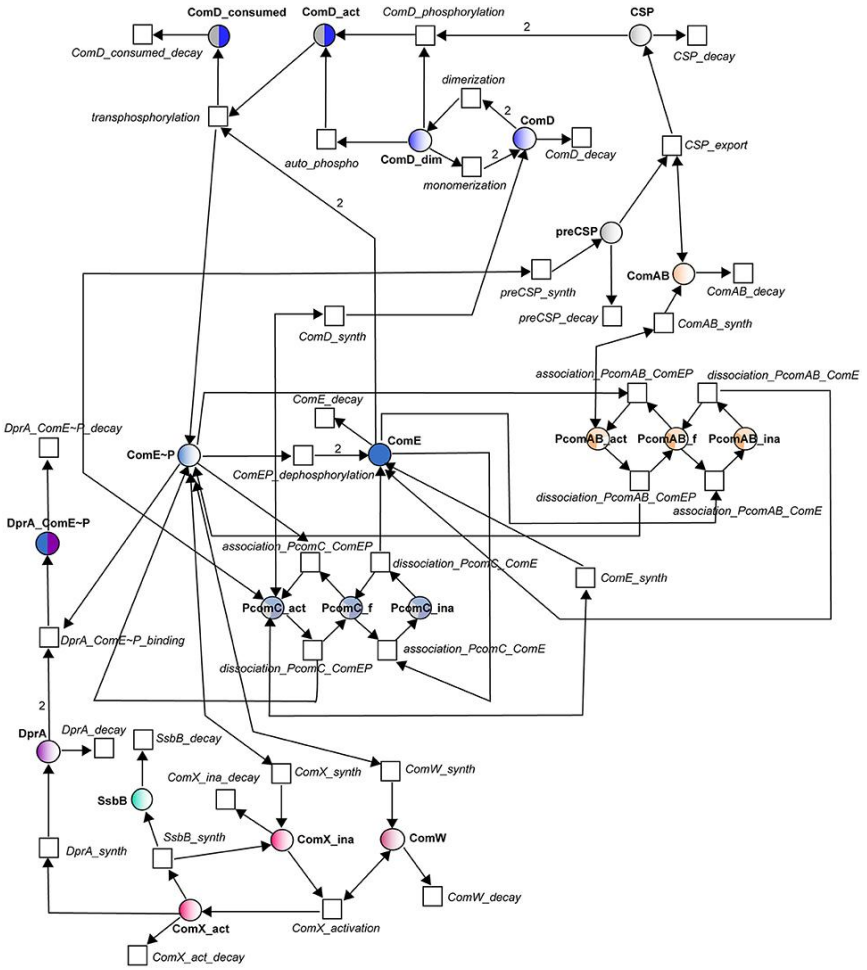


Biologie des systèmes

Intégration et synthèse des connaissances

- modélisation d'un système
  - processus biologique (respiration)
  - organite (mitochondrie)
  - cellule
  - population
  - écosystème





A terme : simulation d'une cellule/organisme/population/écosystème et prédiction de son comportement

Des observations au modèle

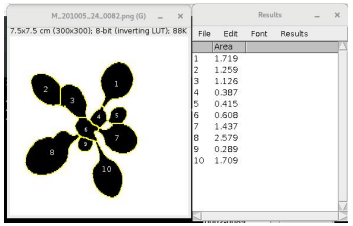
- observations → mesures
- stockage → élaboration d'un schéma de base de données et création de la base de données
- analyses statistiques + recherche d'informations complémentaires (Web)

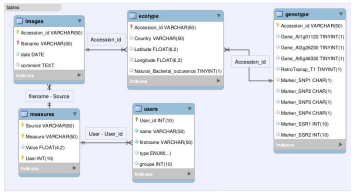
  → interprétation
- modélisation → modèle mathématique (réseau de pétri)
- hypothèses et simulations

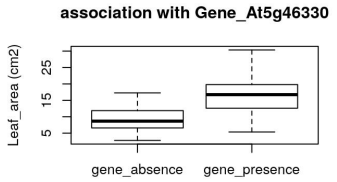
Concepts et méthodes abordés dans cette UE

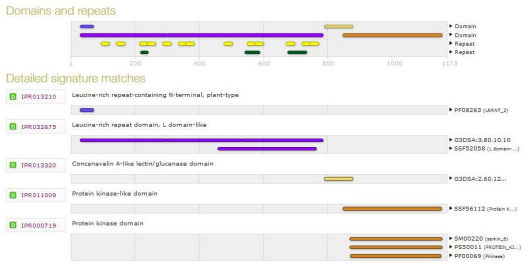
- Traitement d'images
  - introduction au traitement d'images
  - segmentation
  - **macros**
- Bases de données
  - introduction aux bases de données
  - conception d'un schéma de base de données
  - utilisation (**requêtes SQL**)
- Traitement de données
  - des statistiques pour des questions biologiques
  - environnement **R**
  - graphiques - corrélation - tests statistiques
- Bioanalyse
  - apperçu de ressources disponibles
  - **banques** de données publiques (ex: UniProt)

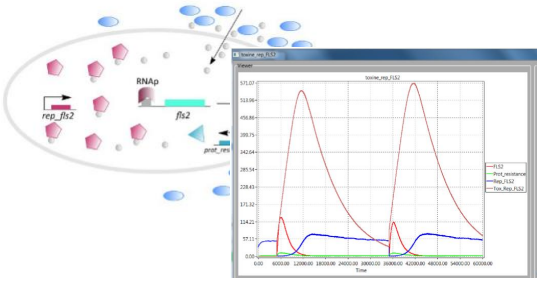
  + serveurs d'**analyses** (ex: BLAST)
  - annotations existantes, recherche de domaines sur une séquence, recherche de séquences similaires
  - synthèse des connaissances et observations → esprit critique
- Modélisation
  - introduction à la biologie des systèmes et à la modélisation
  - validation ou remise en question d'une hypothèse / prédictio
  - propriétés émergentes
  - **réseau de pétri**



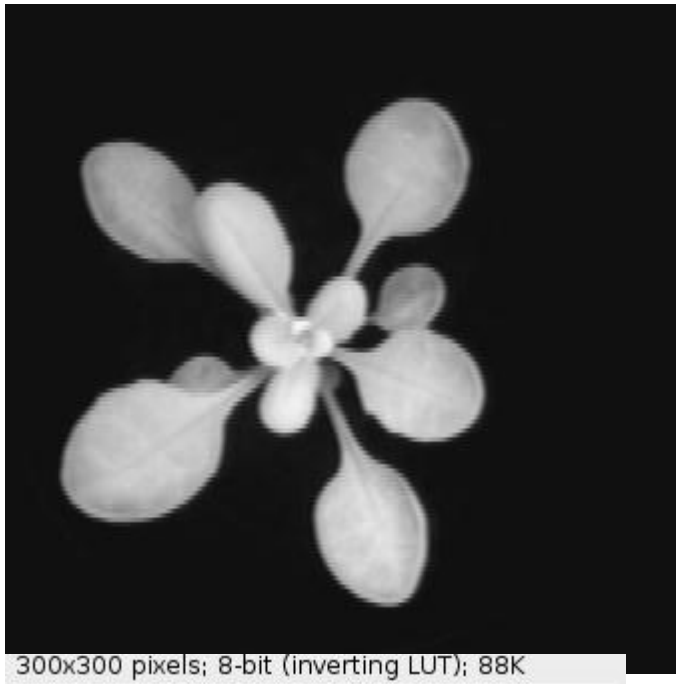


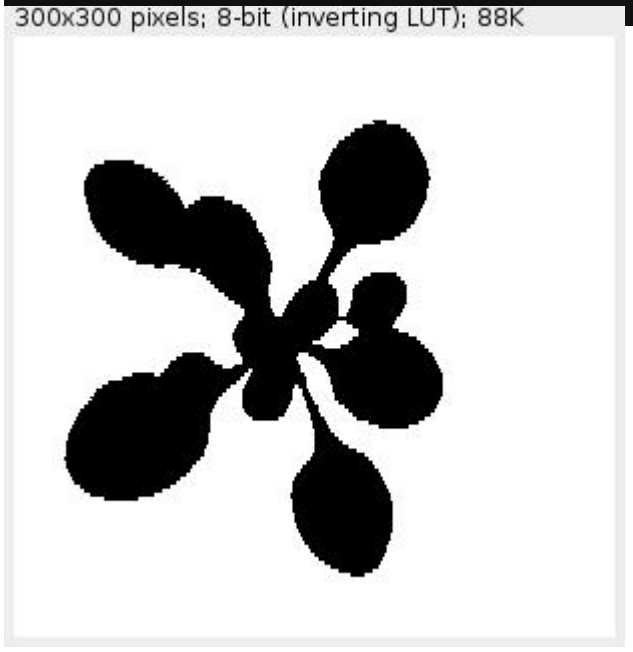


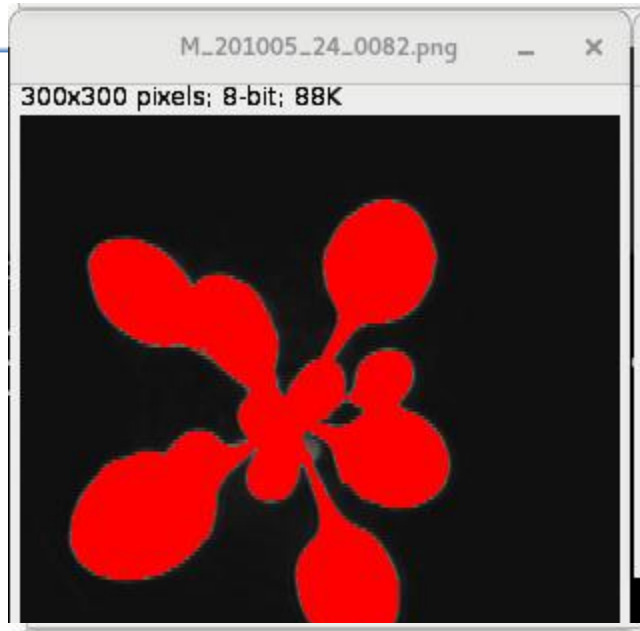


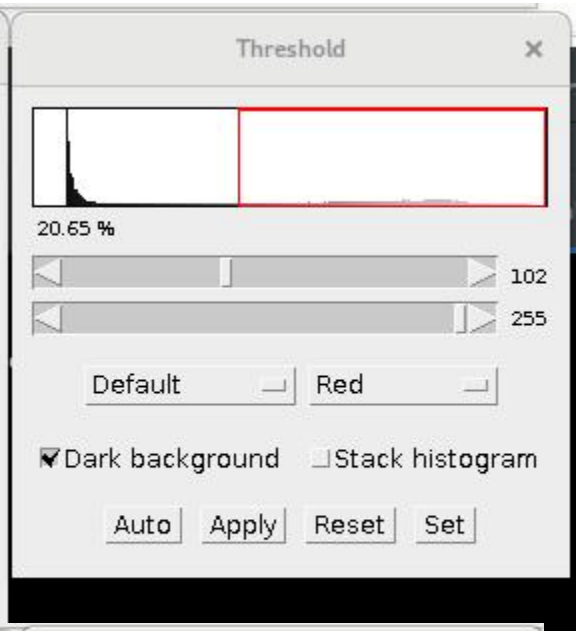


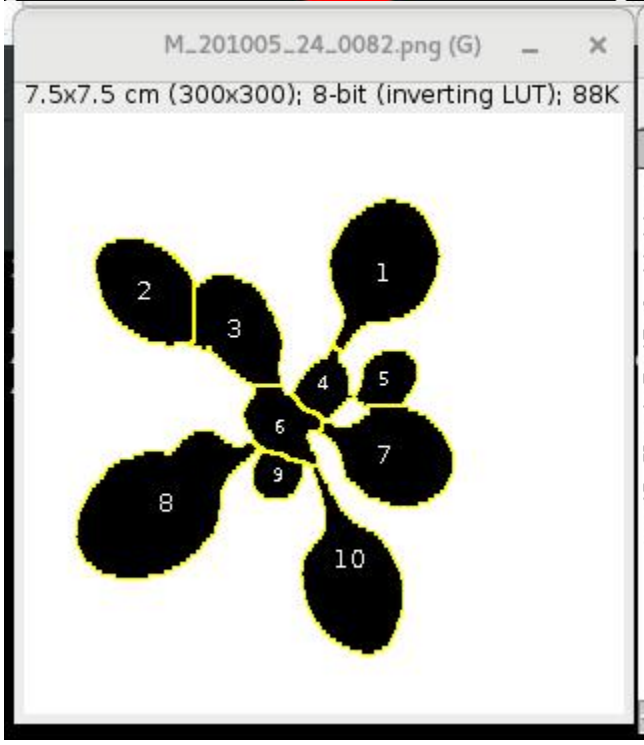
Traitements d'image

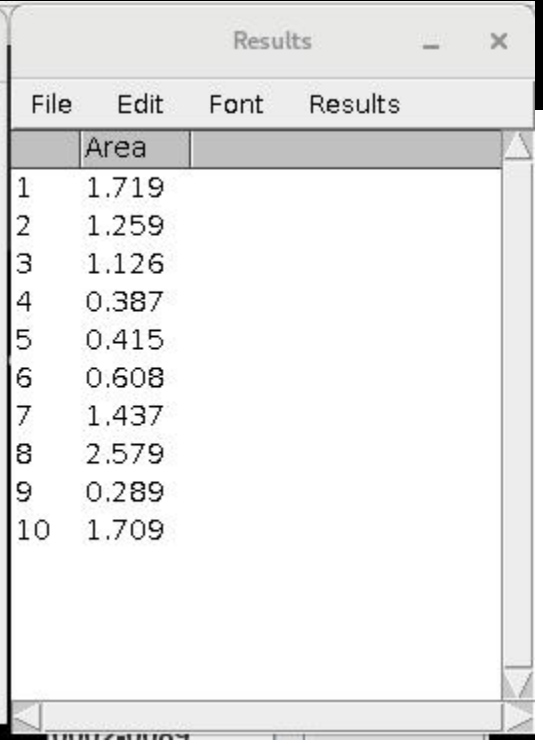




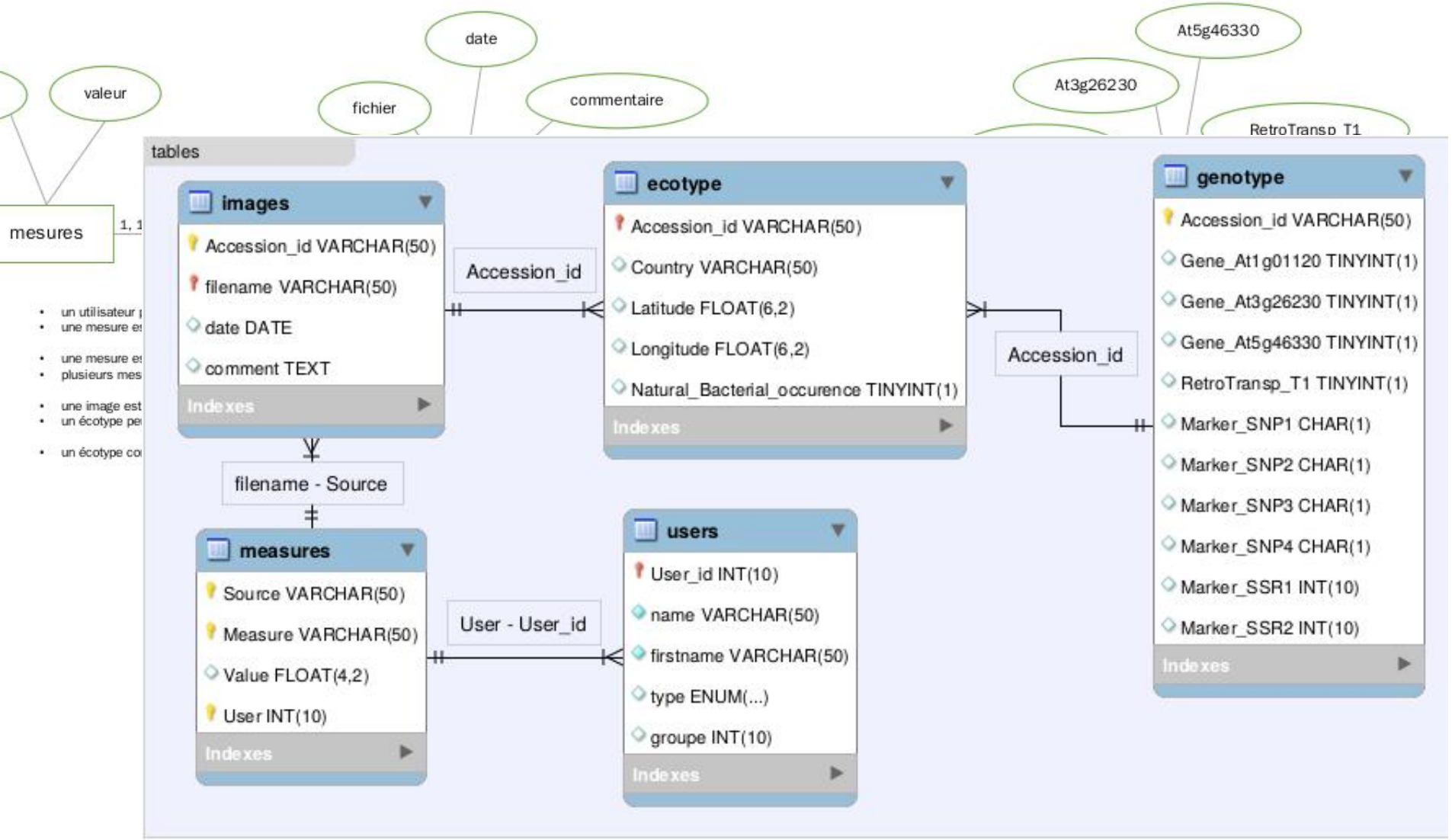






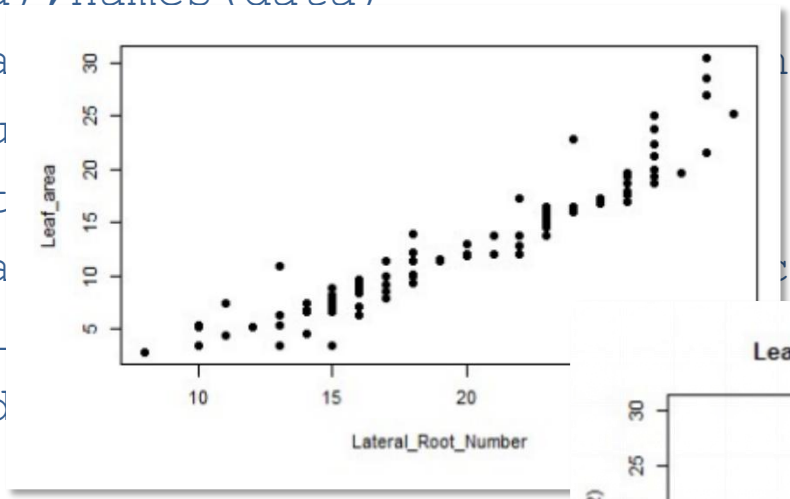


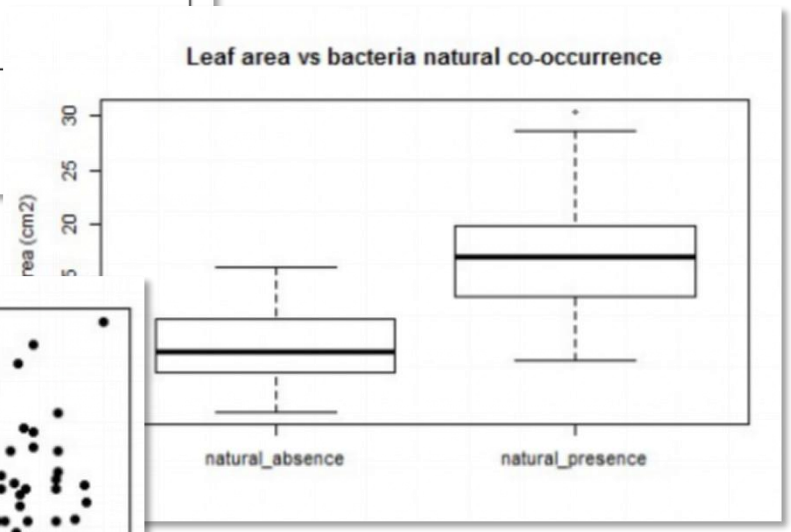
Bases de données

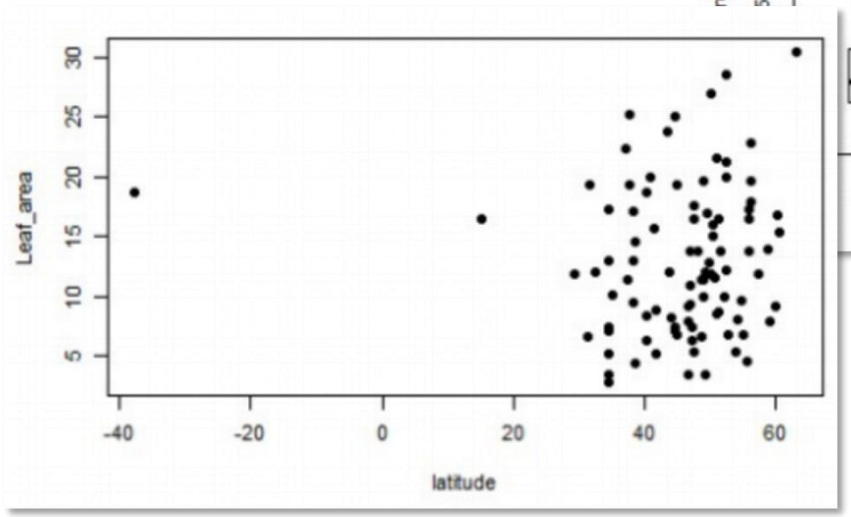


Traitement de données - statistiques

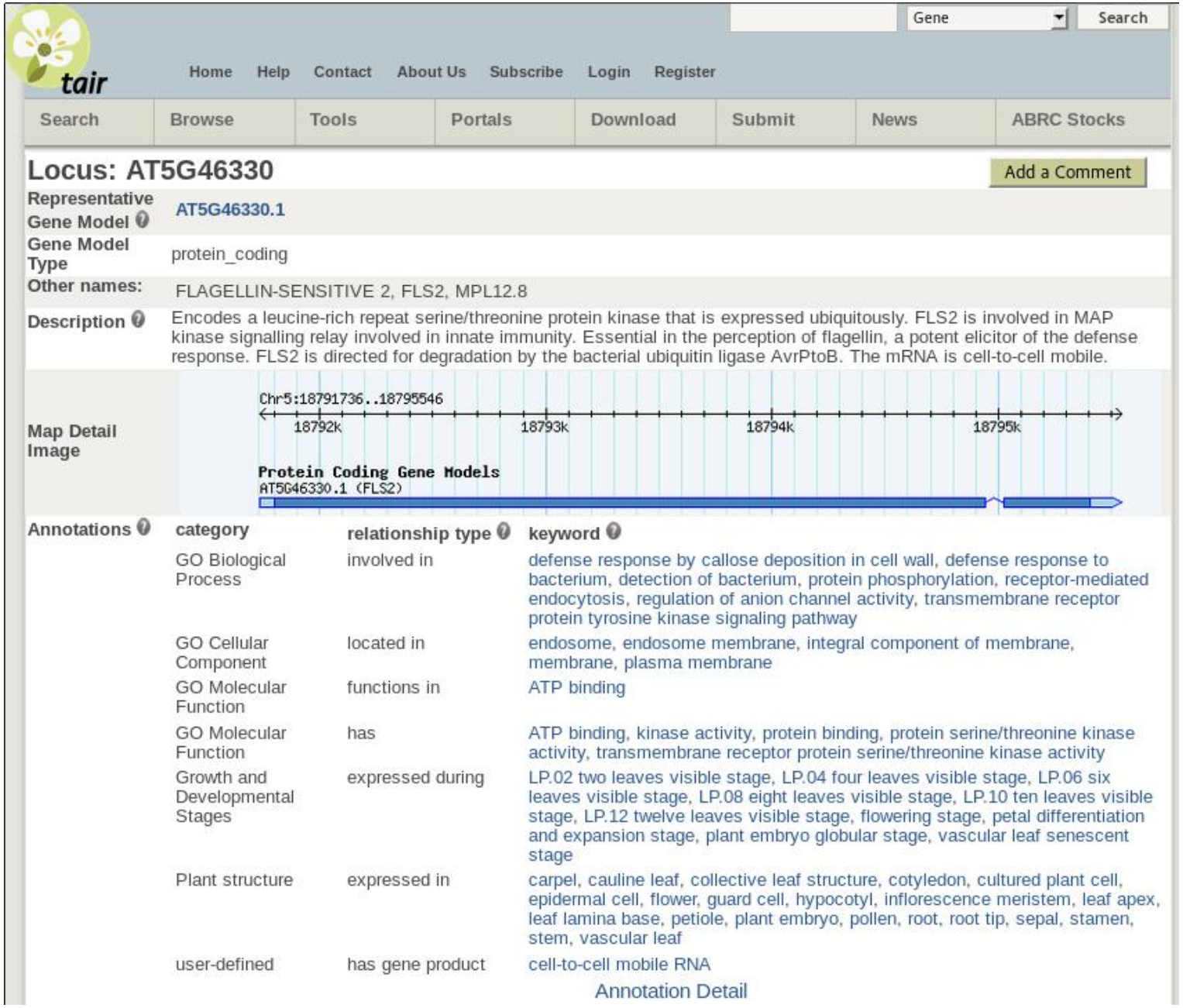
```
data=read.table("data_for_TP3.txt",h=T)
attach(data);names(data)
plot(Latera                    =16)
plot(Latitu
plot(Longit
boxplot(Lea                    curence)
cor(Lateral
cor(Latitud
```



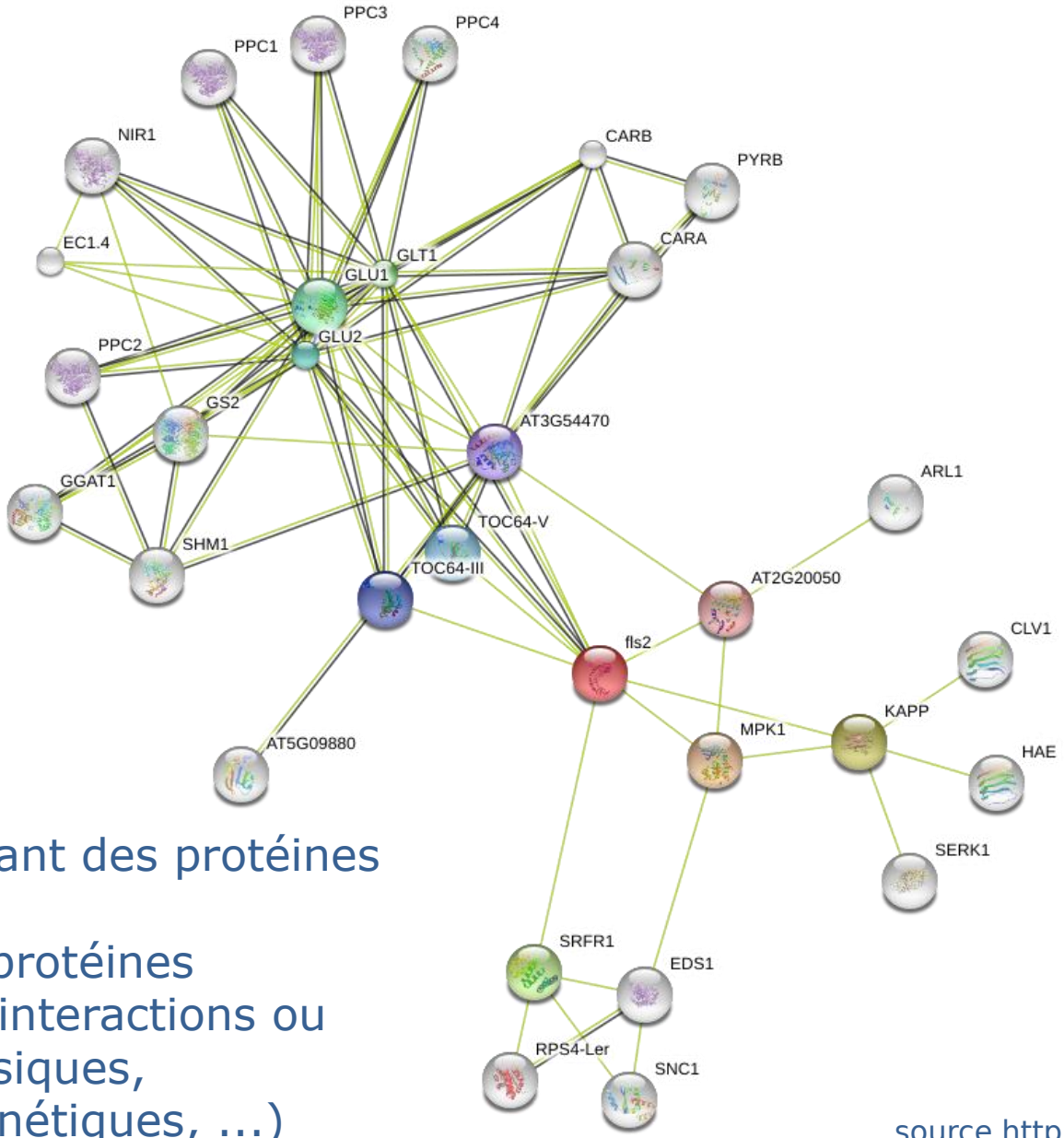




Banques et sites Web



Banques et sites Web

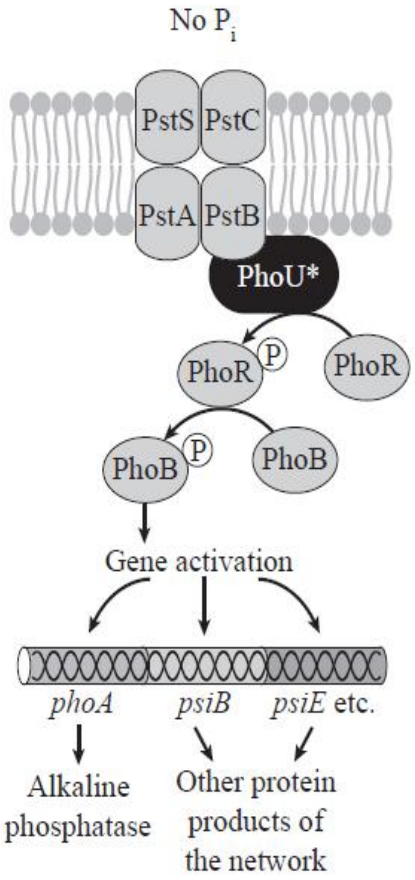


Graphe représentant des protéines

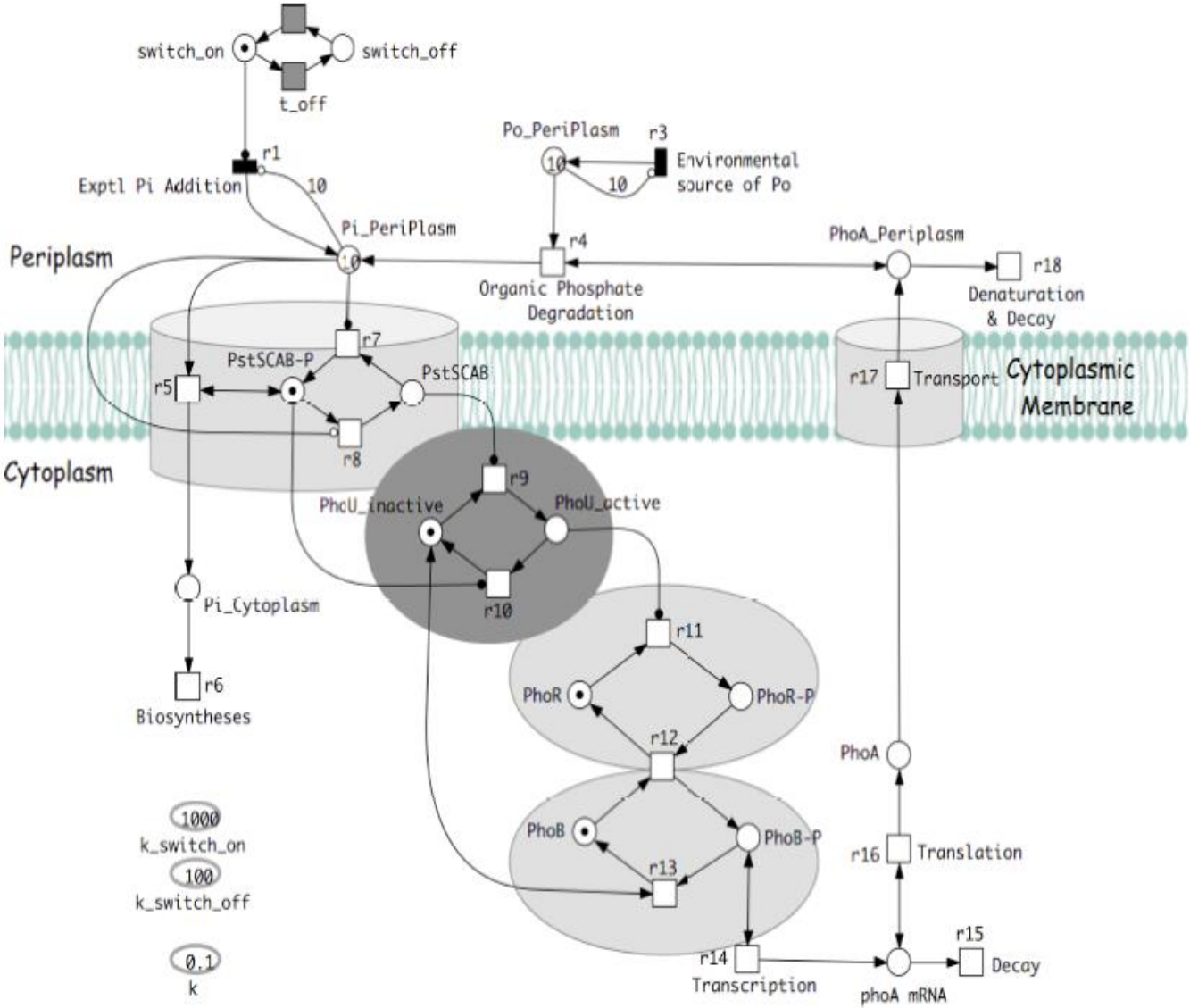
Les liens entre 2 protéines représentent des interactions ou associations (physiques, fonctionnelles, génétiques, ...)

source http://string-db.org

Modélisation de systèmes biologiques



Neidhardt *et al.* 1990



Durzinsky *et al.*, 2011

Homologie, orthologie et orthologie 1:1

Homologie : deux gènes sont homologues s'ils ont divergé à partir d'une même séquence ancêtre

Orthologie

- définition : deux gènes sont ortholgues si leur divergence a commencé après un évènement de **spéciation** (le gène ancêtre se trouvait dans l'organisme ancêtre).
- remarque : deux gènes ortologues ne peuvent pas être présents dans le même génome.
- La relation d'orthologie est souvent abusivement utilisée pour déterminer la présence ou l'absence d'un gène dans un génome.

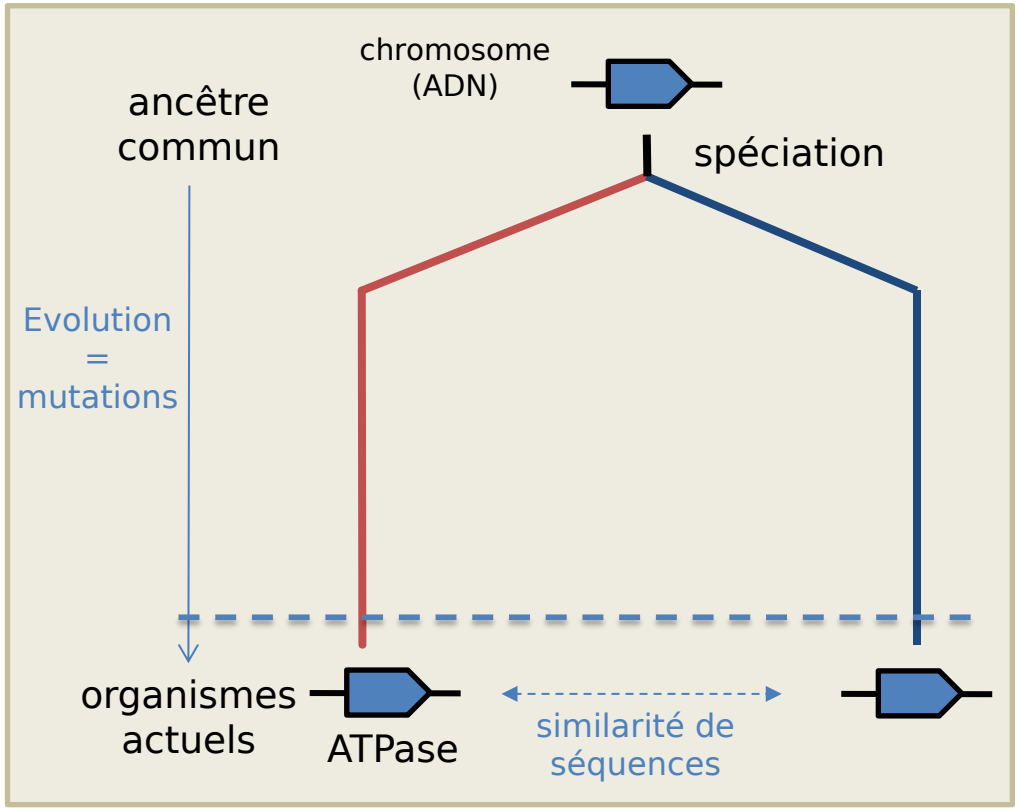
Paralogie

- définition : deux gènes sont paralogues si leur divergence a commencé après la **duplication** du gène ancêtre.
- hypothèse : mécanisme évolutif d'acquisition de nouvelles fonctions par accumulation de mutations sur une des deux "copies" du gène ancêtre non soumis à une pression de sélection ?
- remarque : les deux gènes paralogues sont au départ dans le même génome.

Orthologie 1:1

- définition : deux gènes orthologues sont orthologues 1:1 s'il n'y a pas eu de duplication (apparition de paralogue) après l'évènement de spéciation.
- remarque : la chronologie des évènements de spéciation et de duplication est importante.
- Deux orthologues 1:1 ont une forte probabilité d'avoir conservé la même fonction.

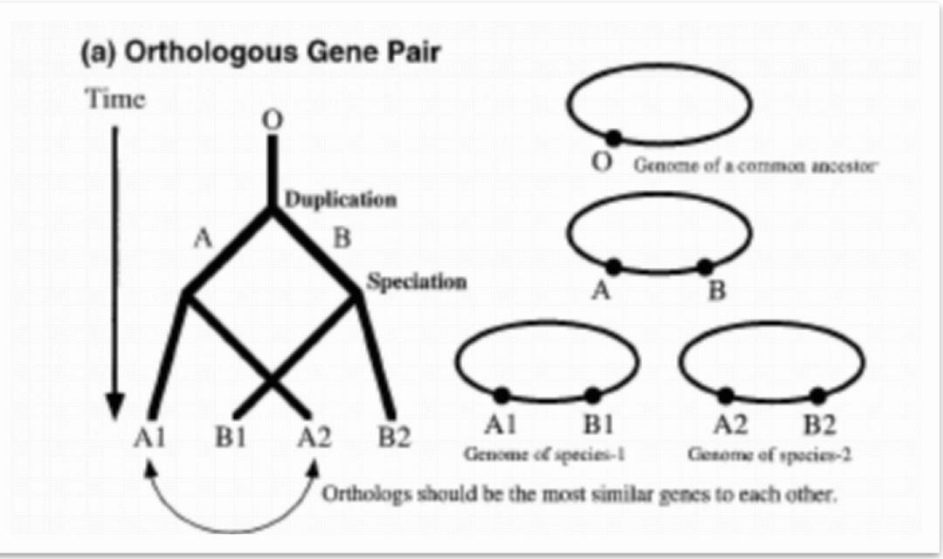
Homologie et transfert d'annotation

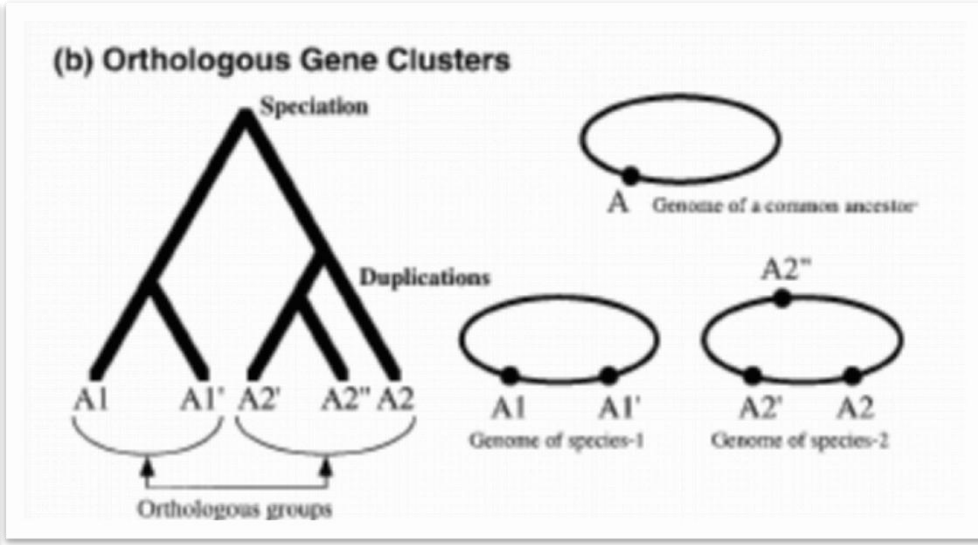


Du temps du séquençage des premiers gènes et génomes

- la conservation/similarité des séquences impliquait des fonctions similaires
- les annotations des gènes caractérisés expérimentalement étaient transférées aux nouveaux gènes/génomes séquencés

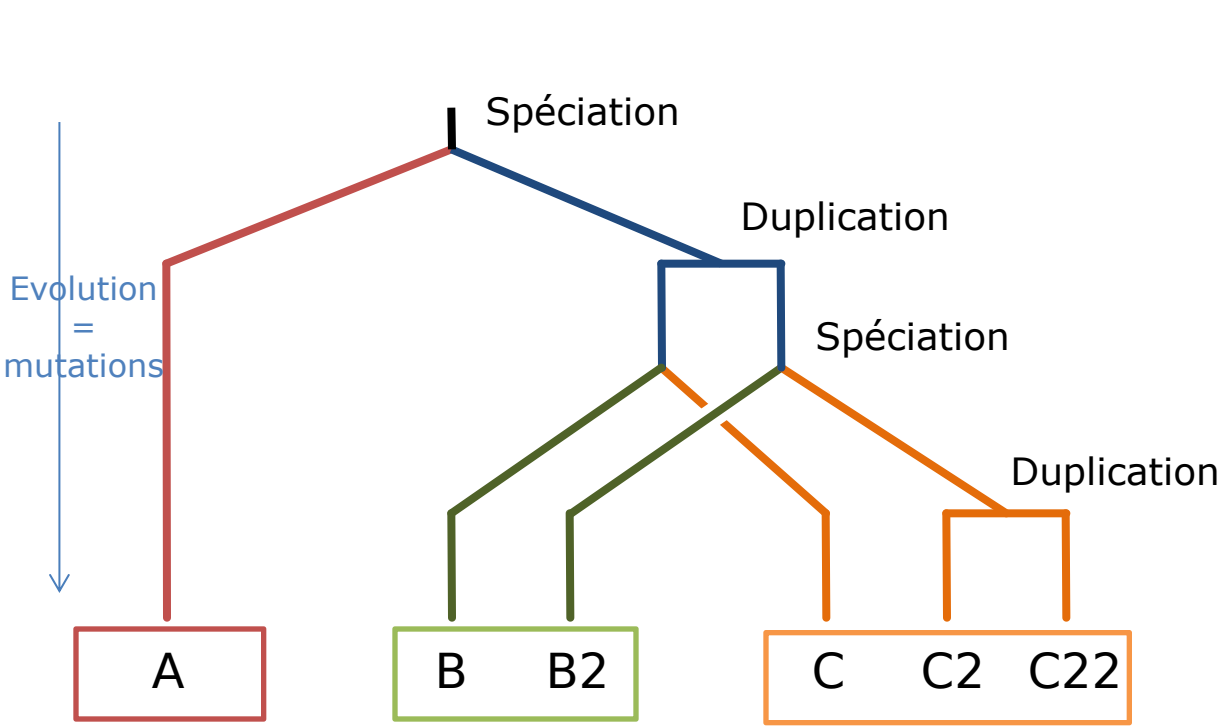
Chronologie des évènements



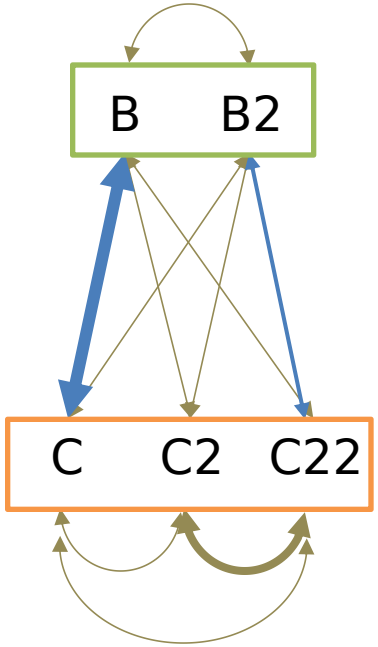


- L'orthologie 1:1 est essentielle pour la prédiction de la fonction des gènes (par transfert d'annotations)
- Problèmes :
  - la chronologie est nécessaire pour différencier paralogues, orthologues et orthologues 1:1
  - la chronologie n'est pas connue
- Deux principales approches pour identifier les orthologues 1:1
  - analyse et reconstruction de l'histoire évolutive
  - représentation sous forme de graphe des relations d'orthologie

Prédiction des orthologues 1:1

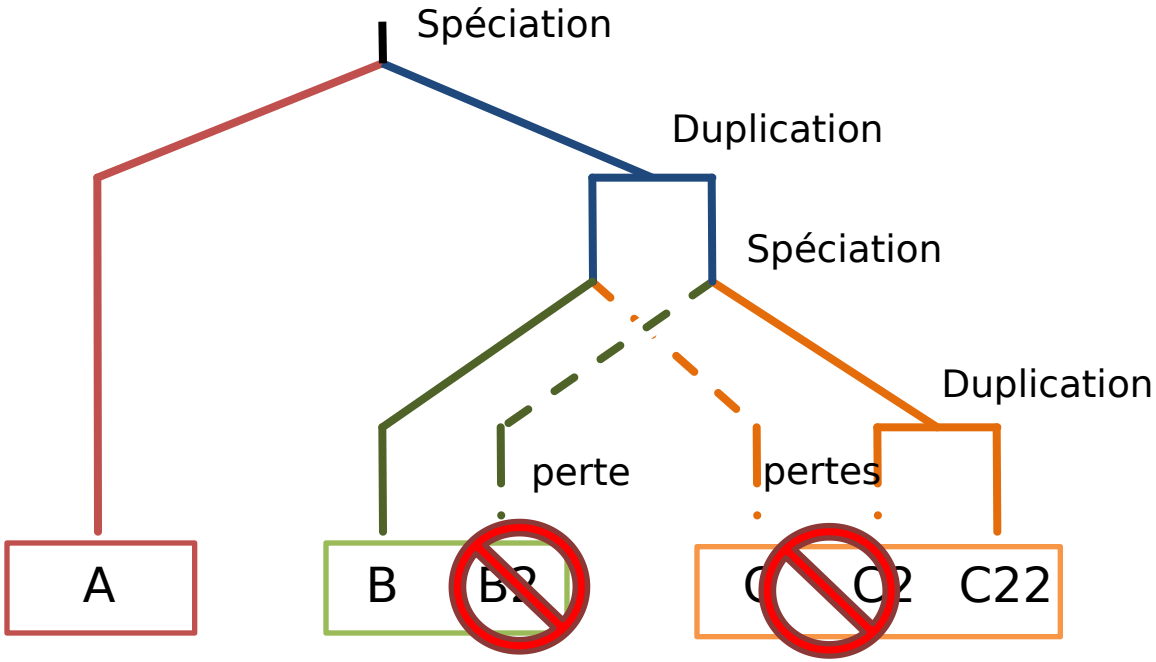


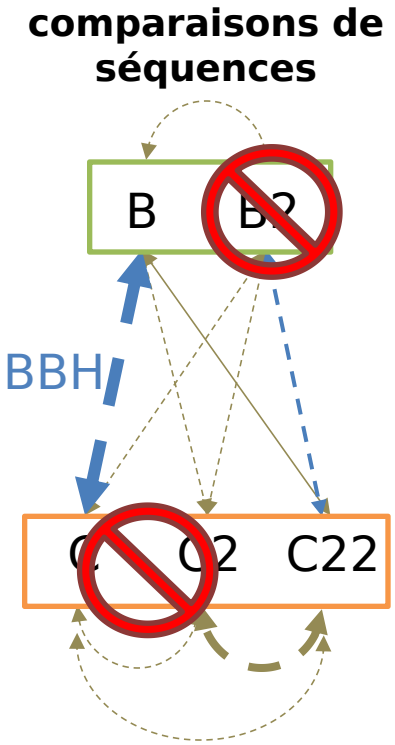
comparaisons de séquences



- comparaisons de toutes les séquences 2 à 2
- B↔C prédit sim(B,C) > max(sim(B,B2), sim(C, C2), sim(C,C22))
- B2↔C22 non prédit sim(B2,C22) < sim(C2,C22)
- Remarque : calcul avec 7,223,104 séquences de 2 355 génomes =>52,000.10^9 comparaisons

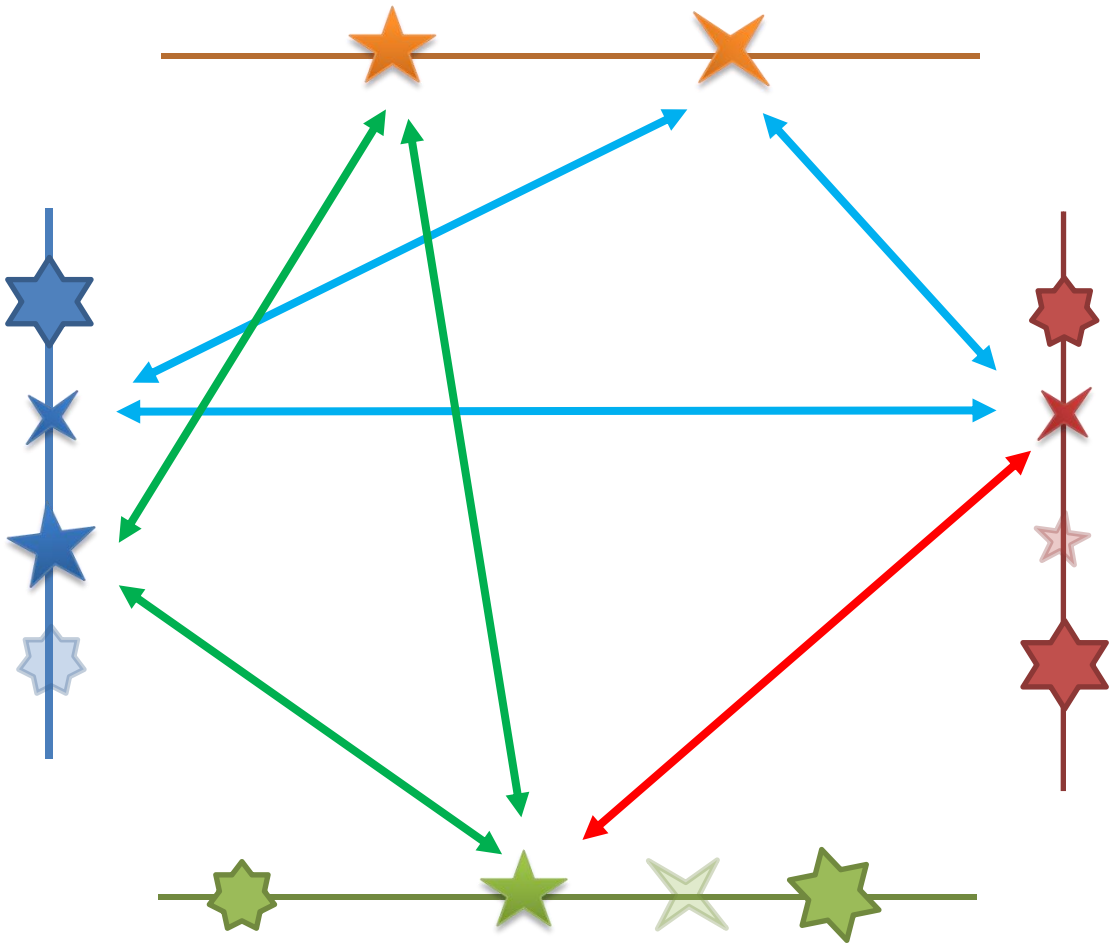
Autre évènement évolutif : perte de gènes





des séries de duplications et de pertes de gènes induisent des erreurs de prédiction

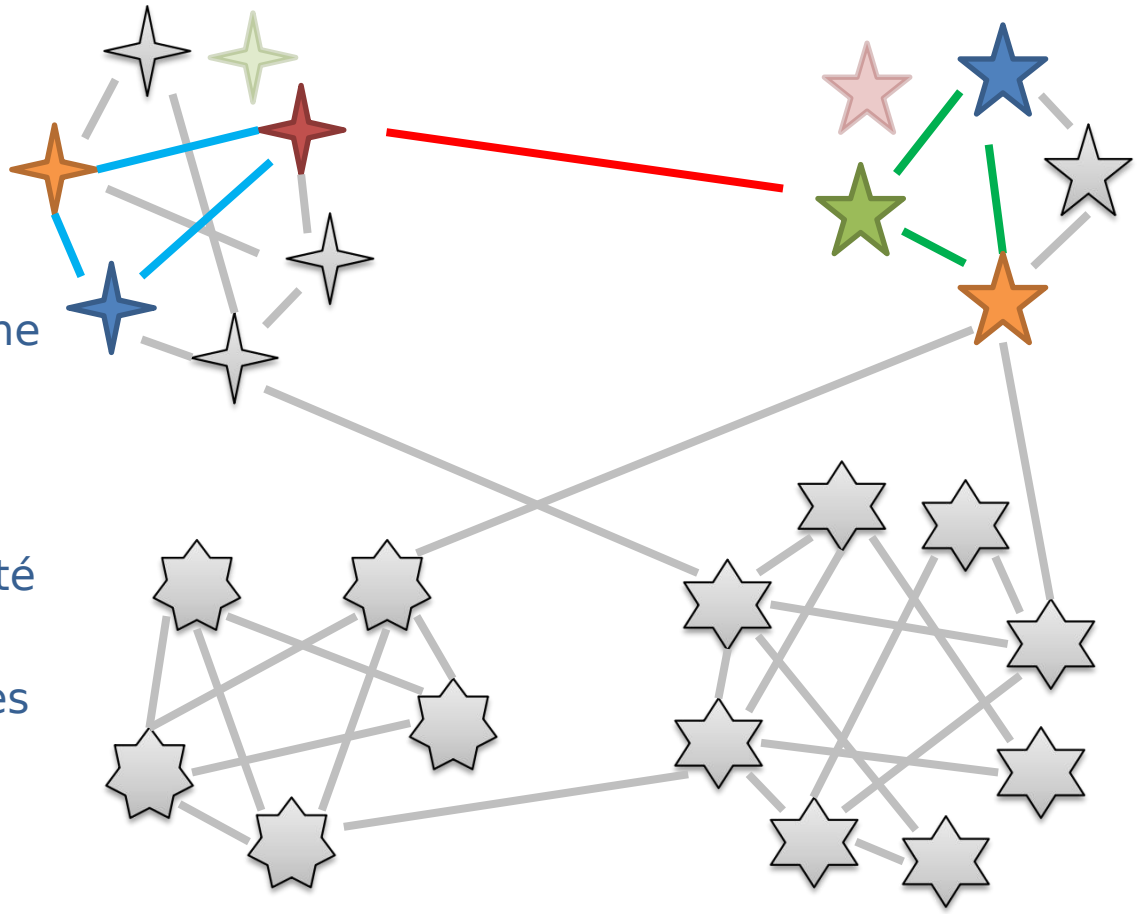
Transitivité de l'orthologie 1:1



Elagage des erreurs de prédiction

Méthodes de traitement de graphes

Répresentation: sommets = gènes, liens = orthologues 1:1



Partitionnement de graphe
=
clustering
=
détection de communauté

principe : les sommets très connectés entre eux sont isolés en cluster