

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%

1^{er} mars 13h45 (durée 1h30)

Contrôle terminal : 70%

12 avril 13h45 (durée 1h30)

Objectifs pédagogiques :

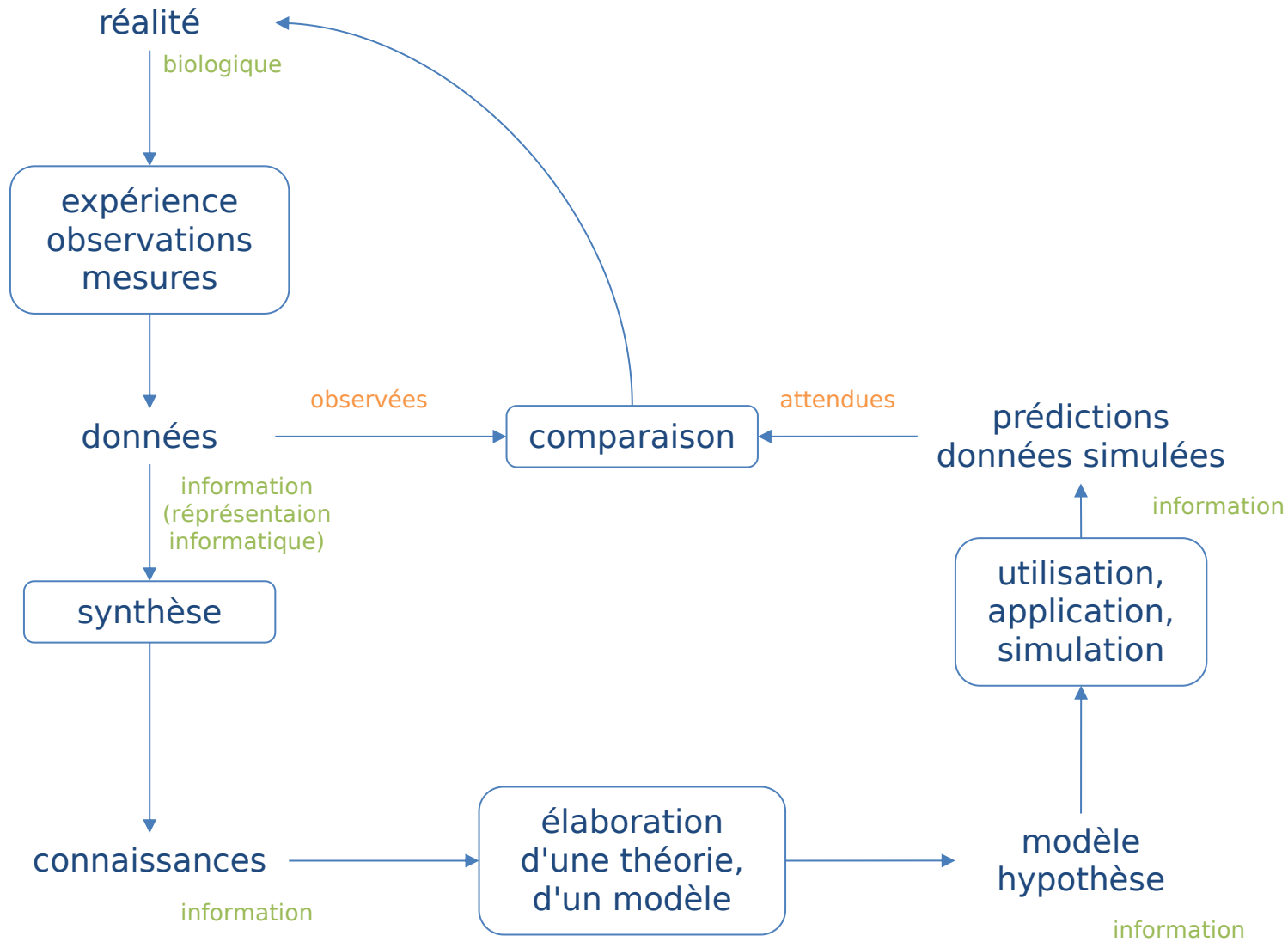
- Aperçu de quelques domaines d'application de la bioinformatique
 - traitement d'image : mesures phénotypiques
 - gestion des données
 - statistiques : corrélation phénotype - génotype
 - banques de données et sites Web publiques
 - (modélisation de systèmes biologiques et simulations)

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

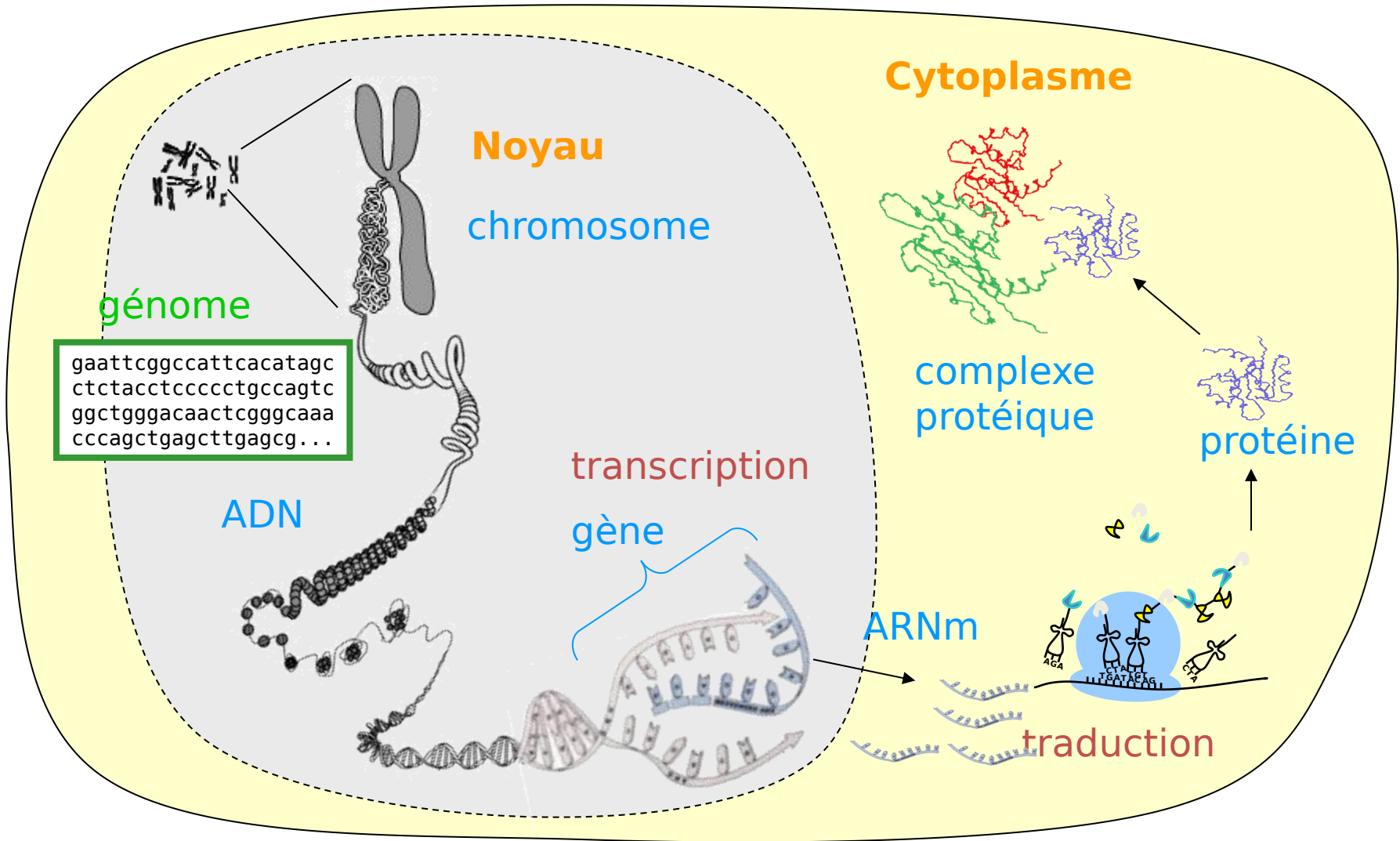
- Qu'est-ce que la bioinformatique ?
- Plusieurs réponses :
 - pas de l'informatique *bio*
 - traitement de l'information biologique
 - domaine spécialisé de la biologie
 - domaine de recherche multidisciplinaire : biologie, santé, mathématique, statistique, informatique, physique, éthique
- Une définition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

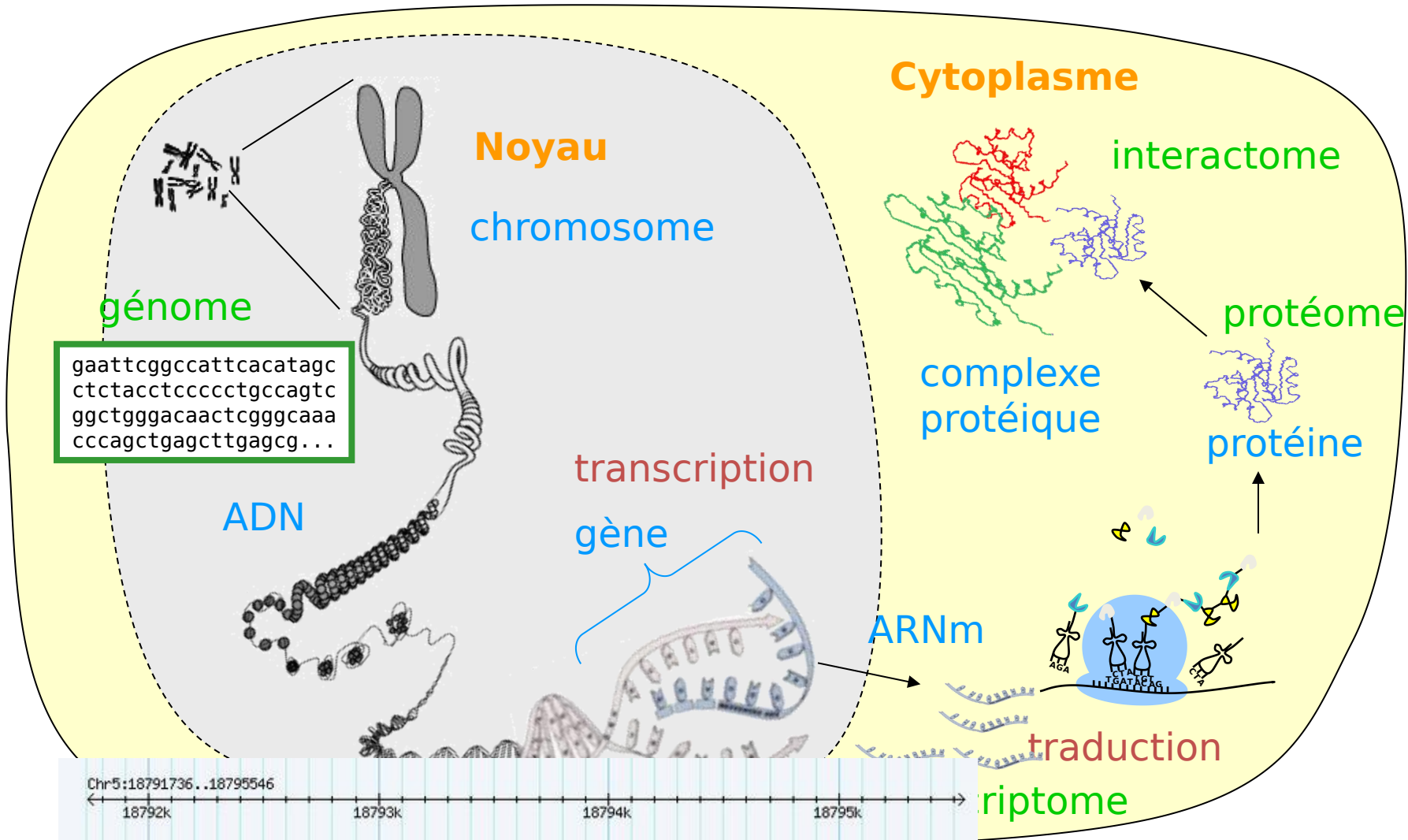


Historique et domaines d'application liés à la bioinformatique

- 1970
- Apparition du terme bioinformatique
 - Disponibilité de séquences de gènes (et du 1^{er} 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 (1^{er} 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

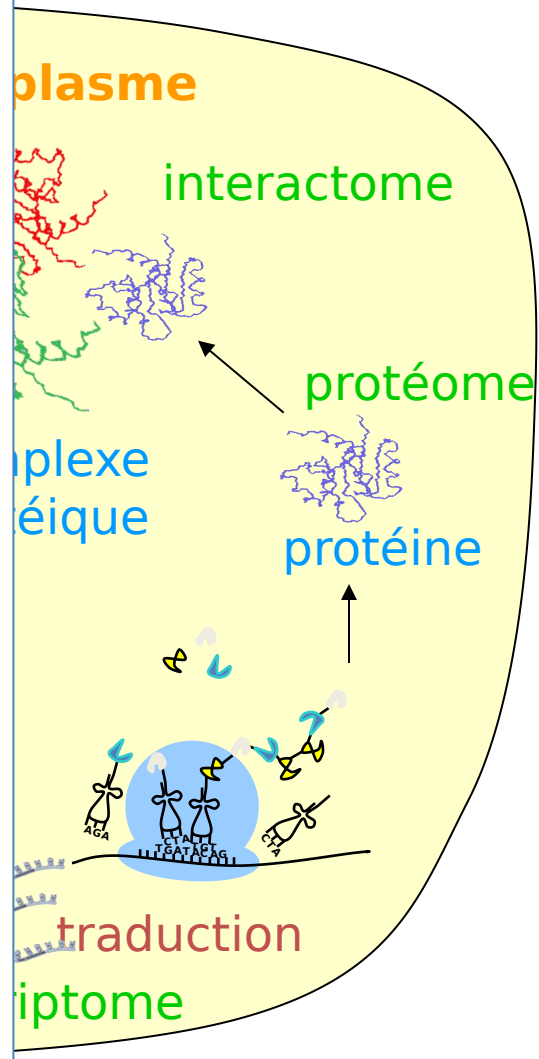
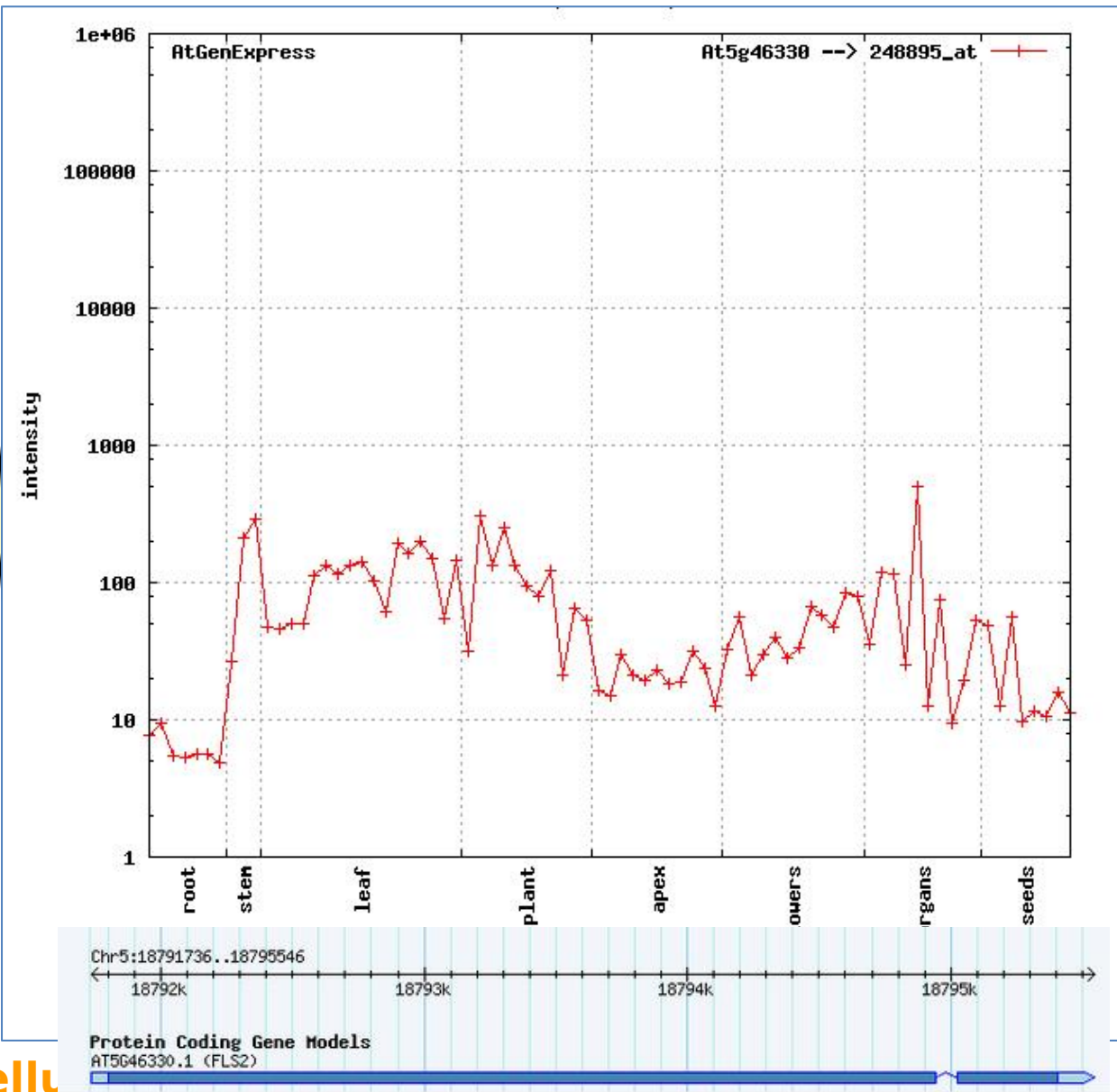


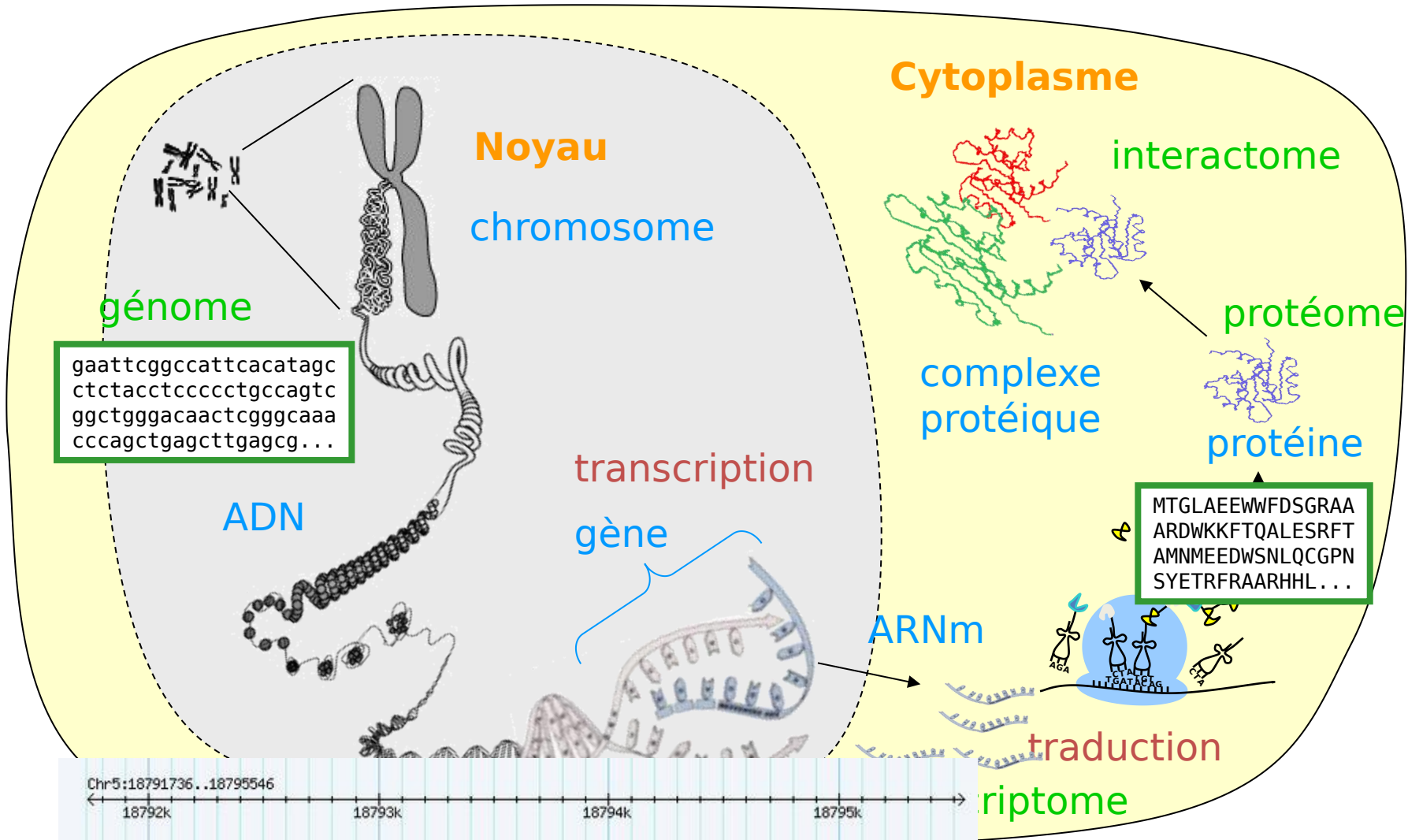
Cellule eucaryote



```
gaattcggccattcacatagc
ctctacctccccctgccagtc
ggctgggacaactcgggcaa
cccagctgagcttgagcg...
```





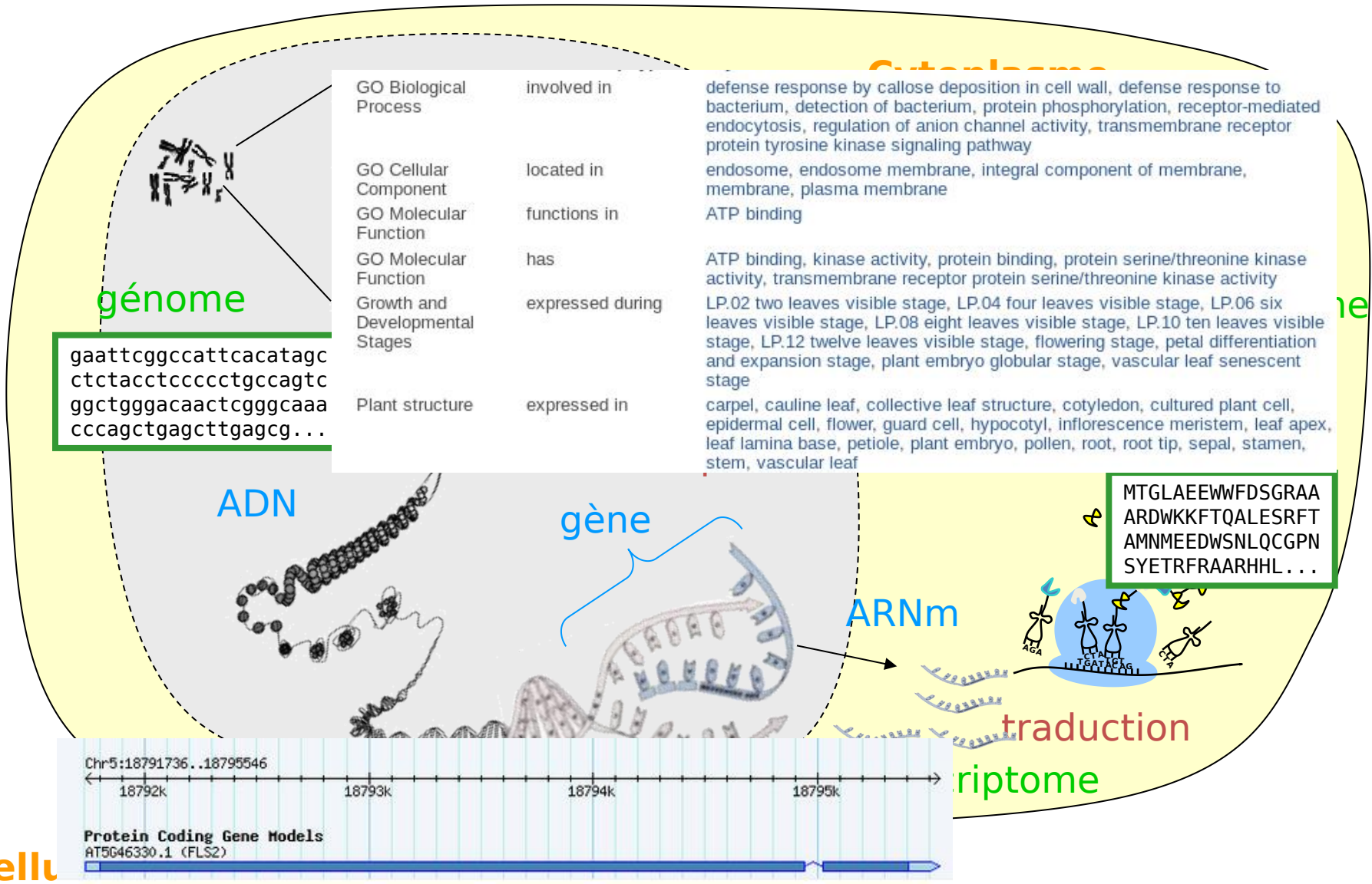


```
gaattcggccattcacatagc
ctctacctccccctgccagtc
ggctgggacaactcgggcaa
cccagctgagcttgagcg...
```

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

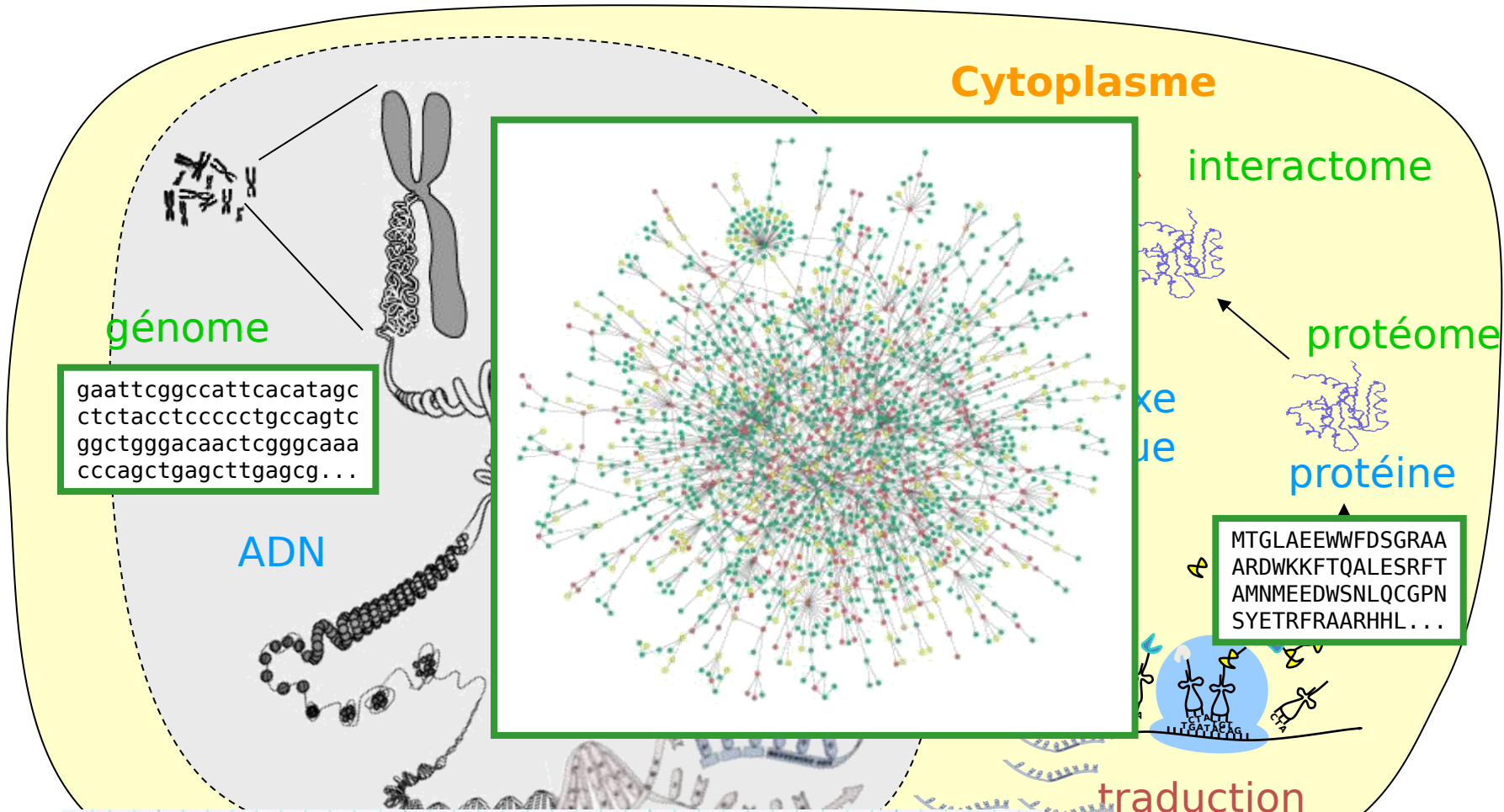


(Quelques) données (mesures et prédictions) et connaissances disponibles



métabolisme régulation

Cytoplasme



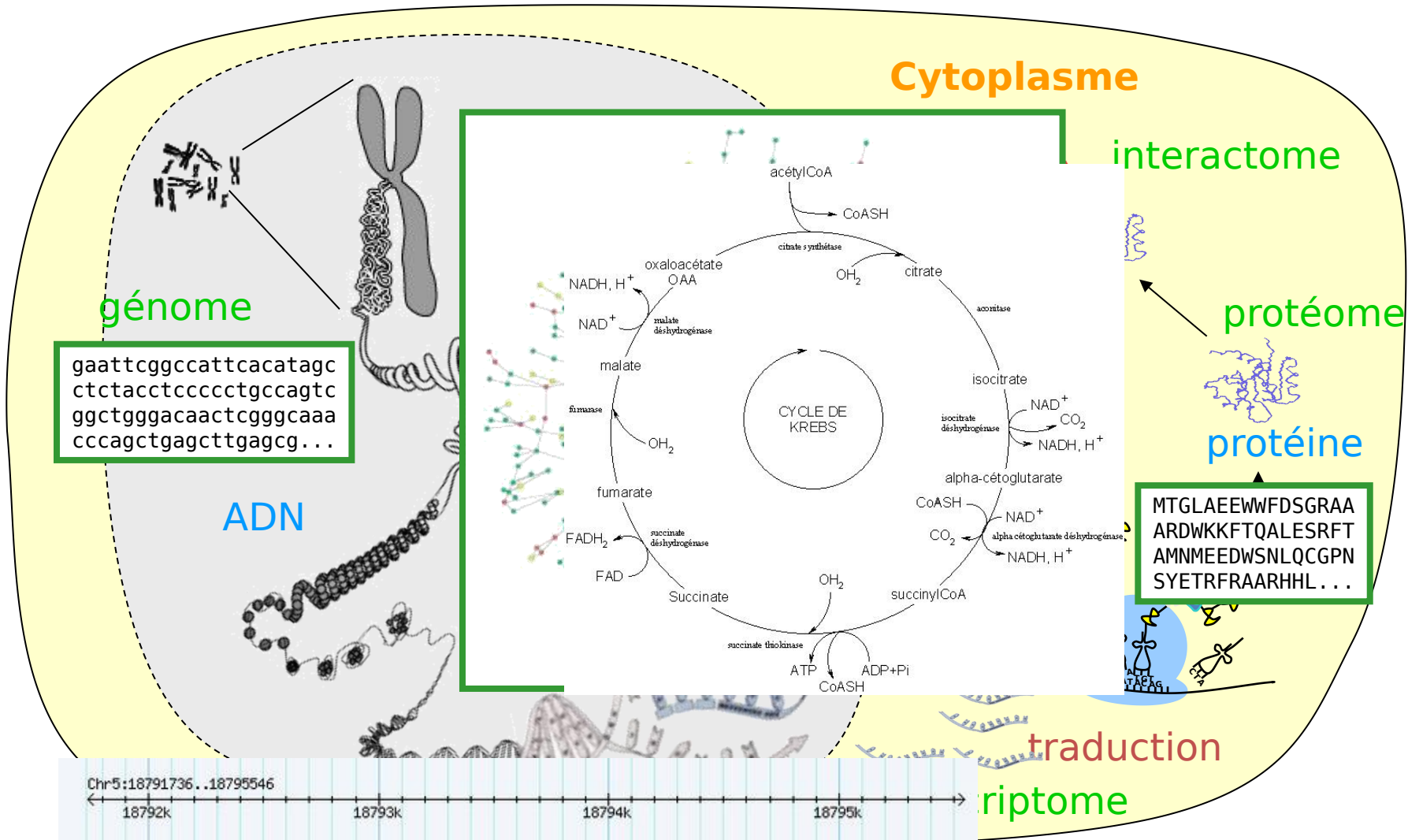
```
gaattcggcattcacatagc
ctctacctccccctgccagtc
ggctgggacaactcgggcaaa
cccagctgagcttgagcg...
```

ADN

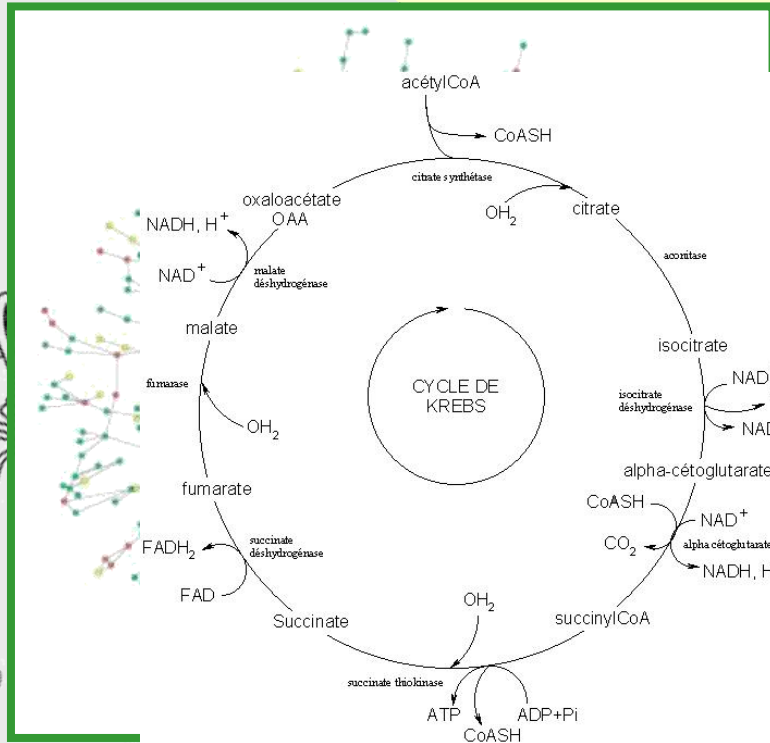
```
MTGLAEEWWFDSGRAA
ARDWKKFTQALESRFT
AMNMEEDWSNLQCGPN
SYETRFRAARHHL...
```



traduction riptome



```
gaattcggcattcacatagc
ctctacctccccctgccagtc
ggctgggacaactcgggcaaa
cccagctgagcttgagcg...
```

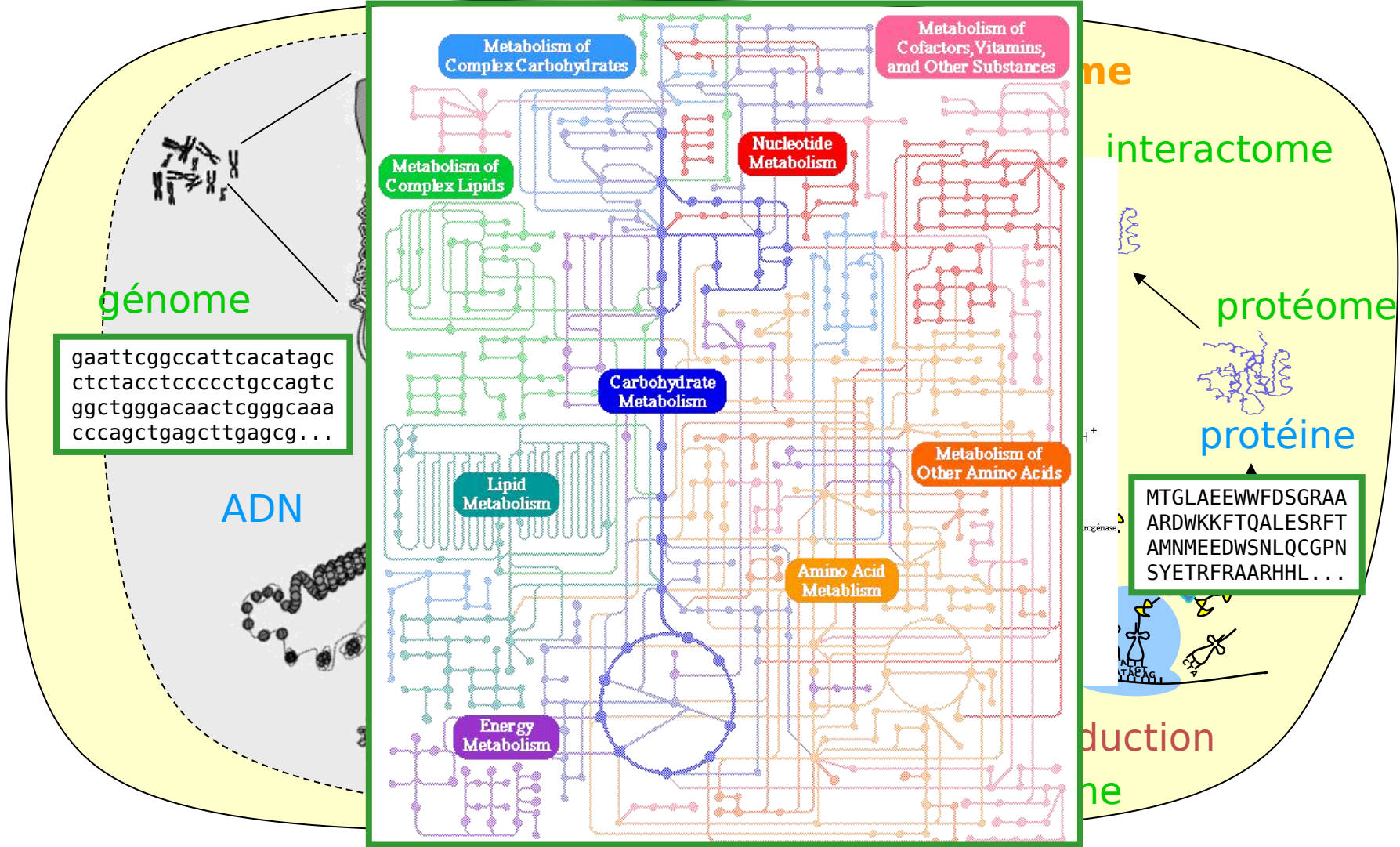


```
MTGLAEEWWFDSGRAA
ARDWKKFTQALESRFT
AMNMEEDWSNLQCGPN
SYETRFRAARHHL...
```



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

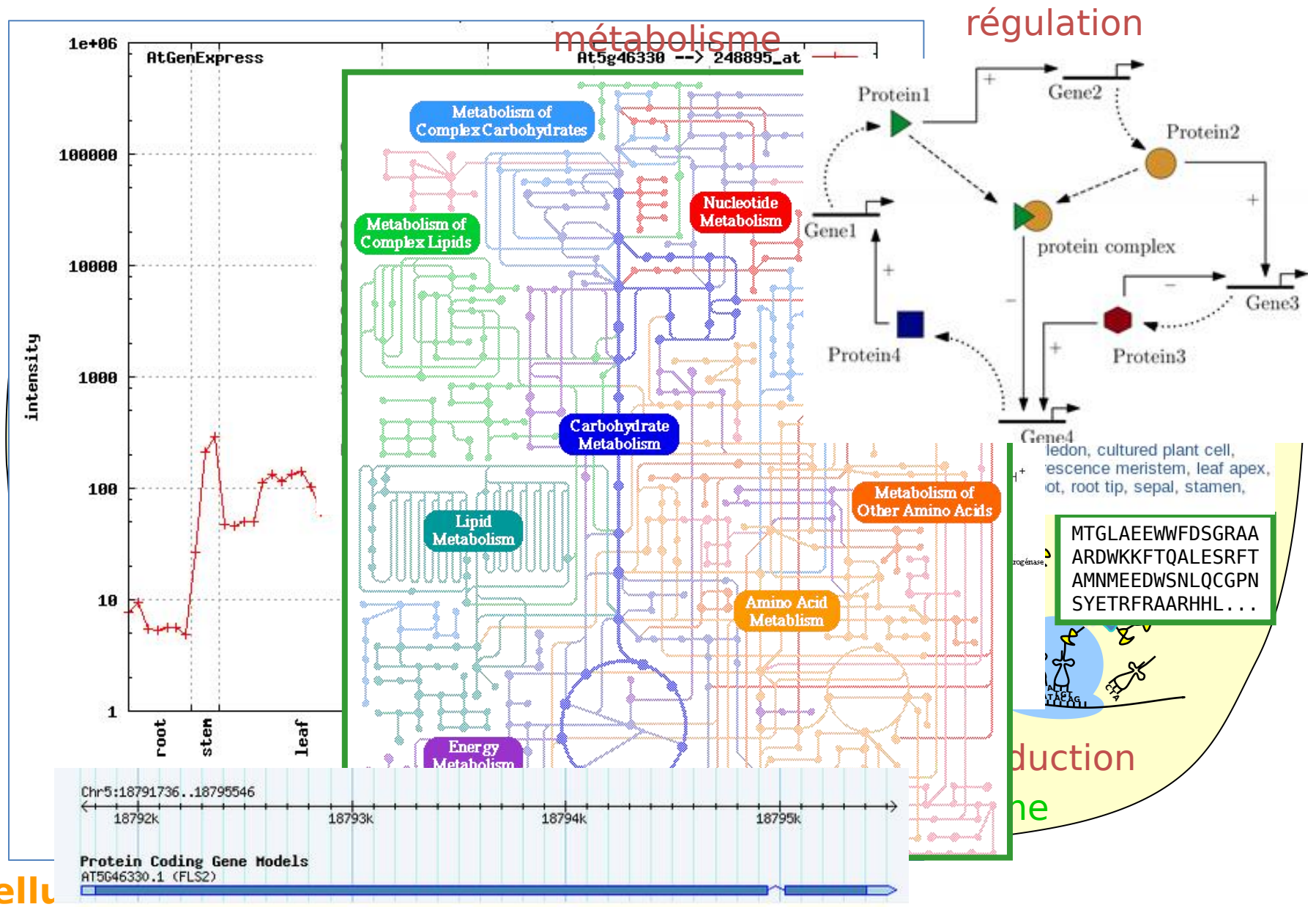
métabolisme



Cellule eucaryote

métabolome

production



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

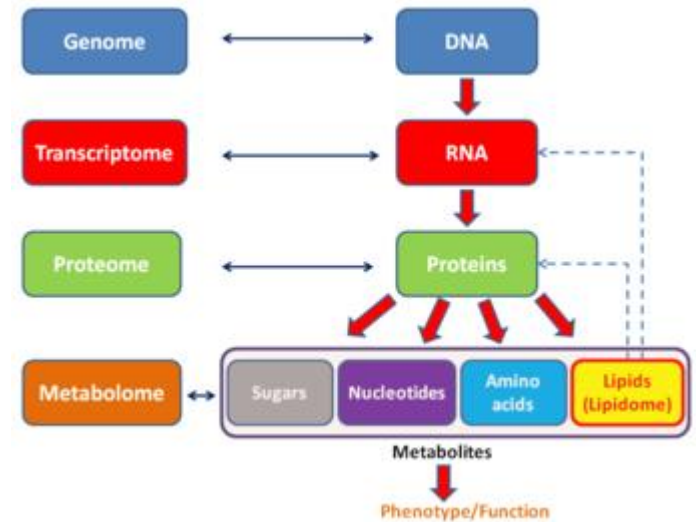
- **Interactome**

- ensemble des interactions moléculaires pouvant survenir *in vivo*
- ensemble des interactions moléculaires dans des conditions données
- ensemble des interactions 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, ...

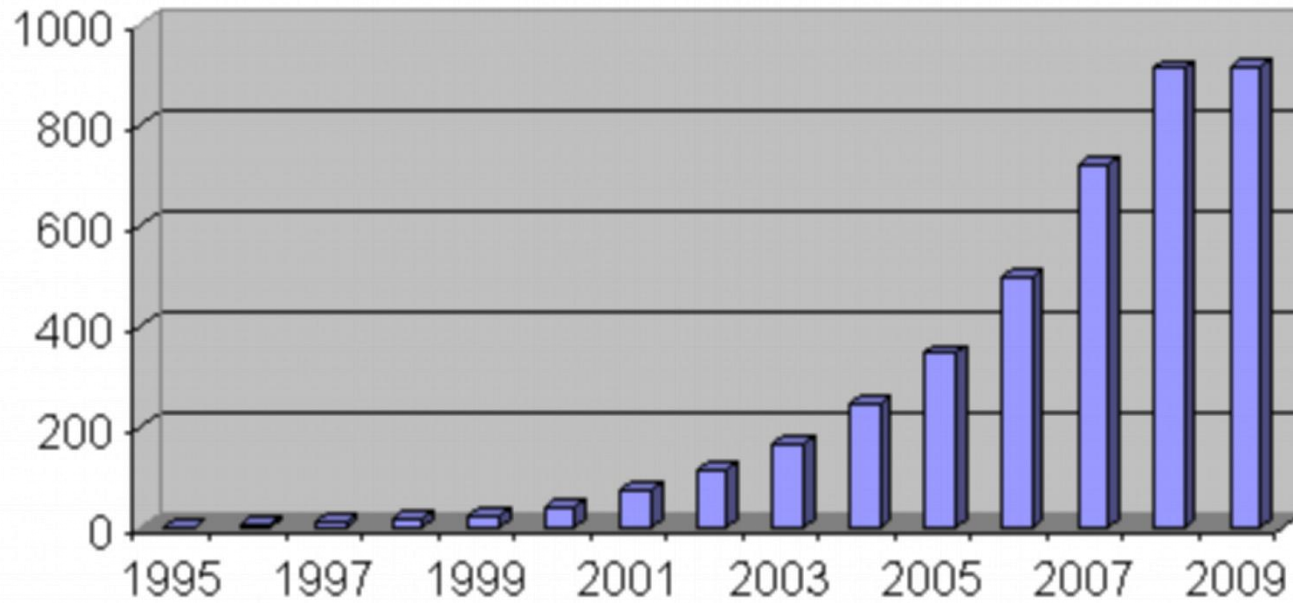


source : Wikipedia

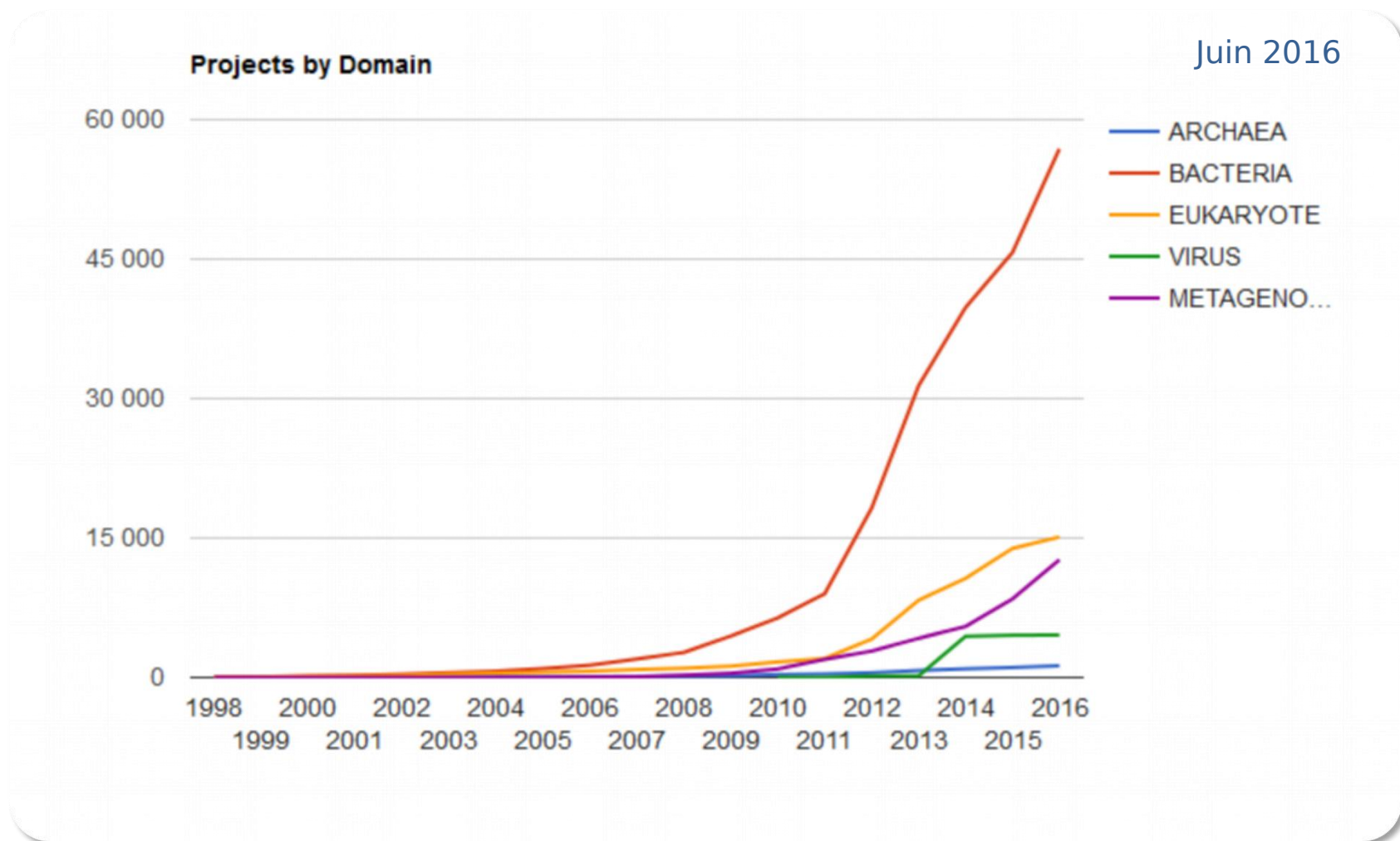
Séquences disponibles : quelques chiffres

Completely Sequenced Genomes ©

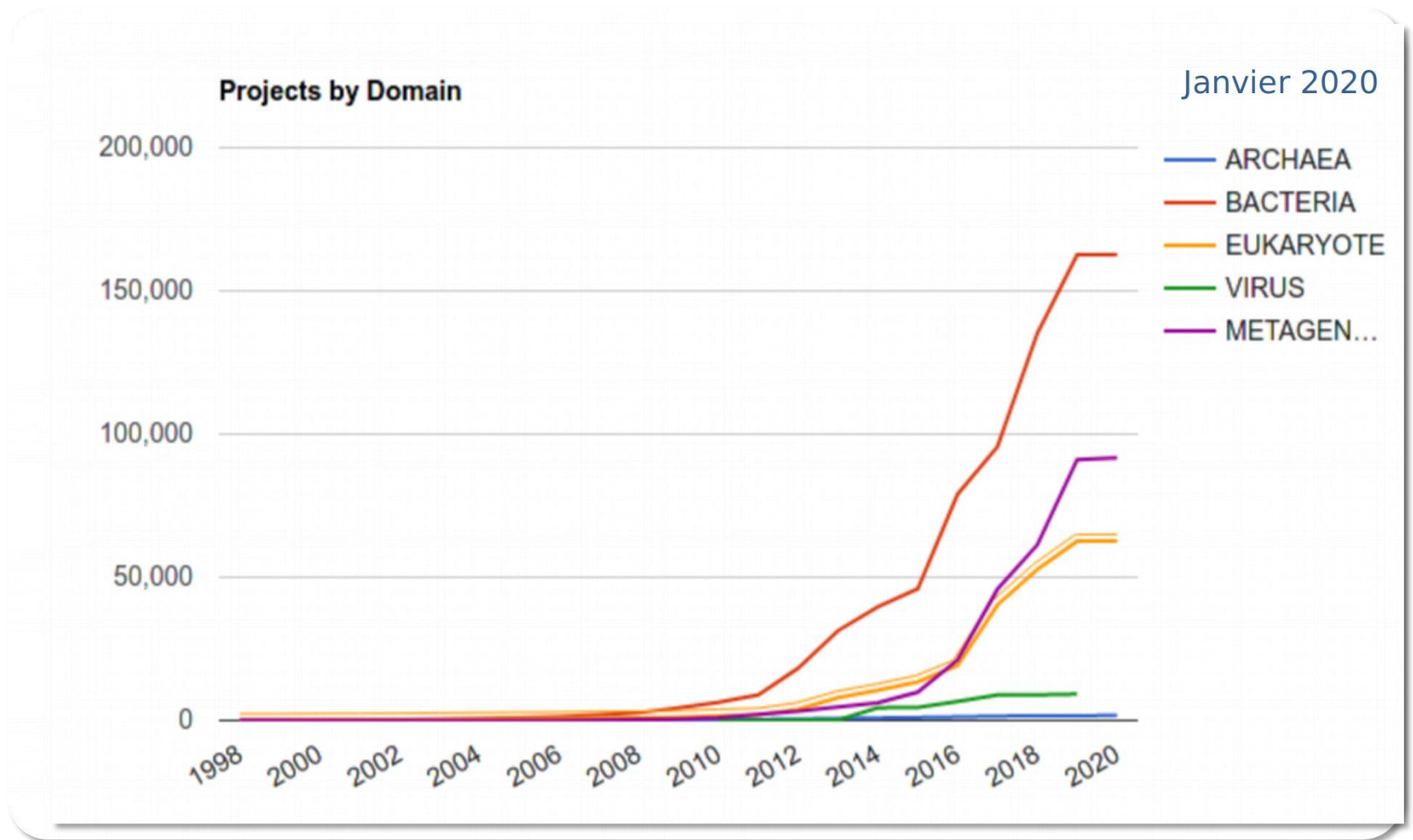
January 2009



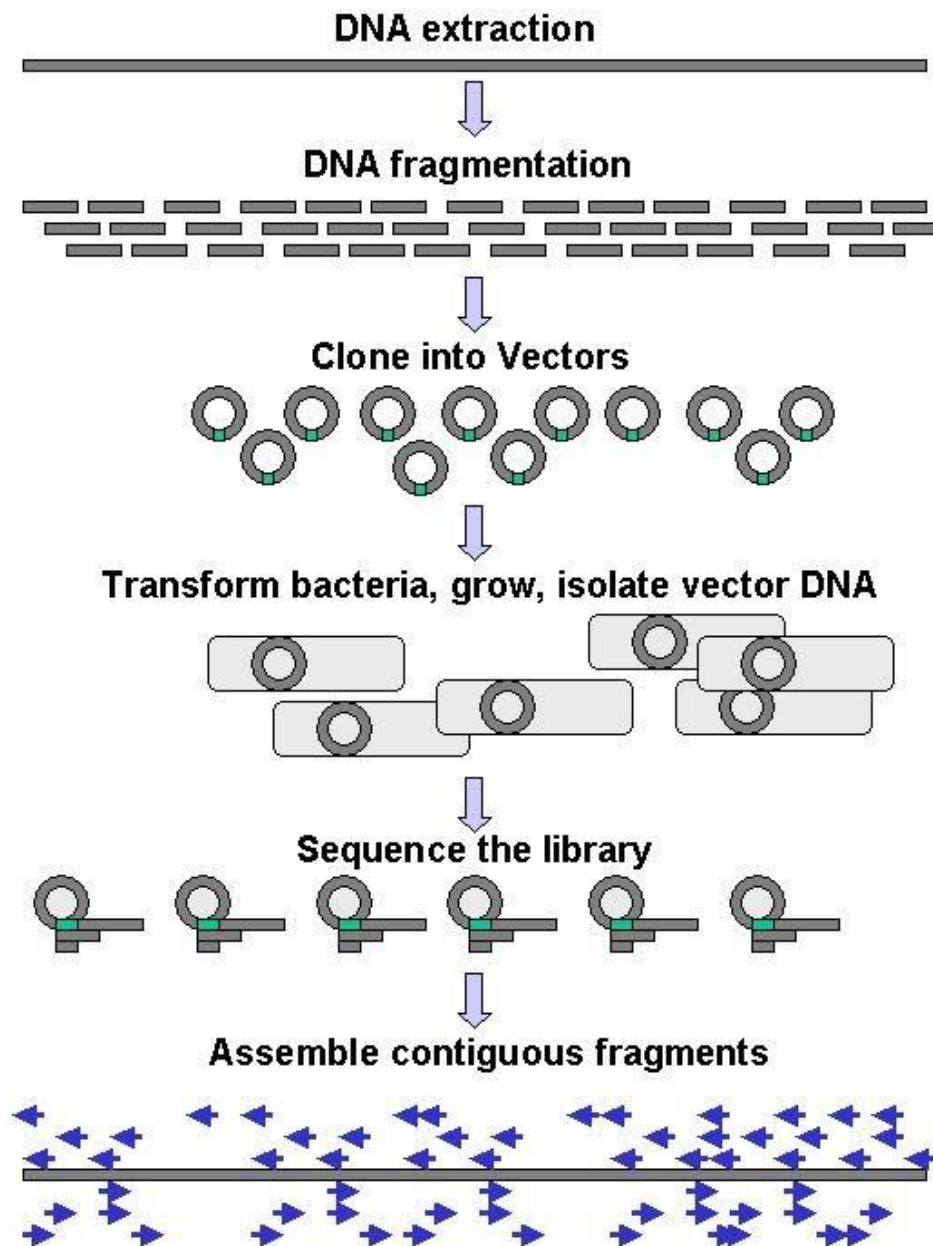
Séquences disponibles : quelques chiffres



Séquences disponibles : quelques chiffres



Séquençage d'un génome



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

TCCTGGCCTACATGTTCTTTGGCAAAGGATCTTCAAATCAACGGCTCCCGGTGCGGCGATCATCCATTTCTTCGGAGGGATTACGAGATTT
ACTTCCCCTACATTTCTGATGAAACCTGGCCCTGATTCTCGCAGCCATTGCCGGCGGAGCAAGCGGACTCTAACATTACGATCTTTAATGCCG
GACTTGTGCGCGCAGCGTCACCGGGAAGCATTATCGCATTGATGGCAATGACGCCAAGAGGAGGCTATTTTCGGCGTATTGGCGGGTGTATTGG
TCGCTGCAGCTGTATCGTTCATCGTTTTAGCAGTGATCCTGAAATCCTCTAAAGCTAGTGAAGAAGACCTGGCTGCCGCAACAGAAAAATGC
AGTCCATGAAGGGGAAGAAAAGCCAAGCAGCAGCTGCTTTAGAGGCGGAACAAGCCAAGCAGAGAAGCGTCTGAGCTGTCTCCTGAAAGCGC
GAACAAAATTATCTTTTCGTGTGATCCGGGATGGGATCAAGTGCCATGGGGGCATCCATCTTAAGAAACAAAGTGAAAAAGCGGAGCTTGACA
TCAGTGTGACCAACACGGCCATTAACAATCTGCCAAGCGATGCGGATATTGTATCACCCACAAAGATTTAACAGACCGCGCGAAAGCAAAGC
TGCCGAACGCGACGCACATATCAGTGGATAACTTCTTAAACAGCCGAAATACGACGAGCTGATTGAAAAGCTGAAAAGTAATCTTATAGAAA
GAGAGTATTGTCATGCAAGTACTCGCAAAGGAAACATTAACCTCAATCAAACGGTATCATCAAAGAAGAGGCTATCAAATTGGCAGGCCAGA
CGCTGATTGACAACGGCTACGTGACAGAGGATTACATTAGCAAAATGTTTGACCGTGAAGAAACGTTCTTACGTTTTATGGGGAATTTTCATTG
CCATTCCACACGGCACAGAAGAAGCGAAAAGCGAGGTGCTTCACTCAGGAATTTCAATCATAACAGATTCCAGAGGGCGTTGAGTACGGAGAAG
GCAACACGGCAAAAGTGGTATTCGGCATTGCGGGTAAAAATAATGAGCATTTAGACATTTTGTCTAACATCGCCATTATCTGTTTCAAGAAG
AAACATTGAACGCCTGATCTCCGCTAAAGCGAAGAAGATTTGATCGCCATTTCAACGAGGTGAACTGACATGATCGCCTTACATTTTCGGTTCG
GGAAATATCGGGAGAGGATTTATCGGCGCGCTGCTTACCCTCCGGCTATGATGTGGTGTTCGCGGATGTGAACGAAACGATGGTCAGCCTC
CTCAATGAAAAAAGAATAACACAGTGGAACTGGCGGAAGAGGGACGTTTCATCGGAGATCATTGGCCCGGTGAGCGCTATTAACAGCGGCAGT
CAGACCGAGGAGCTGTACCGCTGATGAATGAGGCGGCGCTCATCACAACAGCTGTGCGCCCGAATGTCCTGAAGCTGATTGCCCCGTCTATC
GCAGAAGGTTTAAAGACGAAGAAATACTGCAACACACTGAATATCATTGCCTGCGAAAATATGATTGGCGGAAGCAGCTTCTTAAAGAAAGAA
ATATACAGCCATTTAACGGAAGCAGAGCAGAAATCCGTGAGTGAACGTTAGGTTTTCCGAATTTCTGCCGTTGACCGGATCGTCCCGATTGAG
CATCATGAAGACCCGCTGAAAGTATCGGTTGAACCATTTTTTCGAATGGGTCATTGATGAATCAGGCTTTAAAGGAAACACCAGTCATAAAC
GGCGCACTGTTTTGTTGATGATTTAACGCCGTACATCGAACGGAAGCTGTTTACGGTCAATACCGGACACGCGGTACAGCGTATGTCGGCTAT
CAGCGCGGACTCAAACGGTCAAAGAAGCAATTGATCATCCGGAATCCGCCGTGTTGTTTATTTCGGCGCTGCTTGAACTGGTGACTATCTC
GTCAAATCGTATGGCTTTAAGCAAACCTGAACACGAACAATATATTAATAATCAGCGGTCGCTTTTAAATCCTTTTCAATTTTCGGACGATGTGAC
CCGCGTAGCGAGGTCACCTCTCAGAAAACCTGGGAGAAAATGTAGACTTGTAGGCCCGGCAAAGAAAATAAAAGAACC GAATGCACTGGCTGAA
GGAATTGCCGCAGCACTGCGCTTCGATTTACCGGTGACCCTGAAGCGGTTGAACTGCAAGCGCTGATCGAAGAAAAGGATAACAGCGGCGTAC
TTCAAGAGGTGTGCGGCATTCAGTCCCATGAACCGTTGCACGCCATCATTTTTAAAGAACTTAATCAATAACCGACCACCCGTGACACAATGT
CACGGGCTTTTTACTATCTCGCAATCTAGTATAATAGAAAGCGCTTACGATAACAGGGGAAGGAGAATGACGATGAAACAATTTGAGATTGCG
GCAATACCGGGAGACGGAGTAGGAAAGAGGTTGTAGCGGCTGCTGAGAAAGTGCTTATACAGCGGCTGAGGTACACGGAGGTTTGTCAATTTCT
CATTCACAGCTTTTCCATGGAGCTGTGATTATTACTTGGAGCACGGCAAAAATGATGCCCGAAGATGGAATACATACGTTACTCAATTTGAA
GCAGTTTTTGGGAGCTGTCGAAATCCGAAGCTGGTTCCCGATCATATATCGTTATGGGGCTGCTGCTGAAATCCGGAGGGAGCTTGAGCTTT
CCATTAATATGAGACCCGCCAAACAAATGGCAGGCATTACGTCGCCGCTTCTGCATCCAAATGATTTTTGACTTCGTGGTGTATTGCGGAGAAC
AGTGAAGGTGAATACAGTGAAGTTGTGCGGCGCATTACAGAGGCGATGATGAAATCGCCATCCAGAATGCCGTGTTTACGAGAAAAGCGACA
GAACGTGTCATGCGCTTTGCCTTCGAATT

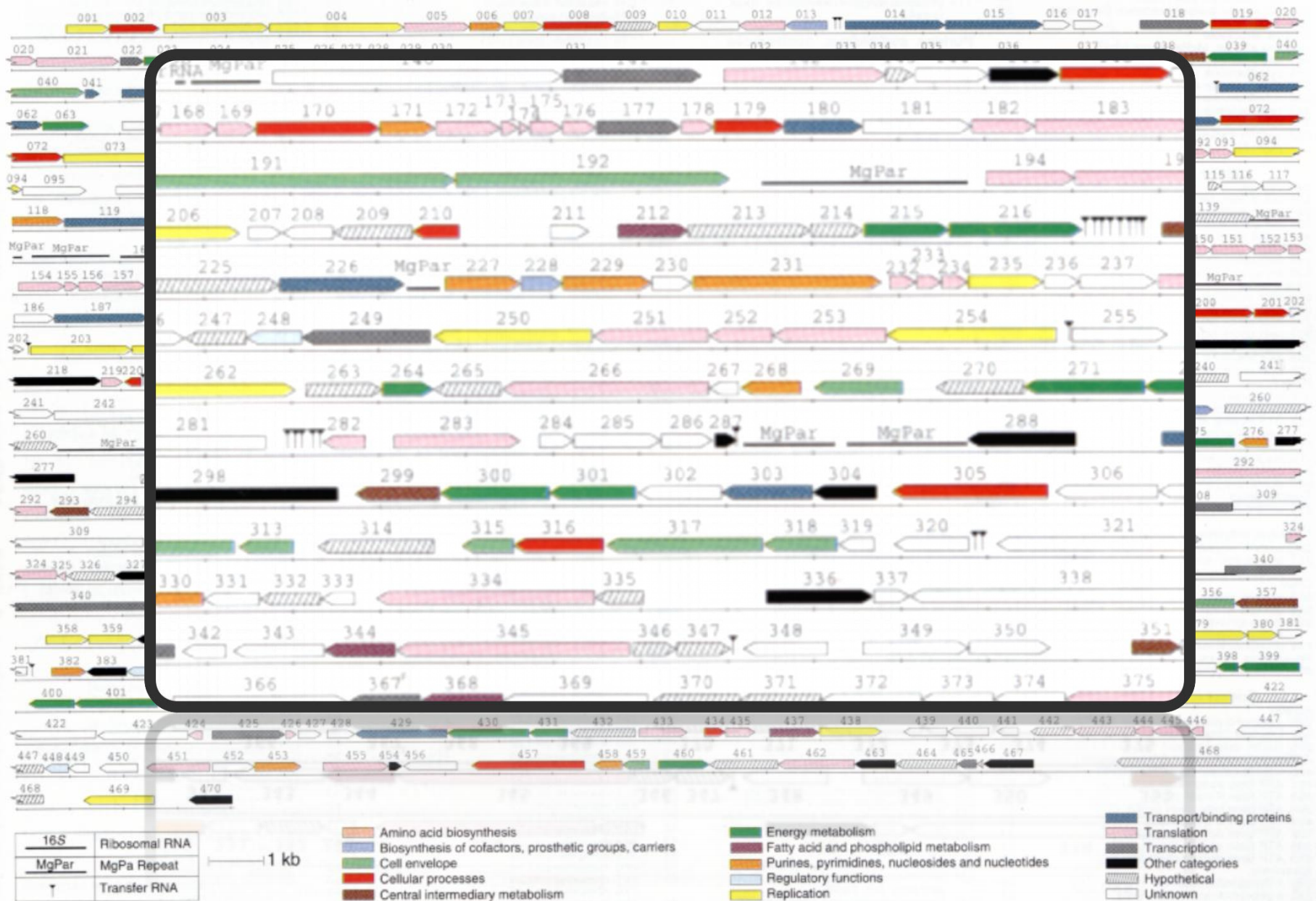
Des séquences... et après ?

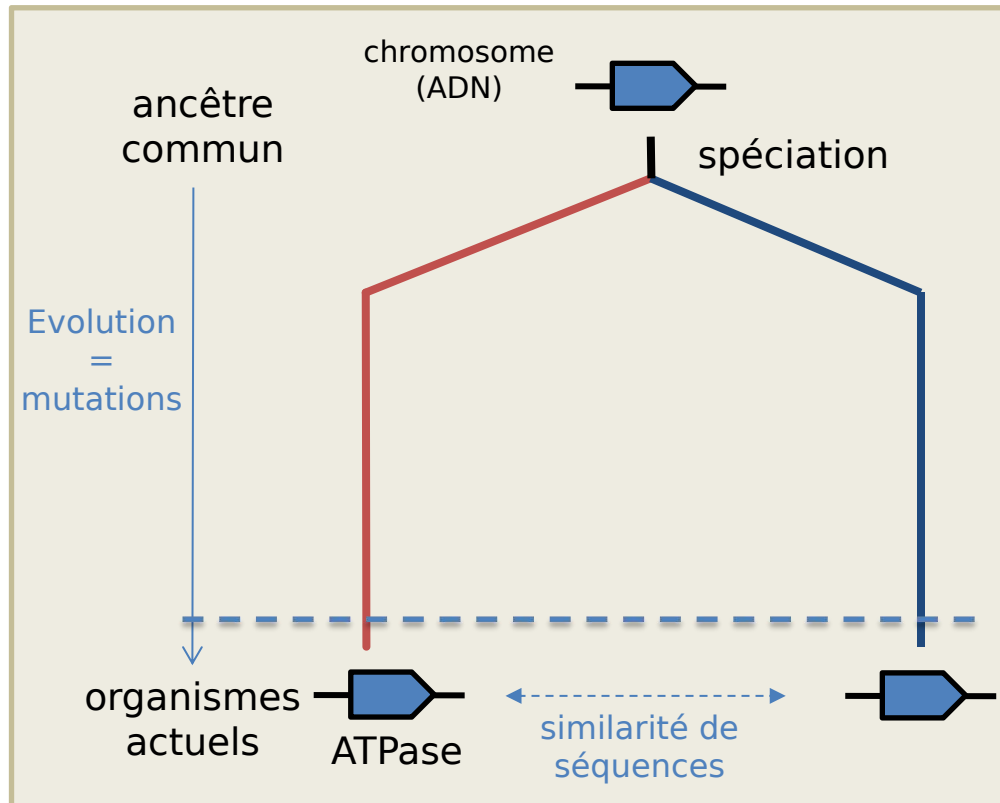
- Annotation
 - régions codantes, régions régulatrice, ...
 - prédiction fonctionnelle
- Identification de protéine/Prédiction de structure
- Analyse des relations génotype/phénotype
- Analyses évolutives
- Analyses d'expression des gènes/protéines
- Réseau d'interaction protéine-protéine
- Reconstruction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





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

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

BLAST+ 2.10.0 is released - Improved Composition-based statistics.

We have updated the BLAST process to improve the stability of BLAST results against changes in the number of results requested.

Mon, 23 Dec 2019 16:00:00 EST [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide




Protein BLAST
protein ► protein

BLAST formulaire de requête

BLAST [®] >> blastp suite


blastn **blastp** blastx tblastn tblastx


Enter Query Sequence


Enter accession number(s), gi(s), or FASTA sequence(s)  Clear  Query subrange 

>OsProt
 MERNKFASKMSQHYTKTICIAVVLVAVL.FSLSSAAAAGSGAAVSVQLEALLEFKNGVADD
 PLGVLAGNRVKGSGDGAVRGGALPRHCNWTGVACDGAGOVTSIQLPESKLRGALSPFLGN
 ISTLQVIDLTSNAFAGGIPPQLGRLGELEQLVVSSNYFAGGIPSSLCNCSAMWALALNVN
 NLTGAIPTSCIGDLSNLEIFAYLNNLDGELPPSMAKLGIMVVDLSCNQLSGSIPPEIGD
 LSNLQIQLYENRFSGHIPRELGRCKNLTLNIFSNGETGEIPGELGELTNLEVMRLYKN
 ALTSETPRSLRRCVSLNLDLSMNQAGTIPPELGEIPSLQRLSLHANRLAGTVPASLTLN
 IVMII TTI ELCEMHI GCDI DACTGCI DNI DD I TVNMML SCOTDACTMCTOI AMACSEM


From
 To


Or, upload file No file selected. 

Job Title
 Enter a descriptive title for your BLAST search 

Align two or more sequences 

Choose Search Set

Database 


Organism Optional exclude
 Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. 

Exclude Optional Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample s

Program Selection

Algorithm

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterative BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm 

Search database nr using Blastp (protein-protein BLAST)
 Show results in a new window

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

[+ 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 [?](#)

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
<input checked="" type="checkbox"/>	LRR receptor-like serine/threonine-protein kinase FLS2 [Oryza sativa Japonica Group]	2372	2372	100%	0.0	100.00%	XP_015634951.1
<input checked="" type="checkbox"/>	OSJNBa0058K23.7 [Oryza sativa Japonica Group]	2357	2357	99%	0.0	100.00%	CAE02151.2
<input checked="" type="checkbox"/>	hypothetical protein OsJ_16186 [Oryza sativa Japonica Group]	2352	2352	99%	0.0	99.91%	EAZ32006.1
<input checked="" type="checkbox"/>	H0313F03.16 [Oryza sativa]	2326	2326	99%	0.0	98.89%	CAH68341.1
<input checked="" type="checkbox"/>	hypothetical protein OsI_17436 [Oryza sativa Indica Group]	2055	2055	99%	0.0	89.59%	EEC78020.1
<input checked="" type="checkbox"/>	PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Oryza brachyantha]	1855	1855	96%	0.0	85.69%	XP_015691635.1
<input checked="" type="checkbox"/>	PH01B019A14.19 [Phyllostachys edulis]	1638	1638	99%	0.0	73.86%	CCI55350.1
<input checked="" type="checkbox"/>	unnamed protein product [Triticum turgidum subsp. durum]	1615	1615	96%	0.0	73.02%	VAH36634.1

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

[Distance tree of results](#) [Multiple alignment](#)

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

Desc

[hover to see the title](#) [click to show alignments](#) [Show Conserved Domains](#)

Alignment Scores < 40 40 - 50 50 - 80 80 - 200 >= 200

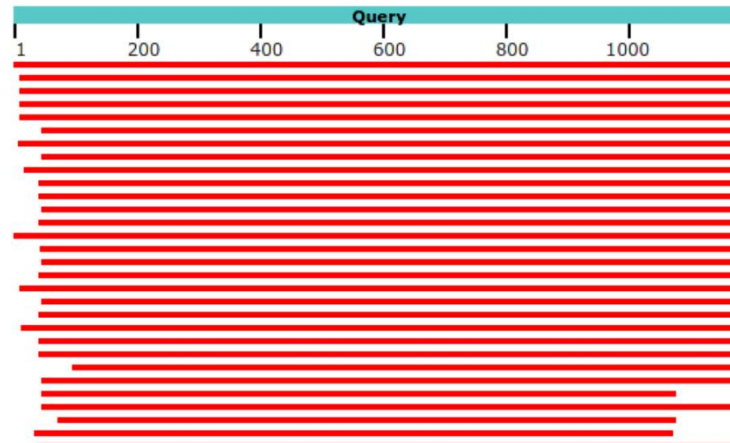
Seq

100 sequences selected [?](#)

Putative conserved domains have been detected, click on the image below for detailed results.

[CD search result summary](#)

Distribution of the top 109 Blast Hits on 100 subject sequences



BLAST résultats

Job Title OsProt
RID [2ANVU5N101R](#)
 Search expires on 01-21 14:32 pm

[Download All](#) ▾

Program BLAS
Database nr
Query ID Icl|Qu
Description OsPro
Molecule type amin
Query Length 1183
Other reports [Distal](#)
[MSA](#)

Descriptions

Sequences produ

select all 100 seq

- [LRR receptor-like s](#)
- [OSJNBa0058K23.7](#)
- [hypothetical protein](#)
- [H0313F03.16 \[Oryz](#)
- [hypothetical protein](#)
- [PREDICTED: LRR](#)
- [PH01B019A14.19 \[](#)
- [unnamed protein pr](#)

Filter Results

Organism only top 20 will appear exclude

Type common name binomial taxid or group name

[Download](#) ▾ [GenPept](#) [Graphics](#)

unnamed protein product [Triticum turgidum subsp. durum]

Sequence ID: [VAH36634.1](#) Length: 1181 Number of Matches: 1

Range 1: 34 to 1181 [GenPept](#) [Graphics](#)

▾ Next Match ▲ Previous Match

	Score	Expect	Method	Identities	Positives	Gaps
	1615 bits(4182)	0.0	Compositional matrix adjust.	850/1164(73%)	944/1164(81%)	41/1164(3%)
Query 45			VQLEALLEFKNGVADDPLGVLAGWRVKGSGDGA VRGGALPRHCNWTGVACDGAGQVTSIQ			104
Sbjct 34			V LEALL FK GV DPLG L+ W G +GD AVRGG +PRHCNWTGVACDGAG+VTSIQ			90
Query 105			LPEskLRGALSPFLGNISTLQVIDLTSNAFAGGIPPQLGRLGELEQLVSSNYFAGGIP-			163
Sbjct 91			LLQTQLQGALTPFLGNISTLQLLDLTENGFTGAIPPQLGRLGQLQLVLAENGFAGGIPP			150
Query 164			-----SSLCNCSAMWALALNVNNTGAIPSCIGDLSNLEIFE			200
Sbjct 151			ELGDLASLQLLDLSNNSLSGGIPSSLCNCSAMWALGVDINNLTGQIPSCIGDLKLQIFE			210
Query 201			AYLNNLDGELPPSMAKLKGMVVDLSCNQLSGSIPPEIGDLSNLQILQLYENRFSGHIPR			260
Sbjct 211			AFMNNLDGELPPSFAKLTQMKSLLSANKLSGSIPQEIGNFSLHWILQMSENRFSGPIPS			270
Query 261			ELGRCKNLTLLNIFSNFTGEIPGELGELTNLEVMRLYKNALTSEIPRSLRRCVSLLNLD			320
Sbjct 271			ELGRCKNLT LNI+SN FTG IP ELGEL NLE +RLY NAL+SEIP SLRRC SLL L			330
Query 321			LSMNQLAGPIPELGE LPSLQRLSLHANRLAGTVPASLTNLVNLTILELSENHLSGPLPA			380
Sbjct 331			LSMNQL G IPPELGE LSLQ L+LHANRL GTVP SLTNLVNLT L L++N LSG LP			390

Banque de séquences protéiques : UniProt

UniProt Advanced

BLAST Align Retrieve/ID mapping Peptide search Help Contact

UniProtKB - Q0JA29 (FLS2_ORYSJ)

Display

-
-
-
-

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 levelⁱ

Functionⁱ

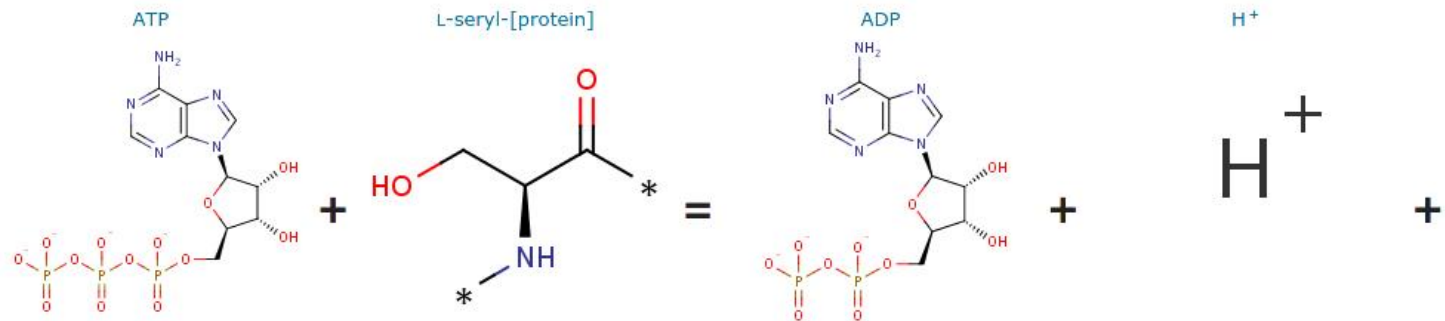
Constitutes the pattern-recognition receptor (PPR) that determines the specific perception of flagellin (flg22), a potent elicitor of the defense response to pathogen-associated molecular patterns (PAMPs). Recognizes flg22 from *Pseudomonas aeruginosa* and *Acidovorax avenae*. flg22 is a peptide derived 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).

Catalytic activityⁱ



• $\text{ATP} + \text{L-seryl-[protein]} = \text{ADP} + \text{H}^+ + \text{O-phospho-L-seryl-[protein]}$

EC:2.7.11.1

Source: Rhea.



Banque de séquences protéiques : UniProt

 BLAST Align  Display Entry Publications Feature viewer Feature table <input checked="" type="checkbox"/> Function <input checked="" type="checkbox"/> Names & Taxonomy <input checked="" type="checkbox"/> Subcellular location <input type="checkbox"/> Pathology & Biophysics <input checked="" type="checkbox"/> PTM / Processing <input checked="" type="checkbox"/> Expression <input checked="" type="checkbox"/> Interaction <input checked="" type="checkbox"/> Structure <input checked="" type="checkbox"/> Family & Domain <input checked="" type="checkbox"/> Sequence <input checked="" type="checkbox"/> Similar proteins <input checked="" type="checkbox"/> Cross-references <input checked="" type="checkbox"/> Entry information <input checked="" type="checkbox"/> Miscellaneous ▲ Top	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=0sFLS2 {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=0s04g0618700 {ECO:0000312 EMBL:BAF15808.1},			
	GN	LOC_0s04g52780 {ECO:0000305};			
	GN	ORFNames=0sJ_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;				

- Pression de sélection → conservation de séquence
- Totalité de la séquence mais plus souvent certaines régions → domaines protéiques
 - 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

Family: *Pkinase* (PF00069)

 9546 architectures 349448 sequences 73 interactions 7104 species 4704 structures

Summary

Domain organisation

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species

Interactions

Structures

Jump to... 

Go

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

There are **202605** sequences with the following architecture: **Pkinase**

[Z4YHY2_DANRE](#) [Danio rerio (Zebrafish) (Brachydanio rerio)] Serine/threonine-protein kinase pim-2 {ECO:0000313|Ensembl:ENSDARP00000076422} (310 residues)



[Show](#) all sequences with this architecture.

There are **11729** sequences with the following architecture: **Pkinase x 2**

[W4GT79_9STRA](#) [Aphanomyces astaci] CMGC/CK2 protein kinase {ECO:0000313|EMBL:ETV82199.1} (659 residues)



[Show](#) all sequences with this architecture.

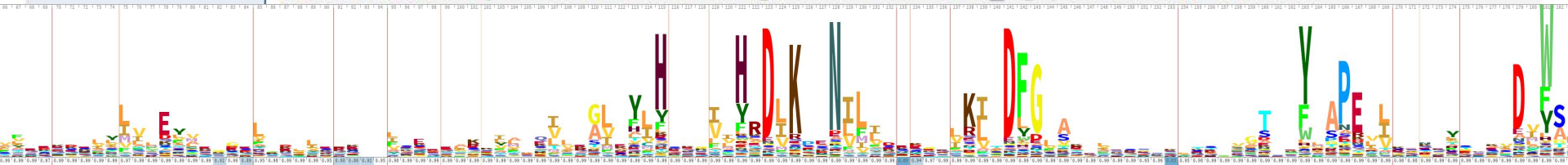
There are **3306** sequences with the following architecture: **Pkinase, Pkinase_C**

[X0KFA3_FUSOX](#) [Fusarium oxysporum f. sp. cubense tropical race 4 54006] Non-specific serine/threonine protein kinase {ECO:0000256|SAAS:SAAS01030073} (635 residues)



[Show](#) all sequences with this architecture.

There are **2855** sequences with the following architecture: **Lectin_legB, Pkinase**



Détection de domaines

LRR receptor-like serine/threonine-protein kinase FLS2 (Q0JA29) - protein - InterPro - Mozilla Firefox

https://www.ebi.ac.uk/interpro/protein/UniProt/C... 90%

InterPro Classification of protein families

Home Search **Browse** Results Release notes Download Help About

Protein family membership

None predicted

Entry matches to this protein

Colour By: Accession

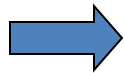
1 100 200 300 400 500 600 700 800

500

IPR000719
Protein kinase domain
D InterPro domain
876 - 1180

- Domain**
 - IPR000719 SM00220 PS50011 PF00069
 - IPR013210 PF08263
- Homologous Superfamily**
 - IPR011009 SSF56112
 - IPR032675 G3DSA:3.80.10.10
- Repeat**
 - IPR001611 PF13855
 - IPR003591 SM00369
- Active Site**
 - IPR008271 PS00108
- Unintegrated**
 - G3DSA:1.10.510.10
 - G3DSA:3.30.200.20
 - PRO00019
 - PTHR27000
 - PTHR27000:SF616
 - SSF52047
 - SSF52058
- Other Features**
 - SIGNAL_PEPTIDE_N_REGION
 - SIGNAL_PEPTIDE
 - SIGNAL_PEPTIDE_C_REGION
 - SIGNAL_PEPTIDE_H_REGION

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

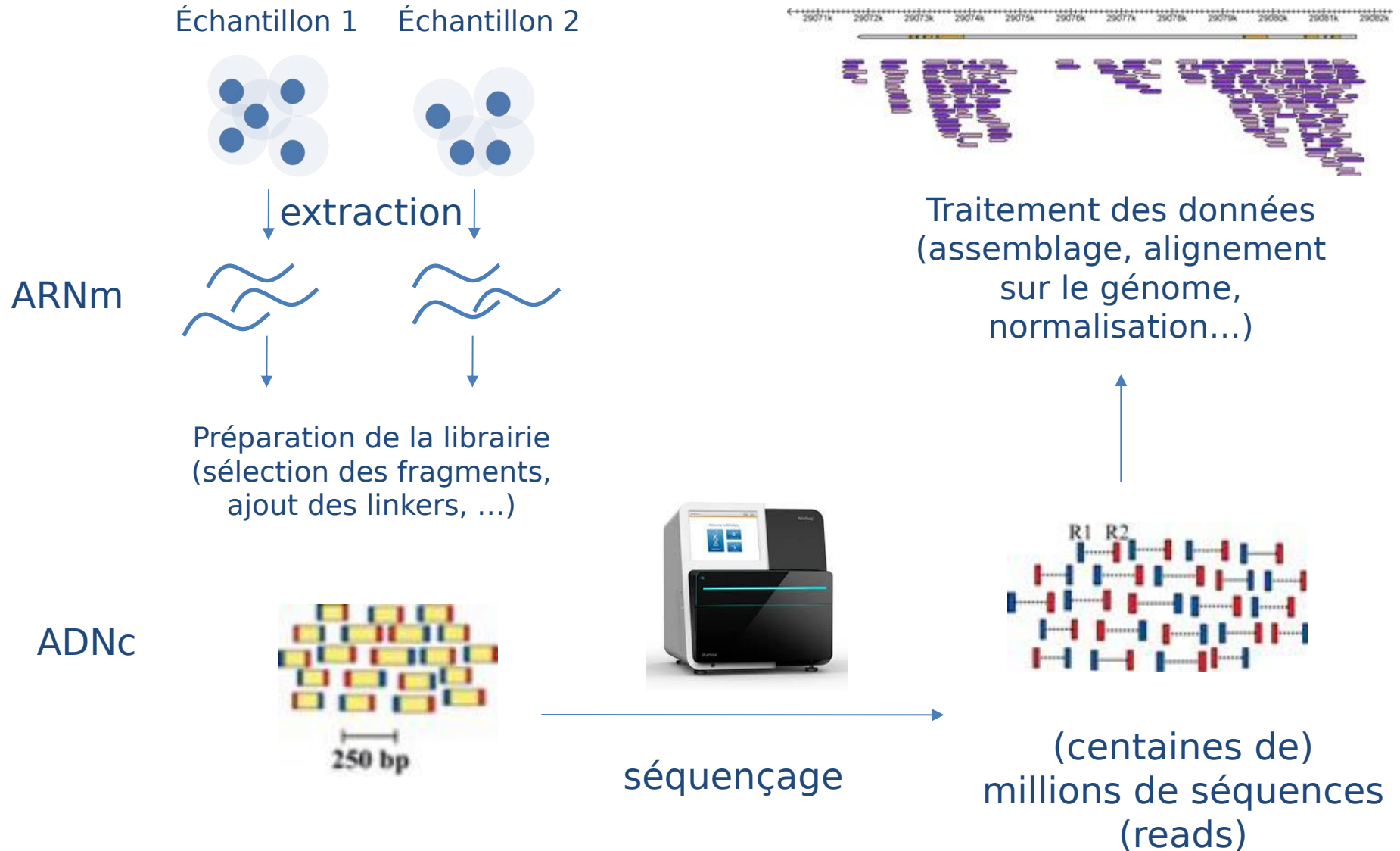


Accès au niveau d'expression de milliers de gènes simultanément (potentiellement l'ensemble des gènes d'un organisme)
= *instantané* de l'état d'une cellule ou d'une population de cellules

Données d'expression des gènes obtenues par :

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

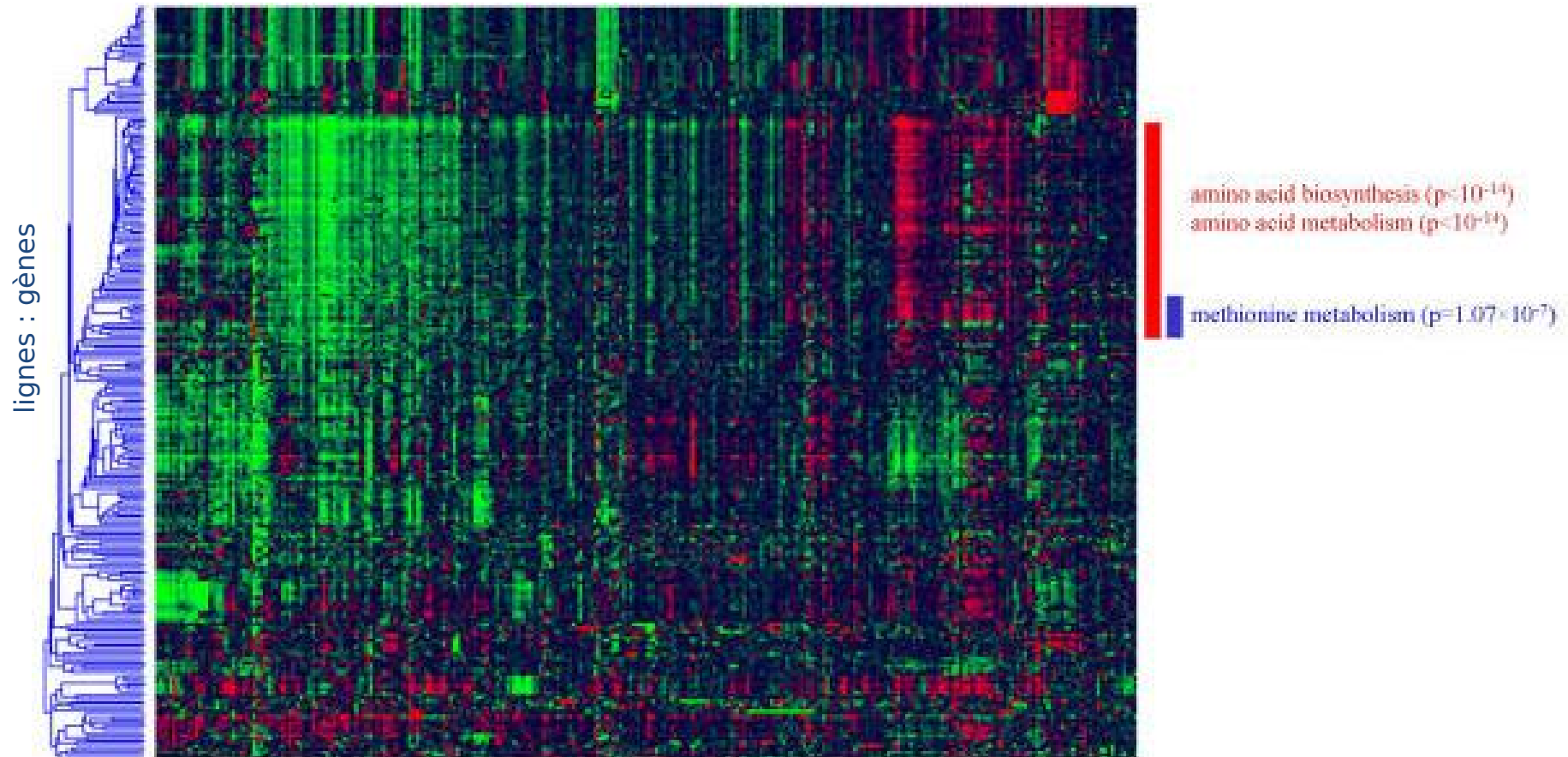
Transcriptome : acquisition des données (RNAseq)



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 expérimentales (ex: mutants)

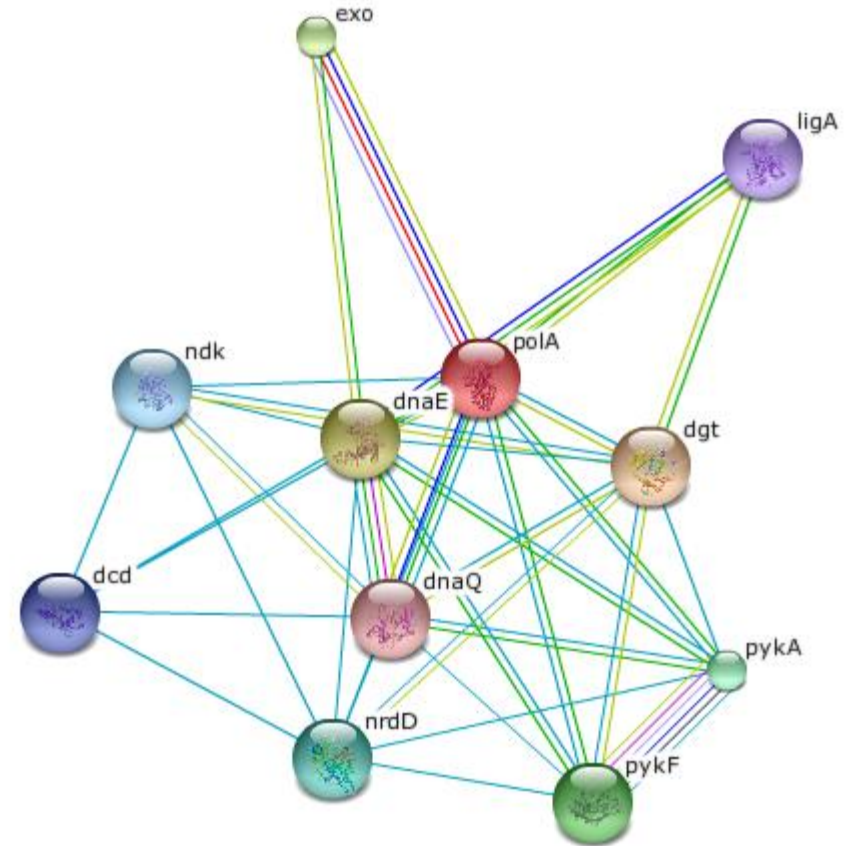
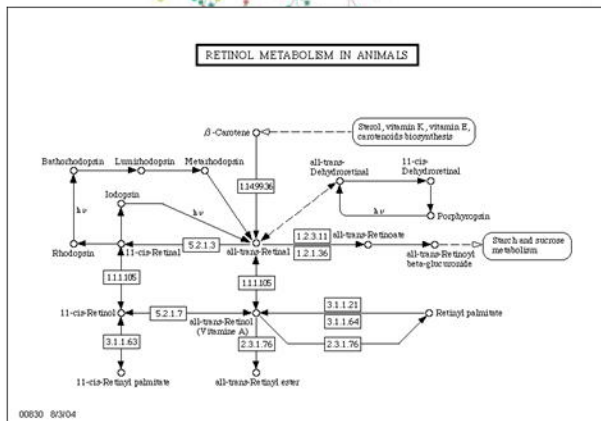


- Métagénomique
 - Echantillon provenant de l'environnement
- ADN ancien · Paléogénomique
- Police scientifique
- Variant de virus
- Biologie intégrative
 - Genome complet
 - Expression
 - Epigénétique (méthylation, modification des histones)
 - Conformation 3D des chromosomes *in vivo*

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

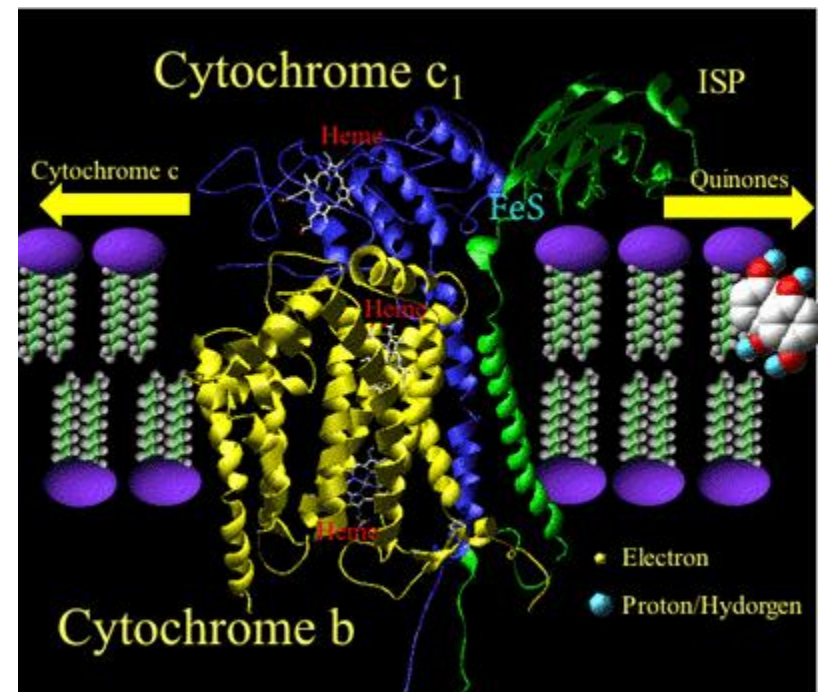
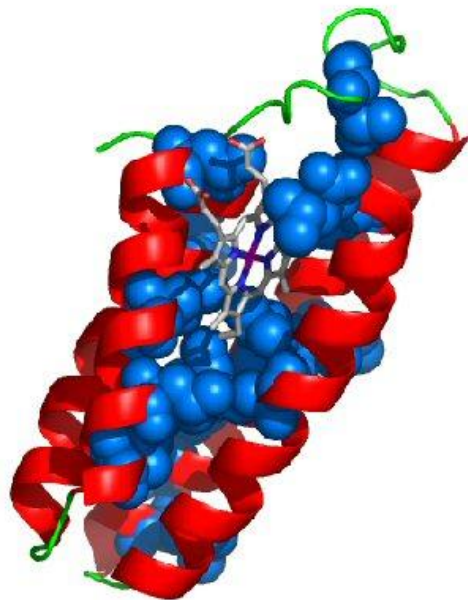
Séquence protéique

>gi|5524211|gb|AAD44166.1| cytochrome b

```
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLLALLSPDMLGDPDNHMPADPLNTPPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFPLPIAGX
IENY
```

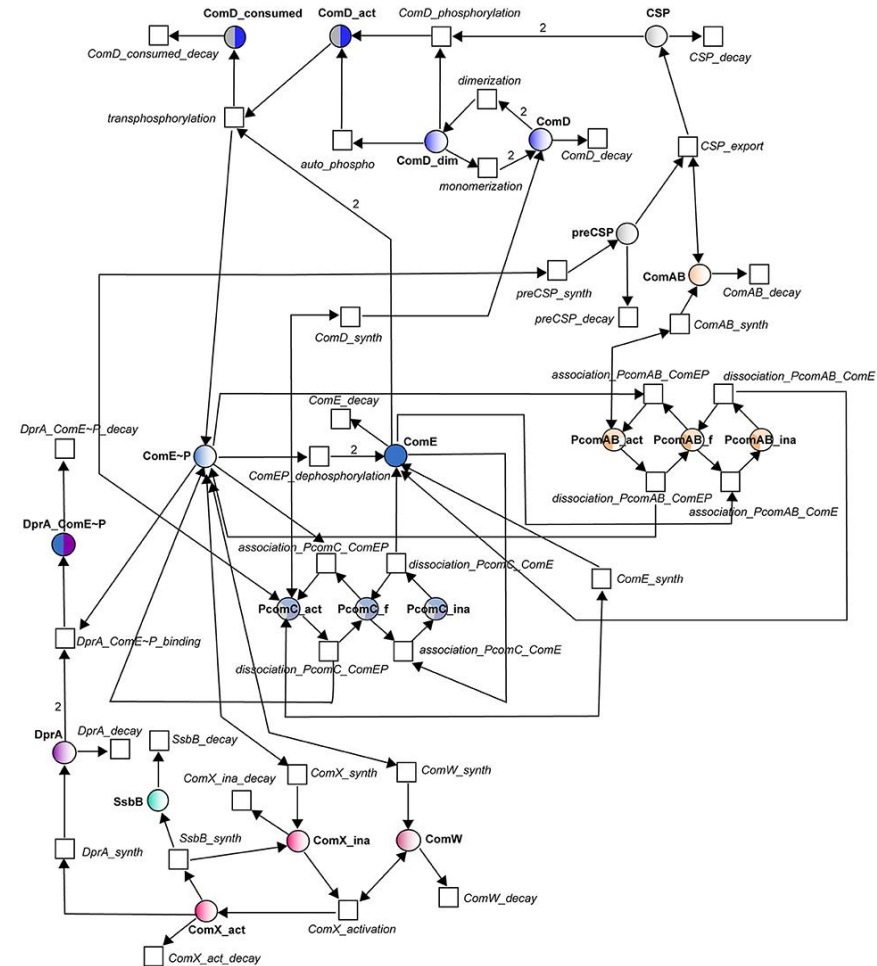
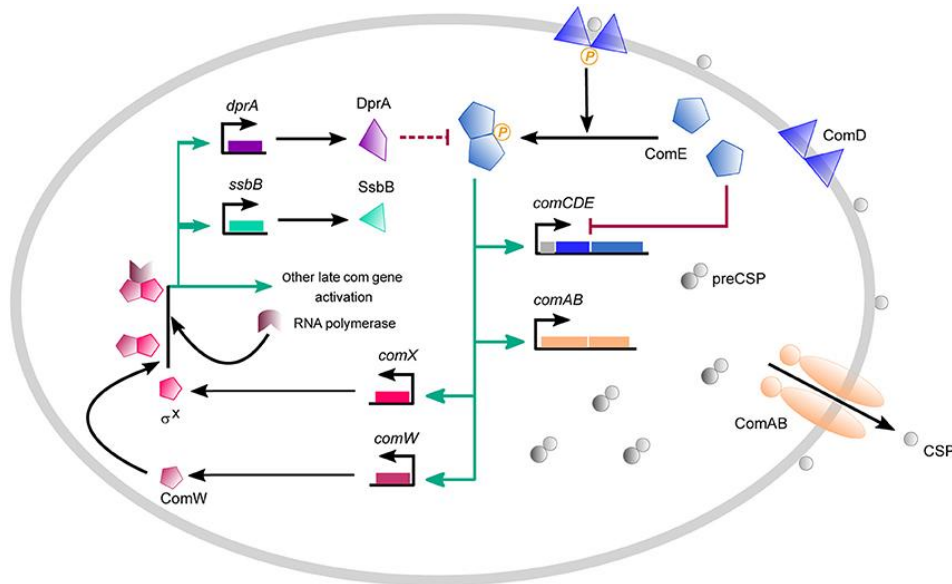


Prédiction ou résolution
de la structure tridimensionnelle



Intégration et synthèse des connaissances

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



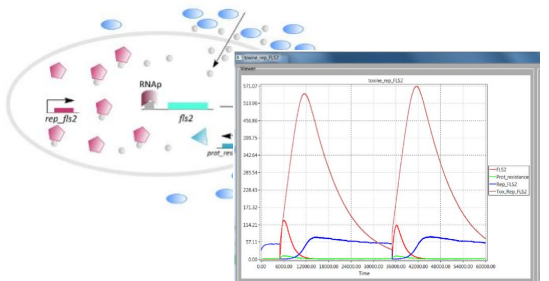
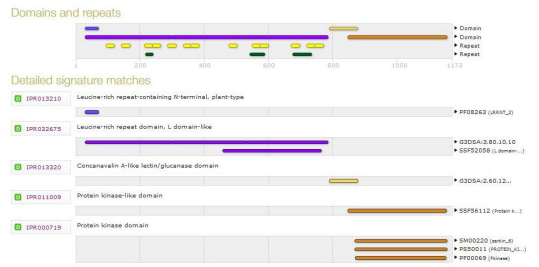
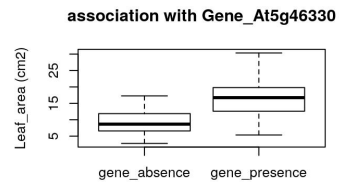
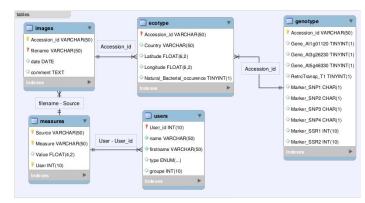
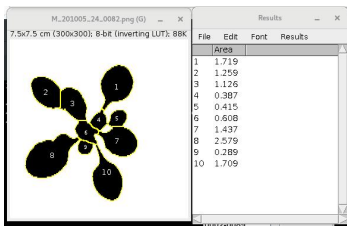
A terme : simulation d'une cellule/organisme/population/écosystème et prédiction de son comportement

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

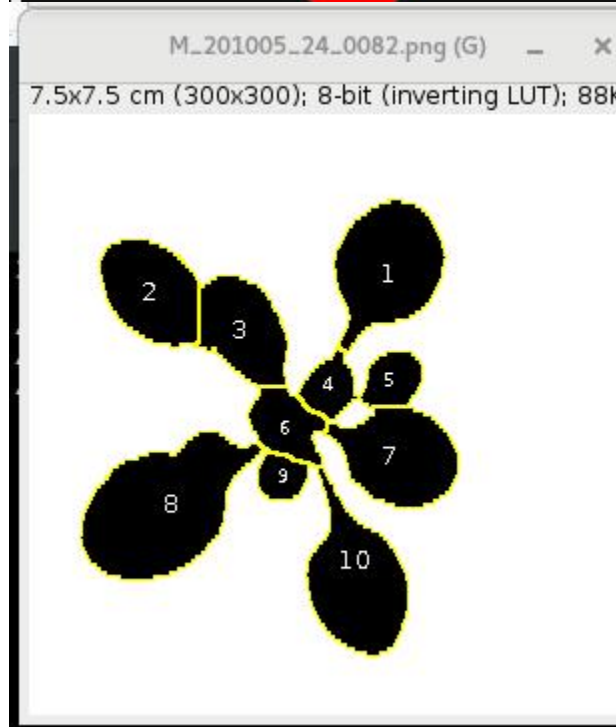
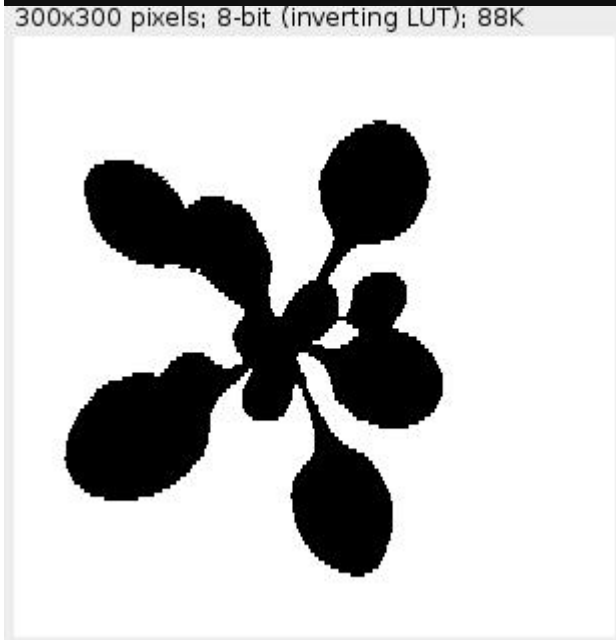
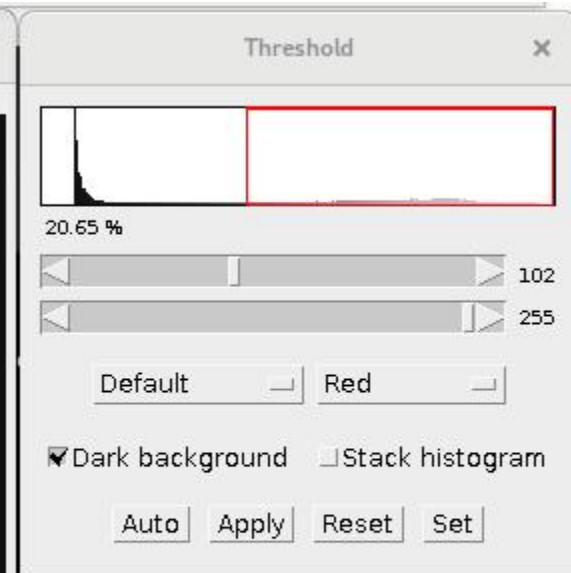
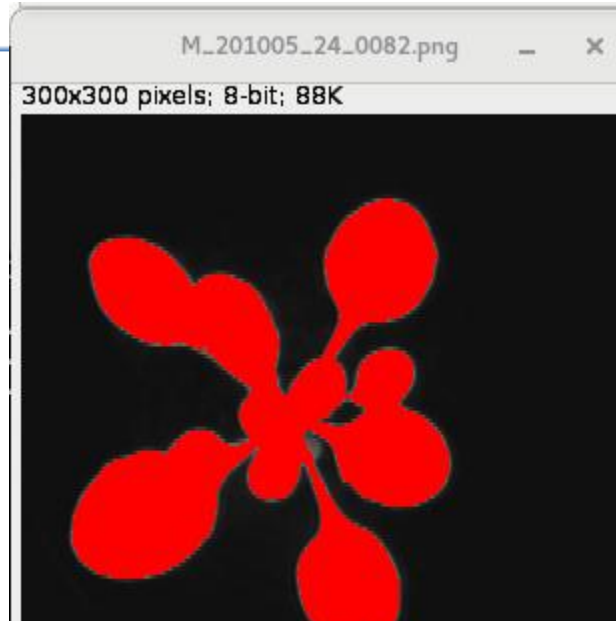
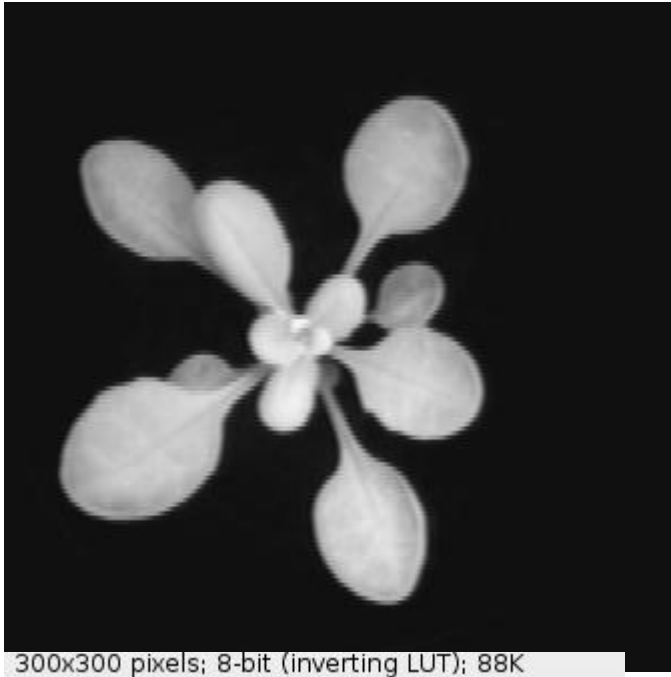
Concepts et méthodes abordés dans cette UE

- 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
 - aperçu 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 / prédiction
 - propriétés émergentes
 - **réseau de pétri**





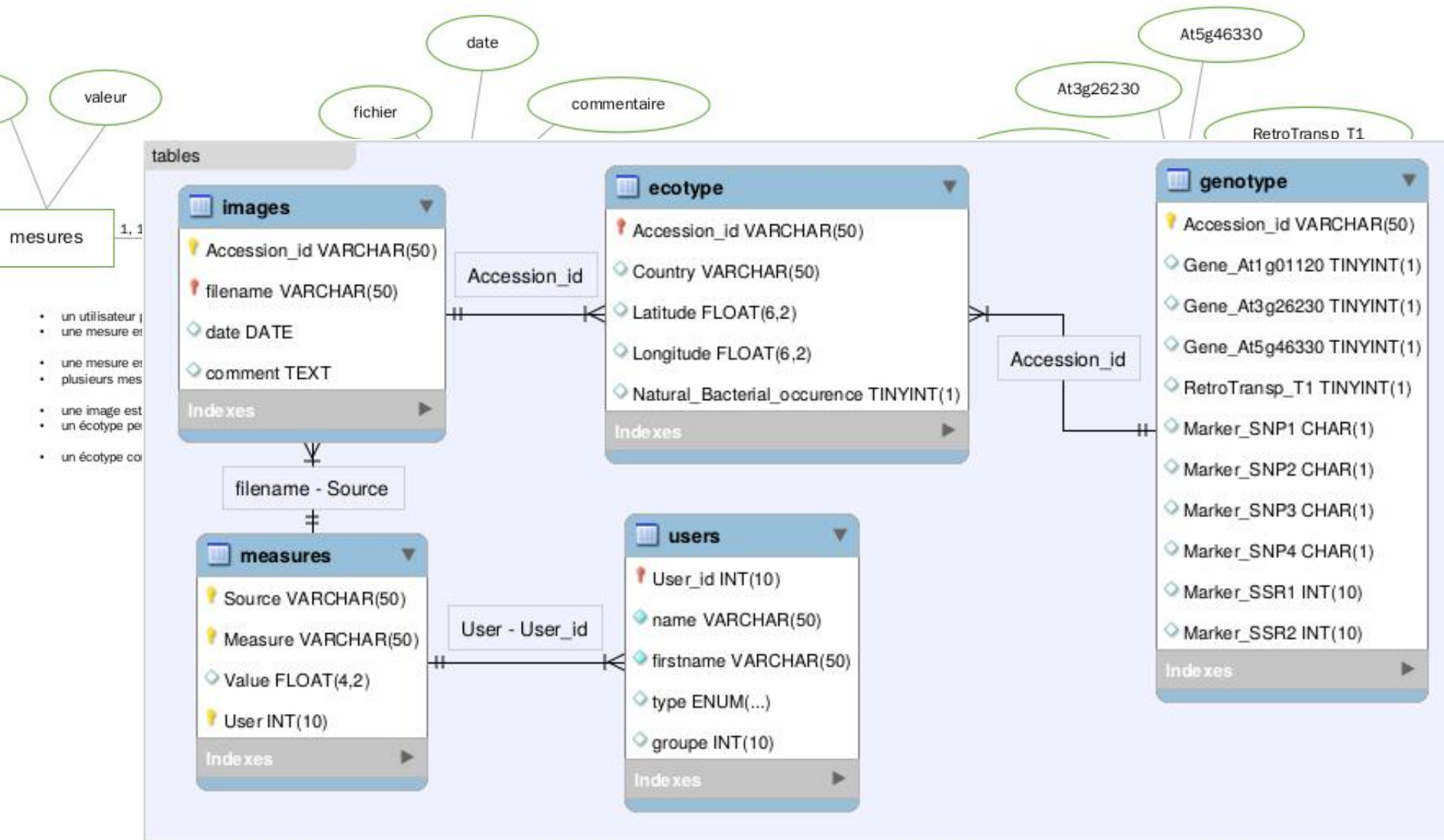
Traitements d'image



Results

	Area
1	1.719
2	1.259
3	1.126
4	0.387
5	0.415
6	0.608
7	1.437
8	2.579
9	0.289
10	1.709

Bases de données



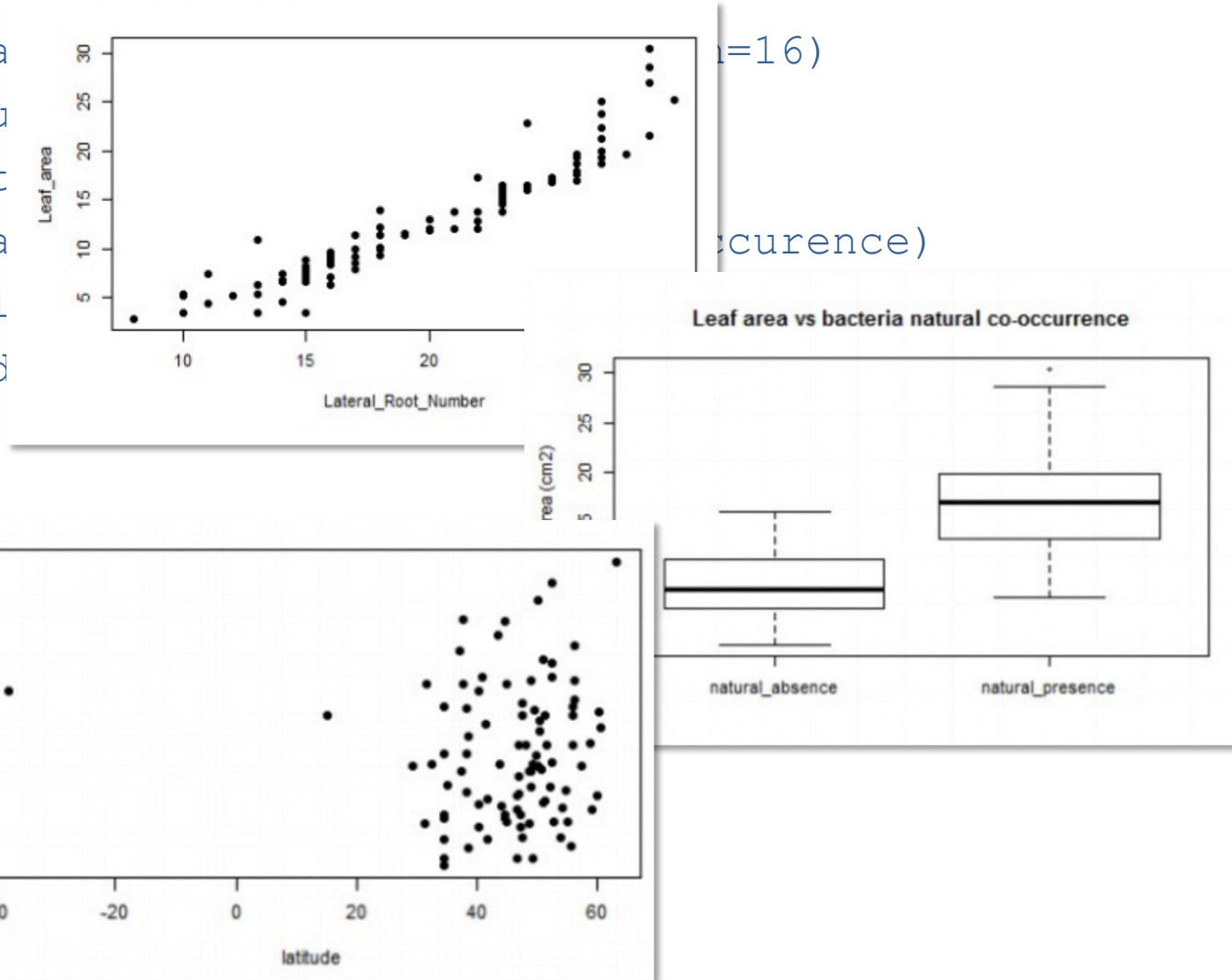
Traitement de données - statistiques


```

data=read.table("data_for_TP3.txt",h=T)
attach(data);names(data)

plot(Lateral_Root_Number,Leaf_area)
plot(Latitude,Leaf_area)
plot(Longitude,Leaf_area)
boxplot(Leaf_area~bacteria_natural_co-occurrence)
cor(Lateral_Root_Number,Leaf_area)
cor(Latitude,Leaf_area)

```





[Home](#) [Help](#) [Contact](#) [About Us](#) [Subscribe](#) [Login](#) [Register](#)

Search Browse Tools Portals Download Submit News ABRC Stocks

Locus: AT5G46330 Add a Comment


Representative Gene Model [AT5G46330.1](#)

Gene Model Type protein_coding

Other names: FLAGELLIN-SENSITIVE 2, FLS2, MPL12.8

Description ? Encodes a leucine-rich repeat serine/threonine protein kinase that is expressed ubiquitously. FLS2 is involved in MAP kinase signalling relay involved in innate immunity. Essential in the perception of flagellin, a potent elicitor of the defense response. FLS2 is directed for degradation by the bacterial ubiquitin ligase AvrPtoB. The mRNA is cell-to-cell mobile.

Map Detail Image

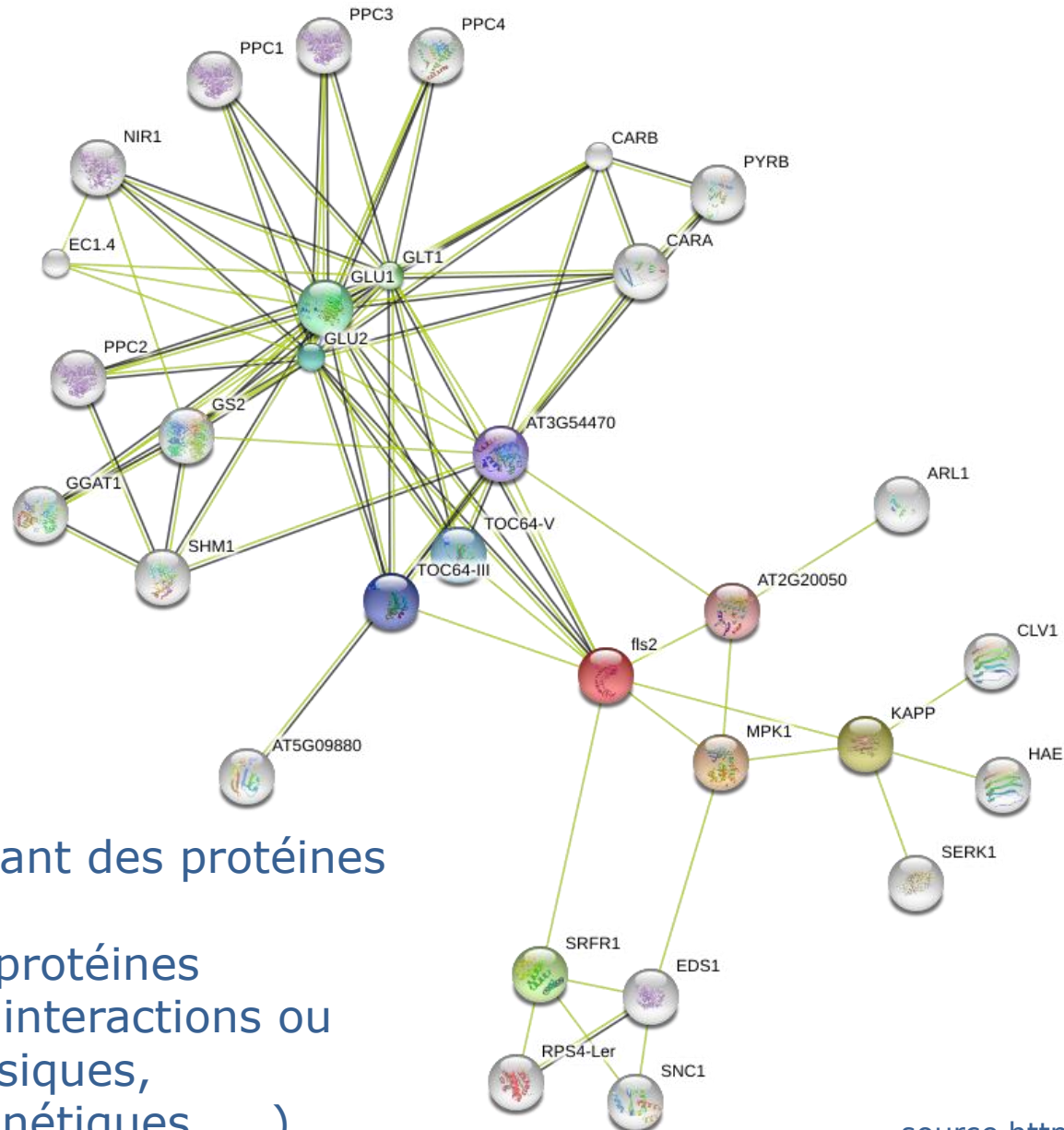


Chr5:18791736..18795546
18792k 18793k 18794k 18795k
Protein Coding Gene Models
AT5G46330.1 (FLS2)

Annotations ?

category	relationship type	keyword
GO Biological Process	involved in	defense response by callose deposition in cell wall, defense response to bacterium, detection of bacterium, protein phosphorylation, receptor-mediated endocytosis, regulation of anion channel activity, transmembrane receptor protein tyrosine kinase signaling pathway
GO Cellular Component	located in	endosome, endosome membrane, integral component of membrane, membrane, plasma membrane
GO Molecular Function	functions in	ATP binding
GO Molecular Function	has	ATP binding, kinase activity, protein binding, protein serine/threonine kinase activity, transmembrane receptor protein serine/threonine kinase activity
Growth and Developmental Stages	expressed during	LP.02 two leaves visible stage, LP.04 four leaves visible stage, LP.06 six leaves visible stage, LP.08 eight leaves visible stage, LP.10 ten leaves visible stage, LP.12 twelve leaves visible stage, flowering stage, petal differentiation and expansion stage, plant embryo globular stage, vascular leaf senescent stage
Plant structure	expressed in	carpel, cauline leaf, collective leaf structure, cotyledon, cultured plant cell, epidermal cell, flower, guard cell, hypocotyl, inflorescence meristem, leaf apex, leaf lamina base, petiole, plant embryo, pollen, root, root tip, sepal, stamen, stem, vascular leaf
user-defined	has gene product	cell-to-cell mobile RNA

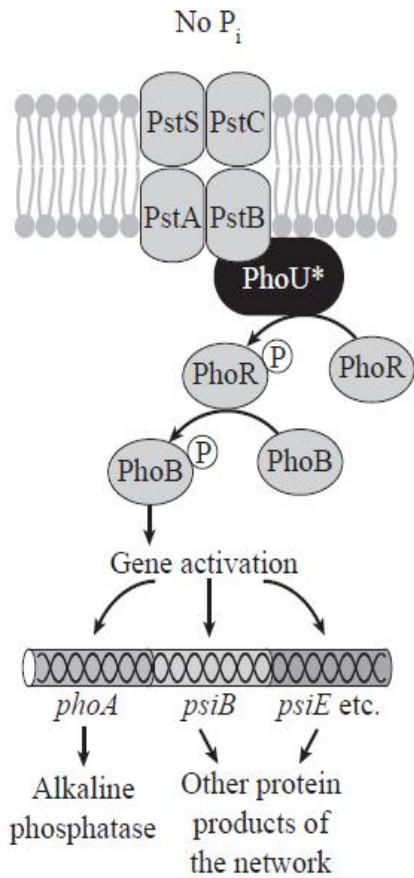
Annotation Detail



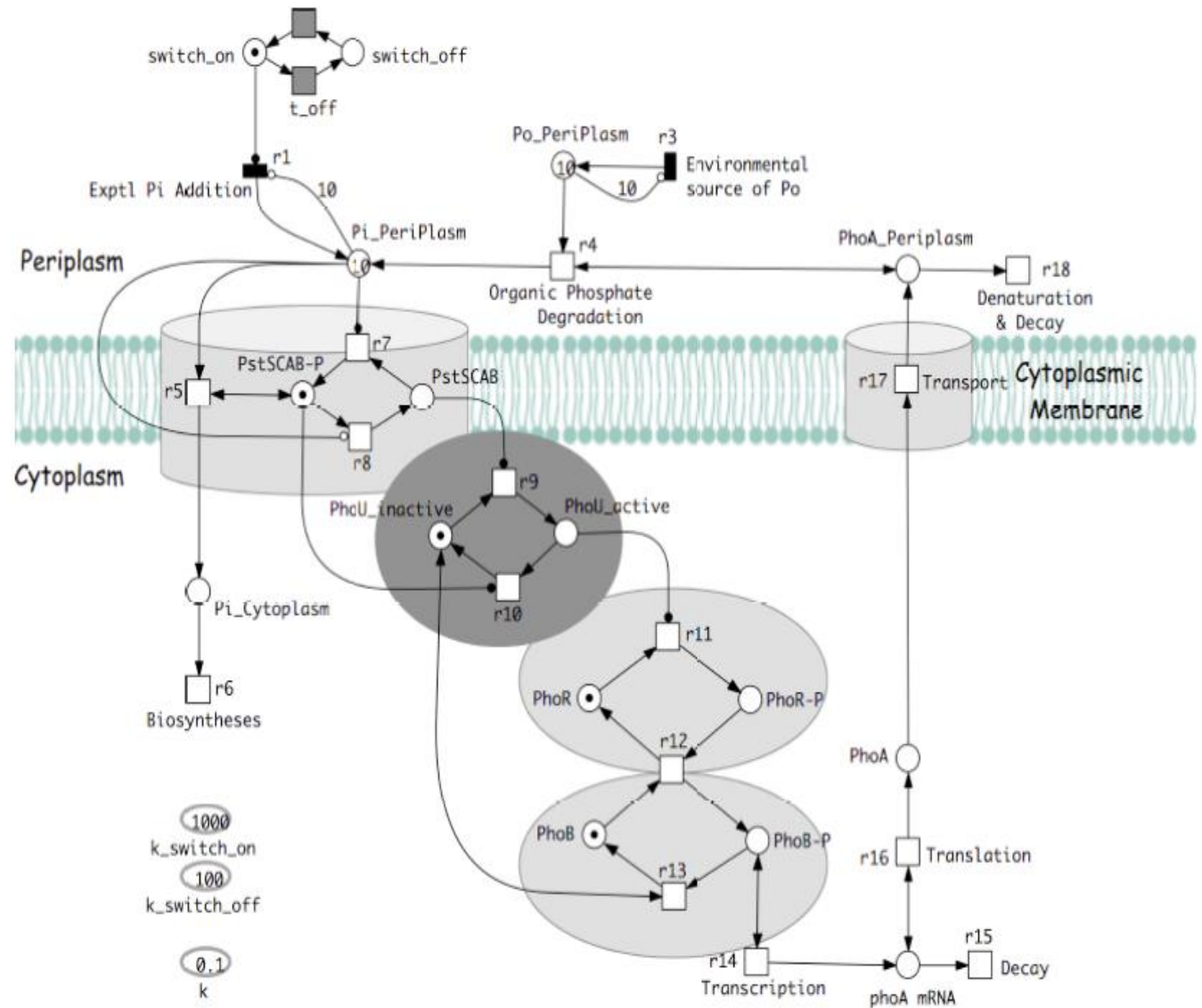
Graphe représentant des protéines

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

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

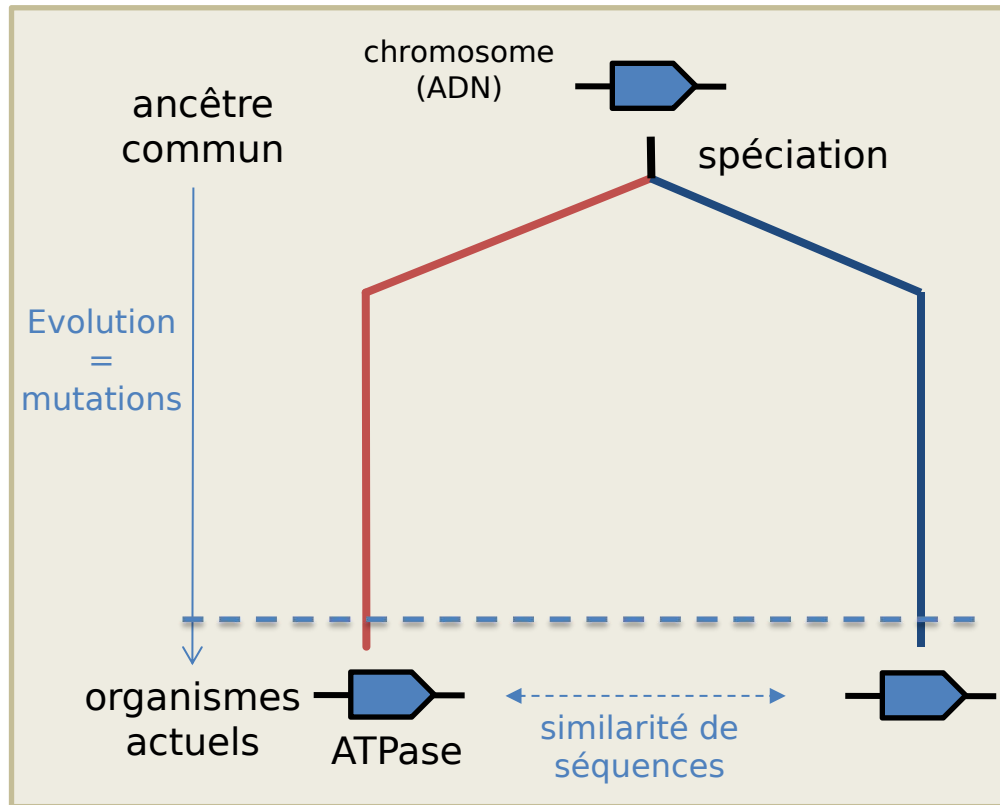
- définition : deux gènes sont orthologues si leur divergence a commencé après un événement de **spéciation** (le gène ancêtre se trouvait dans l'organisme ancêtre).
- remarque : deux gènes orthologues ne peuvent pas être présents dans le même génome.
- La relation d'orthologie est souvent abusivement 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 événements de spéciation et de duplication est importante.
- Deux orthologues 1:1 ont une forte probabilité d'avoir conservé la même fonction.

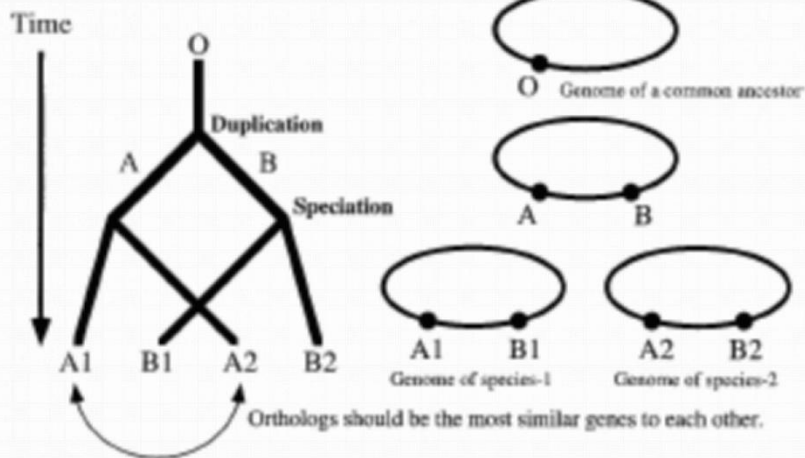


Du temps du séquençage des premiers gènes et génomes

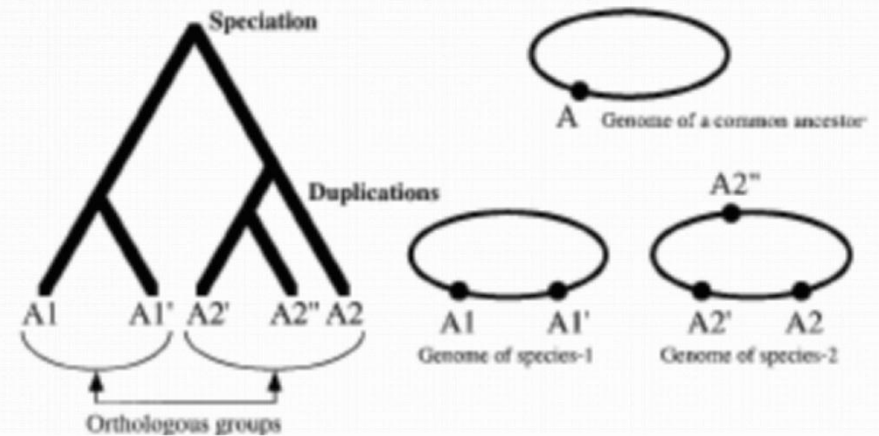
- la conservation/similarité des séquences impliquait des fonctions similaires
- les annotations des gènes caractérisés expérimentalement étaient transférées aux nouveaux gènes/génomes séquencés

Chronologie des évènements

(a) Orthologous Gene Pair

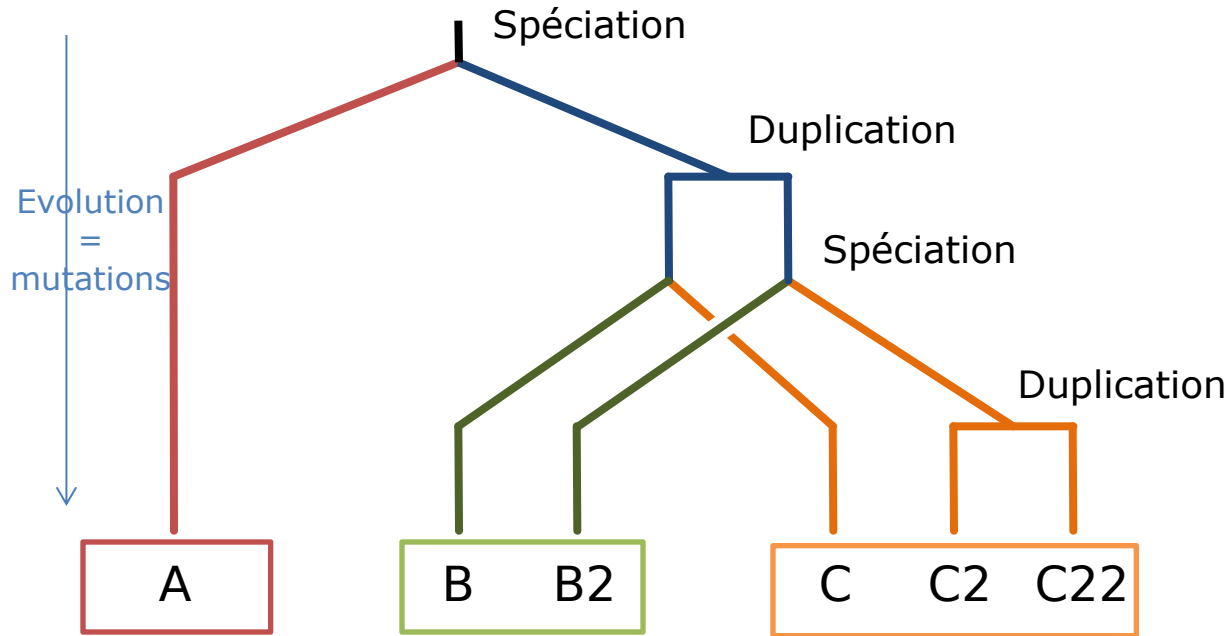


(b) Orthologous Gene Clusters

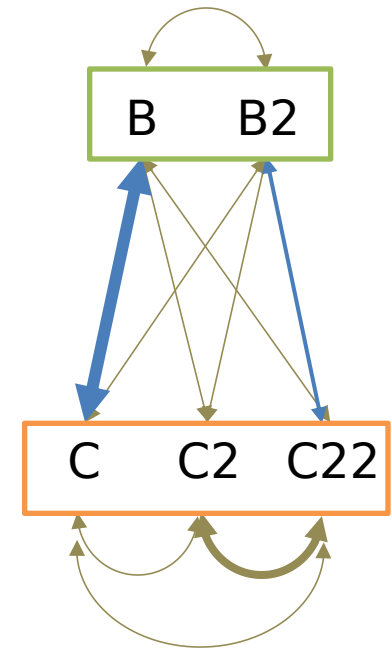


- L'orthologie 1:1 est essentielle pour la prédiction de la fonction des gènes (par transfert d'annotations)
- Problèmes :
 - la chronologie est nécessaire pour différencier paralogues, orthologues et orthologues 1:1
 - la chronologie n'est pas connue
- Deux principales approches pour identifier les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

Prédiction des orthologues 1:1

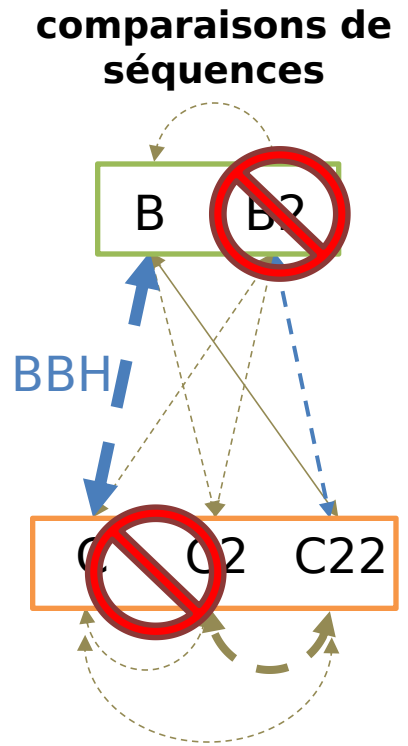
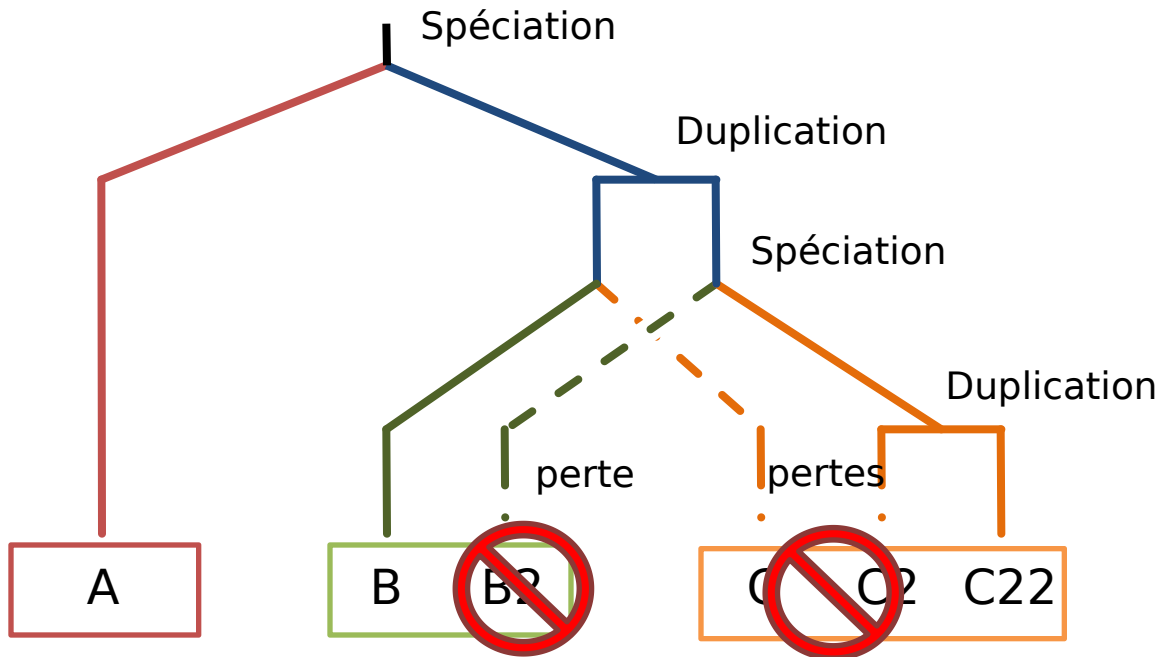


comparaisons de séquences

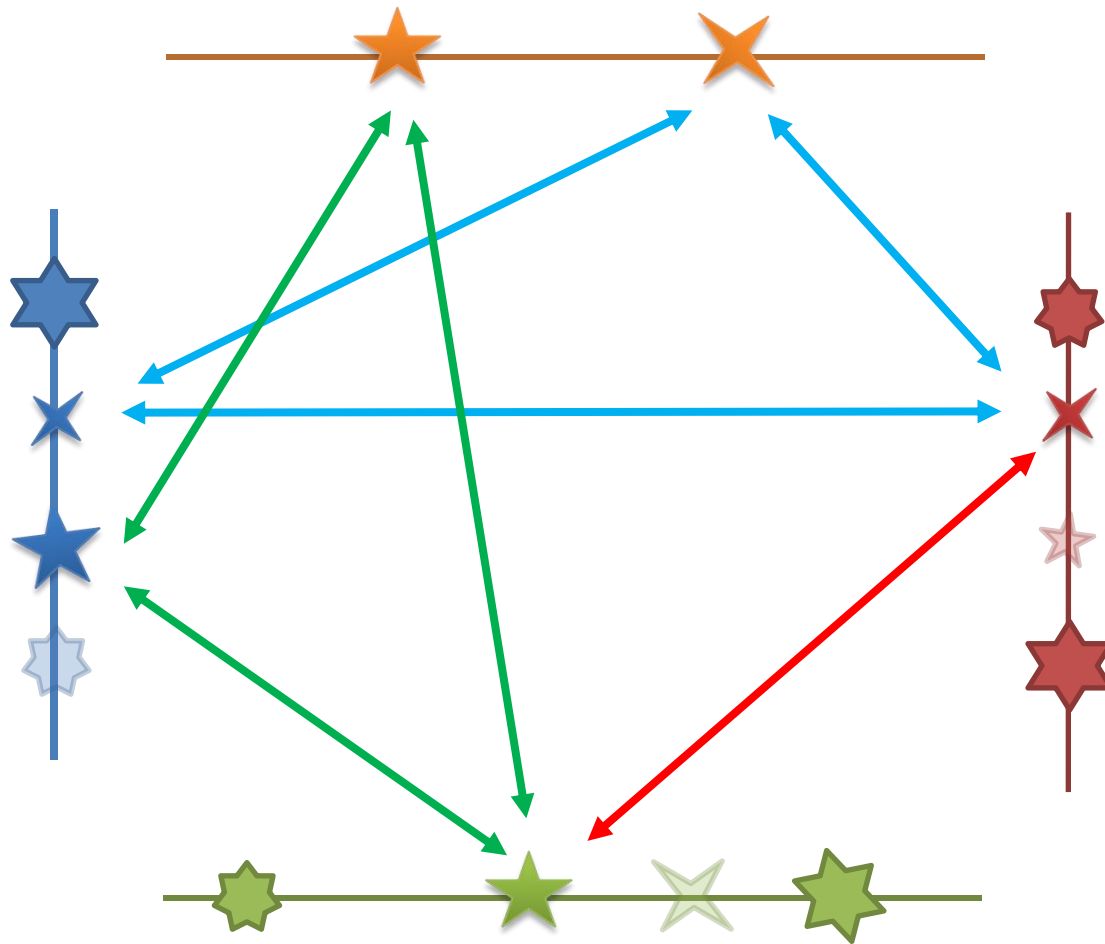


- comparaisons de toutes les séquences 2 à 2
- **B↔C prédit** $\text{sim}(B,C) > \max(\text{sim}(B,B2), \text{sim}(C, C2), \text{sim}(C,C22))$
- **B2↔C22 non prédit** $\text{sim}(B2,C22) < \text{sim}(C2,C22)$
- Remarque : calcul avec 7,223,104 séquences de 2 355 génomes =>52,000.10⁹ comparaisons

Autre évènement évolutif : perte de gènes



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



Elagage des erreurs de prédiction

Méthodes de traitement de graphes

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

