Bioinformatique

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

Contrôle continu : 30% Contrôle terminal : 70%

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

Intervenants

- Roland Barriot
- Maxime Bonhomme
- Franck Delavoie
- Gwennaele Fichant
- Elodie Gaulin
- Florie Gosseau
- Silvia Kocanova

Définition

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

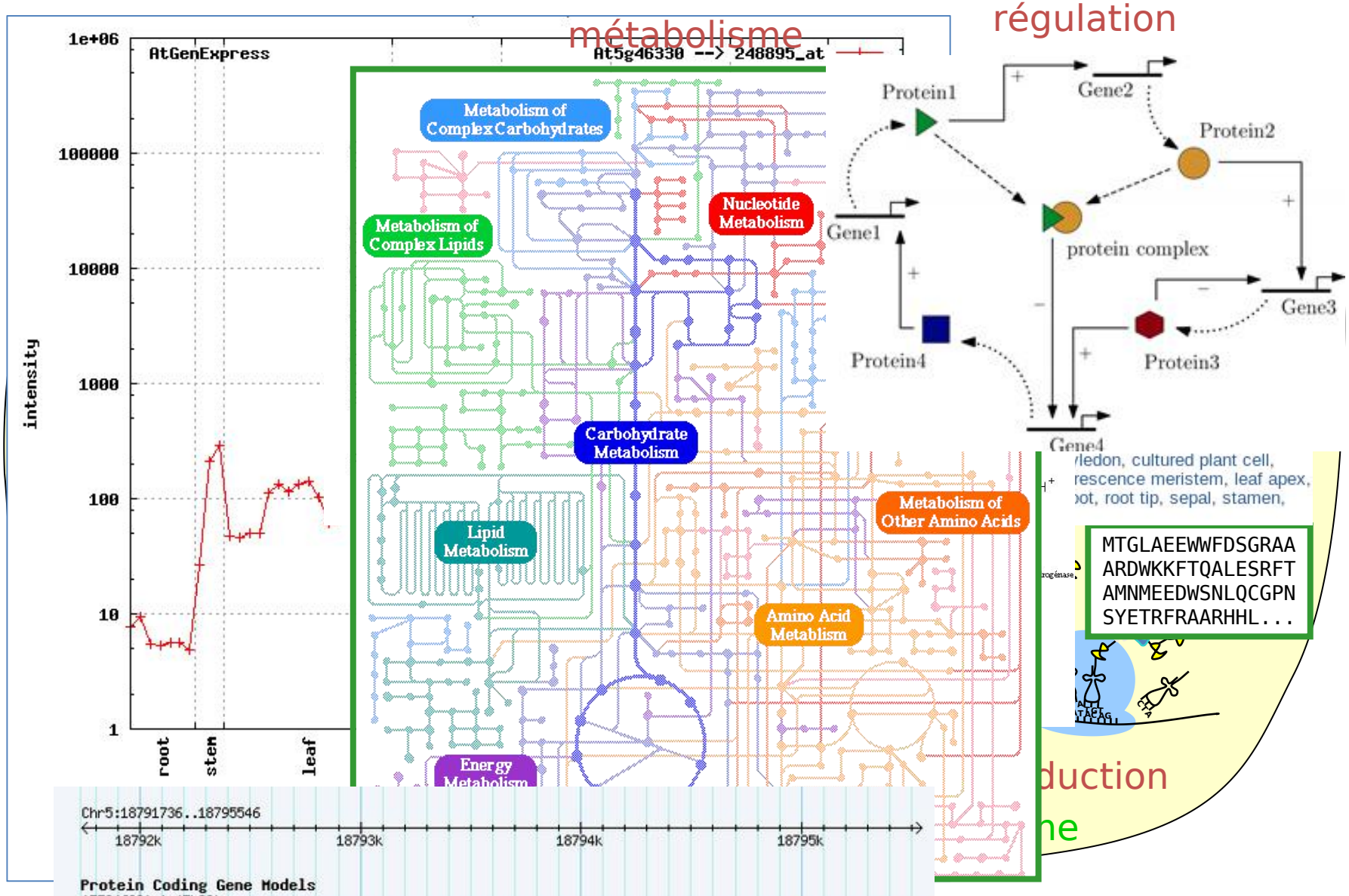
- Une définition : traitement des informations biologiques par des méthodes informatiques et/ou mathématiques.

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

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 Prédiction de fonction Banques de données
1990	Débuts de la génomique et génomique comparative ; analyse du contenu d'un génome 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 métaboliques 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 Modélisation d'une cellule, d'un organisme, d'une population, d'un écosystème Biologie de synthèse

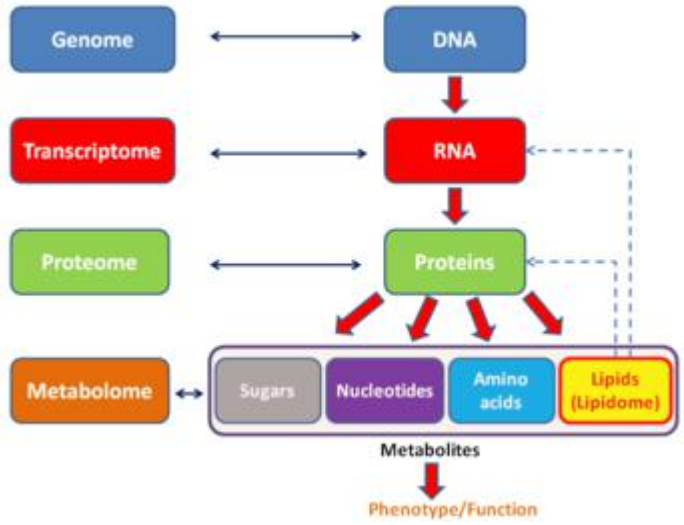
(Quelques) données et connaissances disponibles



Cellu

Données omics

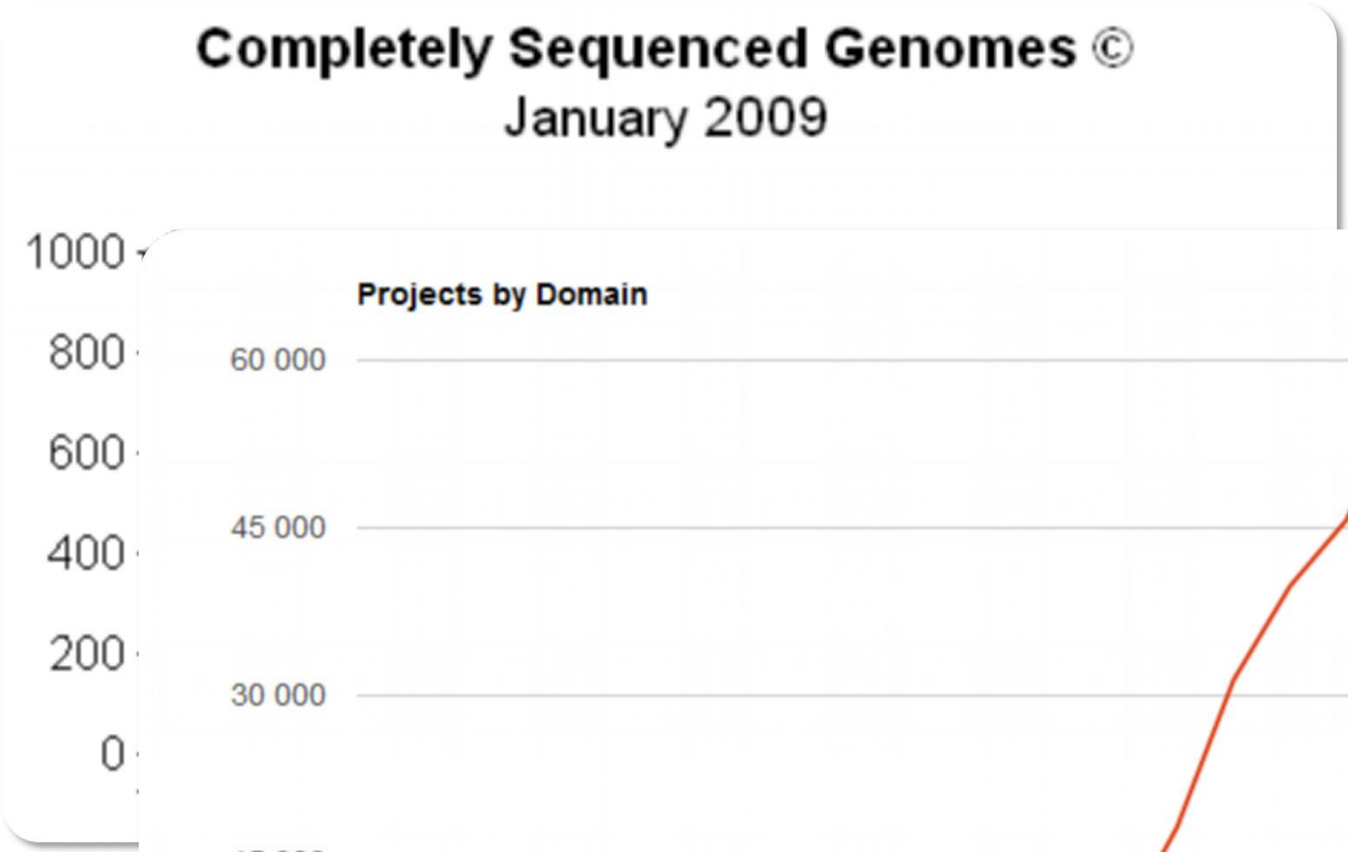
- **Génome**
  - séquence(s) nucléique(s) de l'ensemble des chromosomes d'un organisme
  - ensemble des gènes d'un organisme
- **Transcriptome**
  - ensemble des ARNm ou transcrits présents dans une cellule ou une population de cellules dans des conditions données



source : Wikipedia

- **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, …

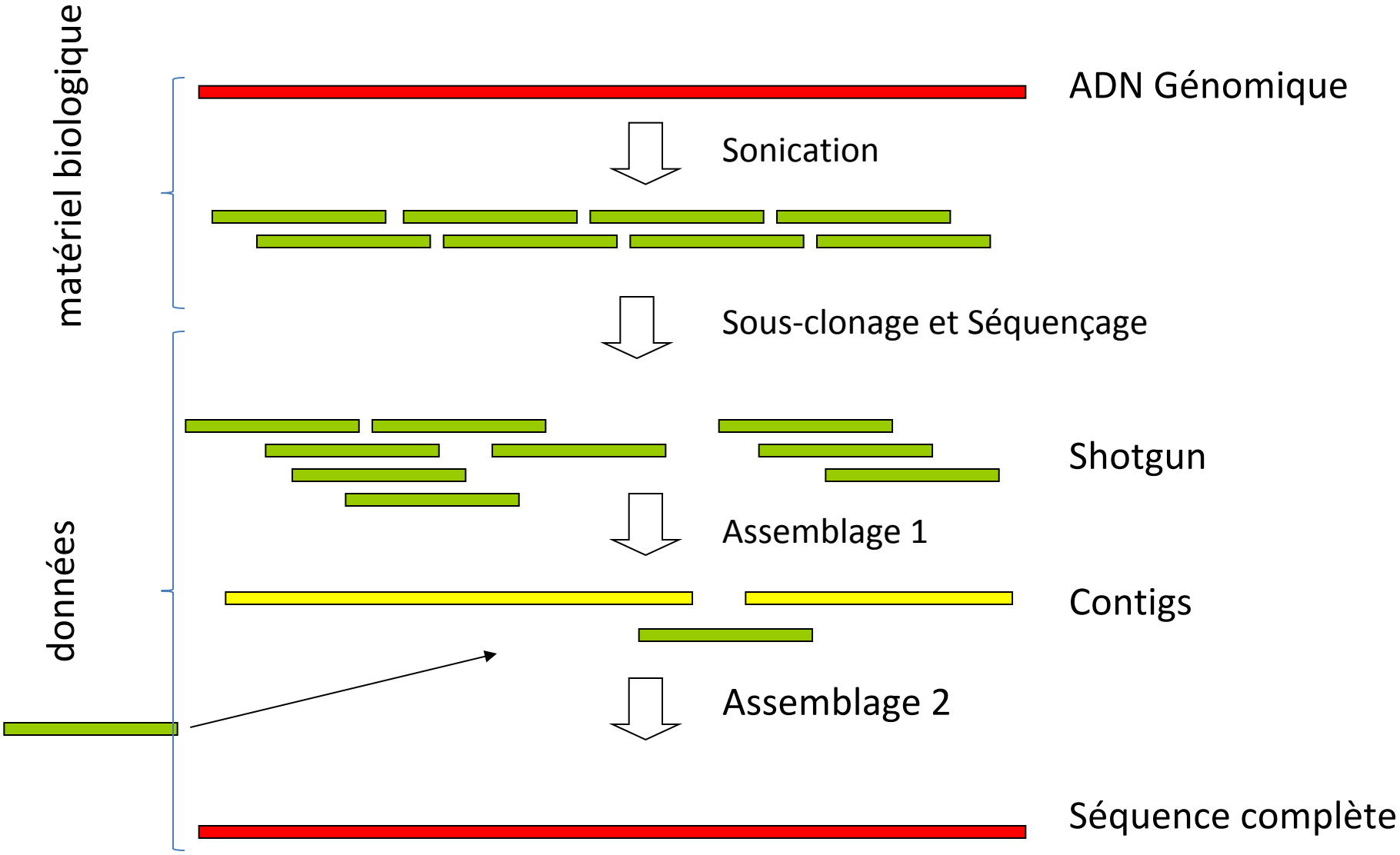
Séquences disponibles : quelques chiffres



Des séquences… et après ?

- Annotation
  - régions codantes, régions régulatrice, …
  - prédiction fonctionnelle
- Reconstruction du réseau métabolique
- Analyse des relations génotype/phénotype
- Analyses évolutives
- Conception de puces d’expression
- Identification de protéine
- Prédiction de structure

Assemblage d'un génome

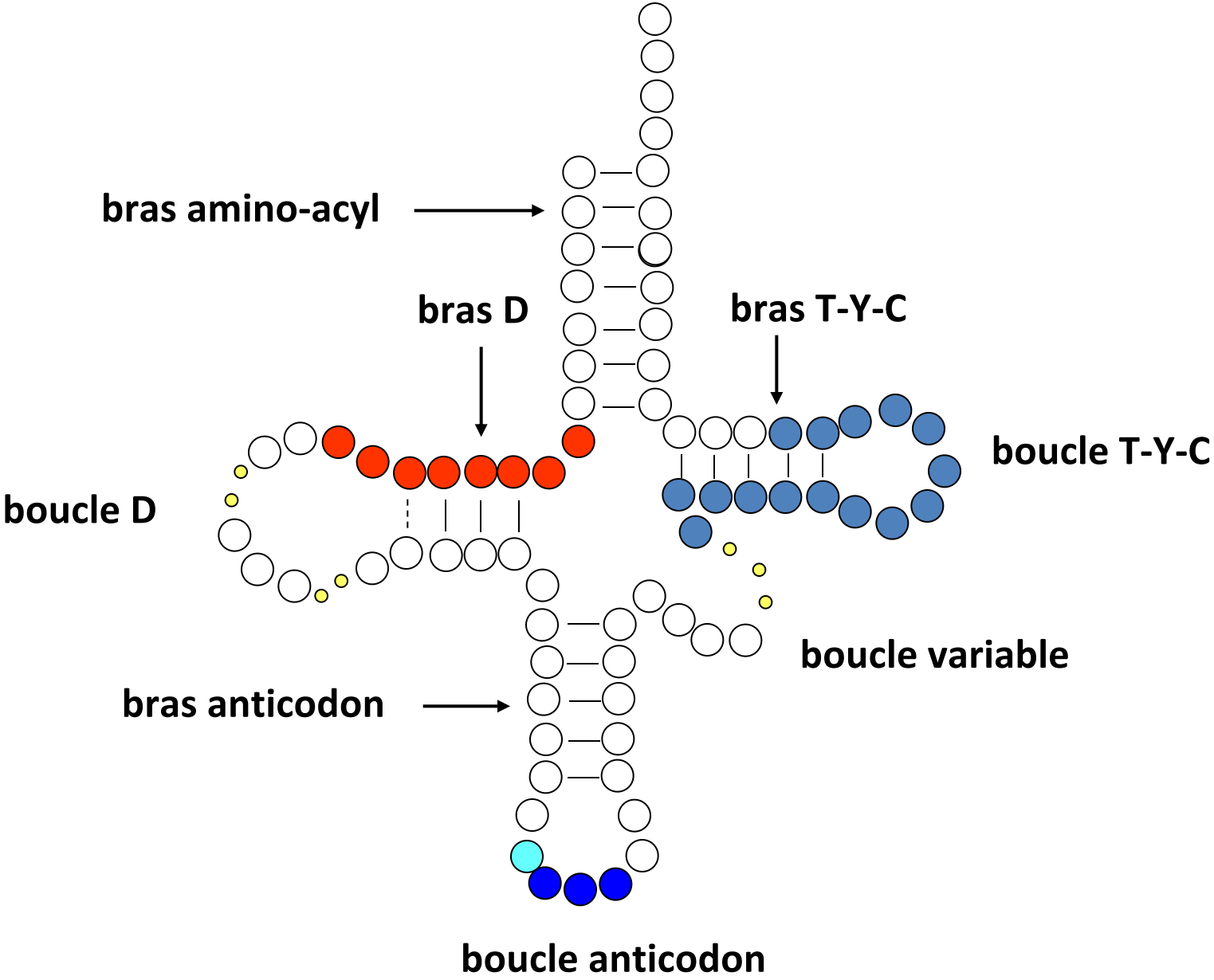


L’analyse manuelle d’une séquence peut s’avérer laborieuse

TCCTGGCCTACATGTTCTTTGGCAAAGGATCTTCAAAATCAACGGCTCCCGGTGCGGCGATCATCCATTTCTTCGGAGGGATT
CACGAGATTTACTTCCCGTACATTCTGATGAAACCTGGCCCTGATTCTCGCAGCCATTGCCGGCGGAGCAAGCGGACTCTTA
ACATTACGATCTTTAATGCCGGACTTGTCGCGGCAGCGTCACCGGGAAGCATTATCGCATTGATGGCAATGACGCCAAGAGG
AGGCTATTTCGGCGTATTGGCGGGTGTATTGGTCGCTGCAGCTGTATCGTTCATCGTTTCAGCAGTGATCCTGAAATCCTCTA
AAGCTAGTGAAGAAGACCTGGCTGCCGCAACAGAAAAAATGCAGTCCATGAAGGGGAAGAAAAGCCAAGCAGCAGCTGC
TTTAGAGGCGGAACAAGCCAAAGCAGAGAAGCGTCTGAGCTGTCTCCTGAAAGCGCGAACAAAATTATCTTTTCGTGTGAT
CCGGGATGGGATCAAGTGCCATGGGGGCATCCATCTTAAGAAACAAAGTGAAAAAGCGGAGCTTGACATCAGTGTGACCA
ACACGGCCATTAACAATCTGCCAAGCGATGCGGATATTGTCATCACCCACAAAGATTTAACAGACCGCGCGAAAGCAAAGCT
GCCGAACGCGACGCACATATCAGTGGATAACTTCTTAAACAGCCCGAAATACGACGAGCTGATTGAAAAGCTGAAAAGTAAT
CTTATAGAAAGAGAGTATTGTCATGCAAGTACTCGCAAAGGAAACATTAAACTCAATCAAACGGTATCATCAAAAGAAGAGGC
TATCAAATTGGCAGGCCAGACGCTGATTGACAACGGCTACGTGACAGAGGATTACATTAGCAAAATGTTTGACCGTGAAGAA
ACGTCTTCTACGTTTATGGGGAATTTCATTGCCATTCCACACGGCACAGAAGAAGCGAAAAGCGAGGTGCTTCACTCAGGAA
TTTCAATCATACAGATTCCAGAGGGCGTTGAGTACGGAGAAGGCAACACGGCAAAAGTGGTATTCGGCATTGCGGGTAAAA
ATAATGAGCATTTAGACATTTTGTCTAACATCGCCATTATCTGTTCAGAAGAAGAAACATTGAACGCCTGATCTCCGCTAAAGC
GAAGAAGATTTGATCGCCATTTCAACGAGGTGAACTGACATGATCGCCTTACATTTCGGTGCGGGAAATATCGGGAGAGGAT
TTATCGGCGCGCTGCTTCACCACTCCGGCTATGATGTGGTGTTTGCGGATGTGAACGAAACGATGGTCAGCCTCCTCAATGA
AAAAAAAGAATACACAGTGGAACTGGCGGAAGAGGGACGTTCATCGGAGATCATTGGCCCGGTGAGCGCTATTAACAGCG
GCAGTCAGACCGAGGAGCTGTACCGGCTGATGAATGAGGCGGCGCTCATCACAACAGCTGTCGGCCCGAATGTCCTGAAG
CTGATTGCCCCGTCTATCGCAGAAGGTTTAAGACGAAGAAATACTGCAAACACACTGAATATCATTGCCTGCGAAAATATGAT
TGGCGGAAGCAGCTTCTTAAAGAAAGAAATATACAGCCATTTAACGGAAGCAGAGCAGAAATCCGTCAGTGAAACGTTAGG
TTTTCCGAATTCTGCCGTTGACCGGATCGTCCCGATTCAGCATCATGAAGACCCGCTGAAAGTATCGGTTGAACCATTTTTCG
AATGGGTCATTGATGAATCAGGCTTTAAAGGGAAAACACCAGTCATAAACGGCGCACTGTTTGTTGATGATTTAACGCCGTA
CATCGAACGGAAGCTGTTTACGGTCAATACCGGACACGCGGTCACAGCGTATGTCGGCTATCAGCGCGGACTCAAAACGGT
CAAAGAAGCAATTGATCATCCGGAAATCCGCCGTGTTGTTCATTCGGCGCTGCTTGAAACTGGTGACTATCTCGTCAAATCGT
ATGGCTTTAAGCAAACTGAACACGAACAATATATTAAAAATCAGCGGTCGCTTTTAAAATCCTTTCATTTCGGACGATGTGACC
CGCGTAGCGAGGTCACCTCTCAGAAAACTGGGAGAAAATGTAGACTTGTAGGCCCGGCAAAGAAAATAAAAGAACCGAAT
GCACTGGCTGAAGGAATTGCCGCAGCACTGCGCTTCGATTTCACCGGTGACCCTGAAGCGGTTGAACTGCAAGCGCTGAT
CGAAGAAAAGGATACAGCGGCGTACTTCAAGAGGTGTGCGGCATTCAGTCCCATGAACCGTTGCACGCCATCATTTTAAAG
AAACTTAATCAATAACCGACCACCCGTGACACAATGTCACGGGCTTTTTACTATCTCGCAATCTAGTATAATAGAAAGCGCTTA
CGATAACAGGGGAAGGAGAATGACGATGAAACAATTTGAGATTGCGGCAATACCGGGAGACGGAGTAGGAAAGAGGTTGT
AGCGGCTGCTGAGAAAGTGCTTCATACAGCGGCTGAGGTACACGGAGGTTTGTCATTCTCATTCACAGCTTTTCCATGGAGC
TGTGATTATTACTTGGAGCACGGCAAAAATGATGCCCGAAGATGGAATACATACGCTTACTCAATTTGAAGCAGTTTTTGGGA
GCTGTCGGAAATCCGAAGCTGGTTCCCGATCATATATCGTTATGGGGCTGCTGCTGAAATCCGGAGGGAGCTTGAGCTTTCC

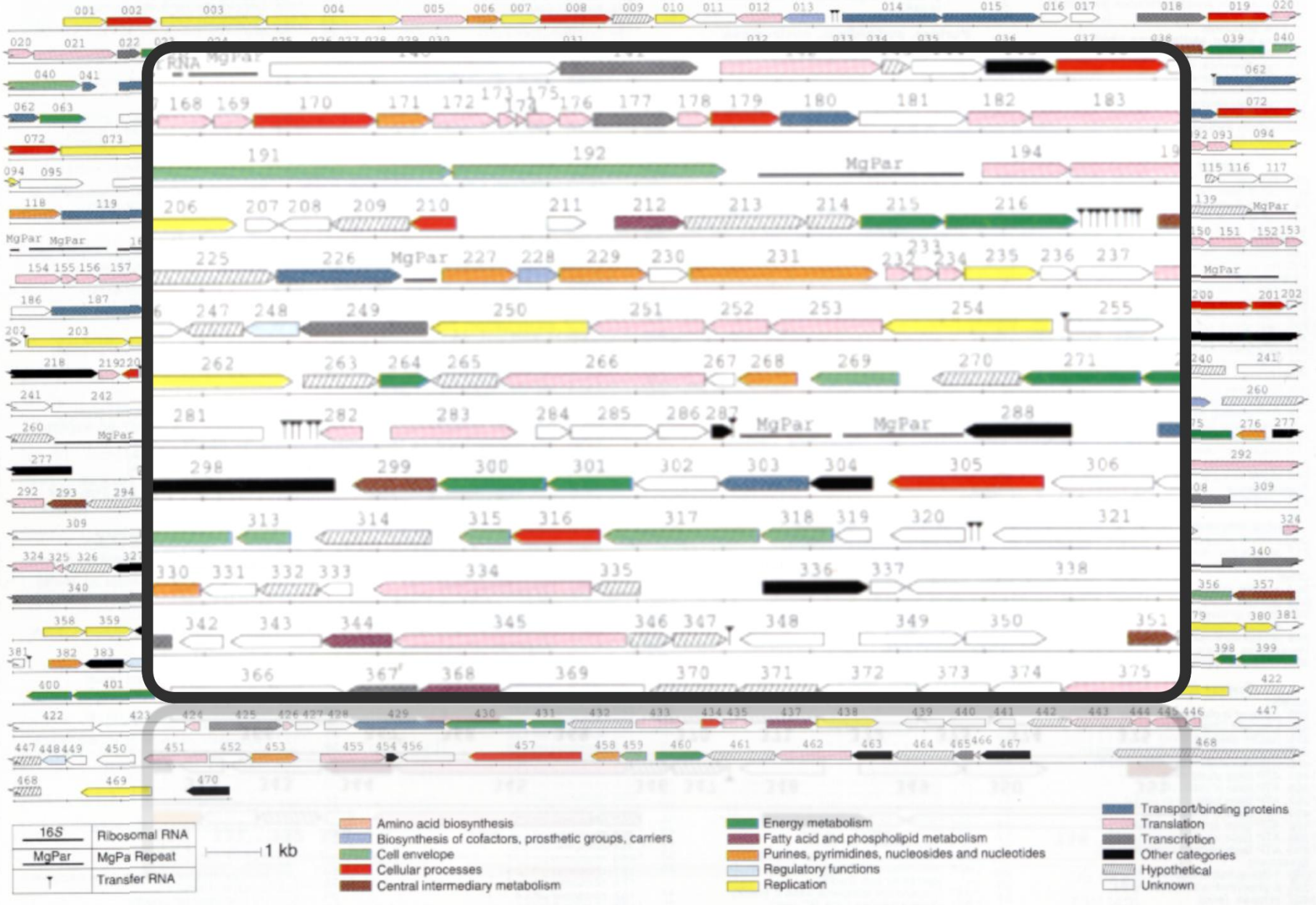
- Identification des gènes codant pour :
  - les ARNr
  - les ARNt
  - les protéines
- Identification des unités de traduction
- Identification des unités de transcription (promoteur et terminateur)
- Pour les gènes codant pour les protéines, prédiction fonctionnelle par recherche de similarité de séquences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

Structure secondaire canonique d'une séquence d'ARNt



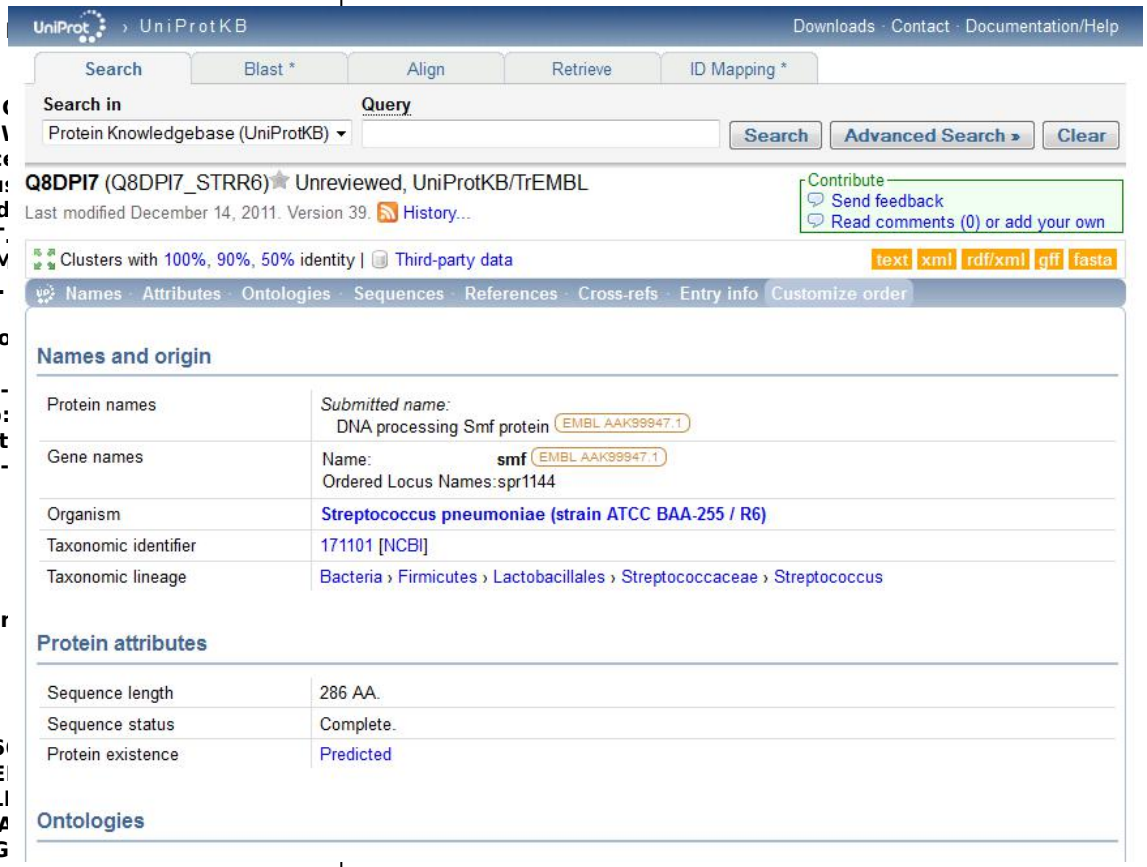
Génome de *Mycoplasma genitalium*

Distribution des unités de traduction et classification fonctionnelle



Stockage des séquences : Banques de séquences

```
ID   Q8DPI7_STRR6   PRELIMINARY;   PRT;   286 AA.
AC   Q8DPI7;
DT   01-MAR-2003, integrated into UniProtKB/TrEMBL.
DT   01-MAR-2003, sequence version 1.
DT   02-MAY-2006, entry version 10.
DE   DNA processing Smf protein.
GN   Name=smf; OrderedLocusNames=spr1144;
OS   Streptococcus pneumoniae (strain ATCC BAA-255 / R6).
OC   Bacteria; Firmicutes; Lactobacillales; Streptococcaceae;
OC   Streptococcus.
OX   NCBI_TaxID=171101;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC
RX   MEDLINE=21429245; PubMed=11544234;
RX   DOI=10.1128/JB.183.19.5709-5717.2001;
RA   Hoskins J., Alborn W.E. Jr., Arnold J., Blaszczak L.
RA   DeHoff B.S., Estrem S.T., Fritz L., Fu D.-J., Fuller
RA   Gilmour R., Glass J.S., Khoja H., Kraft A.R., Lagace
RA   LeBlanc D.J., Lee L.N., Lefkowitz E.J., Lu J., Matsus
RA   McAhren S.M., McHenney M., McLeaster K., Mund
RA   Norris F.H., O'Gara M., Peery R.B., Robertson G.T.
RA   Sun P.-M., Winkler M.E., Yang Y., Young-Bellido M
RA   Zook C.A., Baltz R.H., Jaskunas S.R., Rosteck P.R.
RA   Glass J.I.;
RT   "Genome of the bacterium Streptococcus pneumo
RL   J. Bacteriol. 183:5709-5717(2001).
CC   ---------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http:
CC   Distributed under the Creative Commons Attribut
CC   ---------------------------------------------------------------
DR   EMBL; AE008487; AAK99947.1; -; Genomic_DNA.
DR   PIR; A95147; A95147.
DR   PIR; G98014; G98014.
DR   GenomeReviews; AE007317_GR; spr1144.
DR   BioCyc; SPNE1313:SPR1144-MONOMER; -.
DR   GO; GO:0009294; P:DNA mediated transformatior
DR   InterPro; IPR003488; SMF.
DR   Pfam; PF02481; SMF; 1.
DR   TIGRFAMs; TIGR00732; dprA; 1.
KW   Complete proteome.
SQ   SEQUENCE   286 AA;  31583 MW;  CF12DB83AE36
     MELFMKITNY EIYKLKKSGL TNQQILKVLE YGENVDQE
     FQIDDAHLSK EFQKFPSFSI LDDCYPWDLS EIYDAPVL
     CSKQGAKSVE KVIQGLENEL VIVSGLAKGI DTAAHMAA
     NKRLQDYIGN DHLVLSEYGP GEQPLKFHFP ARNRIIAG
     AMEEGRDVFA IPGSILDGLS DGCHHLIQEG AKLVTSGQDV LAEFEF
//
```



Exemple d'une entrée protéique dans la banque de données UniProt

Transcriptome

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

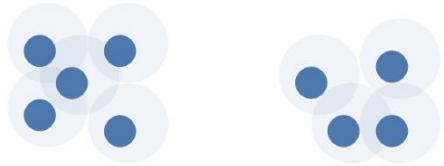
➡ Accès au niveau d'expression de milliers de gènes simultanément (potentiellement l'ensemble des gènes d'un organisme)
= *instantané* de l'état d'une cellule ou d'une population de cellules

Données d'expression des gènes obtenues par :

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

Acquisition des données (RNAseq)

Échantillon 1 Échantillon 2

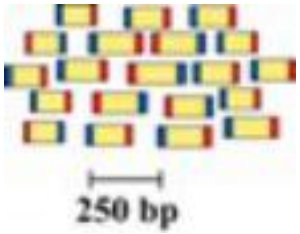


extraction

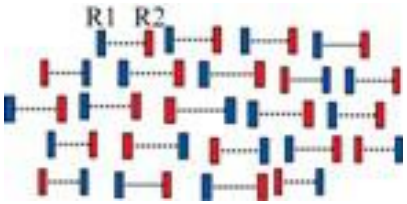
ARNm

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

ADNc

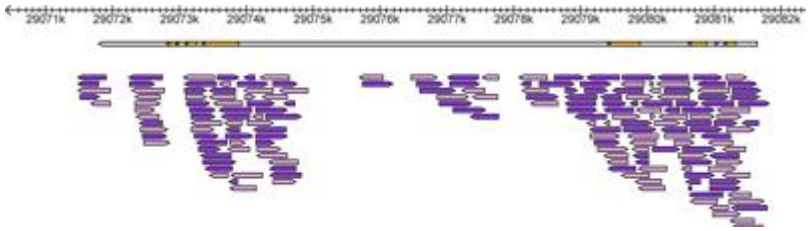


séquençage

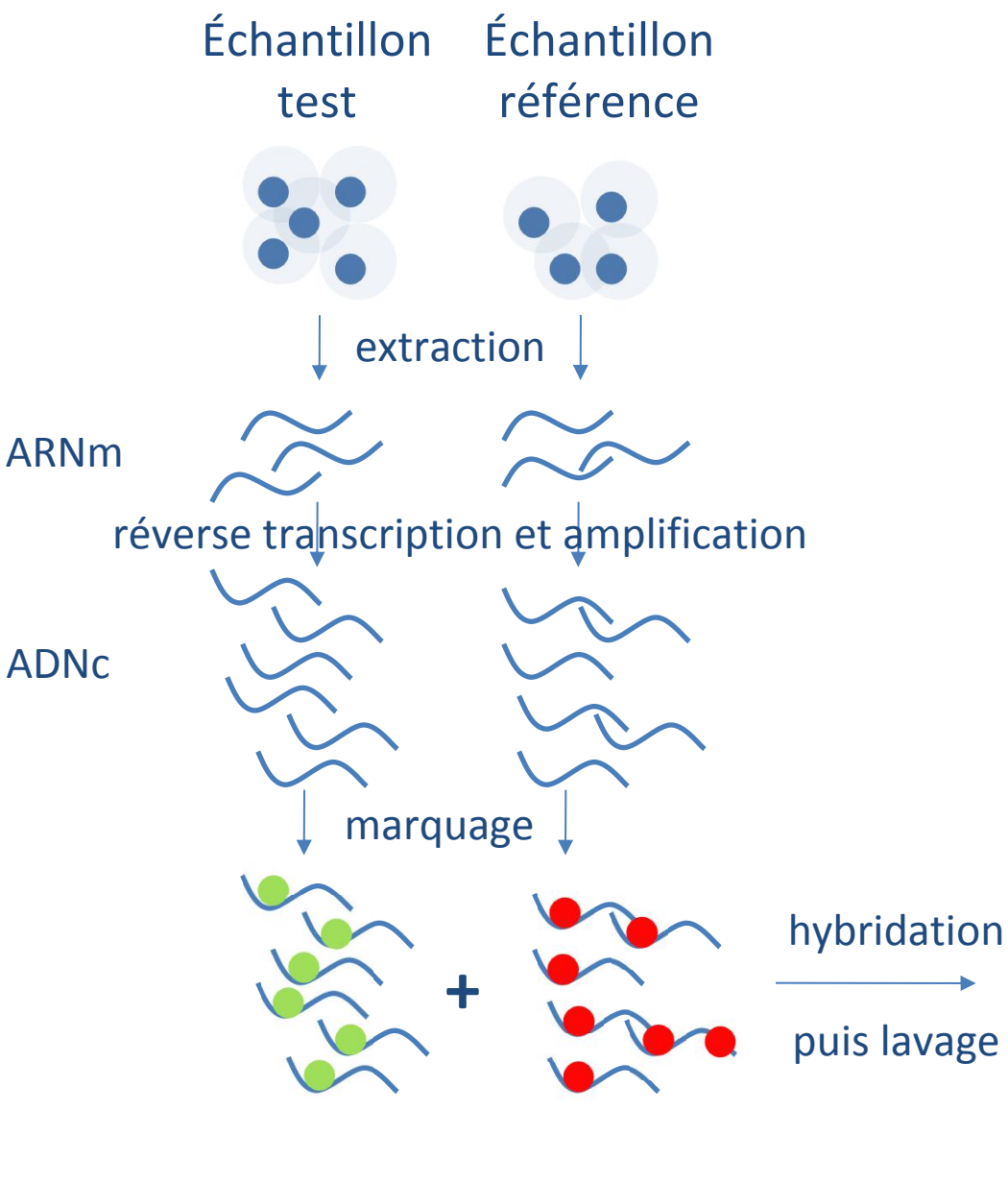


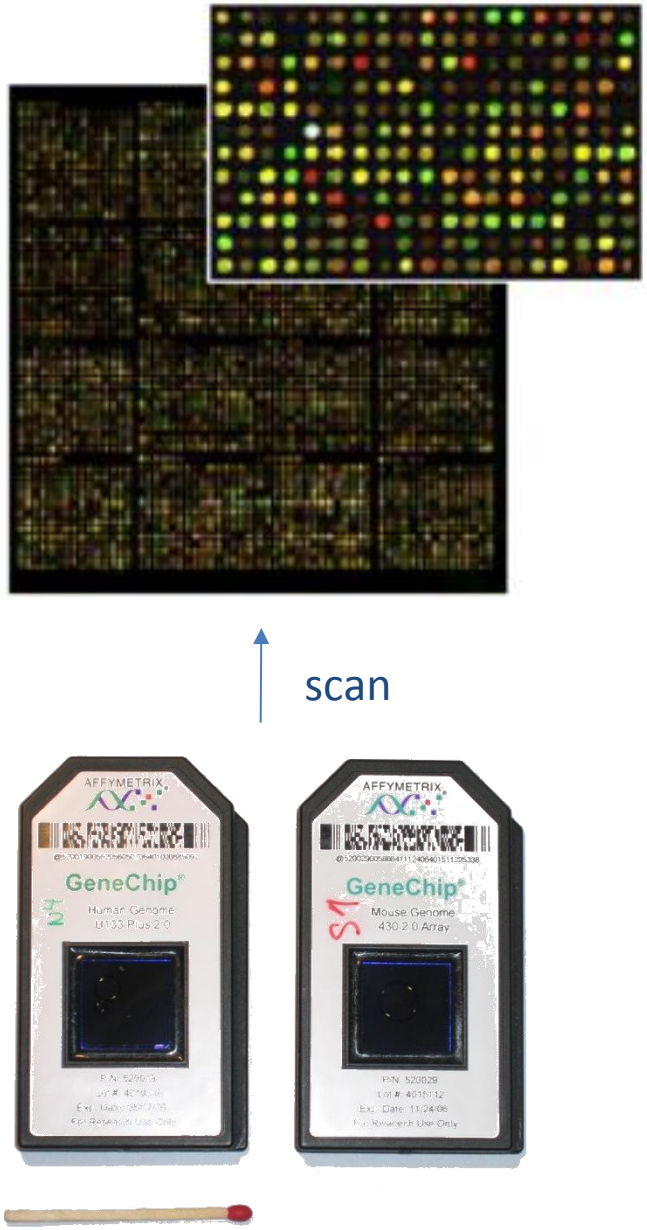
(centaines de)
millions de séquences
(reads)

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

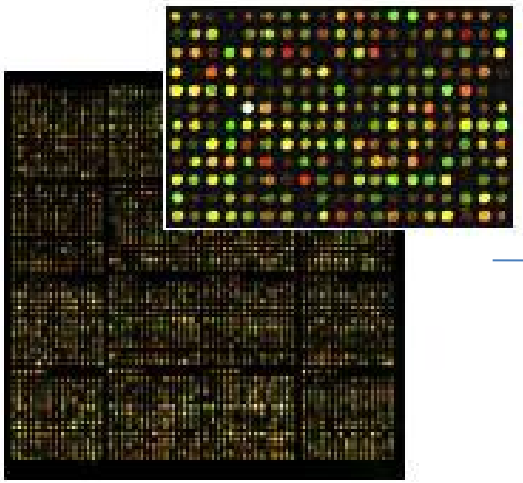


Transcriptome : acquisition des données





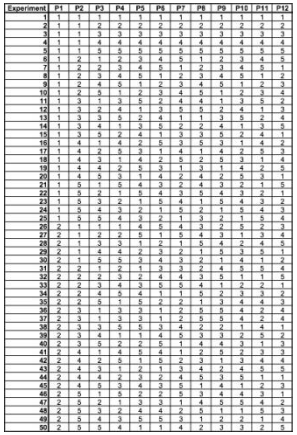
Transcriptome : analyse et interprétation des données



acquisition

mesures

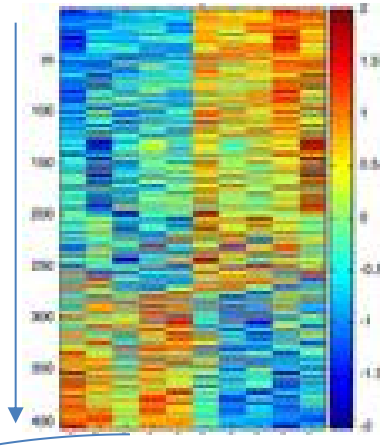
probesets



normalisation
filtrage

conditions

gènes



interprétation

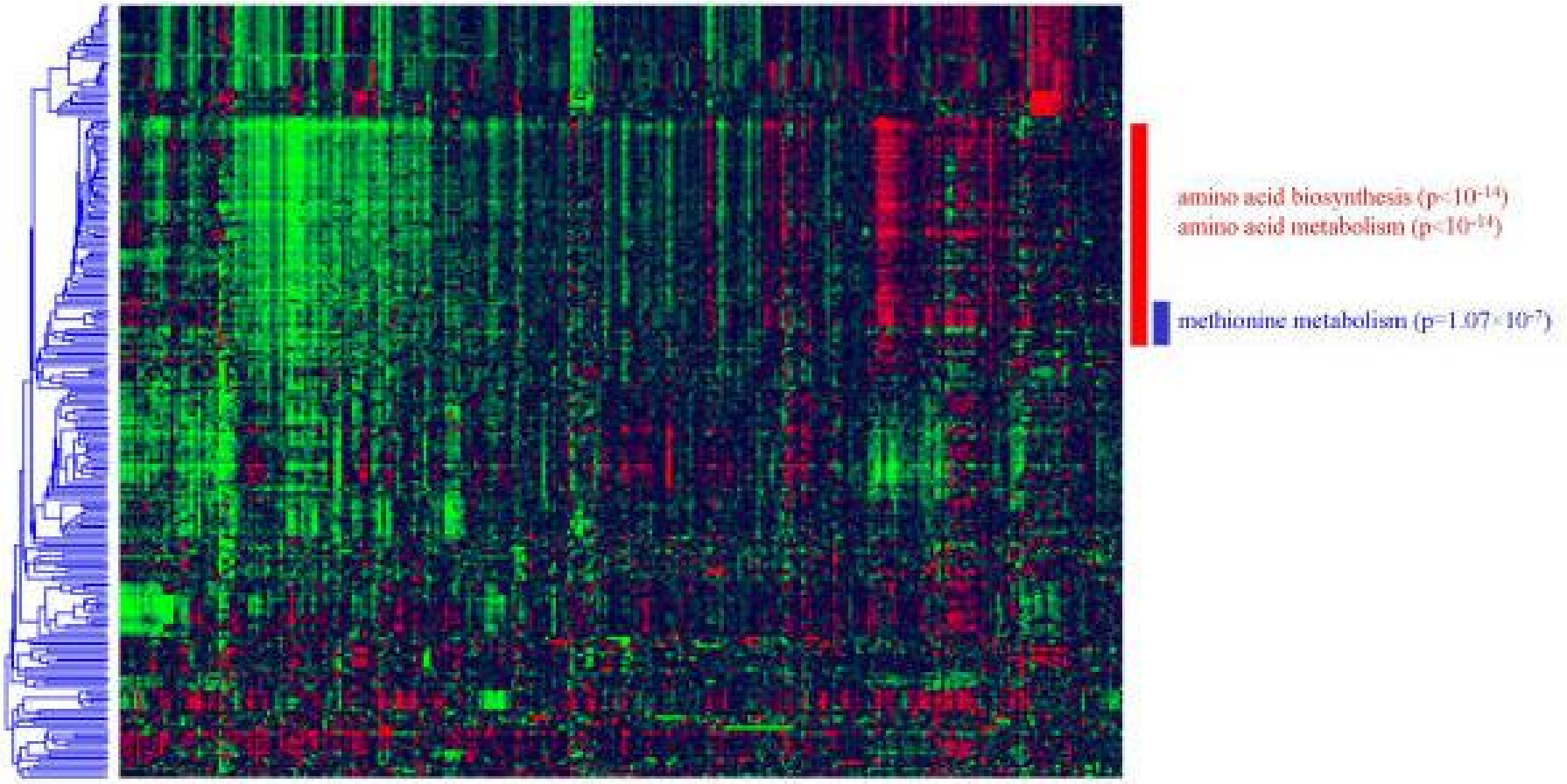
Identification des **gènes différentiellement exprimés**

Identification des ensembles de **gènes co-exprimés**

Caractérisation d'un ensemble de gènes

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



Robinson *et al.* 2002

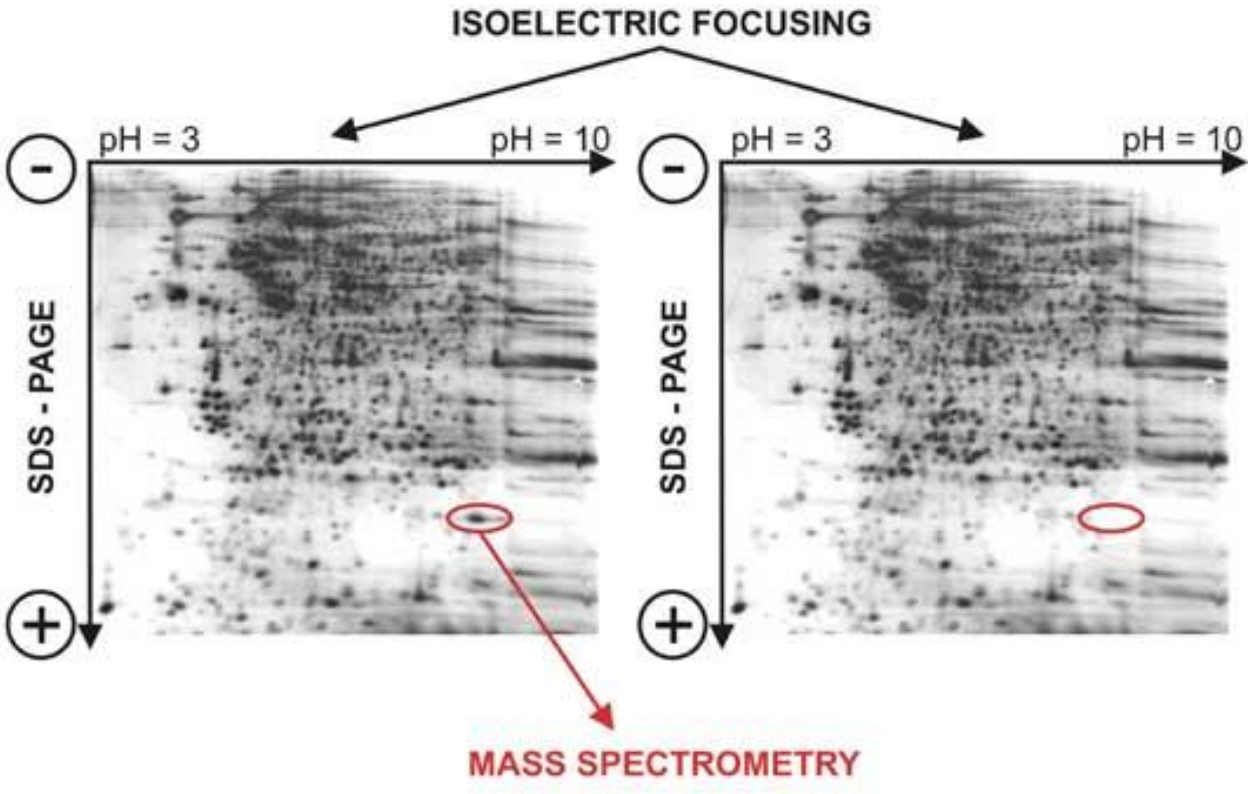
Protéomique

Protéome : ensemble des protéines exprimées dans une cellule, une partie d'une cellule (membranes, organites) ou un groupe de cellules (organe, organisme, groupe d'organismes) dans des conditions données et à un moment donné.

= *instantané* de l'état d'une cellule ou d'une population de cellules

Séparation des protéines par gels d'électrophorèse (1D, 2D) puis identification des spots par spectrométrie de masse

Échantillon test



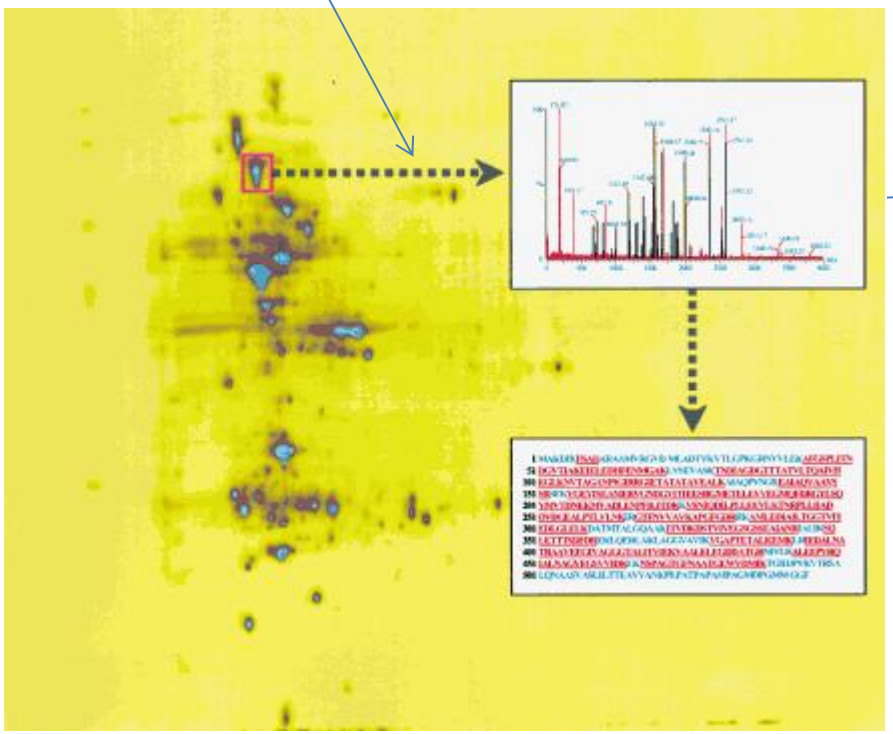
Échantillon référence

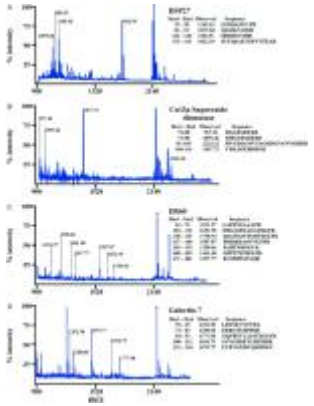
Protéomique : Identification de protéine

Digestion du spot par une enzyme (ex: trypsine) et mesure du poids des peptides obtenus

Digestion *in silico* du protéome

Recherche des protéines correspondant au profil observé

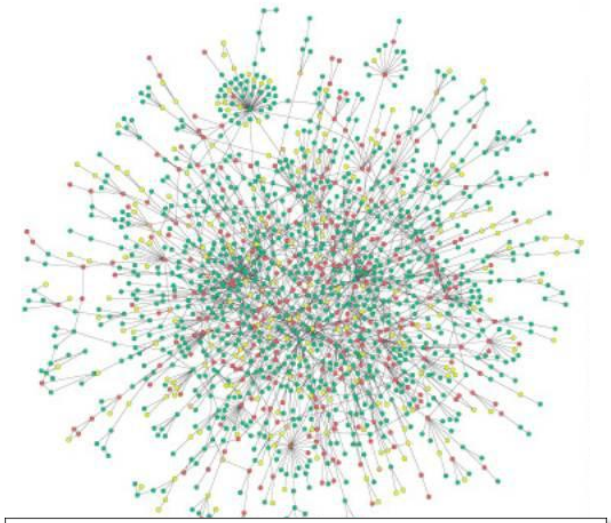


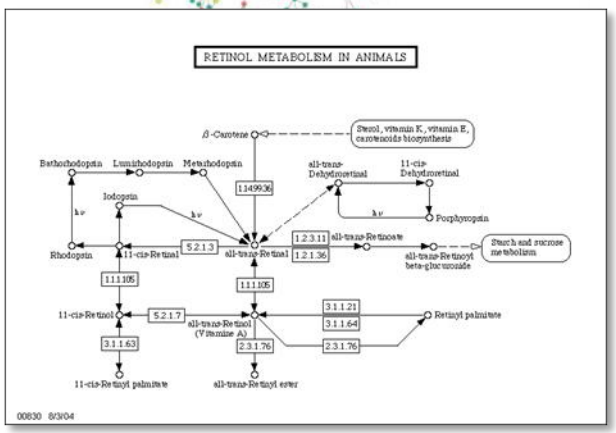


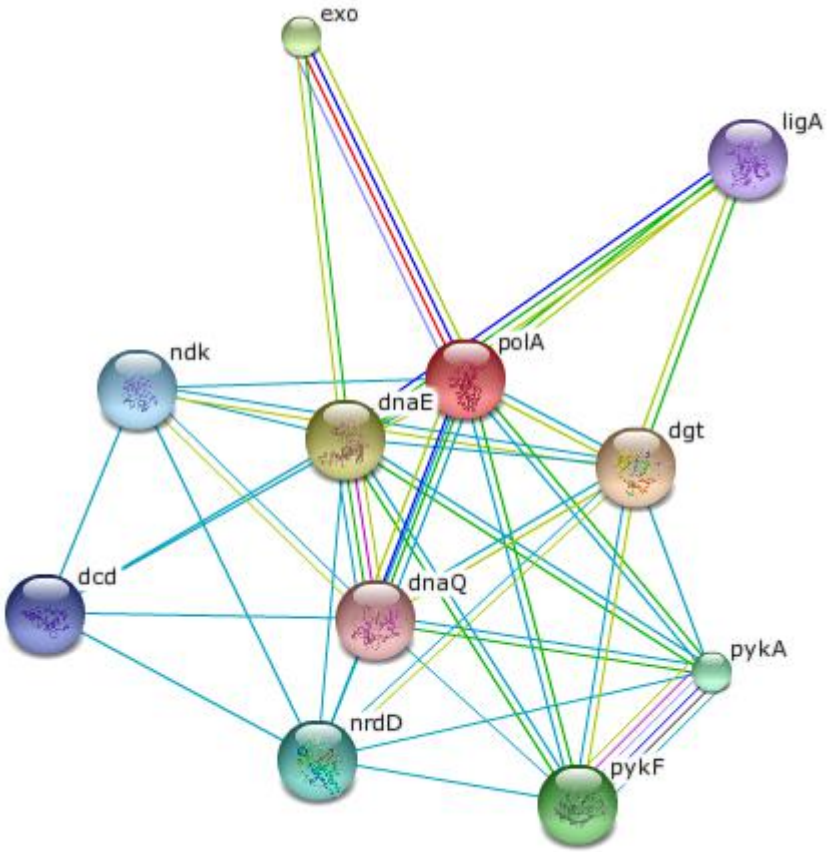
Réseaux de gènes et de protéines

Réseaux :

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





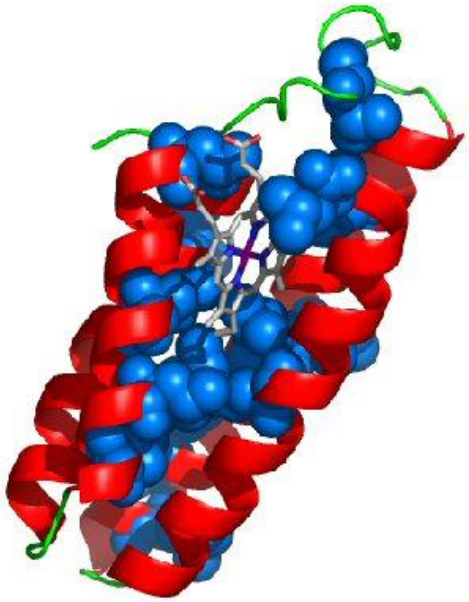


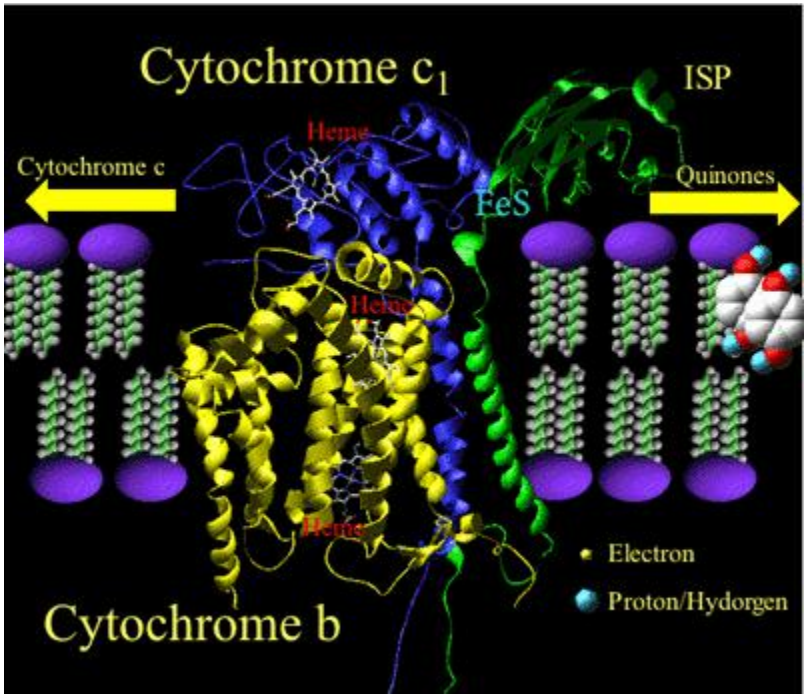
Prédiction de structure

Séquence protéique

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

Prédiction ou résolution de la structure tridimensionnelle

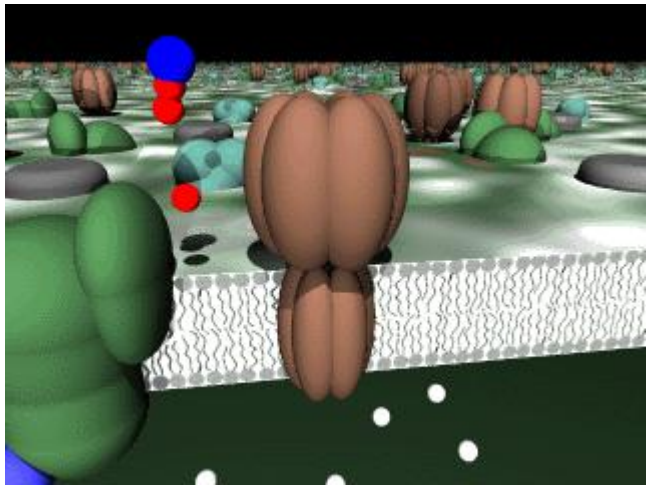




Biologie des systèmes

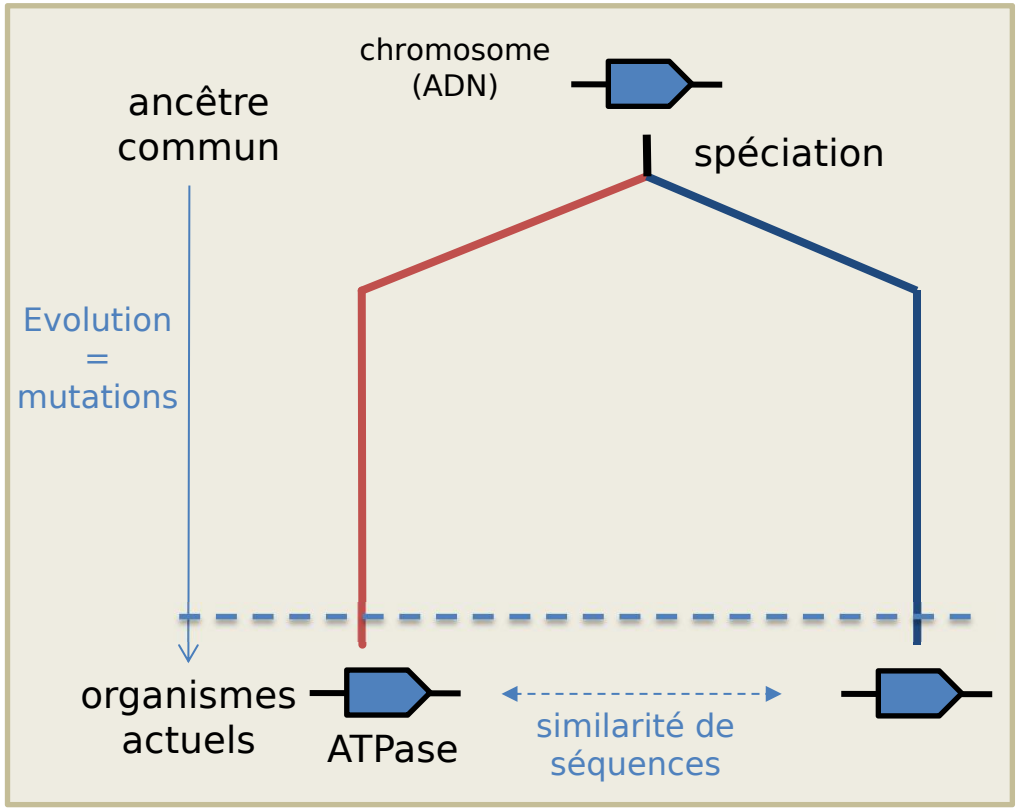
Intégration et synthèse des connaissances

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



À terme : simulation d'une cellule virtuelle et prédiction de son comportement

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 transférés aux nouveaux gènes/génomes séquencés

Homologie, orthologie et orthologie 1:1

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

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.

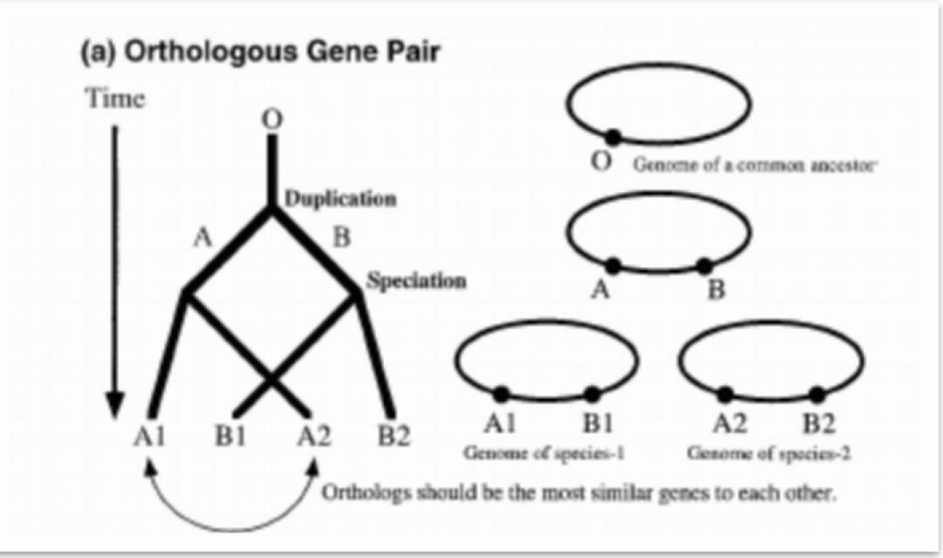
Orthologie

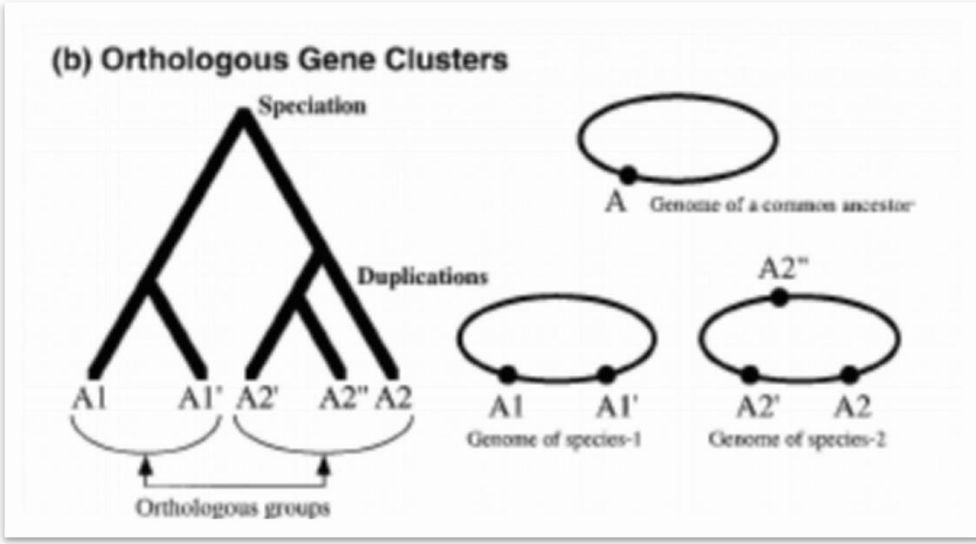
- définition : deux gènes sont ortholgues si leur divergence a commencé après un évènement de **spéciation** (le gène ancêtre se trouvait dans l'organisme ancêtre).
- remarque : deux gènes ortologues ne peuvent pas être présents dans le même génome.
- La relation d'orthologie est souvent abusivement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

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.

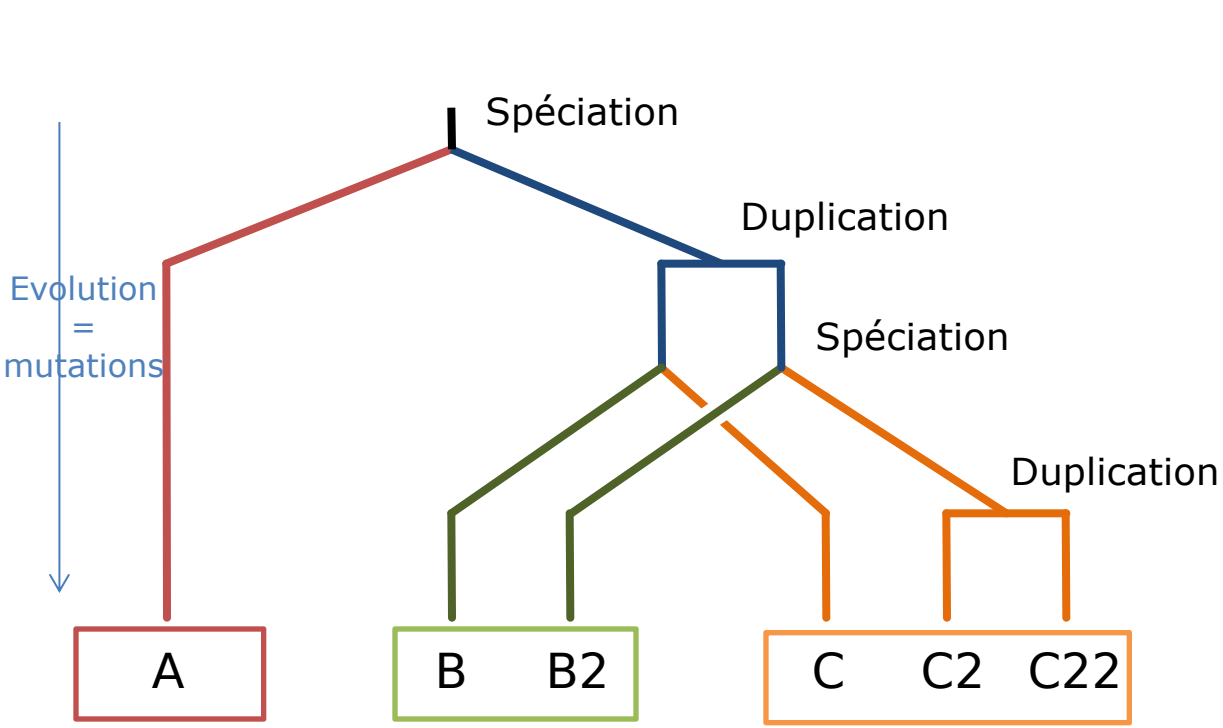
Chronologie des évènements

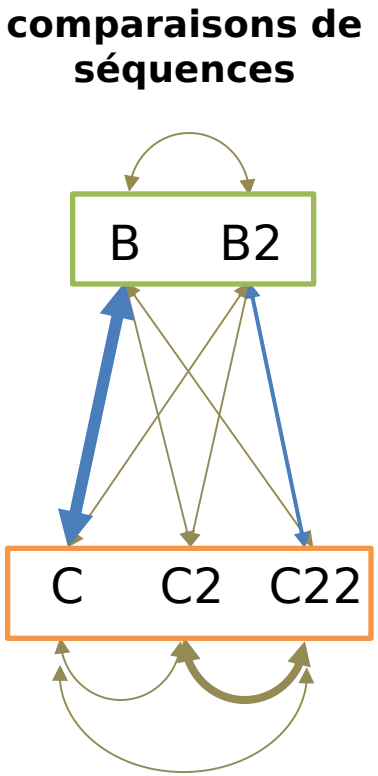




- L'orthologie 1:1 est essentielle pour la prédiction de la fonction des gènes (par transfert d'annotations)
- Problèmes :
  - la chronologie est nécessaire pour différencier paralogues, orthologues et orthologues 1:1
  - la chronologie n'est pas connue
- Deux principales approches pour identifier les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

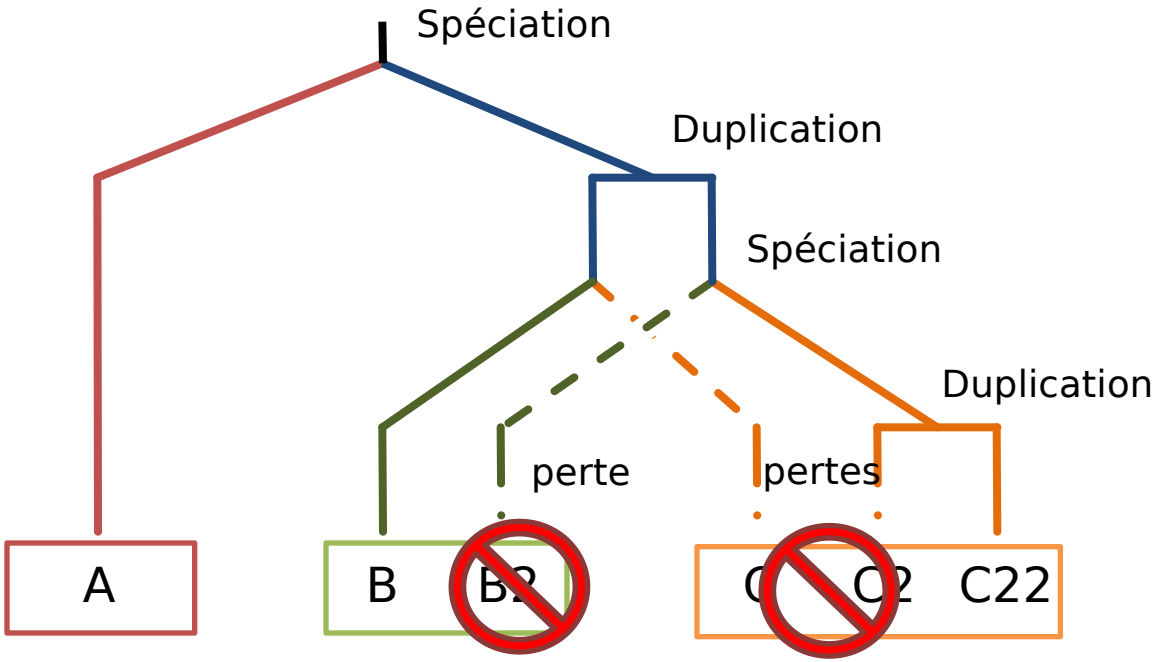
Prédiction des orthologues 1:1

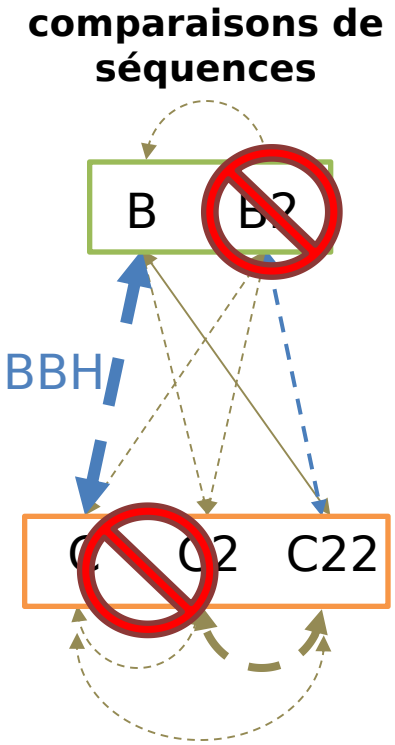




- comparaisons de toutes les séquences 2 à 2
- B↔C prédit sim(B,C) > max(sim(B,B2), sim(C, C2), sim(C,C22))
- B2↔C22 non prédit sim(B2,C22) < sim(C2,C22)
- Remarque : calcul avec 7,223,104 séquences de 2 355 génomes =>52,000.10^9 comparaisons

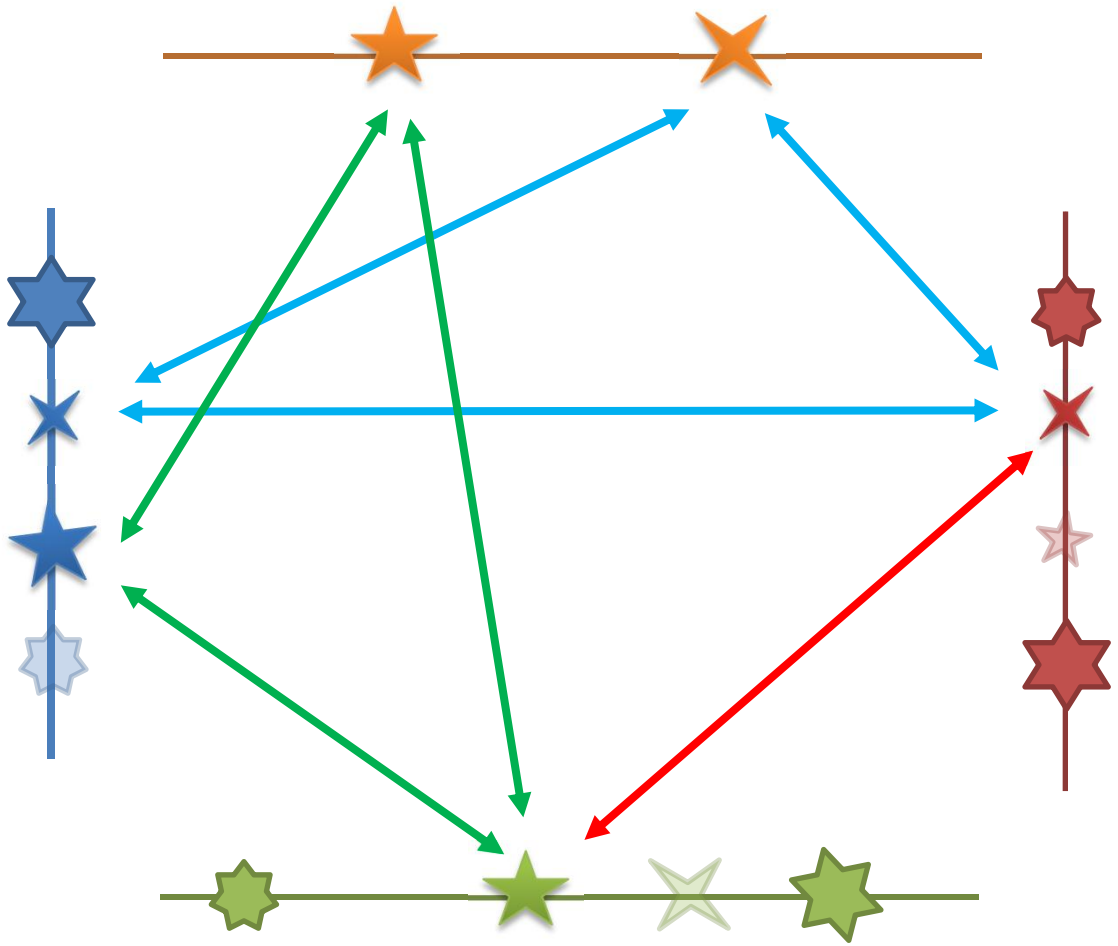
Autre évènement évolutif : perte de gènes





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

Transitivité de l'orthologie 1:1



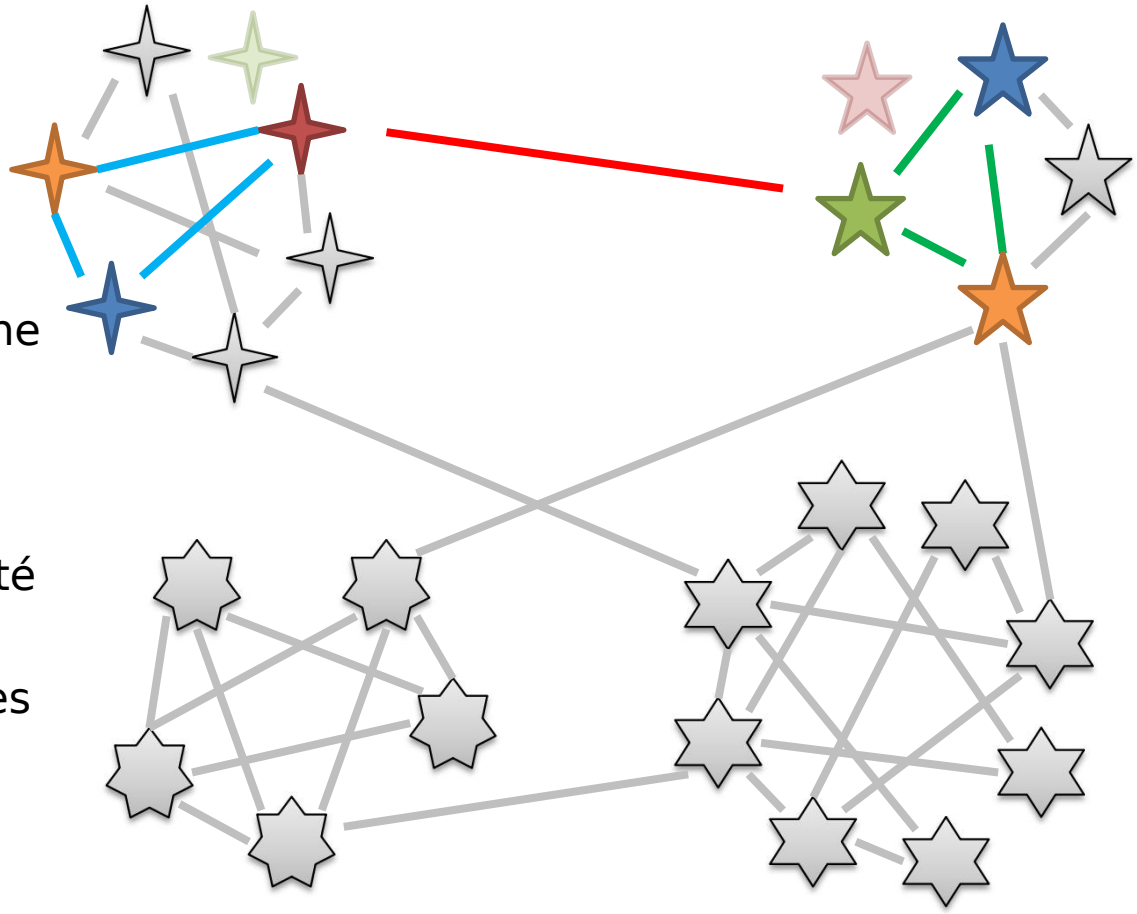
Elagage des erreurs de prédiction

Méthodes de traitement de graphes

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

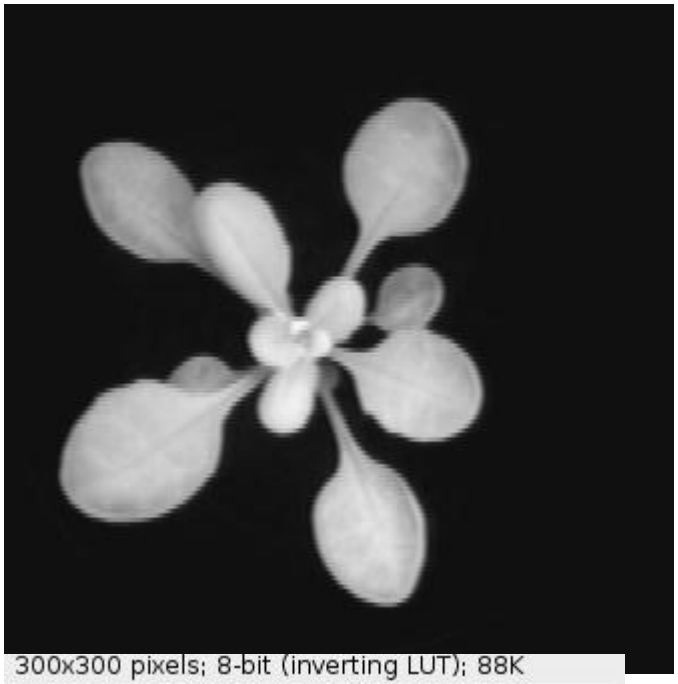
Partitionnement de graphe
=
clustering
=
détection de communauté

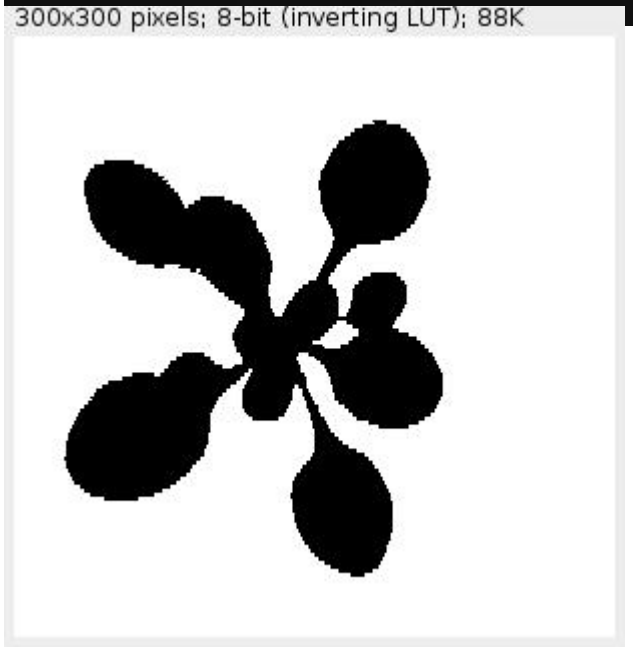
principe : les sommets très connectés entre eux sont isolés en cluster

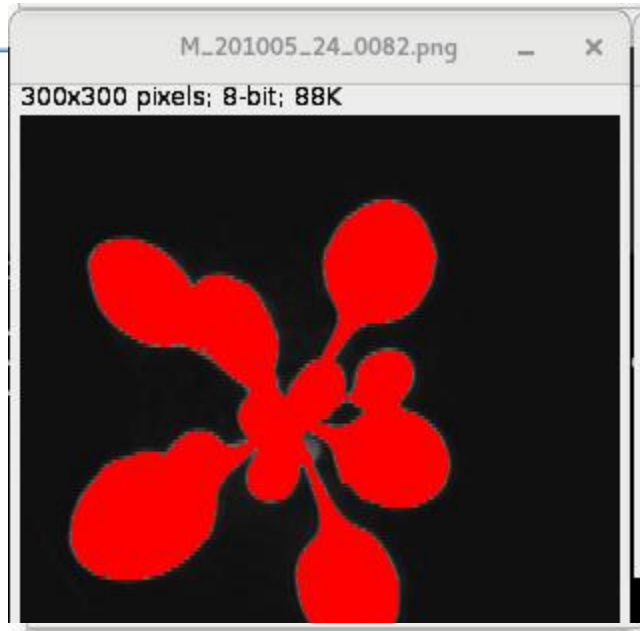


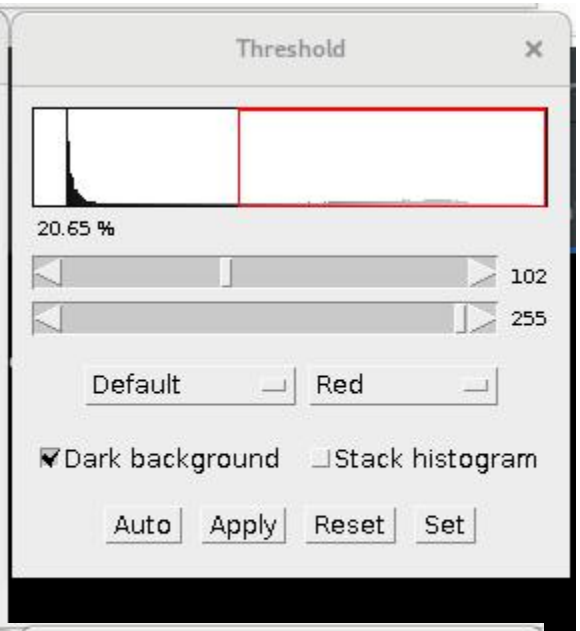
Concepts et méthodes abordés dans cet UE

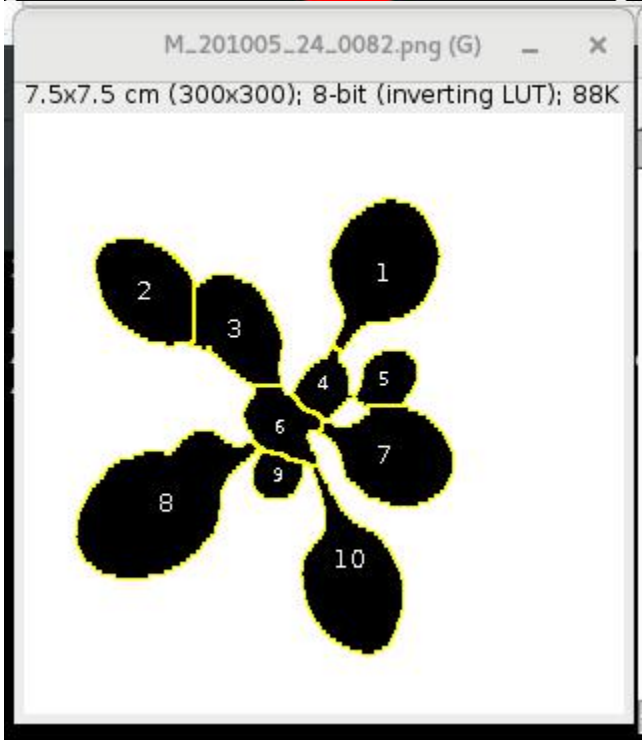
Traitements d'image

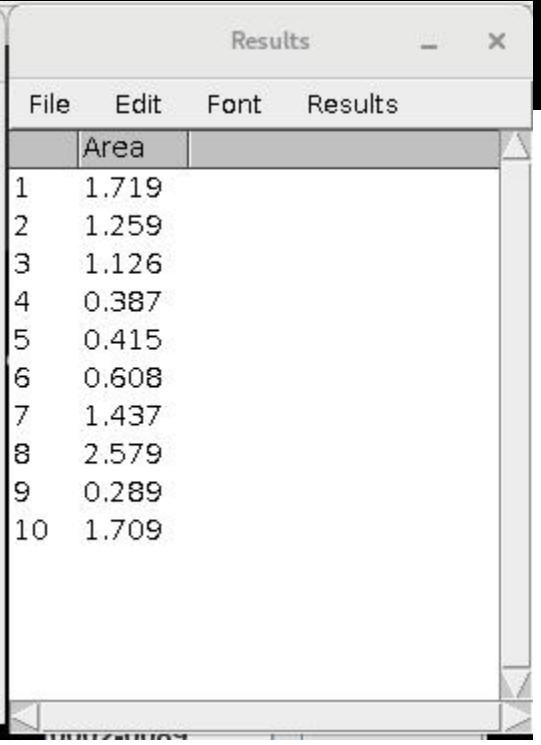


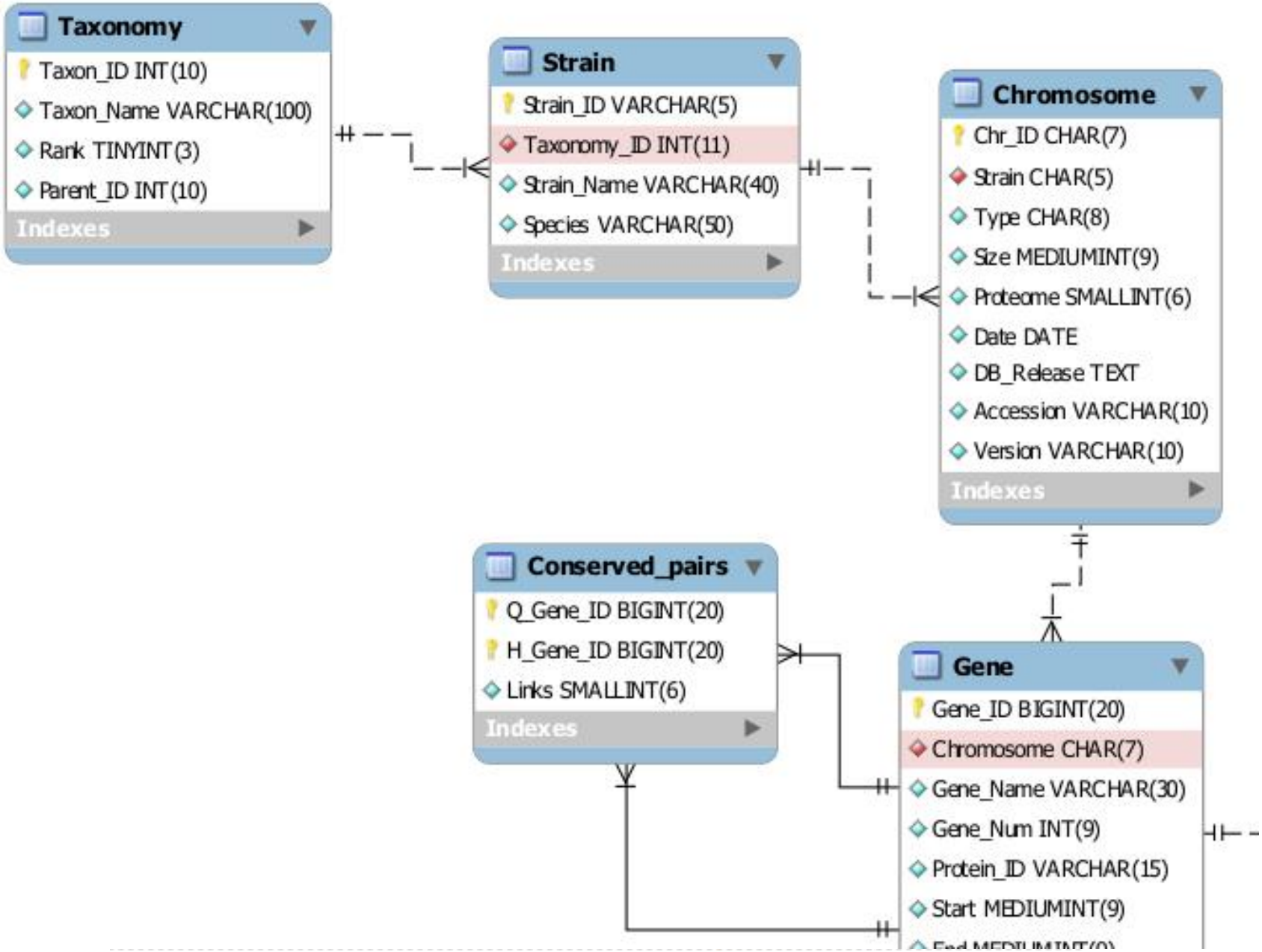




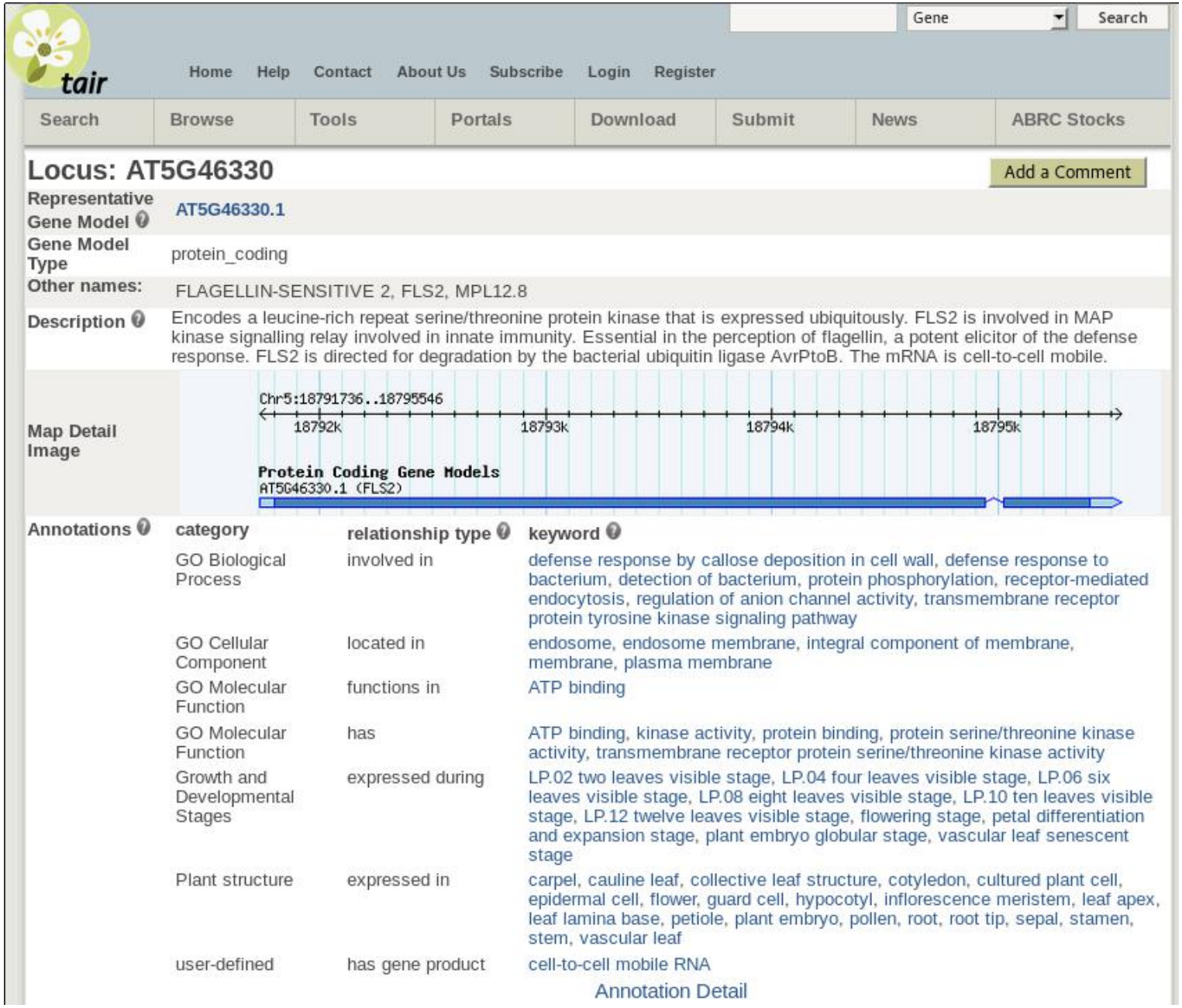




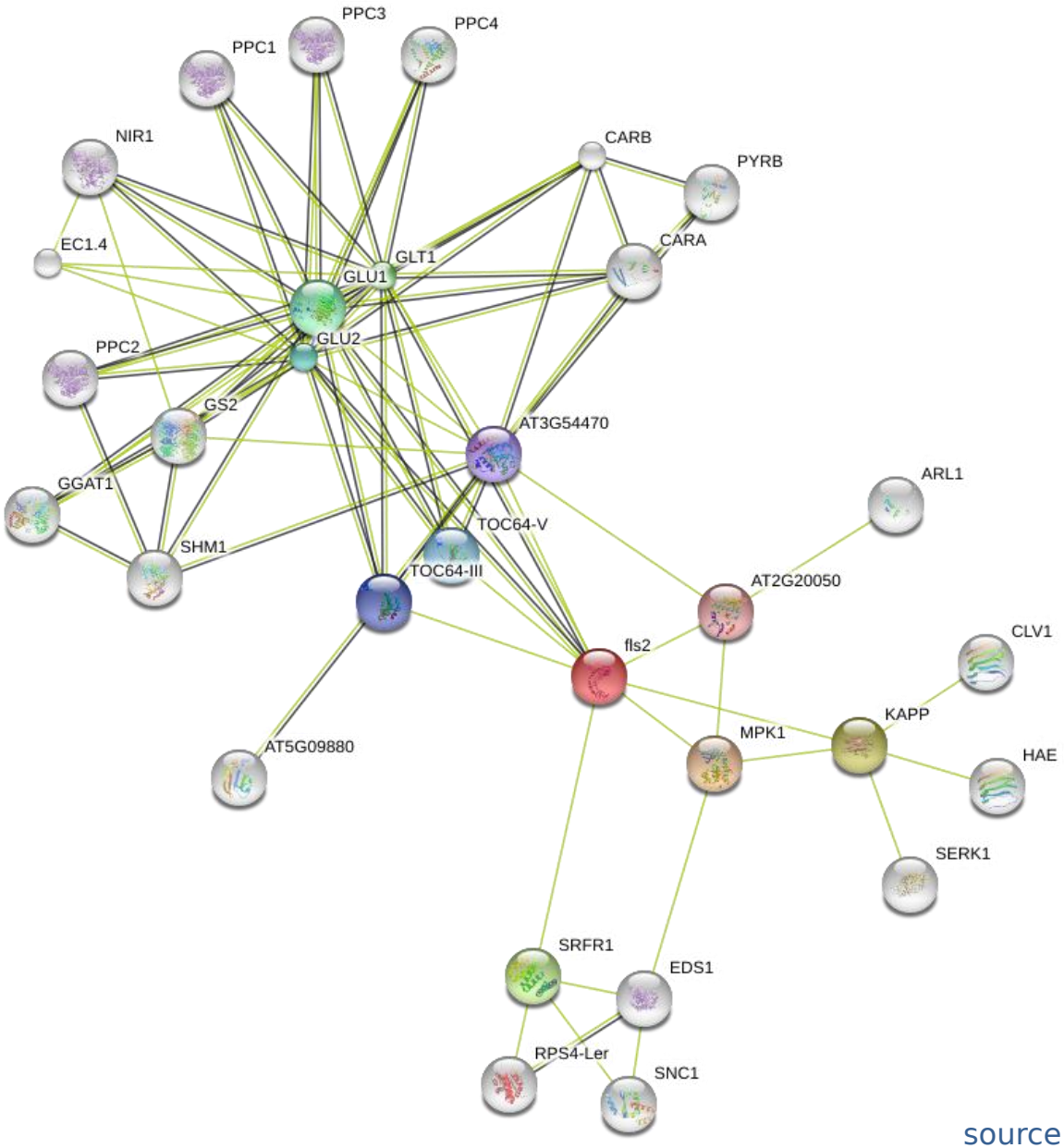



Taxonomy
Taxon_ID INT(10)
Taxon_Name VARCHAR(100)
Rank TINYINT(3)
Parent_ID INT(10)
Indexes
Strain
Strain_ID VARCHAR(5)
Taxonomy_ID INT(11)
Strain_Name VARCHAR(40)
Species VARCHAR(50)
Indexes
Chromosome
Chr_ID CHAR(7)
Strain CHAR(5)
Type CHAR(8)
Size MEDIUMINT(9)
Proteome SMALLINT(6)
Date DATE
DB_Release TEXT
Accession VARCHAR(10)
Version VARCHAR(10)
Indexes
Conserved_pairs
Q_Gene_ID BIGINT(20)
H_Gene_ID BIGINT(20)
Links SMALLINT(6)
Indexes
Gene
Gene_ID BIGINT(20)
Chromosome CHAR(7)
Gene_Name VARCHAR(30)
Gene_Num INT(9)
Protein_ID VARCHAR(15)
Start MEDIUMINT(9)

Banques et sites Web

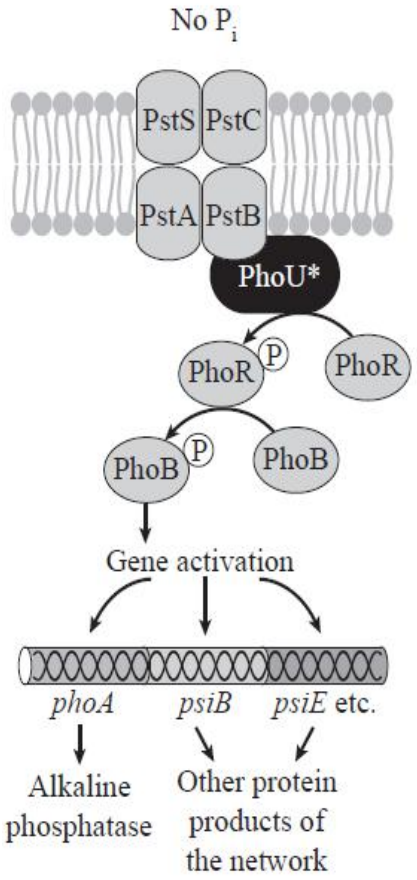


Banques et sites Web

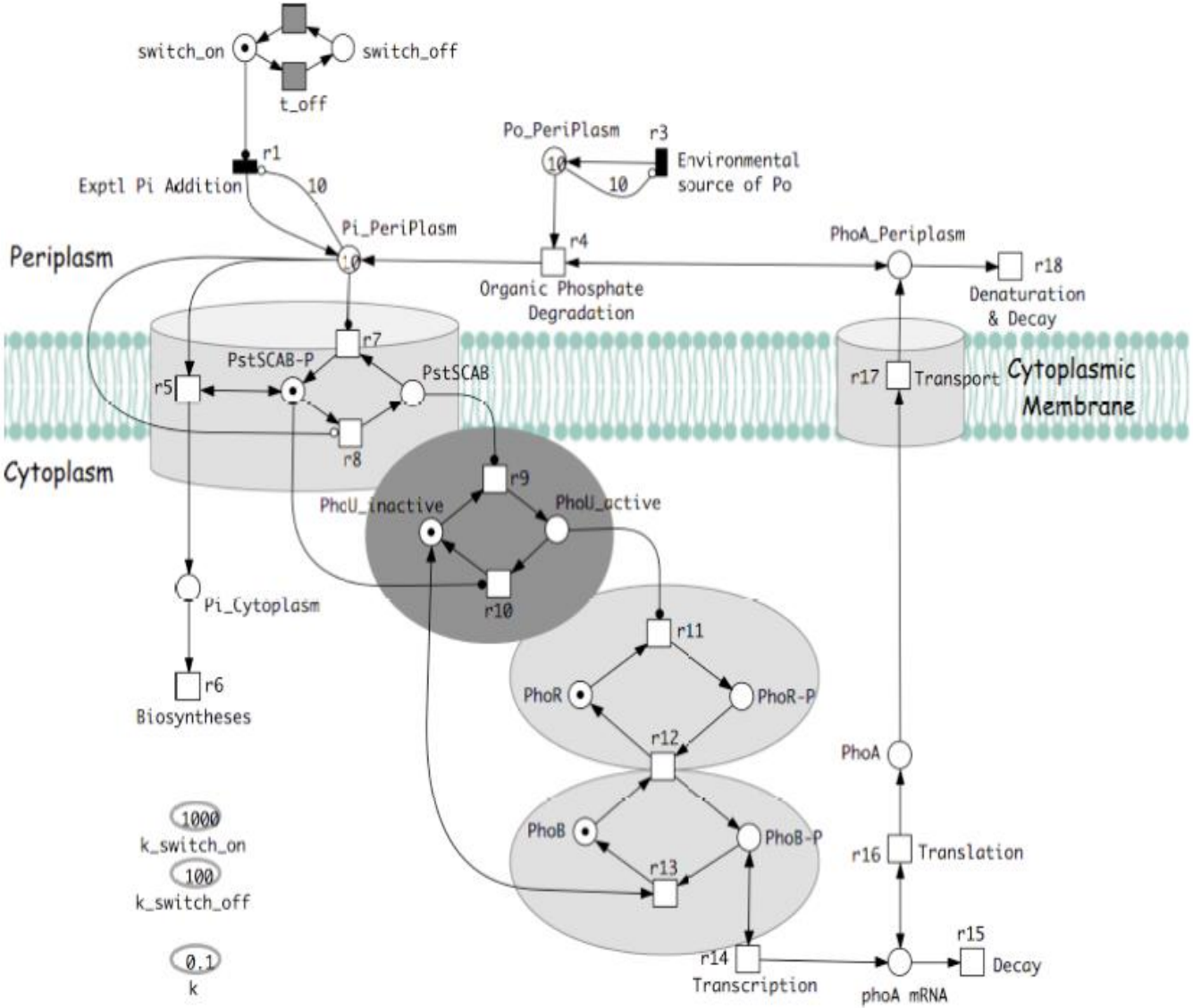


source http://string-db.org

Modélisation de systèmes biologiques



Neidhardt *et al.* 1990



Durzinsky *et al.*, 2011