

# Introduction

# Introduction

La bioinformatique : Traitement des informations biologiques par des méthodes informatiques et/ou mathématiques.

Interdisciplinaire par nature, la bioinformatique est fondée sur les acquis de la biologie, des mathématiques et de l'informatique. En cela, elle constitue une branche nouvelle de la biologie : c'est l'approche *in silico*, qui vient compléter les approches classiques *in situ* (dans le milieu naturel), *in vivo* (dans l'organisme vivant) et *in vitro* (en éprouvette) de la biologie traditionnelle.

# Introduction

Plusieurs domaines d'application (liste non exhaustive) :

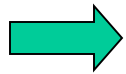
- la génétique des populations
- l'environnement (données écologiques)
- la biologie structurale
- la biologie moléculaire et la génétique
- l'évolution

Le cours portera sur les approches en analyse de séquences, donc les deux derniers domaines d'application.

# Introduction

Développement de méthodes et de logiciels permettant :

- **de gérer et d'organiser les informations génétiques et génomiques**
- **d'analyser ces informations (par approches comparatives ou exploratrices)**



**prédire et produire des connaissances nouvelles dans le domaine  
ainsi qu'élaborer de nouveaux concepts**

approche théorique qui permet :

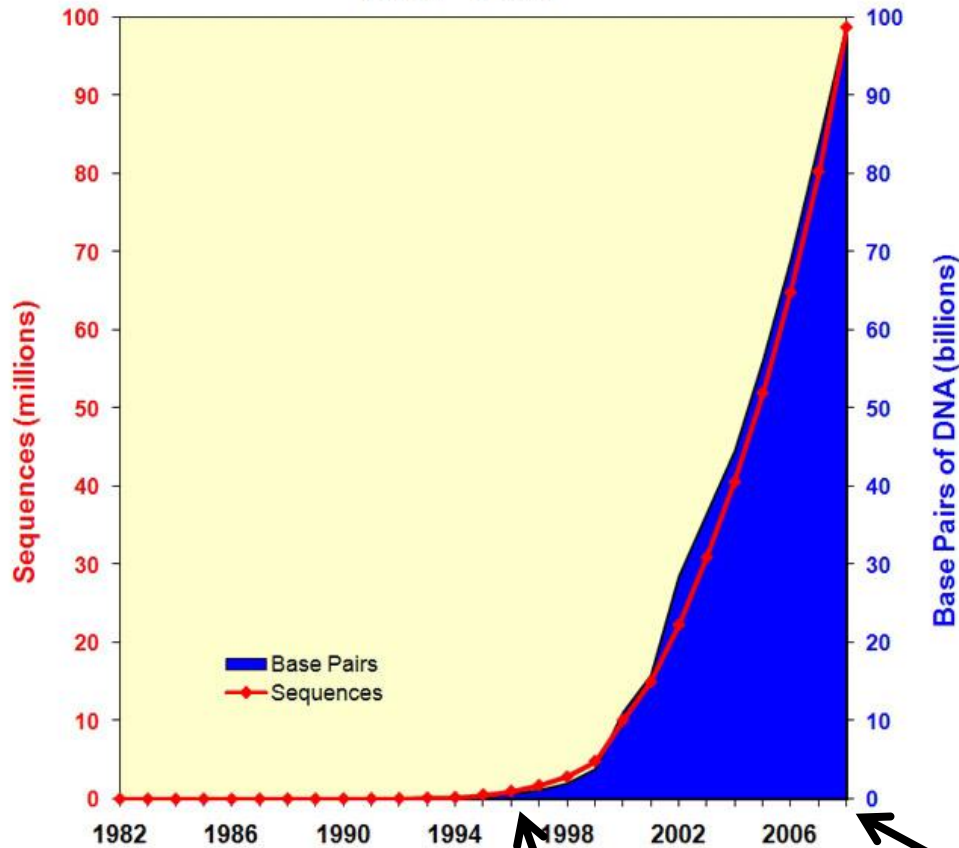
- d'effectuer la synthèse des données disponibles (à l'aide de modèles et de théories)
- d'énoncer des hypothèses généralisatrices (ex: comment les protéines se replient ou comment les espèces évoluent)
- de formuler des prédictions, à partir d'une approche par modélisation appliquée à des objets formalisés.

# Historique rapide de la bioinformatique

- Années 70 : Premières comparaisons de séquences.
  - Années 80 : Premières méthodes de prédiction.  
Premières méthodes d'alignement.  
Banques de données.  
Méthodes de recherche dans les banques de données (Fasta et Blast).
  - Années 90 : Perfectionnement des méthodes.  
Approches intégrées.
- Fin des années 1990 : premiers génomes complets procaryotes et premier génome complet eucaryote (levure, 1996)
- Années 2000 : Génomique  
Début des approches globales, (transcriptomique et protéomique)  
Prédiction de la structure 3D des protéines
  - Aujourd'hui : Génomique (voir les données de Genome Online database (GOLD)).  
Post-génomique : approches omiques (protéome, transcriptome, interactome, métabolome, ...)  
Début de la biologie des systèmes : réseau de régulation, réseau d'interaction, modélisation de la cellule.
  - Demain : Biologie des systèmes et biologie synthétique

# Séquences disponibles : quelques chiffres

**Growth of GenBank**  
(1982 - 2008)



**2021:**

**234 557 297 seq**

**1 053 275 115 030 bp**

**2019:**

**215 333 020 seq**

**388 417 258 009 bp**

**2016:**

**198 565 475 seq**

**224 973 060 433 bp**

**1982: 606 seq**  
**680 338 bp**

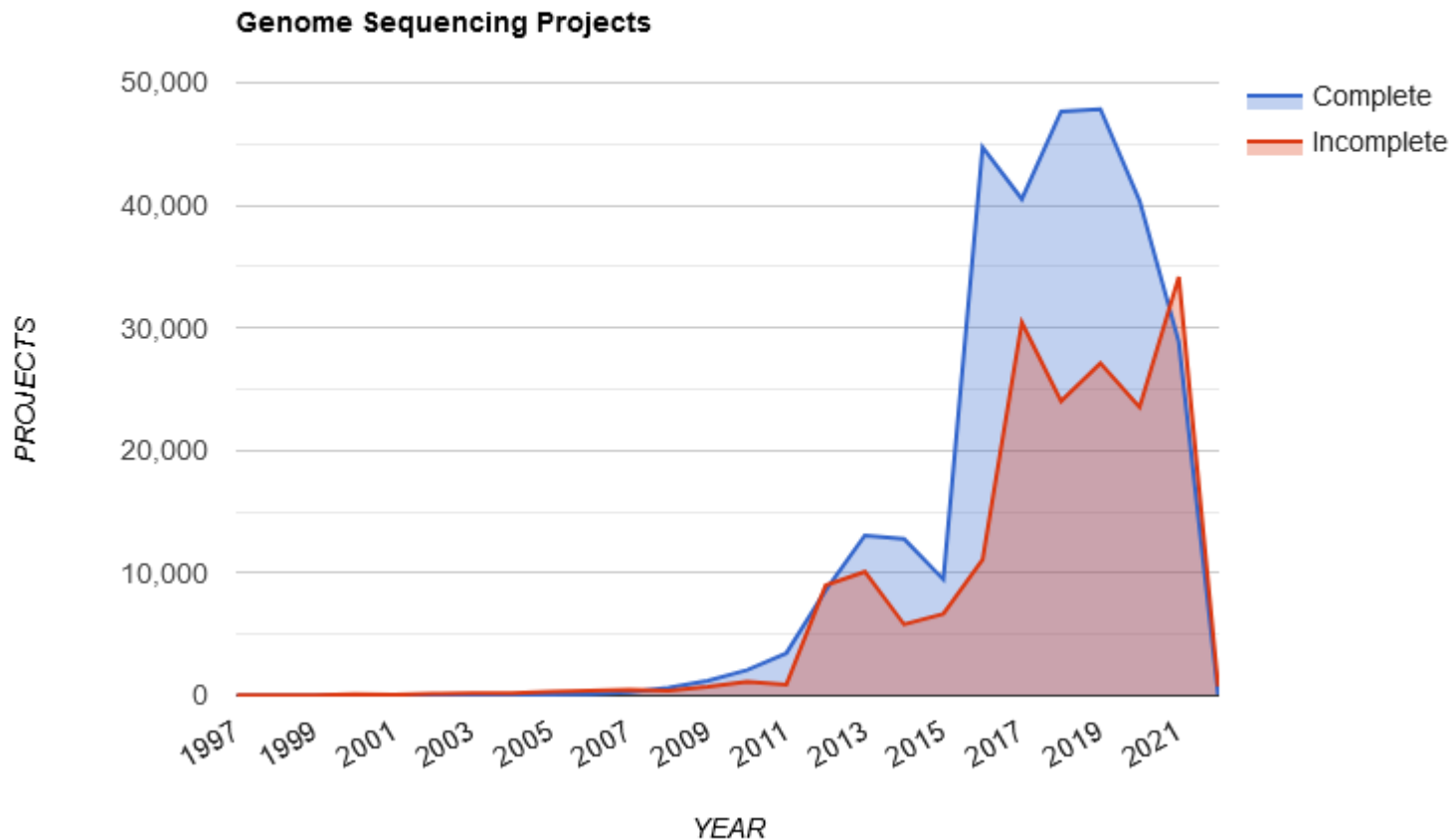
**1996: 1 021 211 seq**  
**651 972 984 bp**

**2008: 98 868 465 seq**  
**99 116 431 942 bp**

# Génomes : quelques chiffres

## GOLD : Genomes Online Database

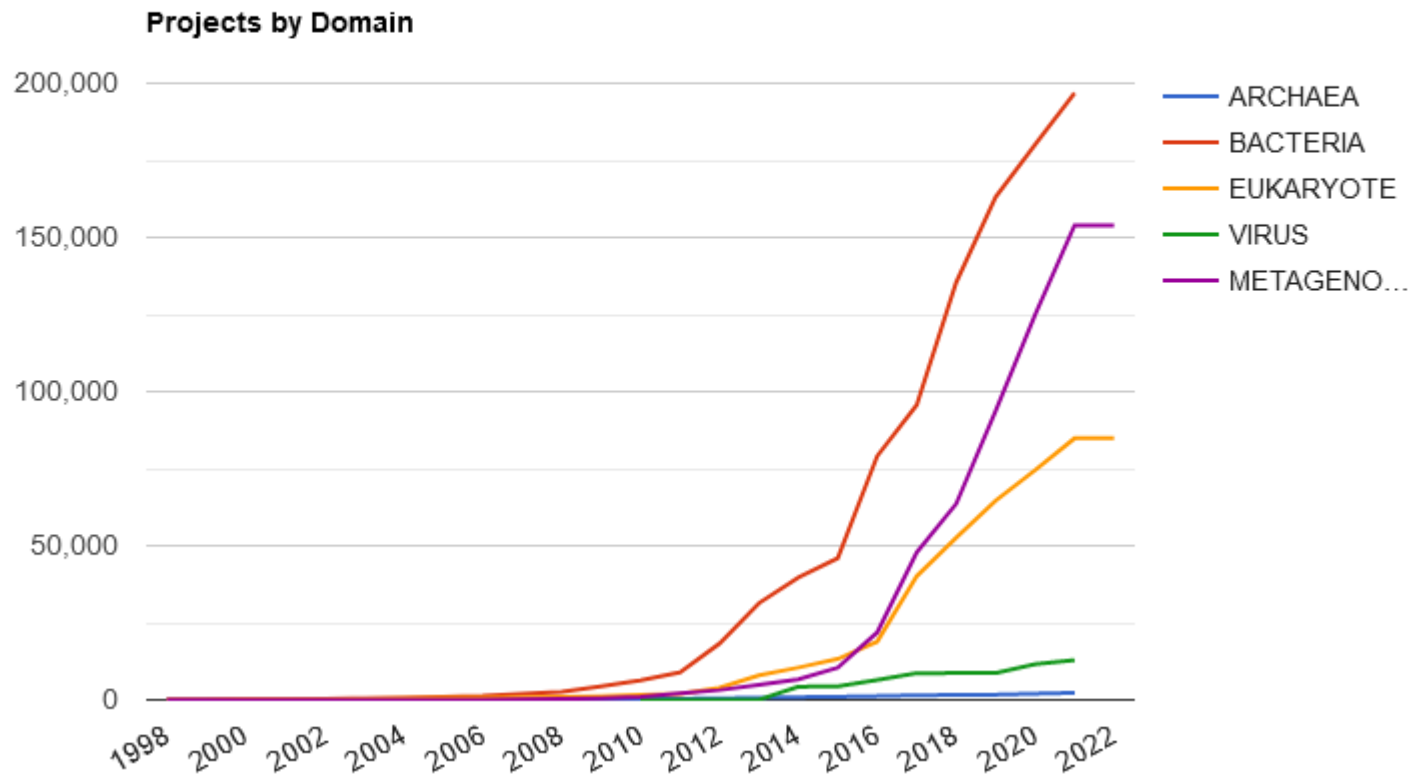
Genome Totals in GOLD (by year and status)



# Génomes : quelques chiffres

## GOLD : Genomes Online Database

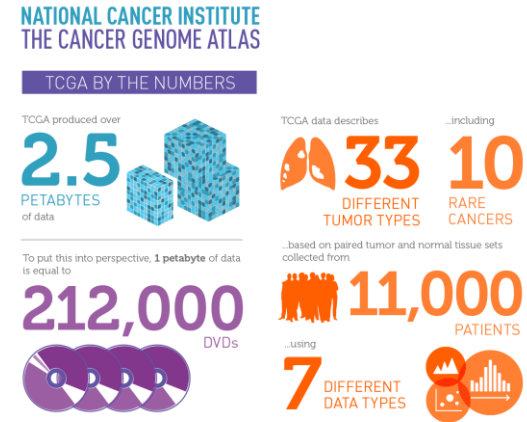
Project Totals in GOLD (by year and Domain Group)





# Exemple de projets ambitieux

**Cancer Genome Atlas:** un programme historique de génomique du cancer, a permis de caractériser moléculairement plus de 20 000 échantillons primaires de cancer couvrant 33 types de cancers.

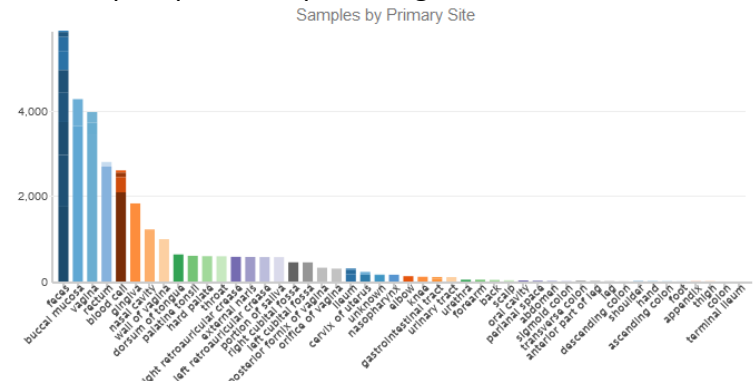


Source : <https://www.cancer.gov/>

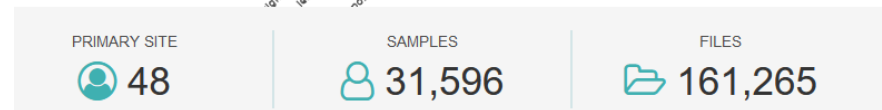
**Encyclopedia of DNA Elements (ENCODE):** Le consortium ENCODE est une collaboration internationale de groupes de recherche dont l'objectif est de construire une liste exhaustive des éléments fonctionnels du génome humain (<https://www.encodeproject.org/>)

Source : <https://portal.hmpdacc.org//>

**Human Microbiome Project:** un des projets visant à caractériser le microbiome humain de différents endroits du corps (ci contre 31 596 échantillons prélevés à partir de 48 parties différentes du corps)

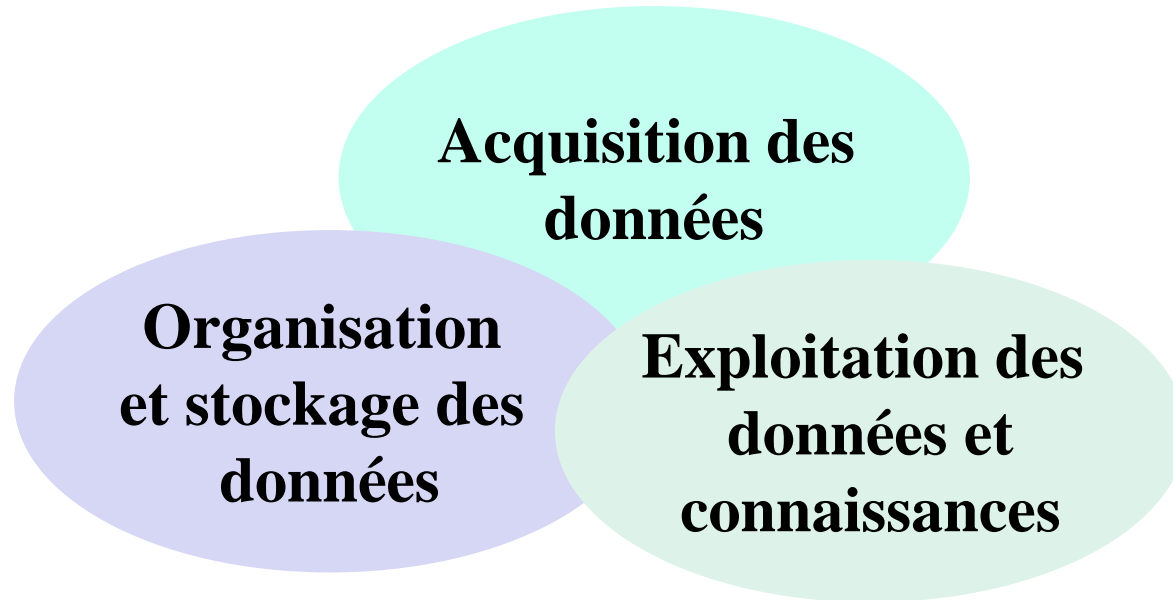


**Earth Microbiome Project:** Caractérisation des communautés microbiennes sur la terre (<https://earthmicrobiome.org/>)

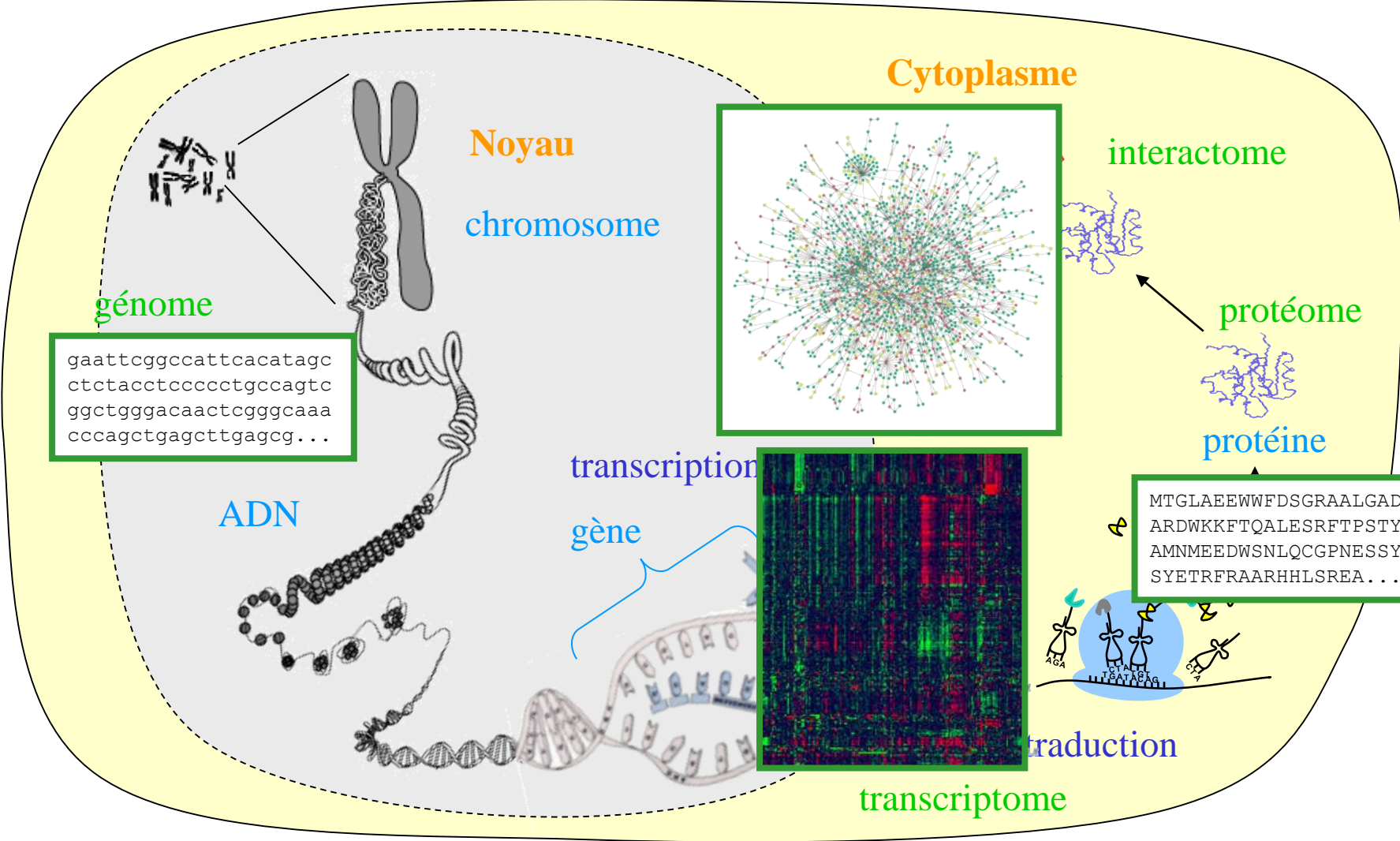


**Genome 10K:** Projet de séquençage de 10,000 espèces (<https://genome10k.soe.ucsc.edu/>)

# Trois grands domaines où intervient la bioinformatique

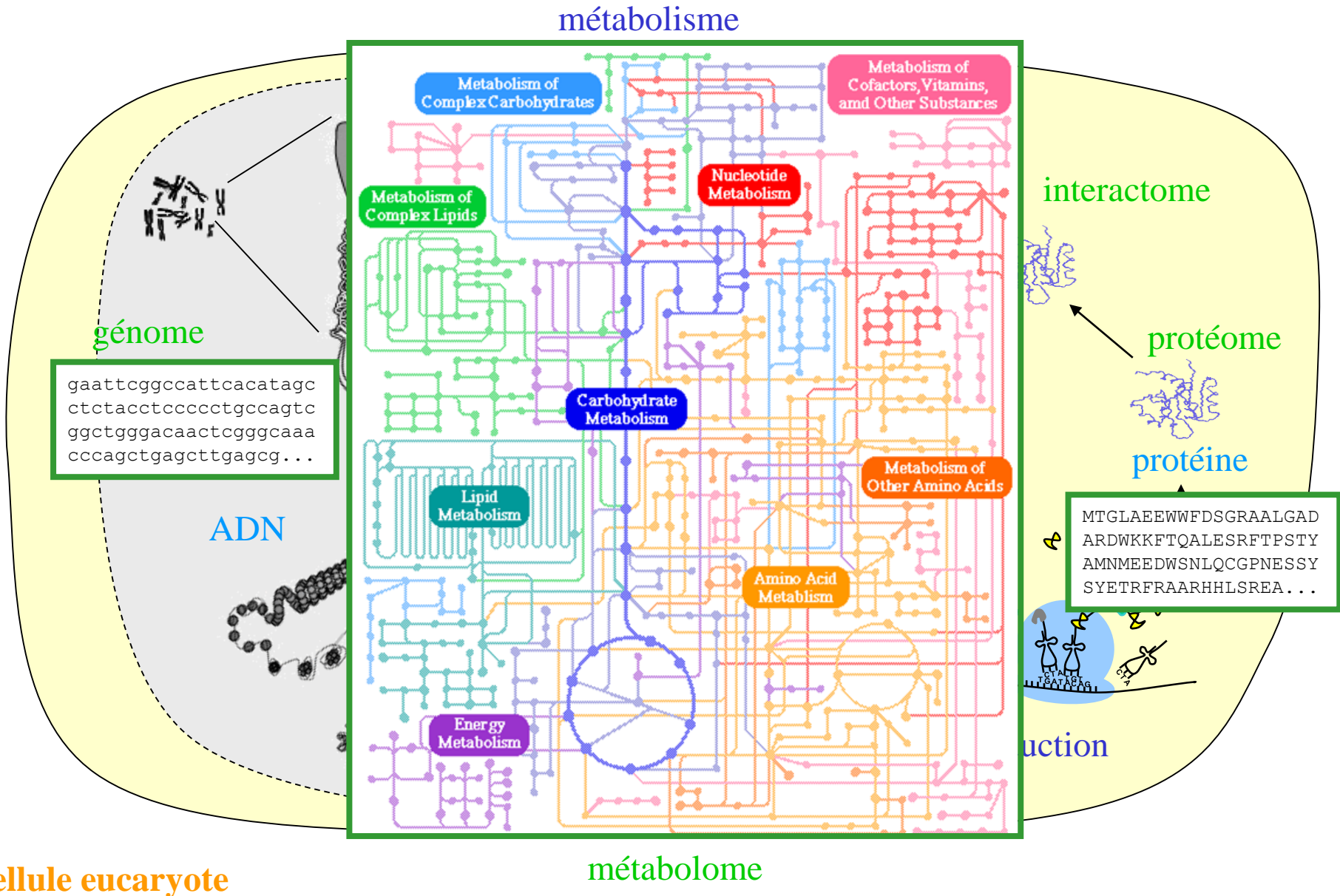


# (Quelques) données et connaissances disponibles



**Cellule eucaryote**

# (Quelques) données et connaissances disponibles



métabolisme

génom

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

ADN

Metabolism of Cofactors, Vitamins, and Other Substances

Metabolism of Complex Lipids

Nucleotide Metabolism

Carbohydrate Metabolism

Lipid Metabolism

Metabolism of Other Amino Acids

Amino Acid Metabolism

Energy Metabolism

interactome

protéome

protéine

```
MTGLAEWWFDSGRAALGAD
ARDWKKFTQALESRFTPSTY
AMNMEEDWSNLQCGPNESSY
SYETRFRAARHLSREA...
```

action

métabolome

Cellule eucaryote

# Métagénomique

La **métagénomique** est une méthode d'étude du contenu génétique d'échantillons obtenus à partir de prélèvements réalisés dans des environnements naturels complexes (ex : intestin, océan, sols, air, etc.) par opposition à des échantillons cultivés en laboratoire).

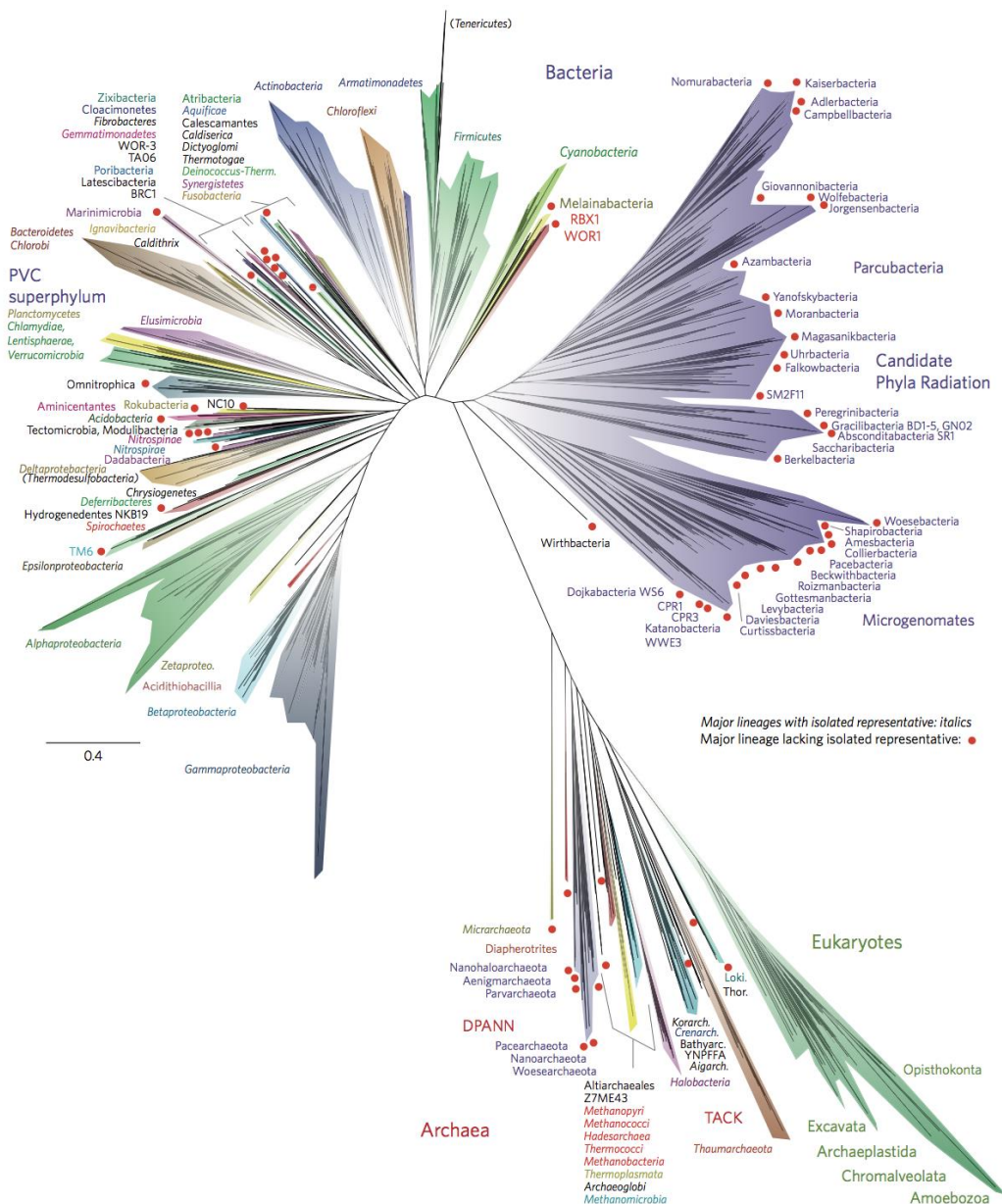
Cette approche, via le séquençage direct de l'ADN présent dans l'échantillon, permet une description génomique du contenu de l'échantillon mais offre aussi un aperçu du potentiel fonctionnel d'un environnement.

Préfixe « méta » → « *ce qui vient après* » : la métagénomique vient après la génomique.



Exemple : études des communautés microbiennes présentes dans ce cours d'eau recevant le drainage acide de mines de charbon en surface.

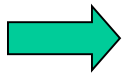
# Métagénomique : nouvel arbre du vivant



Les phyla marqués par des points rouges ont été identifiés par métagénomique et ne possède pas de représentant qui ont été isolés.

# Transcriptome

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



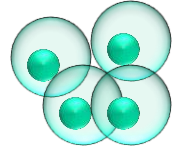
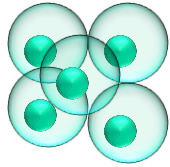
Accès au niveau d'expression de milliers de gènes simultanément  
(potentiellement l'ensemble des gènes d'un organisme)  
= *instantané* de l'état d'une cellule ou d'une population de cellules

Données d'expression des gènes obtenues par :

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

# Acquisition des données

Échantillon test      Échantillon référence



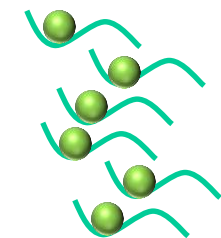
extraction

ARNm

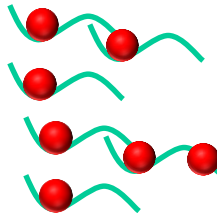
réverse transcription et amplification

ADNc

marquage



+



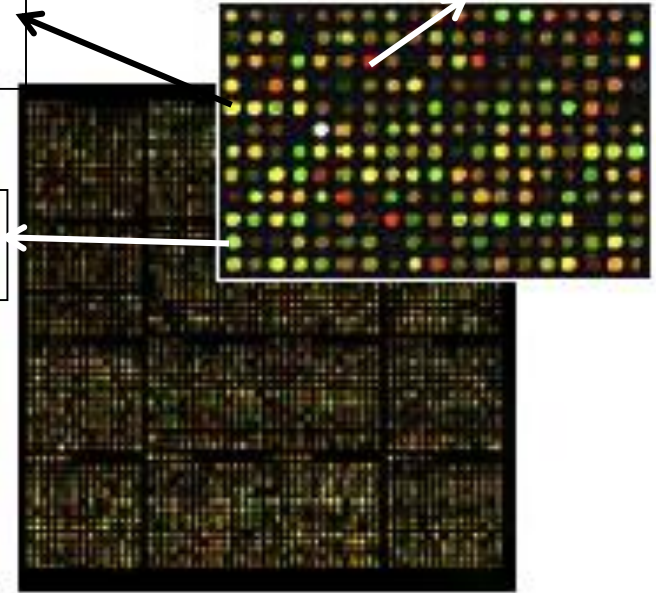
hybridation

puis lavage

Jaune : signal similaire des deux échantillons

Vert: signal fort échantillon test

Rouge : signal fort échantillon de référence

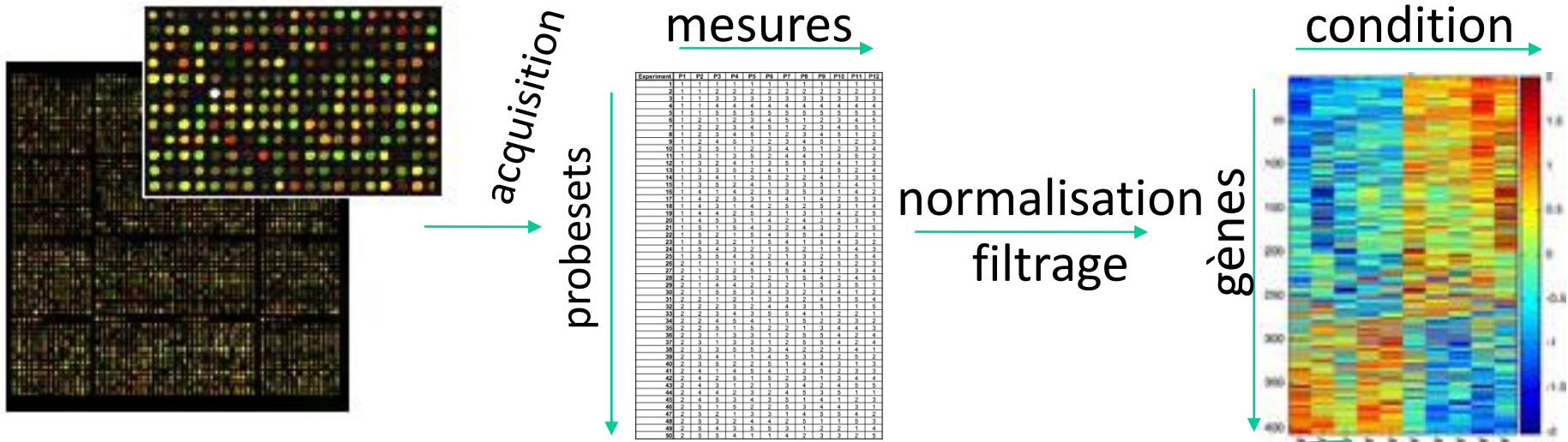


↑ scan





# Analyse et interprétation des données



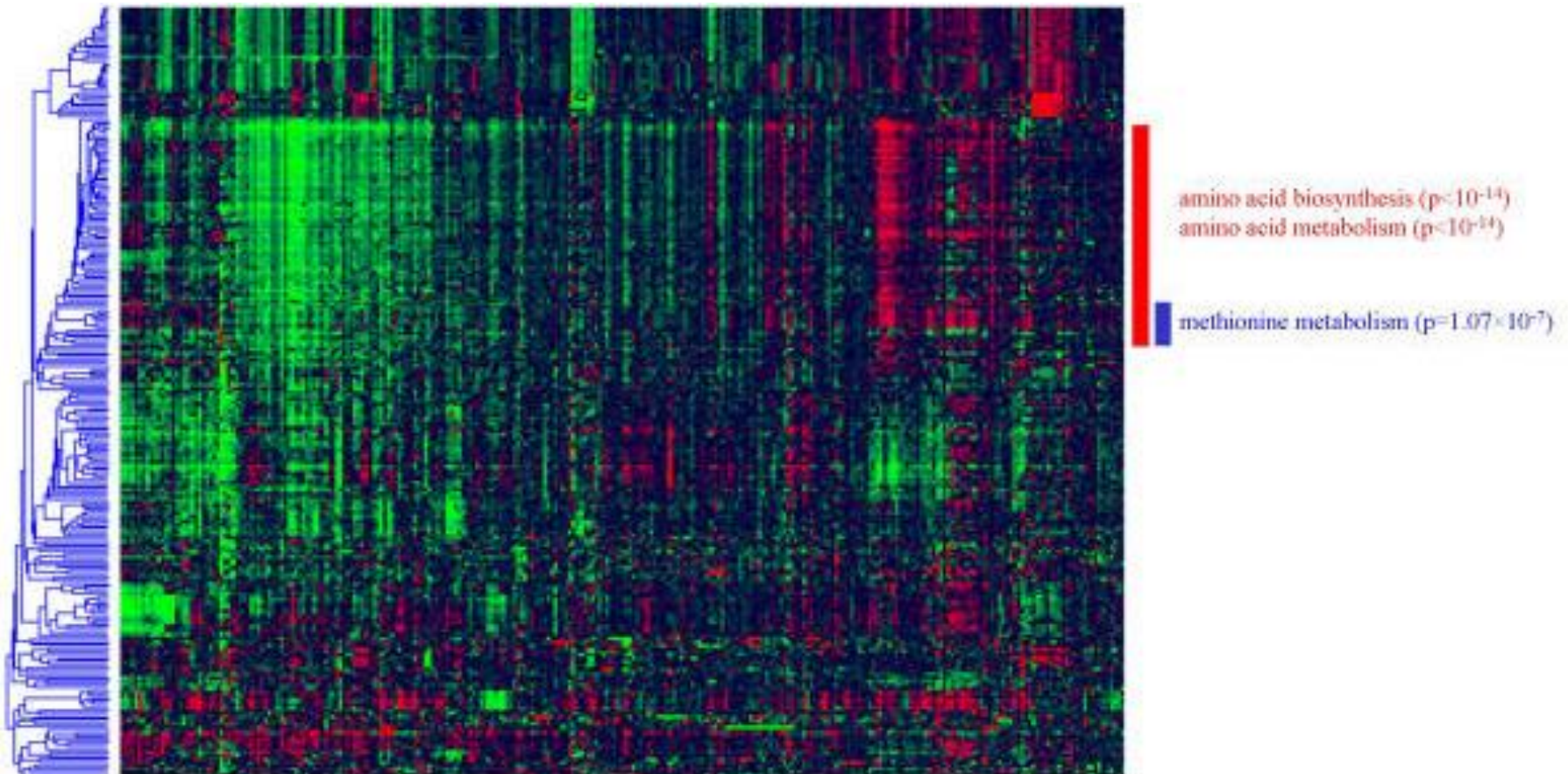
Identification des **gènes**  
**différentiellement**  
**exprimés**

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

**Caractérisation d'un**  
**ensemble de gènes**

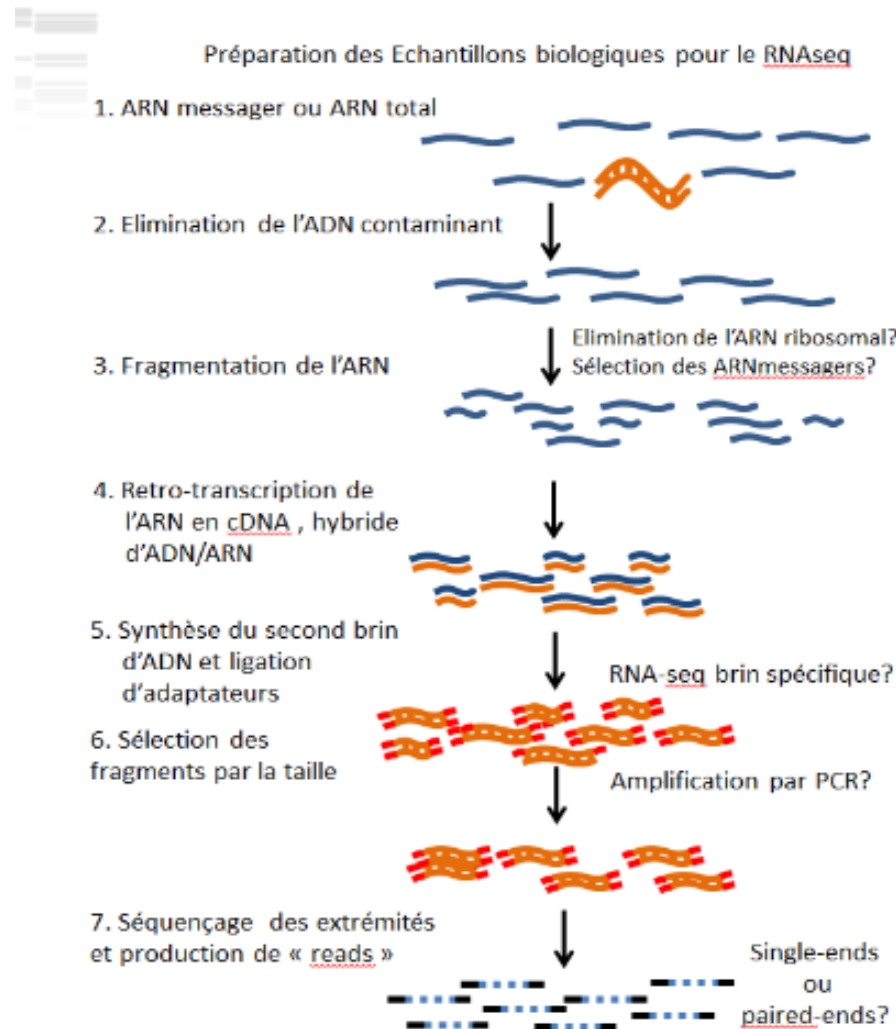
# Gènes co-exprimés

- Motivation : les gènes ayant des profils d'expression similaires sont potentiellement co-régulés et participent à un même processus biologique
- But : regrouper les gènes impliqués dans un même processus biologique



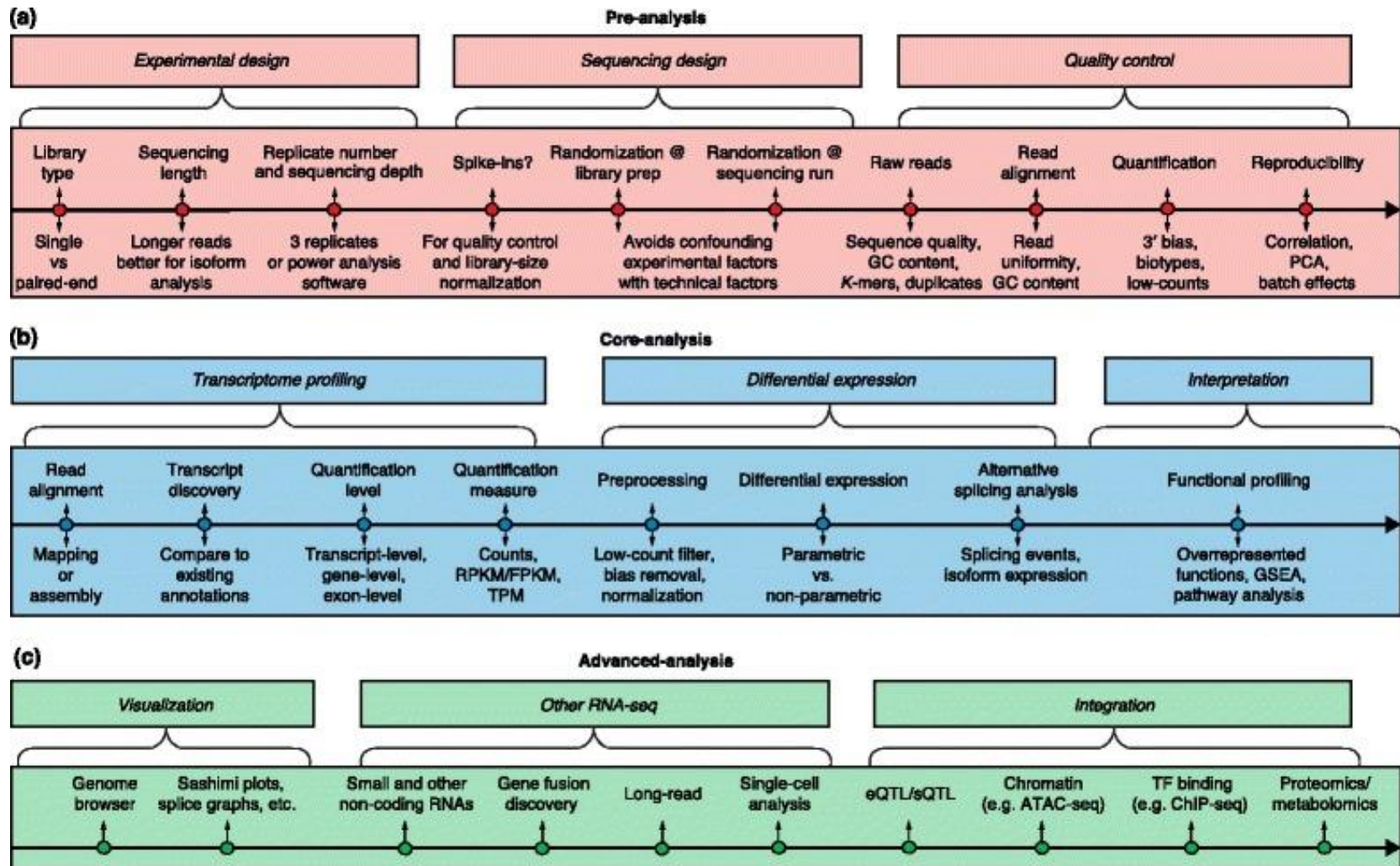
# Etude du transcriptome : RNA-seq

La technologie RNA-seq utilise le séquençage à haut débit (NGS, next-generation sequencing)



# Etude du transcriptome : RNA-seq

Les différentes étapes des analyses bio-informatiques des données RNA-seq

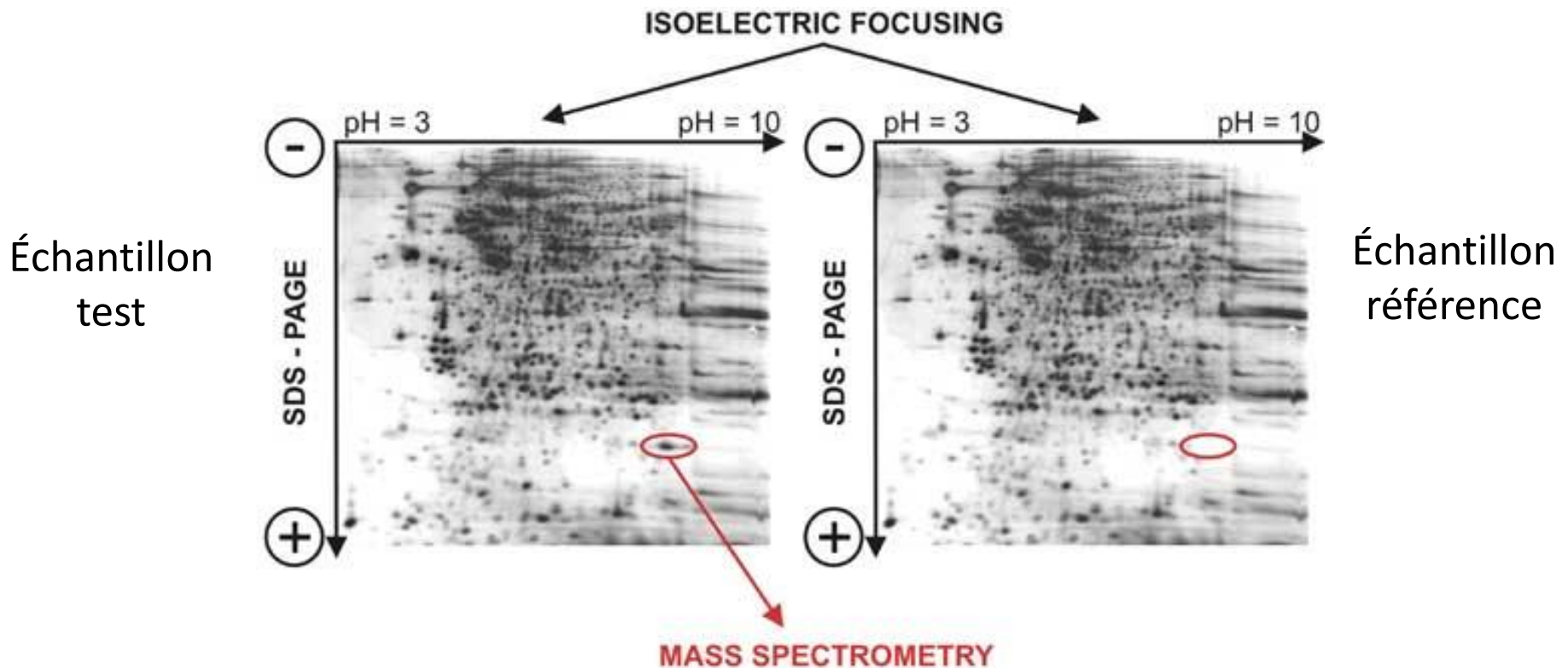


# Protéomique

**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

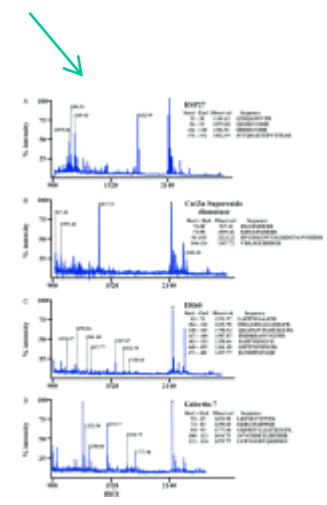
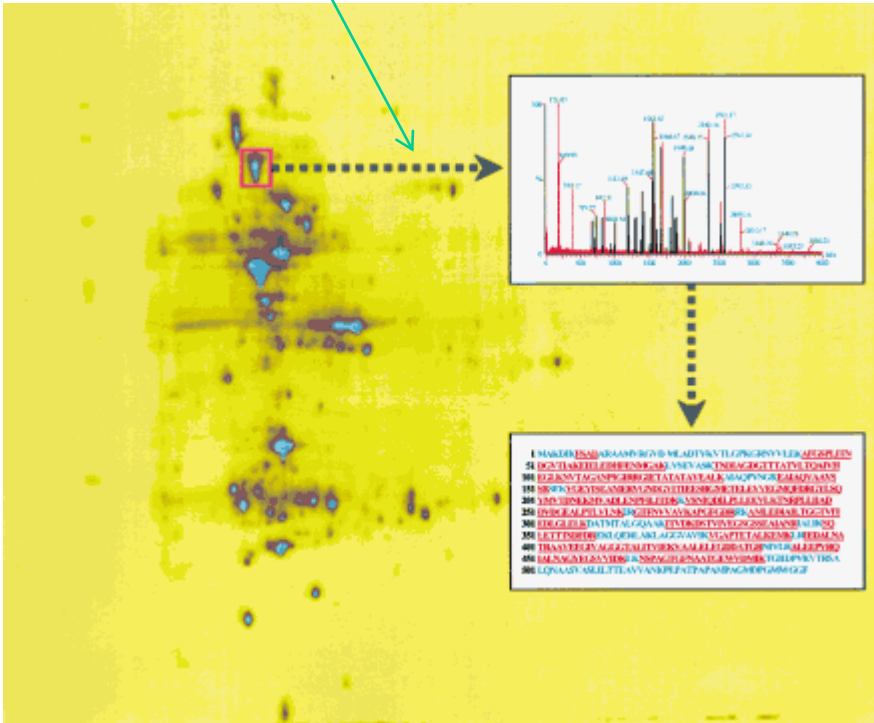


# Identification des protéines

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

Digestion *in silico* du protéome

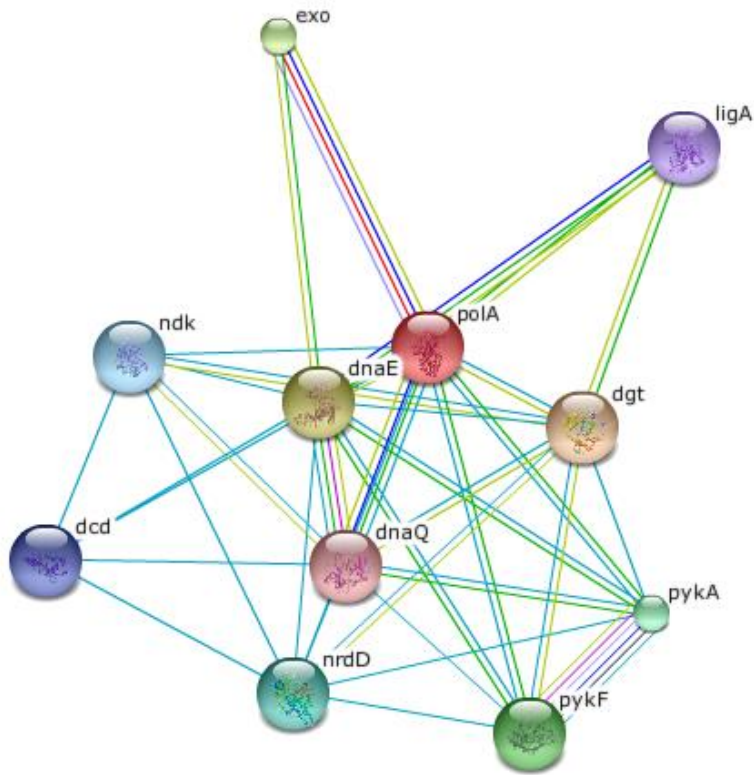
Recherche des protéines correspondant au profil observé



# Réseaux de gènes, de protéines

Réseaux :

- d'interactions protéine - protéine



Exemple de réseau extrait de la base de données STRING

## Edges:

Edges represent protein-protein associations

*associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding each other.*

Known Interactions

- from curated databases
- experimentally determined

Predicted Interactions

- gene neighborhood
- gene fusions
- gene co-occurrence

Others

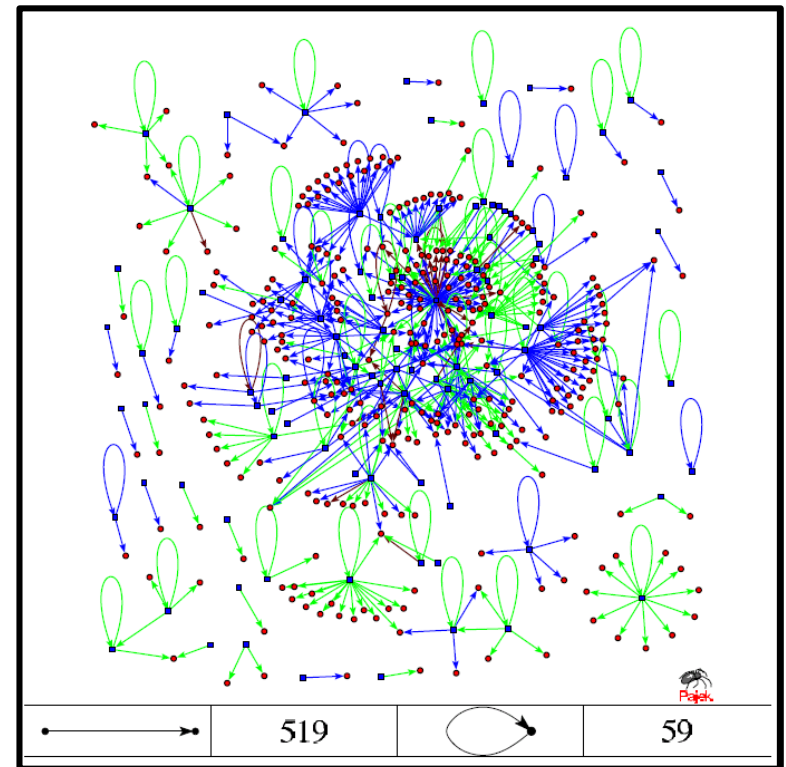
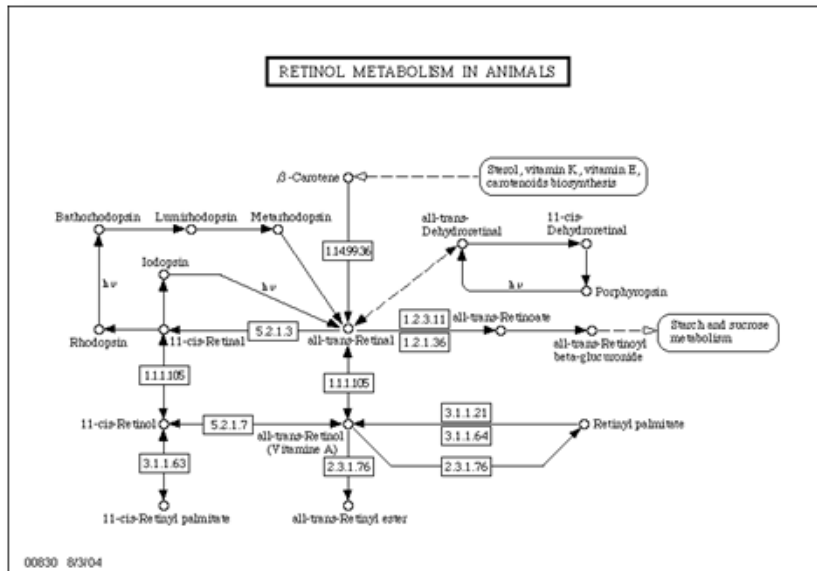
- textmining
- co-expression
- protein homology

# Réseaux de gènes, de protéines

Réseaux :

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

Exemple des réseaux de transcription chez *E. coli*





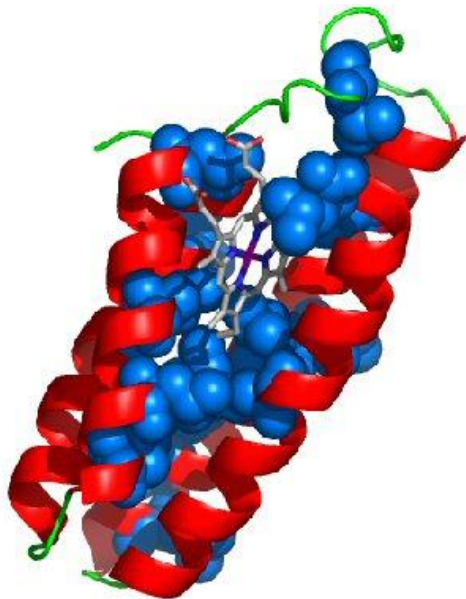
# Biologie structurale

## Séquence protéique

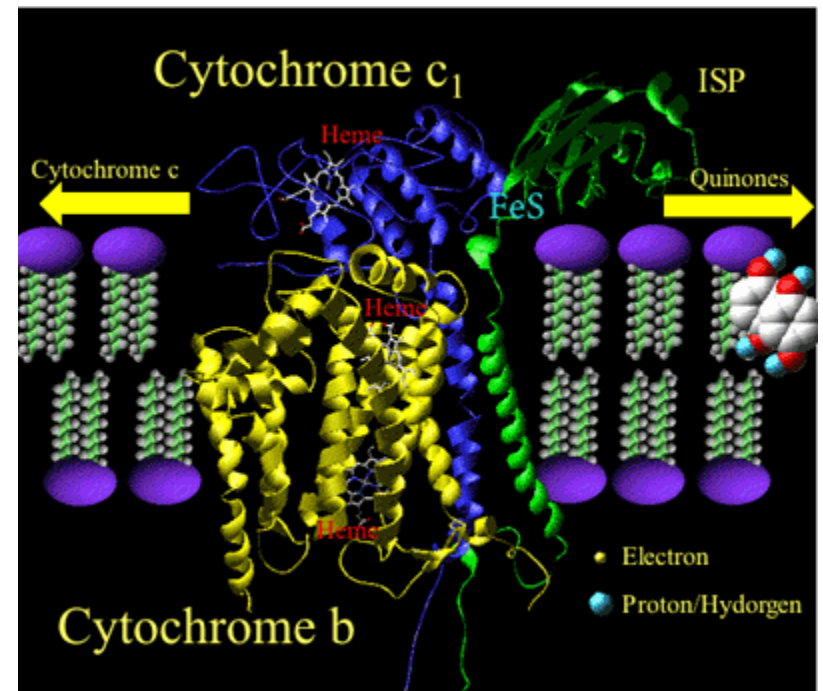
>gi|5524211|gb|AAD44166.1| cytochrome b  
LCLYTHIGRNIYGSYLYSETWNTGIMLLITMATAFMGYVLPWQMSFWGATVITNLFSAIPYIGTNLV  
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG  
LLILLLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVL  
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSILAFPLIAGX  
IENY



Prédiction ou résolution  
de la structure tridimensionnelle



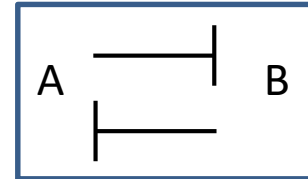
Prédiction des interactions  
protéine – protéine ou  
protéine - ligand



# Biologie des systèmes

But : comprendre comment les réseaux d'interactions complexes contrôlent le comportement de la cellule.

Un exemple d'une répression mutuelle :



Ici le taux de synthèse de A (B) va dépendre de la concentration de B (A) qui a une action d'inhibition. Ceci peut être approximé par une fonction de Hill. On considérera que la dégradation n'est pas régulée et dépend d'une constante de dégradation

Ecriture d'équations différentielles ordinaires permettant d'exprimer le taux de synthèse d'un composé donné en fonction de la concentration des autres composés du système.

Forme générale de l'équation :  $\frac{dx}{dt} = \text{synthesis}(x) - \text{degradation}(x)$

$$\frac{d[A]}{dt} = \beta_{\max} \frac{K_d^n}{[B]^n + K_d^n} - \gamma[A]$$

Avec :

$\beta_{\max}$  : taux maximal de synthèse de A (B) donc en absence de répression

$K_d$  : concentration de B (A) nécessaire pour atteindre la moitié de la répression maximale de A (B) = coefficient de répression.

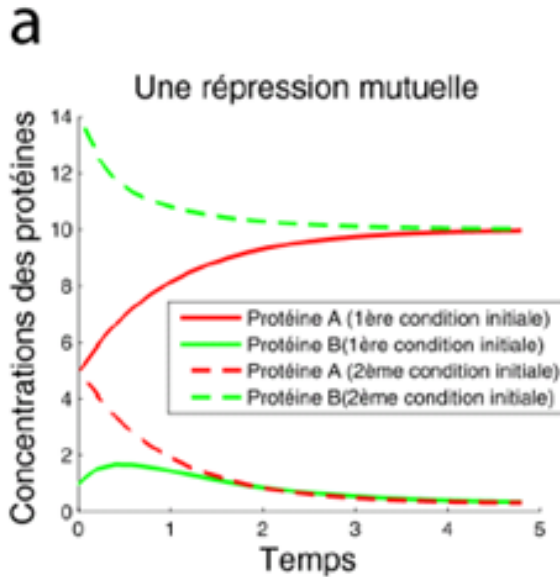
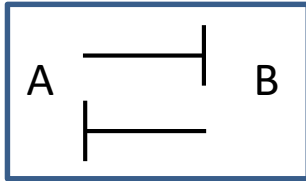
$n$  : coefficient de Hill

$\gamma$  : constante de dégradation

$$\frac{d[B]}{dt} = \beta_{\max} \frac{K_d^n}{[A]^n + K_d^n} - \gamma[B]$$

# Etude du comportement dynamique du réseau (système) : modélisation mathématique

Intégration numérique des équations différentielles et obtention des valeurs simulées des concentrations de protéines au cours du temps (profil dynamique) :

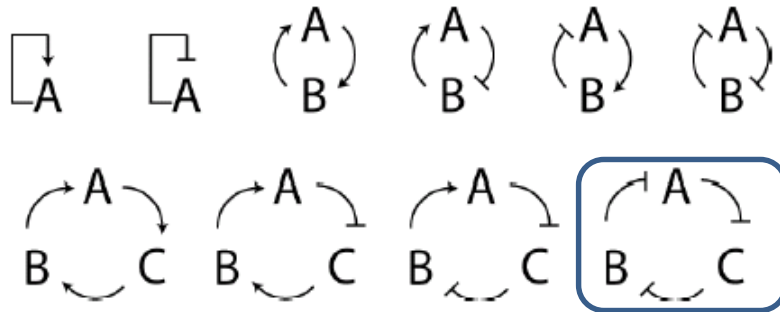


Concentrations initiales des protéines A et B et devenir de l'état du système :

- Si A est présente à haute concentration au début et B à faible concentration, le système atteint un état d'équilibre avec beaucoup de protéines A (courbe rouge trait plein) et peu de protéines B (courbe verte trait plein)
- Si B est présente à une concentration plus élevée que celle de A au début, le système atteint un état d'équilibre inverse avec peu de protéines A (courbe rouge en pointillés) et beaucoup de protéines B (courbe verte en pointillés)
- On dit que le système est **bistable**
- Ce type de motif est appelé un interrupteur (toggle switch) car en perturbant/changeant suffisamment la concentration initiale d'une protéine, on peut faire basculer le système vers un état d'équilibre ou vers l'autre

# Biologie des systèmes

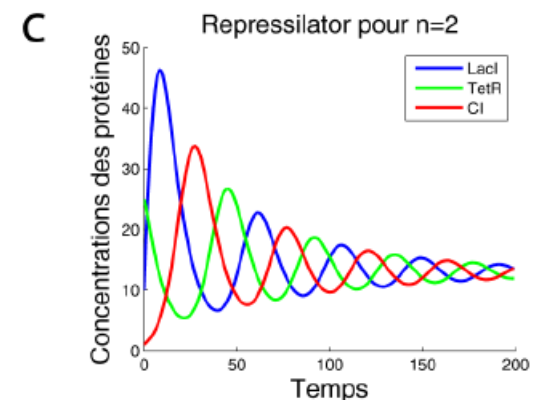
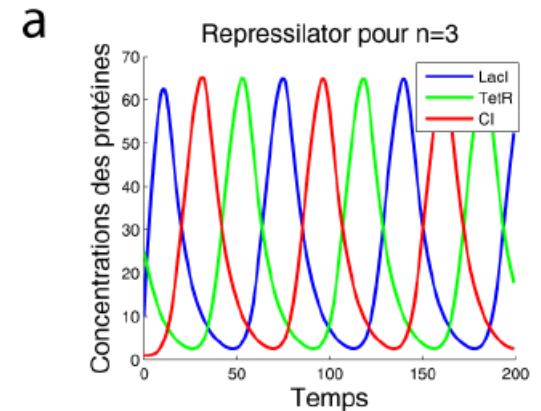
Parmi les réseaux simples ci-dessous lesquels permettent de générer des oscillations des concentrations de protéines ? Pas si facile !



Au moins un, le dernier appelé le « repressilator ».

On observe bien des oscillations des trois protéines en fonction du temps. Cependant, la encore la valeur des paramètres est importante.

- paramètre de Hill  $n = 3$ , les oscillations perdurent et aucun état d'équilibre est atteint.
- paramètre de Hill  $n = 2$ , les oscillations s'atténuent et le système atteint un état d'équilibre stationnaire

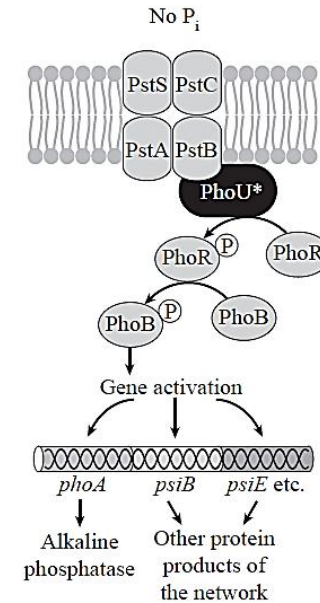
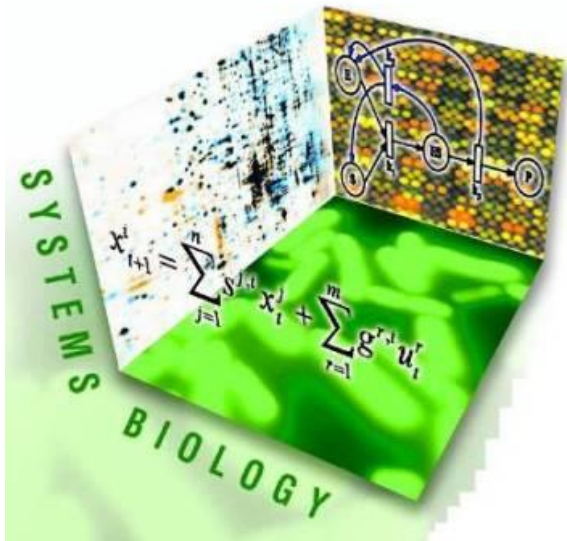


# Biologie des systèmes

Intégration et synthèse des connaissances

- modélisation d'un système
  - circuit de régulation des gènes ou réseaux
  - processus biologique (respiration)
  - organeite (mitochondrie)
  - cellule
  - population
  - écosystème

Représentation d'une cascade de régulation :  
La régulation du phosphate chez les entérobactéries

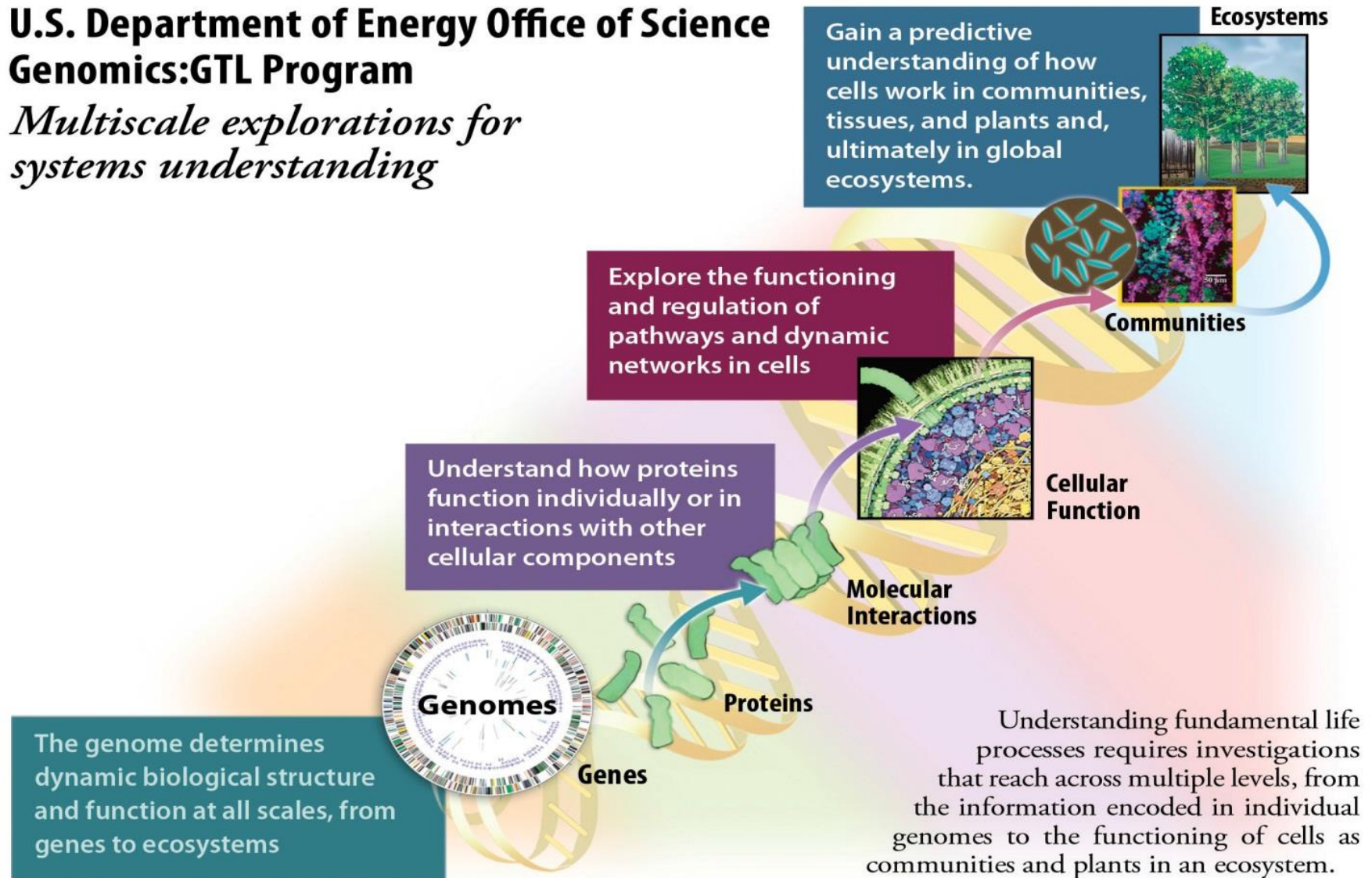


À terme : simulation d'une cellule virtuelle et prédiction de son comportement

# Défis scientifiques

## U.S. Department of Energy Office of Science Genomics:GTL Program

*Multiscale explorations for  
systems understanding*



# Bioinformatique des séquences

# Programme et objectifs

## **Connaître les principales banques de données**

- séquences nucléiques, peptidiques
- structure tridimensionnelle des protéines
- domaines et familles protéiques
- bibliographie

## **Comparaison de deux séquences**

- identification de régions conservées, de répétitions, d'inversions, ...
- matrices de substitutions pour quantifier la similarité des acides aminés
- alignement de deux séquences

## **Analyse de séquences**

- recherche (dans les banques) de séquences similaires à une séquence donnée
  - identification de famille
  - prédiction de fonction
- identification de régions conservées, de domaines
  - alignement multiple de séquences
  - recherche de domaines fonctionnels
  - définition et recherche de motifs/profils correspondant à des régions conservées
  - prédictions fonctionnelles



# Pourquoi comparer des séquences (nucléiques ou protéiques) ?

**Hypothèse 1**: si deux ou plusieurs séquences possèdent des résidus conservés (bases ou acides aminés), cela signifie qu'elles ont une histoire évolutive commune. Elles ont évoluées à partir d'une séquence ancêtre commune.

On dit qu'elles sont **homologues**.

**Hypothèse 2** : si deux séquences sont homologues, alors elles doivent avoir des fonctions similaires.

Le pourcentage de similitude entre deux séquences est considéré comme reflétant la distance évolutive existant entre ces deux séquences. Les différences observées sont dues à l'accumulation de mutations au cours du temps. Les mutations prises en compte sont les substitutions et les insertions/délétions (indels).

# Homologie - orthologie- paralogie

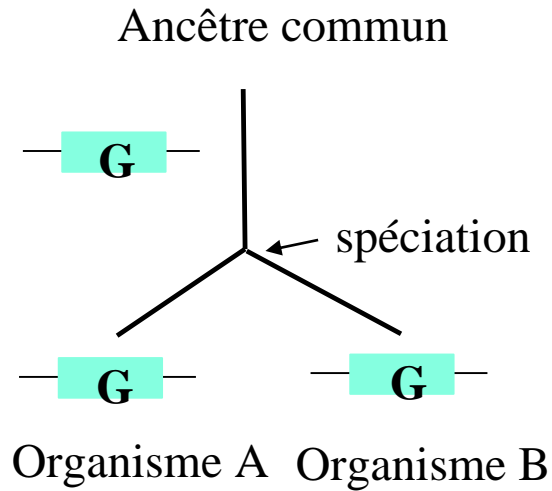
- Deux gènes sont **homologues** s'ils ont divergé à partir d'une séquence ancêtre commune.
- Deux gènes sont **orthologues** si leur divergence est due à la spéciation (le gène ancêtre commun se trouvait dans l'organisme ancêtre).
- Deux gènes sont **paralogues** si leur divergence est due à la duplication du gène ancêtre.

**Donc deux séquences sont ou ne sont pas homologues.**

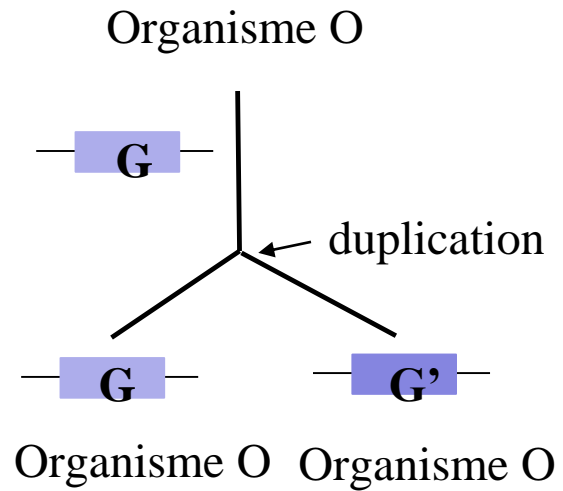
Dire que la protéine X a 80% d'homologie avec la protéine Y est donc **incorrect**:  
soit:

- les deux protéines présentent 80% d'identité (résidus identiques)
- les deux protéines présentent 80% de similarité (résidus similaires)

# Homologie - orthologie- paralogie



**Gènes orthologues**



**Gènes paralogues**

Exemple d'acquisition de données :  
l'annotation d'un génome

TCCTGGCCTACATGTTCTTTGGCAAAGGATCTTCAAAATCAACGGCTCCCGGTGCGGCGATCATCCATTTCTTCGGAGGGATTACAGAGATT  
TACTTCCCGTACATTCTGATGAAACCTGGCCCTGATTCTCGCAGCCATTGCCGGCGGAGCAAGCGGACTCTTAACATTACGATCTTTAATGC  
CGGACTTGTGCGGGCAGCGTCACCGGGAAGCATTATCGCATTGATGGCAATGACGCCAAGAGGAGGCTATTTTCGGCGTATTGGCGGGTGTAT  
TGGTCGCTGCAGCTGTATCGTTCATCGTTTTCAGCAGTGATCCTGAAATCCTCTAAAGCTAGTGAAGAAGACCTGGCTGCCGCAACAGAAAA  
ATGCAGTCCATGAAGGGGAAGAAAAGCCAAGCAGCAGCTGCTTTAGAGGCGGAACAAGCCAAAGCAGAGAAGCGTCTGAGCTGTCTCCTGAA  
AGCGCGAACAAAATTATCTTTTCGTGTGATCCGGGATGGGATCAAGTGCCATGGGGGCATCCATCTTAAGAAACAAAGTAAAAAGCGGAGC  
TTGACATCAGTGTGACCAACACGGCCATTAACAATCTGCCAAGCGATGCGGATATTGTCATCACCCACAAAGATTTAACAGACCGCGCGAAA  
GCAAAGCTGCCGAACGCGACGCACATATCAGTGGATAACTTCTTAACAGCCCCGAAATACGACGAGCTGATTGAAAAGCTGAAAAGTAATCT  
TATAGAAAGAGAGTATTGTCATGCAAGTACTCGCAAAGGAAACATTAAACTCAATCAAACGGTATCATCAAAGAAGAGGCTATCAAATTGG  
CAGGCCAGACGCTGATTGACAACGGCTACGTGACAGAGGATTACATTAGCAAAATGTTTGACCGTGAAGAAACGTCTTCTACGTTTTATGGGG  
AATTTCAATTGCCATTCACACGGCACAGAAGAAGCGAAAAGCGAGGTGCTTCACTCAGGAATTTCAATCATAACAGATTCCAGAGGGCGTTGA  
GTACGGAGAAGGCAACACGGCAAAGTGGTATTCGGCATTGCGGGTAAAAATAATGAGCATTTAGACATTTTGTCTAACATCGCCATTATCT  
GTTCAGAAGAAGAAACATTGAACGCCTGATCTCCGCTAAAGCGAAGAAGATTTGATCGCCATTTCAACGAGGTGAACTGACATGATCGCCTT  
ACATTTTCGGTGCGGGAAATATCGGGAGAGGATTTATCGGCGCGCTGCTTCACTCAGGCTATGATGTGGTGTGTTGCGGATGTGAACGAAA  
CGATGGTCAGCCTCCTCAATGAAAAAAGAATACACAGTGGAACGGCGGAAGAGGGACGTTTCATCGGAGATCATTGGCCCGGTGAGCGCT  
ATTAACAGCGGCAGTCAGACCGAGGAGCTGTACCGGCTGATGAATGAGGCGGCGCTCATCACAACAGCTGTCGGCCCCGAATGTCTGAAGCT  
GATTGCCCGTCTATCGCAGAAGGTTTTAAGACGAAGAAATACTGCAAACACACTGAATATCATTGCCTGCGAAAATATGATTGGCGGAAGCA  
GCTTCTTAAAGAAAGAAATATACAGCCATTTAACGGAAGCAGAGCAGAAATCCGTCAGTGAAACGTTAGGTTTTCCGAATTCTGCCGTTGAC  
CGGATCGTCCCGATTACGATCATGAAGACCCGCTGAAAGTATCGGTTGAACCATTTTTTCGAATGGGTCATTGATGAATCAGGCTTTAAAGG  
GAAAACACCAGTCATAAACGGCGCACTGTTTGTGATGATTTAACGCCGTACATCGAACGGAAGCTGTTTACGGTCAATACCGGACACGCGG  
TCACAGCGTATGTCGGCTATCAGCGCGGACTCAAACGGTCAAAGAAGCAATTGATCATCCGGAAATCCGCCGTGTTGTTTCATTCGGCGCTG  
CTTGAAACTGGTGACTATCTCGTCAAATCGTATGGCTTTAAGCAAACCTGAACACGAACAATATATTAAAAATCAGCGGTGCTTTTTAAATC  
CTTTCAATTTTCGGACGATGTGACCCGCGTAGCGAGGTCACCTCTCAGAAAATGAGACTTGTAGGCCCCGGCAAAGAAAATAA  
AAGAACCGAATGCACTGGCTGAAGGAATTGCCGCAGCACTGCGCTTCGATTTACCGGTGACCCTGAAGCGGTTGAACTGCAAGCGCTGATC  
GAAGAAAAGGATACAGCGGCGTACTTCAAGAGGTGTGCGGCATTCAGTCCCATGAACCGTTGCACGCCATCATTTTAAAGAACTTAATCAA  
TAACCGACCACCCGTGACACAATGTCACGGGCTTTTTACTATCTCGCAATCTAGTATAATAGAAAGCGCTTACGATAACAGGGGAAGGAGAA  
TGACGATGAAACAATTTGAGATTGCGGCAATACCGGGAGACGGAGTAGGAAAGAGGTTGTAGCGGCTGCTGAGAAAGTGCTTCATACAGCGG  
CTGAGGTACACGGAGGTTTTGTCATTCTCATTACAGCTTTTCCATGGAGCTGTGATTATTACTTGGAGCACGGCAAAAATGATGCCCGAAGA  
TGGAATACATACGCTTACTCAATTTGAAGCAGTTTTTTGGGAGCTGTCGGAAATCCGAAGCTGGTTCCCGATCATATATCGTTATGGGGCTGC  
TGCTGAAATCCGGAGGGAGCTTGAGCTTTCCATTAATATGAGACCCGCCAAACAAATGGCAGGCATTACGTCGCCGCTTCTGCATCCAAATG  
ATTTTTGACTTCGTGGTGAATTCGCGAGAACAGTGAAGGTGAATACAGTGAAGTTGTCGGGCGCATTACAGAGGCGATGATGAAATCGCCAT  
CCAGAATGCCGTGTTTACGAGAAAAGCGACAGAACGTGTCATGCGCTTTGCCTTCGAATT

# Annotation d'un génome

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....)

# Exemple d'acquisition de données : l'annotation d'un génome *Mycoplasma genitalium*

Distribution des unités de traduction et classification fonctionnelle



# Exploitation des données : une illustration d'une démarche bioinformatique

## Independent evolution of competence regulatory cascades in streptococci?

Bernard Martin, Yves Quentin, Gwennaele Fichant and Jean-Pierre Claverys

Trends in Microbiology, Volume 14, Issue 8 , August 2006, Pages 339-345

## La transformation génétique naturelle

Étapes :

- capture d'ADN exogène
- internalisation
- intégration dans le génome

Processus **largement répandu** chez les bactéries

- > 80 espèces de bactéries, distribuées dans tous les groupes taxonomiques.

Rôles de la transformation

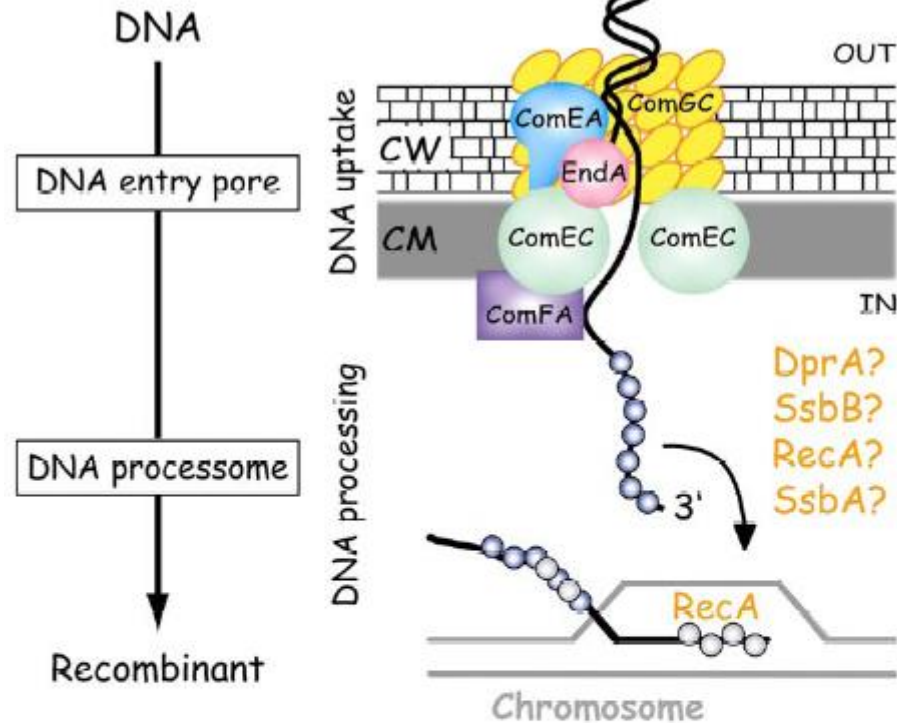
- échanges génétiques (sexualité bactérienne)
- réparation de l'ADN
- nutriments

**La compétence** : état physiologique permettant la transformation, génétiquement programmé et transitoire



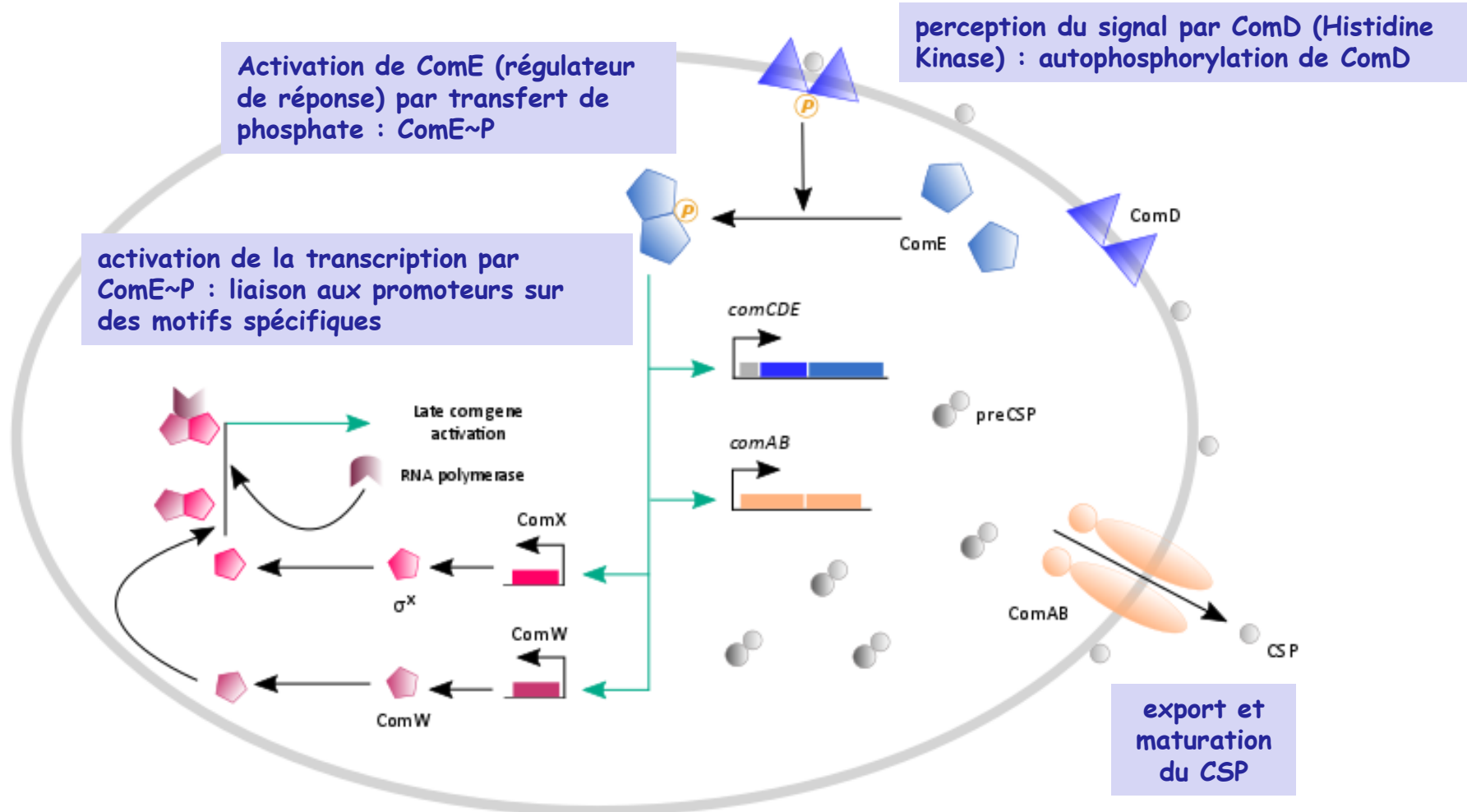
# Le transformasome

DNA transformasome:



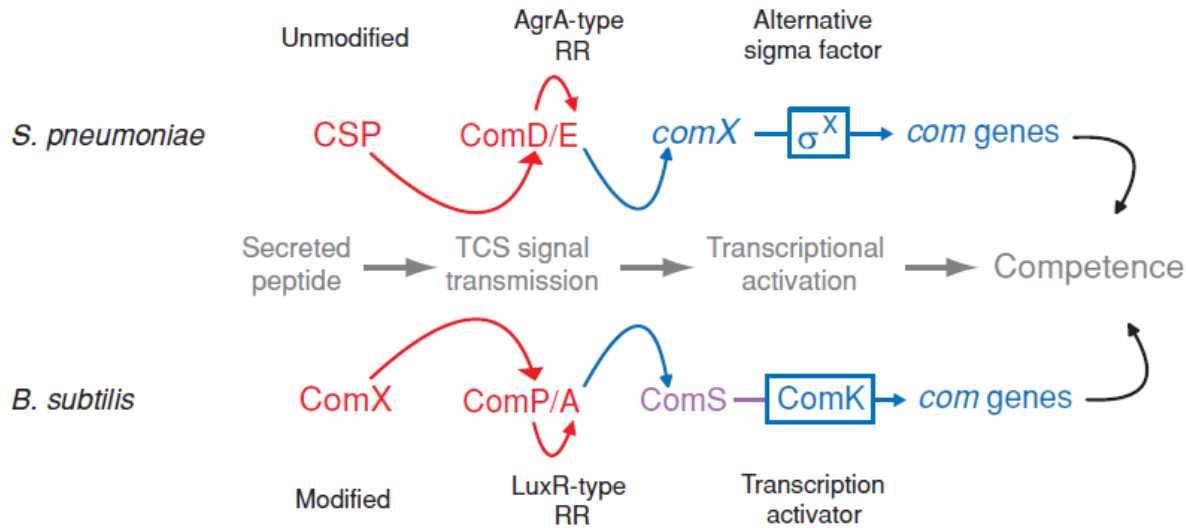
# Régulation de l'état de compétence : modèle *S. pneumoniae*

CSP : Competence Signal Peptide



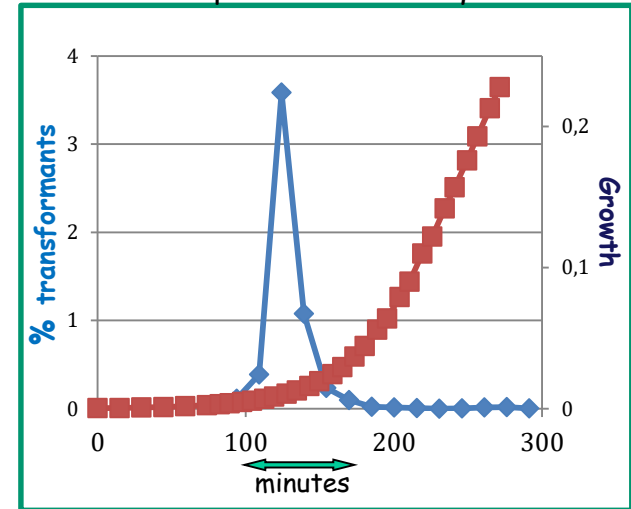
# La cascade de régulation

**Exemple :** cascades similaires mais acteurs différents chez *Streptococcus pneumoniae* et *Bacillus subtilis* : deux bactéries avec des styles de vie différents



Claverys et al. (2006) *Annu. Rev. Microbiol.* 60: 451-75

Etat de compétence chez *S. pneumoniae*



a) *S. pneumoniae*

CSP: peptide non-modifié  
ComAB: export de CSP  
ComD/E: TCS AgrA-type  
ComX: facteur sigma

b) *B. subtilis*

ComX: peptide modifié  
ComQ: export de ComX  
ComP/A: TCS LuxA-type  
ComS: accumulation de ComK  
ComK: facteur de transcription

a) *S. pneumoniae*

Induite en phase exponentielle  
Touche ensemble de la population  
Induction rapide  
Délimitée dans le temps

b) *B. subtilis*

Induite en phase stationnaire  
Touche ~10% de la population  
Induction lente  
Période étendue (~2 heures)

# Questions posées

Chez *S. pneumoniae* apparition du pic de compétence environ 10 minutes après l'ajout de CSP dans le milieu.

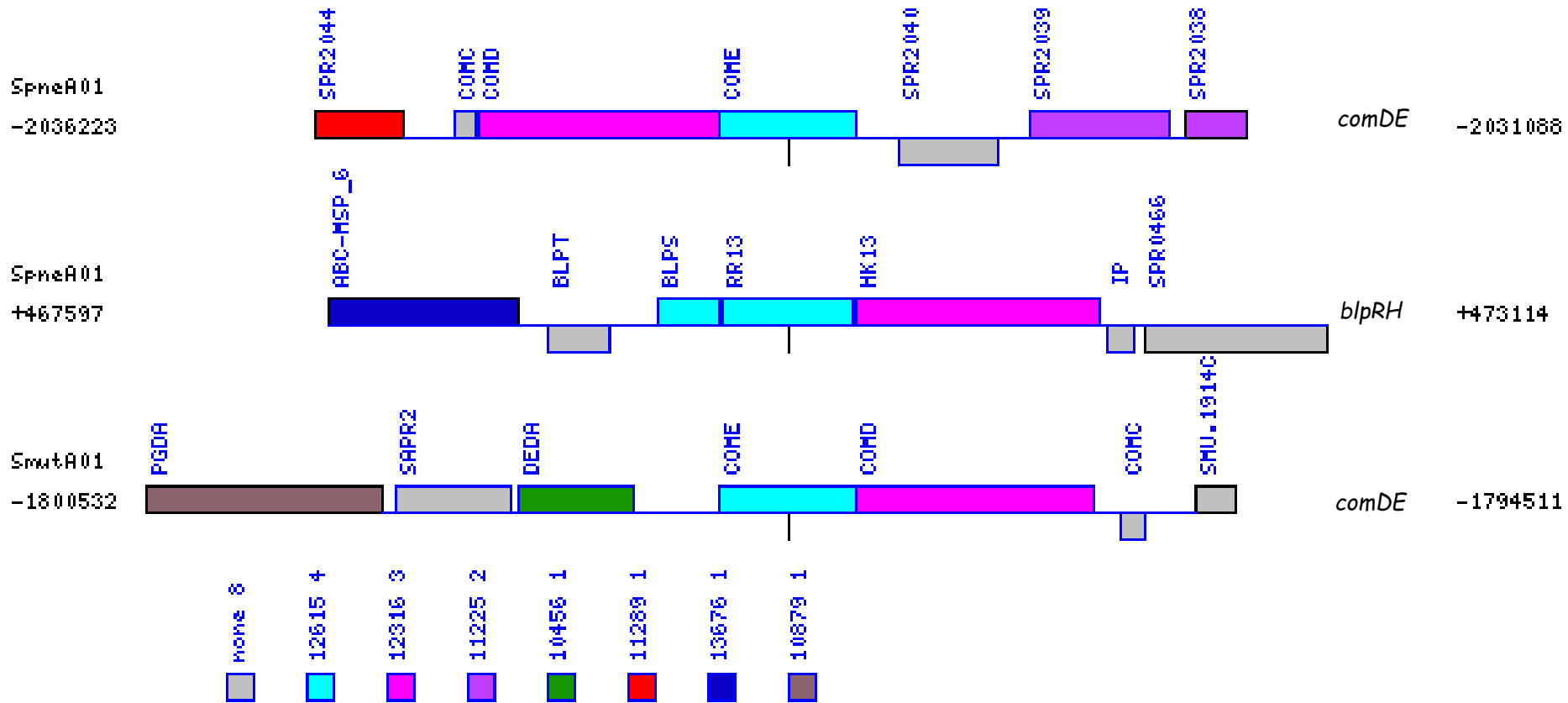
Chez *S. mutans*, ce temps de latence est d'au moins 2 heures.

- Peut-on trouver une explication à cette différence ?
- Peut-on définir qu'elles sont les espèces de Streptocoques pour lesquelles la régulation est du même type que celle de *S. pneumoniae* ?

# Voisinage chromosomique

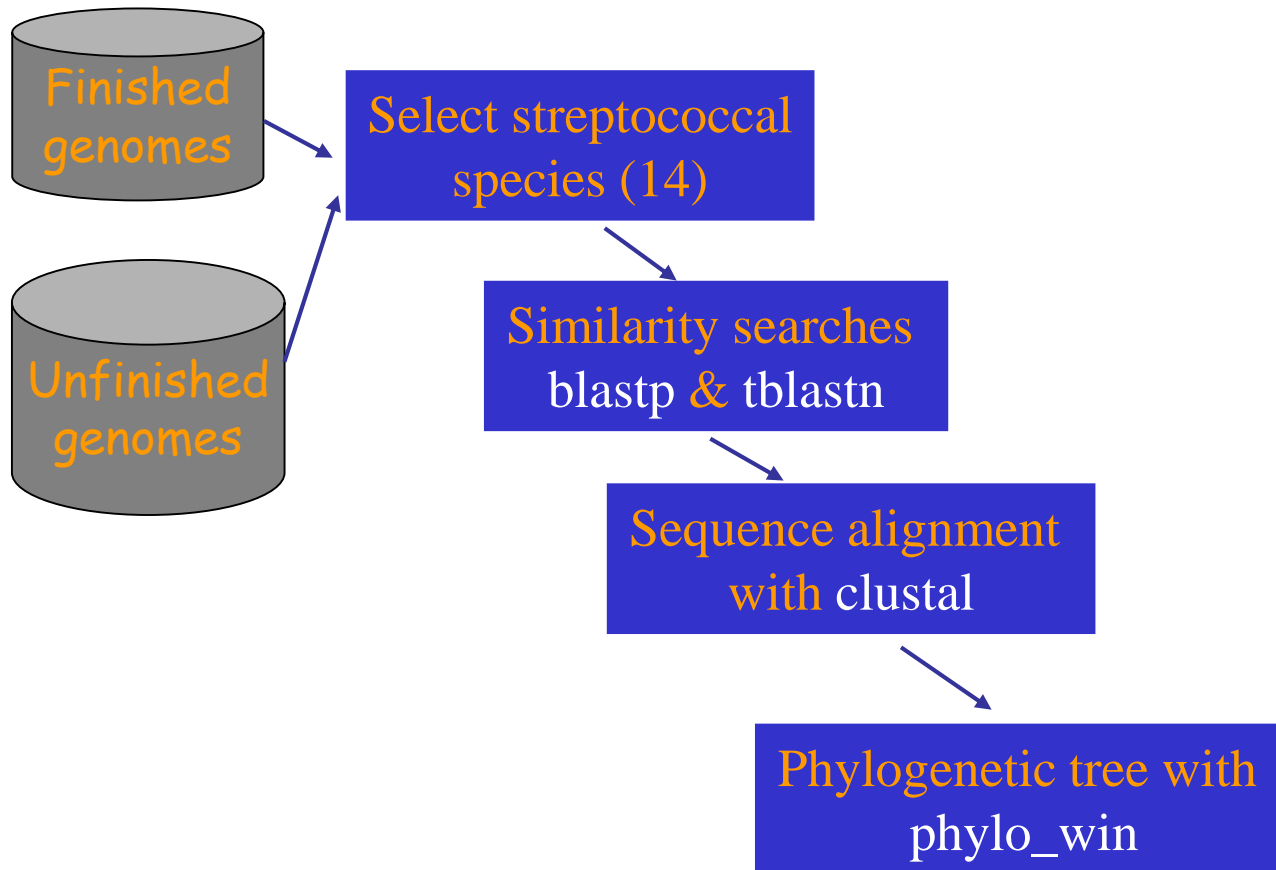
Le génome de *S. pneumoniae* code pour **deux** TCS paralogues ComDE, BlpRH.  
BlpR contrôle l'expression du régulon Blp impliqué dans la production de peptides de type bactériocines.

Le génome de *S. mutans* ne code que pour **un seul** système de ce type (« comDE »).



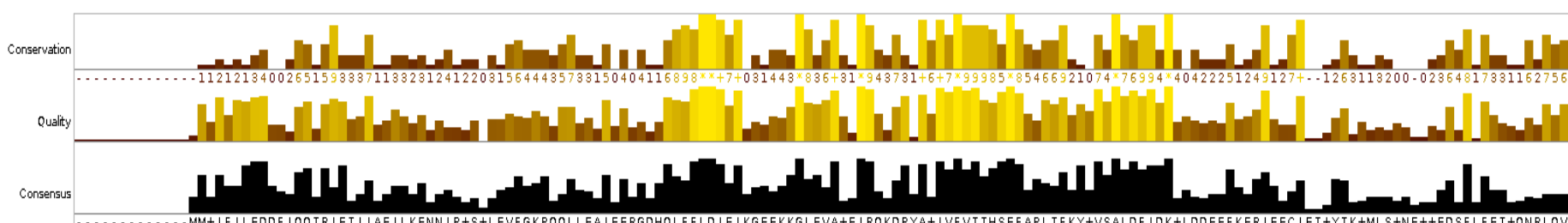
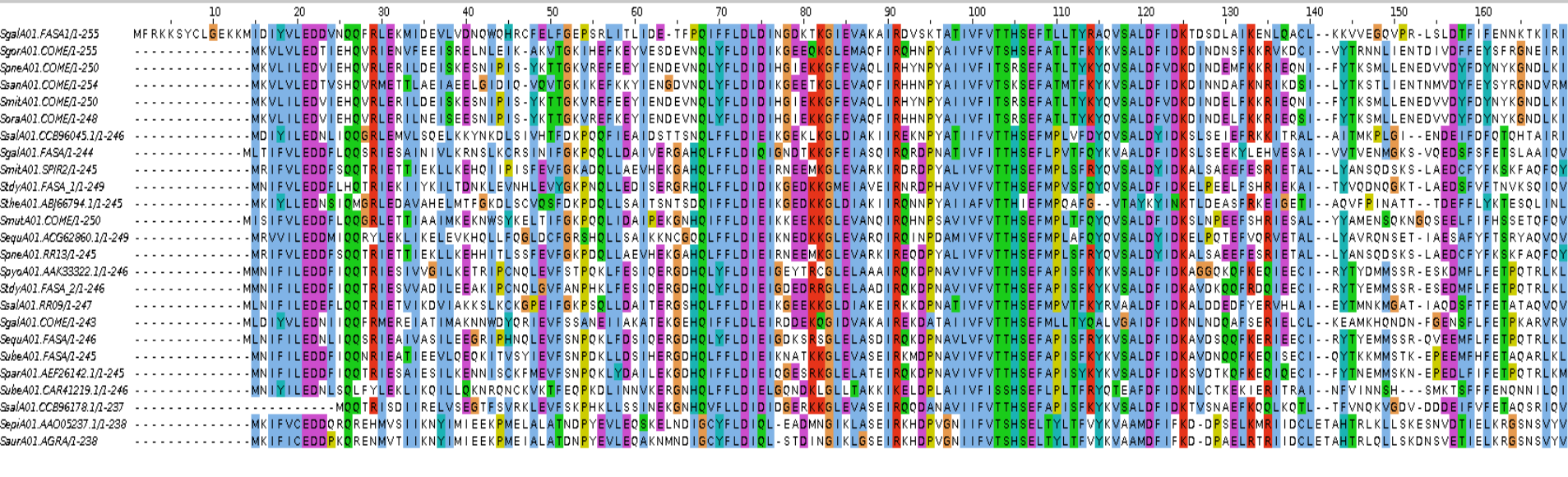
# Distribution des gènes homologues à *comD* et *comE*

Séquences de référence : protéines ComD et ComE de *S. pneumoniae*

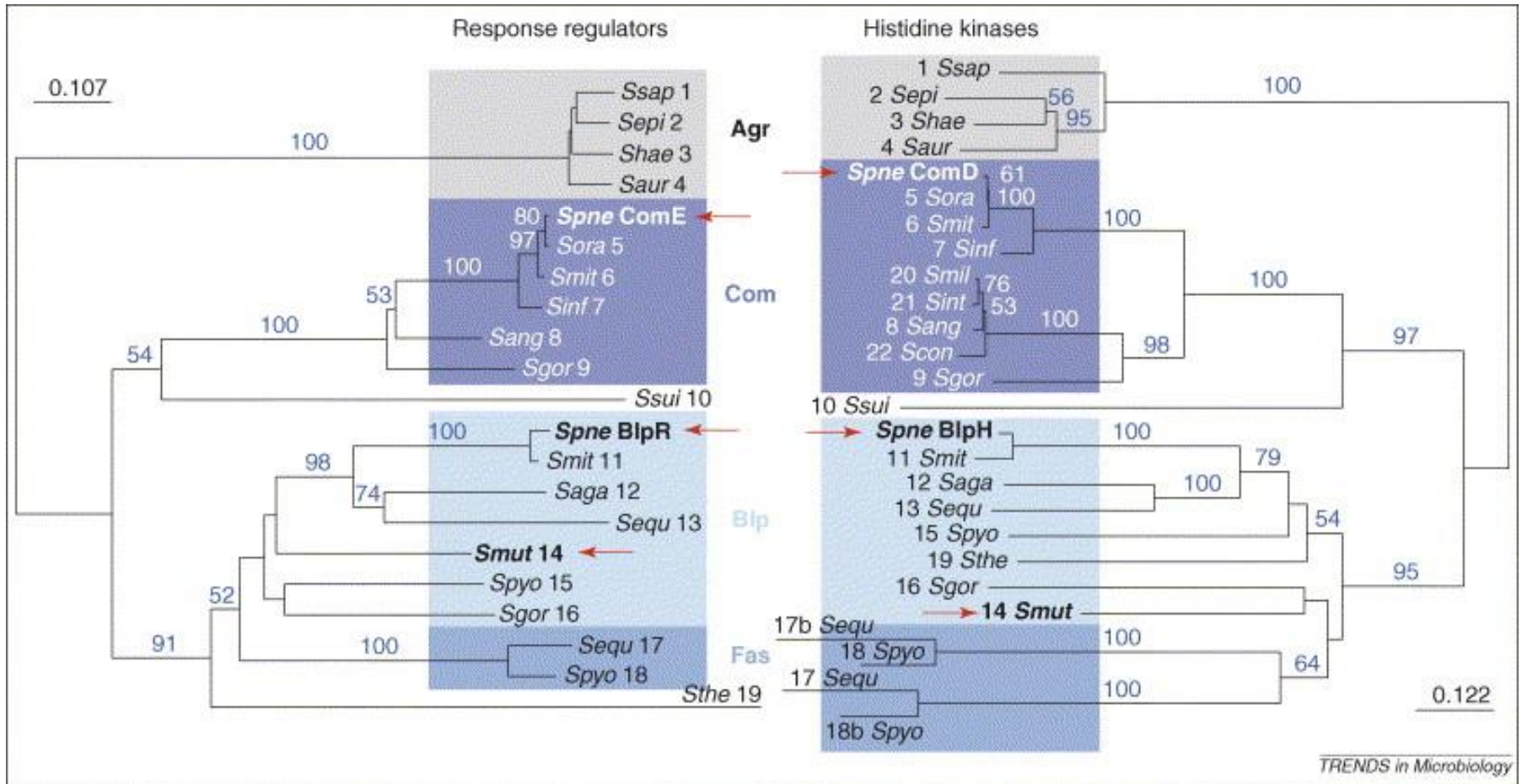


# Exemple de l'alignement multiple des protéines homologues à ComE

File Edit Select View Annotations Format Colour Calculate Web Service



# Relations évolutives entre les protéines homologues à ComD et ComE



Arbre enraciné avec Agr.

Trois groupes : Com, Blp et Fas.

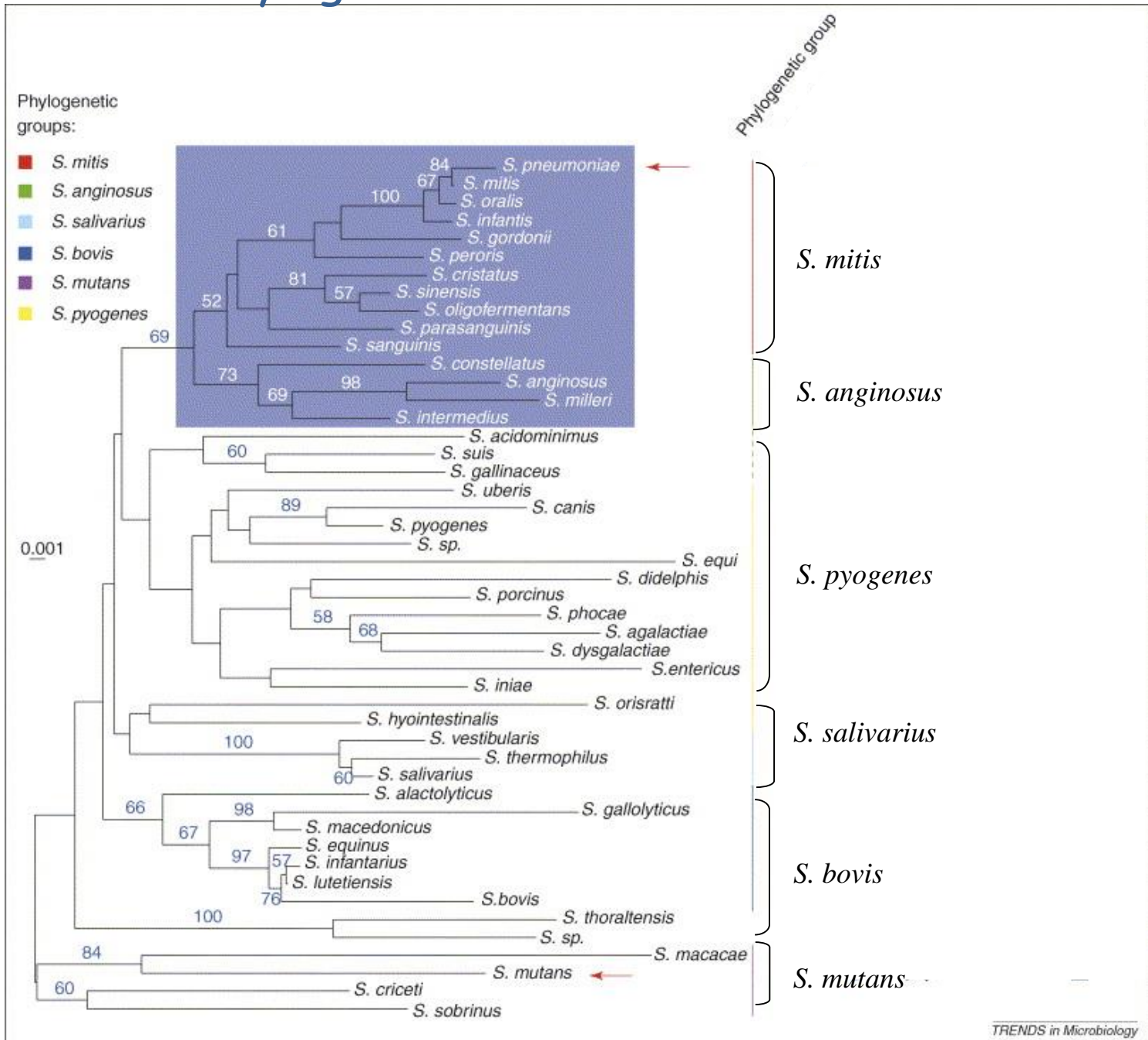
Topologies congruentes avec celle de l'arbre de l'ARNr 16S.

Trajectoires parallèles pour les paires (RR, HK).

*Les gènes comE et comD orthologues uniquement dans les groupes mitis et anginosus.*



# Phylogénie basée sur l'ARNr 16S



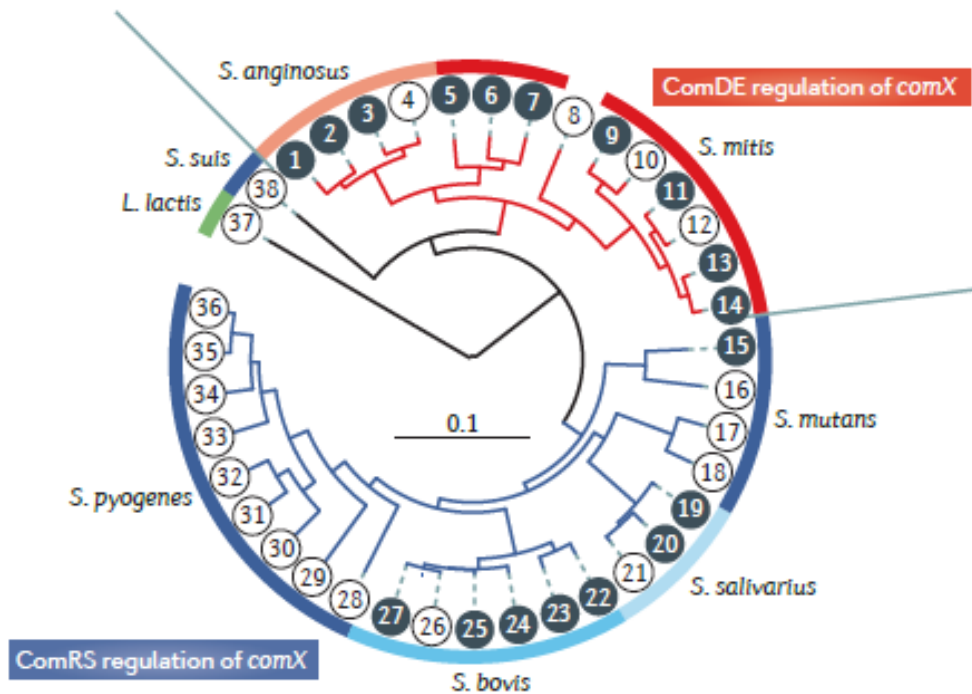
# Origine des gènes

## Observations

- ❑ Opéron *comCDE* présente un %G+C nettement inférieur à celui du génome.
- ❑ L'opéron est encadré par des gènes codant pour des ARNt Arg et Glu.
  - Acquisition de ce morceau d'ADN par **transfert latéral** de gène,
  - juste avant l'émergence du groupe (*mitis*, *anginosus*).
- ❑ Des gènes orthologues à *blpRH* sont observés dans tous les streptocoques analysés.
  - l'ancêtre de *blpRH* déjà présent chez l'ancêtre commun des streptocoques.

# Deux types de régulation au sein des Streptocoques

En 2010, un autre système de régulation ComRS a été identifié dans les espèces de Streptocoques en dehors du groupe mitis-anginosus



Martin *et al.* (2006) *Trends Microbiol.* 14(8):339-45; Gardann *et al.* (2009) *J Bacteriol* 191(14):4647-655; Fontaine *et al.* (2010) *J Bacteriol* 192(5):1444-54; Mashburn-Warren *et al.* (2010) *Mol Microbiol* 78(3):589-606.; Håvarstein (2010) *Mol Microbiol* 78(3):541-44; Mashburn-Warren *et al.* (2012) *J Bacteriol* 194(17):4589-600.