Automatisation d'Analyses de Données Biologiques

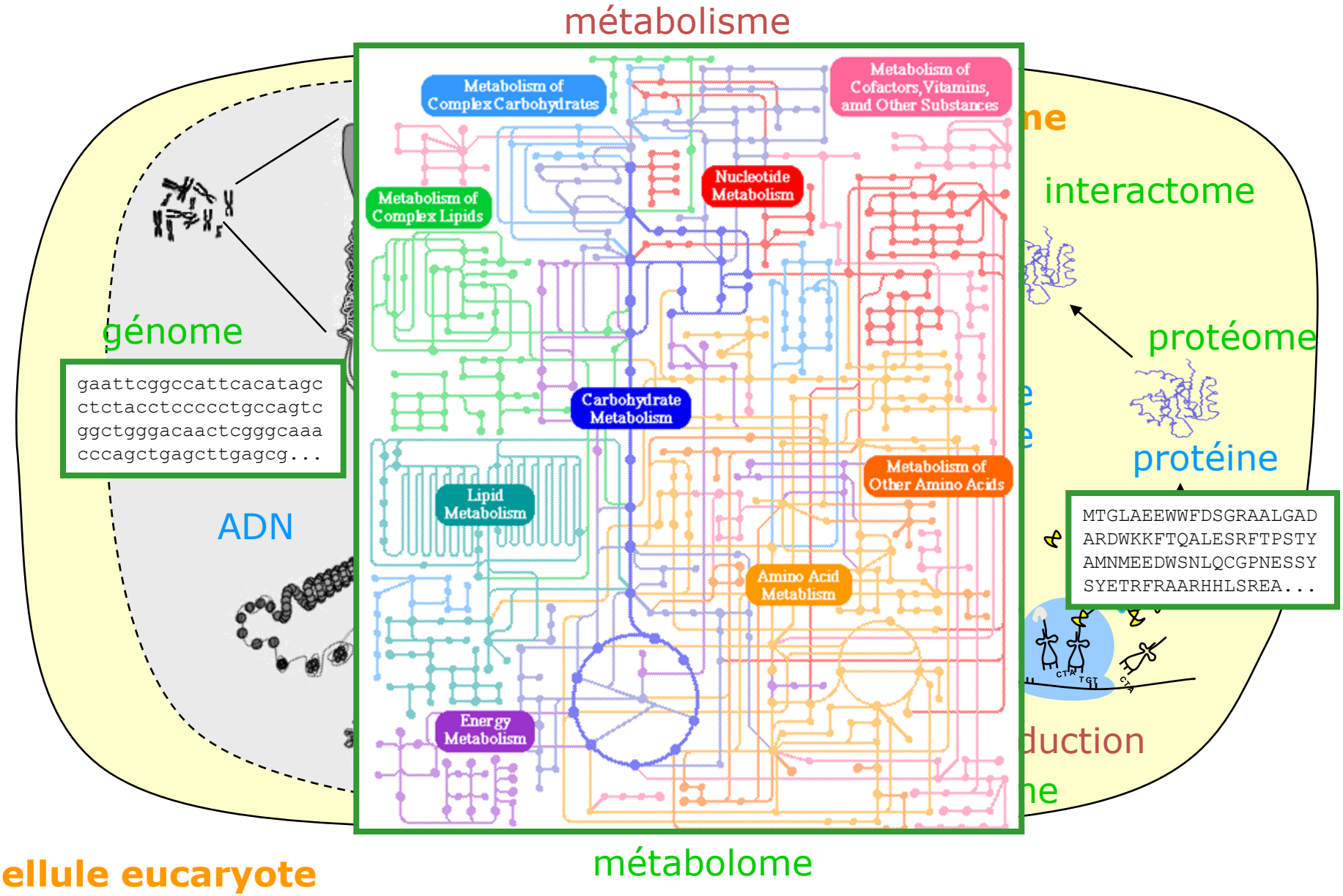
Objectifs pédagogiques :

- Apperçu des données et connaissances biologiques actuellement disponibles
- Distinction entre le protocole d'analyses et les données
- Organisation rationnelle des données
- Outils pour l'automatisation
- Mise en œuvre d'un protocol
- Sélection, synthèse et présentation des résultats

Roland Barriot

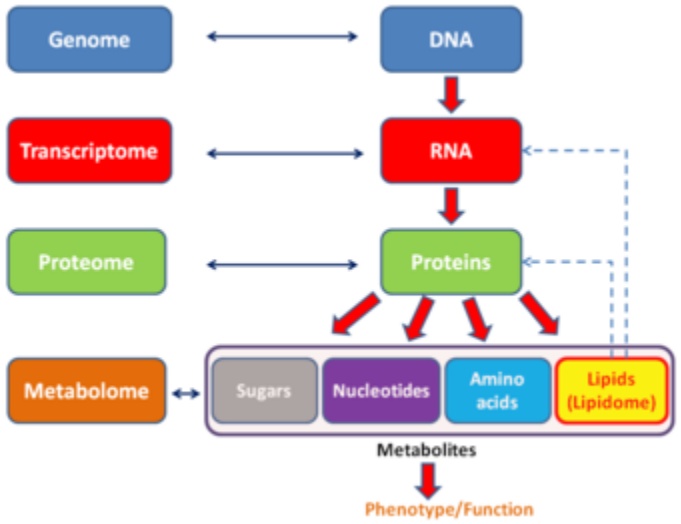
Elodie Gaulin

(Quelques) données et connaissances disponibles



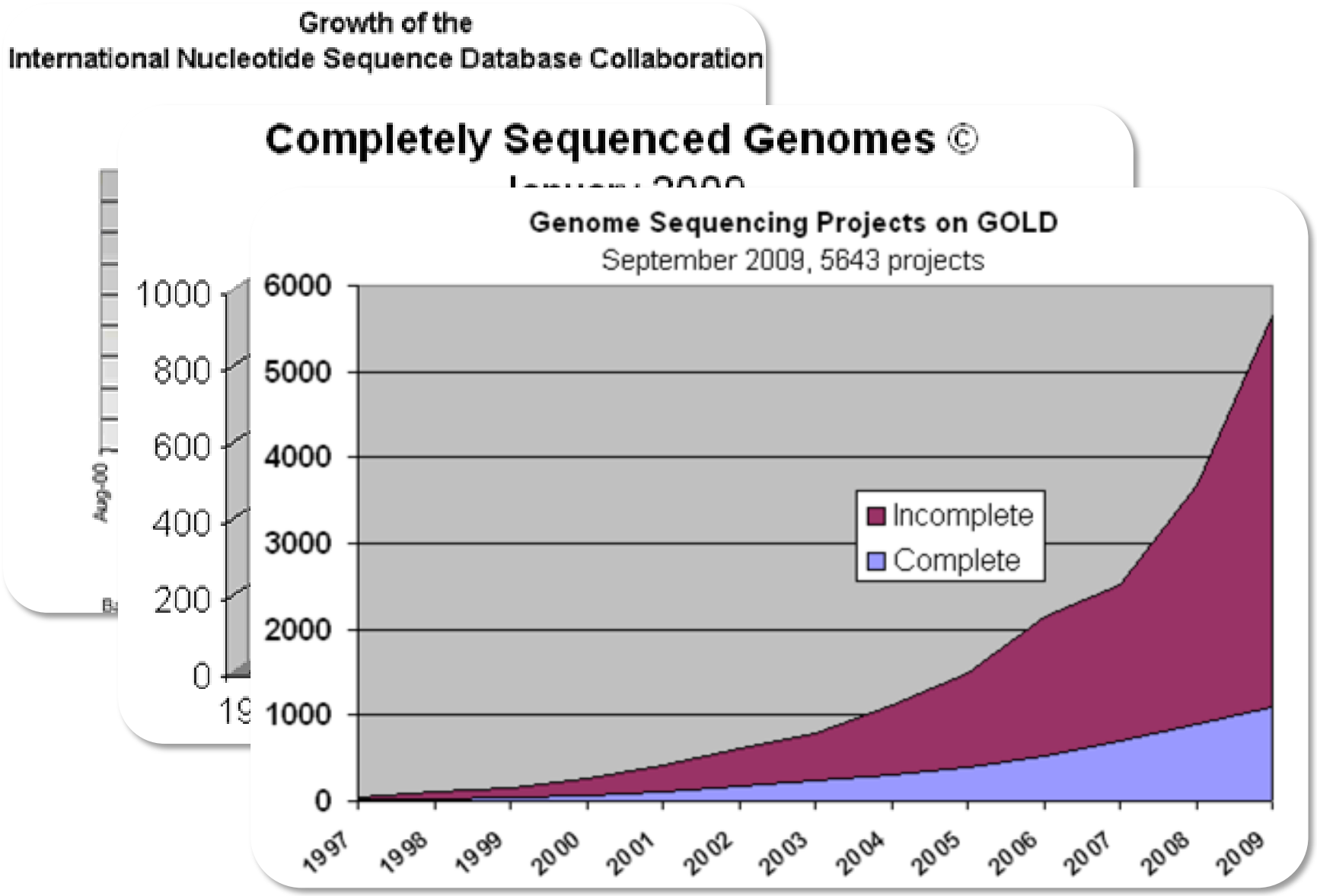
Données omics

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



source : Wikipedia

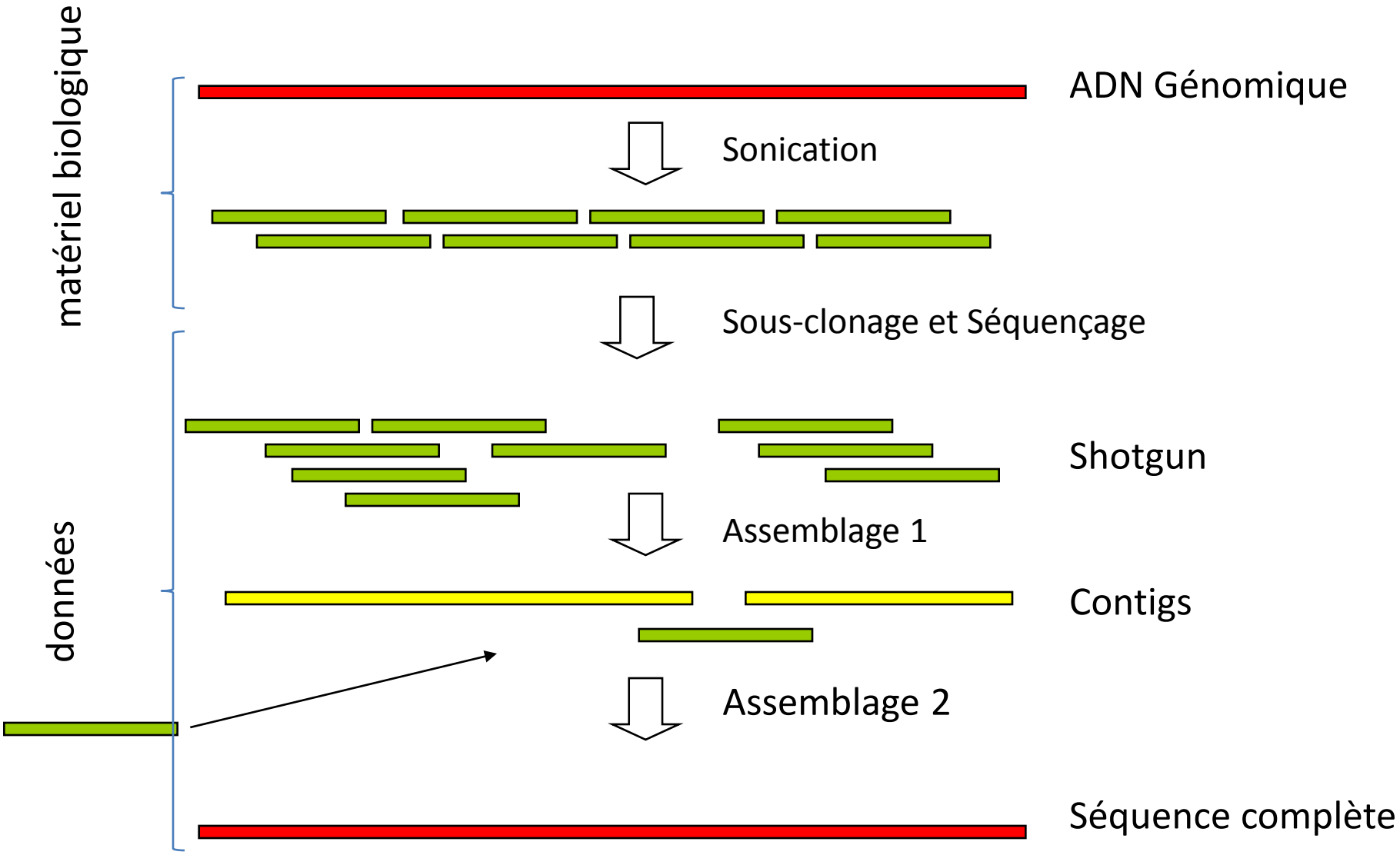
Séquences disponibles : quelques chiffres



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

TCCTGGCCTACATGTTCTTTGGCAAAGGATCTTCAAAATCAACGGCTCCCGGTGCGGCGATCATCCATTTCTTCGGAGGGATTCACGAGATT
TACTTCCCGTACATTCTGATGAAACCTGGCCCTGATTCTCGCAGCCATTGCCGGCGGAGCAAGCGGACTCTTAACATTACGATCTTTAATGC
CGGACTTGTCGCGGCAGCGTCACCGGGAAGCATTATCGCATTGATGGCAATGACGCCAAGAGGAGGCTATTTCGGCGTATTGGCGGGTGTAT
TGGTCGCTGCAGCTGTATCGTTCATCGTTTCAGCAGTGATCCTGAAATCCTCTAAAGCTAGTGAAGAAGACCTGGCTGCCGCAACAGAAAAA
ATGCAGTCCATGAAGGGGAAGAAAAGCCAAGCAGCAGCTGCTTTAGAGGCGGAACAAGCCAAAGCAGAGAAGCGTCTGAGCTGTCTCCTGAA
AGCGCGAACAAAATTATCTTTTCGTGTGATCCGGGATGGGATCAAGTGCCATGGGGGCATCCATCTTAAGAAACAAAGTGAAAAAGCGGAGC
TTGACATCAGTGTGACCAACACGGCCATTAACAATCTGCCAAGCGATGCGGATATTGTCATCACCCACAAAGATTTAACAGACCGCGCGAAA
GCAAAGCTGCCGAACGCGACGCACATATCAGTGGATAACTTCTTAAACAGCCCGAAATACGACGAGCTGATTGAAAAGCTGAAAAGTAATCT
TATAGAAAGAGAGTATTGTCATGCAAGTACTCGCAAAGGAAACATTAAACTCAATCAAACGGTATCATCAAAAGAAGAGGCTATCAAATTGG
CAGGCCAGACGCTGATTGACAACGGCTACGTGACAGAGGATTACATTAGCAAAATGTTTGACCGTGAAGAAACGTCTTCTACGTTTATGGGG
AATTTCATTGCCATTCCACACGGCACAGAAGAAGCGAAAAGCGAGGTGCTTCACTCAGGAATTTCAATCATACAGATTCCAGAGGGCGTTGA
GTACGGAGAAGGCAACACGGCAAAAGTGGTATTCGGCATTGCGGGTAAAAATAATGAGCATTTAGACATTTTGTCTAACATCGCCATTATCT
GTTCAGAAGAAGAAACATTGAACGCCTGATCTCCGCTAAAGCGAAGAAGATTTGATCGCCATTTCAACGAGGTGAACTGACATGATCGCCTT
ACATTTCGGTGCGGGAAATATCGGGAGAGGATTTATCGGCGCGCTGCTTCACCACTCCGGCTATGATGTGGTGTTTGCGGATGTGAACGAAA
CGATGGTCAGCCTCCTCAATGAAAAAAAAGAATACACAGTGGAACTGGCGGAAGAGGGACGTTCATCGGAGATCATTGGCCCGGTGAGCGCT
ATTAACAGCGGCAGTCAGACCGAGGAGCTGTACCGGCTGATGAATGAGGCGGCGCTCATCACAACAGCTGTCGGCCCGAATGTCCTGAAGCT
GATTGCCCCGTCTATCGCAGAAGGTTTAAGACGAAGAAATACTGCAAACACACTGAATATCATTGCCTGCGAAAATATGATTGGCGGAAGCA
GCTTCTTAAAGAAAGAAATATACAGCCATTTAACGGAAGCAGAGCAGAAATCCGTCAGTGAAACGTTAGGTTTTCCGAATTCTGCCGTTGAC
CGGATCGTCCCGATTCAGCATCATGAAGACCCGCTGAAAGTATCGGTTGAACCATTTTTCGAATGGGTCATTGATGAATCAGGCTTTAAAGG
GAAAACACCAGTCATAAACGGCGCACTGTTTGTTGATGATTTAACGCCGTACATCGAACGGAAGCTGTTTACGGTCAATACCGGACACGCGG
TCACAGCGTATGTCGGCTATCAGCGCGGACTCAAAACGGTCAAAGAAGCAATTGATCATCCGGAAATCCGCCGTGTTGTTCATTCGGCGCTG
CTTGAAACTGGTGACTATCTCGTCAAATCGTATGGCTTTAAGCAAACTGAACACGAACAATATATTAAAAATCAGCGGTCGCTTTTAAAATC
CTTTCATTTCGGACGATGTGACCCGCGTAGCGAGGTCACCTCTCAGAAAACTGGGAGAAAATGTAGACTTGTAGGCCCGGCAAAGAAAATAA
AAGAACCGAATGCACTGGCTGAAGGAATTGCCGCAGCACTGCGCTTCGATTTCACCGGTGACCCTGAAGCGGTTGAACTGCAAGCGCTGATC
GAAGAAAAGGATACAGCGGCGTACTTCAAGAGGTGTGCGGCATTCAGTCCCATGAACCGTTGCACGCCATCATTTTAAAGAAACTTAATCAA
TAACCGACCACCCGTGACACAATGTCACGGGCTTTTTACTATCTCGCAATCTAGTATAATAGAAAGCGCTTACGATAACAGGGGAAGGAGAA
TGACGATGAAACAATTTGAGATTGCGGCAATACCGGGAGACGGAGTAGGAAAGAGGTTGTAGCGGCTGCTGAGAAAGTGCTTCATACAGCGG
CTGAGGTACACGGAGGTTTGTCATTCTCATTCACAGCTTTTCCATGGAGCTGTGATTATTACTTGGAGCACGGCAAAAATGATGCCCGAAGA
TGGAATACATACGCTTACTCAATTTGAAGCAGTTTTTGGGAGCTGTCGGAAATCCGAAGCTGGTTCCCGATCATATATCGTTATGGGGCTGC
TGCTGAAATCCGGAGGGAGCTTGAGCTTTCCATTAATATGAGACCCGCCAAACAAATGGCAGGCATTACGTCGCCGCTTCTGCATCCAAATG
ATTTTTGACTTCGTGGTGATTCGCGAGAACAGTGAAGGTGAATACAGTGAAGTTGTCGGGCGCATTCACAGAGGCGATGATGAAATCGCCAT
CCAGAATGCCGTGTTTACGAGAAAAGCGACAGAACGTGTCATGCGCTTTGCCTTCGAATT

Annotation d'un génome

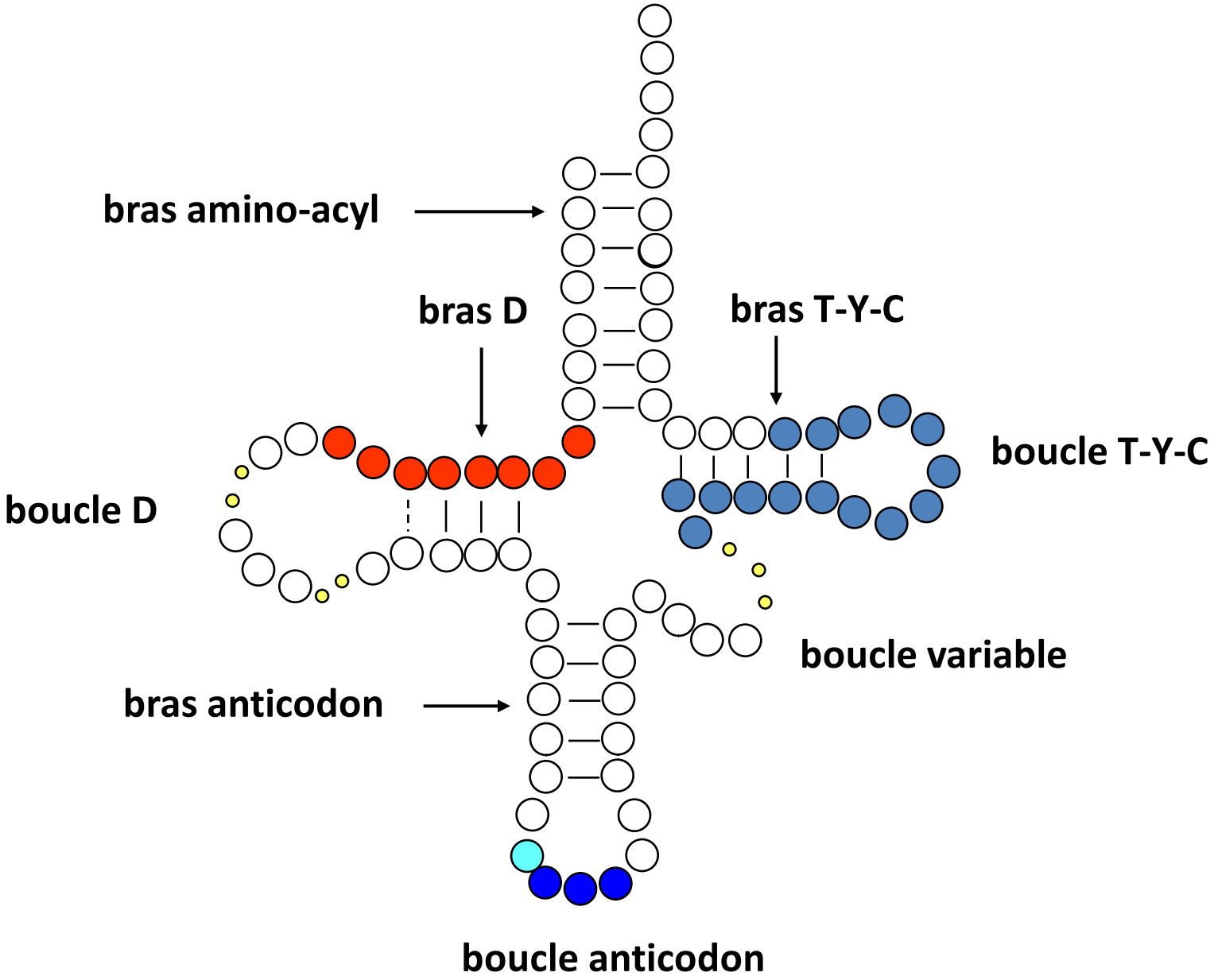
- Identification des gènes codant pour :
  - les ARNr
  - les ARNt
  - les protéines

- Identification des unités de traduction

- Identification des unités de transcription (promoteur et terminateur)

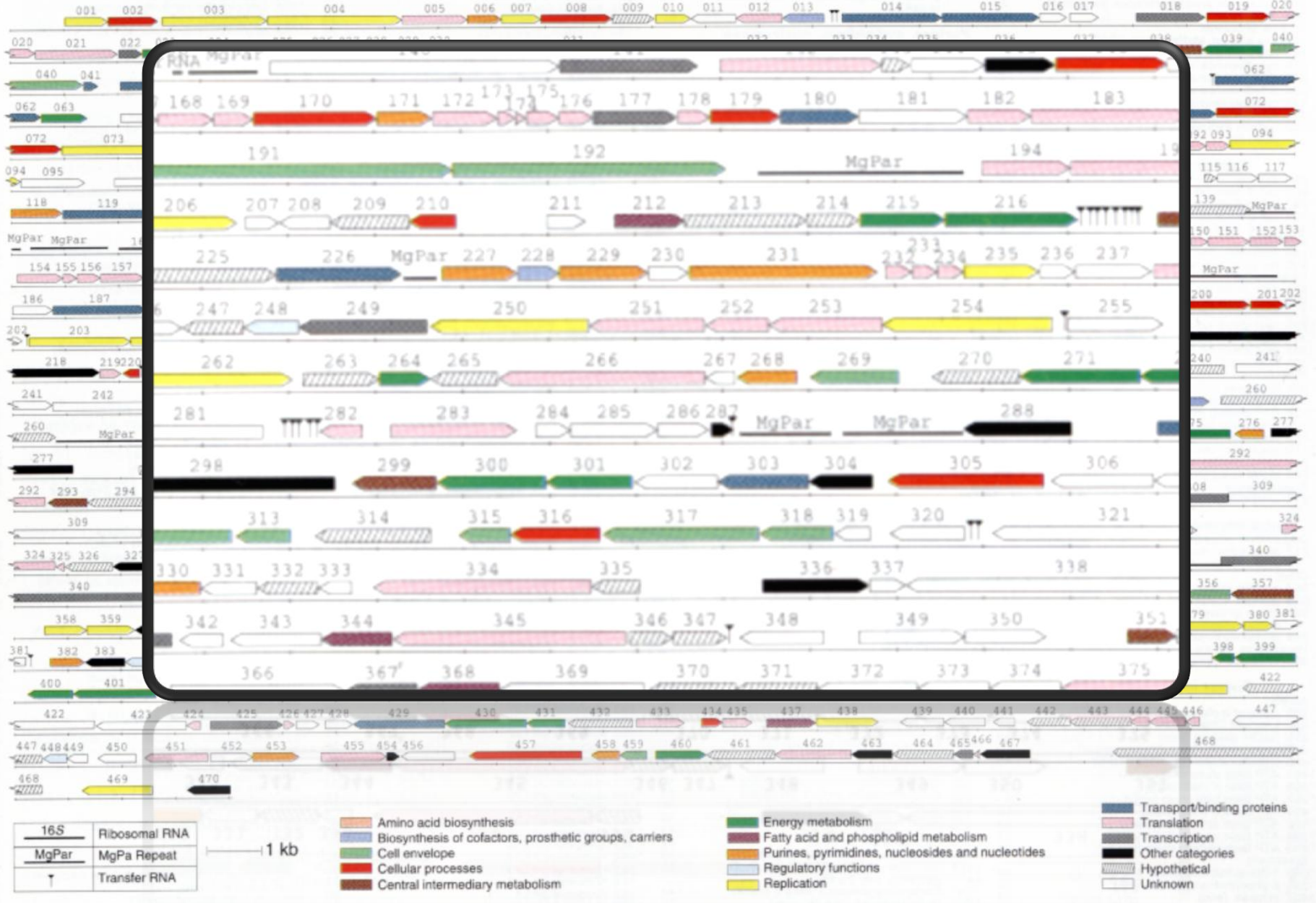
- Pour les gènes codant pour les protéines, prédiction fonctionnelle par recherche de similarité de séquences (Blast) et classification en grandes classes fonctionnelles (ex: biosynthèse des acides aminés, métabolisme énergétique….)

Structure secondaire canonique d'une séquence d'ARNt



Génome de *Mycoplasma genitalium*

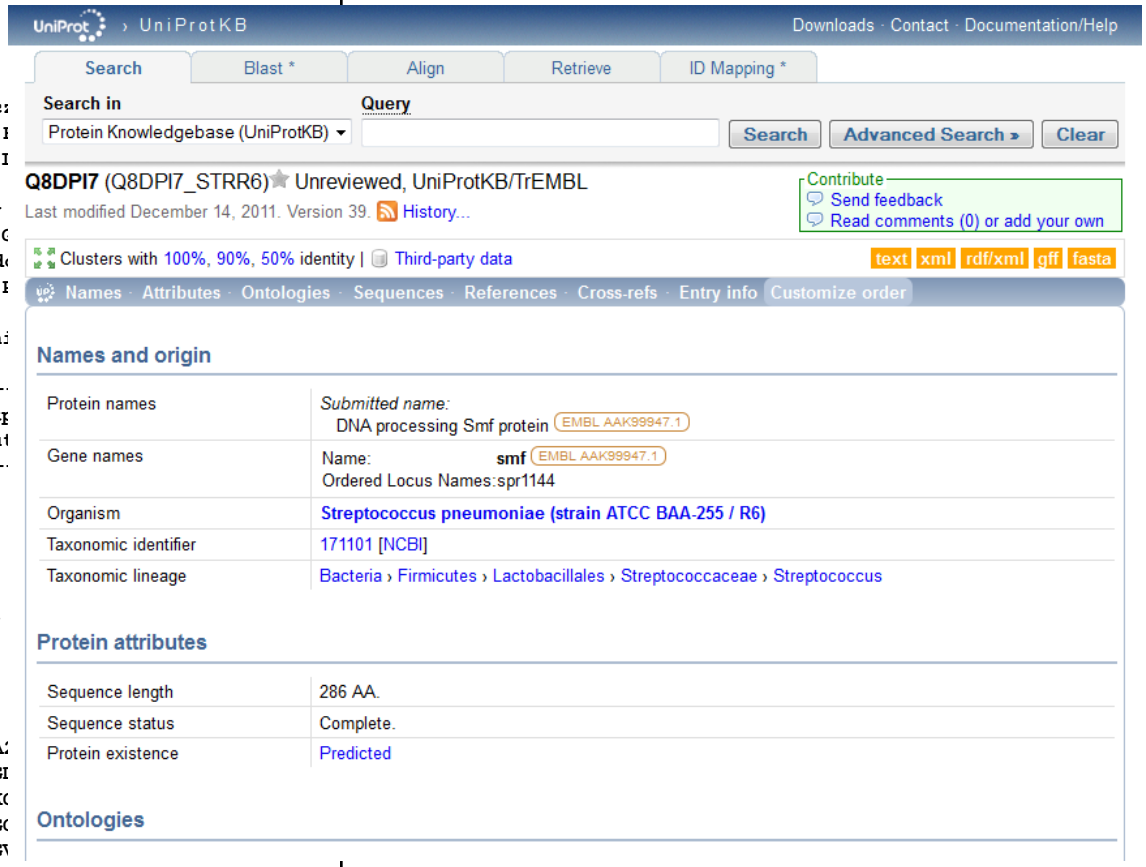
Distribution des unités de traduction et classification fonctionnelle



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 DNA].
RX   MEDLINE=21429245; PubMed=11544234;
RX   DOI=10.1128/JB.183.19.5709-5717.2001;
RA   Hoskins J., Alborn W.E. Jr., Arnold J., Blaszc
RA   DeHoff B.S., Estrem S.T., Fritz L., Fu D.-J.,
RA   Gilmour R., Glass J.S., Khoja H., Kraft A.R.,
RA   LeBlanc D.J., Lee L.N., Lefkowitz E.J., Lu J.,
RA   McAhren S.M., McHenney M., McLeaster K., Mundy
RA   Norris F.H., O'Gara M., Peery R.B., Robertson
RA   Sun P.-M., Winkler M.E., Yang Y., Young-Bellid
RA   Zook C.A., Baltz R.H., Jaskunas S.R., Rosteck
RA   Glass J.I.;
RT   "Genome of the bacterium Streptococcus pneumoni
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 transformation;
DR   InterPro; IPR003488; SMF.
DR   Pfam; PF02481; SMF; 1.
DR   TIGRFAMs; TIGR00732; dprA; 1.
KW   Complete proteome.
SQ   SEQUENCE   286 AA;  31583 MW;  CF12DB83AE3663A
     MELFMKITNY EIYKLKKSGL TNQQILKVLE YGENVDQELL LGI
     FQIDDAHLSK EFQKFPSFSI LDDCYPWDLS EIYDAPVLLF YKC
     CSKQGAKSVE KVIQGLENEL VIVSGLAKGI DTAAHMAALQ NGC
     NKRLQDYIGN DHLVLSEYGP GEQPLKFHFP ARNRIIAGLC RGV
     AMEEGRDVFA IPGSILDGLS DGCHHLIQEG AKLVTSGQDV LAEFEF
//
```

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



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

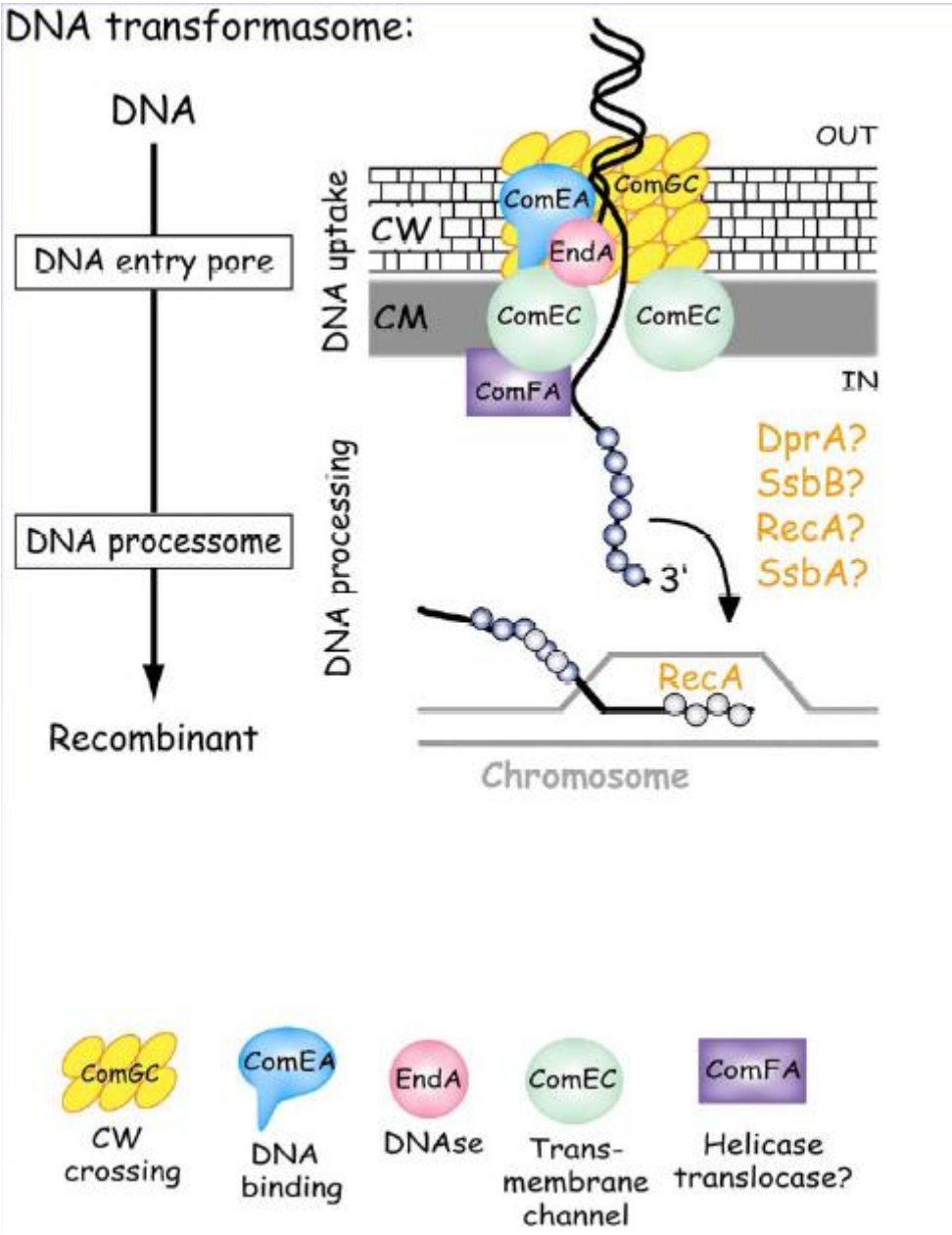
- > 40 espèces de bactéries, distribuées dans tous les groupes taxonomiques.

Rôles de la transformation

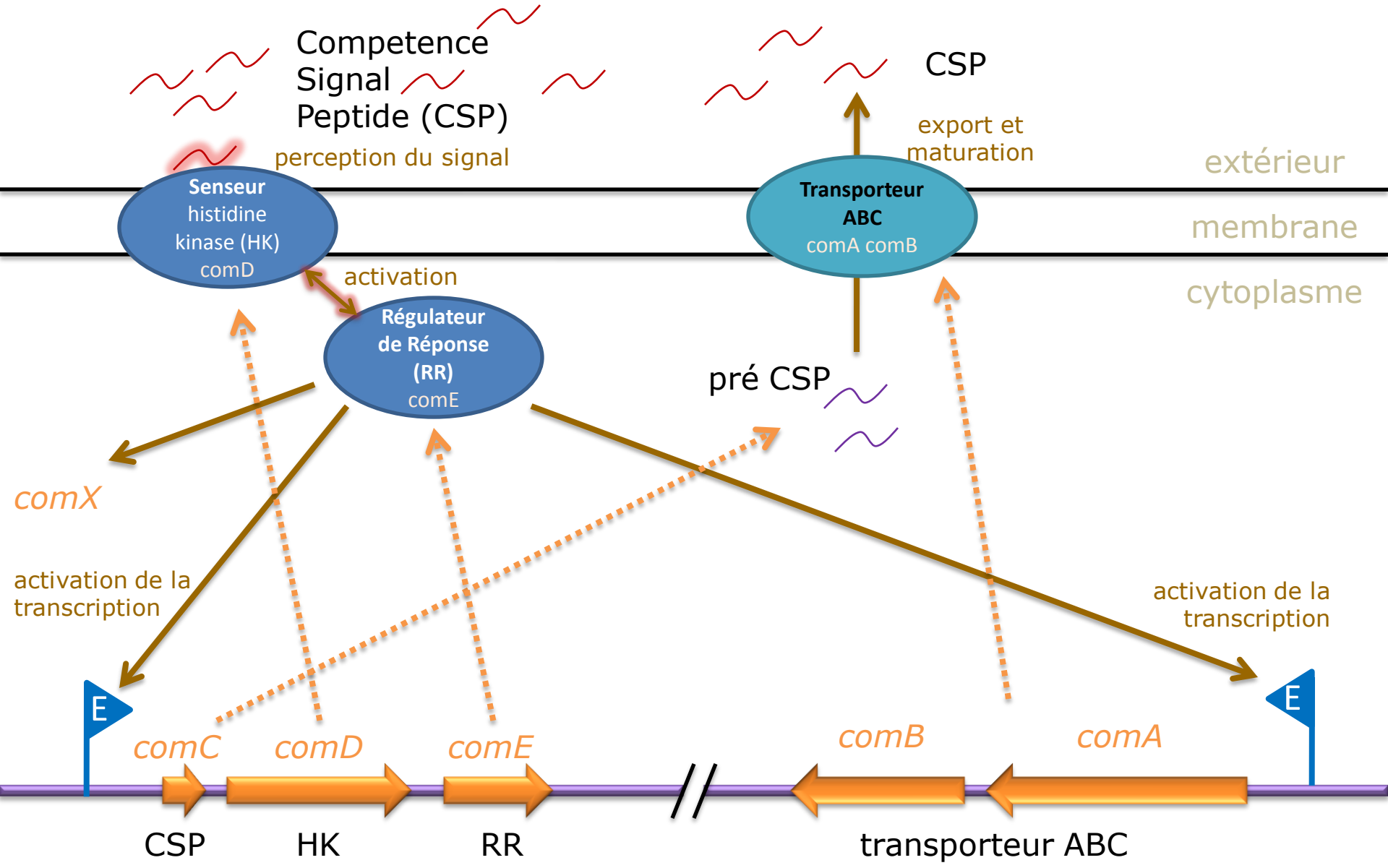
- échanges génétiques (sexualité bactérienne)
- réparation de l'ADN
- nutriments

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

Le transformasome

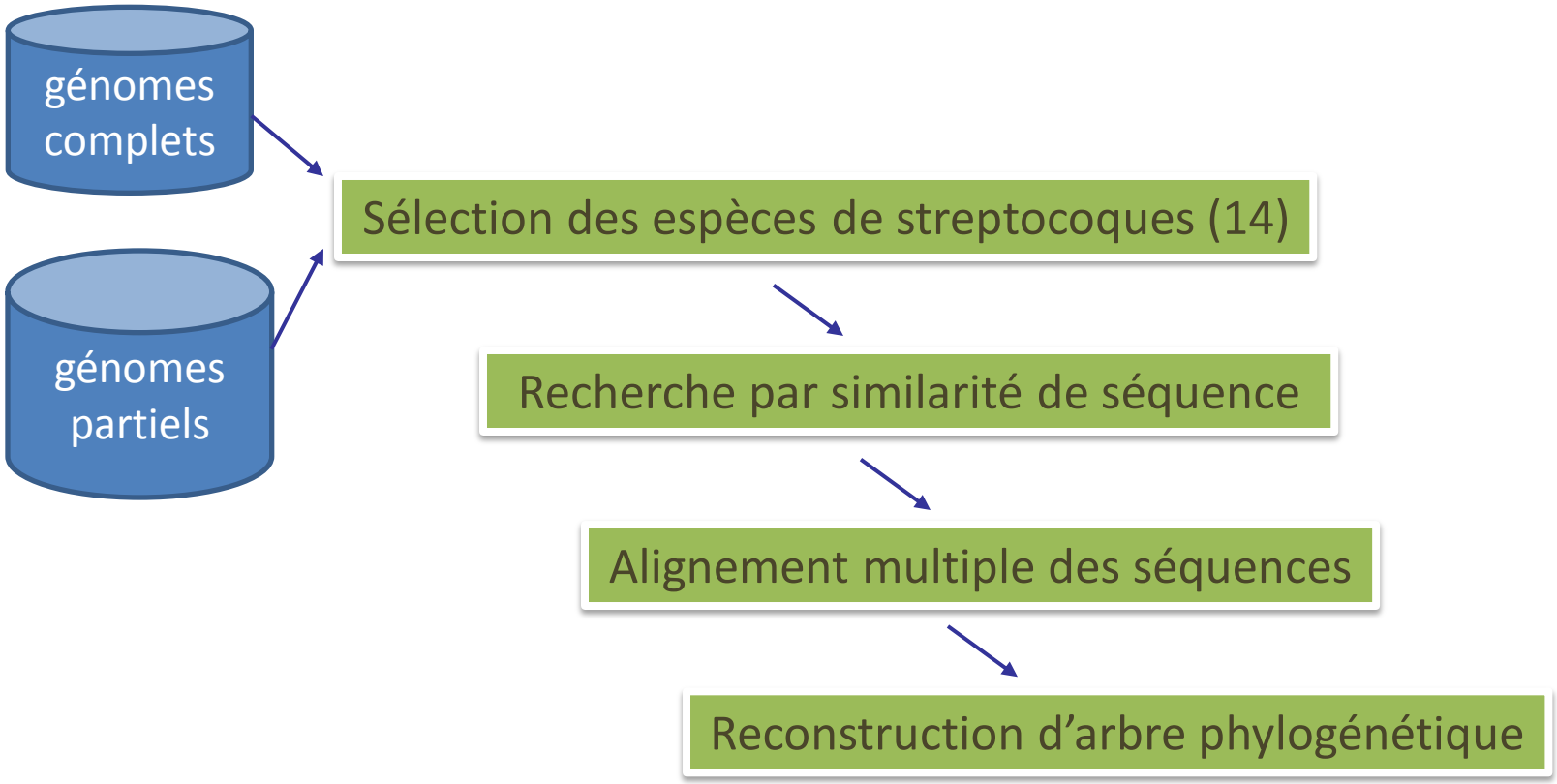


Régulation de l'état de compétence, modèle *Streptococcus pneumoniae*

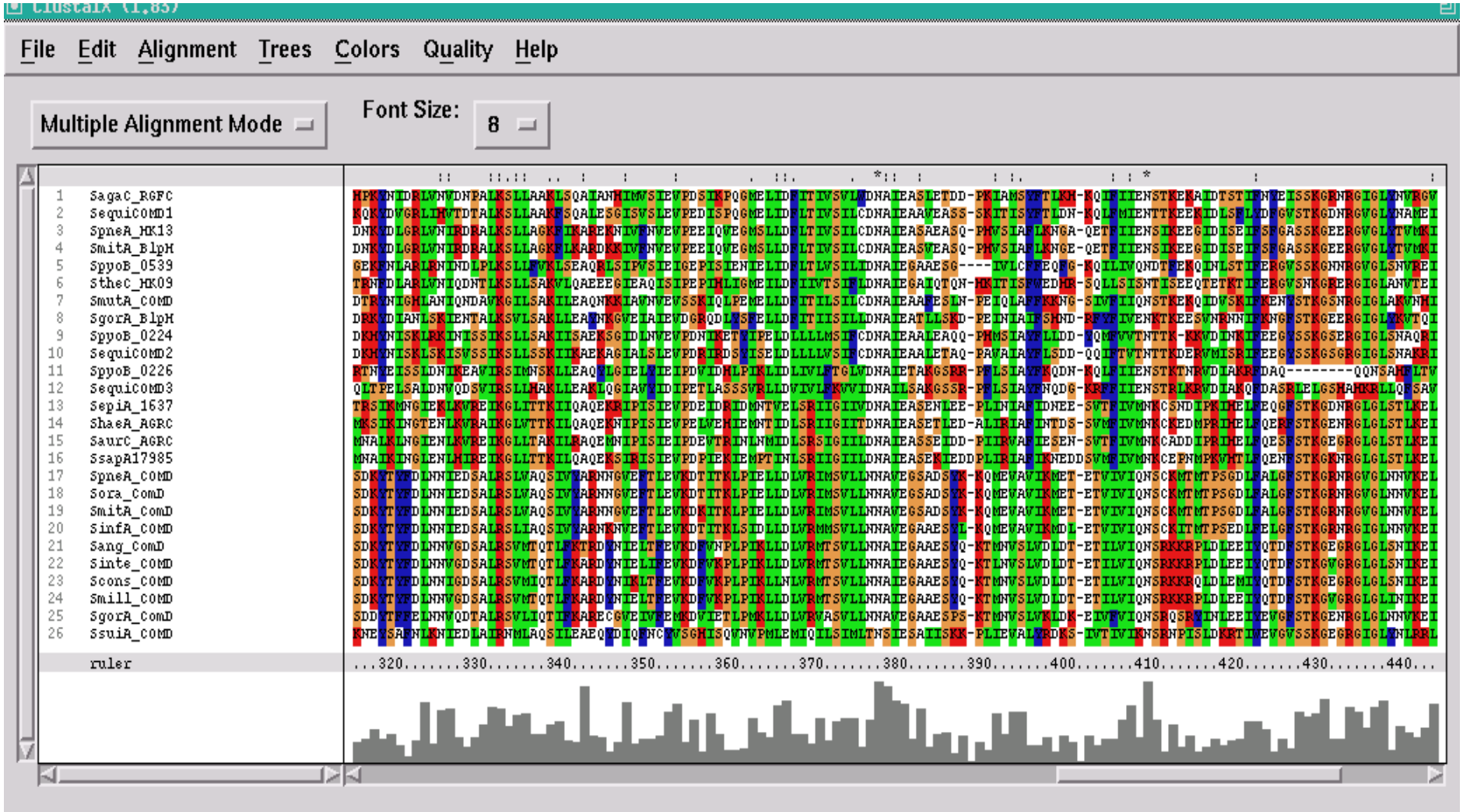


Distribution des gènes homologues à *comD* et *comE*

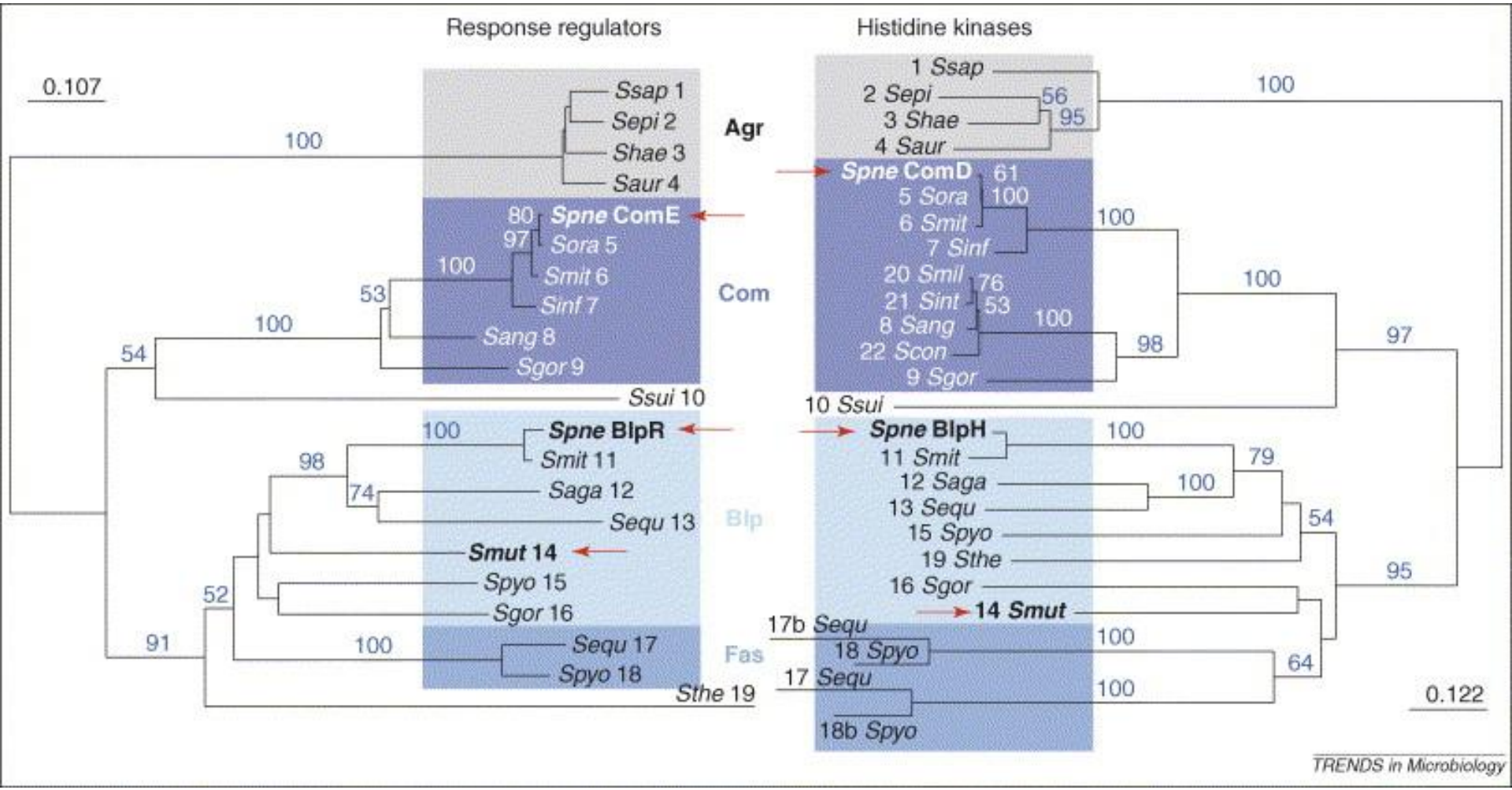
Séquences de référence : protéines ComD et ComE de *S. pneumoniae*



Alignement multiple des séquences des protéines homologues à ComD



Relation évolutive entre les gènes homologues à *comD* et *comE*



Transcriptome

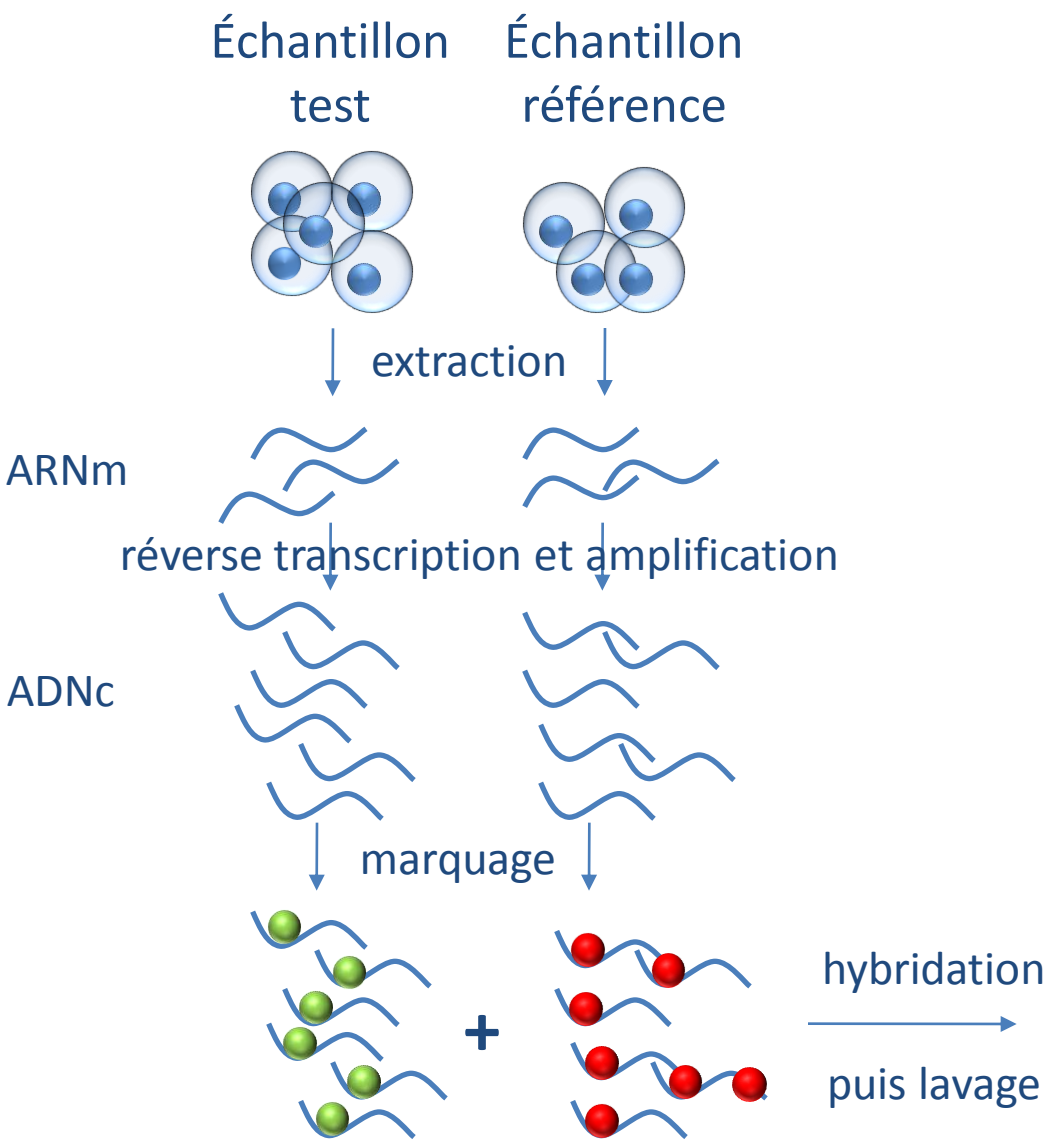
Transcriptome : ensemble des ARNm ou transcrits présents dans une population de cellules dans des conditions données.

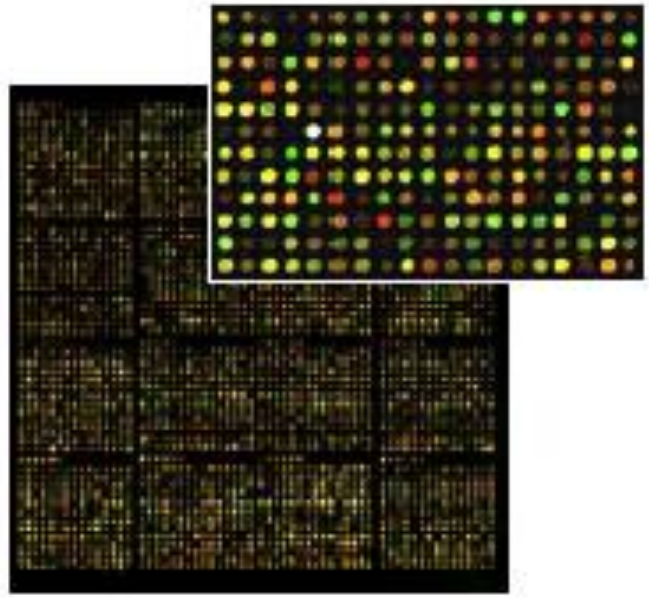
➡ Accès au niveau d'expression de milliers de gènes simultanément (potentiellement l'ensemble des gènes d'un organisme)
= *instantané* de l'état d'une cellule ou d'une population de cellules

Données d'expression des gènes obtenues par :

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

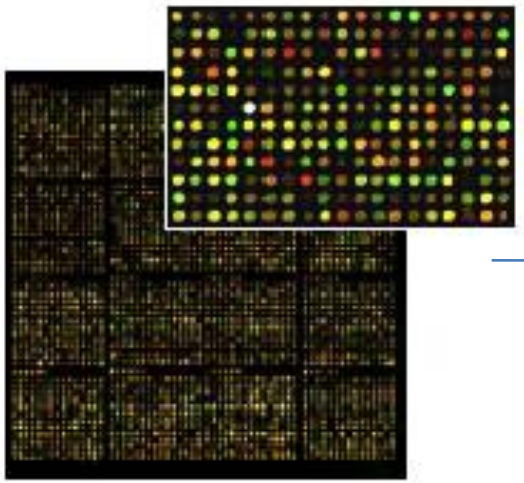
Transcriptome : acquisition des données





scan

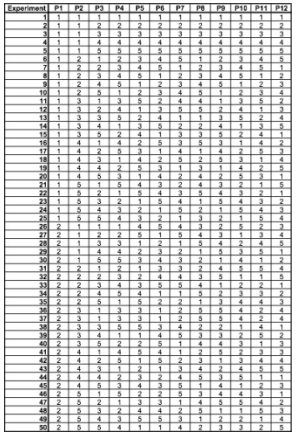
Transcriptome : analyse et interprétation des données



acquisition

mesures

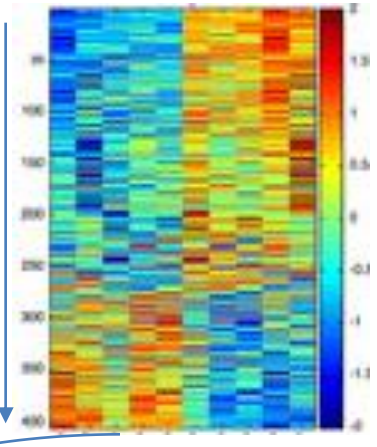
probesets



normalisation
filtrage

conditions

gènes



interprétation

Identification des **gènes différentiellement exprimés**

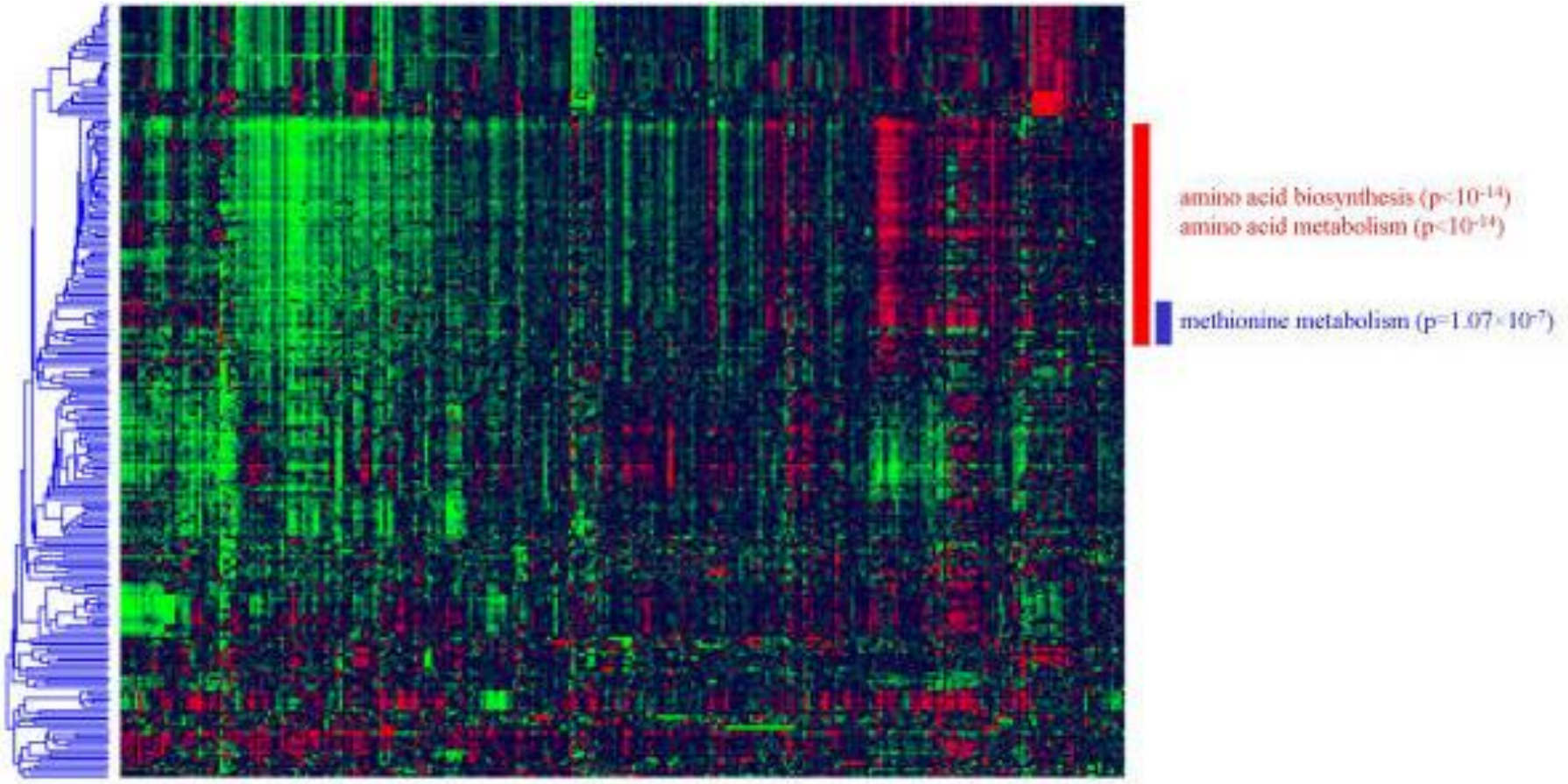
Identification des ensembles de **gènes co-exprimés**

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



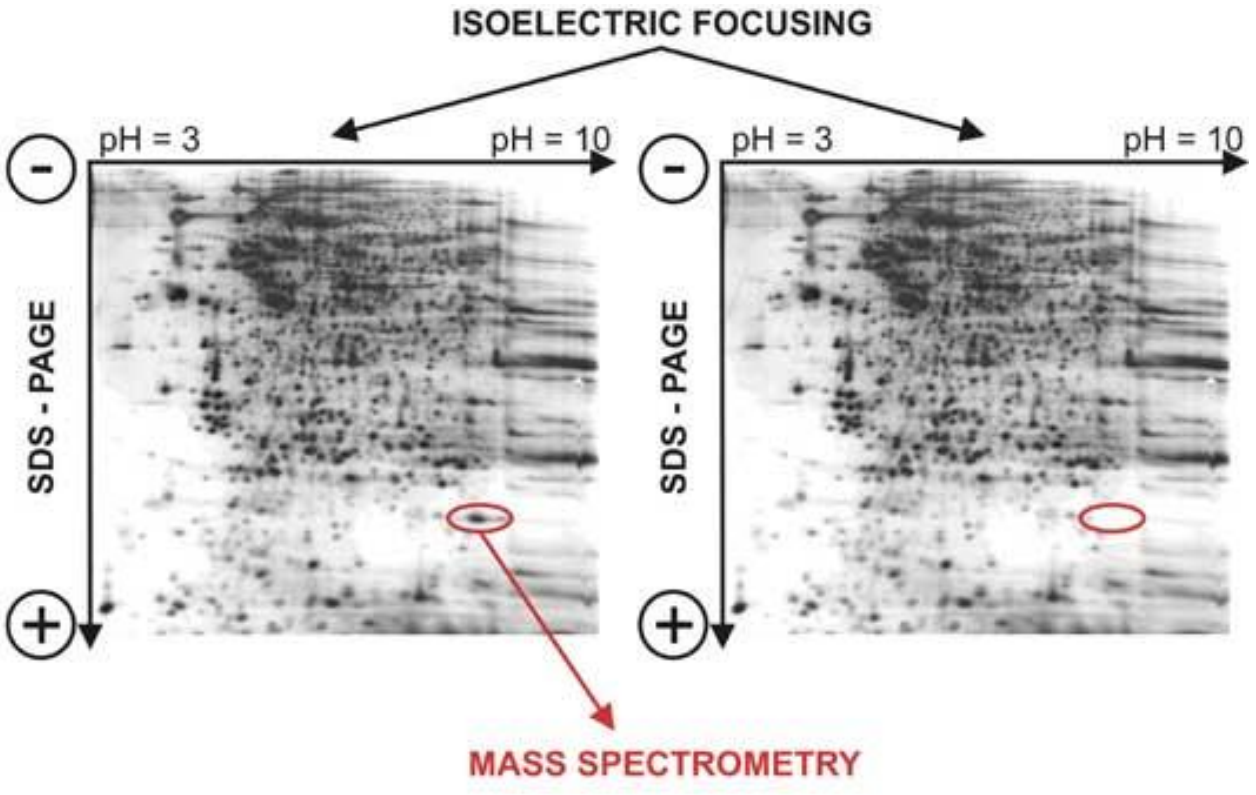
Robinson *et al.* 2002

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

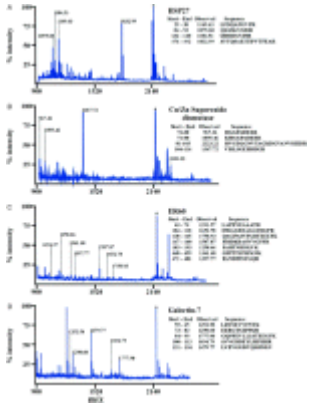


Échantillon référence

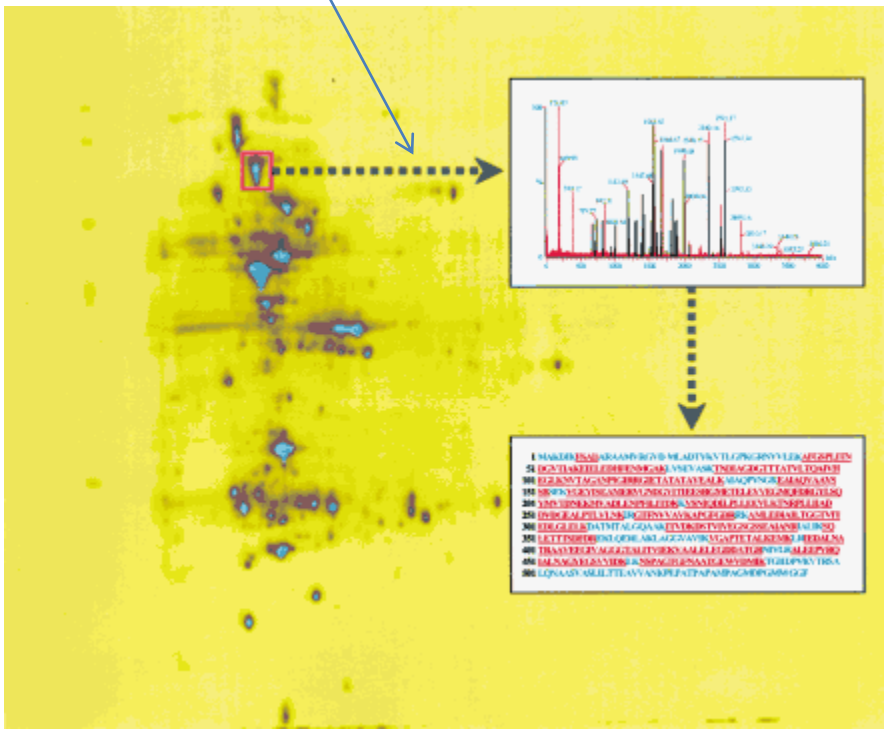
Protéomique : Idenitification de protéine

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

Digestion *in silico* du protéome



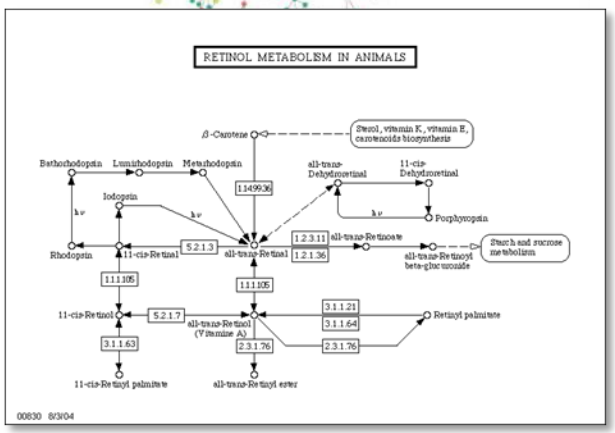
Recherche des protéines correspondant au profil observé

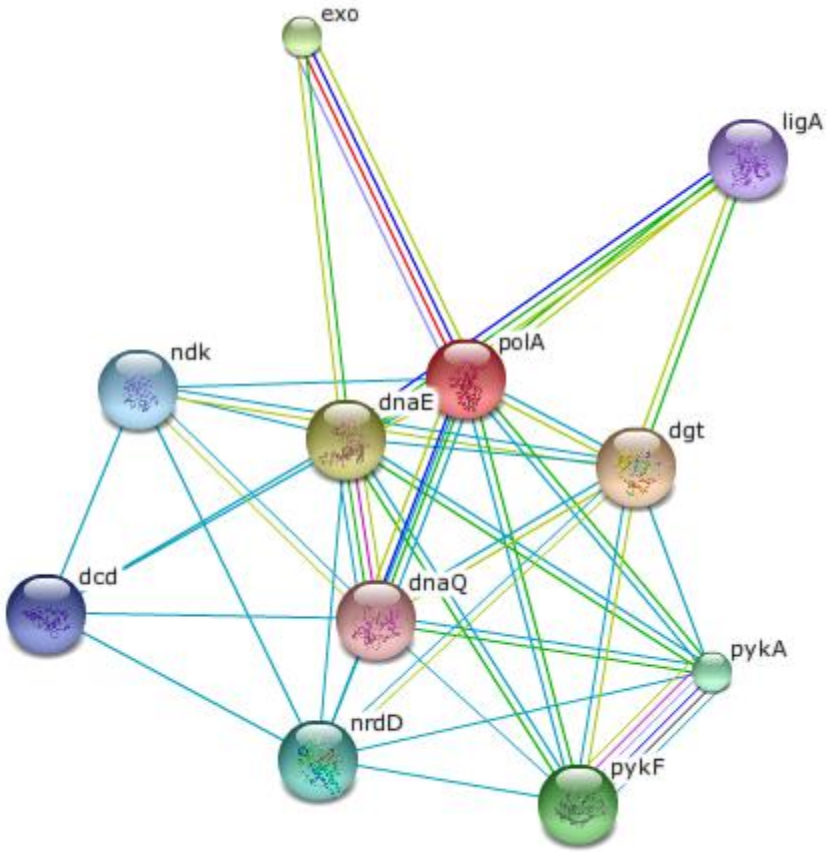


Réseaux de gènes et de protéines

Réseaux :

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



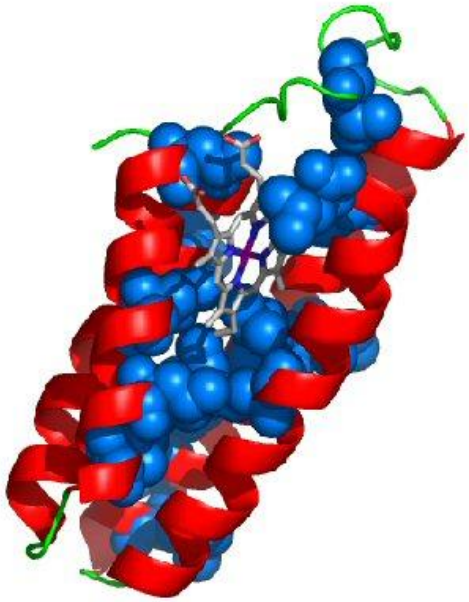


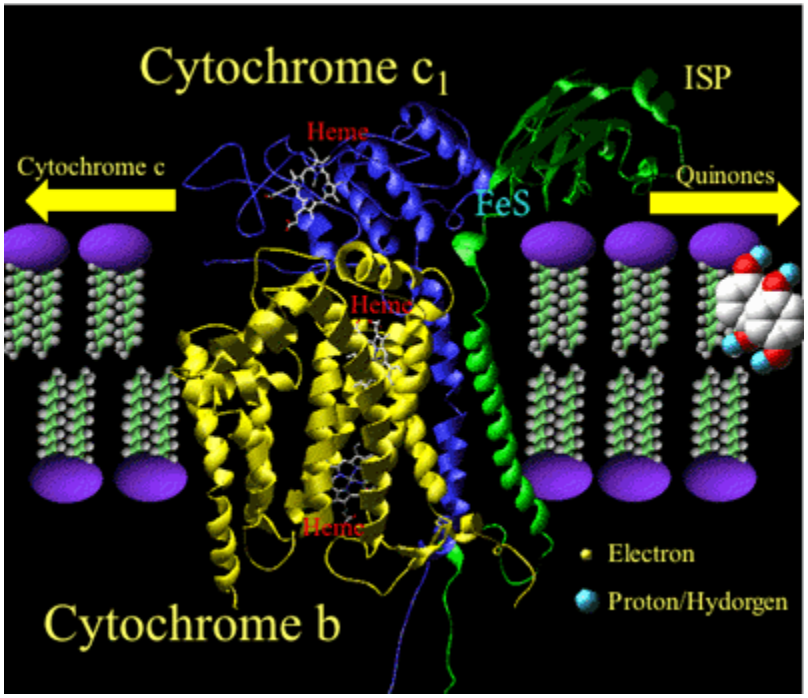
Prédiction de structure

Séquence protéique

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

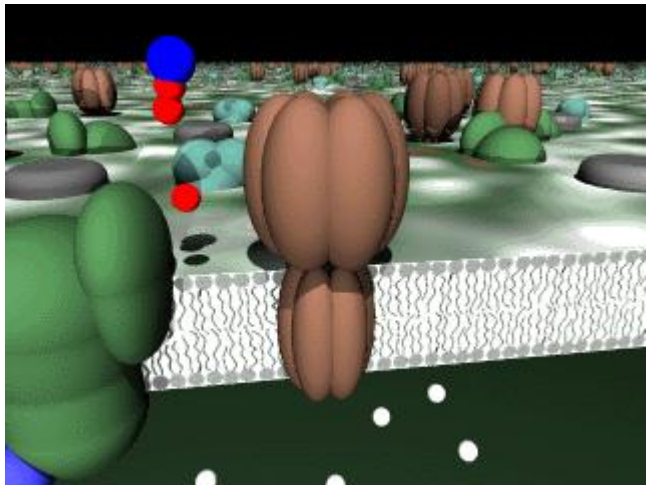
Prédiction ou résolution de la structure tridimensionnelle





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