Introduction

<u>La bioinformatique</u> : 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.

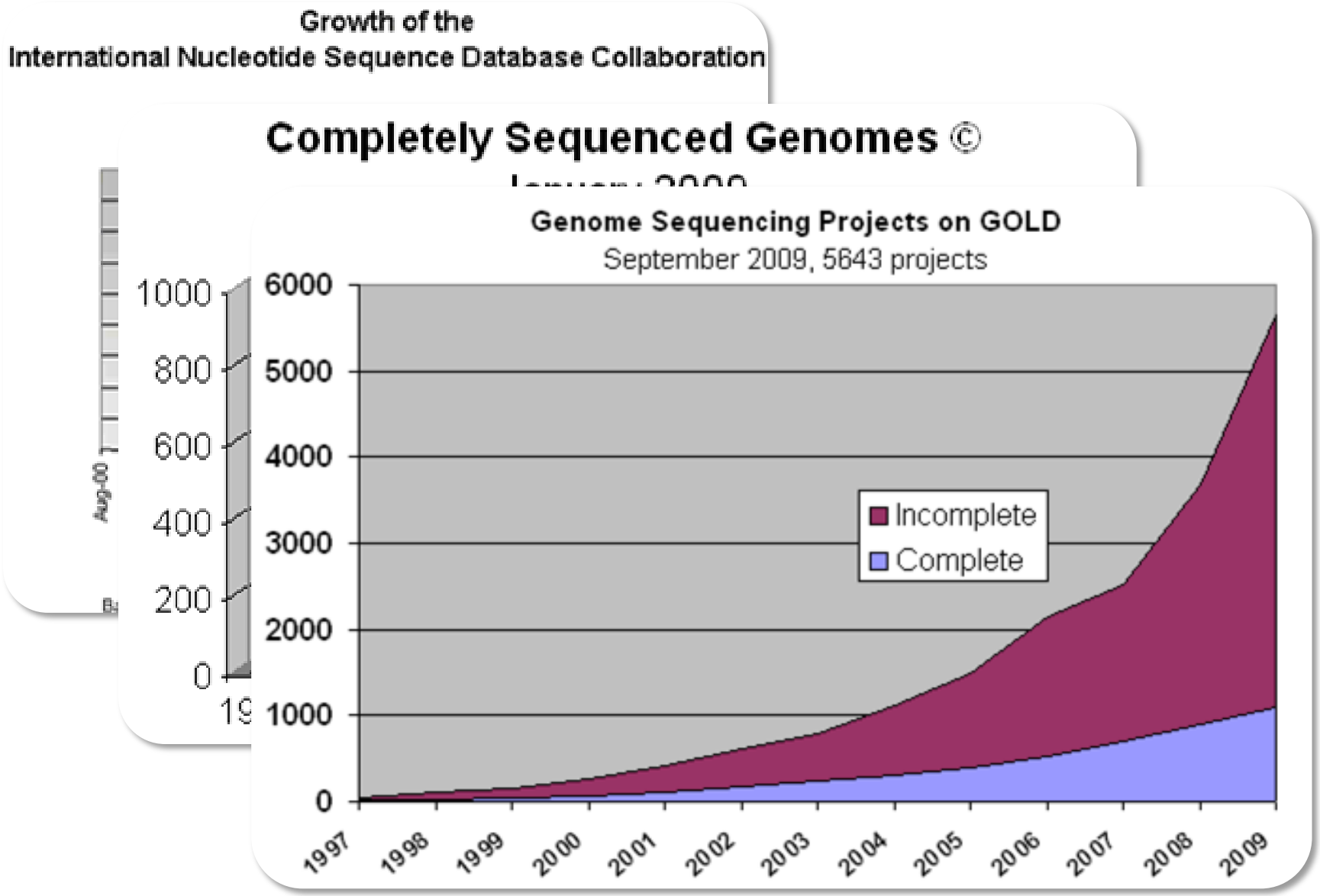
Historique rapide de la bioinformatique

- 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 : 3102 génomes complets (150 archaea, 2784 bactéries, 168 eucaryotes), 7741 projets de séquençage en cours (179 archaea, 5516 bactéries, 2046 eucaryotes). 340 études de métagénomes (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



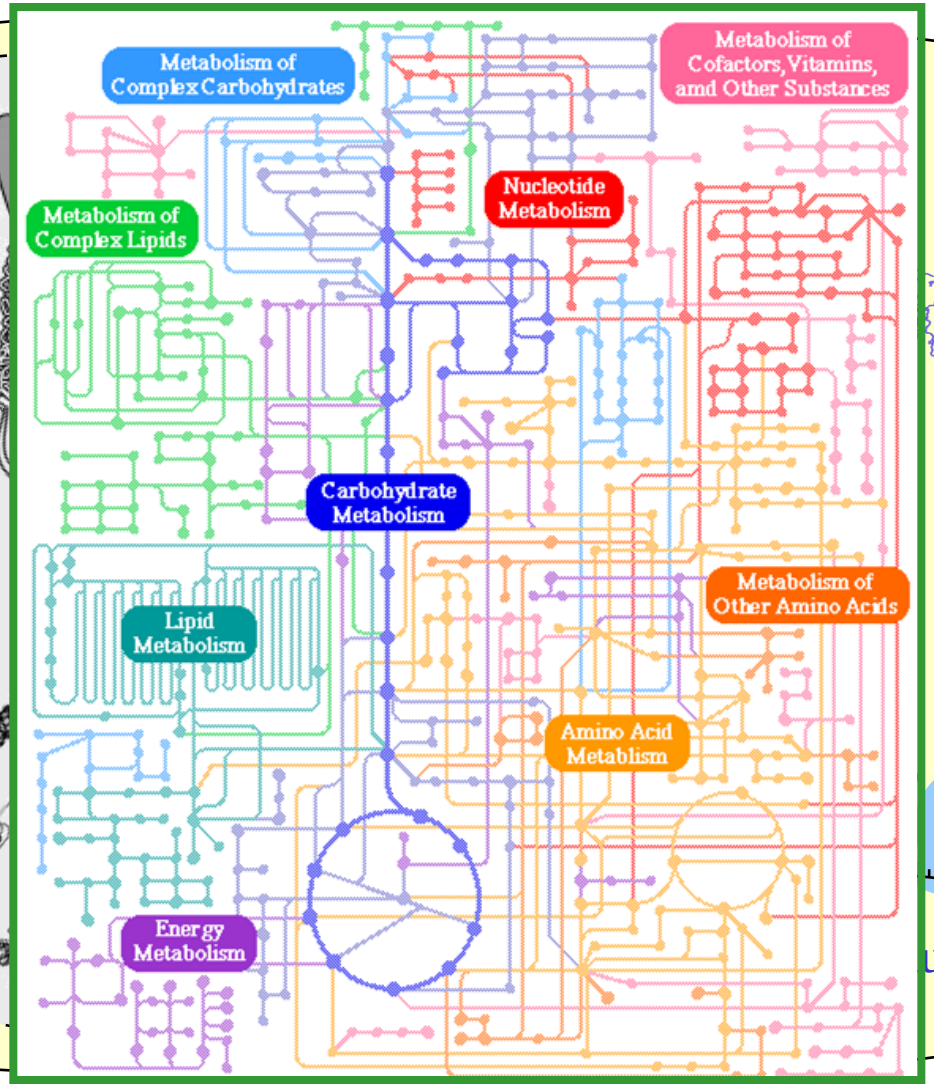
(Quelques) données et connaissances disponibles

métabolisme

génome

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

ADN



interactome

protéome

protéine

```
MTGLAEEWWFDSGRAALGAD
ARDWKKFTQALESRFTPSTY
AMNMEEDWSNLQCGPNESSY
SYETRFRAARHHLSREA...
```

uction

métabolome

Cellule eucaryote

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

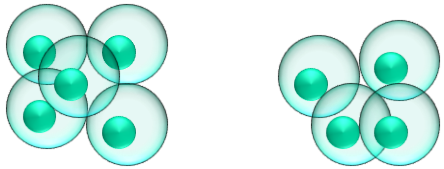
Données d'expression des gènes obtenues par :

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

Acquisition des données

Échantillon test

Échantillon référence



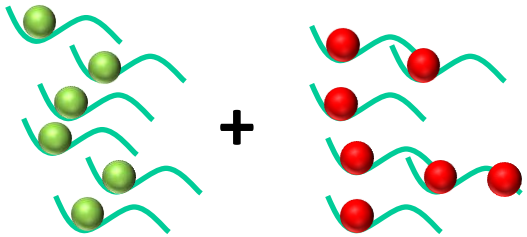
extraction

ARNm

réverse transcription et amplification

ADNc

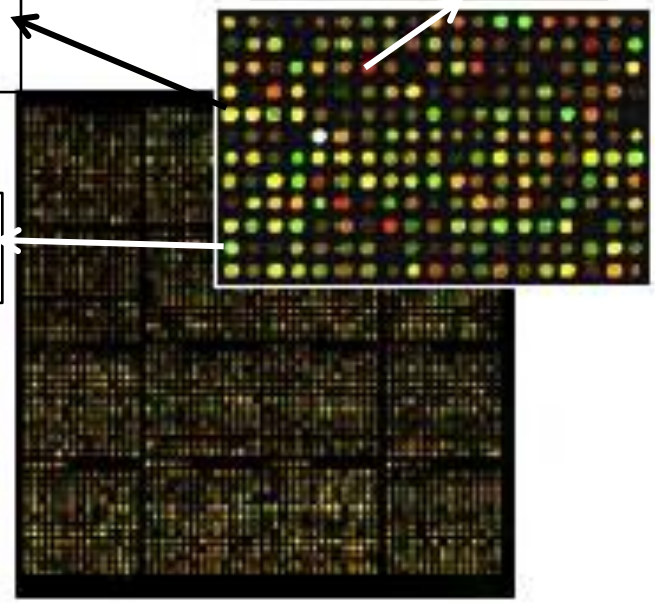
marquage



hybridation

puis lavage

scan

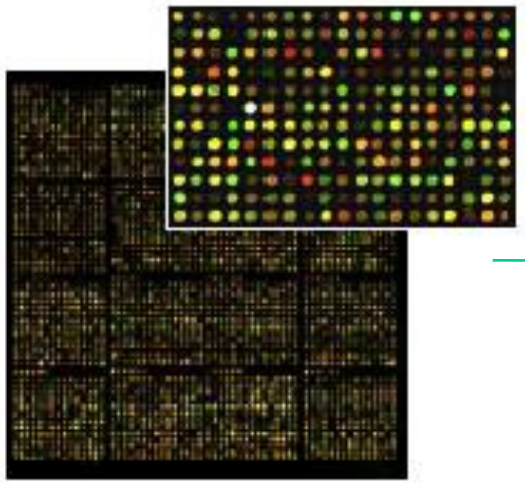


Jaune : signal similaire des deux échantillons

Rouge : signal fort échantillon de référence

Vert: signal fort échantillon test

Analyse et interprétation des données



acquisition

mesures

probesets

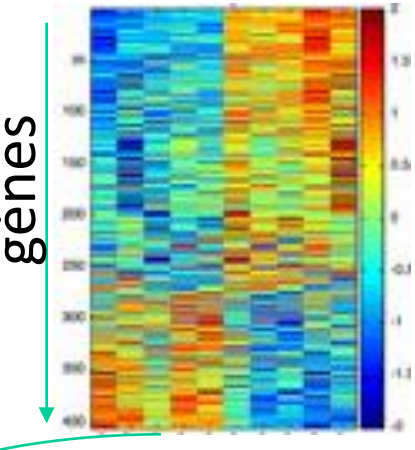
Experiment	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	2	2	2	2	2	2	2	2	2	2
3	1	1	3	3	3	3	3	3	3	3	3	3
4	1	1	4	4	4	4	4	4	4	4	4	4
5	1	1	5	5	5	5	5	5	5	5	5	5
6	1	2	1	2	3	4	5	1	2	3	4	5
7	1	2	2	3	4	5	1	2	3	4	5	1
8	1	2	3	4	5	1	2	3	4	5	1	2
9	1	2	4	5	1	2	3	4	5	1	2	3
10	1	2	5	1	2	3	4	5	1	2	3	4
11	1	3	1	3	5	2	4	4	2	5	5	2
12	1	3	2	4	1	3	5	5	2	4	1	3
13	1	3	3	5	2	4	1	1	3	5	2	4
14	1	3	4	1	3	5	2	2	4	1	3	5
15	1	3	5	2	4	1	3	3	5	2	4	1
16	1	4	1	4	2	5	3	5	3	1	4	2
17	1	4	2	5	3	1	4	1	4	2	5	3
18	1	4	3	1	4	2	5	2	5	3	1	4
19	1	4	4	2	5	3	1	3	1	4	2	5
20	1	4	5	3	1	4	2	4	2	5	3	1
21	1	5	1	5	4	3	2	4	3	2	5	1
22	1	5	2	1	5	4	3	5	4	3	2	1
23	1	5	3	2	1	5	4	1	5	4	3	2
24	1	5	4	3	2	1	5	2	1	5	4	3
25	1	5	5	4	3	2	1	3	2	1	5	4
26	2	1	1	1	4	5	4	3	2	5	2	3
27	2	1	2	2	5	1	5	4	3	1	3	4
28	2	1	3	3	1	2	1	5	4	2	4	5
29	2	1	4	4	2	3	2	1	5	3	5	1
30	2	1	5	5	3	4	3	2	1	4	1	2
31	2	2	1	2	1	3	3	2	4	5	5	4
32	2	2	2	3	2	4	4	3	5	1	1	5
33	2	2	3	4	3	5	5	4	1	2	2	1
34	2	2	4	5	4	1	1	5	2	3	3	2
35	2	2	5	1	5	2	2	1	3	4	4	3
36	2	3	1	3	3	1	2	2	5	4	2	4
37	2	3	1	3	3	1	2	5	5	4	2	4
38	2	3	3	5	1	3	4	2	2	1	4	1
39	2	3	4	1	1	4	5	3	3	2	5	2
40	2	3	5	2	2	5	1	4	4	3	1	3
41	2	4	1	4	5	4	1	2	2	3	3	3
42	2	4	2	5	1	5	2	3	1	3	4	4
43	2	4	3	1	2	1	3	4	2	4	5	5
44	2	4	4	2	3	2	4	5	3	5	1	1
45	2	4	5	3	4	3	5	1	4	1	2	3
46	2	5	1	5	2	4	3	5	3	4	3	1
47	2	5	2	1	3	3	1	4	5	5	4	2
48	2	5	3	2	4	4	2	5	1	1	5	3
49	2	5	4	3	5	5	3	1	2	2	1	4
50	2	5	5	4	1	1	4	2	3	3	2	5

normalisation

filtrage

condition

gènes



interprétation

Identification des **gènes différentiellement exprimés**

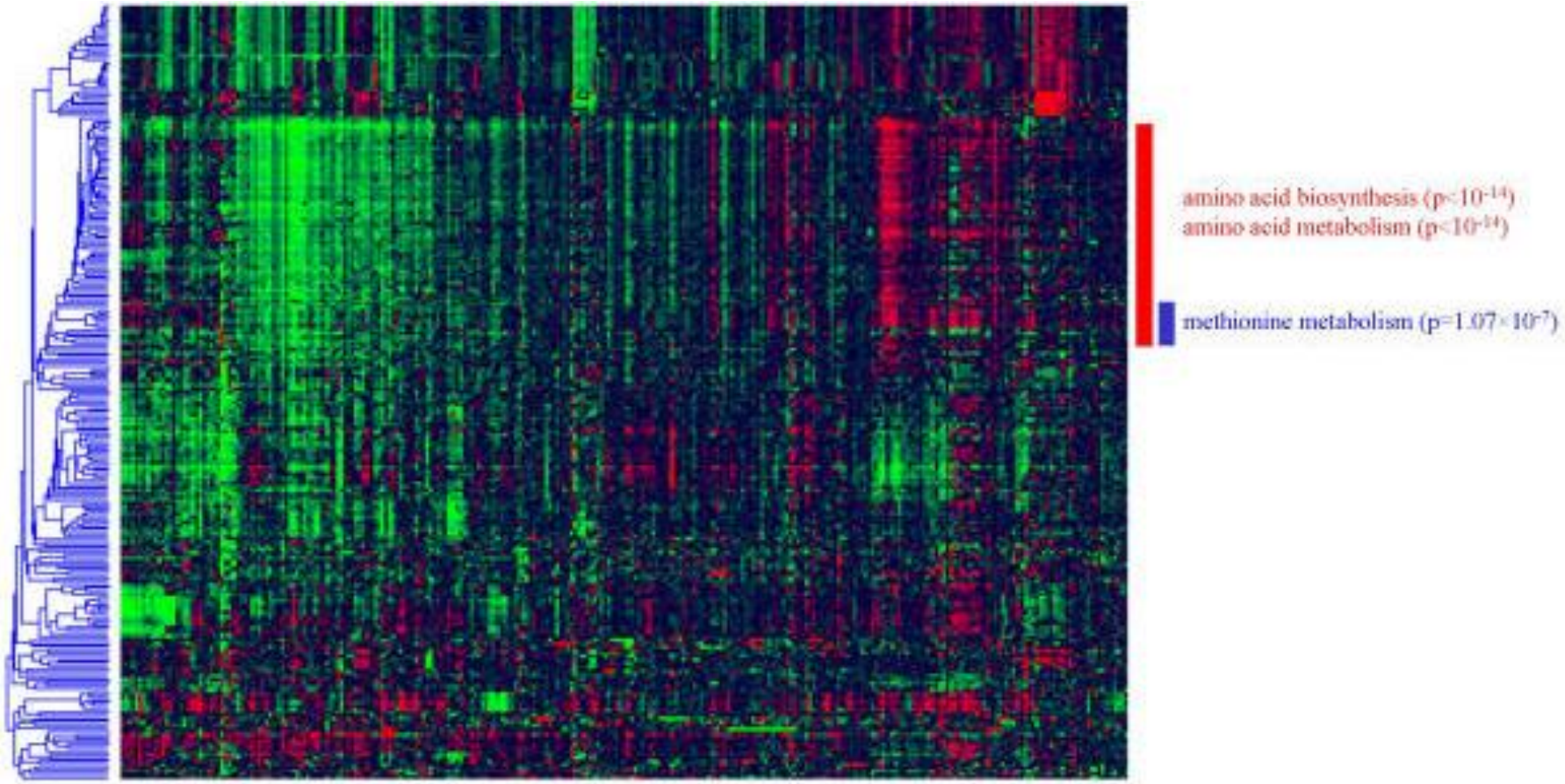
Identification des ensembles de **gènes co-exprimés**

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**



Robinson *et al.* 2002

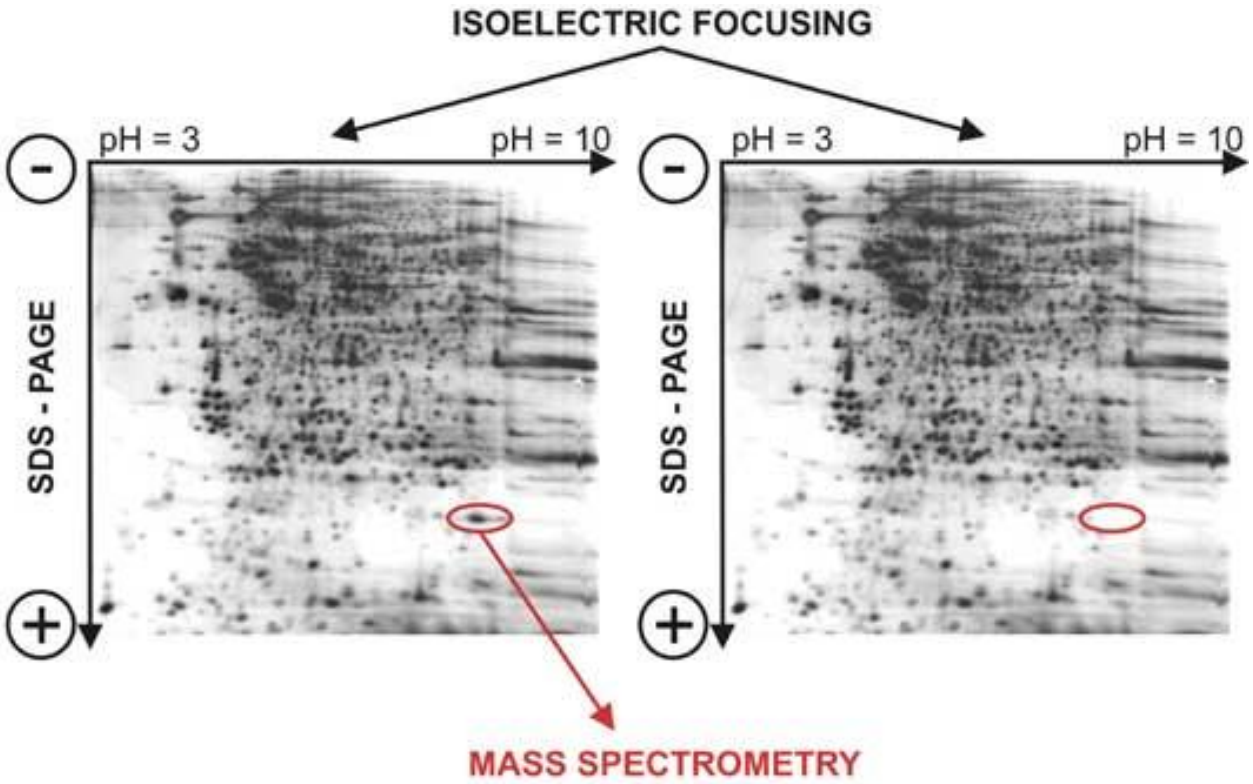
Protéomique

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

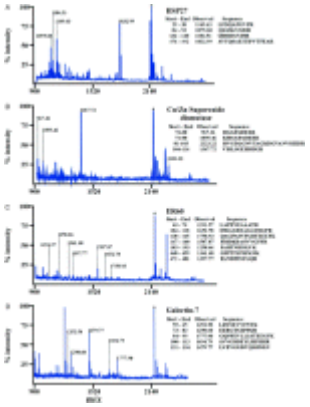


Échantillon référence

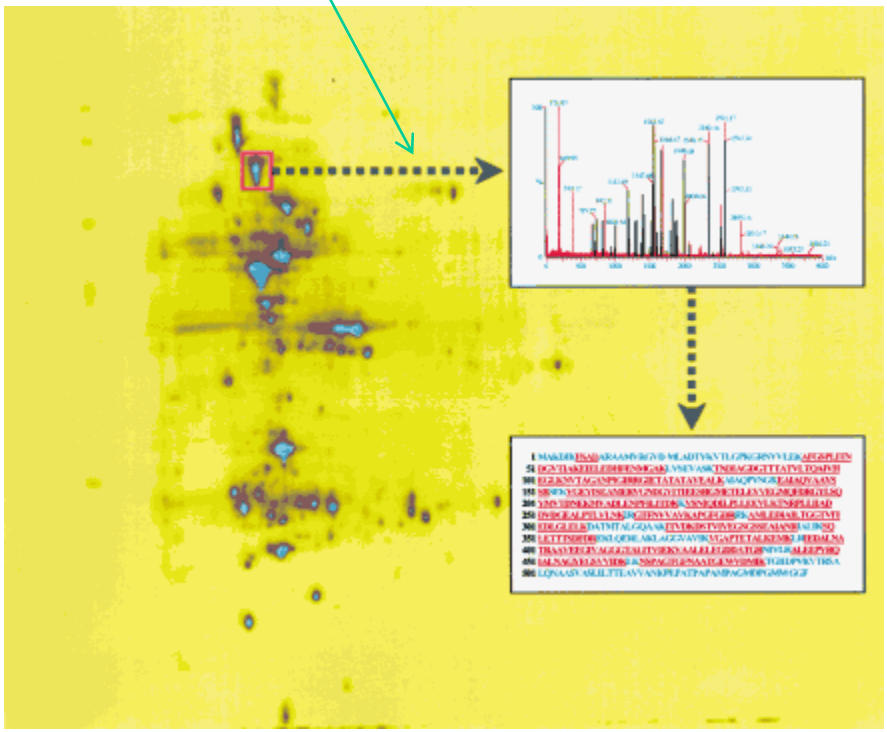
Identification des protéines

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

Digestion *in silico* du protéome



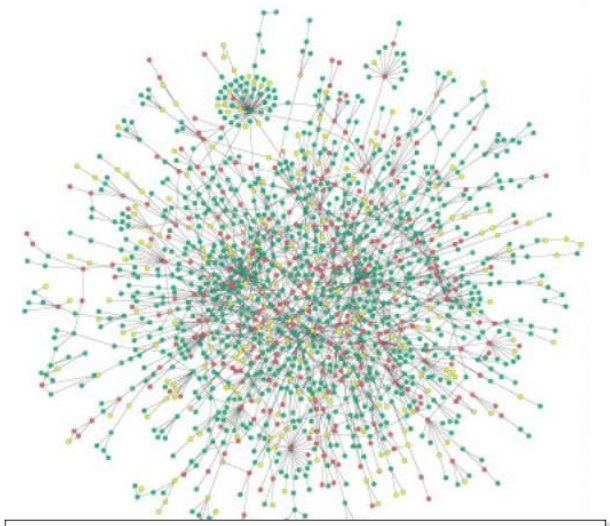
Recherche des protéines correspondant au profil observé

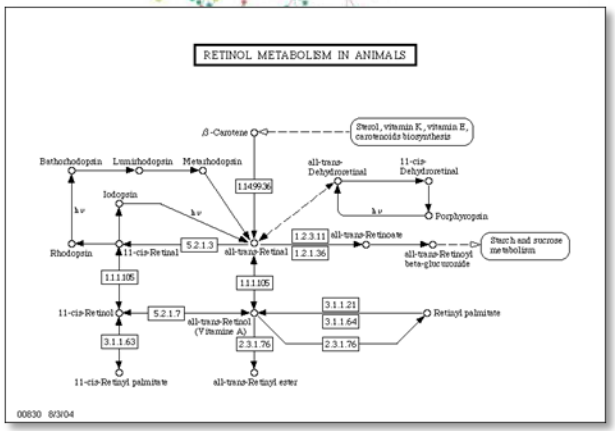


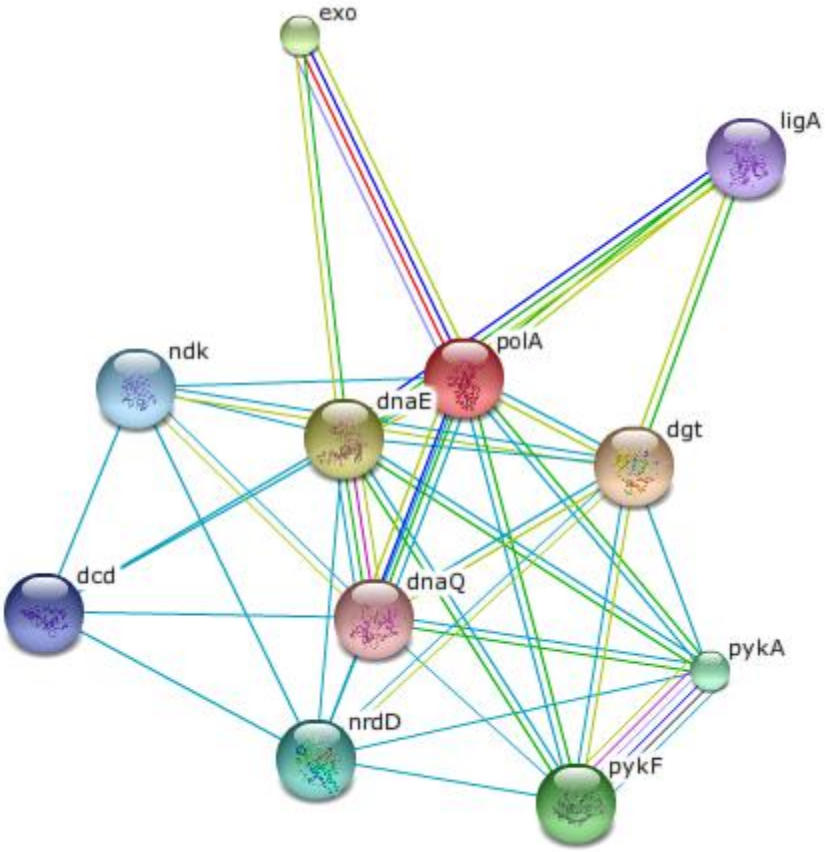
Réseaux de gènes, de protéines

Réseaux :

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





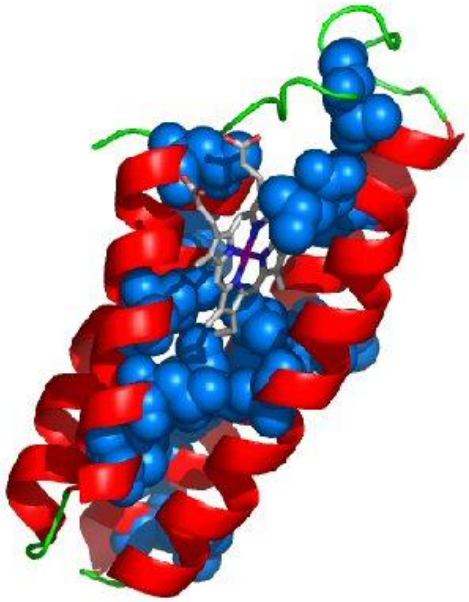


Biologie structurale

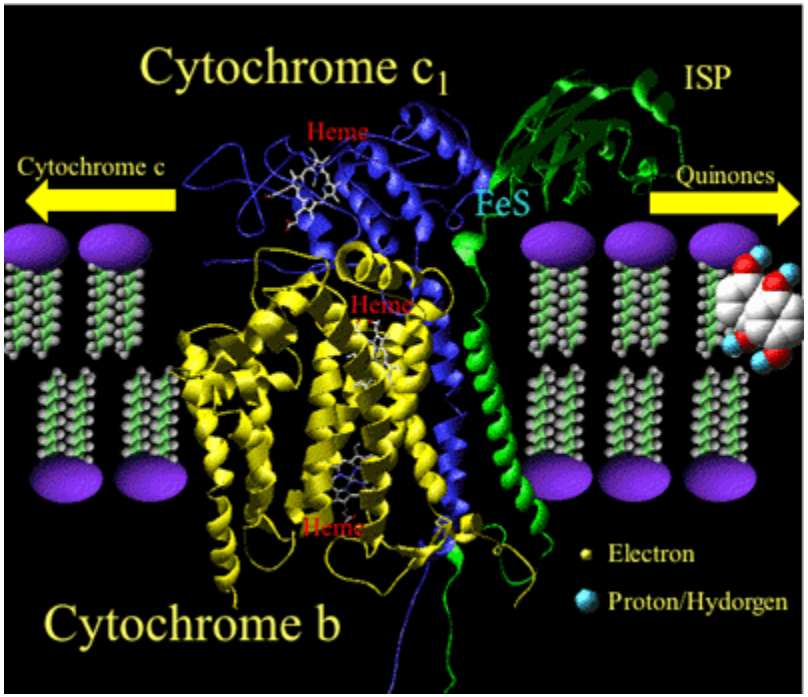
Séquence protéique

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

Prédiction ou résolution
de la structure tridimensionnelle



Prédiction des interactions
protéine – protéine ou
protéine - ligand



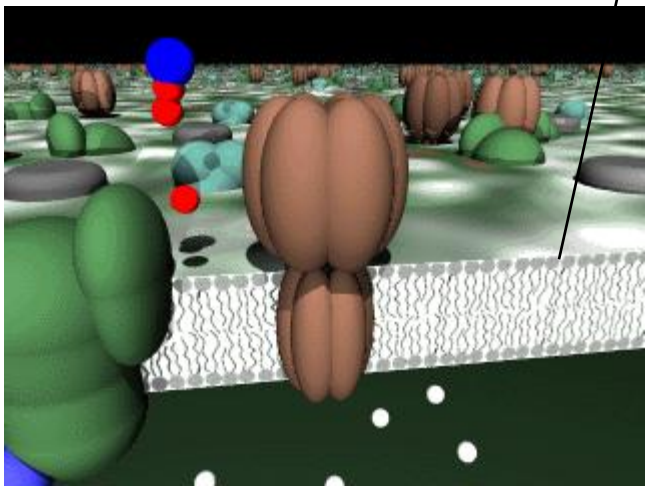
Biologie des systèmes

Intégration et synthèse des connaissances

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

Membrane mitochodriale

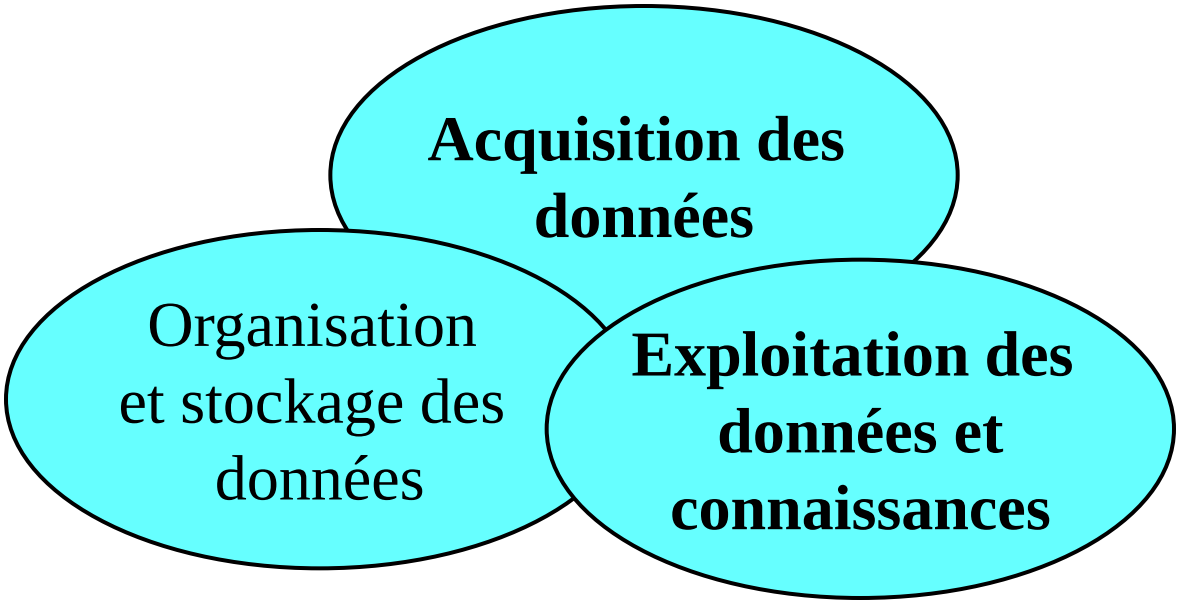
Exemple : F1-F0 ATP synthase



L'énergie proton-motrice est utilisée pour ajouter un groupe phosphate minéral à une molécule d'adénosine diphosphate (ADP) pour former une molécule d'adénosine triphosphate (ATP)

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

Trois grands domaines où intervient la bioinformatique



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
- prédictions de structures secondaires

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).

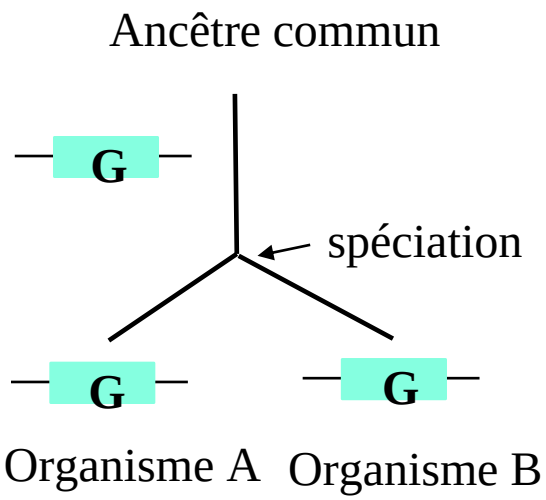
• Deux gènes sont **paralogues** si leur divergence est due à la duplication du gène ancêtre.

Donc deux séquences sont ou ne sont pas homologues.

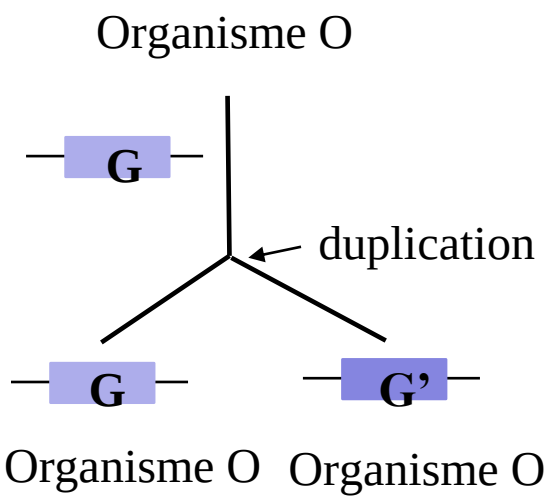
Dire que la protéine X a 80% d'homologie avec la protéine Y est donc incorrect:
soit:

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

Homologie - orthologie- paralogie



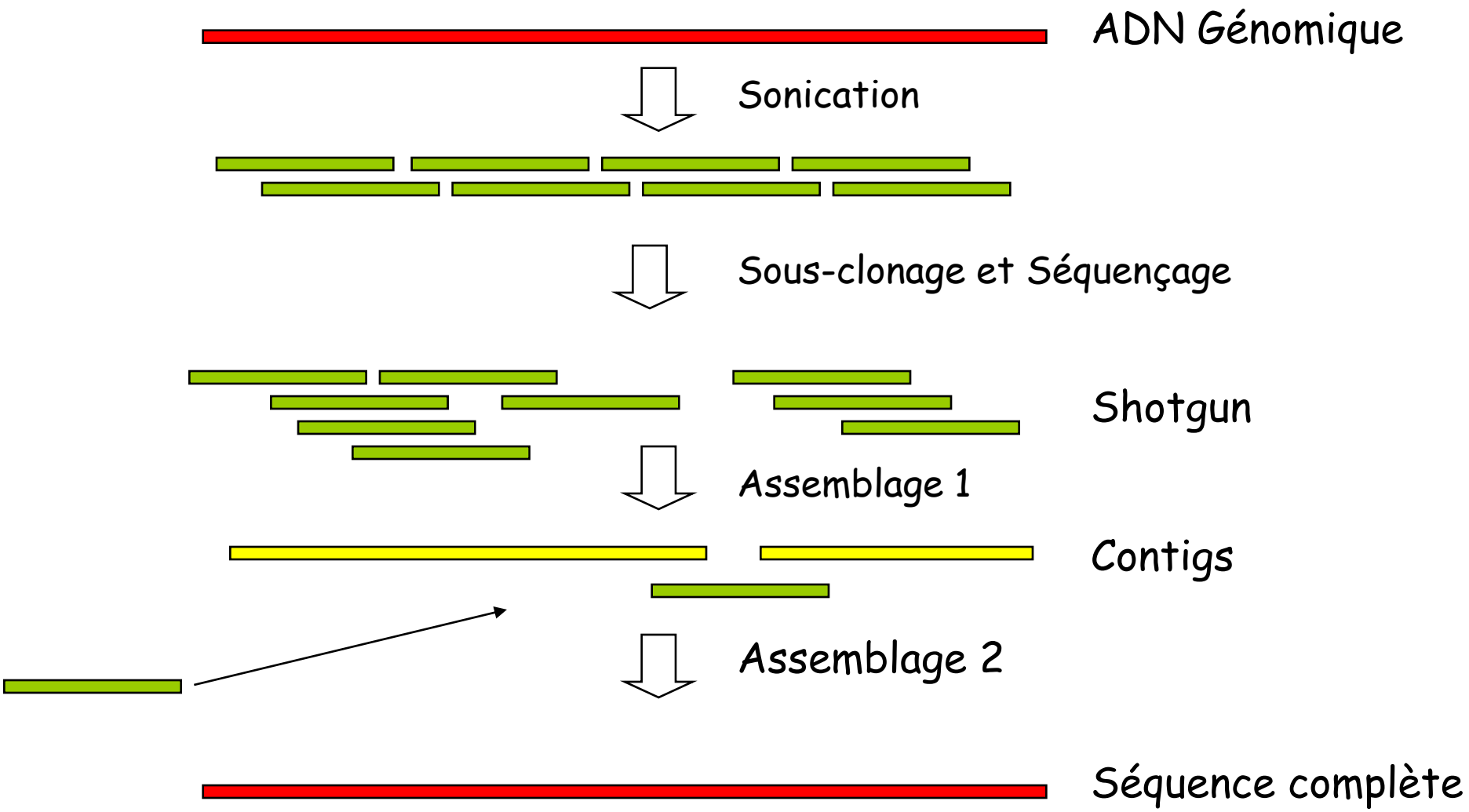
Génes orthologues



Génes paralogues

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

Post-traitement du séquençage



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 :

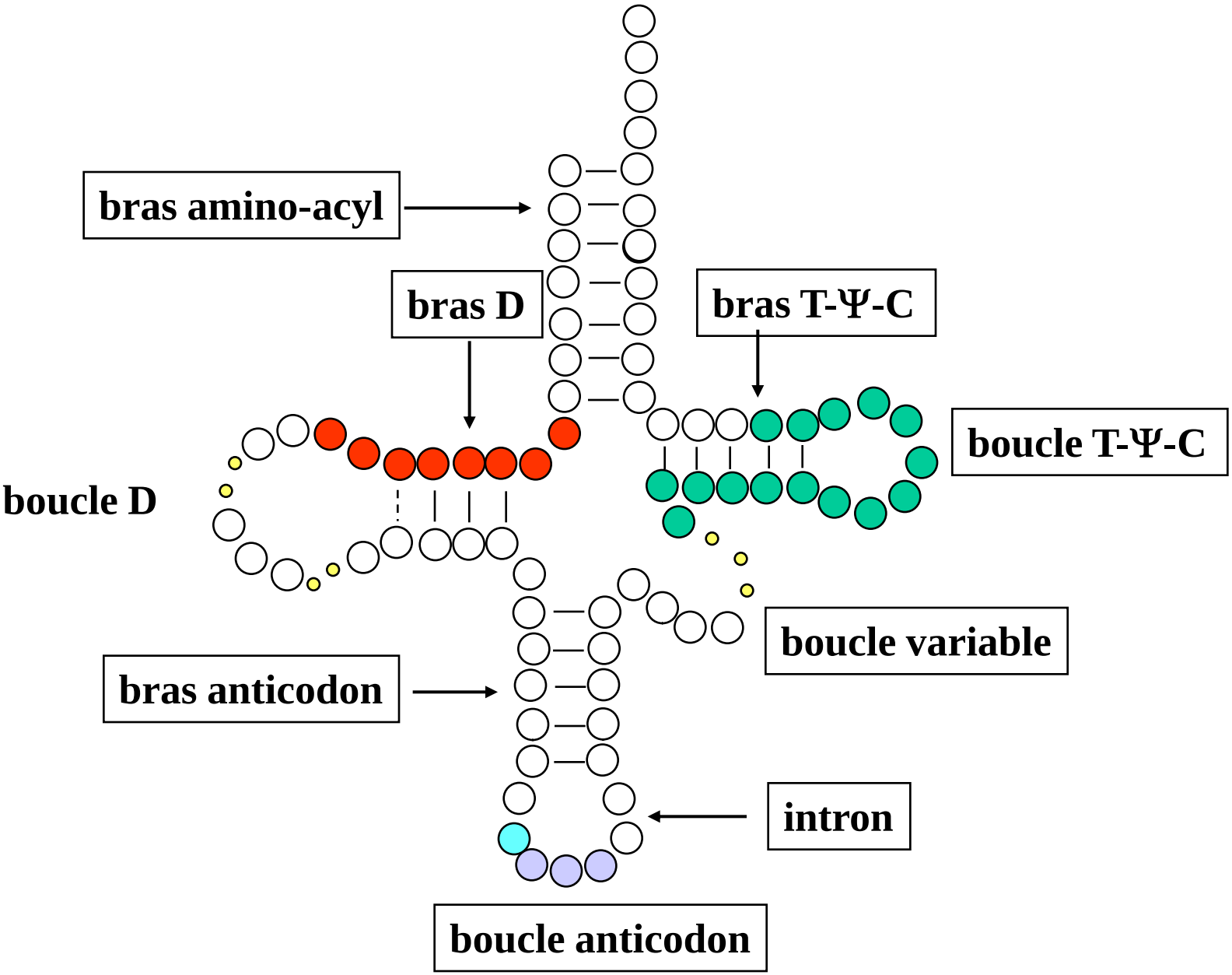
- les ARNr
- les ARNt
- les protéines

Identification des unités de traduction

Identification des unités de transcription (promoteur et terminateur)

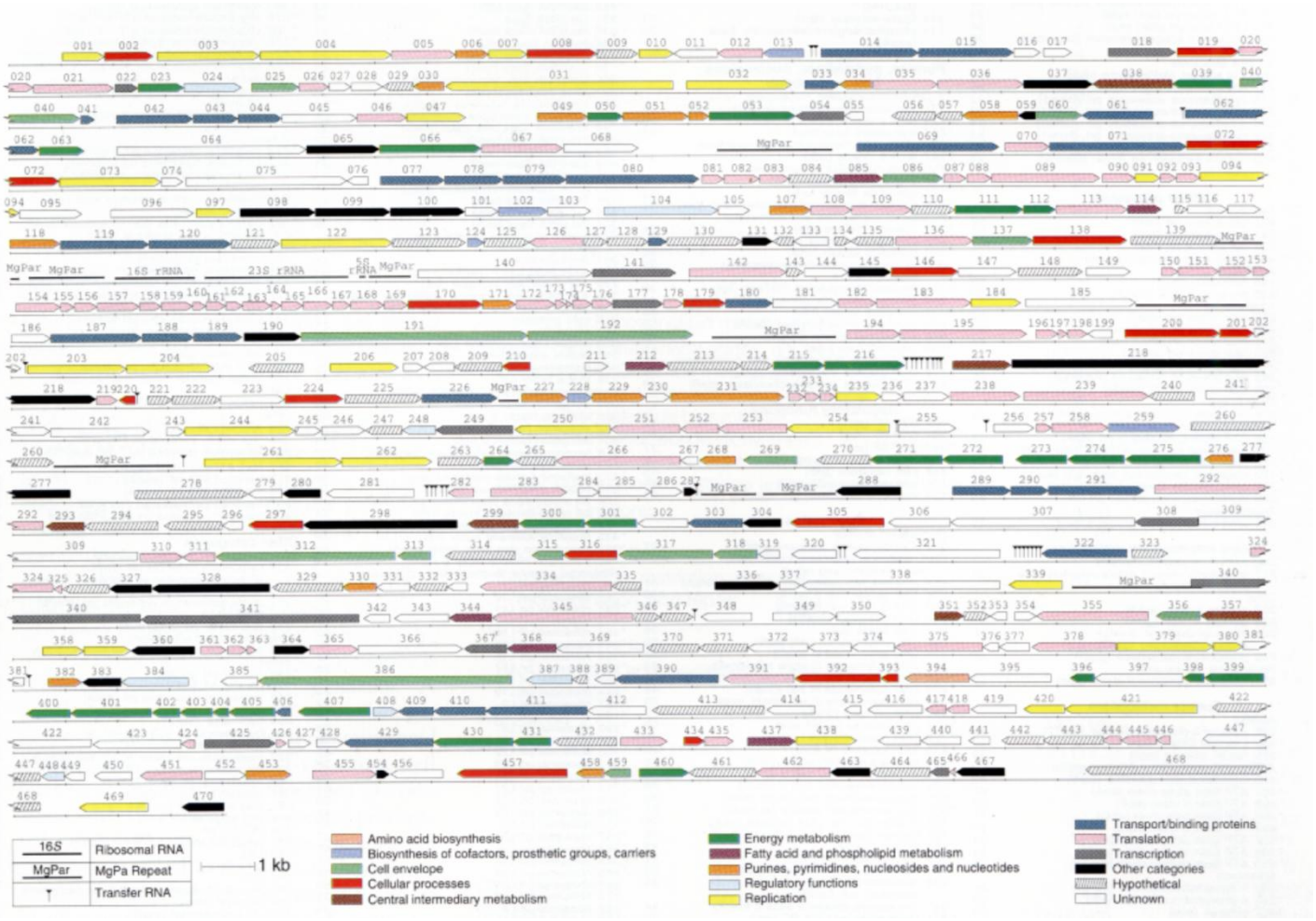
Pour les gènes codant pour les protéines, prédiction fonctionnelle par recherche de similarité de séquences (Blast) et classification en grandes classes fonctionnelles (ex: biosynthèse des acides aminés, métabolisme énergétique….)

Structure secondaire canonique d'une séquence d'ARNt



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

Distribution des unités de traduction et classification fonctionnelle



Stockage et gestion des données : Développement de banques et bases de données

```
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., Blaszczak L.C., Burgett S.,
RA   DeHoff B.S., Estrem S.T., Fritz L., Fu D.-J., Fuller W., Geringer C.,
RA   Gilmour R., Glass J.S., Khoja H., Kraft A.R., Lagace R.E.,
RA   LeBlanc D.J., Lee L.N., Lefkowitz E.J., Lu J., Matsushima P.,
RA   McAhren S.M., McHenney M., McLeaster K., Mundy C.W., Nicas T.I.,
RA   Norris F.H., O'Gara M., Peery R.B., Robertson G.T., Rockey P.,
RA   Sun P.-M., Winkler M.E., Yang Y., Young-Bellido M., Zhao G.,
RA   Zook C.A., Baltz R.H., Jaskunas S.R., Rosteck P.R. Jr., Skatrud P.L.,
RA   Glass J.I.;
RT   "Genome of the bacterium Streptococcus pneumoniae strain R6.";
RL   J. Bacteriol. 183:5709-5717(2001).
CC   ----------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
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; IEA.
DR   InterPro; IPR003488; SMF.
DR   Pfam; PF02481; SMF; 1.
DR   TIGRFAMs; TIGR00732; dprA; 1.
KW   Complete proteome.
SQ   SEQUENCE   286 AA;  31583 MW;  CF12DB83AE3663A2 CRC64;
     MELFMKITNY EIYKLKKSGL TNQQILKVLE YGENVDQELL LGDIADISGC RNPAVFMERY
     FQIDDAHLSK EFQKFPSFSI LDDCYPWDLS EIYDAPVLLF YKGNLDLLKF PKVAVVGSRA
     CSKQGAKSVE KVIQGLENEL VIVSGLAKGI DTAAHMAALQ NGGKTIAVIG TGLDVFYPKA
     NKRLQDYIGN DHLVLSEYGP GEQPLKFHFP ARNRIIAGLC RGVIVAEAKM RSGSLITCER
     AMEEGRDVFA IPGSILDGLS DGCHHLIQEG AKLVTSGQDV LAEFEF
//
```

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

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

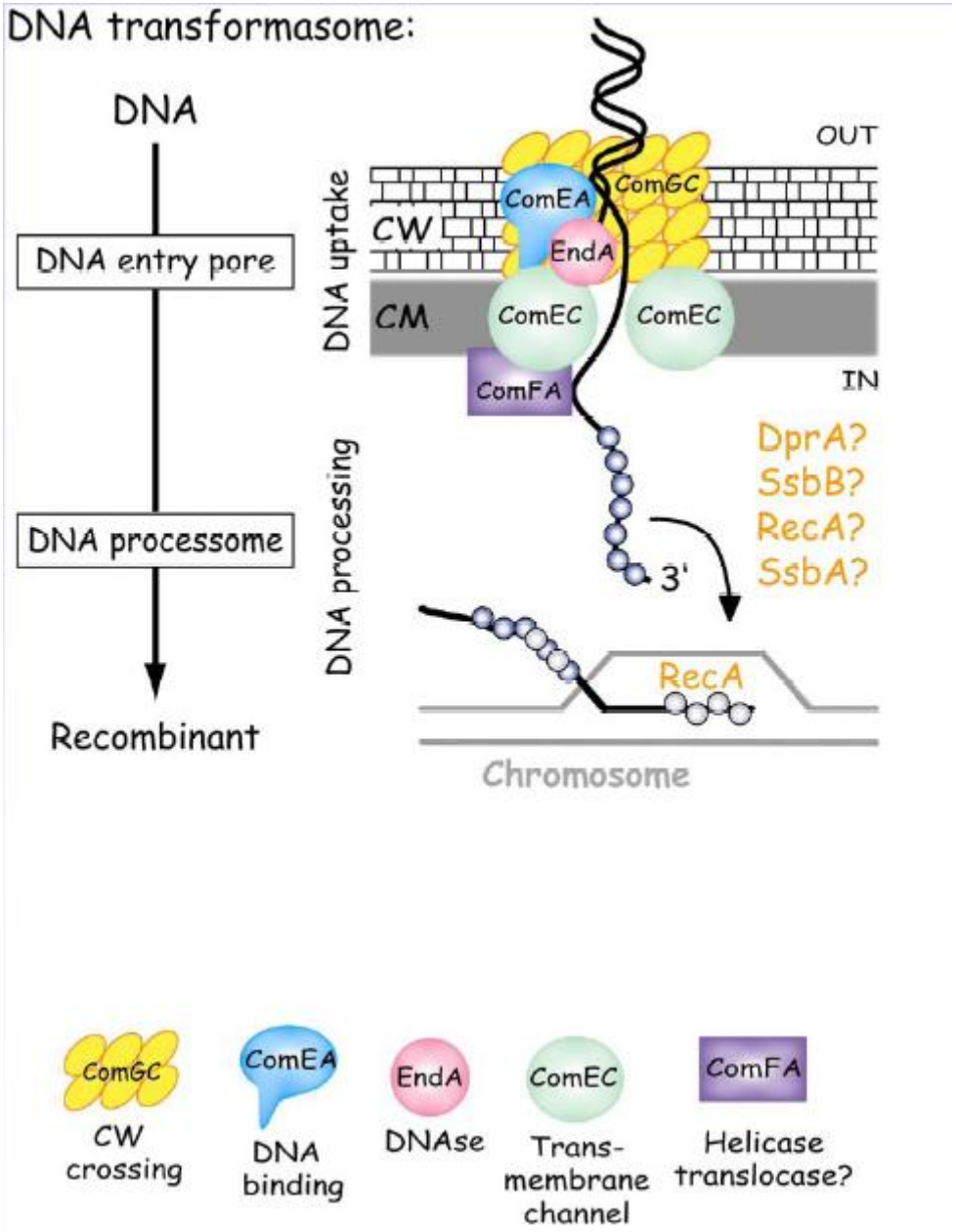
- > 40 espèces de bactéries, distribuées dans tous les groupes taxonomiques.

Rôles de la transformation

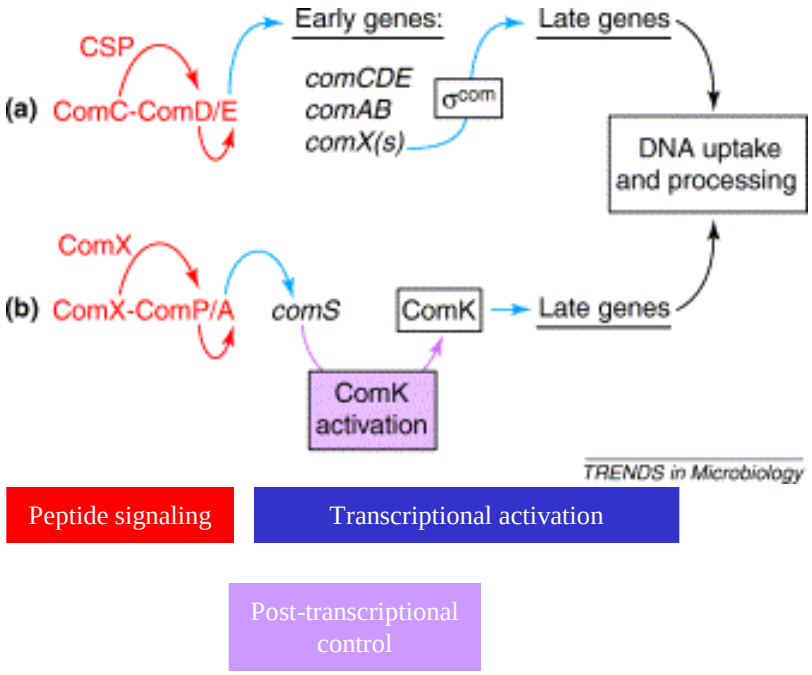
- échanges génétiques (sexualité bactérienne)
- réparation de l'ADN
- nutriments

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

Le transformasome



La cascade de régulation

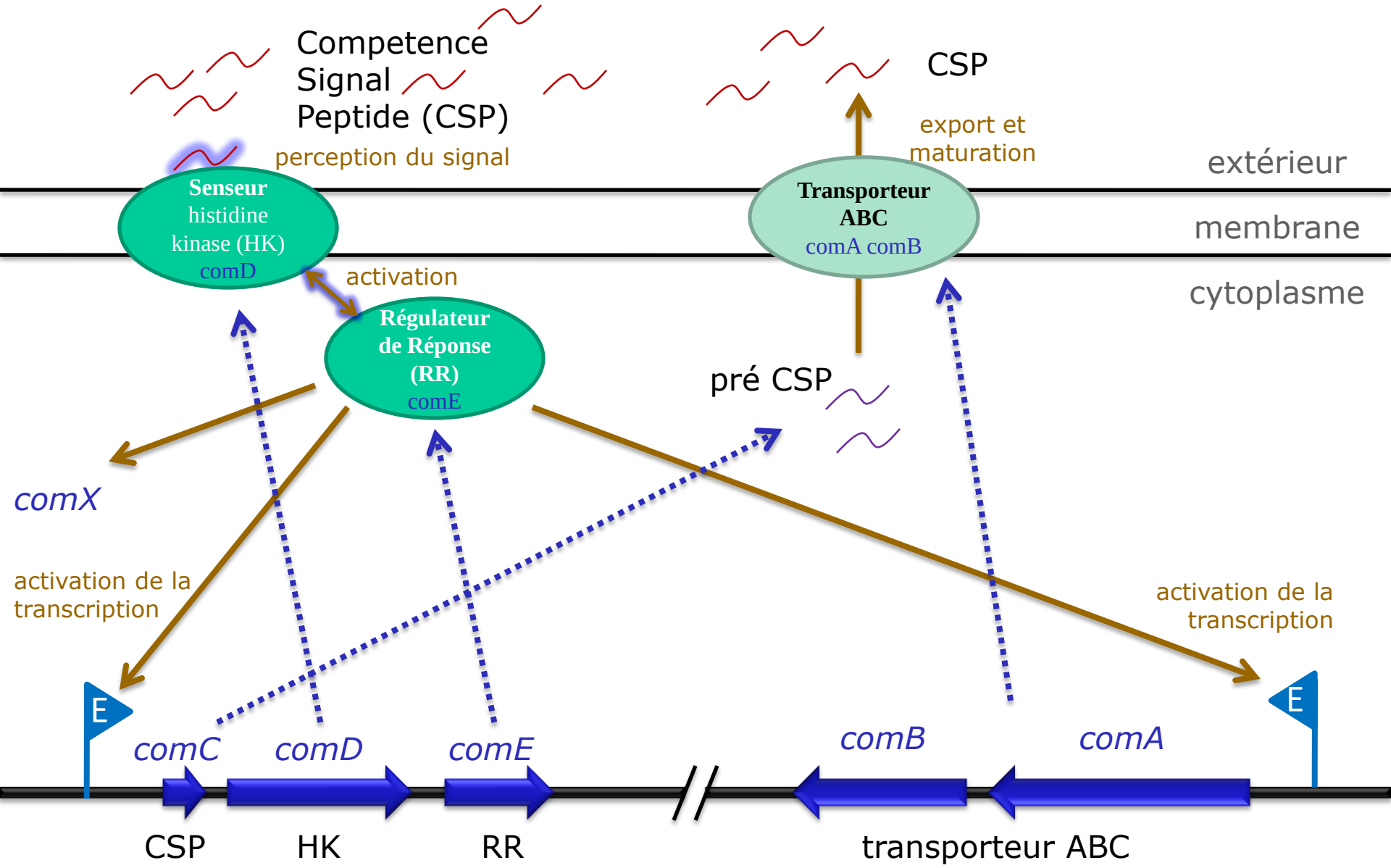


a) *S. pneumoniae*	b) *B. subtilis*
CSP: peptide non-modifié	ComX: peptide modifié
ComAB: export de CSP	ComQ: export de ComX
ComD/E: TCS AgrA-type	ComP/A: TCS LuxA-type
	ComS: accumulation de ComK
ComX: facteur sigma	ComK: facteur de transcription
Inhibée en phase stationnaire	Induite en phase stationnaire
Touche l'ensemble de la population	Touche ~10% de la population
Induction rapide	Induction lente
Délimitée dans le temps	Période étendue

Premières étapes sont similaires mais les acteurs ont des origines différentes

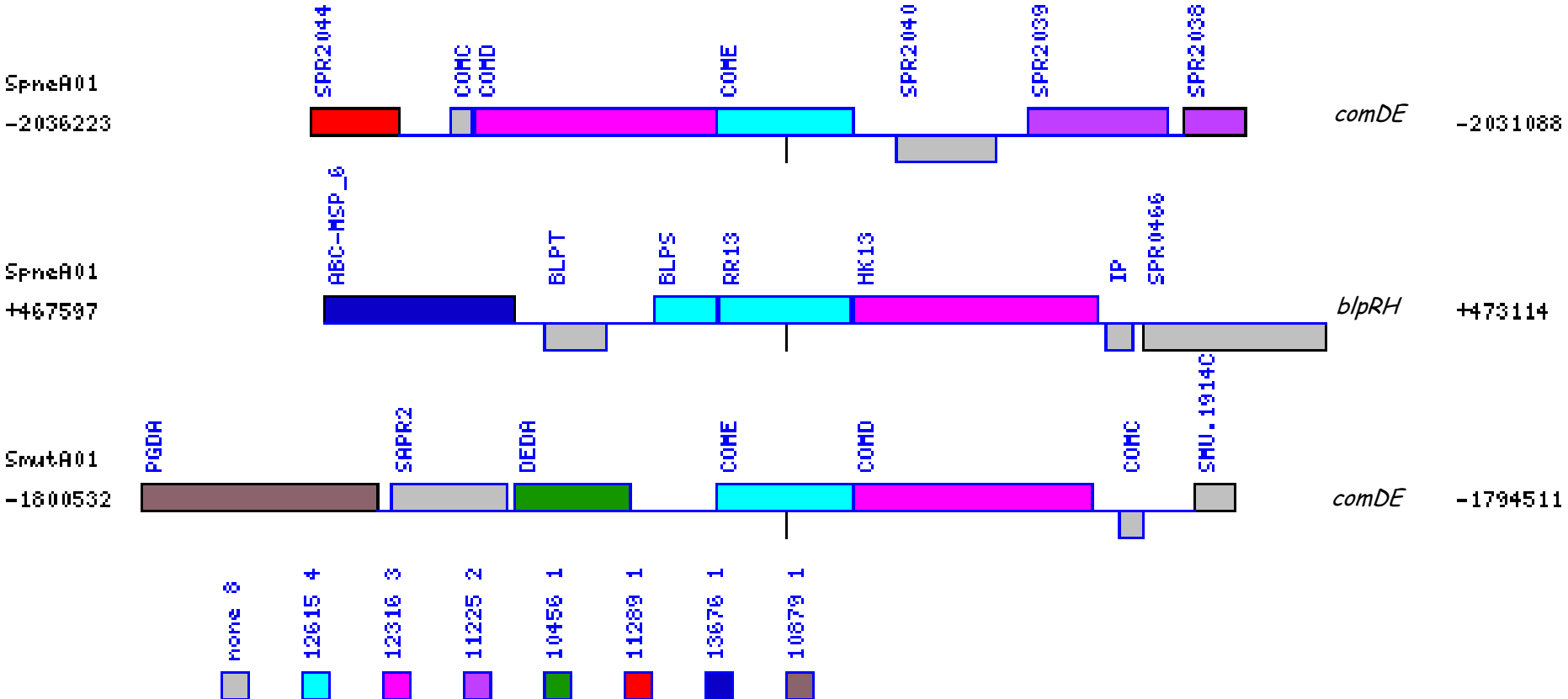
Utiliser le système modèle, *S. pneumoniae,* pour tenter de retracer l'origine et l'évolution des circuits de régulation chez les streptocoques.

Régulation de l'état de compétence, Modèle *S. pneumoniae*



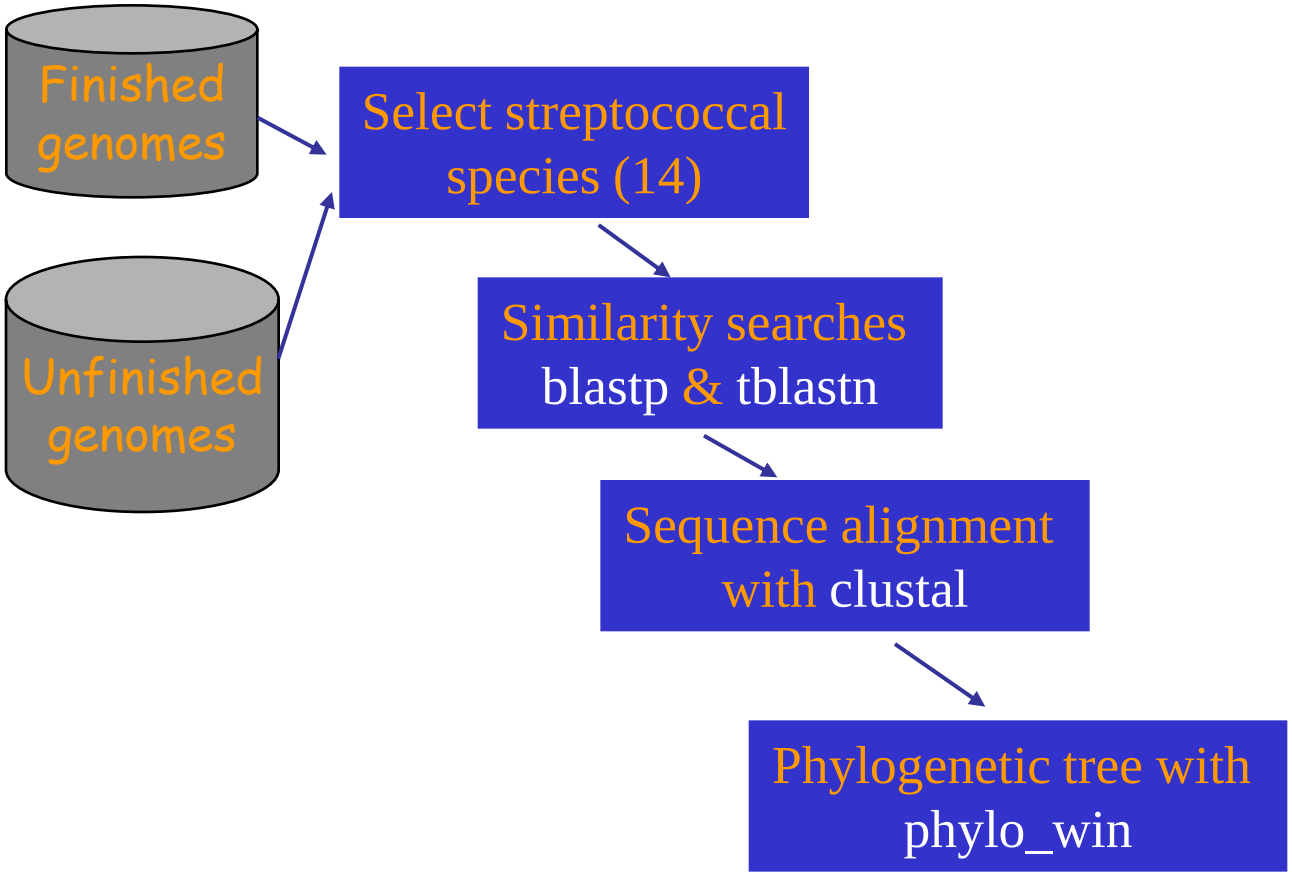
Voisinage chromosomique

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 »).

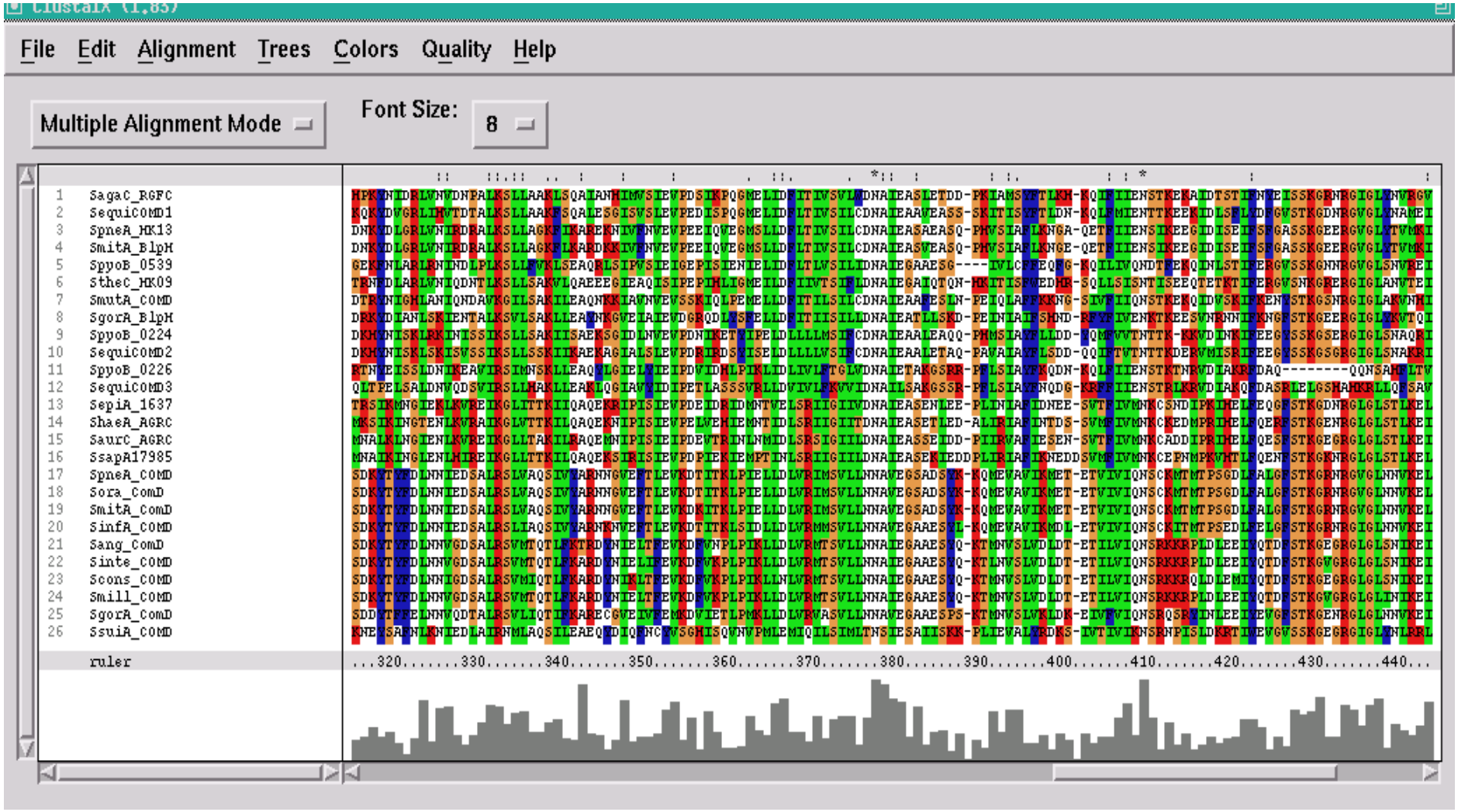


Distribution des gènes homologues à *comD* et *comE*

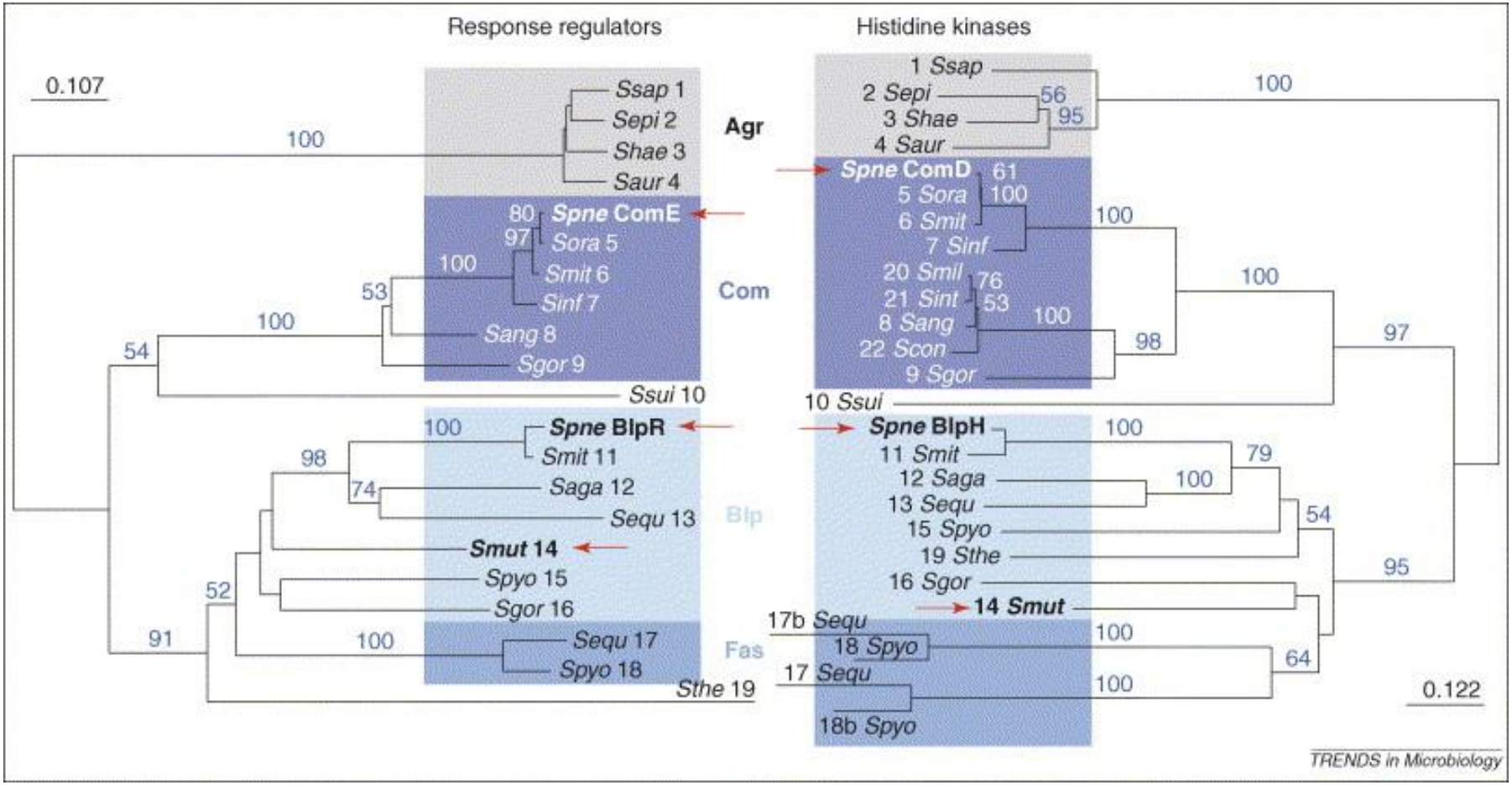
Séquences de référence : protéines ComD et ComE de *S. pneumoniae*



Exemple de l'alignement multiple des protéines homologues à ComD



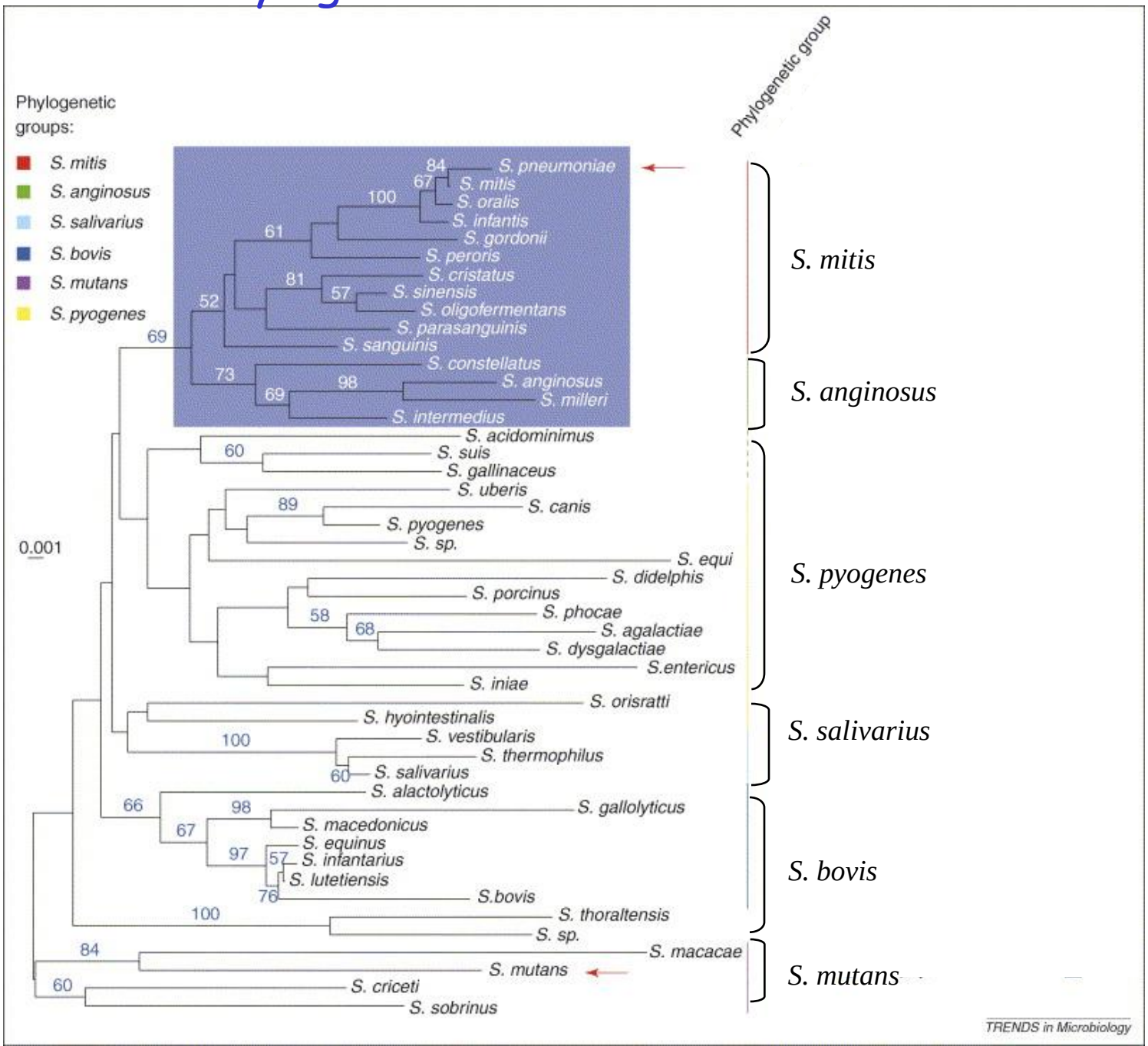
Relations évolutives entre les gènes homologues à *comD* et *comE*



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).

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

Phylogénie basée sur l'ARNr 16S

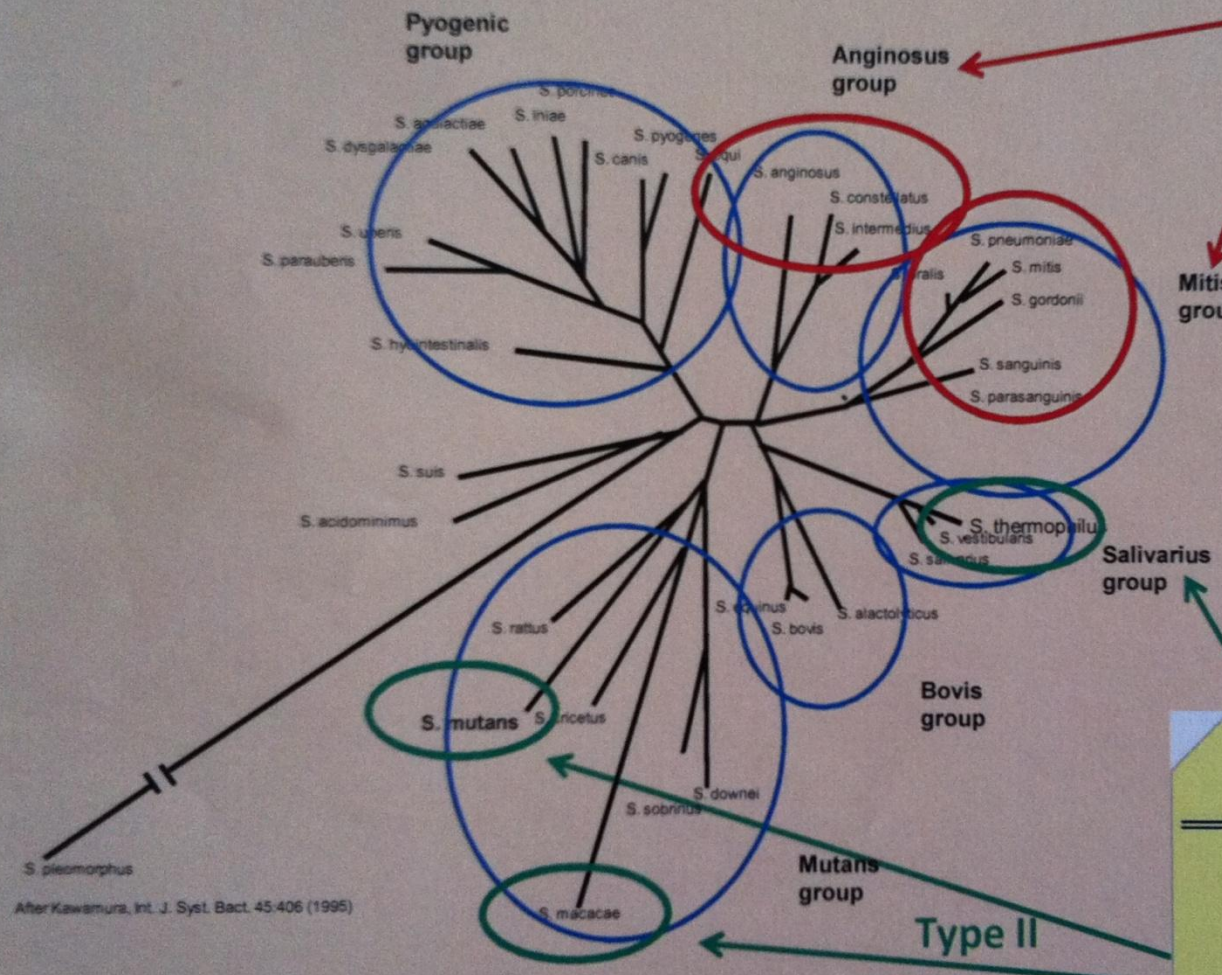


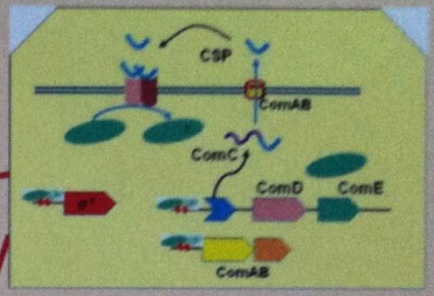
Origine des gènes

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.

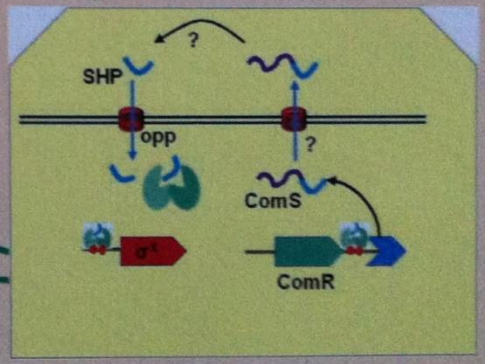
Two types of regulators of *sigX* in different streptococcal groups





TCSTS: ComE

Type I



Rgg: ComR

Type II

Fontaine et al., **2010**. J. Bact. 192:1444.
Mashburn-Warren et al., **2010.** Mol. Microbiol. **78:589**.