Introduction à la Bio-Informatique IFT3295

Nadia El-Mabrouk

DIRO, Université de Montréal

Qu'est-ce que la Bio-informatique?

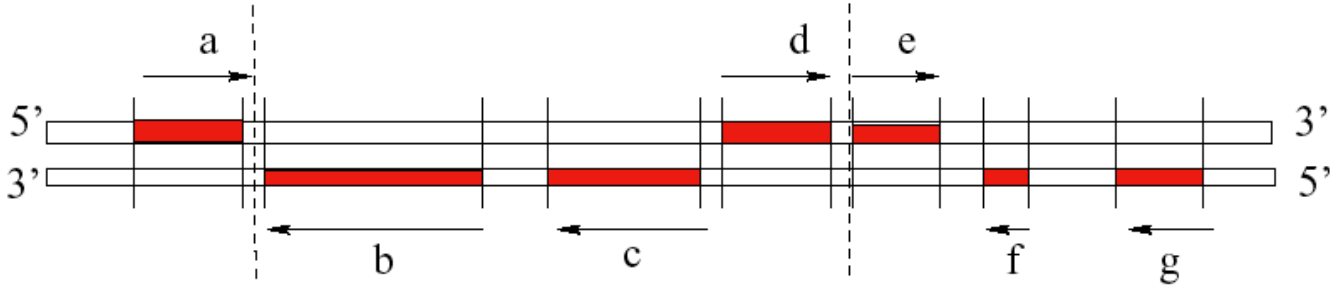
- Champs multi-disciplinaire impliquant la biologie, l'informatique, les mathématiques, les statistiques dont l'objectif est d'analyser les séquences biologiques et de prédire la structure et la fonction des macromolécules.
- Discipline qui évolue en fonction des nouveaux problèmes posés par la biologie.
- Applications à l'agriculture, la pharmacologie, la médecine, la virologie, etc.

Qu'est-ce que la Bio-informatique?

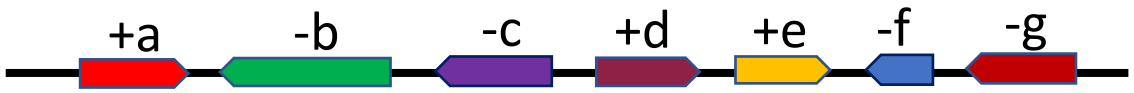
- **Biology "computationnelle"**:
  - Développement d'algorithmes efficaces permettant de résoudre un problème biologique spécifique.
  - Méthodologie générale:
    - Définir le modèle d'évolution;
    - Formaliser le problème biologique;
    - Étudier la complexité théorique du problème;
    - Développer des algorithmes permettant de le résoudre;
    - S'il y a lieu, prouver l'exactitude de l'algorithme
    - Tester l'efficacité de l'algorithme sur des données simulées;
    - L'appliquer à des données biologiques
    - En déduire des hypothèses biologiques.
    - Validation biologique

Exemple : Évolution de l'ordre des gènes

- Segments of DNA transcribed into RNAs are genes.



➢ Gene order : **+a –b –c +d +e –f –g**



➢ Identical to : **+g +f –e –d +c +b –a**

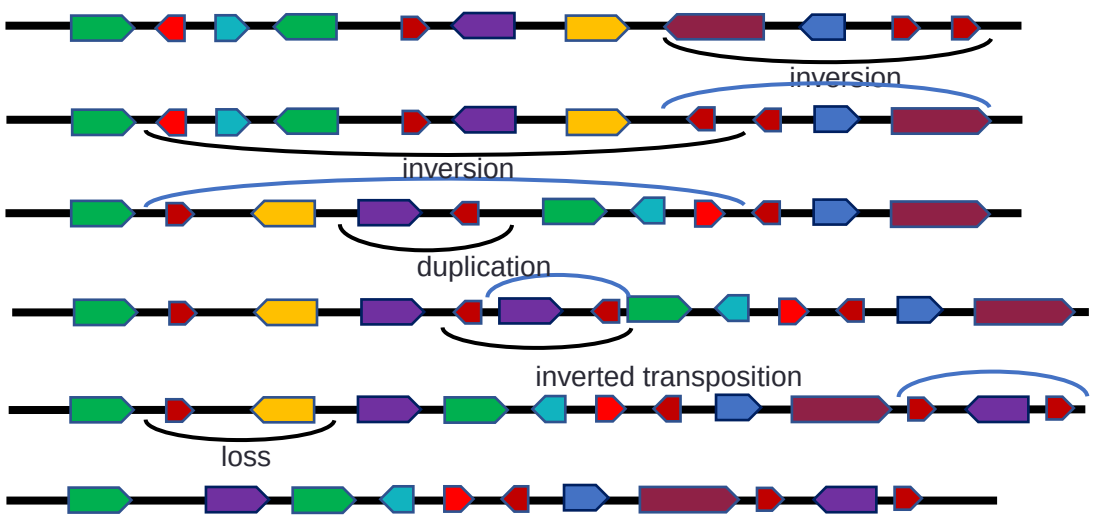
Exemple : Évolution de l'ordre des gènes

- **Données** : Deux génomes bactériens annotés

- **Question biologique** : Déterminer les gènes orthologues. Pour celà, il faut comprendre comment les génomes ont évolué

- **Modèle d'évolution** :
  - Inversions
  - Transpositions
  - Inversions transposées
  - Duplications
  - Pertes

- **Problème informatique:** Trouver le scenario impliquant le minimum de mouvements pour passer d'une configuration à une autre

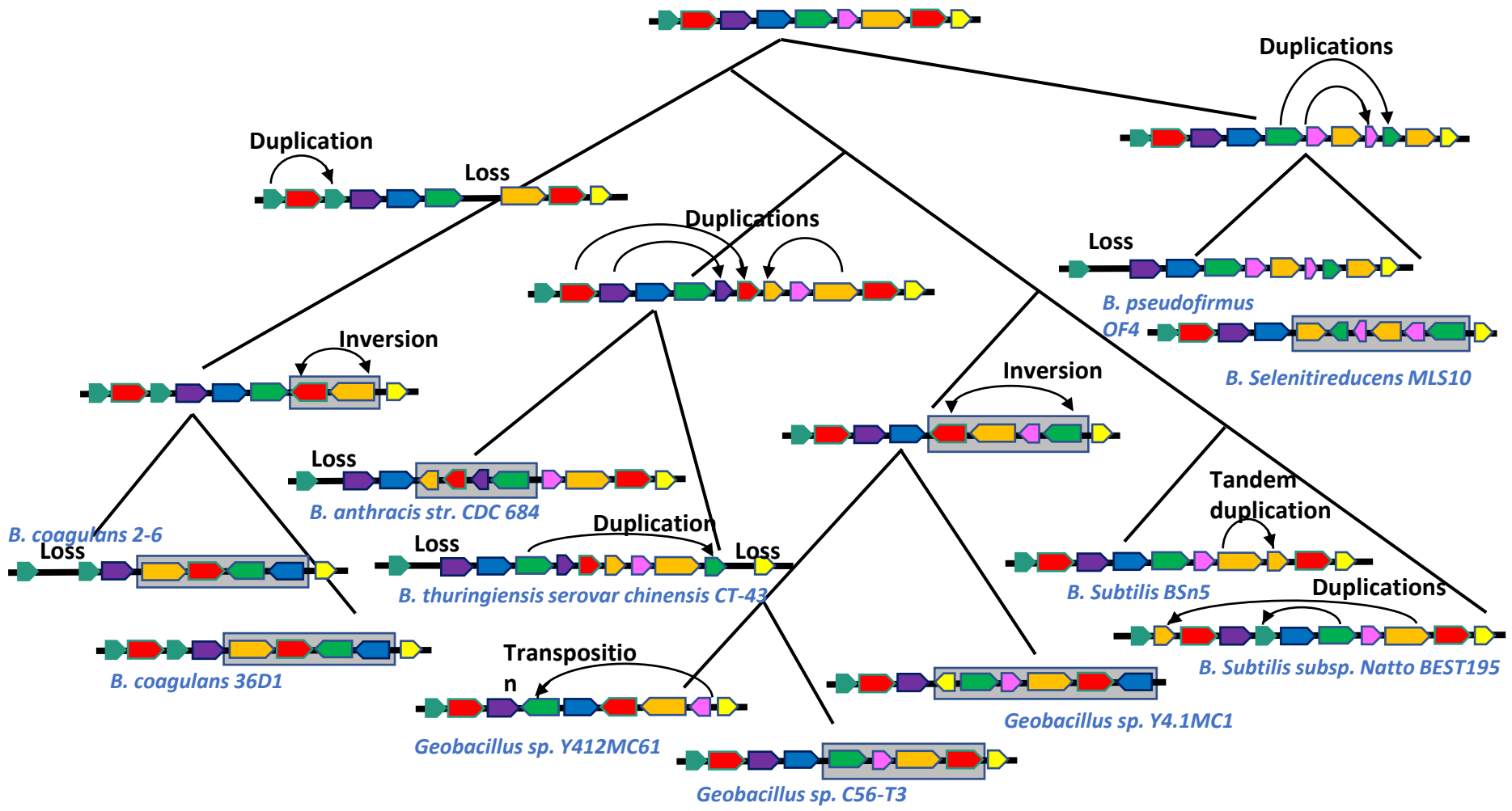


Exemple : Évolution de l'ordre des gènes

Étant donné deux génomes A et B, trouver un scenario de coût minimal pour transformer A en B.

- **Génomes de contenu identique et une seule copie de chaque gène**
  - La plupart des problems peuvent se résoudre en temps polynomial. Par exemple tri de permutations par inversions.
- **Gènes dupliqués**
  - La plupart des problèmes deviennent NP-complet

Représentation simplifiée de l'évolution des operons d'ARN de transfert dans les génomes Bacillus.



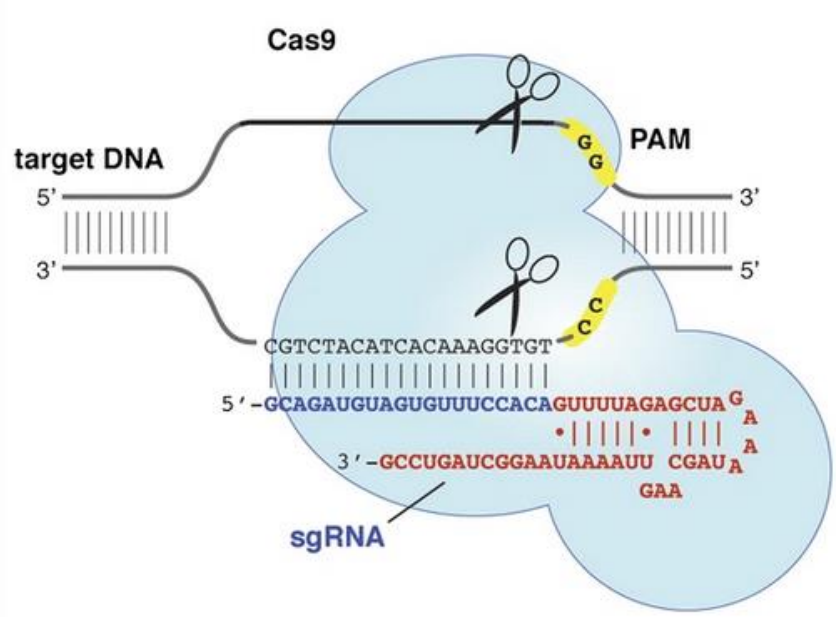
Qu'est-ce que la Bioinformatique?

« Bioinformatics »

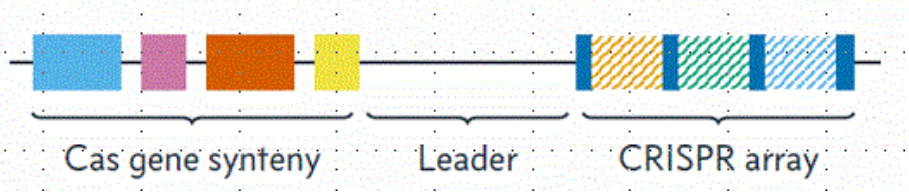
- Difficultés pour les « computational biologists »:
  - Très difficile de définir avec exactitude un modèle adéquat d'évolution des séquences.
  - Les problèmes biologiques sont généralement trop complexes pour pouvoir les résoudre par un algorithme exact en temps raisonnable.
- **Bioinformatique**: Discipline plus pragmatique. Développement d'outils pratiques pour l'analyse et l'organisation des données. Moins d'emphase sur l'exactitude ou l'efficacité de la méthode. Dédiée à des applications pratiques comme l'identification de protéines cible pour la conception de médicaments.

Exemple des systèmes CRISPR-Cas

Système immunitaire adaptatif des bactéries et archée

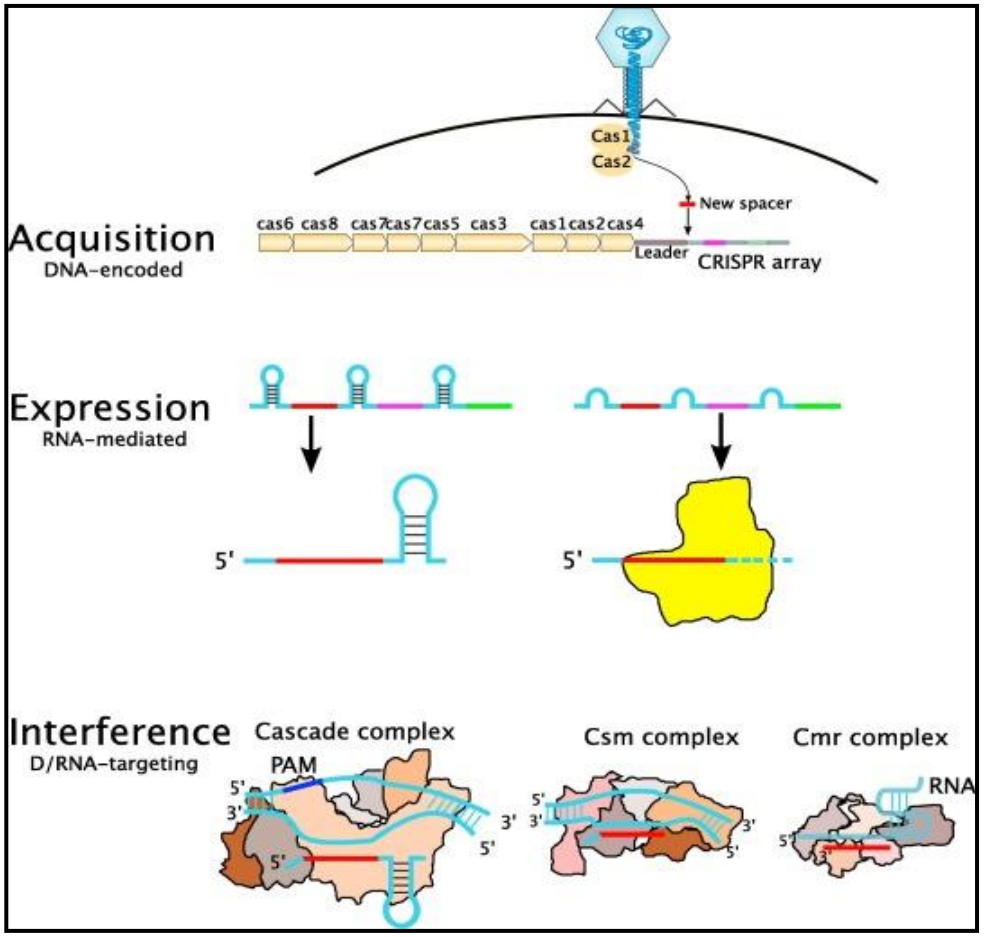


- Système en trois parties:
  - Un réseau de répétitions séparées par des séquences uniques appelées espaceurs (séquences étrangères)
  - Une séquence leader contenant le promoteur.
  - Un ensemble de gènes cas.



CRISPR-Cas

- Trois fonctions :
  - Adaptation
  - Expression - Synthèse des ARN CRISPR (ARNc). La molécule de pré-ARNc est liée à Cas9.
  - Interférence - Clivage de l'ADN.
- CRISPR-Cas de type II, nécessite une seule enzyme Cas9 pour catalyser la coupure de l'ADN



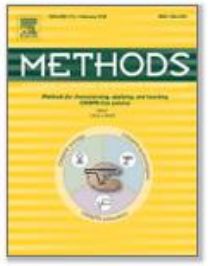
https://www.sciencedirect.com/science/article/abs/pii/S1046202318304717

- La Bioinformatique a joué un rôle clef dans toutes les étapes de la découverte et de l'analyse des séquences CRISPR-Cas

ELSEVIER

Methods

Volume 172, 1 February 2020, Pages 3-11



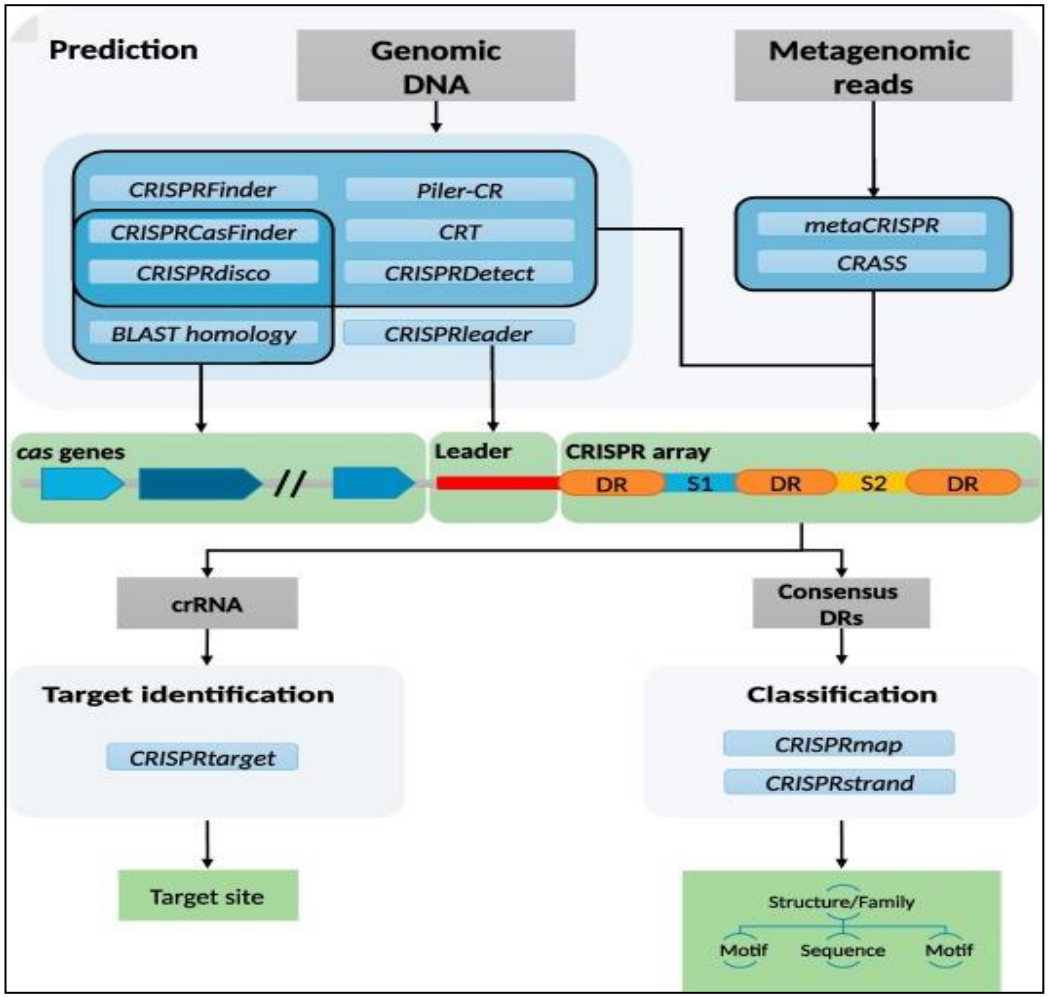
CRISPR-Cas bioinformatics

Omer S. Alkhnbashi [a], Tobias Meier [c], Alexander Mitrofanov [a], Rolf Backofen [a, b], Björn Voß [c]

Show more

Prediction

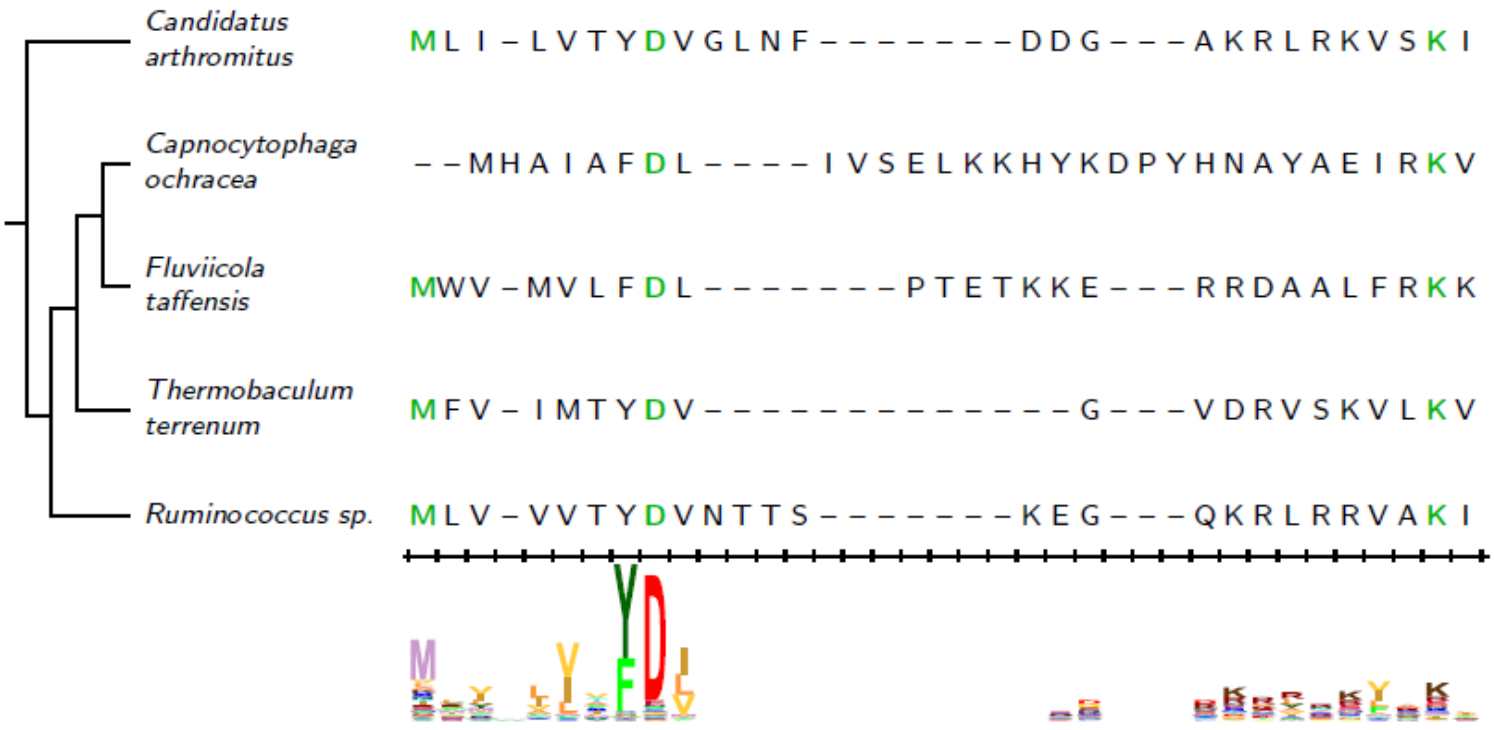
- Identification des genes Cas
  - Par l'outil de recherche BLAST
- Identifier les réseaux de répétitions
  - Outils de Pattern matching
  - Suffix arrays
  - Heuristiques basées sur le principe de fenêtres coulissantes
  - Alignment multiple pour valider la prediction d'espaceurs (CRISPR-arrays). La similarité ne doit pas ête trop grande,
- Leaders: Séquences ADN non-codantes, Peuvent être identifiées par alignement de séquences.



Alignement de la famille Cas2

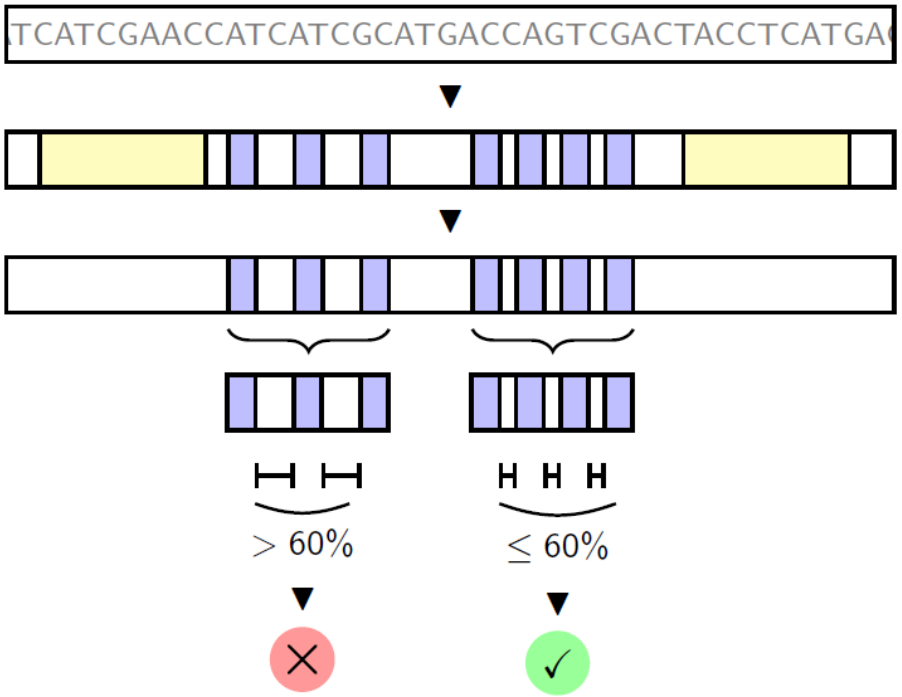
Candidatus arthromitus	M L I L V T Y D V G L N F D D G A K R L R K V S K I
Capnocytophaga ochracea	M H A I A F D L I V S E L K K H Y K D P Y H N A Y A E I R K V
Fluviicola taffensis	M W V M V L F D L P T E T K K E R R D A A L F R K K
Thermobaculum terrenum	M F V I M T Y D V G V D R V S K V L K V
Ruminococcus sp.	M L V V V T Y D V N T T S K E G Q K R L R R V A K I

Inférence evolutive de la famille Cas2



Prédiction des CRISPR Arrays

CRISPRFinder: Using Computational Tools to Detect CRISPR Arrays



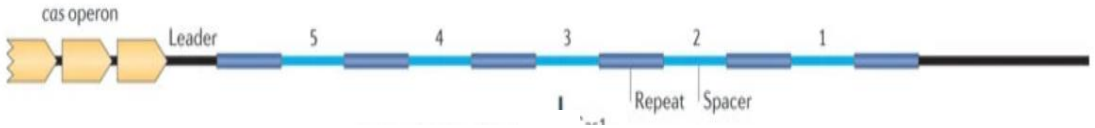
◀ *Input sequence*

1. Find maximal repeats (using suffix arrays)
2. Filter candidates by length
3. Cluster repeats
4. Discard tandem repeats (based on spacer similarity)

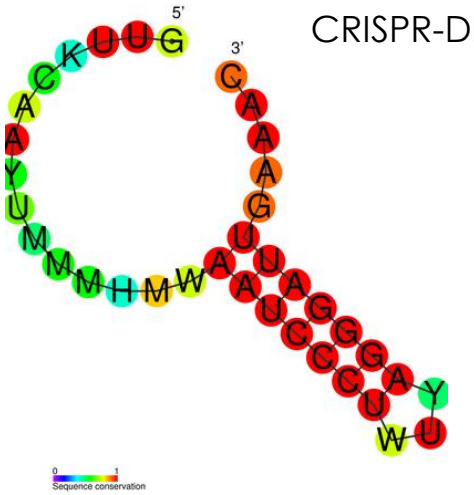
Alignement multiple – Validation des CRISPR arrays

```
Array 1 2969028-2968265                 **** Predicted by CRISPRDetect 2.1 ***
>gi|170079663|ref|NC_010473|-Escherichia coli str. K-12 substr. DH10B chromosome, complete       Array_Or
```

Position	Repeat	%id	Spacer	Repeat_Sequence	Spacer_Sequence
2969028	29	100.0	32	.....	CTTTCGCAGACGCGCGGCGATACGCTCACGCA
2968967	29	100.0	32	.....	CAGCCGAAGCCAAAGGTGATGCCGAACACGCT
2968906	29	100.0	32	.....	GGCTCCCTGTCGGTTGTAATTGATAATGTTGA
2968845	29	100.0	33	.....	TTTGGATCGGGTCTGGAATTTCTGAGCGGTCGC
2968783	29	100.0	33	.....	CGAATCGCGCATACCCTGCGCGTCGCCGCCTGC
2968721	29	100.0	32	.....	TCAGCTTTATAAATCCGGAGATACGGAAACTA
2968660	29	96.6	32	.....A.....	GACTCACCCCGAAAGAGATTGCCAGCCAGCTT
2968599	29	100.0	32	.....	CTGCTGGAGCTGGCTGCAAGGCAAGCCGCCCA
2968538	29	100.0	32	.....	GGGGGCGCATGACCGTAAACATTATCCCCCGG
2968477	29	100.0	32	.....	GGAGTTCAGACATAGGTGGAATGATGGACTAC
2968416	29	93.1	32	.....TT.....	CCCGGTAGCCAGGTTTGCAACGCCTGAACCGA
2968355	29	96.6	32	.....A.....	GCAACGACGGTGAGATTTCACGCCTGACGCTG
2968294	29	89.7	0	.T.....AT.....	|
13	29	98.2	32	GAGTTCCCCGCGCCAGCGGGGATAAACCG	

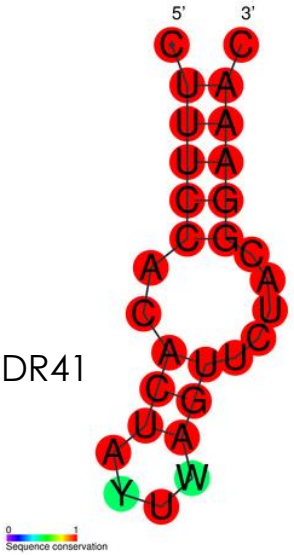


Répétitions palindromiques qui doivent se replier en structures secondaires



CRISPR-DR5

…MWAAUCCCUWUYAGGGAUUGA…

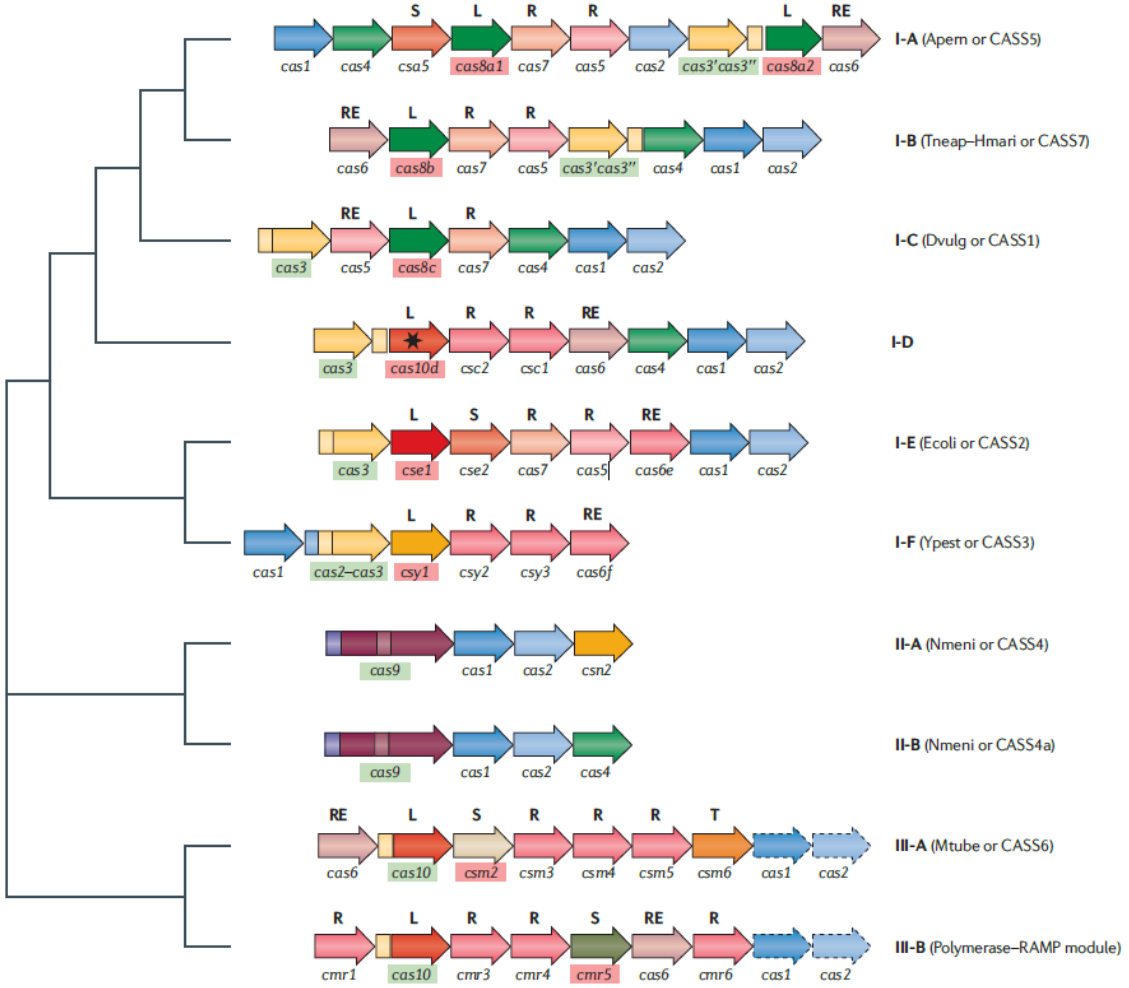


CRISPR-DR41

…CUUUCCACACUAYUWAGUUCUACGGAAAC…

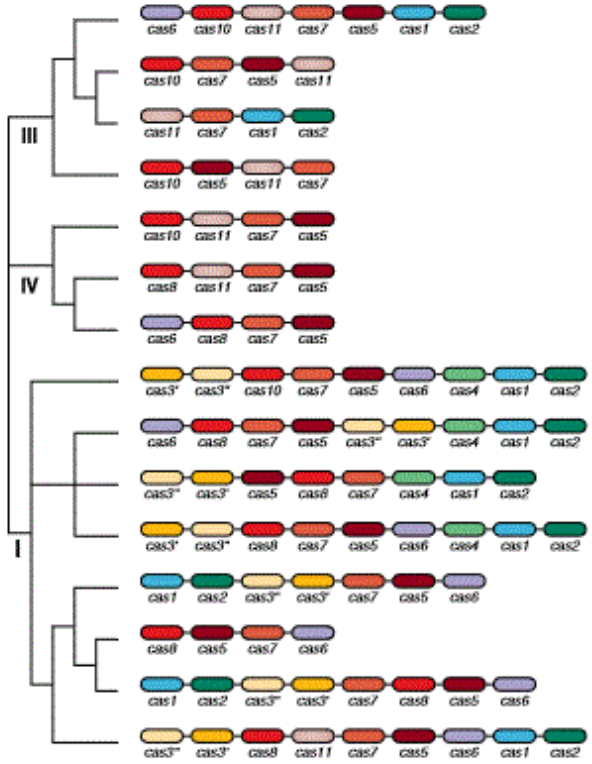
https://en.wikipedia.org/wiki/CRISPR#Locus_structure

Classification des CRISPR-Cas



Systèmes CRISPR-Cas

Architecture de l'opéron des gènes Cas de la classe I

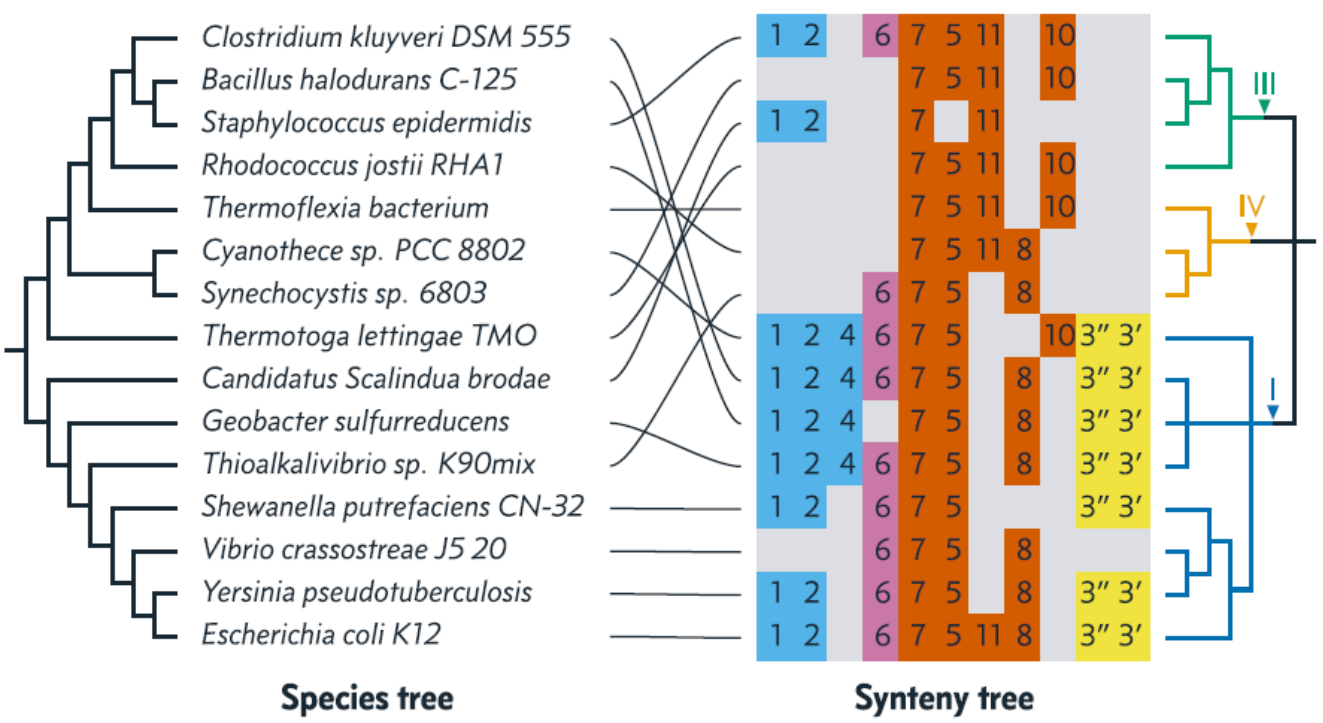


Arbre de synténies G

QUESTION:
Comment les gènes Cas ont-ils évolué à l'intérieur de l'arbre des espèces?

Version ordonnée

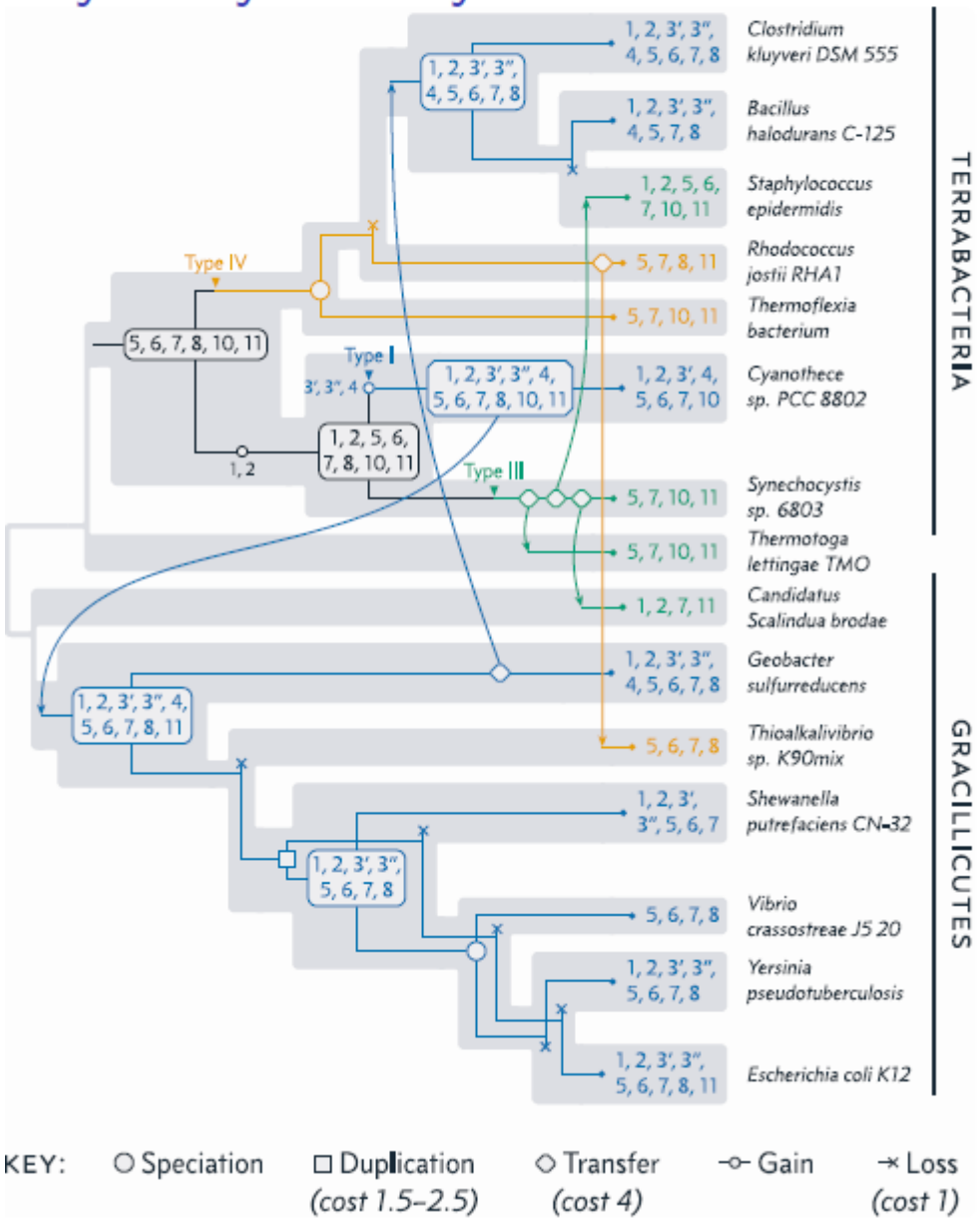
Systèmes CRISPR-Cas



QUESTION:
Comment les gènes Cas ont-ils évolué à l'intérieur de l'arbre des espèces?

Version non-ordonnée

Synteny history scenario inferred with SuperDTL



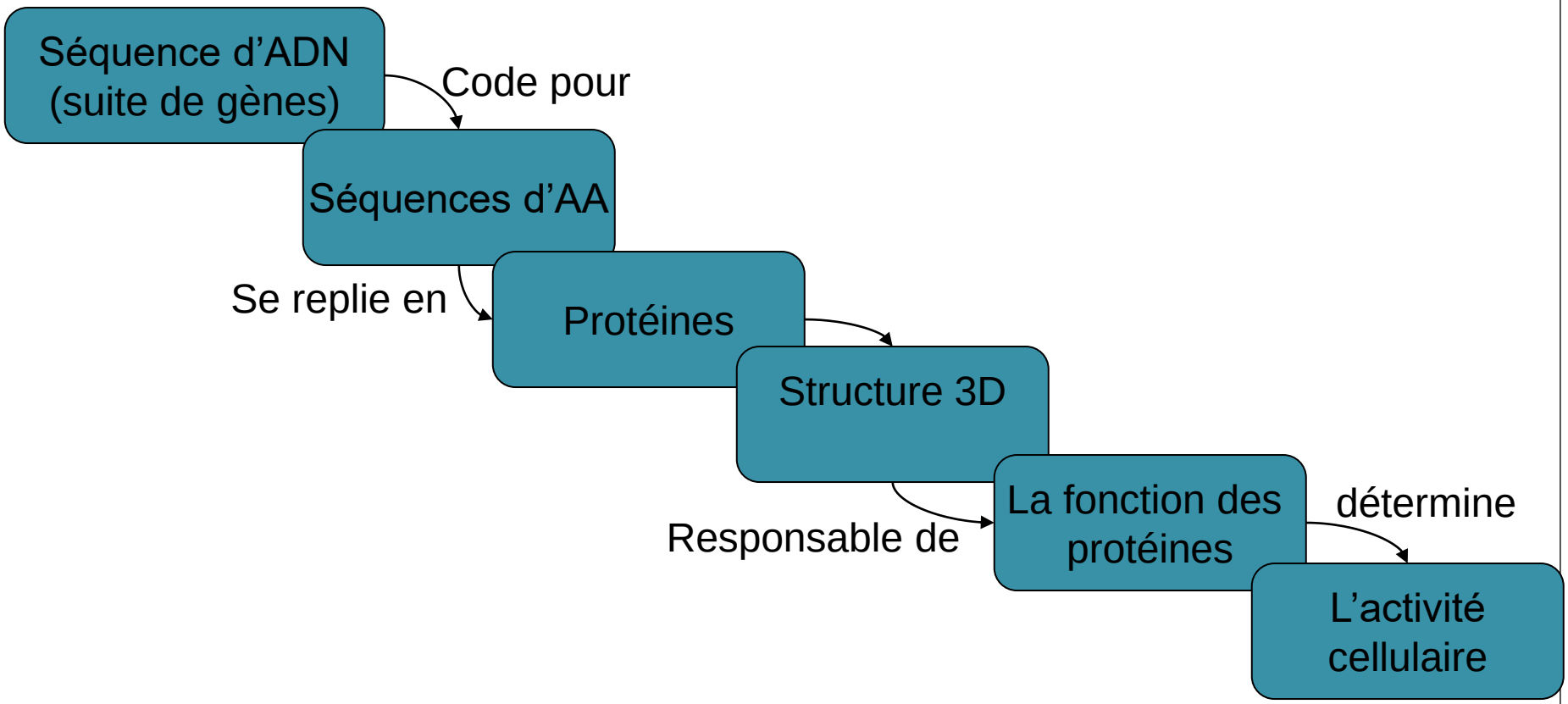
- Emergence of Cas synteny at the root of the Terrabacteria
- Emergence of Types I and III in the Cyanobacteria ancestor after inclusion of Cas1 and Cas2
- Emergence of Type I with inclusion of Cas3
- Mostly in line with evolutionary scenario from [Koonin, E. V. Makarova, K. S. **Evolutionary plasticity and functional versatility of CRISPR systems.** PLOS Biology 20, e3001481 (2022)]

Retour à l'histoire de la Bioinformatique Événements majeurs

- Bioinformatique: Apparue dans les années 1960, après que les biologistes aient découverts comment séquencer de l'ADN et les protéines.
  - Dans les années 1950, Frédérick Sanger détermine la séquence des acides aminés de l'insuline.
  - 1965, Margaret Dayhoff: Premier atlas de séquences de protéines
  - Dans les années 1970, Russel F. Doolittle: l'un des premiers à avoir utilisé l'ordinateur pour analyser les protéines.
  - Quelques autres pères fondateurs: Walter M. Fitch, Michael S. Waterman, David Sankoff, etc.

Qu'est-ce que la Bioinformatique?

De l'ADN à la fonction cellulaire



Slide inspiré de http://www.macdevcenter.com

Qu'est-ce que la Bioinformatique?
La séquence code pour la fonction

Bonne nouvelle:

- De plus en plus de génomes complètement séquencés
- Il existe une correspondance directe entre la séquence et la fonction
  - Séquence d'ADN d'un gène → structure de la protéine

Malheureusement:

- Pas d'algorithme universel permettant de faire le lien entre la séquence et la fonction.

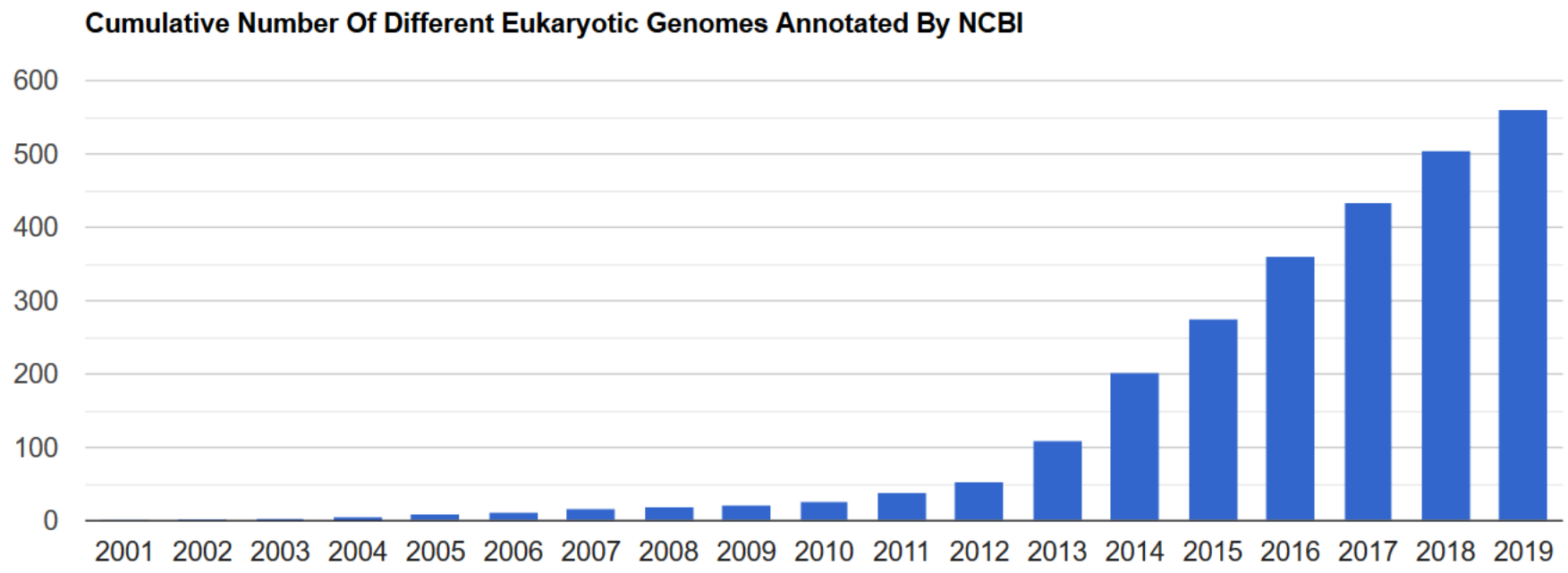
Qu'est-ce que la Bioinformatique?
Défis

- Décoder l'information contenue dans les séquences d'ADN, i.e.
  - Trouver les gènes
  - Prédire la séquence d'AA produite par un gène
  - Identifier les régions régulatrices du génome
  - Étudier l'évolution des génomes …
- Génomique structurale:
  - Prédire les structures 2D et 3D des protéines et des ARN structurels…
- Génomique fonctionnelle
  - Étudier la régulation des gènes
  - Étudier le niveau d'expression des gènes (microarrays)
  - Déterminer les réseaux d'interaction entre les protéines…

Qu'est-ce que la Bioinformatique?

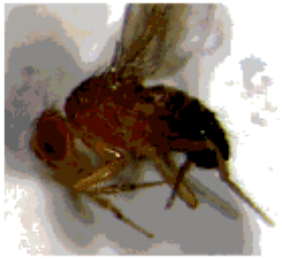
Défi

- Croissance exponentielle des séquences de nucléotides et d'AA dans les banques de données biologiques.
- Croissance exponentielle de génomes séquencés.



http://www.ncbi.nlm.nih.gov/genome/annotation_euk/

Whole Genomes

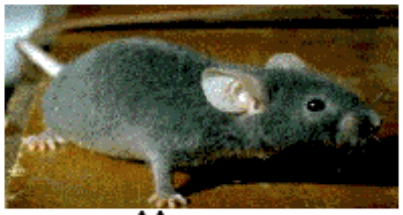


Drosophila

C. elegans

Rat

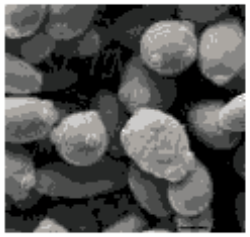
Human



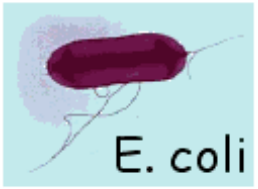
Mouse

Rice

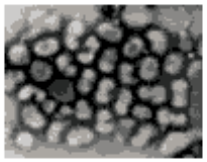
Mosquito



Yeast



E. coli



H. influenza

Arabidopsis

http://bip.weizmann.ac.il/education/course/introbioinfo/04/lect1/introbioinfo04/sld016.htm

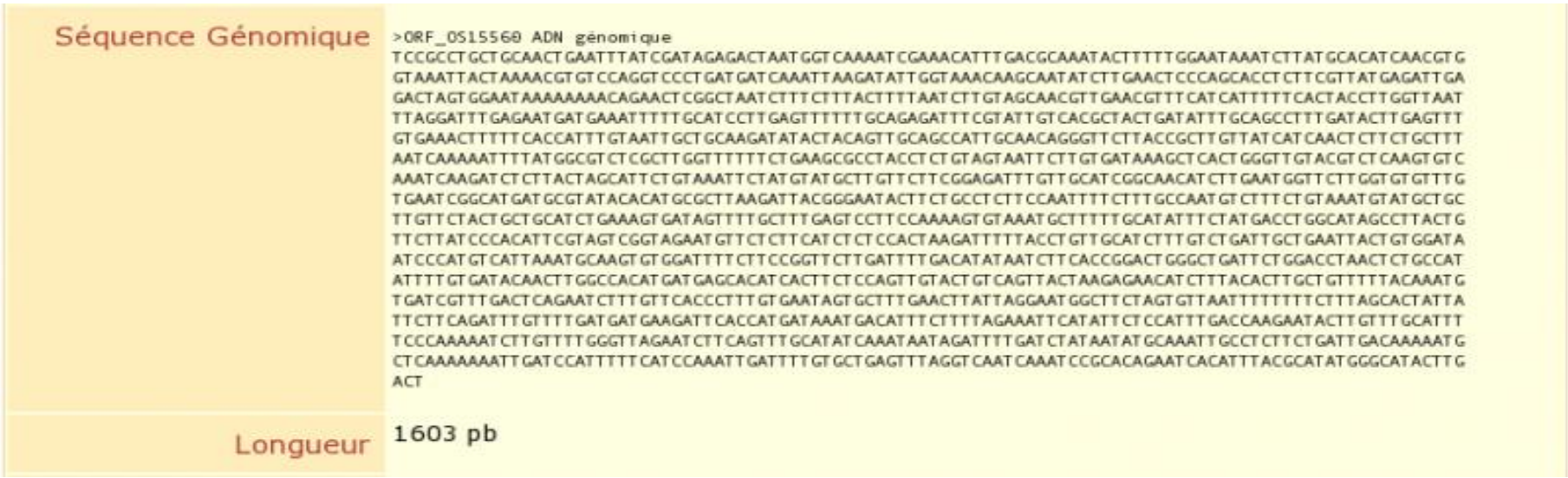
Séquençage d'ADN

- 1977, Frédérick Sanger: Premier génome séquencé: Virus bactérien.
- 1995, J. Craig Venter: Premier génome bactérien: *H. Influenzae*
- 1996: Pemier génome eukaryote (levure *S. cerevisiae*).
- 1997: Bactérie E. coli, modèle important en microbiologie.
- 1998: Premier génome animal: le ver plat *C. elegans*
- 2000: Premier génome végétal: *A. Thaliana*; 1ère plante alimentaire: le riz.
- 2001: Génome humain …

Qu'est-ce que la Bioinformatique?

Pour les informaticiens

- Malgré sa complexité, l'ADN peut être représenté comme un texte de 4 caractères A,C,G,T, et les protéines comme des mots sur un alphabet de 20 lettres.
- Décoder le texte de l'ADN: une manne de problèmes mathématiques, statistiques, algorithmiques, combinatoires …



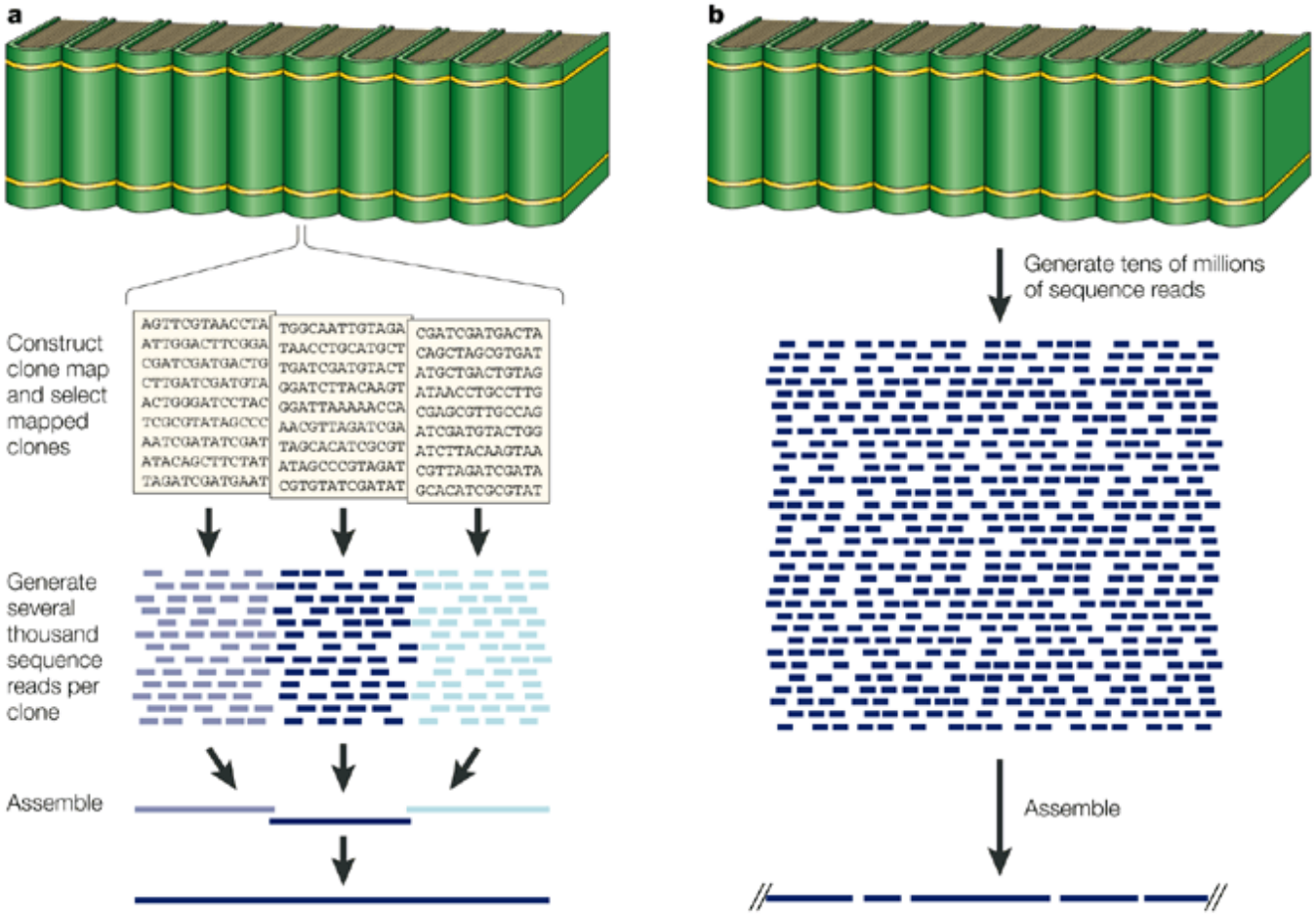
Séquence Génomique	>ORF_OS15560 ADN génomique TCCGCCTGCTGCAACTGAATTTATCGATAGAGACTAATGGTCAAAATCGAAACATTTGACGCAAATACTTTTTGGAATAAATCTTATGCACATCAACGTG GTAAATTACTAAAACGTGTCCAGGTCCCTGATGATCAAATTAAGATATTGGTAAACAAGCAATATCTTGAACTCCCAGCACCTCTTCGTTATGAGATTGA GACTAGTGGAATAAAAAAAACAGAACTCGGCTAATCTTTCTTTACTTTTAATCTTGTAGCAACGTTGAACGTTTCATCATTTTTCACTACCTTGGTTAAT TTAGGATTTGAGAATGATGAAATTTTTGCATCCTTGAGTTTTTTGCAGAGATTTCGTATTGTCACGCTACTGATATTTGCAGCCTTTGATACTTGAGTTT GTGAAACTTTTTCACCATTTGTAATTGCTGCAAGATATACTACAGTTGCAGCCATTGCAACAGGGTTCTTACCGCTTGTTATCATCAACTCTTCTGCTTT AATCAAAAATTTTATGGCGTCTCGCTTGGTTTTTTCTGAAGCGCCTACCTCTGTAGTAATTCTTGTGATAAAGCTCACTGGGTTGTACGTCTCAAGTGTC AAATCAAGATCTCTTACTAGCATTCTGTAAATTCTATGTATGCTTGTTCTTCGGAGATTTGTTGCATCGGCAACATCTTGAATGGTTCTTGGTGTGTTTG TGAATCGGCATGATGCGTATACACATGCGCTTAAGATTACGGGAATACTTCTGCCTCTTCCAATTTTCTTTGCCAATGTCTTTCTGTAAATGTATGCTGC TTGTTCTACTGCTGCATCTGAAAGTGATAGTTTTGCTTTGAGTCCTTCCAAAAGTGTAAATGCTTTTTGCATATTTCTATGACCTGGCATAGCCTTACTG TTCTTATCCCACATTCGTAGTCGGTAGAATGTTCTCTTCATCTCTCCACTAAGATTTTTACCTGTTGCATCTTTGTCTGATTGCTGAATTACTGTGGATA ATCCCATGTCATTAAATGCAAGTGTGGATTTTCTTCCGGTTCTTGATTTTGACATATAATCTTCACCGGACTGGGCTGATTCTGGACCTAACTCTGCCAT ATTTTGTGATACAACTTGGCCACATGATGAGCACATCACTTCTCCAGTTGTACTGTCAGTTACTAAGAGAACATCTTTACACTTGCTGTTTTTACAAATG TGATCGTTTGACTCAGAATCTTTGTTCACCCTTTGTGAATAGTGCTTTGAACTTATTAGGAATGGCTTCTAGTGTTAATTTTTTTTCTTTAGCACTATTA TTCTTCAGATTTGTTTTGATGATGAAGATTCACCATGATAAATGACATTTCTTTTAGAAATTCATATTCTCCATTTGACCAAGAATACTTGTTTGCATTT TCCCAAAAATCTTGTTTTGGGTTAGAATCTTCAGTTTGCATATCAAATAATAGATTTTGATCTATAATATGCAAATTGCCTCTTCTGATTGACAAAAATG CTCAAAAAAATTGATCCATTTTTCATCCAAATTGATTTTGTGCTGAGTTTAGGTCAATCAAATCCGCACAGAATCACATTTACGCATATGGGCATACTTG ACT
Longueur	1603 pb

ADN - Séquençage

- Action de déterminer la suite de nucléotides d'un fragment d'ADN.
- Petite histoire du séquençage:
  - 1977: Technique Maxam et Gilbert: 1.5kb /personne/année
  - 1988: séquençage par capillaire: 10Mb / personne/année
  - 2008: SOLiD ABI: 150Gb /personne/année
  - 2010: environ 2000Gb/personne/année
    - Petit fragments: routinier au laboratoire
    - Génomes complets: de plus en plus commun (génome en moins d'un mois)
- Impossibilité de séquencer plus de mille bases par réaction

ADN - Séquençage

Séquençage par « shotgun »

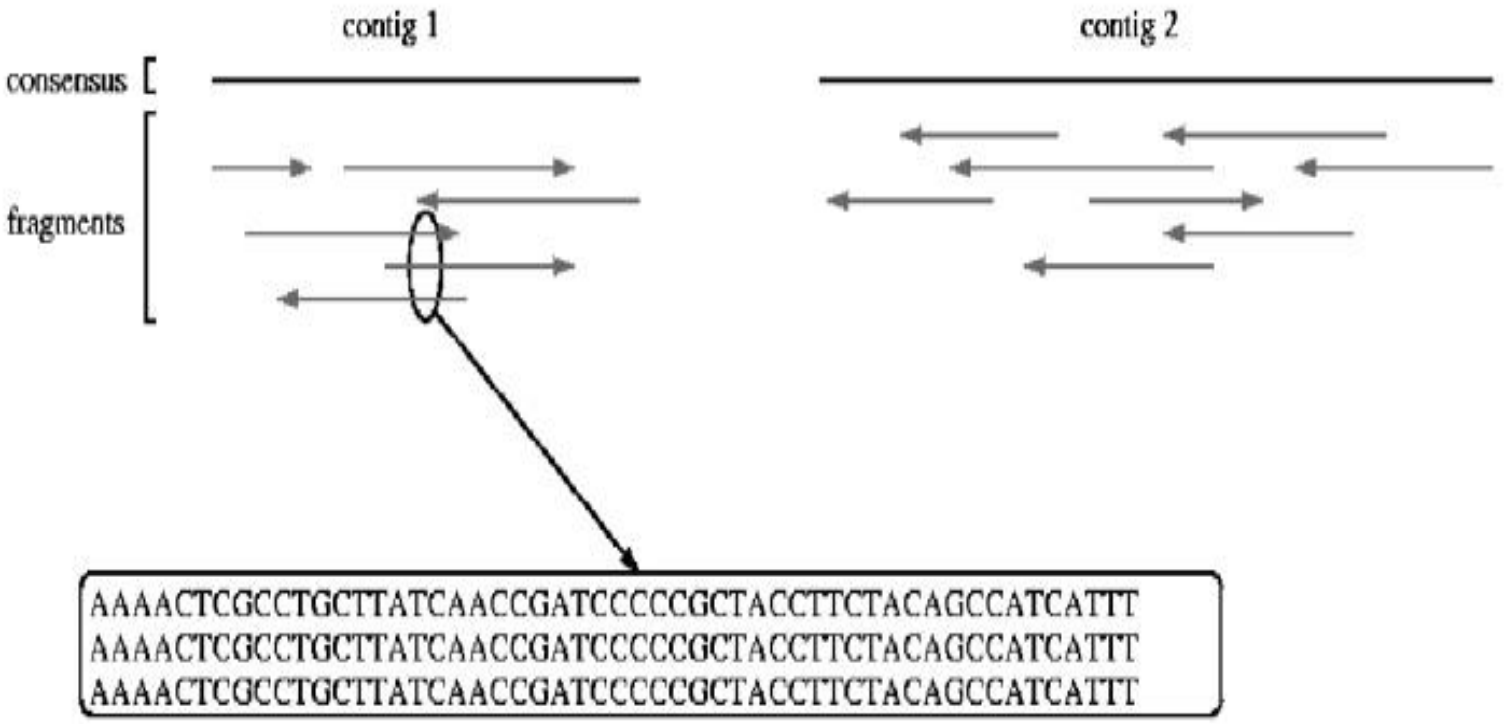


http://www.nature.com/nrg/journal/v2/n8/fig_tab/nrg0801_573a_F6.html

ADN - Séquençage

Assemblage

- Problème: Reconstruire le génome à partir des fragments (reads) provenant du séquençage.



ADN - Séquençage

Assemblage

- Difficultés :
  - Les fragments séquencés peuvent contenir des erreurs

Exemple:

```
          T
TTGCTGAC        TTACATTT
```

1) si on prend les fragments avec erreur on obtient

```
    TTGCTGAC
TTACATTT
```

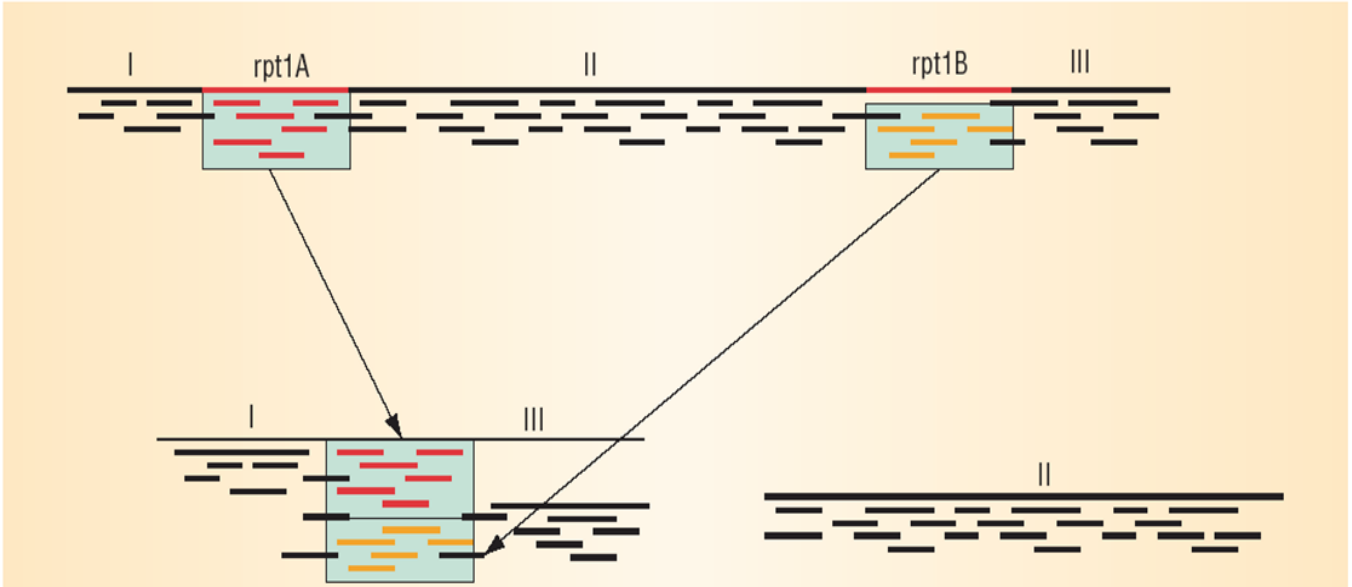
2) si on n'avait pas fait d'erreur lors du séquençage, on aurait plutôt:

```
TTGCTTAC
    TTACATTT
```

ADN - Séquençage

Assemblage

- Difficultés :
  - Les fragments séquencés peuvent contenir des erreurs
  - Les génomes peuvent contenir beaucoup de séquences répétées



© M.Pop, S.L. Salzberg and M. Shumway, *Genome Sequence Assembly, Algorithms and Issues*, IEEE Computer, 2002

Figure 2. Repeat sequence. The top represents the correct layout of three DNA sequences. The bottom shows a repeat collapsed in a misassembly.

ADN - Séquençage

Assemblage

- Difficultés :
  - Les fragments séquencés peuvent contenir des erreurs
  - Les génomes peuvent contenir beaucoup de séquences répétées
  - Certains fragments d'ADN sont impossibles à séquencer (le fragment est toxique pour l'organisme dans lequel il est cloné)
    - assemblage partiel

ADN - Séquençage

Algorithmes d'assemblage

- **Première approximation du problème:**
  - Trouver la plus courte super-séquence commune à un ensemble de séquences:

Étant donné un ensemble de séquences $\{s_1, s_2, \ldots, s_n\}$, trouver la plus courte séquence T, telle que chaque séquence s_i est une sous-séquence de T

- **Problème NP-difficile.**
- **Heuristiques** pour résoudre ce problème

ADN - Séquençage

Algorithmes d'assemblage

- **Solutions heuristiques :**
  1. Calculer un score de chevauchement pour chaque paire de séquences;
  2. Coller les deux séquences ayant un score maximal en une nouvelle séquence;
  3. Recommencer les étapes 1) et 2) avec le nouvel ensemble de séquences (qui contient maintenant une séquence en moins) jusqu'à ce qu'il ne reste qu'une séquence.

ADN - Séquençage
Algorithmes d'assemblage

- **Facteur d'approximation:**
  - L'algorithme glouton trouve une séquence T qui est, dans le pire des cas, 4 fois plus longue que la séquence optimale. Une optimisation de l'algorithme glouton permet de réduire à *3n*.

[A. Blum, T. Jiang, M Li, J. Tromp and M. Yannakakis, Journal of the ACM, 41:630-647, 1994]
https://www.cs.cmu.edu/~avrim/Papers/superstring.pdf

  - Amélioration a un ratio de 2.596. Aucun exemple n'a été trouvé produisant une solution plus de 2 fois la longueur d'une séquence optimale. Il est conjecturé que l'algorithme glouton est une 2-approximation.

[D. Breslauer, T. Jiang and Z. Jiang. Journal of Algorithms, 24:340-353, 1997]
https://pure.mpg.de/rest/items/item_1827386/component/file_2574031/content

ADN - Séquençage

Algorithmes d'assemblage

- Algo glouton facile à implémenter, utilisé dans les premiers assembleurs (TIGR assembler, CAP3, Phrap)
- **Problèmes:**
  - Travaille localement et ignore les relations de longues portées entre les séquences → problèmes avec les répétitions
  - Demande beaucoup de temps de calcul.
- Autres approches utilisant des graphes : nœuds représentent les morceaux de séquences et les arêtes indiquent les chevauchement entre les séquences.

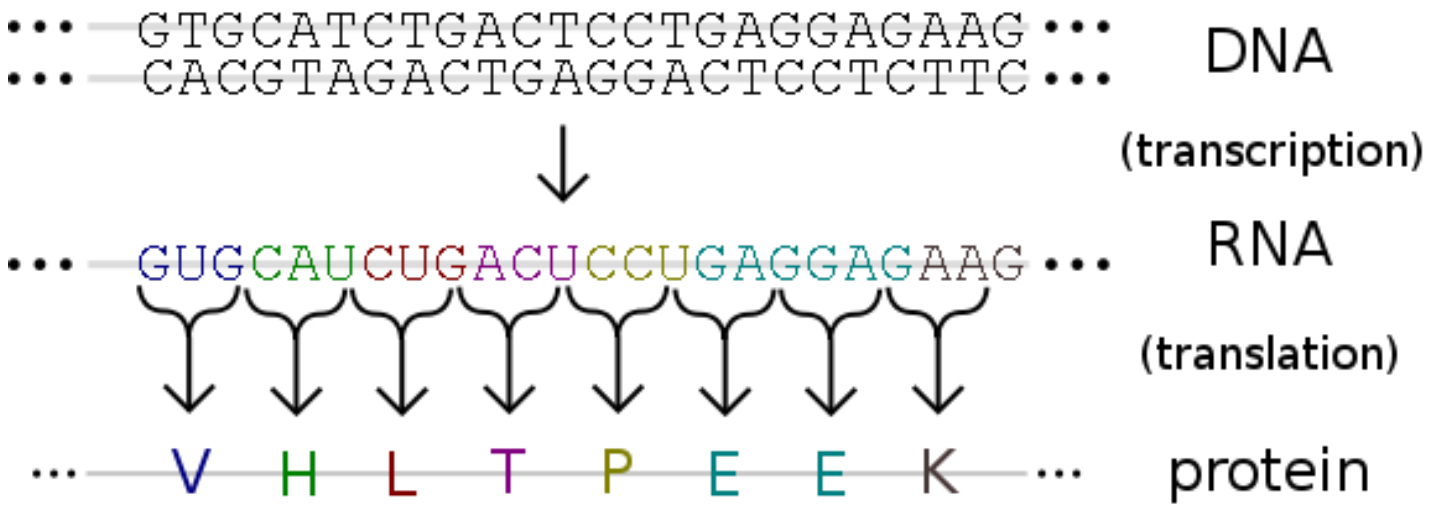
ADN - Annotation

- Une Séquence d'ADN:
  - Est-ce que cette séquence a déjà été complètement ou partiellement déposée dans les banques de données?
  - Codant? Non-codant?
  - Y a-t-il des gènes?…

tcacaaattgttactgaaatagttgagattg
tagttataagagtttagtgcgaagcctttgg
cagtaatgcttactacgtatttgctaaagta
actataatctttgaggaattagaagtagcta
tgtccttgttatcagttcaatgatatagctaa
ttattgtatttagcagcaacggtataatgat
ctgttaatacttaatatgatagagagtggtt
gttgtgaattgcatagtgtgattgccgaggc
cttaaactagaggaattaccaagtcatctcc
taaatctgaatatgtcaaatattcttcgctca
ttaataaataagtggattatagaaggcata
ttgacttatggacggattacttaacgggtga
gaaatttgaagtggaatatgcccaatattta
gactaataccgatctagtcagattgagaaa
tgttctaactgtatcattgctaagaattactt
aatataagtctaaatatcttgttgtatgggg
ggtggtctttcccctaccaatagtaaatgta
aatctagctcaatttggctttattgtcttgtta
aatccgtaattagttaatatgatggtattaa
agttacaatatttagactaataccgatctag

ADN – Annotation

Transcription et traduction (suite)

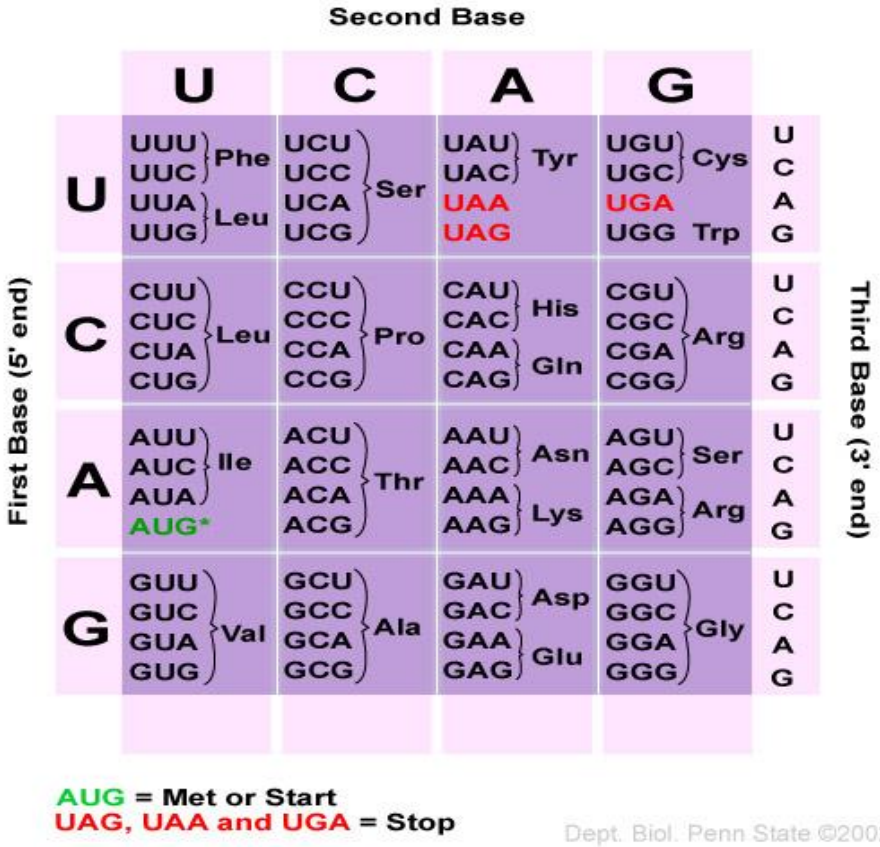


http://fr.wikipedia.org/wiki/Fichier:Genetic_code.svg, 01/2011

ADN – Annotation

Open Reading Frame

Code génétique universel:

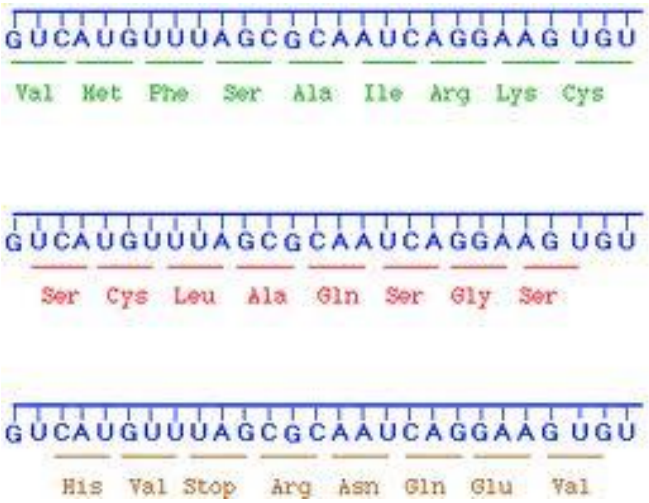


- Un « cadre de lecture » est une région d'une taille suffisante (ex. 90 nuc.) située entre un codon START (ou Methyanine) et un codon STOP.

ADN – Annotation

Open Reading Frame (ORF)

- Chaque région d'ADN a 6 cadres de lecture différents: 3 sur un brin et 3 sur le brin complémentaire



http://www.uic.edu/classes/phar/phar331/lecture3/
01/2011

- Habituellement, un seul cadre de lecture est utilisé pour la traduction d'un gène.

ADN – Annotation

Détection des ORF chez les Procaryotes

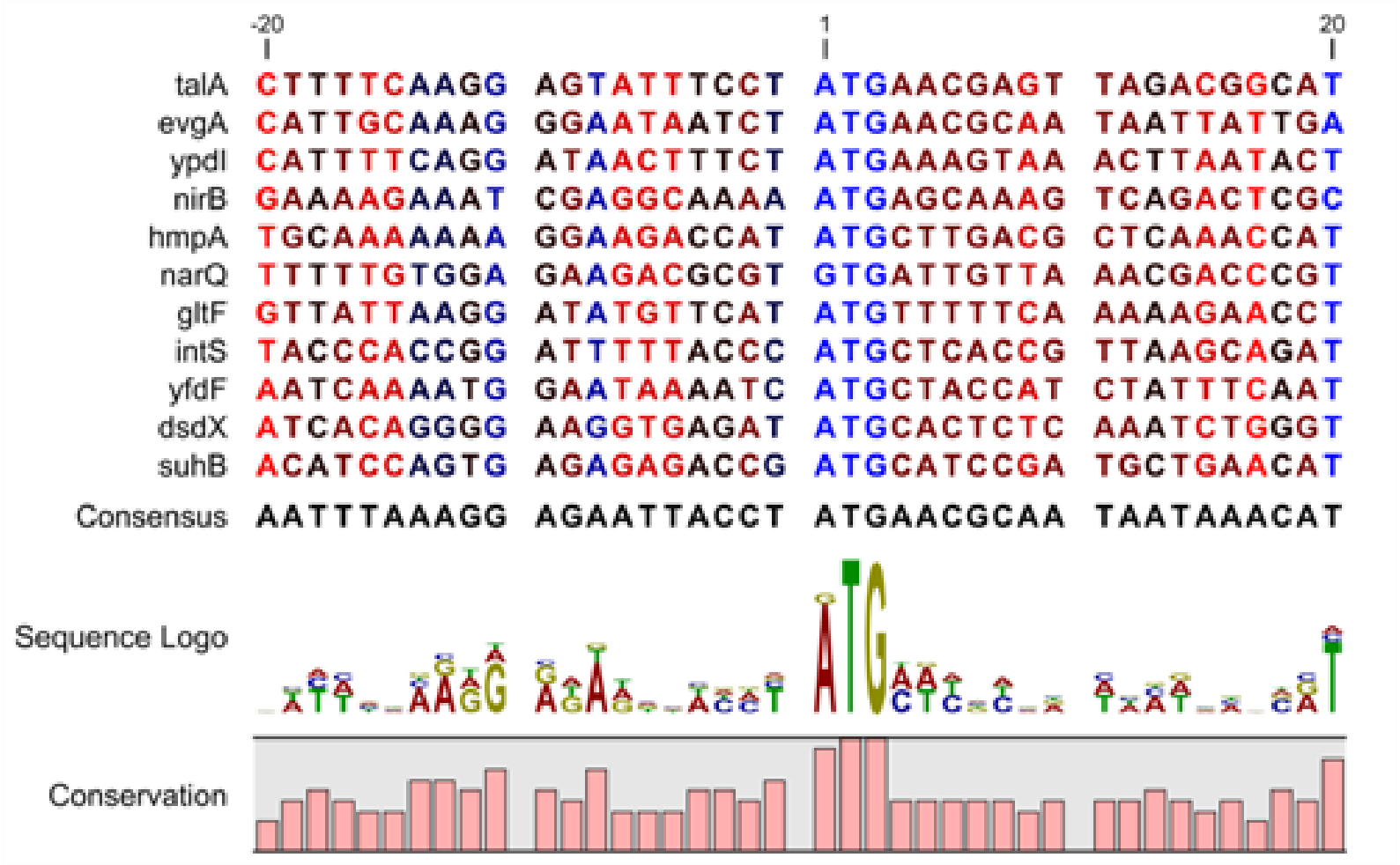
- Gènes généralement sans introns.
- Grande densité en gènes (environ 1 gène par kb)
- Séquence de gène, typiquement: séquence d'une taille dépassant un certain seuil (ex. 90 nuc.), commençant par un codon START et finissant par un codon STOP, et ne contenant que des triplets codant pour des AA.
- En général, plusieurs « débuts » possibles pour un ORF. Typiquement, prendre le plus long ORF.

ADN – Annotation
Difficultés de la prédiction de gènes

- Un ORF n'est pas nécessairement un gène.
- Le code génétique peut varier du code génétique universel.
- La structure des gènes est compliquée:
  - Présence d'introns
  - Épissage alternatif
- Comment choisir le bon codon START dans un ORF?
- Gènes incomplets, pseudogènes, erreurs de séquençage
- Gènes chevauchants (cadres de lecture différents)
- …

ADN - Annotation

Concept de similarité

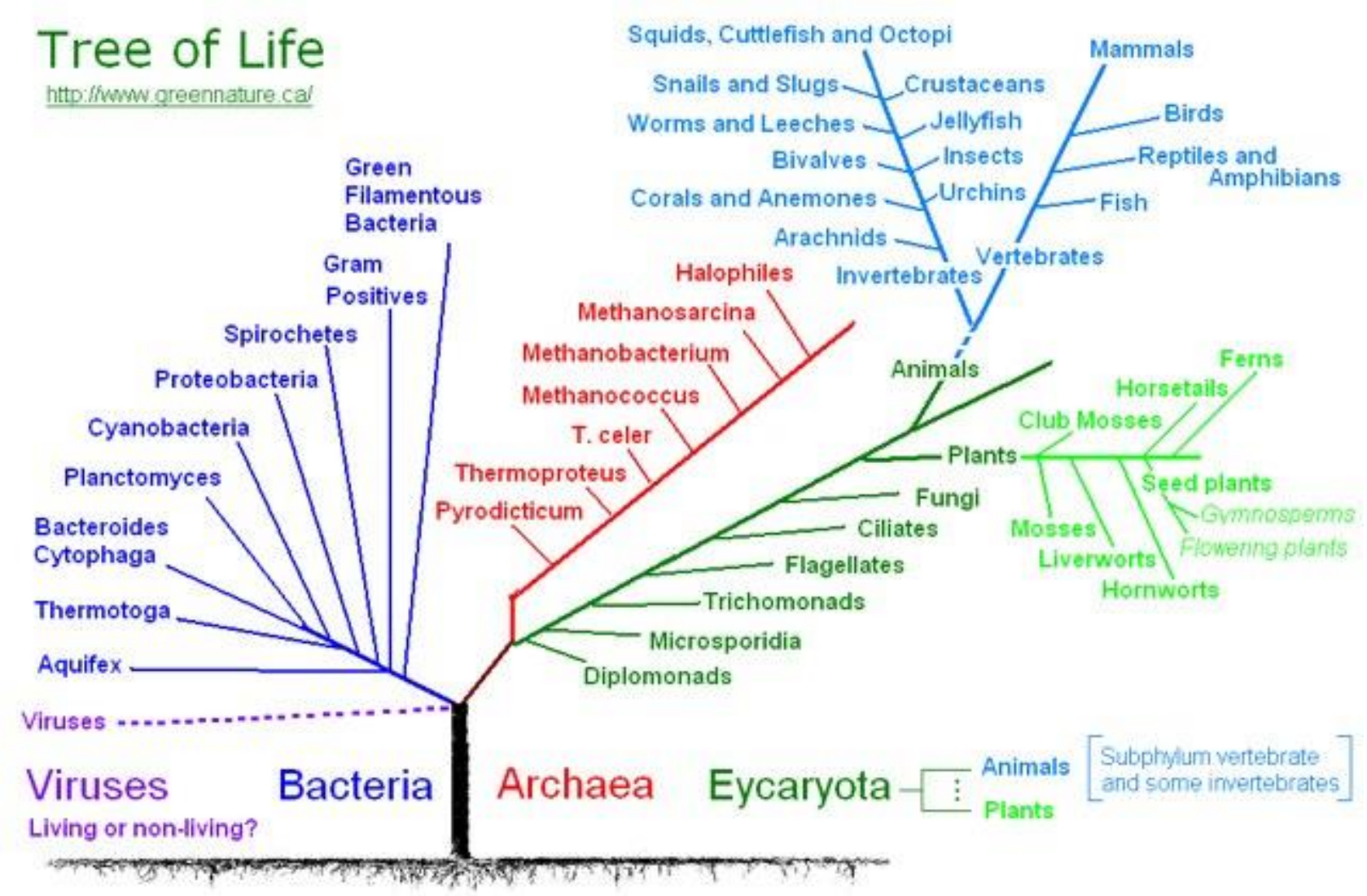


Évolution

- 99% des gènes humains sont conservés chez tous les mammifères
  - Souris 2.1 x10^9 pb versus 2.9 x 10^9 pour l'humain.
  - Environ 95% du matérial génétique partagé.
  - 99% des gènes communs sur un total d'environ 30,000.
- La fonction des gènes est pratiquement la même dans tous les organismes.
- L'innovation fonctionnelle se fait par duplication suivie de mutations
- La plupart des découvertes en biologie moléculaires se font à la lumière de l'évolution.

Évolution

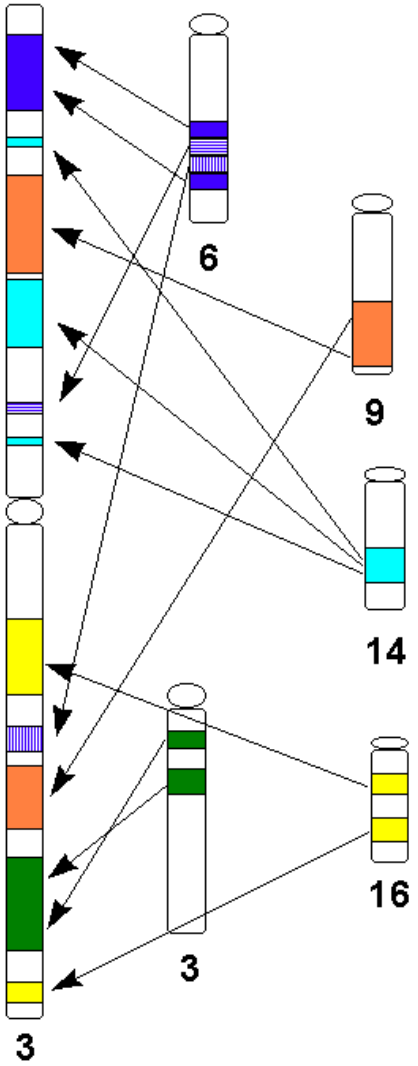
- **<u>Postulat</u>:** Tous les êtres vivants descendent d'un ancêtre commun.
- Tout au long de l'évolution, les gènes accumulent des mutations. Lorsqu'elle sont neutres ou bénéfiques à l'organisme elles sont transmises d'une génération à l'autre
- L'isolement d'une population et l'adaptation à son environnement peut entrainer la création d'une nouvelle espèce.



https://korben.info/il-nous-reste-tant-a-decouvrir.html

Réarrangements génomiques

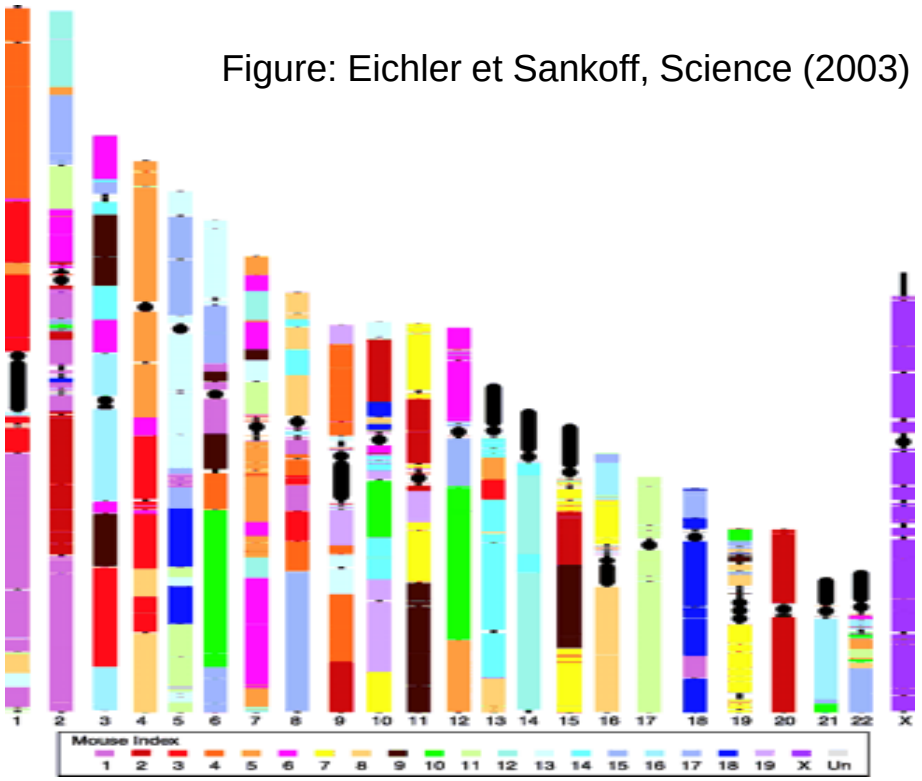
Mapping du chromosome 3 de l'homme avec les chromosomes de la souris.



Problématiques Bio-Informatiques
Génomique évolutive

- Comment les génomes ont-ils évolués par réarrangements, duplications et pertes?
- Permet de comprendre ce qui fait la spécificité d'une espèce: gènes spécifiques, mécanismes évolutifs spécifiques

Figure: Eichler et Sankoff, Science (2003)



Conserved synteny blocks from the mouse genome (MGSCv. 3.0) are overlaid on human chromosomes (April 2003, assembly). All conserved sytenic blocks >10 kb are shown.