BIN702
ALGORITHMES POUR LA BIOINFORMATIQUE

Manuel Lafond

Qu'est-ce que la bioinformatique?

- Utilisation de méthodes informatiques pour approfondir notre compréhension des processus biologiques.

Qu'est-ce que la bioinformatique?

- Utilisation de méthodes informatiques pour approfondir notre compréhension des processus biologiques.
  - **Comparaison** de séquences biologiques
  - Découverte de **nouveaux gènes**
  - **Séquençage** et assemblage de génomes
  - Analyse d'**expression** des gènes
  - Prédiction de **structure des ARN** et des **protéines**
  - Prédiction de **fonctions** des gènes
  - Reconstruction de **l'évolution des espèces**
  - Recherche **d'interactions** entre protéines
  - ...

Qu'est-ce que la bioinformatique?

- Utilisation de méthodes informatiques pour approfondir notre compréhension des processus biologiques.
- Combine des expertises en biologie, en informatique, en mathématiques/statistiques.

Qu'est-ce que la bioinformatique?

- Utilisation de méthodes informatiques pour approfondir notre compréhension des processus biologiques.
- Combine des expertises en biologie, en informatique, en mathématiques/statistiques.
- Applications à l'agriculture, la pharmacologie, la médecine, la virologie, etc.

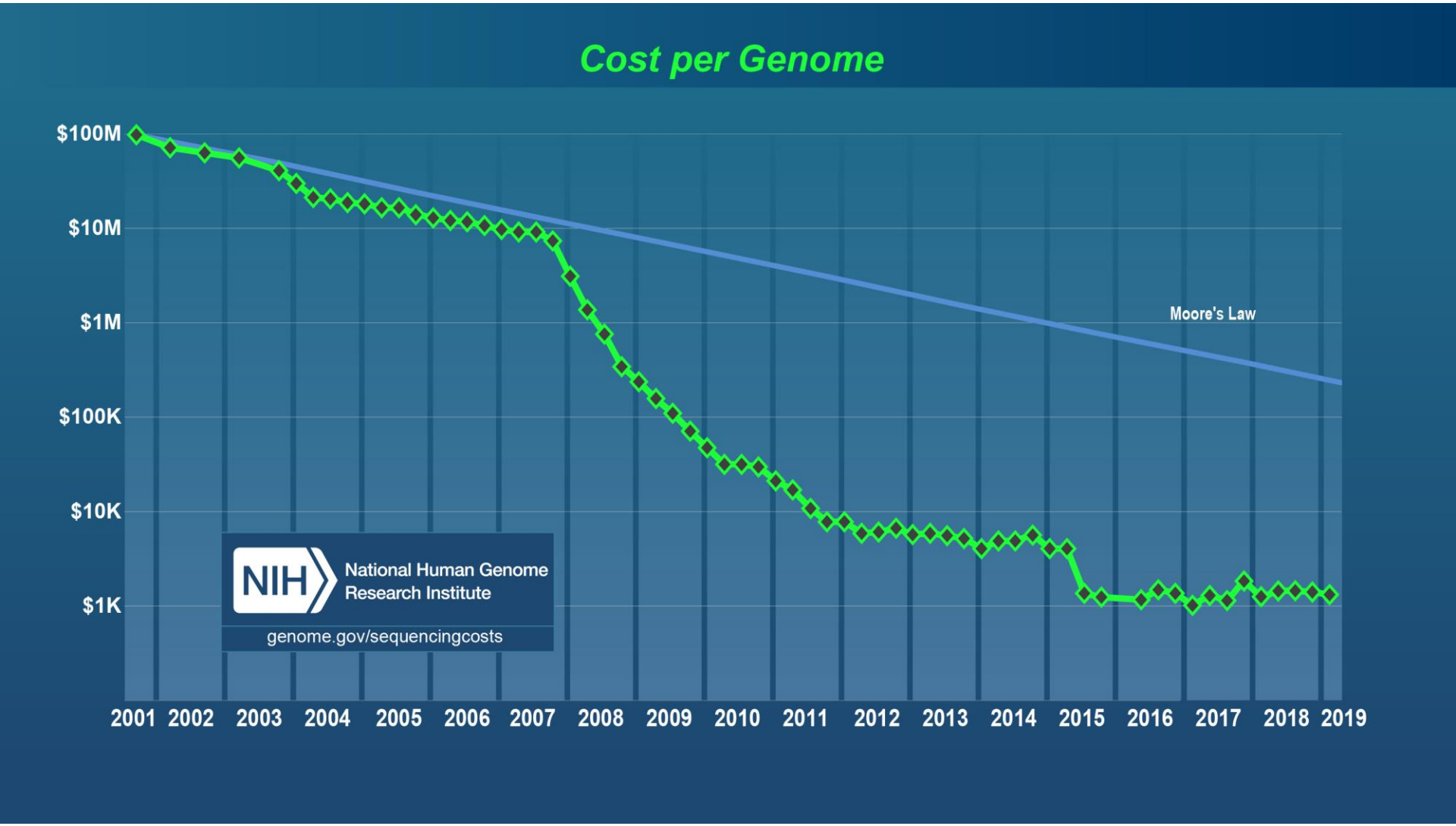
Micro-historique

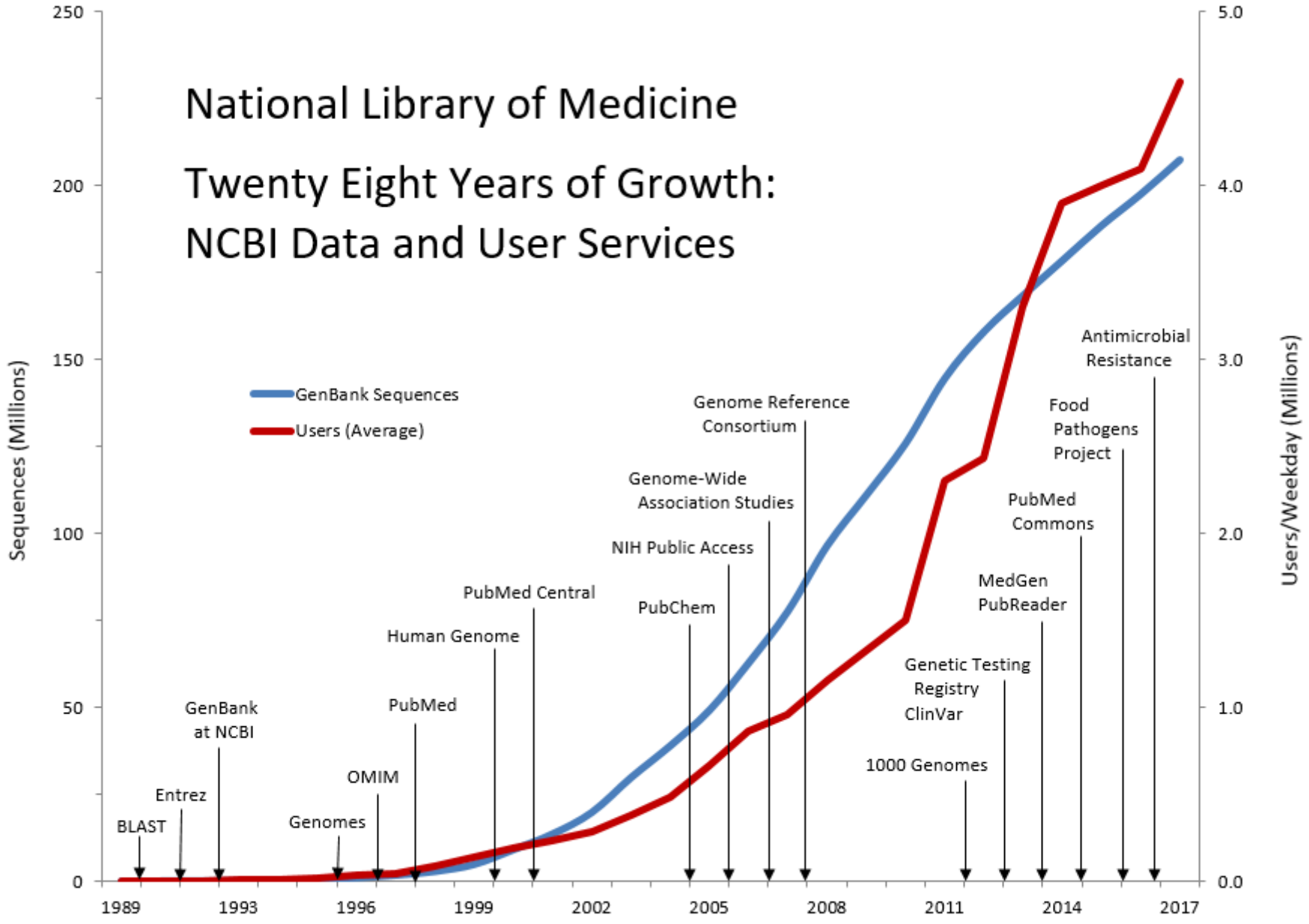
- Années 1950, Frédérick Sanger détermine la séquence des acides aminés de l'insuline.
- En 1965, Margaret Dayhoff: premier atlas de séquences de protéines.
- De 1990 à 2003: *Human Genome Project :* séquençage de l'ADN de l'homme. Coûts estimés: $2.7 milliards.

Micro-historique

- Années 1950, Frédérick Sanger détermine la séquence des acides aminés de l'insuline.
- En 1965, Margaret Dayhoff: premier atlas de séquences de protéines.
- De 1990 à 2003: *Human Genome Project :* séquençage de l'ADN de l'homme. Coûts estimés: $2.7 milliards.

- Black Friday de 2018: Dante Labs peuvent séquencer votre génome pour $200.


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019


National Library of Medicine
Twenty Eight Years of Growth:
NCBI Data and User Services
Sequences (Millions)
Users/Weekday (Millions)
250
200
150
100
50
0
5.0
4.0
3.0
2.0
1.0
0.0
1989
1993
1996
1999
2002
2005
2008
2011
2014
2017
GenBank Sequences
Users (Average)
BLAST
Entrez
GenBank at NCBI
Genomes
OMIM
PubMed
Human Genome
PubMed Central
PubChem
NIH Public Access
Genome-Wide Association Studies
Genome Reference Consortium
1000 Genomes
Genetic Testing Registry ClinVar
MedGen PubReader
PubMed Commons
Food Pathogens Project
Antimicrobial Resistance

Des données, des données

- Croissance **exponentielle** du nombre de séquences de gènes, d'ARN et de protéines.
- Cette croissance ne va probablement pas s'arrêter bientôt
  - de plus en plus d'espèces sont séquencées
  - génomes d'individus différents => multiplication des séquences

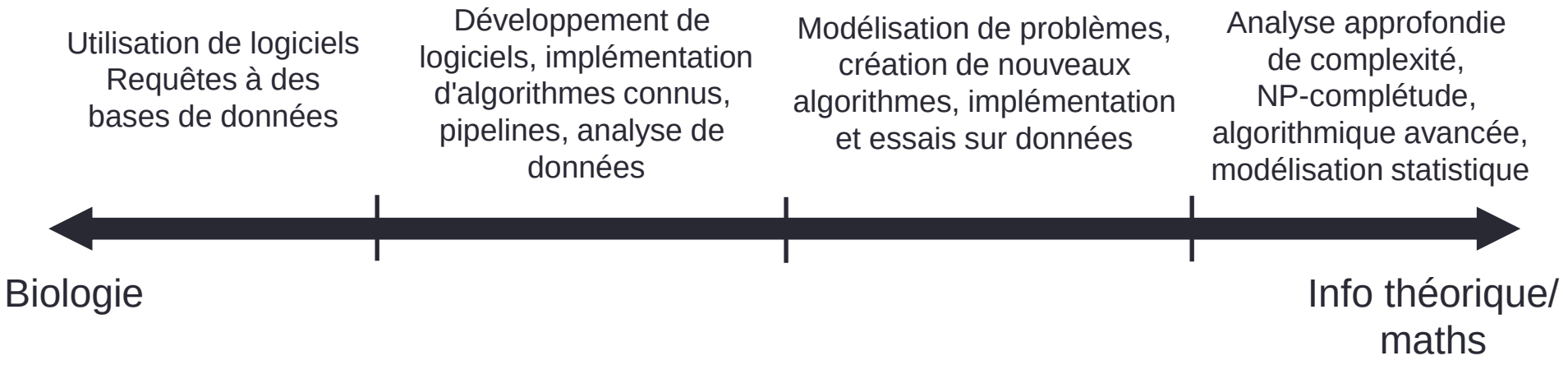
- Uniprot:
  - 167,761,270 séquences sur 1,175,367 espèces
  - 56,302,024,981 acides aminés

- GenBank:
  - 213,383,758 séquences, 329,835,282,370 paires de bases

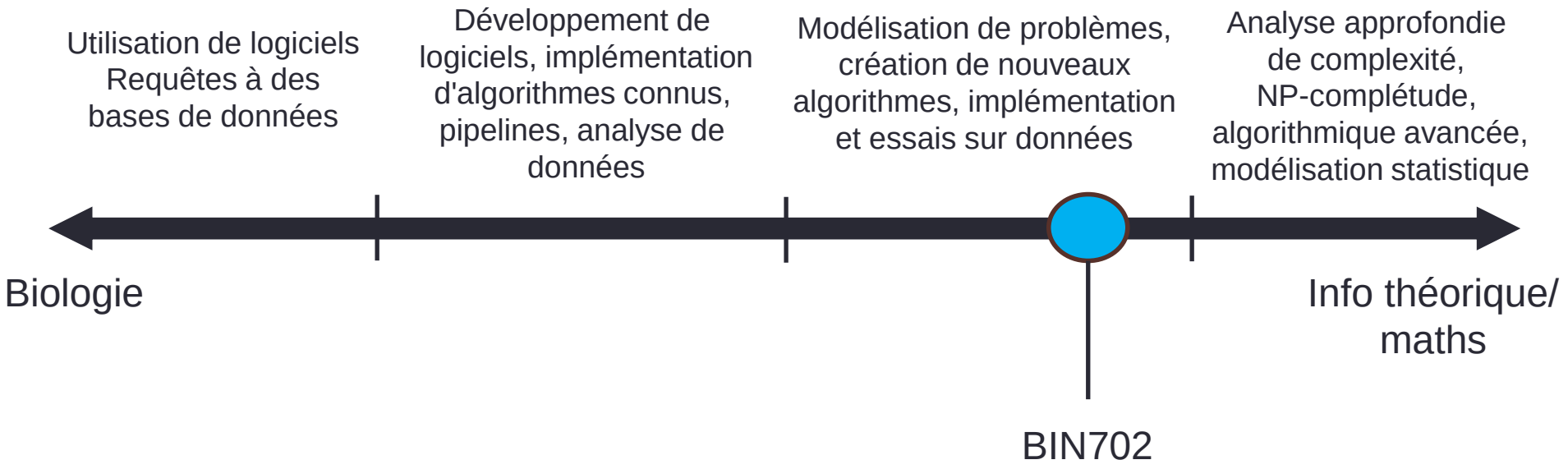
https://www.ebi.ac.uk/uniprot/TrEMBLstats, https://en.wikipedia.org/wiki/GenBank

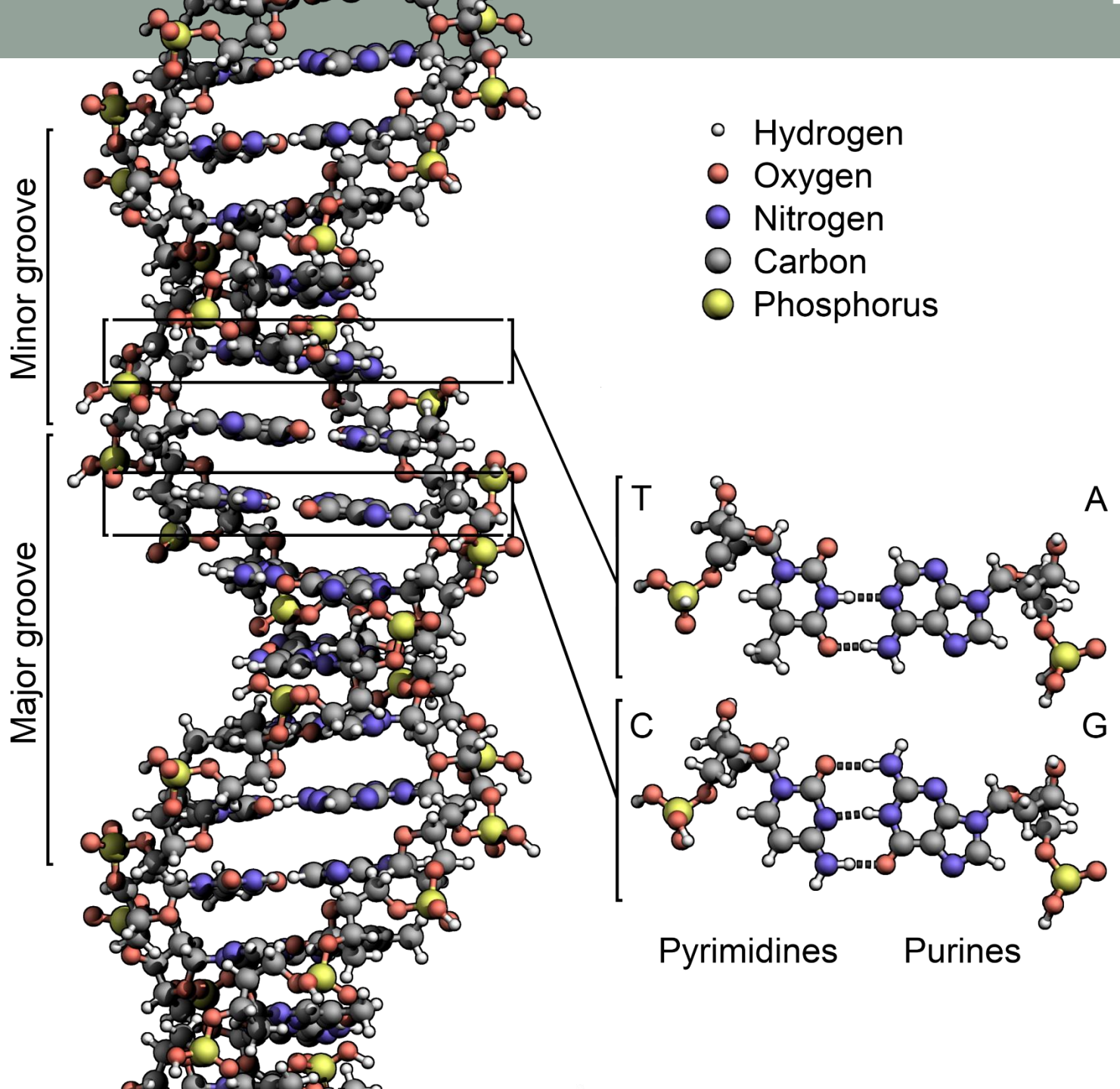
Le spectre de la bioinformatique

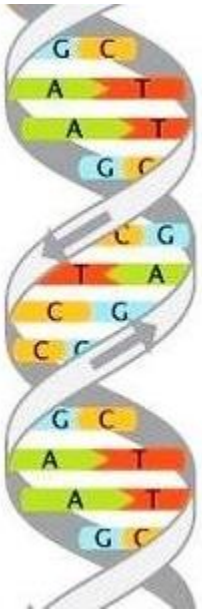
Le spectre de la bioinformatique



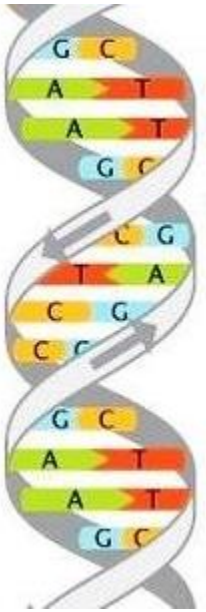
Le spectre de la bioinformatique




Minor groove
Major groove
Hydrogen
Oxygen
Nitrogen
Carbon
Phosphorus
T
A
C
G
Pyrimidines
Purines


G C
A T
A T
G C
C G
T A
C G
C G
G C
A T
A T
G C

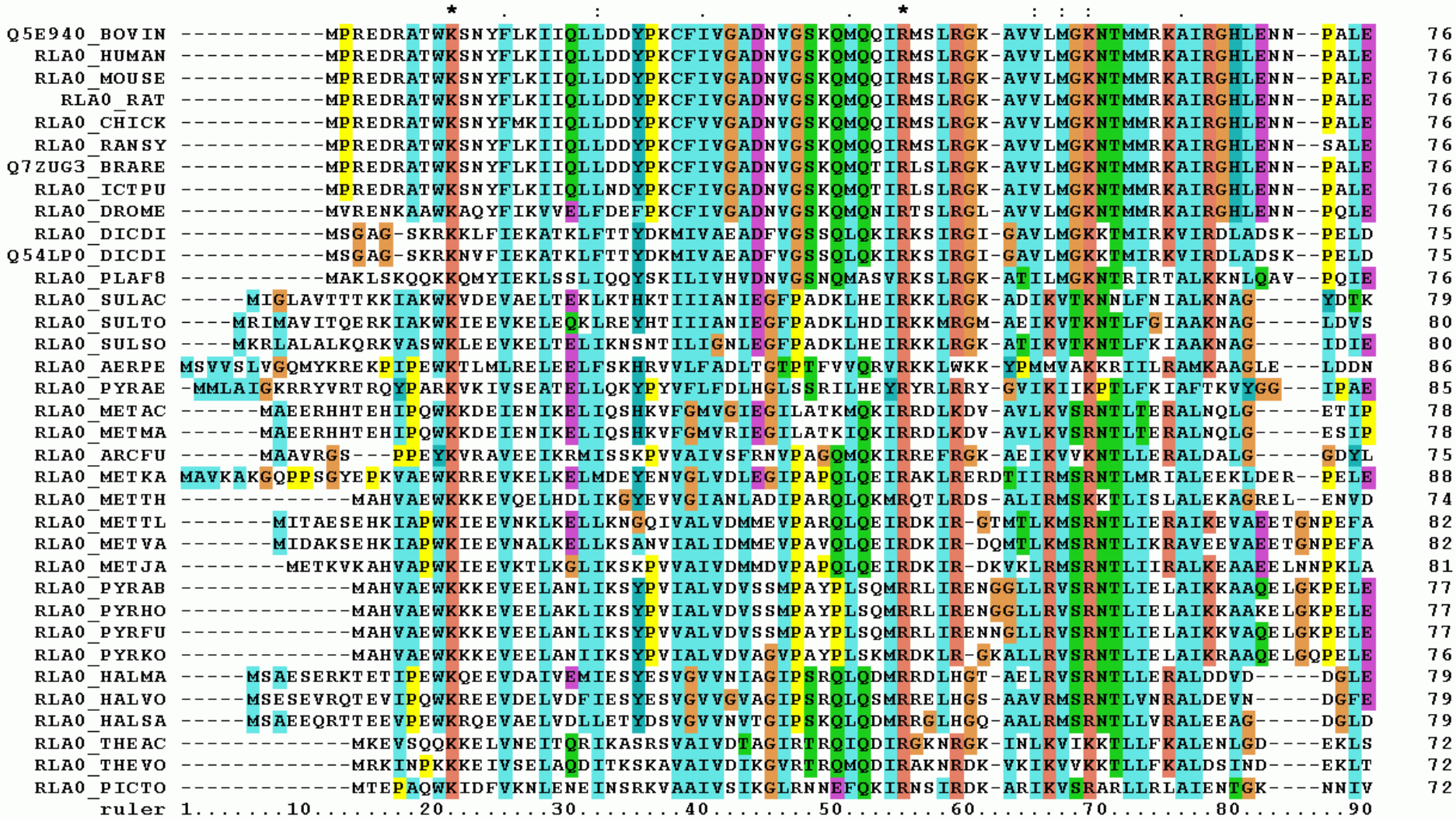
- Pour nous informaticiens, l'ADN n'est qu'une chaîne de caractères sur un alphabet de 4 lettres.
  - (plus tard, nous verrons que les protéines sont des séquences sur un alphabet de taille 20)
- Cette conversion "entité bio" => "concept informatique" est très fréquente.



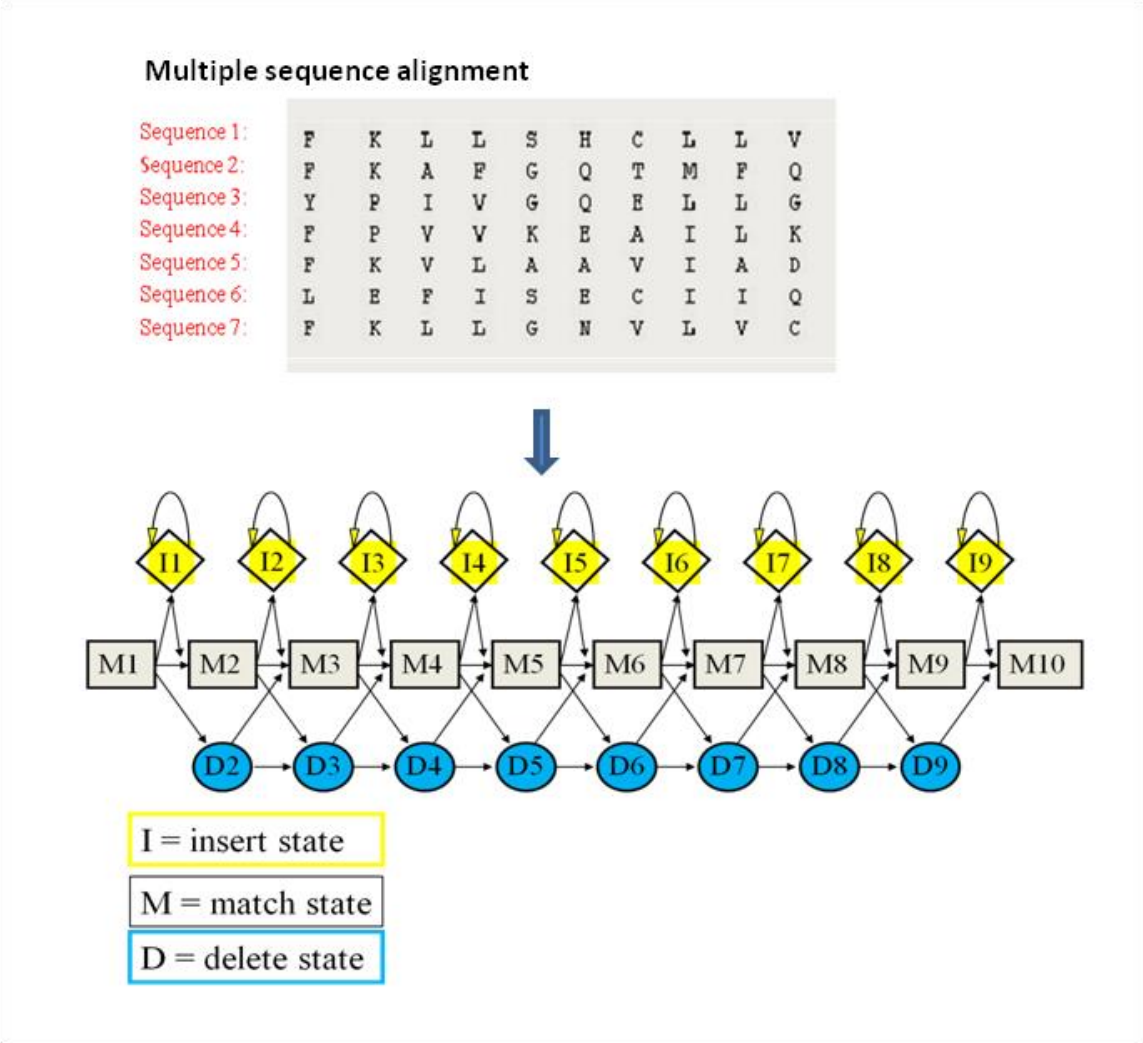
> seq1
GAAGCTCCGAAG...

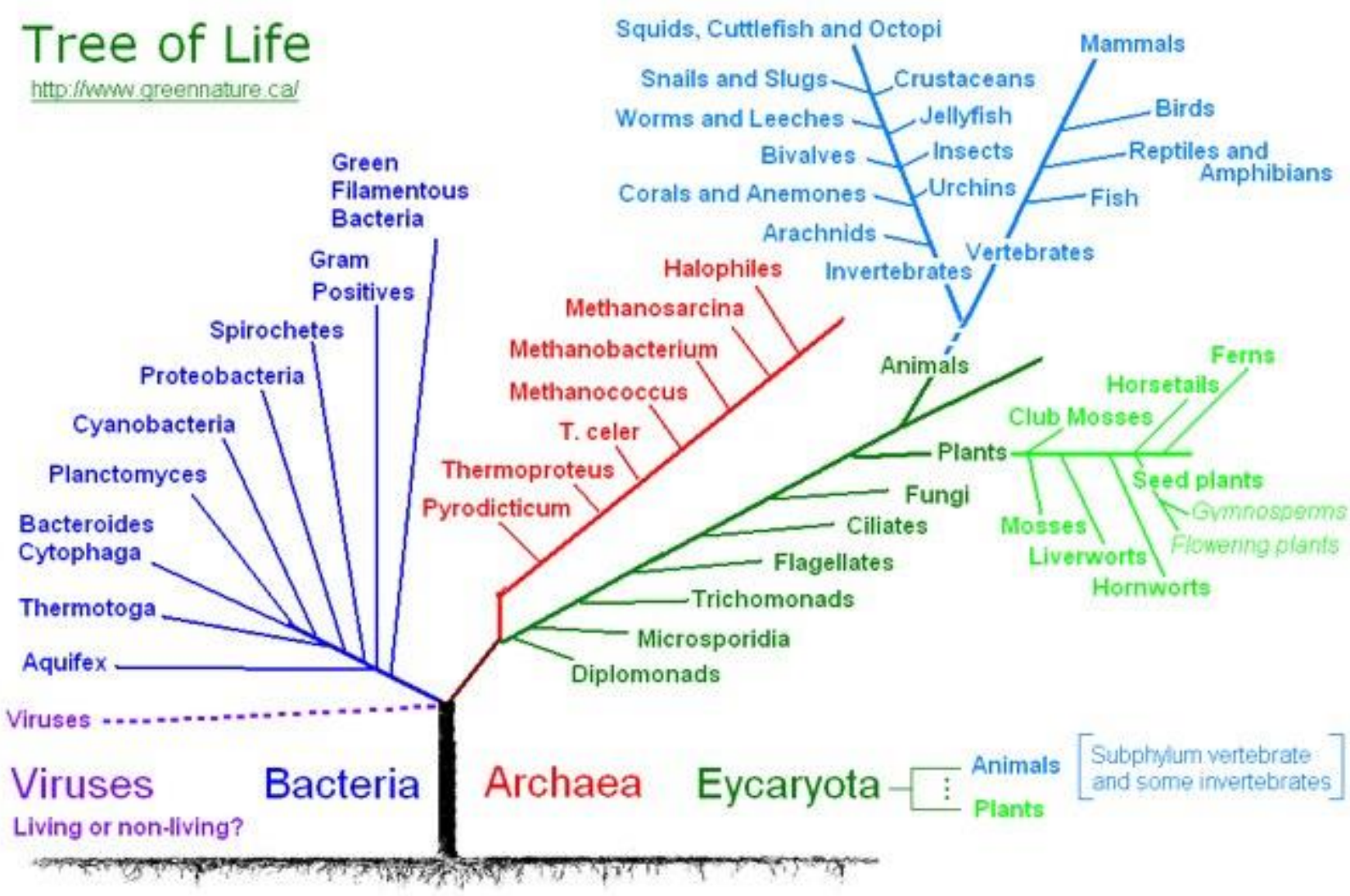
>>agp1
>S.cerevisiae
CTTATGTTAATACTATTATTACTACTATCATTACATTTTAATATCCTGTACATATATATACTCTTTGAAGAAGTTTGGATTTTTCTTTCCCATTATAGATCACTT
TTTTTTTTTACCTTATCTTTTCTGTAAGTTCTTTTTGTGGCTATATTTCCTCTTTGGCGGCTAACTCAAGAAGCAGTAAGGGAATAATTCGTTACACATTA
CAGCTAGCTAAAATCATCAAGAGGCAGGAACACACGAGATATACTGATAATGTTGGAATGTCTAAGCGCTCTGCTGGTTCTCTTTGCCGGTGGTGGT
GGGAGCGTTCTCGCAGCGGTTCAAAGCAAAACTGTGGCAGATCCTAACTTATGTCCAGGTTACAATTCACAACTGATTTCCCCCTTCCTCTCCAGTT
GCGTATGTACAACAAATTTACCGAGGTACTAATATTCTCTAGATTTTATTACTGACAAGATCCATGTGAAGAAACGAAATTTAAGTGAATGCGTGTCCCG
TTATTTTGATGAACAATATGCATTTTGTCGGTCATGCGTTACTGTGAATAATGAAACA
>S.paradoxus
CCTATGTTAATACTATCATTACATTAGTATCCTGTACATATATATACTCTATTCGAAGAACTCTGAATTTTTCCCCCTTATGGATCATCTTTTTCTATTTTTTT
TTTTTTTTGCTAGTCCTTTTTGTGGCTATATTTCCTCTTTGGCGGCTAACTTAAGAAGTAGCATGGGAATAATTTGTTACATATTGCAGTTATCGAAAGT
CATCAAGAAGCTGGAACACACGAAATACACTAATAATGTTGGGATGTTTAAGCGCTCTGCTGGTTCTCTTGGCTGGCGGGGGTAGGAGCATTCTCG
CGGCGGTTCAAAGTAAAACCGTGACAGATCCTAACTTATGCCCAGGTTACAATTCACAACTAATTTCCCCTTTCCTTTCTAGTTGCGTATGTACAATAA
ACTTGCTGAGGTACTAATTTAATCTTTAAATTTATTACTGACAAGATCCACTTGAAGAAACGAGATTTGAATGAATGCATATTCCATTATTTTGATGAACA
ATATGCATTTTGTCGGTCATGTGTTACTGTGAATAACGAAACA
>S.mikatae
CTTACGTTAATACTGTTATTATAATTACTATTACTACATTAGTATCTTGTACATATATATACTCTACTTGAAGAACTTTGAATATTTCTATATCCTGTTATGGTT
TAGAATTTTTTTCTTTGATCATTTCTTTTTTTGGTTAGTCATTTTTGTGGTCATATCGTCTCTTTGGCGGCTAACTTGTGAAATGTCGAGGGAGACACCT
GTTATATATTAGAACTACCAAAATTTATCAAGAAACTGGAACACAAGAAATATACCAGCAATGTTGGGATGTTTAAGCACACTGCTGGTTCTCCTAGCT
GGCGGAGGTAGTTTGATTCTCCCAGCAGTTCAAAGCAAAACCGTGTCAGATCCTAACTTATGTCCAGGTTACAATTCACGAAGAGTTTCTCCCTTCC
TTTCCAGTTGTGTATGTACAATGAACCTACTGATATACTAGTGTAATTTCTAGATTTTATTACTAACAAGAGCCACGTTTAGAAACAAAATCTGAATAAAT
GC
>S.kudriavzevii
CCACAGCCATTGATTGCCCAAGCAGCTTTCTTTTTCTTCTCCCATAACAGCAGCCCAACAGATAAACACCGAAGTTATGCCCAATCGAAATGCATTAA
CGACATCAACAACGATCCTAATACGATGATCTTTCATCCTCGATTCCGTAACCATGTCTTTTCGTTGTTTTTTTTCCATACGAAATCGTTTGCTAAAGAC
AGTTGATAGCATCCGTTGATGTATCTTCTAGAAAAACAATTCTGCAGAAGTGTTTATTTTATCTCCTTTTTTATAGCAAGCGGCACGGCCTCGGCGGCT
AACTCCTTAAGGCCTCGGCGGCTAAAAACGGGTCCTTCCATTTCACACAGACTCCGTACCTCCCACGGCATTTTTTTGCATCTTTTTTTTTTTTCTTG
AGACATAATAGAAGGAAGGGGGAACGCGCGCCATTAGACGGCGCCGCTCAGAAGGCGCCGTCTAAGCGGCACCGATCTGCTGCGAAACATGCAG
CTTTGCCGTTAATGGCGCTCCAGTGGCACCATCGCAGTAATTGCTGGCCTGATCGACCCATAGATAAGCTGCCGCTACCACCAGGCAGTTCTCAAT
GCCCCCCCACGTCCACTTGATTGCTGCTCTTTGGTAGTCACACGCTTCTCGTTTTAGCTGCCGTAAGGCTCTTTTTCTTCTTTTCTCTTGCTTTTTTTA
GTGCCTTCTGTCCATCTCTCCAGCTTGGCAATGGCTATCTTTTTCACTAAATCGAAAATCAGTGCTGATAAGATGGCTCAGGTGCATGGGATAGGTCG
GCATACATCTGTTTTACGCTATCTTCTCCTCGTCCGCTAGCGCGACTCACTTATTCTGTAATAATTTCTTTTTTTTTCAAGAGTAACCTTTGTATTTTACA
TACGTATATAAGTGGTTTGTTTCTGGATTTGCTGCTTTACAAGGGTGAGGAATAGCATTCGTCATGTATTTTTTTTTTAGAACGCAAGTAGTATAGACTA
AACAAAAAAGCTTCGCATA
>>spo16
>S.cerevisiae
GGATCGTATATTTACCAGTATGCTACCTCCTTGAAATCTTGCCCAATAACCTAATGACTTCCATTGAATATACAATATAGAACATTGTGATGTATTTATTTA
TTGGCAAAAAATGTGACGCGAAAAACAACATTAGATGAGTGAAAAAGAAAAGTCACATCTGATATAGCTTACCCTGTTAACGCGTGAAAAGTGGGCG
GCTAAAACCGAGAAAATACGAAATAGTGTAAAGGTTCAAGAGAATAAACTTCTGGAACAACATTTCTGGAATGGTTCAATAGAA
>S.paradoxus
GGAGTATATATTGACCAGTACACTACCTCCTTGGAATCTTATCCGATGACCCAATTACTTACATTGAATACAAACTATAGAACATCTATGATGTATTTATTT
ATTGGCAAAAAAATGTGACGCGAAAAACAGAACATTAGGTGAGTGAAAAAGAAAACTCAGATCTCATGTCCTTACCCTGTTAACGCGTAAAAAGTGG
GCGGCTAAAACCGGGAAAATATGAAAGAGTAAAGGTTCAAGAGAATTAACTTCTGGGAGAACGTTTCTCGAATGCTTAACTGGAC
>S.mikatae
AGCTTATATCTCGACCAGAACACTATTTCCTGGAAATCTTACTTGGAGTGTTAATAACCTTGCATTGAGTATAAACTATAGCATGTATATGATGTATTTATT
TATTAGCAGAGAAATGTGACGCGAAAAATCCAACACTTCAGTGAGTCAAAAATAAAACCACGTTCTAAATTTCTTACCCTGTTGACGCGTGAATAAATG
TGGGCGGCTAAAACCAAAAATCTTGAAACAGGATAAAGAGAAGATGAGCTACAAGTATAAAGCTACAGAAAACCGCTTGTGAAATAACCGGAATATAC
>S.kudriavzevii

Alignement multiple de séquences



Modélisation par Modèle de Markov




Tree of Life
http://www.greennature.ca/
Squids, Cuttlefish and Octopi
Snails and Slugs
Worms and Leeches
Bivalves
Corals and Anemones
Arachnids
Crustaceans
Jellyfish
Insects
Urchins
Invertebrates
Mammals
Birds
Reptiles and Amphibians
Fish
Vertebrates
Green Filamentous Bacteria
Gram Positives
Spirochetes
Proteobacteria
Cyanobacteria
Planctomyces
Bacteroides Cytophaga
Thermotoga
Aquifex
Halophiles
Methanosarcina
Methanobacterium
Methanococcus
T. celer
Thermoproteus
Pyrodicticum
Animals
Fungi
Ciliates
Flagellates
Trichomonads
Microsporidia
Diplomonads
Plants
Ferns
Horsetails
Club Mosses
Seed plants
Gymnosperms
Flowering plants
Mosses
Liverworts
Hornworts
Viruses
Viruses
Living or non-living?
Bacteria
Archaea
Eycaryota
Animals
Plants
Subphylum vertebrate and some invertebrates

Qu'est-ce que BIN702?

- Nous nous intéressons aux fondements des méthodes utilisées par les biologistes et bioinformaticien(ne)s.
  - Quels algorithmes sont implémentés? [IFT436]
  - Quelle est leur complexité? [IFT339/IFT436]
  - Quelles structures de données sont utilisées? [IFT339]

Qu'est-ce que BIN702?

- Nous nous intéressons aux fondements des méthodes utilisées par les biologistes et bioinformaticien(ne)s.
  - Quels algorithmes sont implémentés? [IFT436]
    - **Préalable souhaité**: concepts de base en algorithmique (approches diviser-pour-régner, programmation dynamique, algorithmes gloutons, graphes, ...)
  - Quelle est leur complexité? [IFT339/IFT436]
    - **Préalable requis**: savoir analyser le temps d'un algorithme en notation O
  - Quelles structures de données sont utilisées? [IFT339]
    - **Préalable requis**: concepts de base en structures de données (listes, tableaux, arbres binaires de recherche, tables de hachage)

Qu'est-ce que BIN702 n'est pas?

Qu'est-ce que BIN702 n'est pas?

- Nous n'utilisons **pas d'outils concrets** - ce sont les méthodes derrière ces outils qui nous intéressent.
  - des algorithmes des années 1970 sont encore très utilisés
  - très peu de logiciels ont une durée de vie de plus de 10 ans
  - compétences utiles dans d'autres domaines
- Nous ne discutons **pas de sources de données** biologiques (ex: Uniprot, Ensembl, Cancer Genome Atlas, ...).
- Nous n'entrons pas dans des **détails d'implémentation**.
- Nous n'utilisons **pas l'IA ni l'apprentissage-machine** appliqué à la bioinformatique.

- Quelques exceptions peuvent survenir.
- Vous pouvez explorer le côté pratique au cours de votre **projet**.

Projet en BIN702

- Vous avez un maximum de liberté!
  - Une banque d'idée sera proposée
  - Si vous amenez votre propre idée, le correcteur sera content :)
- Projet pratique
  - Implémenter un algorithme qui n'existe qu'en pseudo-code
  - Améliorer les performances d'un algorithme connu (ou du moins, essayer)
  - Tester plusieurs logiciels accomplissant la même tâche (*benchmarking)*
  - Développer un algo + programme pour un nouveau problème
  - Développer un *pipeline* pour une tâche commune en bioinformatique

Projet en BIN702

- Vous avez un maximum de liberté!
  - Une banque d'idée sera proposée
  - Si vous amenez votre propre idée, le correcteur sera content :)
- Projet théorique
  - Résoudre un problème ouvert
  - Revue approfondie de litérature sur un sujet donné
  - Améliorer la complexité à un problème déjà étudié
  - Montrer qu'un problème est NP-complet

Projet en BIN702

- Rapport de projet
  - Mise en contexte - où se place votre projet par rapport à la litérature? Quelle est sa pertinence?
  - Qu'est-ce qui rend votre projet original?
  - Quel est son niveau de difficulté?
  - Quels défis devez-vous surmonter?
  - Quelles solutions avez-vous apportées?
  - Projets pratiques: comment évaluez-vous votre approche? Quels résultats attendiez-vous, et qu'avez-vous obtenu?
  - Projets théoriques: si vous n'avez pas résolu votre problème, quelles pistes avez-vous tentées? Pourquoi ont-elles échoué?