Introduction à la Bio-Informatique IFT3295/IFT6291/BIN6000

Nadia El-Mabrouk

DIRO, Université de Montréal

Qu'est-ce que la Bio-informatique?

Informatique et recherche opérationnelle

Du génome de la banane à celui de l'humain

Apparue dans les années 60, après que les biologistes eurent découvert comment séquencer l'ADN et les protéines, la bio-informatique permet aujourd'hui de documenter l'histoire de l'évolution de la vie sur Terre et de comprendre les processus d'adaptation. PHOTO : ISTOCKPHOTO.

Les 50 ans de contributions du bio-informaticien David Sankoff sont soulignés au colloque MAGE

« Gérer, traiter, comparer et archiver la quantité phénoménale de données issues de la biologie, de la biochimie et de l'écologie serait impossible sans la bio-informatique, soutient Nadia El-Mabrouk, professeure au Département d'informatique et de recherche opérationnelle (DIRO) de l'Université de Montréal. Notre connaissance de la fonction des gènes, des relations évolutives entre les espèces et de leurs spécificités génétiques provient essentiellement de cette jeune discipline qui se trouve au carrefour de l'informatique, des mathématiques et des sciences de la vie. »

Pour la chercheuse spécialisée en biologie computationnelle, il n'est donc pas étonnant que la majorité des recherches en bio-informatique soient menées par des mathématiciens, des informaticiens, des statisticiens et des biologistes. Une quarantaine d'entre eux, dont David Sankoff, mondialement reconnu comme l'un des piliers de la biologie computationnelle, étaient présents au colloque MAGE (Modèles et algorithmes pour la génomique évolutive), qui a eu lieu du 23 au 26 aout à l'hôtel Château-Bromont, dans les Cantons-de-l'Est.

Organisé par le Centre de recherches mathématiques (CRM) de l'UdeM et le DIRO, en collaboration avec les universités Claude-Bernard Lyon 1 (UCLB) et Simon Fraser, ce colloque international a réuni 38 conférenciers qui travaillent dans différentes sphères de la biologie computationnelle, un volet plus théorique de la bio-informatique qui vise l'élaboration d'algorithmes efficaces afin de permettre la résolution d'un problème biologique particulier, par exemple la désignation de protéines cibles pour la mise au point de médicaments.

Nadia El-Mabrouk

Venant d'une quinzaine d'universités des quatre coins du monde, ces chercheurs ont présenté des communications qui s'articulent autour de la génomique évolutive. « Les sujets abordés ont concerné l'étude évolutive d'une grande variété de génomes, de la banane à l'être humain, et les diverses approches de la bio-informatique », souligne la professeure d'origine tunisienne, l'un des maitres d'œuvre du colloque avec ses collègues Éric Tannier (UCLB) et Cédric Chauve (Simon Fraser).

« La rencontre a permis de dresser le bilan des acquis, d'explorer des directions futures et, surtout, de créer une synergie entre les chercheurs et les étudiants permettant de faire émerger de nouvelles collaborations et de concevoir des méthodes et des outils qui tiennent compte de toute la complexité des données biologiques. »

50 ans de contributions scientifiques

Les chercheurs en sciences de la vie produisent une quantité croissante de nouvelles données portant sur les génomes, les biomolécules, les organismes, leurs interactions et leur évolution. Sur NCBI, la bible des scientifiques qui s'intéressent à la génomique comparative, on trouve quelque 10 millions de séquences de protéines, des données biologiques concernant au-delà de 10 000 espèces ainsi que les séquences génomiques complètes de plus de 1000 espèces.

Pour Nadia El-Mabrouk, la conception d'approches informatiques pour la manipulation, l'archivage, la visualisation et l'analyse de ces données revêt une importance fondamentale. « Chaque problème, chaque type de mutation nécessite une modélisation spécifique et donne lieu à des développements algorithmiques, statistiques et mathématiques différents », dit-elle.

Par la qualité des invités, le colloque MAGE constituera un point de repère pour les études en génomique comparative, estime Nadia El-Mabrouk. Celle-ci relate la contribution essentielle de David Sankoff, dont le premier article scientifique a été publié en 1963. Professeur au Département de mathématiques et de statistique de l'UdeM de 1984 à 2002, il a favorisé l'essor de la biologie computationnelle au CRM avec Robert Cedergren, du Département de biochimie. M. Sankoff est présentement titulaire d'une chaire de recherche du Canada en génomique mathématique à l'Université d'Ottawa.

Dominique Nancy

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.
- De plus en plus, la bioinformatique est développée dans un but d'application à l'agriculture, la pharmacologie, la médecine.
- Discipline qui évolue en fonction des nouveaux problèmes posés par la biologie.

Qu'est-ce que la Bio-informatique?

- **Biology "computationnelle"**:
  - Développement d'algorithmes efficaces permettant de résoudre un problème biologique spécifique.
  - Méthodologie générale:
    - Définir un modèle d'évolution;
    - Formaliser le problème;
    - 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

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.
- **Bioinformatics**: 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.

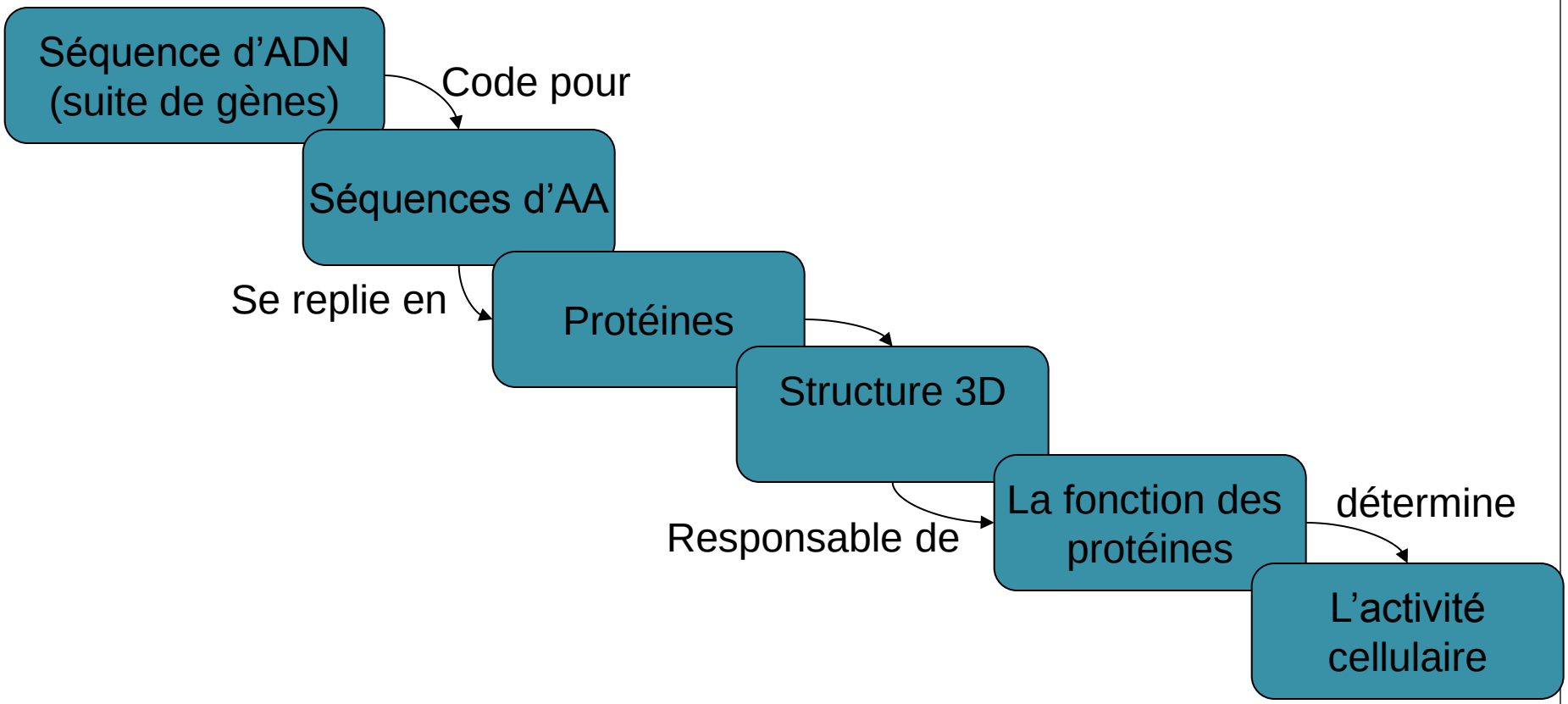
Origine

- 1953: Structure en double hélice de l'ADN
- 1963: Découverte du code génétique, de l'ARNm.
- Bioinformatique: Apparue dans les années 1960, après que les biologistes aient découverts comment séquencer de l'ADN et des protéines.
  - Margaret O. Dayhoff: répertoire de protéines,
  - 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.


HONORING 50 YEARS OF
Models and Algorithms for Genome Evolution
MAGE
AUGUST 23–26 2013
BY DAVID SANKOFF
Bromont, Québec
ORGANIZERS
Cédric Chauve (Simon Fraser University, Canada)
Dannie Durand (Carnegie Mellon University, USA)
Nadia El-Mabrouk (Université de Montréal, Canada)
Laxmi Parida (IBM Research, USA)
Eric Tannier (INRIA, Université de Lyon, France)
COMMITTEE
Anne Bergeron (Université du Québec à Montréal, Canada)
Mathieu Blanchette (McGill University, Canada)
Guillaume Bourque (McGill University, Canada)
David Bryant (University of Otago, New Zealand)
Vincent Ferretti (Ontario Inst. For Cancer Research, Canada)
Anthony Labarre (Université Paris-Est Marne-la-Vallée, France)
Manuel Lafond (Université de Montréal, Canada)
Krister Swenson (Université de Montréal, Canada)
Marcel Turcotte (University of Ottawa, Canada)
Tandy Warnow (The University of Texas at Austin, USA)
Chunfang Zheng (University of Ottawa, Canada)
Binhai Zhu (Montana State University, USA)
SPEAKERS
Victor A. Albert (University of Buffalo, USA)
Marilia Braga (Inmetro, Brazil)
Miklos Csuros (Université de Montréal, Canada)
Joe Felsenstein (University of Washington, USA)
Jotun Hein (University of Oxford, UK)
Tao Jiang (UC Riverside, USA)
John Kececioglu (University of Arizona, USA)
Jens Lagergren (Science for Life Laboratory, Sweden)
Eric Lyons (University of Arizona, USA)
Aoife McLysaght (University of Dublin, Ireland)
Joao Meidanis (University of Campinas, Brazil)
Bernard Moret (EPFL, Switzerland)
Gene Myers (Max Planck Inst. CBG, Germany)
Joseph H. Nadeau (Institute for Systems Biology, USA)
Pavel Pevzner (UCSD, USA)
David Sankoff (University of Ottawa, Canada)
Ron Shamir (University of Tel Aviv, Israel)
Jens Stoye (University of Bielefeld, Germany)
Olga Troyanskaya (Princeton University, USA)
Sophia Yancopoulos (Feinstein Inst. for Medical Research, USA)
Liqing Zhang (Virginia Tech, USA)
Louxin Zhang (National University of Singapore)
A volume of the Computational Biology series of Springer untitled "Models and Algorithms for Genome Evolution", gathering contributions from MAGE participants, will be published following the conference (editors: Cédric Chauve, Nadia El-Mabrouk and Éric Tannier).
www.crm.umontreal.ca/2013/MAGE13

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
  - Par exemple, le génome humain (3.2 milliard de bases)
- 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.

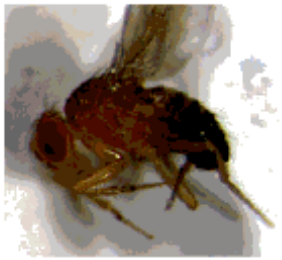
Qu'est-ce que la Bioinformatique?
Défis

- 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.
- Présentement dans RefSeq (NCBI):
  - 10.640.515 protéines
  - 10.728 espèces
- Plus de 1200 génomes de procaryotes et 460 génomes d'eucaryotes complètement séquencés.

Whole Genomes



Drosophila

C. elegans

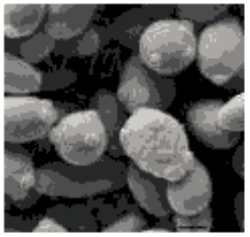
Rat

Human

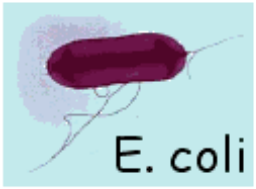
Mouse

Rice

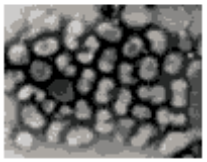
Mosquito



Yeast



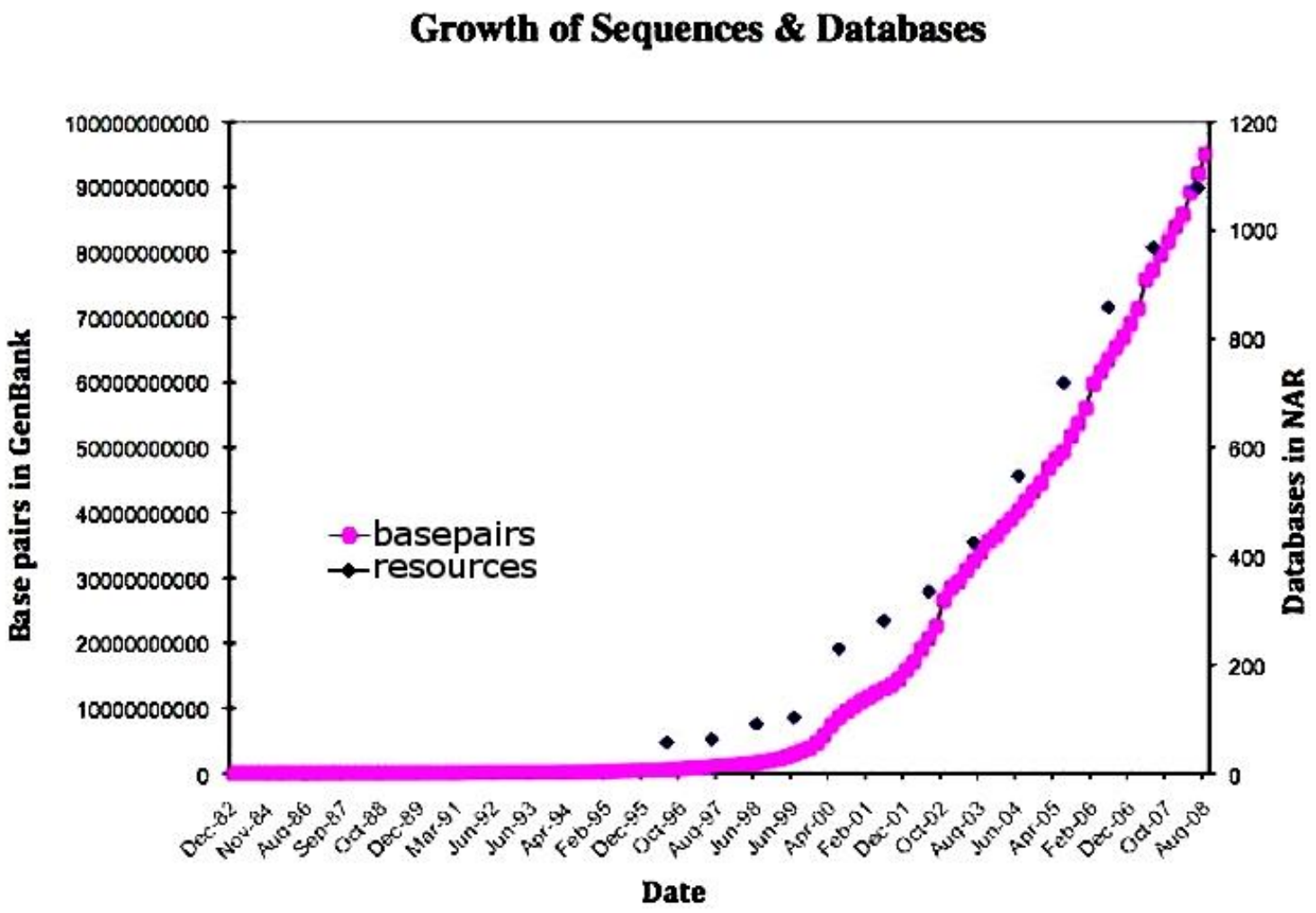
E. coli



H. influenza

Arabidopsis

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



GenBank release notes: ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt and list size obtained from NAR archives: http://nar.oxfordjournals.org/archive/

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

Qu'est-ce que la Bioinformatique?

Information manipulée

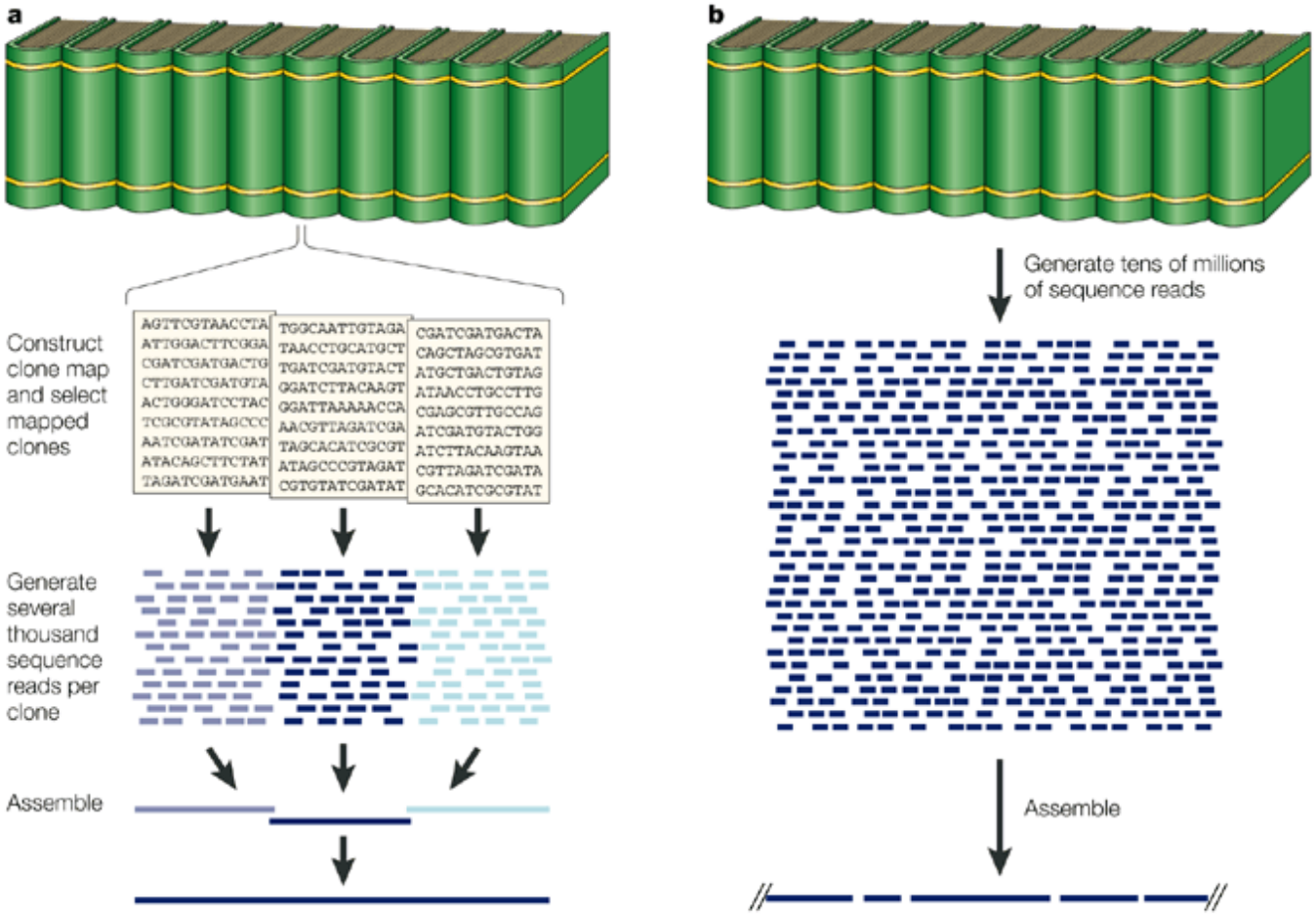
- ADN (Génome)
  - Séquences de nucléotides
  - Séquence de gènes
  - Banques de données
- ARN (Transcriptome)
  - Séquence
  - Structure
- Protéines (Protéome)
  - Séquence
  - Structure
  - Réseaux d'intéraction

ADN - Séquençage

- Action de déterminer la suite de nucléotides d'un fragment d'ADN.
- Taille fragment: 100nt - 10^9nt (génome)
- 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 »

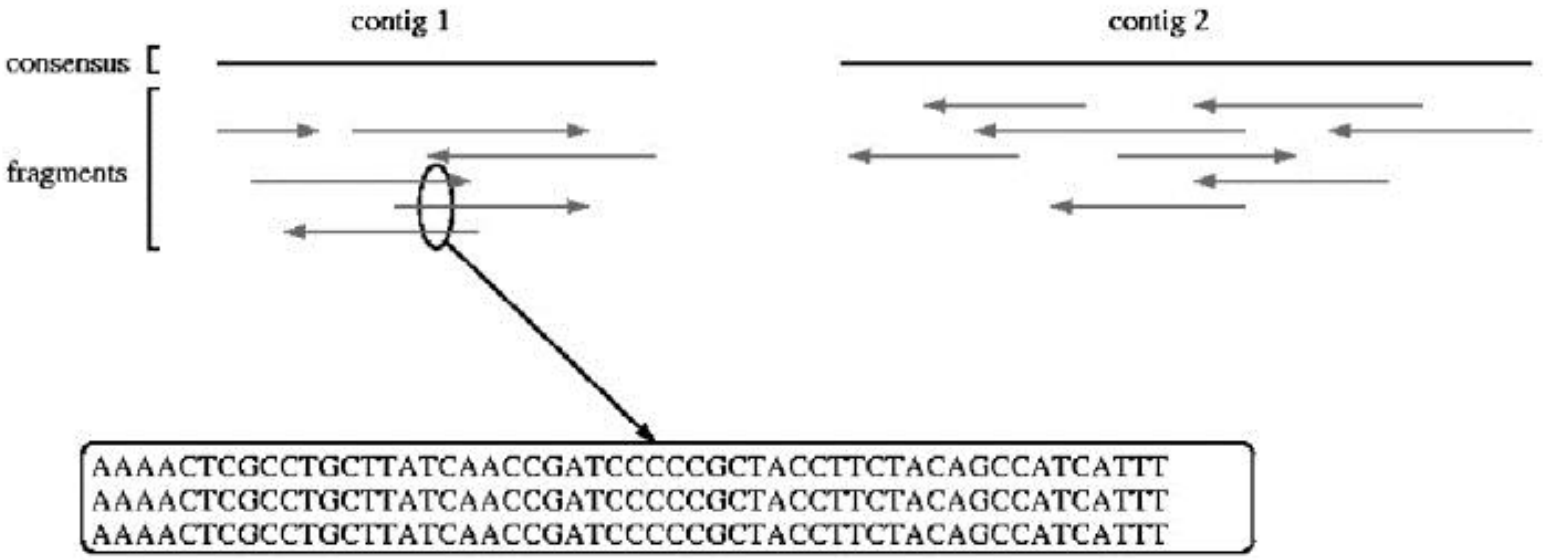


http://www.nature.com/nrg/journal/v2/n8/fig_tab/nrg0801_573a_F6.html

ADN - Séquençage

Assemblage

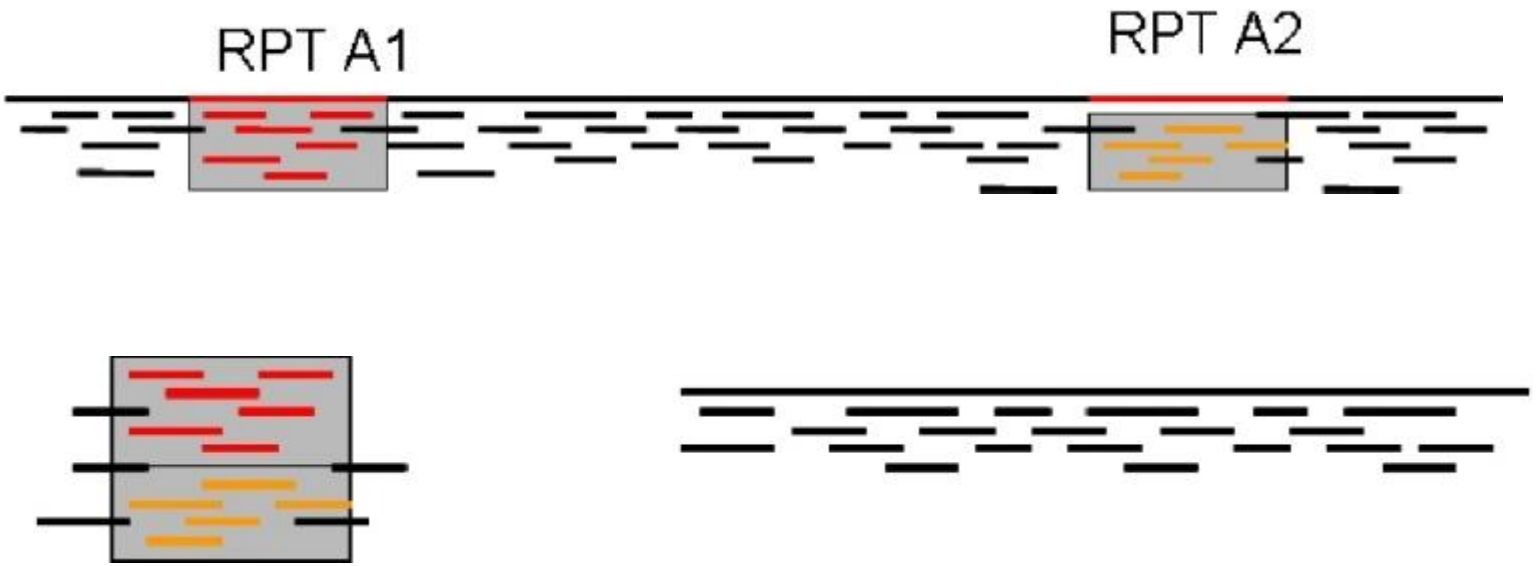
- Problème: Reconstruire la séquence cible à partir des fragments obtenus.
- Difficultés supplémentaires: présence d'erreurs, régions répétées



ADN - Séquençage

Assemblage

Difficultés causées par les régions répétées:



http://www.cbcb.umd.edu/research/assembly_primer.shtml#challenges

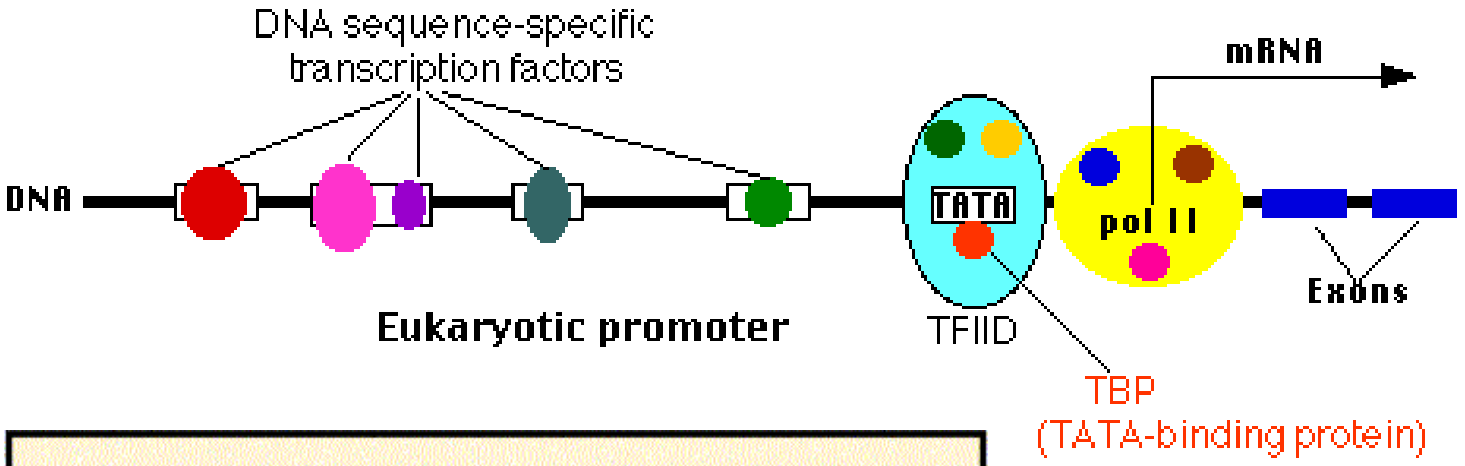
ADN - Annotation

- Une Séquence d'ADN:
  - Est-ce que cette séquence a déja 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

Structure des gènes eucaryotes



CTCCGCGTATTGCTGTCACCCCGCTGCCCTGCATCCGTTTGTT
CGTCGCCGTTTGTCATTTGCCCTGCGCTCATGCCCCGCAC
CTCGCCGCCCGCCCCATTTCCTCATGCCCCGCACCCGCGC
TACTGTCGTCCATTTGCCCTGCGCTCATGCCCCGCACCTCG
TTTGCTTGCTCCATTTGCCTCATGCCCCGCACTGCCGCTCA
CTGTCGTCCATTTGCCCTGCGCTCACGCCCTGC
GCTCGTCTTACTCCGCCGCCCTGCCGTCGTTCATGCCCCG
CCGTCGTTCATGCCCCGCTGTATTGTTTGCCCTGCGCCCAC
CTGCTTCGTTTGTCATGCCCCGCACGCTGCTCGTGCCCC

GENE COLOR KEY

- Untranslated region
- TATT = promoter sequence
- ATG = initiation codon
- Open reading frame

ADN - Annotation

Concept de similarité

Les séquences codant pour des protéines importantes ou impliquées dans la régulation des gènes sont conservées au cours de l'évolution.

Étant donnée une nouvelle séquence de gène, prédire la structure de la protéine codée par cette séquence.

Méthode basée sur la similarité de séquence:

- Trouver une séquence connue ayant une similarité significative avec la nouvelle séquence;
- Utiliser la structure connue pour prédire la nouvelle.

ADN - Annotation

Concept de similarité

Comment aligner deux séquences?

- Comment aligner cette séquence:

`gattcagacctagct`

- Avec cette séquence:

`gtcagatcct`

ADN - Annotation

Concept de similarité

Réponses possibles:

1. Sans insertions/suppressions. Distance de Hamming:

```
gattcagacctagct
 gtcagatcct
```

2. Minimum insertions, suppressions, substitutions: distance d'édition.

```
gattcaga-cctagct
g-t-cagatcct----
```

3. Minimiser gaps+subs :

```
gattcaga-cctagct
g--tcagatcct----
```

ADN - Annotation

Concept de similarité

Différents types d'alignements de deux séquences :

Alignement Global:

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

Alignement Local:

CAGCACTT – G G A T TCTCGG
TAGTTTA G G - T GGCAT

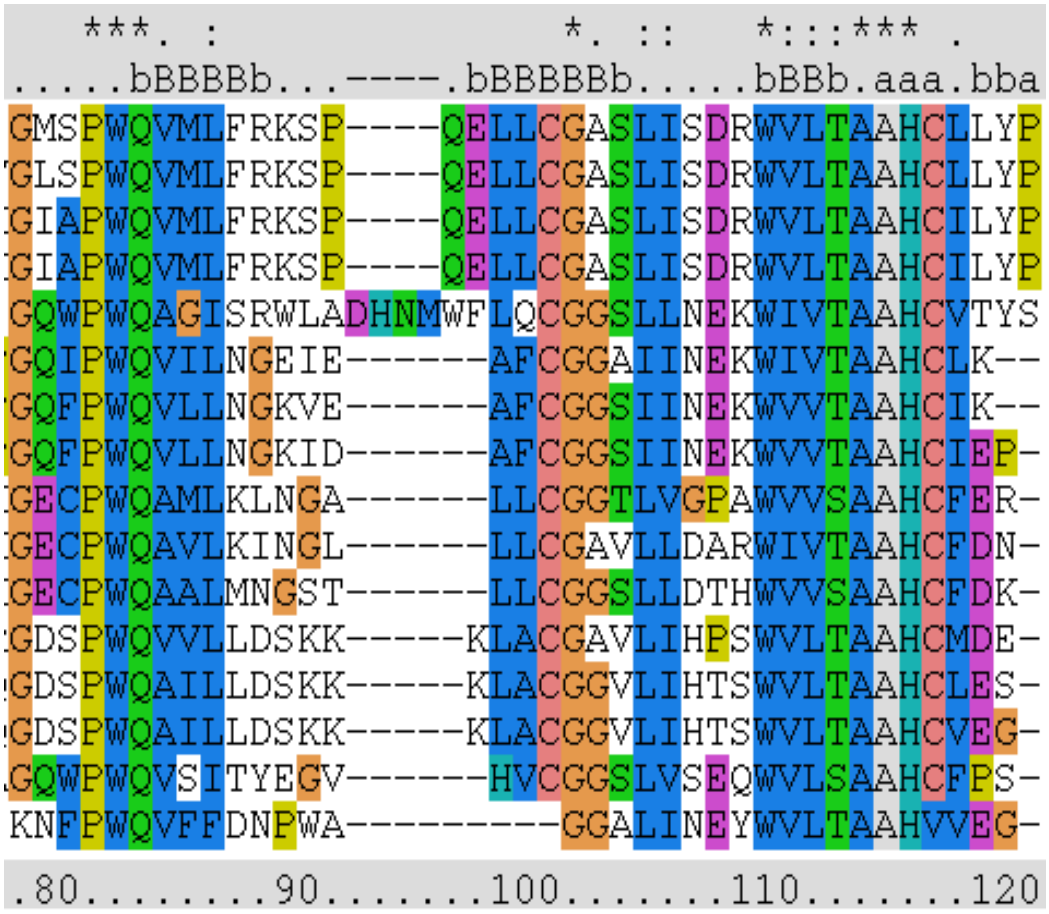
Recherche:

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

ADN - Annotation

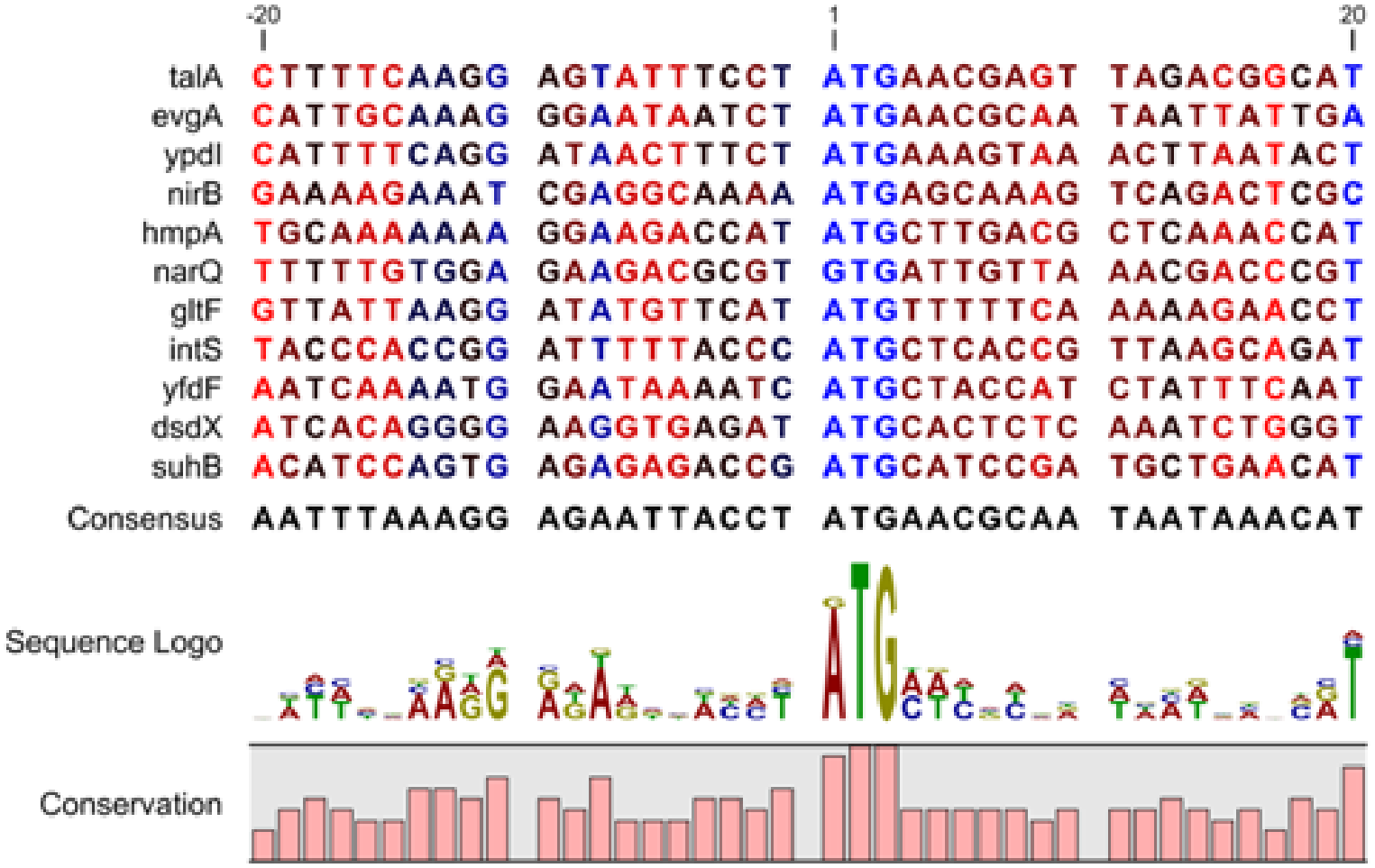
Concept de similarité

- L'évolution procède par duplications et mutations.
- Les protéines sont regroupées en familles :
- Alignement multiple de séquences homologues: permet d'extraire les caractéristiques importantes d'une famille de protéines.



ADN - Annotation

Concept de similarité



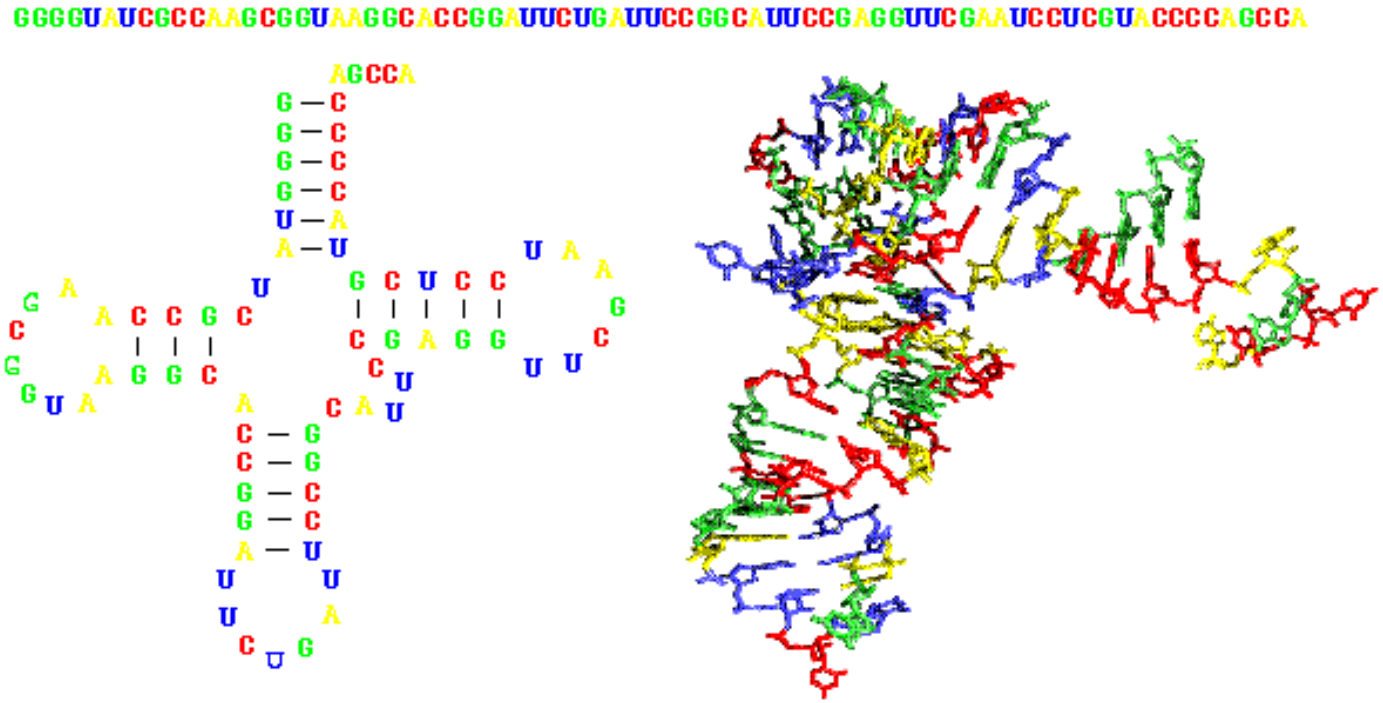
ARN
ARN non-codants

- Un ARN non-codant (ncRNA) est un ARN fonctionnel qui n'est pas traduit en protéine: tous les ARN autres que les ARNm.
- ARN non-codants inclu:
  - Les familles d'ARN ayant un rôle fondamentale dans la synthèse des protéines:
    - ARN de transfert (ARNt ou tRNA)
    - ARN ribosomique (ARNr)
  - Beaucoup d'autres familles découvertes plus récemment, dont les fonctions ne sont pas toujours connues: snoRNAs, microRNAs, siRNAs, piRNAs, RNase P…

ARN

ARN non-codants

- Prédire les structure 2D et 3D des ARN à partir de leur séquence primaire



http://www.mpi-inf.mpg.de/departments/d1/projects/CompBio/align.html

ARN

Transcriptome

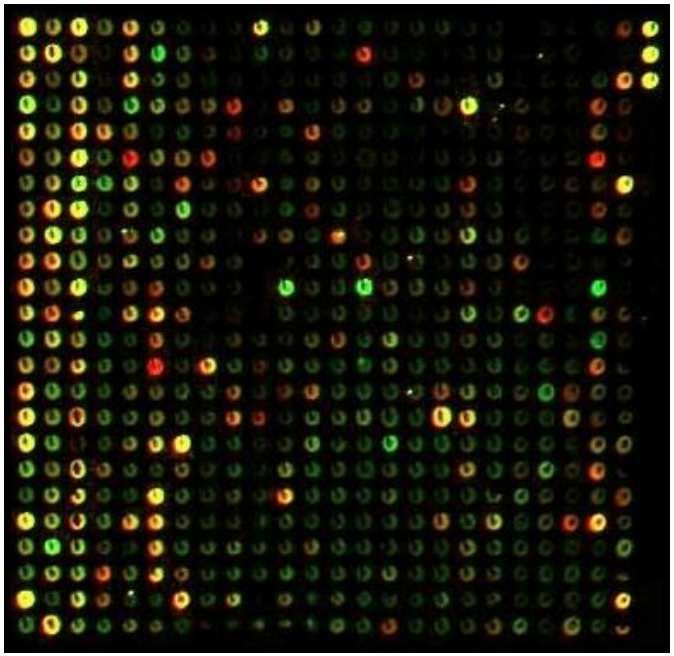
- Ensemble des ARNm issu de l'expression d'une partie du génome d'un tissu cellulaire ou d'un type de cellule.
- Caractérisation et quantification du transcriptome dans un tissu donné et dans des conditions données permettent:
  - D'identifier les gènes actifs,
  - De déterminer les mécanismes de régulation d'expression des gènes
  - De définir les réseaux d'expression des gènes.

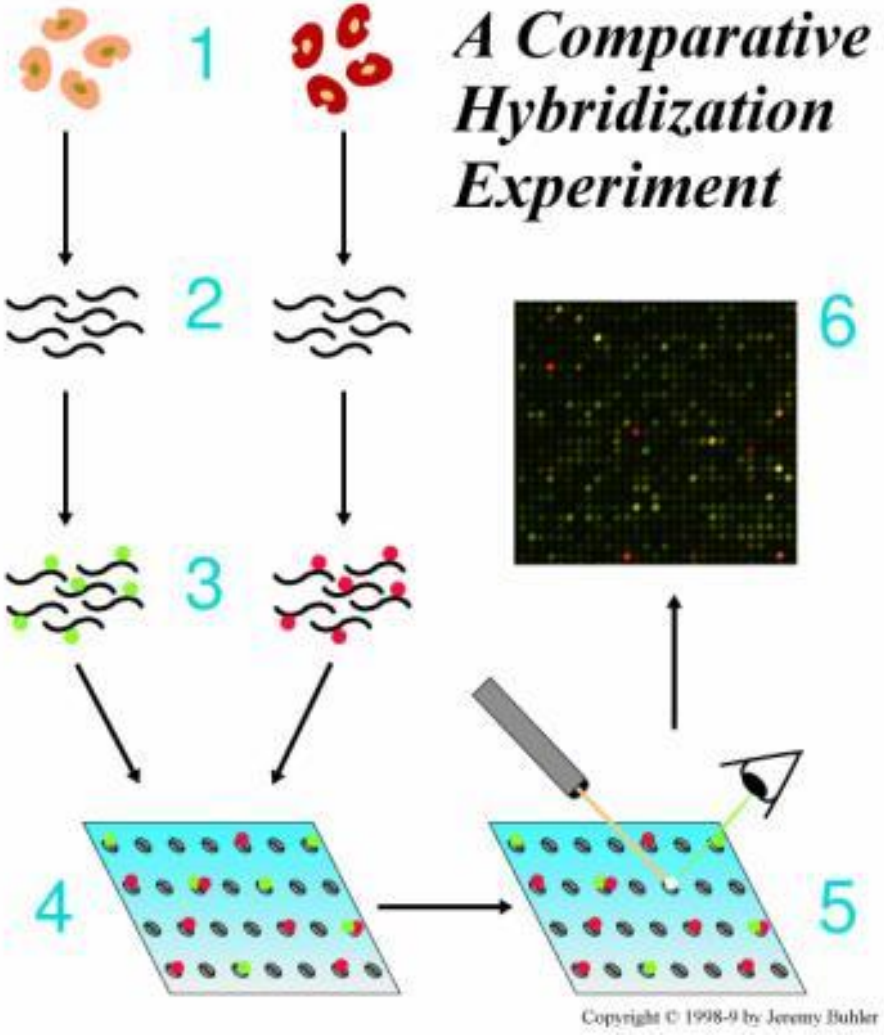
ARN

Puces à ADN « *DNA chip, DNA-microarray, biochip* ».

Mesure les niveaux d'expressions des gènes.

À partir de l'ARNm recueilli dans une cellule

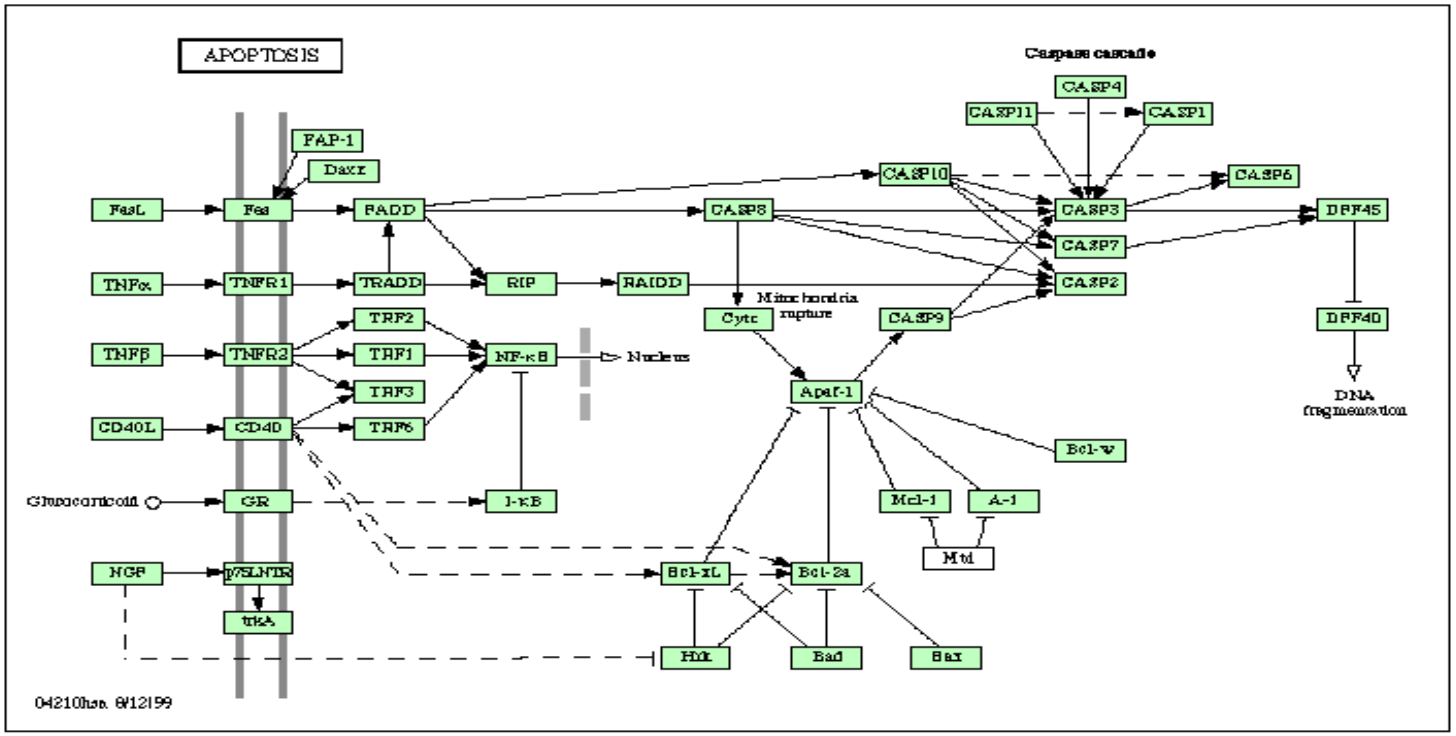




ARN

Chemins métaboliques, réseaux de régulation

Pathway of Apoptosis in Homo sapiens

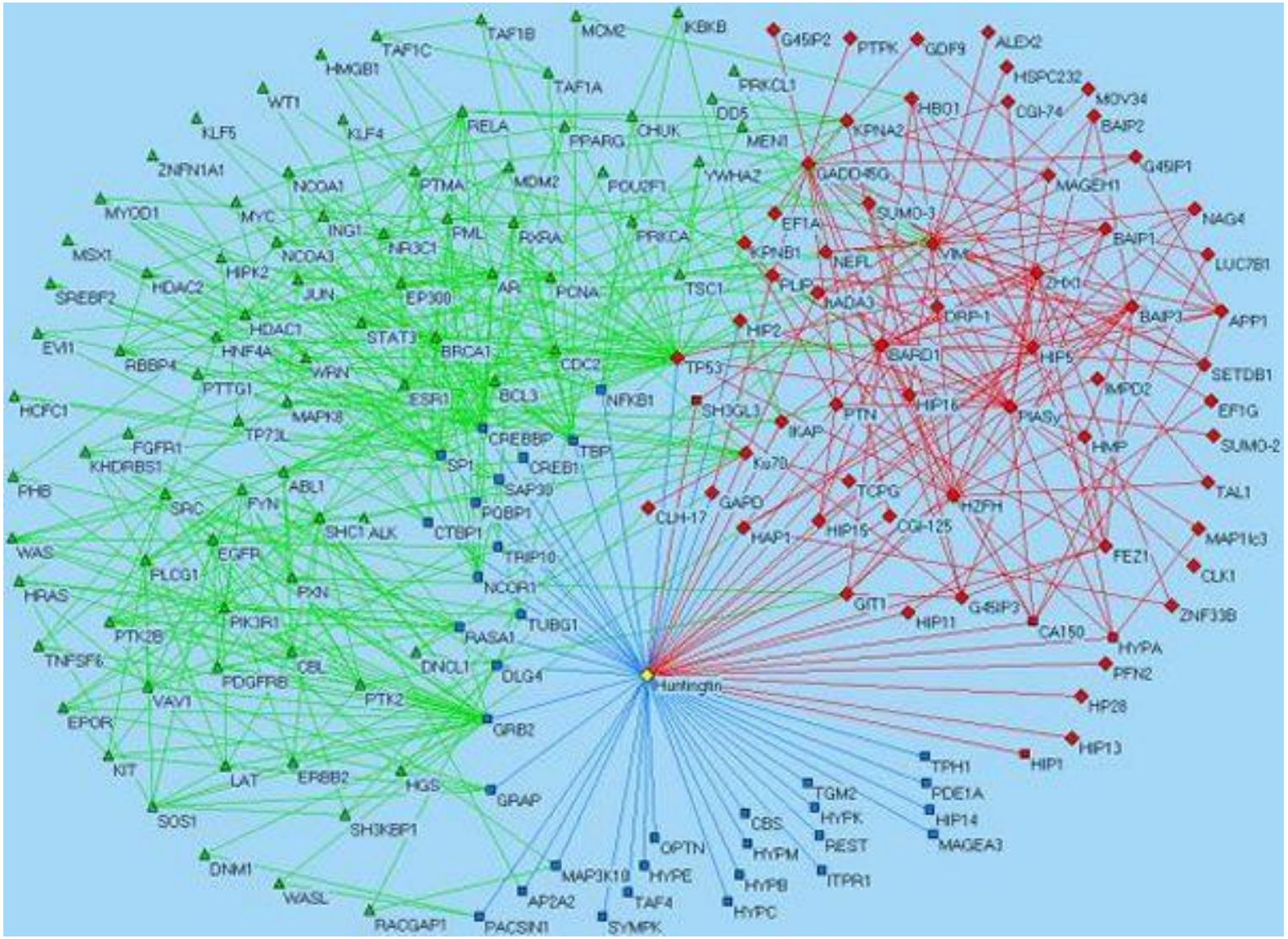


http://bip.weizmann.ac.il/education/course/introbioinfo/04/lect1/introbioinfo04/sld021.htm

"Apoptosis": process of programmed cell death.

Protéines

- Prédire la structure d'une protéine à partir de sa séquence
- Protéome: Ensemble des protéines exprimées dans une cellule dans des conditions données et à un moment donné.
- Protéomique: Étude du protéome.
  - Réseaux d'intéractions protéines-protéines
  - Chemins métaboliques



View of a protein-interaction-network in Huntington's disease

http://www.mdc-berlin.de/en/research/research_teams/proteomics_and_molecular_mechanisms_of_neurodegenerative_diseases/research/research1/index.html

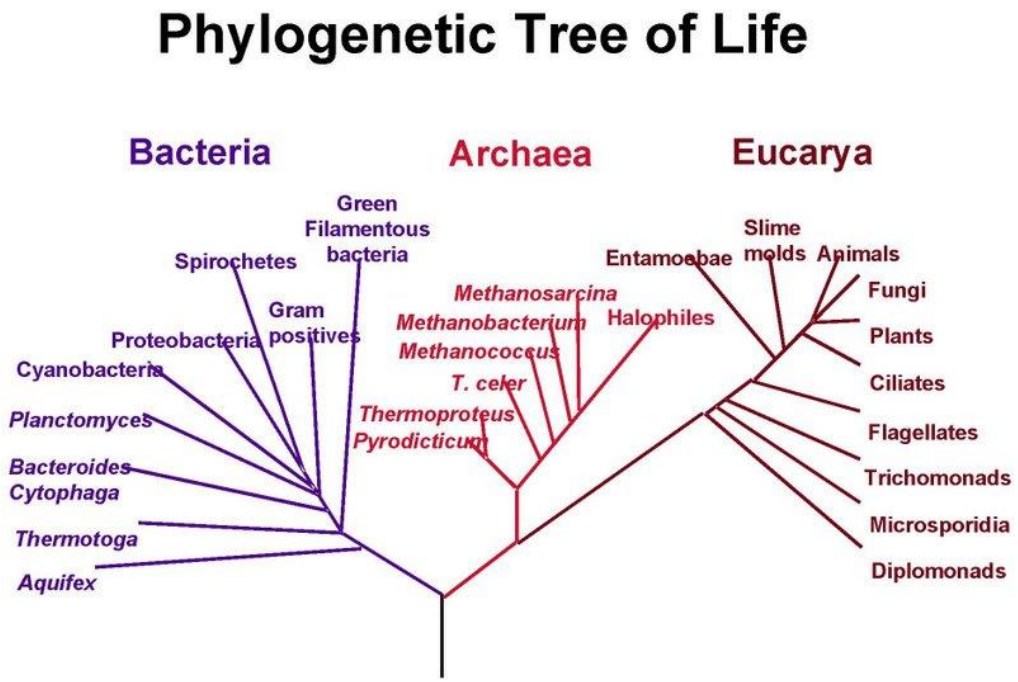
Problématiques Bio-Informatiques
Phylogénie

- Étude des relations d'évolution entre les espèces.

- **<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.

Problématiques Bio-Informatiques
Arbre de Phylogénie

- Premier objectif des études phylogénétiques: Reconstruire l'arbre de vie de toutes les espèces vivantes à partir des données génétiques observées.

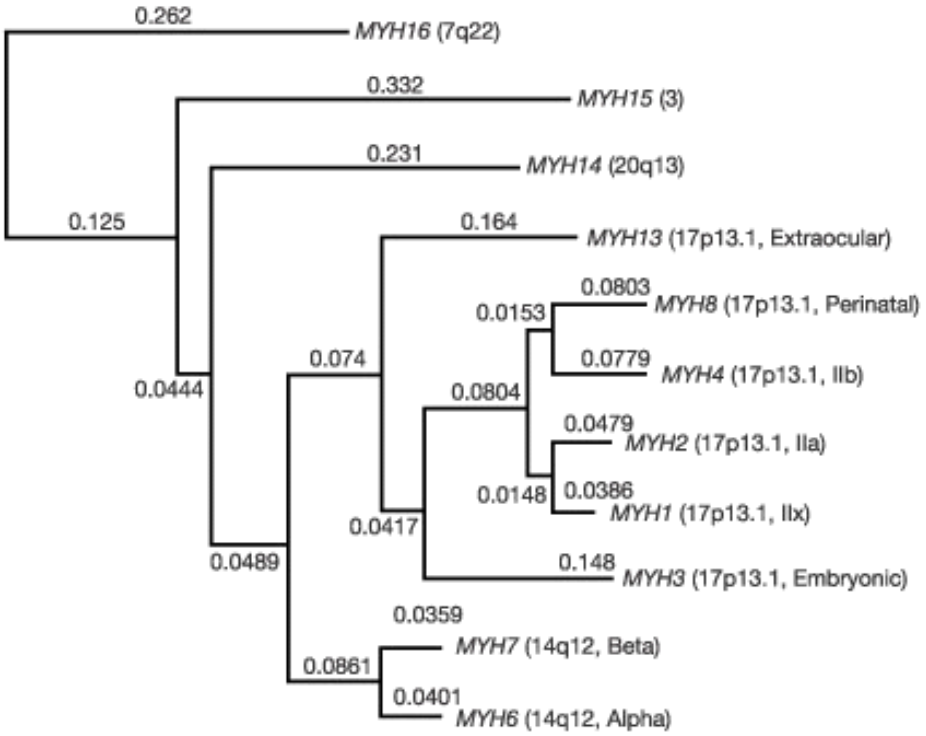


NASA:http://www.nasa.gov

Problématiques Bio-Informatiques

Arbre de Phylogénie

- Les arbres de phylogénie sont également utilisés pour représenter l'évolution commune d'une famille de gènes, ou de virus comme le HIV ou l'influenza.



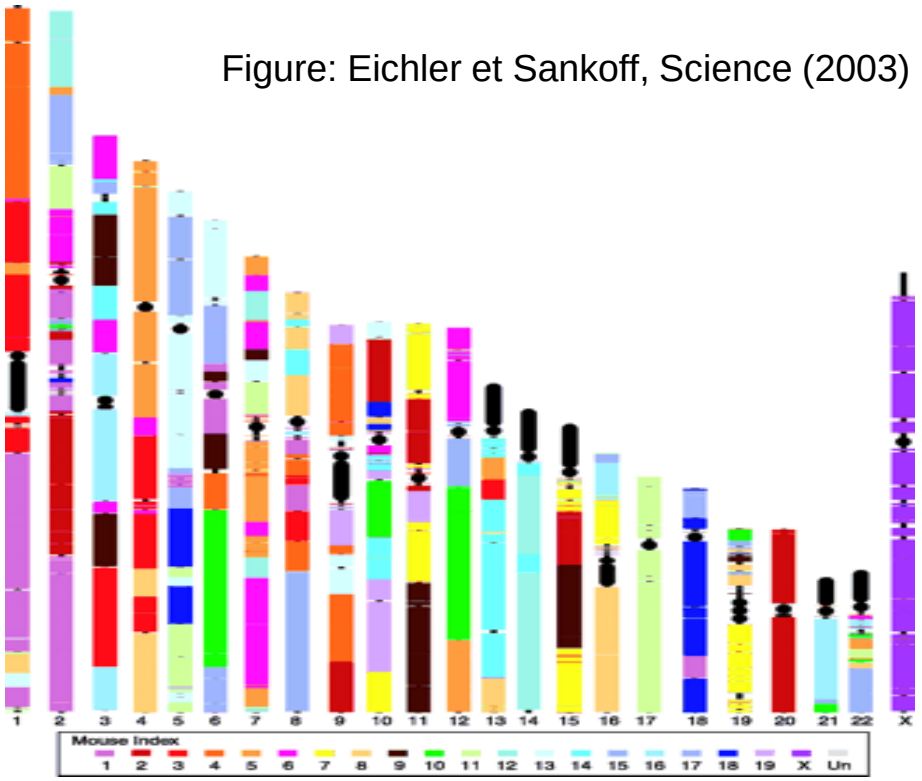
Observation de corrélations entre les mutations du gène Myosin avec certains changements anatomiques dans la lignée humaine. MYH16 chez l'humain très divergeant des autres copies du gène.

http://bio.nyk.ch/Myosin

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.