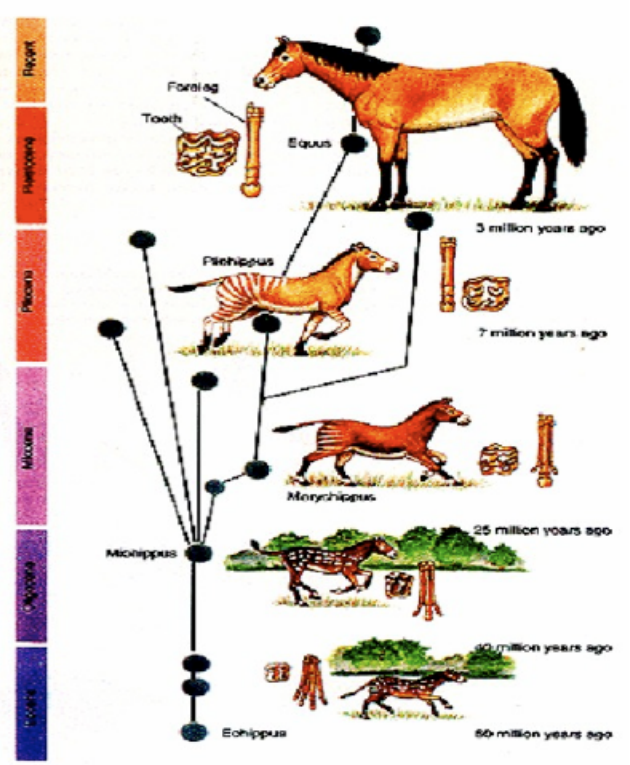
PARTIE I : Phylogénie moléculaire

PARTIE I : Phylogénie moléculaire

POURQUOI RÉALISER UNE PHYLOGÉNIE ?

Pourquoi réaliser une phylogénie ?

- Histoire de l'évolution
  - reconstruire l'histoire des espèces (→ paléontologie)



Pourquoi réaliser une phylogénie ?

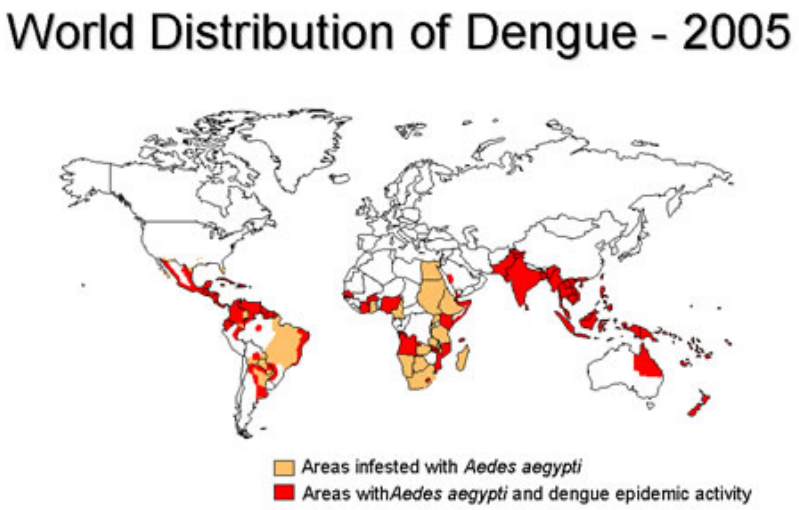
- Histoire de l'évolution
- Evolution des caractères
  - comprendre la mise en place des plans d'organisation

  → évolution-développement

Pourquoi réaliser une phylogénie ?

- Histoire de l'évolution
- Evolution des caractères
- Bio- écologie
  - Déplacement des espèces
  - Relations hôtes-parasites
  - mesurer/appréhender la biodiversité

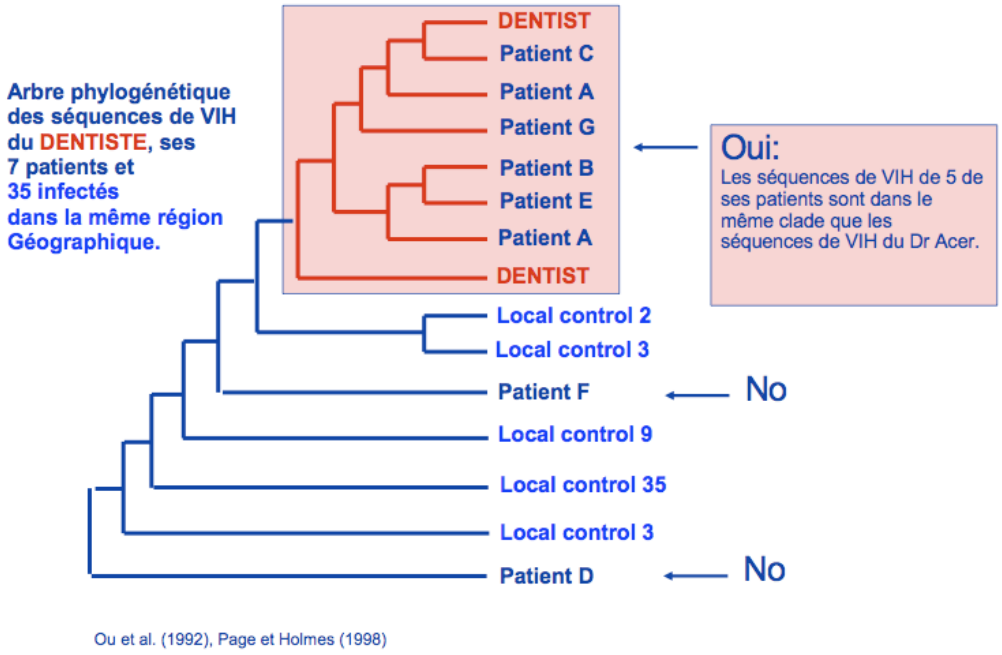
  → gestion, conservation des écosystèmes



Pourquoi réaliser une phylogénie ?

- Histoire de l'évolution
- Evolution des caractères
- Bio- écologie
- Epidémiology, microbiologie, virologie

→ caractériser la dynamique d'interactions durables

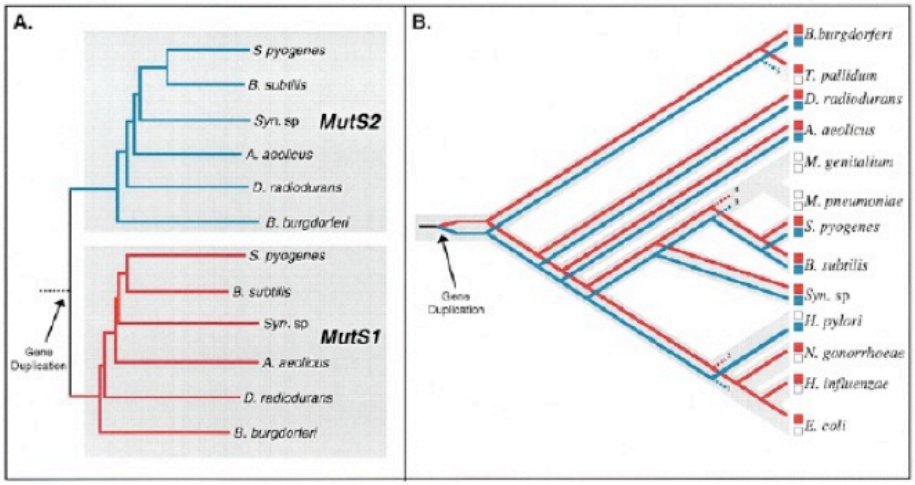


Ou et al. (1992), Page et Holmes (1998)

Pourquoi réaliser une phylogénie ?

- Histoire de l'évolution
- Evolution des caractères
- Bio- écologie
- Epidémiology
- Evolution des gènes au sein des espèces
  - caractériser les gènes de l'adaptation

  → amélioration des espèces domestiquées

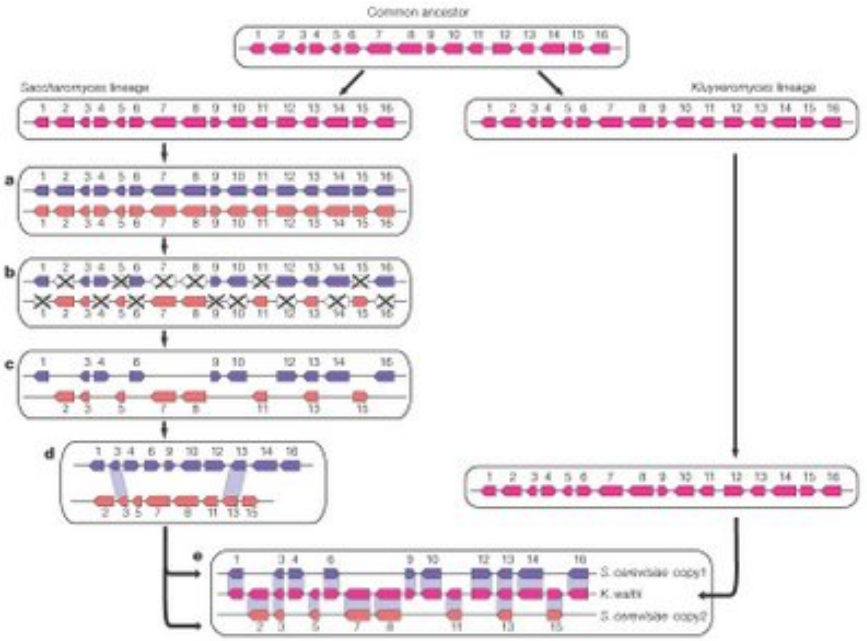


A. Arbre de gène

B. L'arbre de gènes superposé à un arbre d'espèces pour identifier les pertes de gènes.

Pourquoi réaliser une phylogénie ?

- Histoire de l'évolution
- Evolution des caractères
- Bio- écologie
- Epidémiology
- Evolution des gènes au sein des espèces
- Annotation des génomes (Genomique fonctionelle)
  - comprendre les mécanismes de l'évolution moléculaire (→ génomique structurale)



Pourquoi réaliser une phylogénie ?

Deux domaines d'application majeurs :

- Reconstruire l'histoire évolutionnaire de taxons, caractères ou de gènes.

- Analyse de caractères et de vitesses d'évolution

PARTIE I : Phylogénie moléculaire

LES ÉLÉMENTS D'UNE PHYLOGENIE

- Les caractères
- Le jeu de données
- Les banques
- Les alignements
- Les arbres

Pour faire une bonne phylogénie

Cinq étapes de l'analyse phylogénétique

1. Choix du jeu de données
   - Une bonne connaissance des séquences que l'on analyse
   - S'assurer de la validité du jeu de donnée (qualité des séquences et cohérence du JDD)

2. Alignement des séquences
   - Obtenir un bon alignement
   - Tester différents méthodes et revenir à la main sur les résultats

3. Détermination du modèle de substitution
4. Construction des arbres
5. Evaluation des arbres

PARTIE I : Phylogénie moléculaire

LES ÉLÉMENTS D'UNE PHYLOGENIE

- Les caractères
- Le jeu de données
- Les banques
- Les alignements
- Les arbres

Les caractères

- **1 Description et codage des états.**

  Présence absence : +/-; 0/1; a/b

  Etats multiples :

  Les 20 acides aminés

  Les 4 nucléotides A,T,C,G

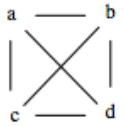
  Nb de répétitions en tandem (microsatellites)

  Morphologie (a, b, c, d, …)

Les caractères

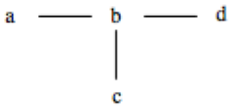
- **2 Transformation des états**
  - Non additives (1)
  - Additives (ou ordonnées)
    - Linéaires (2)
    - Non linéaires (3)
    - Complexes (4)
  - Polarisées (ou orientées) (5)

(1)

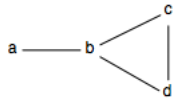


(2) a — b — c — d

(3)

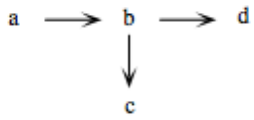


(4)

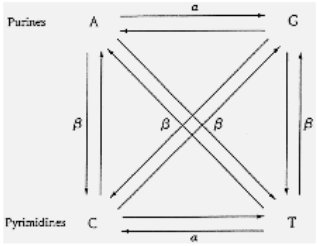


(5) a → b → c → d

(5)



- **3 Pondérations**
  - Des transformations (PAM/JTT ou ts/tv)
  - Des caractères



Les caractères

Au début les modes de classifications des espèces étaient:

- Les comparaisons morphologiques
- Les comparaisons comportementales
- Les répartitions géographiques

Les carractères

Morphologie vs. Données moléculaires

African white-backed vulture
(old world vulture)

Andean condor
(new world vulture)

Les vautours du vieux et du nouveau monde semblent être étroitement liés sur la base de leur morphologie

Les données moléculaires indiquent que les vautours du vieux monde sont liées à des oiseaux de proie (faucons, éperviers, etc), tandis que les vautours du Nouveau Monde sont plus étroitement liés à des cigognes

C'est un exemple de convergence évolutive

Les caractères

Au début les modes de classifications des espèces étaient:

- Les comparaisons morphologiques
- Les comparaisons comportementales
- Les répartitions géographiques

Aujourd'hui les phylogénies sont obtenues à partir:

- des séquences moléculaires (phylogénie moléculaire) : ADN, ARN, Protéines, Codons
- des caractères discrets (présence, absence, 0, 1)
- des fréquences des gènes
- des traits quantitatifs
- des sites de restriction, RFLP
- des microsatellites, SNP

PARTIE I : Phylogénie moléculaire

LES ÉLÉMENTS D'UNE PHYLOGENIE

- Les caractères
- Le jeu de données
- Les banques et formats
- Les alignements
- Les arbres

Le jeu de données

Les séquences d'ADN présentent beaucoup d'avantages face aux caractères de taxonomie morphologiques:

- L'état des caractères peut être déterminé sans ambiguïté
- Un grand nombre de caractère peuvent être pris en compte pour chaque individu

Inconvénients :

- Peu d'états donc possibilité de mutations silencieuses (homoplasie)
- Arbre de gène vs Arbre de génome (cad arbre de espèces)
- Alignements de qualité difficile à obtenir

Utilisation de méthodes mathématiques permettant de nous donner la représentation la moins fausse possible.

Le jeu de données

- Cas de la phylogénie d'espèces

  il faut choisir le marqueur moléculaire approprié au groupe taxonomique étudié.

- Critères du choix d'un marqueur:
  - universalité
  - structure conservée
  - absence de transfert génétique
  - taux d'évolution approprié
  - absence de biais sélectif

- Quelques exemples:
  - phylogénie de bactéries (16S rDNA)
  - phylogénie d'eucaryotes (18S rDNA, actine, EF1, RPB1(RNA polymerase))
  - phylogénie de plantes (rbcL(ribulose carboxylase), 18S rDNA)

- Phylogénie d'animaux
  - niveau phylum, classe, ordre : (18S rDNA, génome mt)
  - niveau famille : (RAG2(recombination activating gene 2), 12S, 16S mt)
  - niveau genre : (ITS, protéines mt)
  - niveau intra-spécifique : (D-Loop, introns)

Le jeu de données

- JDD :
  - Eviter les séquences incomplètes
  - Eviter les xénologues (transfert latéraux)
  - Eviter les séquences recombinantes (2 ancêtres)
  - Eviter les grandes familles complexes (répétitions et nombres de domaines importants)
  - Ajouter un groupe externe (outgroup)

- ADN ou protéines :
  - Quand cela est possible travailler préférentiellement avec des alignement de protéines en particulier lorsque les séquences d'ADN diffèrent de plus 70%. Si les séquences protéiques sont trop proches revenir à l'ADN.

- Intégralité ou partie du JDD:
  - Préférer des JDD réduits (peu de taxons) -> moins source d'erreur et moins temps pour l'analyse.
  - Ne pas obligatoirement inclure tous les caractères :
    - Supprimer les caractères incertains
    - Supprimer les zones trop variables: parties 3', 5', zones riches en gap.

Le jeu de données

- Constituer son jeu de données
  - Plan d'expérience (prélèvements, amplification, sequençage …)
  - Collecte dans les banques de données

PARTIE I : Phylogénie moléculaire

LES ÉLÉMENTS D'UNE PHYLOGENIE

- Les caractères
- Le jeu de données
- Les banques et formats
- Les alignements
- Les arbres

Les banques

Différentes catégories de bases de données :

Bases de données bioinformatiques

Bases généralistes
-banque N
-banque P

Bases spécialisées

Banques génomiques

Banques fonctionnelles :
-transcriptome
-protéome
-métabolome…

Difficultés !!
-Harmoniser le classement des informations
-Utiliser un langage commun pour échanger des informations entre toutes ces bases

Les banques

Types de données

Bases de séquences	Adresse
• **Bases génériques (multi-organismes)**	
EMBL / trEMBL	http://www.ebi.ac.uk/embl/
Genbank / GenPept	http://www.ncbi.nlm.nih.gov/entrez
DDBJ (DNA Data Bank of Japan)	http://www.ddbj.nig.ac.jp/
SwissProt	http://www.expasy.org/sprot/
• **Bases spécialisées (organisme)**	
GenoList	http://genodb.pasteur.fr/cgi-bin/WebObjects/GenoList
Cyanobase	http://www.kazusa.or.jp/cyano/
TAIR (The Arabidopsis Information Resource)	http://www.arabidopsis.org
FlyBase (Database of the Drosophila Genome)	http://flybase.bio.indiana.edu/
MGD (Mouse Genome Database)	http://www.informatics.jax.org/
GDB (Human Genome data Base)	http://gdbwww.gdb.org/
• **Bases spécialisées (thématique)**	
PROSITE (proteins families and domains)	http://prosite.expasy.org
PRINTS (protein fingerprints : group of motifs**)**	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php
EPD (Eukaryotic Promoter Database)	http://epd.vital-it.ch

Les banques

Types de données

Nom	adresse
• **Métabolisme**	
KEGG (Kyoto Encyclopedia of Genes and Genomes)	http://www.genomes.ad.jp/kegg
BRENDA (BRaunschweig ENzyme DAtabase)	http://www.brenda-enzymes.org
Enzyme	http://www.expasy.ch/enzyme
EcoCyc	http://ecocyc.org
• **Régulation transcriptionnelle**	
RegulonDB	http://regulondb.ccg.unam.mx
• **Interactions protéine-protéine**	
DIP (Database of Interacting Proteins)	http://dip.doe-mbi.ucla.edu/
BIND (The Biomolecular Interaction Network Database)	http://www.bind.ca/
• **Données structurales (3D)**	
PDB (Protein Data Bank)	http://www.rcsb.org/pdb
ModBase	http://modbase.compbio.ucsf.edu
• **Famille de gènes ou de protéines**	
The Protein Kinase Resource (PKR)	http://pkr.genomics.purdue.edu
5S Ribosomal RNA Database	http://biobases.ibch.poznan.pl/5Sdata/

Les banques

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genline, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GF
HAEMB, HAMST
HIDB, HIDC, HIV
HSC-2DPAGE,
KDNA, KEGG, Kiotho, LGIC, MAD, MaizeDb, MDB,
Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5
Micado, MitoDat, MITOMAP
MPDB, MRR, MutBase, MycD
OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene,Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, YEPD, YPD, YPM, **etc, etc, etc..... !!!!**

Nucleic Acids Research

http://www.oxfordjournals.org/nar/database/c/

Les banques

Les banques de données généralistes

- Ces banques contiennent des données hétérogènes
  - Collecte la plus exhaustive possible
  - Banques de séquences nucléiques
  - Banques de séquences protéiques
  - Banques de structure 3D de macromolécules
  - Banques d'articles scientifiques
- **Avantage :** tout est consultable en une fois

Les banques

Faiblesses des banques généralistes

- Hétérogénéité dans la nature des séquences
  - ADN nucléaire ou mitochondrial, ARN (t, r, m, ….), génome
- Variabilité de l'état des connaissances
  - caractérisation biologique beaucoup plus lente que le séquençage
- Erreurs dans les séquences
  - Liées à l'origine du fragment, à la technologie, à la méthodologie.
- Biais d'échantillonnage
  - des espèces
  - des gènes
  - redondance des données

→ création de banques spécialisées

Les banques

Les banques de données spécialisées

- Ces banques contiennent des données homogènes
  - Collecte établie autour d'une thématique particulière
- **Avantages :** facilité pour mettre à jour les données, vérifier leur intégrité, offrir une interface adaptée, …
- **Inconvénients :** ne cible pas toujours ce que l'on veut; toutes les banques possibles n'existent pas
- **Exemples :** banques spécialisées pour un génome, banques de séquences d'immunologies, banques sur des séquences validées, …

Les formats

Banques nucléiques, format d'une entrée

EMBL

```
ID   MET_MOUSE      STANDARD;      PRT;  1379 AA.
AC   P16056; Q62125;
DT   01-APR-1990 (Rel. 14, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   10-MAY-2005 (Rel. 47, Last annotation update)
DE   Hepatocyte growth factor receptor precursor (EC 2.7.1.112) (Met proto-
DE   oncogene tyrosine kinase) (c-met) (HGF receptor) (HGF-SF receptor).
GN   Name=Met;
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi;
OC   Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   NUCLEOTIDE SEQUENCE.
RX   MEDLINE=88262253; PubMed=2838789;
RA   Chan A.M.-L., King H.W.S., Deakin E.A., Tempest P.R., Hilkens J.,
RA   Kroezen V., Edwards D.R., Wills A.J., Brookes P., Cooper C.S.;
RT   "Characterization of the mouse met proto-oncogene.";
RL   Oncogene 2:593-599(1988).
RN   [2]
RP   NUCLEOTIDE SEQUENCE OF 1199-1270.
RX   MEDLINE=90152381; PubMed=2482828; DOI=10.1016/0378-1119(89)90465-4;
RA   Wilks A.F., Kurban R.R., Hovens C.M., Ralph S.J.;
RT   "The application of the polymerase chain reaction to cloning members
RT   of the protein tyrosine kinase family.";
RL   Gene 85:67-74(1989).
RN   [3]
RP   NUCLEOTIDE SEQUENCE OF 924-935.
RX   PubMed=8384622;
RA   Weidner K.M., Sachs M., Birchmeier W.;
RT   "The Met receptor tyrosine kinase transduces motility, proliferation,
RT   and morphogenic signals of scatter factor/hepatocyte growth factor in
RT   epithelial cells.";
RL   J. Cell Biol. 121:145-154(1993).
CC   -!- FUNCTION: Receptor for hepatocyte growth factor. Has a tyrosine-
CC       protein kinase activity.
CC   -!- CATALYTIC ACTIVITY: ATP + a protein tyrosine = ADP + a protein
CC       tyrosine phosphate.
```

Genbank

```
LOCUS       SCU49845               5028 bp    DNA     linear   PLN 21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
   PUBMED   7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
   PUBMED   8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
```

Banques nucléiques, format d'une entrée

- 3 parties :

Description générale de la séquence

« Features »

Description des objets biologiques présents sur la séquence

- Chaque ligne commence par un mot-clé
  - Deux lettres pour EMBL
  - Maximum 12 lettres pour Genbank et DDBJ
- Fin d'une entrée : //

La séquence

```
ctccggcagc ccgaggtcat cctgctagac tcagacctgg atgaacccat agacttgcgc        60
tcggtcaaga gccgcagcga ggccggggag ccgcccagct ccctccaggt gaagcccgag       120
acaccggcgt cggcggcggt ggcggtggcg gcggcagcgg cacccaccac gacggcggag       180
```

Les formats

Exemple d'une entrée Genbank :

Identificateur de l'entrée dans la banque (nom de l'entrée, longueur (pb), molécule (ADN, ARN, ADNc), date)

Informations descriptives sur la séquence

Numéro d'accession de la séquence

Un numéro d'accession/entrée + version
Un Genbank Identifier par version de séquence

Mots clefs qui peuvent être associés à une fiche Genbank pour retrouver plus facilement une entrée dans la base

Organisme d'où provient la séquence

1ière ligne : nom scientifique de l'organisme
2ième ligne : Classification taxonomique

Attribue un numéro unique à chaque référence bibliographique liée à la séquence
Pour chaque référence, on retrouve la liste des auteurs, le titre de la publication et le journal

```
LOCUS       JX402632               16170 bp    DNA     linear   INV 22-AUG-2012
DEFINITION  Crassostrea gigas protein kinase AMP-activated alpha-catalytic
            subunit gene, complete cds.
ACCESSION   JX402632
VERSION     JX402632.1  GI:401665878
KEYWORDS    .
SOURCE      Crassostrea gigas (Pacific oyster)
  ORGANISM  Crassostrea gigas
            Eukaryota; Metazoa; Lophotrochozoa; Mollusca; Bivalvia;
            Pteriomorphia; Ostreoida; Ostreoidea; Ostreidae; Crassostrea.
REFERENCE   1  (bases 1 to 16170)
  AUTHORS   Guevelou,E., Corporeau,C. and Huvet,A.
  TITLE     Regulation of truncated AMP-activated protein kinase alpha (AMPK
            alpha) in response to hypoxia in the muscle of Pacific oyster
            Crassostrea gigas
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 16170)
  AUTHORS   Guevelou,E., Corporeau,C. and Huvet,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (19-JUL-2012) Departement Ressources Biologiques et
            Environnement, IFREMER, Z.I. Pointe du Diable Technopole
            Brest-Iroise, Plouzane, Bretagne 29280, France
```

Les formats

Feature Table partagée entre Genbank/DDBJ/EMBL

Séquence complète décrite (commence tjrs à 1)

Description de la séquence complète : Organisme, type moléculaire, référence(s) dans d'autres banques (ici description du taxon associé)

mRNA : portions de la « source » qui sont transcrites en ARN messager

/product : annotation associée

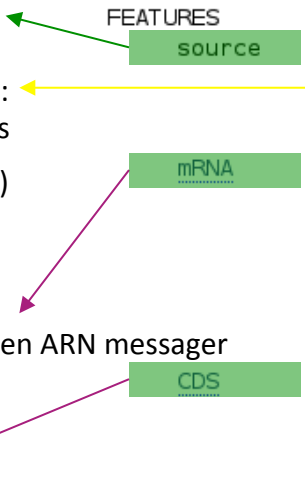
CDS : portions de la « source » qui correspondent à la séquence codante

/codon_start : 1, 2 ou 3 selon la phase ouverte de lecture

/protein_id : lien vers la fiche de la protéine correspondante au CDS

/db_xref : GI correspondant

/translation : séquence protéique correspondante au CDS




```
Location/Qualifiers
1..16170
/organism="Crassostrea gigas"
/mol_type="genomic DNA"
/db_xref="taxon:29159"
join(<1..100,771..912,1152..1245,2478..2622,2724..2811,
3118..3187,3306..3354,3664..3769,4089..4194,4436..4550,
5963..6027,7325..7369,10242..10281,10533..10619,
10853..10923,11161..11226,11680..11839,13594..13626,
15195..15233,16073..>16170)
/product="protein kinase AMP-activated alpha-catalytic
subunit"
join(1..100,771..912,1152..1245,2478..2622,2724..2811,
3118..3187,3306..3354,3664..3769,4089..4194,4436..4550,
5963..6027,7325..7369,10242..10281,10533..10619,
10853..10923,11161..11226,11680..11839,13594..13626,
15195..15233,16073..16170)
/codon_start=1
/product="protein kinase AMP-activated alpha-catalytic
subunit"
/protein_id="AFP95933.1"
/db_xref="GI:401665879"
/translation="MAEKSSSSQNAQVKIGHYILGDTLGIGTFGKVKIATHQLTNHK
AVKILNRQKIKSLDVVSKIKREIQNLKLFRHPHIIKLYQVISTPTDIFMVMEYVSGG
LFDYIVKHGKLKEPEARRFFQQIISGVDYCHRHMVVHRDLKPENLLLDSSLNVKIAD
GLSNMMHDGEFLRTSCGSPNYAAPEVISGKLYAGPEVDIWSCGVILYALLCGTLPFD
EHVPTLFRKIKSGIFAVPDYLNKEVVSLLCLMLQVDPLKRATIAQIRDHDWFQKDLP
YLFPSPQDQDASIVEMDVIREICEKFGVTEYEVQRALLSNDPHDQLNIAYHLIVDNR
LAGEVTDVELQEFYLASSPPPDSFLLAKQSHILGTPTPQEHASSPMRPHPERMPEMK
TTHTLEPVSSAKQLGAQAKKAKWHLGIRSQSKPLDIMHEVFRAMKTLDYEWKIVTPY
VRVRRKNPVSGRFSKMSLQLYQVDQKSYLLDFKSLSNVEIHESMSSSSSLEGGRMPL
PPSSCSDLDPVTDTVLLMPESASTSESFCSNLDEKMDIDEEQPRQHQTLEFFEMCAS
ITTLAR"
```

Les formats

Séquence complète de la base 1 à 16170

ORIGIN

```
    1 atggcggaga agtcctcctc ctctcagaac gcacaagtca agattggaca ttacattttg
   61 ggggataccc taggaatagg aacgtttggc aaagttaaaa gtaagtacaa gagaaattca
  121 catgtattag gctgtaaacg acaagcatgg tcaagtctca agagataggc ccatgaacaa
  181 tgaaagtagt ctgaagttca ttacccttta gccctatgaa tatacaaaca cagtgcttaa
  241 ttctaaacag atcaatatag aatgaataag aatttgccgt gctttgaata tctcaggtga
  301 atgaccacaa tcttgtacaa aggaatcaaa attaagcttt cactgcccca gtaagcactg
  361 gttatgaatt acatgatctg tatttctaat gaagaaaata atagtgtggt attcaataag
  421 tagatgactt acaatcttta aatttatatt taatgtgtaa aatgttacgt ggaatactac
  481 aaatacaatt ggcgtggatc tgataatttt tgtgtataaa cctgtatgca tgtcgtaata
  541 tgaattgcat caaaaagatt taaacttta aacctttaa ccacattgtg atacatatat
  601 tgtaagctac tgttagcaag gttatacttg ttaattaagc ttgattatat attaaaatct
  661 tttgattttt aaagaaaaca tccttttttg gaaaaatctt ttttacatca tatgaaataa
  721 ctgtaatatt ttgtcatgat atggtatgaa accctcttcc tttttctcag ttgccaccca
  781 tcagctgacc aatcataagg tggcggtaaa gatcctcaac aggcagaaga tcaagagtct
  841 cgatgtcgtc agtaaaatca agagagaaat tcagaatctc aagctttttc gtcacccaca
  901 cattatcaaa ctgtaagtca tggcatgcag gtagtcttta atggttaagg caatgaatga
  961 caatttttca ttaaattaca acaatcagac gatgatgttt ataatgaaat gacccttata
 1021 caatataaag tacatcaaac tatttatgaa ctgatatttt tatacttata gacctattgc
 1081 ctgtttttttt acttaccggt atacacctgt tacctgattg ctttgtttat ttggtttctc
 1141 tgtttggaca ggtatcaggt gatcagtacc cccacagata tcttcatggt gatggagtat
 1201 gtgtctgggg gagaattgtt tgattacatt gtcaaacacg gcaaggtact cctttatctc
 1261 cttacttgct gctgaacaaa tgactttgac cttttgtatg tgcatcattc cttatactta
 1321 ggccaaaaaa aattattcct gtttcctgtt gcccgaccga ccctatcttt taccctccga
 1381 ccctaaaagt ttttttgtca tgatggtggt gatcggtaac ttcgccagaa tttcctcaga
 1441 aagagaagtg aagatgacca agtcttctga gttcataatc gtgtaaatta acctcactgg
 1501 cgaaagaatg atagaaaatt aatctacaga cattttact gcataaatat accttgtggt
 1561 gaataacttg cgactatctt accttcaggg gttataaata tgggtgatgc aacacattct
 1621 gtatattgga aattacaaaa acaatgttta ttactttaaa tcctgatatt acagttagtc
 1681 aataacctgt tattagttaa taactgcaat aaatgttaaa attatcataa aactgctttc
 1741 tatattatct tgagaattaa acagatcctc tacataacgt acggtagtaa gtgggtcaaa
 1801 gggttggact aataccccaa atacatggaa gtacagagag tgatttaatt aatcatgttg
 1861 atattgacat taagctaata caatacaata atatttatcc agaaaaggat aaatacattc
 1921 atacacatag attaaatttt gcaaataata ttagttttta tttttaatta aaaaattaat
 1981 ctctatttta tttttctttg cctgttgttt tctatacatt acatgcatca acaatacaaa
 2041 gacctttgga aattttaaat catctactta catactgaac actgggaata tacatctctc
```

...

Les formats

Feature Table

5.1 Eukaryotic gene

```
source          1..1509
                /organism="Mus musculus"
                /strain="CD1"
                /mol_type="genomic DNA"
promoter        <1..9
                /gene="ubc42"
mRNA            join(10..567,789..1320)
                /gene="ubc42"
CDS             join(54..567,789..1254)
                /gene="ubc42"
                /product="ubiquitin conjugating enzyme"
                /function="cell division control"
                /translation="MVSSFLLAEYKNLIVNPSEHFKISVNEDNLTEGPPDTLY
                QKIDTVLLSVISLLNEPNPDSPANVDAAKSYRKYLYKEDLESYPMEKSLDECS
                AEDIEYFKNVPVNVLPVPSDDYEDEEMEDGTYILTYDDEDEEEDEEMDDE"
exon            10..567
                /gene="ubc42"
                /number=1
intron          568..788
                /gene="ubc42"
                /number=1
exon            789..1320
                /gene="ubc42"
                /number=2
polyA_signal    1310..1317
                /gene="ubc42"
```

Site de fixation de l'ARN polymérase

2 exons du CDS

5.2 Bacterial operon

```
source          1..9430
                /organism="Lactococcus sp."
                /strain="MG1234"
                /mol_type="genomic DNA"
operon          160..6865
                /operon="gal"
-35_signal      160..165
                /operon="gal"
-10_signal      179..184
                /operon="gal"
CDS             405..1934
                /operon="gal"
                /gene="galA"
                /product="galactose permease"
                /function="galactose transporter"
CDS             2003..3001
                /operon="gal"
                /gene="galM"
                /product="aldose 1-epimerase"
                /EC_number="5.1.3.3"
                /function="mutarotase"
CDS             3235..4537
                /operon="gal"
                /gene="galK"
                /product="galactokinase"
                /EC_number="2.7.1.6"
mRNA            189..6865
                /operon="gal"
```

3 gènes de l'opéron

Gène avec activité enzymatique

Plus d'informations sur www.insdc.org/files/feature_table.html

Les formats

Format Fasta

- Format standard de stockage des séquences nucléiques et protéiques
- En entrée de la plupart des outils de bioinformatique

Ligne de description commence toujours par un **>Identifiant Description**

Convention dans les banques généralistes :
Genbank :

>gi|*numéro gi*|**gb**|*numéro d'accession*|locus

EMBL :

>gi|*numéro gi*|**emb**|*numéro d'accession*|locus

DDBJ :

>gi|*numéro gi*|**dbj**|*numéro d'accession*|locus

Local Sequence identifier :
>lcl|identifiant

```
>gi|401665878|gb|JX402632.1| Crassostrea gigas protein kinase AMP-activated
alpha-catalytic subunit gene, complete cds
ATGGCGGAGAAGTCCTCCTCCTCTCAGAACGCACAAGTCAAGATTGGACATTACATTTTGGGGGATACCC
TAGGAATAGGAACGTTTGGCAAAGTTAAAAGTAAGTACAAGAGAAATTCACATGTATTAGGCTGTAAACG
ACAAGCATGGTCAAGTCTCAAGAGATAGGCCCATGAACAATGAAAGTAGTCTGAAGTTCATTACCCTTTA
GCCCTATGAATATACAAACACAGTGCTTAATTCTAAACAGATCAATATAGAATGAATAAGAATTTGCCGT
GCTTTGAATATCTCAGGTGAATGACCACAATCTTGTACAAAGGAATCAAAATTAAGCTTTCACTGCCCCA
GTAAGCACTGGTTATGAATTACATGATCTGTATTTCTAATGAAGAAAATAATAGTGTGGTATTCAATAAG
TAGATGACTTACAATCTTTAAATTTATATTTAATGTGTAAAATGTTACGTGGAATACTACAAATACAATT
GGCGTGGATCTGATAATTTTTGTGTATAAACCTGTATGCATGTCGTAATATGAATTGCATCAAAAAGATT
TAAACTTTTAAACCTTTTAACCACATTGTGATACATATATTGTAAGCTACTGTTAGCAAGGTTATACTTG
TTAATTAAGCTTGATTATATATTAAAATCTTTTGATTTTTAAAGAAAACATCCTTTTTTGGAAAAATCTT
TTTTACATCATATGAAATAACTGTAATATTTTGTCATGATATGGTATGAAACCCTCTTCCTTTTTCTCAG
TTGCCACCCATCAGCTGACCAATCATAAGGTGGCGGTAAAGATCCTCAACAGGCAGAAGATCAAGAGTCT
CGATGTCGTCAGTAAAATCAAGAGAGAAATTCAGAATCTCAAGCTTTTTCGTCACCCACACATTATCAAA
CTGTAAGTCATGGCATGCAGGTAGTCTTTAATGGTTAAGGCAATGAATGACAATTTTTCATTAAATTACA
ACAATCAGACGATGATGTTTATAATGAAATGACCCTTATACAATATAAAGTACATCAAACTATTTATGAA
CTGATATTTTTATACTTATAGACCTATTGCCTGTTTTTTTACTTACCGGTATACACCTGTTACCTGATTG
CTTTGTTTATTTGGTTTCTCTGTTTGGACAGGTATCAGGTGATCAGTACCCCCACAGATATCTTCATGGT
GATGGAGTATGTGTCTGGGGGAGAATTGTTTGATTACATTGTCAAACACGGCAAGGTACTCCTTTATCTC
CTTACTTGCTGCTGAACAAATGACTTTGACCTTTTGTATGTGCATCATTCCTTATACTTAGGCCAAAAAA
AATTATTCCTGTTTCCTGTTGCCCGACCGACCCTATCTTTTACCCTCCGACCCTAAAAGTTTTTTTGTCA
TGATGGTGGTGATCGGTAACTTCGCCAGAATTTCCTCAGAAAGAGAAGTGAAGATGACCAAGTCTTCTGA
GTTCATAATCGTGTAAATTAACCTCACTGGCGAAAGAATGATAGAAAATTAATCTACAGACATTTTTACT
GCATAAATATACCTTGTGGTGAATAACTTGCGACTATCTTACCTTCAGGGGTTATAAATATGGGTGATGC
AACACATTCTGTATATTGGAAATTACAAAAACAATGTTTATTACTTTAAATCCTGATATTACAGTTAGTC
AATAACCTGTTATTAGTTAATAACTGCAATAAATGTTAAAATTATCATAAAACTGCTTTCTATATTATCT
```

Les formats

Description générale de la séquence

```
ID   AF226511   standard; genomic DNA; PRO; 948 BP.
AC   AF226511;
SV   AF226511.1
DT   15-MAR-2000 (Rel. 63, Created)
DT   04-JAN-2006 (Rel. 86, Last updated, Version 2)
DE   Neisseria meningitidis strain 1000 membrane protein GNA1220 (gna1220) gene,
DE   complete cds.
OS   Neisseria meningitidis
OC   Bacteria; Proteobacteria; Betaproteobacteria; Neisseriales; Neisseriaceae;
OC   Neisseria.

RP   1-948
RX   DOI; 10.1126/science.287.5459.1816.
RX   PUBMED; 10710308.
RA   Pizza M., Scarlato V., Masignani V., Giuliani M.M., Arico' B., …
RT   "Identification of vaccine candidates … "
RL   Science 287(5459):1816-1820(2000).
RL   Submitted (19-JAN-2000) to the EMBL/GenBank/DDBJ databases.
RL   IRIS Immunobiological Research Institute in Siena, Chiron SpA, Via
RL   Fiorentina, 1, Siena 53100, Italy
```

Banques nucléiques, les différentes lignes (1/2)

- ID : nom de l'entrée , …
  - Unique (propre à une entrée)
  - Non permanent (peut changer au cours des versions)
- AC : numéro d'accession
  - parfois plusieurs pour une même entrée (fusion d'entrées)
  - Permanent (ne disparaît jamais de la banque)
- SV : version de la séquence (Acc.version)
- DT : date d'incorporation dans la banque et de dernière mise à jour
- DE : description du contenu de l'entrée

Banques nucléiques, la ligne ID

```
ID entryname dataclass; molecule; division; sequencelength BP.
```

Exemple: ID AB000263 standard; RNA; PRI; 368 BP.

- **Entryname :** nom de l'entrée
  - *en général numéro d'accession*
- **Dataclass :** toujours le mot « standard »
- **Molecule :** type de la molécule de l'entrée
  - *DNA, RNA, circular DNA, …*
- **Division :** essentiellement basé sur la taxonomie
  - *HUM (Human), MUS (Souris), MAM (Other Mammals), ...*
- **Taille** : en paires de bases

Banques nucléiques, les différentes lignes (2/2)

- **KW** : liste de mots-clés (désuet)
- **OS** : organisme d'où provient la séquence (nom latin)
- **OC** : taxonomie (ou « artificial sequence »)
  - *Exemple : Eukaryota; Planta; Phycophyta; Euglenophyceae.*
- **OG** : localisation de séquences non nucléaires
  - Exemple : *Mito, Plasmid …*
- **RA, RT, RN, RC, RX, RP, RL** : réf. bibliographiques
- **DR** : liaison avec d'autres banques de données
- **FH, FT** : caractéristiques d'une entrée (Features)
- **SQ** : séquence (termine par **//**)

Les formats

« Features » : Description des objets biologiques présents sur la séquence

```
FH   Key             Location/Qualifiers
FH
FT   source          1..948
FT                   /db_xref="taxon:487"
FT                   /mol_type="genomic DNA"
FT                   /note="serogroup: B"
FT                   /organism="Neisseria meningitidis"
FT                   /strain="1000"
FT   gene            1..948
FT                   /gene="gna1220"
FT   CDS             1..948
FT                   /codon_start=1
FT                   /db_xref="GOA:Q9JPH5"
FT                   /db_xref="InterPro:IPR001107"
FT                   /db_xref="InterPro:IPR001972"
FT                   /db_xref="UniProtKB/TrEMBL:Q9JPH5"
FT                   /note="similar to stomatin-like proteins; Genome-derived
FT                   Neisseria Antigen GNA1220"
FT                   /transl_table=11
FT                   /gene="gna1220"
FT                   /product="membrane protein GNA1220"
FT                   /protein_id="AAF42660.1"
FT                   /translation="MEFFIILLVAVAVFGFKSFVVIPQQEVHVVERLGRFHRALTAGLN
FT                   ILIPFIDRVAYRHSLKEIPLDVPSQVCITRDNTQLTVDGIIYFQVTDPKLASYGSSNYI
FT                   MAITQLAQTTLRSVIGRMELDKTFEERDEINSTVVSALDEAAGAWGVKVLRYEIKDLVP
FT                   PQEILRSMQAQITAEREKRARIAESEGRKIEQINLASGQREAEIQQSEGEAQAAVNASN
FT                   AEKIARINRAKGEAESLRLVAEANAEAIRQIAAALQTQGGADAVNLKIAEQYVAAFNNL
FT                   AKESNTLIMPANVADIGSLISAGMKIIDSSKTAK"
XX
```

Les formats

Banques nucléiques, Features

But : Mettre à disposition un vocabulaire étendu pour décrire les caractéristiques biologiques des séquences.

Format :

- **Key :** indique un groupe fonctionnel
  - Vocabulaire contrôlé, hiérarchique
- **Location :** instructions pour trouver l'objet sur la séquence de l'entrée
- **Qualifiers :** informations complémentaires
  - /qualifier= ' 'commentaires libres' '

Les formats

Banques nucléiques, Key

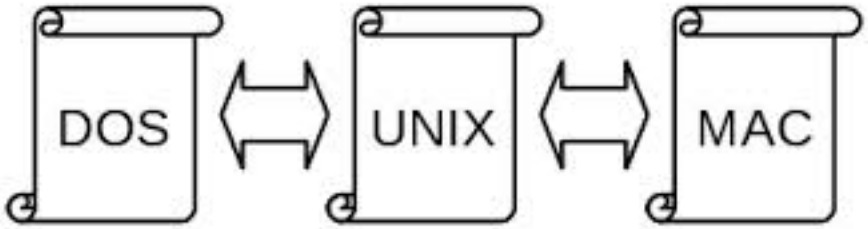
- Mot-clé le plus général : misc_feature
- Changements dans la séquence : misc_difference, ...
- Régions répétées : repeat_region, ...
- Régions des Ig : immunoglobulin_related, ...
- Structures secondaires : misc_structure
  - stem_loop
  - D-loop
- Régions impliquées dans la recombinaison : misc_recomb, ...

Les formats

Manipulation des fichiers.

Utilisez des éditeurs de texte et pas des traitements de texte !!

Méfiez vous des formats!!



PARTIE I : Phylogénie moléculaire

LES ÉLÉMENTS D'UNE PHYLOGENIE

- Les caractères
- Le jeu de données
- Les banques et formats
- Les alignements
- Les arbres

L'alignement

L'alignement est une étape cruciale qui permet de choisir les sites qui seront utilisés dans les analyses phylogénétiques.

But : S'assurer que chacun des sites choisis est homologue :

- hérité d'un ancêtre commun par descendance directe

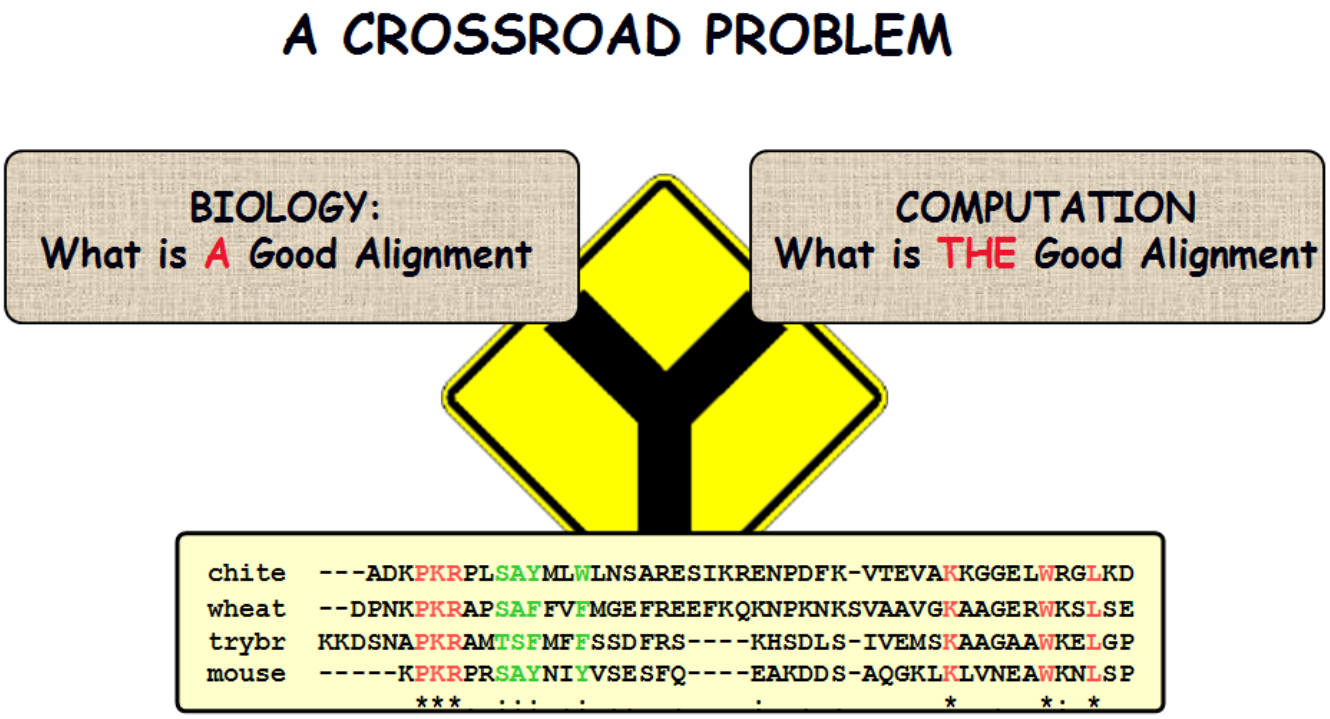
⇨ contient une information phylogénétique sur cette histoire évolutive.

L'alignement est une étape cruciale pour la phylogénie mais pas seulement.

Main applications of multiple sequence alignments

Application	*Procedure*
Extrapolation	A good multiple alignment can help convincing you that an uncharacterized sequence is really a member of a protein family.
Phylogenetic analysis	If you carefully chose the sequences to include in your multiple alignment, you can reconstruct the history of these proteins.
Pattern Identification	By discovering very conserved positions you can identify a region that is characteristic of a function (in proteins or in nucleic acid sequences).
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain. You can use this profile to scan databases for new members of the family.
DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potential similar binding sites.
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for proteins or RNA. Sometimes it can also help building a 3-D model.
PCR analysis	A good multiple alignment can help you identifying the less degenerated portions of a protein family
nsSNP	Identify the nsSNP that are the most likely to alter the function

L'alignement



Une nouvelle demi-journée de formation

Laurent Duret : Pôle Bioinformatique Lyonnais : http://pbil.univ-lyon1.fr/alignment.html
Cedric Notredame

L'alignement

Critères structurels
Les résidus sont agencés de sorte que ceux qui jouent un rôle similaire se retrouvent dans la même colonne.

Critères évolutifs
Les résidus sont agencés de sorte que ceux ayant le même ancêtre se retrouvent dans la même colonne.

Critères de similarité
Les résidus similaires seront autant que possible dans la même colonne

Pour s'assurer de l'homologie d'un site:

- la structure primaire des séquences (ordre des nucléotides)
- la structure secondaire des séquences (gènes ribosomiques : tiges boucles)
- la séquence en acides aminés (gènes codant pour des protéines)

L'alignement

Représentation

- Les résidus (nucléotides, acides-aminés) sont superposés de façon à maximiser la similarité entre les séquences.

```
G T T A A G G C G – G G A A A
G T T – – – G C G A G G A C A
* * *       * * *   * * *   *
```

- Mutations :
  - Substitution (*mismatch*)
  - Insertion
  - Délétion

    -> Insertions ou délétions : indels (***gap***).

L'alignement

Quel est bon ?

```
G T T A C G A        G T T A C G A
G T T - G G A        G T T G - G A
* * *   * *          * * *     * *
```

OU

```
G T T A C - G A
G T T - - G G A
* * *       * *
```

- Pour le biologiste, généralement, le bon alignement est celui qui représente le scénario évolutif le plus probable

L'alignement

Exercice

- Calculer le pourcentage d'identité et le pourcentage de similarité entre ces deux séquences protéiques :

```
Query  1   MAPWMHLLTVLALLAVWGPNSVQVYSSQHLCGSNLVEALYMTC--RSGFYRPHDRRELED  58
           MA WM LL +LALLA+WGP+  Q + +QHLCGS+LVEALY+ C  R  FY P  RRE ED
Sbjct  1   MALWMRLLPLLALLALWGPDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60

Query  59  LQVEQAEL--GLEAGGLQPSALEMILQKRGIVDQCCNNIC  96
           LQV Q EL  G  AG LQP ALE  LQKRGIV+QCC +IC
Sbjct  61  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIC  100
```

L'alignement

Exercice

- Calculer le pourcentage d'identité et le pourcentage de similarité entre ces deux séquences protéiques :

Score	Expect	Method	Identities	Positives	Gaps
106 bits(265)	1e-27	Compositional matrix adjust.	64/100(64%)	73/100(73%)	4/100(4%)

```
Query  1    MAPWMHLLTVLALLAVWGPNSVQVYSSQHLCGSNLVEALYMTC--RSGFYRPHDRRELED  58
            MA WM LL +LALLA+WGP+  Q + +QHLCGS+LVEALY+ C  R  FY P  RRE ED
Sbjct  1    MALWMRLLPLLALLALWGPDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60

Query  59   LQVEQAEL--GLEAGGLQPSALEMILQKRGIVDQCCNNIC  96
            LQV Q EL  G  AG LQP ALE  LQKRGIV+QCC +IC
Sbjct  61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIC  100
```

Utilisation de % identitée et similarité pas suffisant -> utilisation de score

L'alignement

Score de similarité

```
G T T A A G G C G - G G A A A
G T T - - - G C G A G G A C A
* * *       * * *   * * *   *
```

Score = $\sum_{début}^{fin} pondération\_substitution - \sum_{début}^{fin} pénalité\_gap$

Exemple:

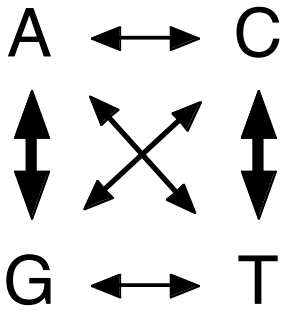
- identité = 1
- mismatch = 0
- gap = -1

Score = 10 - 4 = 6

L'alignement

Modèles d'évolution

ADN

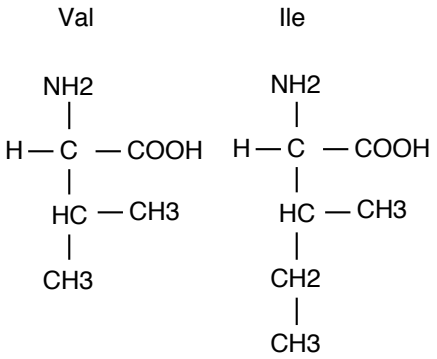


- Transition: A <-> G T <-> C
- Transversions : autres substitutions
- p(transition) > p(transversion)

```
G T T A C G A
G T T - G G A
* * *     * *
```

```
G T T A C G A
G T T G - G A
* * * .   * *
```

Modèles d'évolution



Proteines

Substitutions conservatrices

- Code génétique
  - Asp (GAC, GAU) ↔ Tyr (UAC, UAU) : 1 mutation
  - Asp (GAC, GAU) ↔ Cys (UGC, UGU) : 2 mutations
  - Asp (GAC, GAU) ↔ Trp (UGG) : 3 mutations

- Propriétés physico-chimiques des acides-aminés (acidité, hydrophobicité, encombrement stérique, etc.)

- Matrices de Dayhoff (PAM), BLOSUM: mesures des fréquences de substitutions dans des alignements de protéines homologues
  - PAM 60, PAM 120, PAM 250 (extrapolations à partir de PAM 15)
  - BLOSUM 80, BLOSUM 62, BLOSUM 40 (basé sur des alignements de blocs)

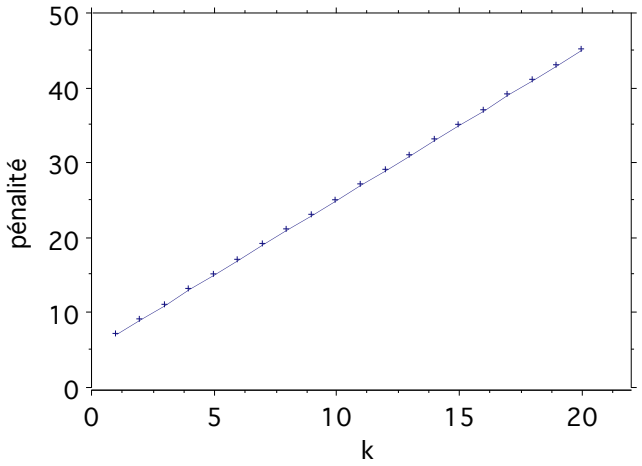
Pondération des gaps

```
TGATATCGCCA        TGATATCGCCA
TGAT---TCCA        TGAT-T--CCA
****    ***        **** *  ***
```

- Gap de longueur *k*: Pénalités linéaires: $w = \delta_o + \delta_e k$

  δ_o : pénalité pour l'ouverture d'un gap

  δ_e : pénalité pour l'extension d'un gap



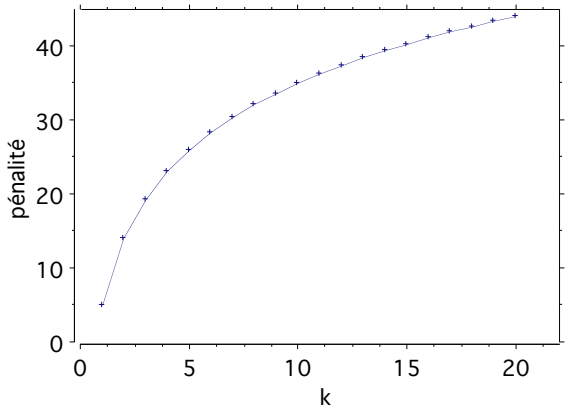
L'alignement

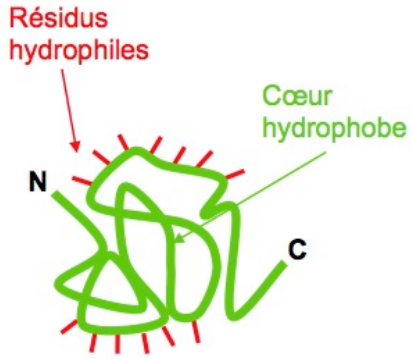
Pondération des gaps

Estimation des paramètres sur des alignements "vrais" (par exemple basés sur l'alignement de structures connues)

Gap de longueur *k*:

- Pénalités logarithmiques: $w = \delta_o + \delta_e \log(k)$





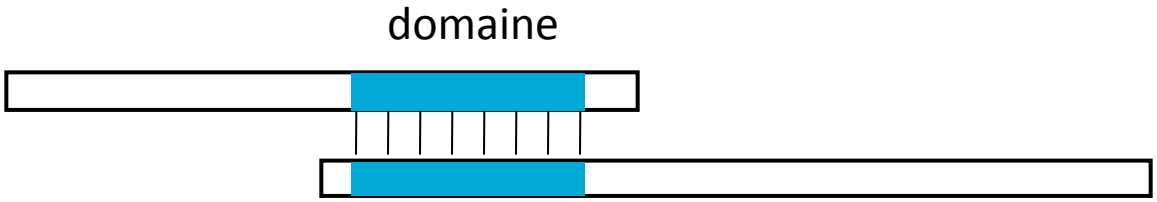
- *w* = f(*log(k), log(PAM), résidus, structure*)
  - *PAM: la probabilité d'un gap augmente avec la distance évolutive*
  - *Résidus, structure: la probabilité d'un gap est plus forte dans une boucle (hydrophile) que dans le cœur hydrophobe des protéines*

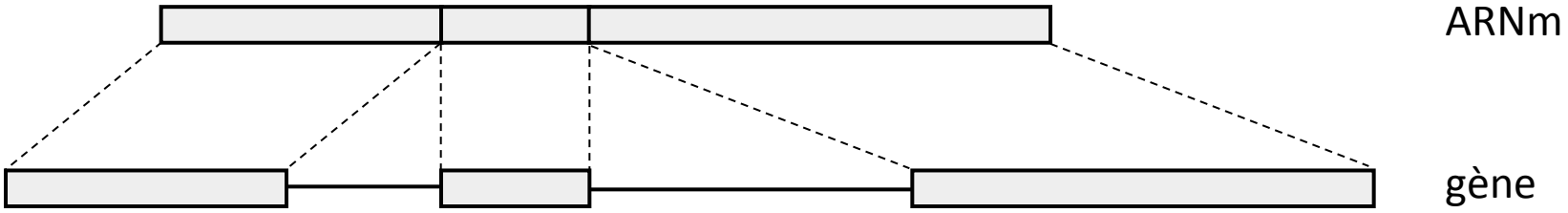
L'alignement

Local vs. Global

- Alignement global (Needlman & Wunsch, 1970)

- Alignement local (Smith & Waterman, 1981 ; FASTA, 1988 ; BLAST, 1990)

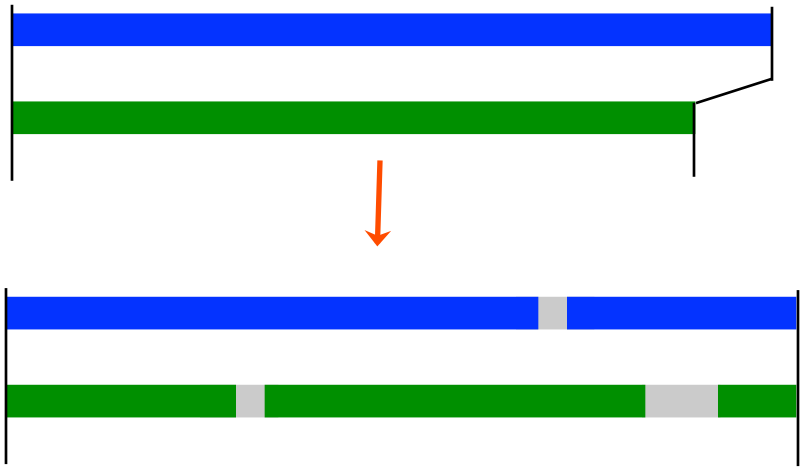




L'alignement

Local vs. **Global**

Alignement de 2 séquences sur la totalité de leur longueur



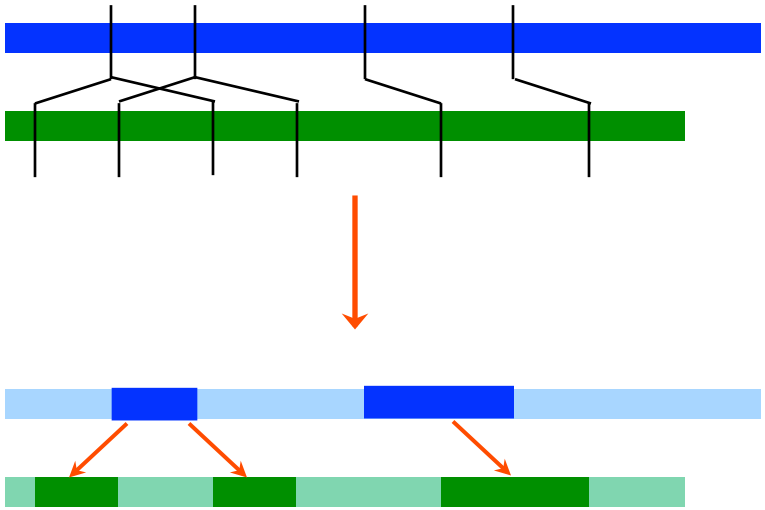
Déterminer le degré de similarité de 2 séquences

Logiciels : ALIGN, LALIGN, Needle

L'alignement

Local vs. Global

Alignement sur des segments de séquences



Comparer une séquence inconnue avec une banque de séquences

FASTA (Fast Alignment Search Tool)
BLAST (Basic Local Alignment Search Tool)

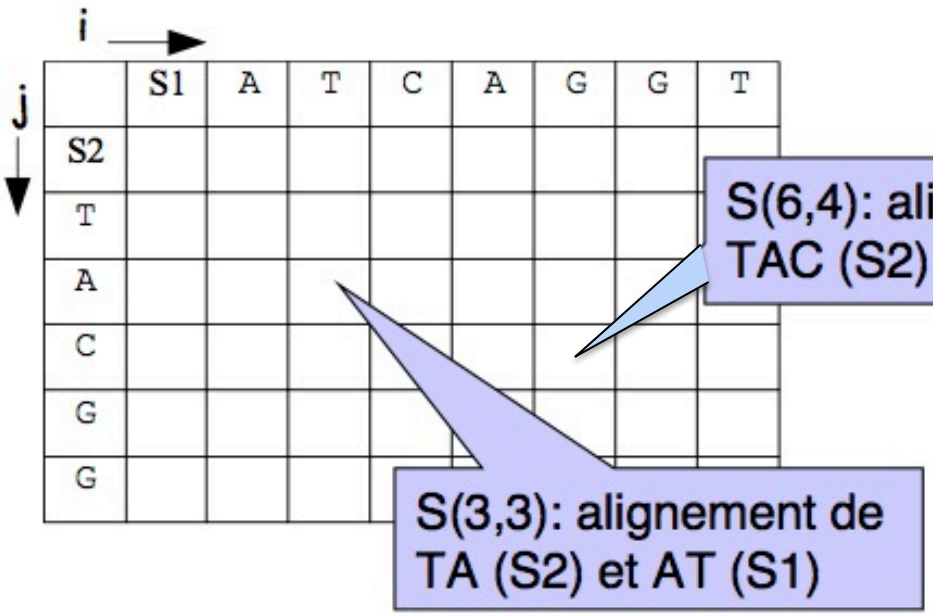
L'alignement : deux séquences

Similarité entre deux séquences

=> Alignement par paire: programmation dynamique

Principe = combinaison des solutions de sous-problèmes

1. **Construction d'une matrice initiale (dimension (n,m))**
2. Remplissage de la matrice
3. Recherche du chemin de score maximum



S(i,j) = coût optimal pour les i (j) premières lettres de S1 (S2)

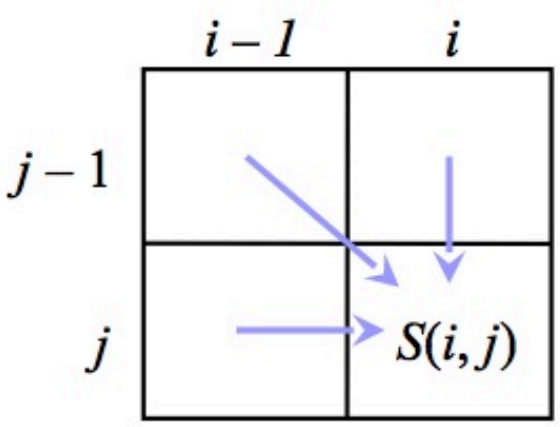
L'alignement : deux séquences

Similarité entre deux séquences

=> Alignement par paire: programmation dynamique

Principe = combinaison des solutions de sous-problèmes

1. Construction d'une matrice initiale (dimension (n,m))
2. **Remplissage de la matrice**
3. Recherche du chemin de score maximum



$$S(i,j) = \max \begin{pmatrix} S(i-1,j) + \delta(a_i, -), \\ S(i-1,j-1) + \delta(a_i, b_j), \\ S(i,j-1) + \delta(-, b_j) \\ [\,0\,] \end{pmatrix}$$

$$[\, S(i,j) < 0 \Rightarrow S(i,j) = 0 \,]$$

L'alignement : deux séquences

Fonction de récurrence :

Sim(i,j) : score optimal entre Seq1(1..i) et Seq2(1..j)

Formule de récurrence :

- **Sim(0,0) = 0**
- **➔ Sim(0,j) = Sim(0,j-1) + Indel**
- **🡻 Sim(i,0) = Sim(i-1,0) + Indel**
- **🡾 : Sim(i,j) = max [**

	j-1	j
i-1		
i		?

Sim(i-1,j-1) + Corresp.ou Substi.
(Corresp. si Seq1(i) = Seq2(j) sinon Substi.)
Sim(i-1,j) + Indel
Sim(i,j-1) + Indel
]

L'alignement : deux séquences

Exemple : on veut aligner 2 séquences

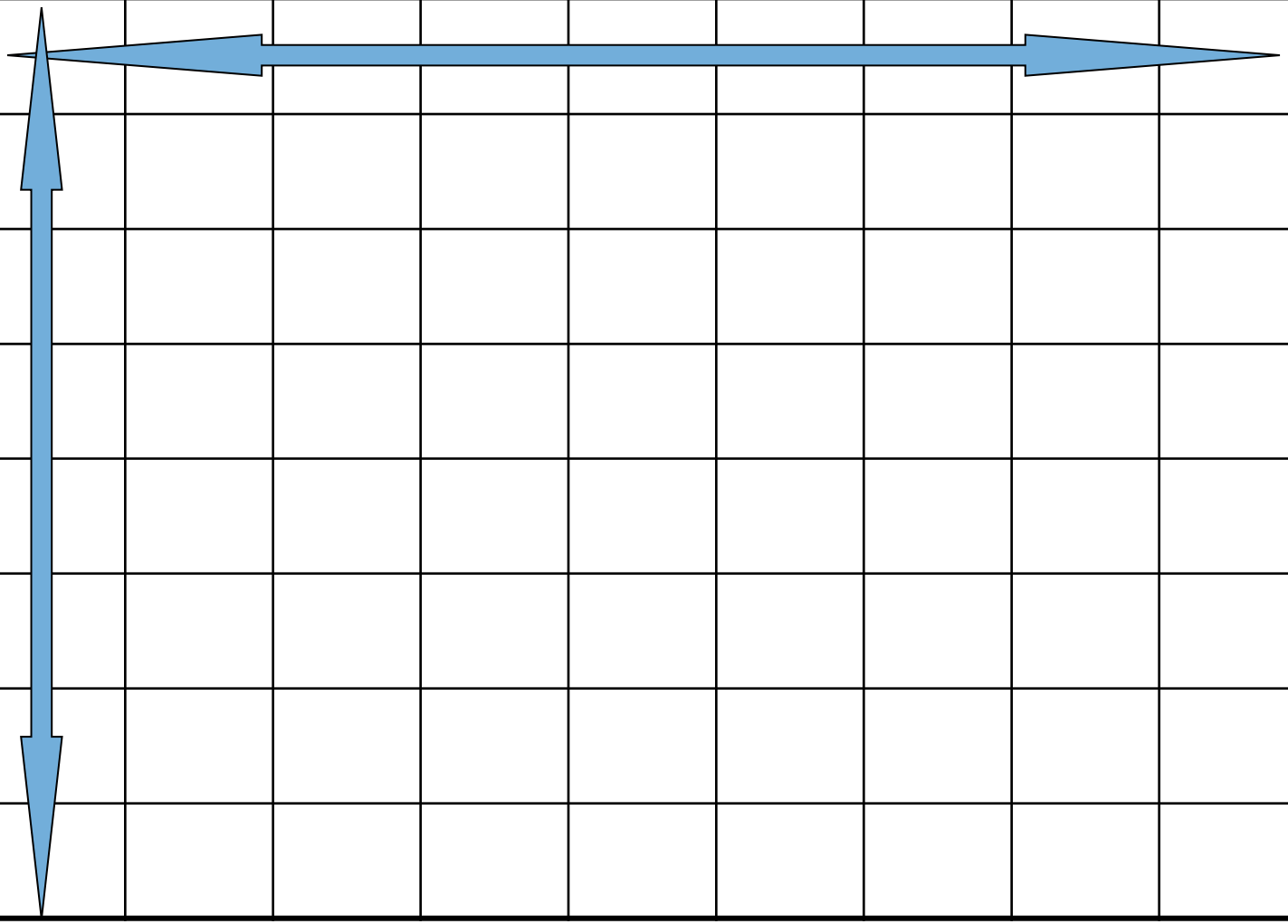
- ATGGCTTA et ATCGCTA

Calcul du score :

- Identité (ou concordance) : 0
- Substitution : -1
- INDEL : -3

L’alignement : deux séquences

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



L’alignement : deux séquences

		A	T	G	G	C	T	T	A
	0								
A									
T									
C									
G									
C									
T									
A									

Sim(0,0) = 0

L'alignement : deux séquences

		A	T	G	G	C	T	T	A
	0	-3	-6	-9	-12	-15	-18	-21	-24
A	-3								
T	-6								
C	-9								
G	-12								
C	-15								
T	-18								
A	-21								

➔Sim(0,j) = Sim(0,j-1) + Indel

🡻 Sim(i,0) = Sim(i-1,0) + Indel

L'alignement : deux séquences

		A	T	G	G	C	T	T	A
	0	-3	-6	-9	-12	-15	-18	-21	-24
A	-3	?							
T	-6								
C	-9								
G	-12								
C	-15								
T	-18								
A	-21								

↘: **Sim(i,j) = max [**

Sim(i-1,j-1) + Corresp.ou Substi.
(Corresp. si Seq1(i) = Seq2(j) **sinon Substi.)**

Sim(i-1,j) + Indel

Sim(i,j-1) + Indel

]

L'alignement : deux séquences

		A	T	G	G	C	T	T	A
	0	-3	-6	-9	-12	-15	-18	-21	-24
A	-3	?							
T	-6								
C	-9								
G	-12								
C	-15								
T	-18								
A	-21								

↓ : **Sim(i-1,j) + Indel**= -3 -3 = -6

→ : **Sim(i,j-1) + Indel**=-3 -3 = -6

↘ : **Sim(i-1,j-1) + Corresp.ou Substi.**
(Corresp. si Seq1(i) = Seq2(j) **sinon Substi.)**
= 0 + 0 (puisque A = A) = 0

Max [-6, -6, 0]

Le max : 0

L’alignement : deux séquences

		A	T	G	G	C	T	T	A
	0	-3	-6	-9	-12	-15	-18	-21	-24
A	-3	0							
T	-6								
C	-9								
G	-12								
C	-15								
T	-18								
A	-21								

L'alignement : deux séquences

		A	T	G	G	C	T	T	A
	0	-3	-6	-9	-12	-15	-18	-21	-24
A	-3	0	?						
T	-6								
C	-9								
G	-12								
C	-15								
T	-18								
A	-21								

↓ : **Sim(i-1,j) + Indel**= -6 -3 = -9

→ : **Sim(i,j-1) + Indel**= 0 -3 = -3

↘ : **Sim(i-1,j-1) + Corresp.ou Substi.**
(Corresp. si Seq1(i) = Seq2(j) **sinon Substi.)**
= -3 - 1 (puisque T ≠ A) = -4

Le max : -3

L’alignement : deux séquences

		A	T	G	G	C	T	T	A
	0	-3	-6	-9	-12	-15	-18	-21	-24
A	-3	0	-3						
T	-6								
C	-9								
G	-12								
C	-15								
T	-18								
A	-21								

L'alignement : deux séquences

		A	T	G	G	C	T	T	A
	0	-3	-6	-9	-12	-15	-18	-21	-24
A	-3	0	-3	-6	-9	-12	-15	-18	-21
T	-6	-3	0	-3	-6	-9	-12	-15	-18
C	-9	-6	-4	-1	-4	-6	-9	-12	-15
G	-12	-9	-7	-4	-1	-4	-7	-10	-13
C	-15	-12	-10	-7	-4	-1	-4	-7	-10
T	-18	-15	-13	-10	-7	-4	-1	-4	-7
A	-21	-18	-16	-13	-10	-7	-4	-2	-4

L'alignement : deux séquences

Principe = combinaison des solutions de sous-problèmes

1. Construction d'une matrice initiale (dimension (n,m))
2. Remplissage de la matrice
3. **Recherche du chemin de score maximum**

Identité (ou concordance) : 0
Substitution : -1
INDEL : -3

2 Alignements possibles :

ATGGC**TTA**
ATCGC**-TA**

ATGGCTTA
ATCGCT-A

SCORE = -4

		A	T	G	G	C	T	T	A
	0	-3	-6	-9	-12	-15	-18	-21	-24
A	-3	0	-3	-6	-9	-12	-15	-18	-21
T	-6	-3	0	-3	-6	-9	-12	-15	18
C	-9	-6	-4	-1	-4	-6	-9	-12	-15
G	-12	-9	-7	-4	-1	-4	-7	-10	-13
C	-15	-12	-10	-7	-4	-1	-4	-7	-10
T	-18	-15	-13	-10	-7	-4	-1	-4	-7
A	-21	-18	-16	-13	-10	-7	-4	-2	-4

L'alignement : deux séquences

Algorithme de programmation dynamique

Alignement global: Needleman & Wunsh (1970)

- Trouver le meilleur alignement sur toute leur longueur
- Utilisé pour des séquences trés similaires et ~ de même longueur

Alignement local: Smith & Waterman (1981)

- Chercher les régions les plus similaires dans les 2 séquences
- Permet de trouver des sous-séquences ayant des relations biologiques
- Plus adapté pour des séquences avec un faible degré de similarité ou de tailles différentes

Heuristiques :

- FASTA - BLAST

L'alignement multiple

- Alignements multiples plus informatifs que les alignements de 2 séquences

- Alignement = point de départ pour les analyses phylogénétiques => qualité de l'alignement très important

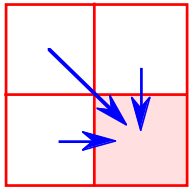
- Alignement multiple global : alignement 2 à 2 est étendu pour inclure 3 seq. ou plus
  - Logiciels : CLUSTALW, MUSCLE, MAFFT, T-COFFEE, DIALIGN, PRANK, . . .

- Alignement multiple local : recherche de domaines/régions conservés
  - Logiciels : BLOCKS Web site, eMOTIF, GIBBS, HMMER, . . .

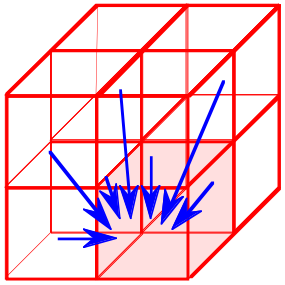
L'alignement multiple

Programmation dynamique

- La généralisation de l'algorithme N&W au traitement simultané de plus de deux séquences est théoriquement possible mais **inexploitable** en pratique.



Alignement de deux séquences : trois choix



Alignement de trois séquences : sept choix

- Pour un alignement de *n* séquences le nombre de chemins possibles pour chaque case est de $2^n - 1$.

- On a une croissance **exponentielle** du temps de calcul et de l'espace mémoire requis en fonction du nombre de séquences.
  - ⇨ Utilisation de méthodes **heuristiques**.

L'alignement multiple

Programmation dynamique

- Utilisation d'heuristiques : approximation de l'alignement optimal

- Méthodes d'alignement global:
  - Approche progressif (ex : CLUSTALW)
  - Amélioration de l'approche progressive : raffinement itératif, notion de consistance (ex : MUSCLE, MAFFT, PROBCONS, T-COFFEE)
  - Autres approches d'alignement multiple : structurelle (ex : STAMP), itérative (ex : HMMER, SAGA), principe « diviser pour régner » (ex : DCA), utilisation d'informations externes (ex : PRALINE, Expresso)

- Propriétés des méthodes progressives
  - Rapide
  - Besoin de peu de mémoire
  - Bonnes performances avec des séquences conservées

L'alignement multiple

Alignement progressif

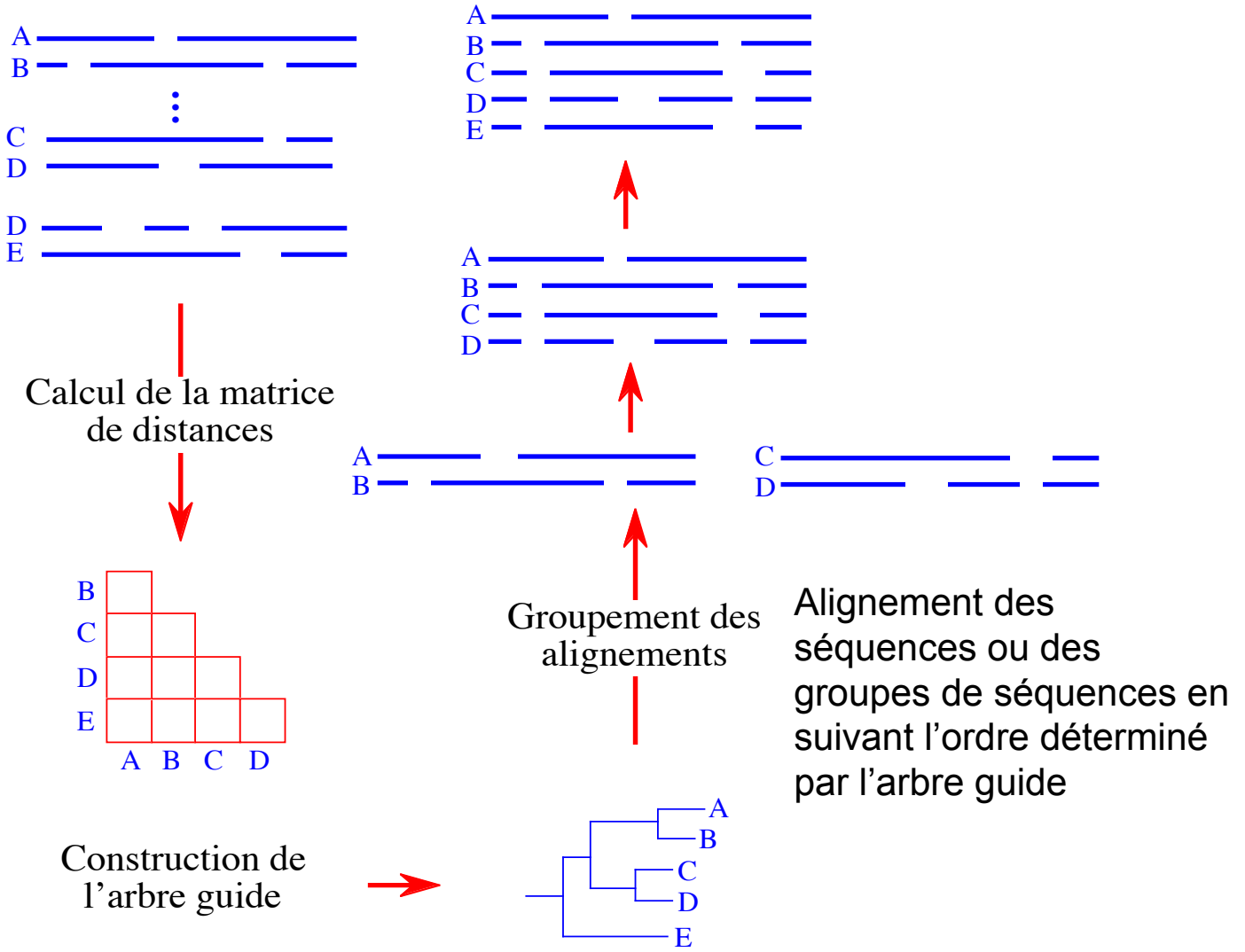
- Approche consistant à construire **itérativement** l'alignement multiple en groupant des alignements de paires de séquences.

- Ce genre de méthodes comporte trois étapes :
  - L'alignement des paires de séquences.
  - Le groupement des séquences.
  - Le groupement des alignements (alignement progressif).

- **CLUSTAL** (Higgins, Sharp 1988, Thompson *et al.*, 1994), l'un des programmes d'alignements multiples le plus utilisé à l'heure actuelle utilise cette approche.
- MULTALIN, PILEUP, T-Coffee

L'alignement multiple

Alignement progressif



L'alignement multiple

Pénalités en fonction de la position

CLUSTAL introduit des pondérations qui sont dépendantes de la **position** des gaps.

- Diminution de la pénalité à l'**emplacement** de gaps préexistants.
- Augmentation de la pénalité **au voisinage** (8 résidus) de gaps préexistants.
- Réduction de la pénalité au niveau de régions contenant des suites d'acides aminés **hydrophiles** (≥ 5 résidus).
- Modification spécifiques en fonction des **acides aminés** présents (*e.g.*, la pénalité est plus faible avec Gly, Asn, Pro).

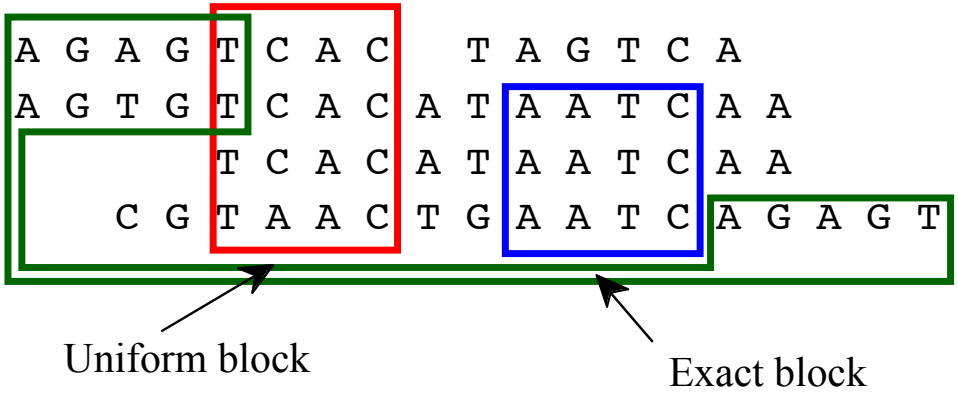
Ces pondérations sont prises en compte au moment du groupement des alignements.

L'alignement multiple

Dialign

Morgenstern et al. 1996 PNAS 93:12098

- Recherche de blocs similaires (≠ exact) sans gap entre les séquences



- Sélection de la meilleure combinaison possible de blocs similaires (uniformes ou non) consistents : heuristique (Abdeddaim 1997)
- Alignement ancré sur les blocs
- Plus lent que alignement progressif, mais meilleur alignement quand les séquences contiennent de grands indels; ne cherche pas à aligner des régions non-alignables

L’alignement multiple

Amélioration des alignements progressifs

- Alignement progressif
- Lors des alignements intermédiaire, prise en compte de tous les alignements deux à deux (globaux et locaux)
- Possibilité d'incorporer d'autres informations (structure, etc.)

L'alignement multiple

Amélioration des alignements progressifs

Les schémas de score utilisés par les algorithmes d'alignements par paires sont un élément très influent dans l'algorithme progressif :

- méthodes basées sur les matrices (ClustalW, Muscle, Kalign, . . .)
- méthodes basées sur la consistance (T-Coffee, MAFFT, ProbCons, . . .)

Contraintes de consistance:

- prises en compte de combinaisons consistantes de séquences

```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT
SeqC GARFIELD THE VERY FAST CAT
SeqD THE FAT CAT
```

L'alignement multiple

Notredame, Higgins, Heringa (2000) JMB 302:205

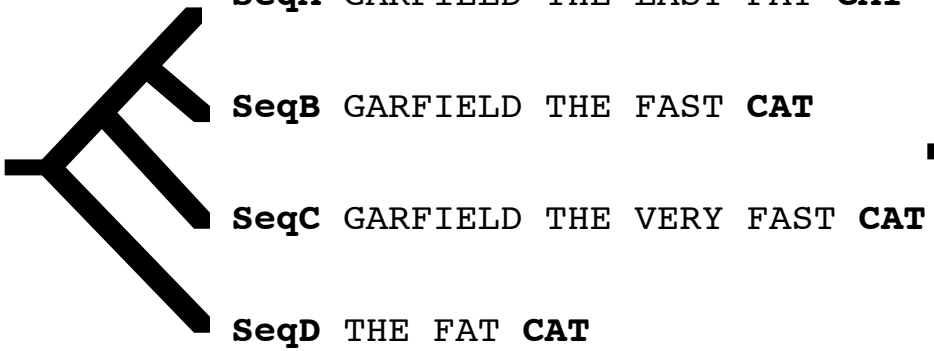
Amélioration des alignements progressifs

Alignements par paire

```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT
SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FAT CAT
SeqD ---------THE ---- FAT CAT
```

```
SeqB GARFIELD THE ---- FAST CAT
SeqC GARFIELD THE VERY FAST CAT

SeqB GARFIELD THE FAST CAT
SeqD ---------THE FA-T CAT

SeqC GARFIELD THE VERY FAST CAT
SeqD ---------THE ---- FA-T CAT
```

Alignement progressif



Alignement A-B retenu:

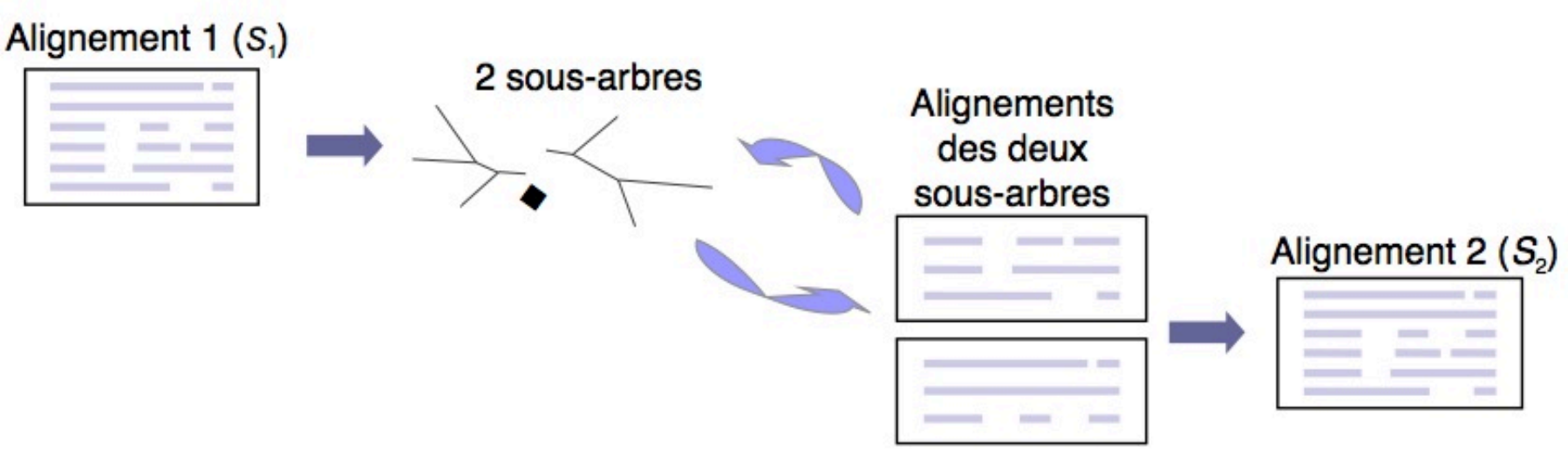
```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST ---- CAT
SeqC GARFIELD THE LAST FAST CAT
SeqB -------- THE ---- FA-T CAT
```

L'alignement multiple

Amélioration des alignements progressifs

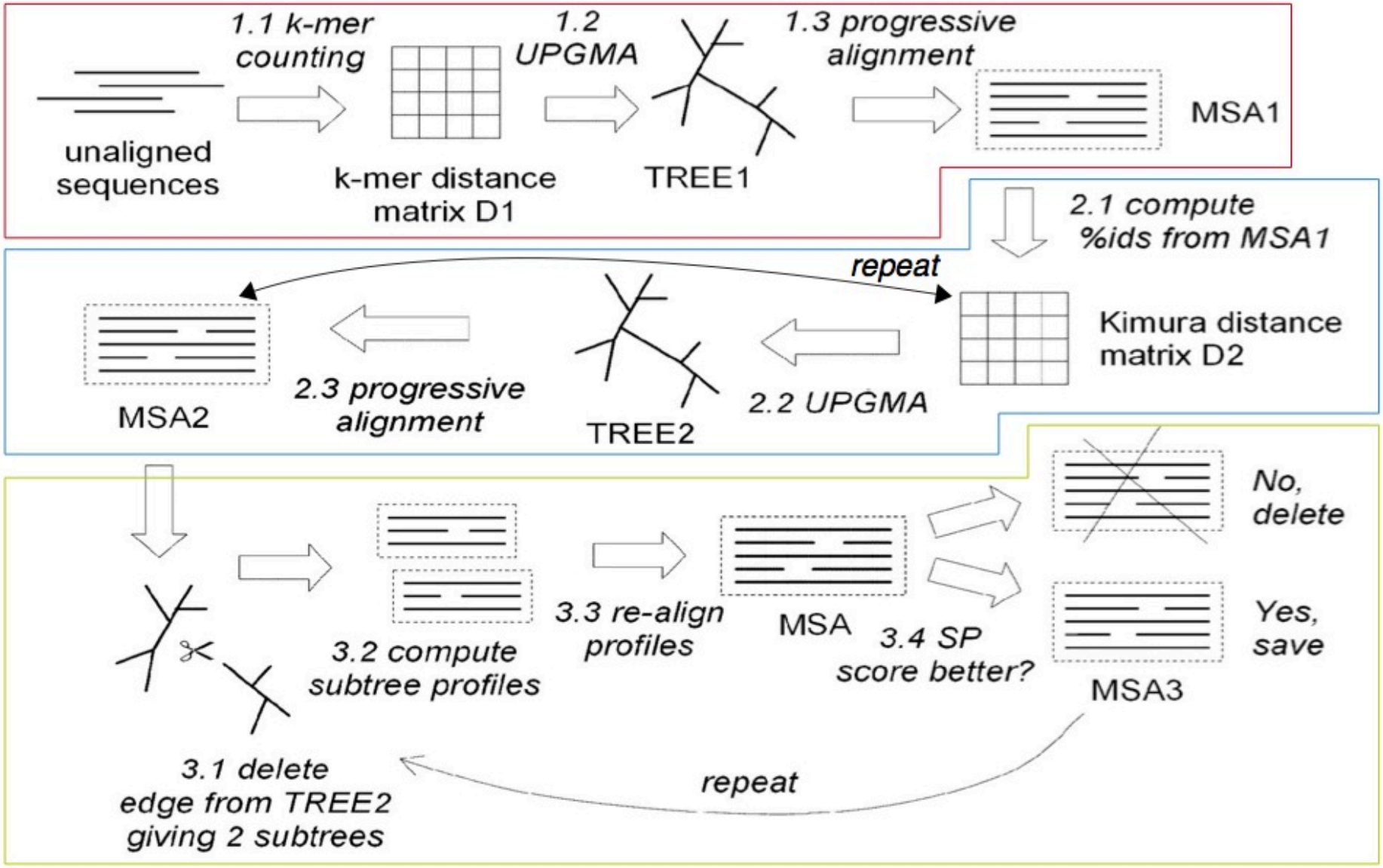
Le raffinement itératif

- Pb majeur de l'alignement progressif = une erreur faite au début de l'alignement ne peut être corrigée par la suite
  => étape de raffinement ajoutée (construction itérative de l'arbre guide, division de l'arbre guide en sous-arbre, ...)
- Objectif = améliorer le score d'alignement global en réalignant des sous- groupes de séquences de manière répétée puis en alignant ces sous-groupes dans l'alignement global de toutes les séquences



L'alignement multiple

Amélioration des alignements progressifs : exemple MUSCLE



L'alignement multiple

Amélioration des alignements progressifs

Ex MAFFT

Même principe que MUSCLE

- Etape d'alignement par paire : utilisation de l'algorithme Transformée de Fourrier rapide (FFT) et d'un algorithme de programmation dynamique
- 3 types de versions de MAFFT
  - Méthode progressive (I)
  - Méthode de raffinements itératif utilisant le score WSP (weighted sum-of- pairs) (II)
  - Méthode de raffinements itératif utilisant le score WSP et des scores bases sur la consistance (III)
- Méthodes pour construire les alignements par paire
  - Algorithme FFT (I et II)
  - Algorithme d'alignement global de Needleman et Wunsch (III)
  - Algorithme d'alignement local de Smith et Waterman (III)

L'alignement multiple

Published online March 23, 2006

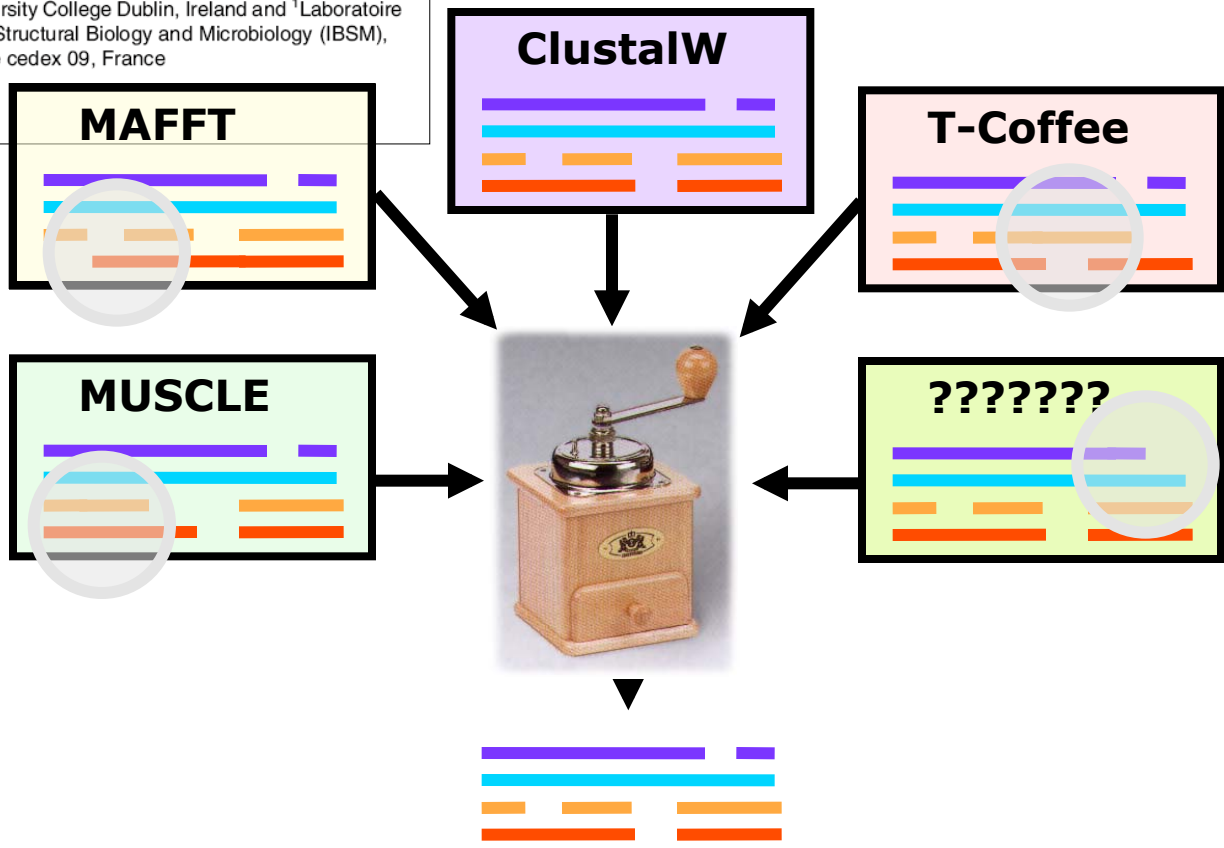
1692–1699 *Nucleic Acids Research, 2006, Vol. 34, No. 6*
doi:10.1093/nar/gkl091

M-Coffee: combining multiple sequence alignment methods with T-Coffee

Iain M. Wallace, Orla O'Sullivan, Desmond G. Higgins and Cedric Notredame[1,*]

The Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Ireland and [1]Laboratoire Information Génomique et Structurale, CNRS UPR2589, Institute for Structural Biology and Microbiology (IBSM), Parc Scientifique de Luminy, 163 Avenue de Luminy, 13288, Marseille cedex 09, France

Received January 19, 2006; Revised February 7, 2006; Accepted March 7, 2006



L'alignement multiple

Nettoyage de l'alignement

- Alignement pouvant contenir des grandes zones de gap
- Séquences de longueurs différentes
- ....

Nettoyage de l'alignement pour avoir un alignement de meilleur qualité
= suppression des régions non informatives, mal conservées

-> Subjectivité du nettoyage manuel

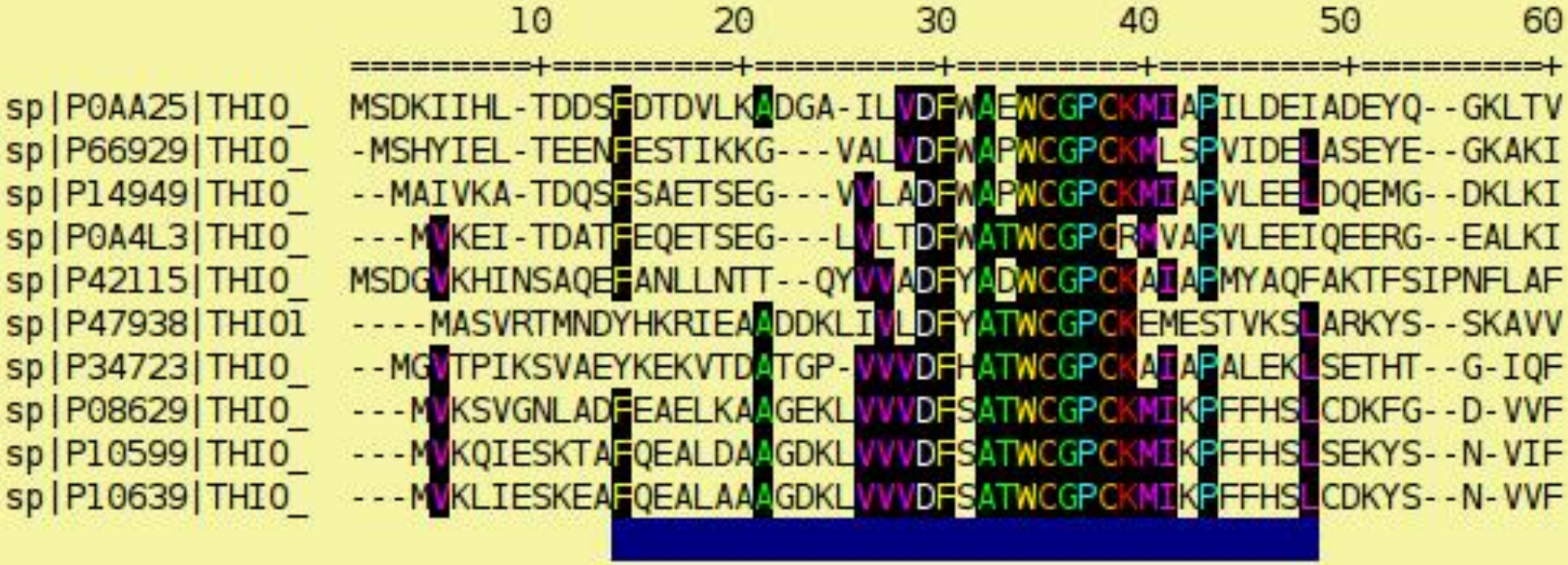
Logiciels: Trimal, Gblocks, ...

trimAl
A tool for automated alignment trimming

L'alignement multiple

Gblocks 0.91b Results

Processed file: **gblocks**
Number of sequences: **10**
Alignment assumed to be: **Protein**
New number of positions: **60** (selected positions are underlined in blue




```
                      130
                    =========+
sp|P0AA25|THIO_  ----------
sp|P66929|THIO_  ----------
sp|P14949|THIO_  ----------
sp|P0A4L3|THIO_  ----------
sp|P42115|THIO_  KEKAAAAGSS
sp|P47938|THIO1  ----------
sp|P34723|THIO_  ----------
sp|P08629|THIO_  ----------
sp|P10599|THIO_  ----------
sp|P10639|THIO_  ----------
```

Parameters used

Minimum Number Of Sequences For A Conserved Position: 6
Minimum Number Of Sequences For A Flanking Position: 8
Maximum Number Of Contiguous Nonconserved Positions: 8
Minimum Length Of A Block: 5
Allowed Gap Positions: With Half
Use Similarity Matrices: Yes

Flank positions of the 2 selected block(s)

Flanks: [14 48] [65 89]

New number of positions in gblocks-gb: **60** (46% of the original 130 positions)

Blast

- BLAST = Basic Local Alignment Search Tool

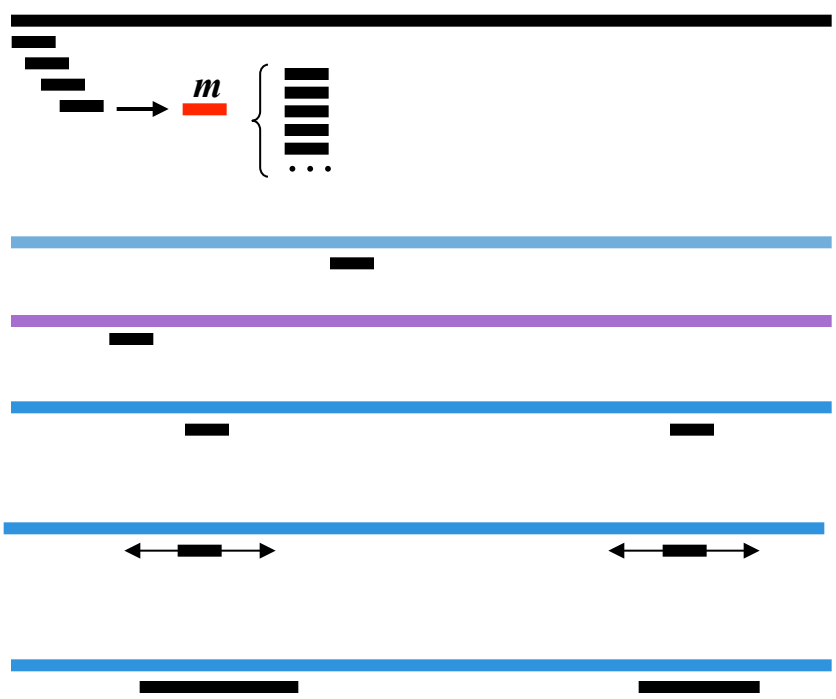
- **Objectif** : retrouver, dans une banque de référence, une ou des séquences homologues à ma séquence en entrée

- Vocabulaire de Blast :
  - **Query** : la séquence requête
  - **Subject** : la séquence dans la banque de référence
  - **Hit** : séquence de la banque similaire à la query
  - **HPS** (High Scoring Pairs) : résultats de blasts : meilleurs alignements locaux par paires pour le hit
  - **Score S** : Détermine le degré de ressemblance entre deux séquences (basé sur la matrice utilisée et sur les pénalités de gaps). Plus le score est élevé, plus les séquences sont proches
  - **Bit-Score S'** : score normalisé (permet de comparer différents résultats de blast entre eux)

Blast

BLAST : Basic Local Alignment Search Tool (Altschul *et al*, 1990)

- Détermination d'une longueur de mot : *w* = 2 ou 3 acides aminés pour les protéines , 11 pour les nucléotides ;
- Hachage de la séquence « requête » en mot de taille *w*



Le programme effectue une extension sans gap de l'alignement dans les deux sens. L'extension s'arrête quand le score du mot étendu diminue de plus qu'un seuil *X* fixé.

BLAST trouve donc des alignements locaux, appellés HSP (Hight-scoring Segment Pair).

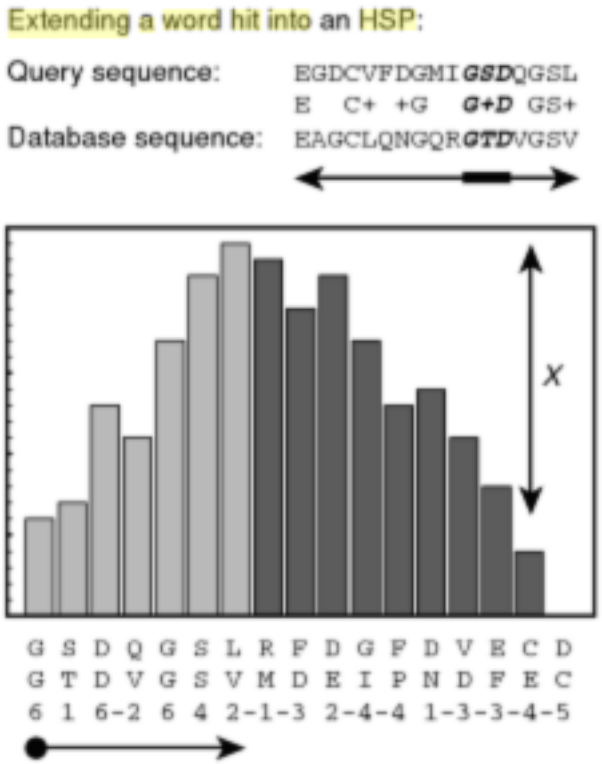
Blast

- Recherche de mots (oligonucléotides/oligopeptides) partagés entre la **query** et la séquence de la banque :
  - Taille par défaut des mots : ADN : 11 /Protéines : 3
  - Pour les protéines, blast recherche à la fois les mots exacts mais aussi les mots synonymes (voisins)

- Quand un mot est trouvé : blast essaie d'étendre le **hit** en ajoutant des paires de bases/ acides aminés. L'alignement grandissant obtient un **score** :
  - Score nucléotides calculé selon la matrice de substitution (par défaut : +1/-3)
  - Score protéines calculé selon la matrice BLOSUM62 par défaut

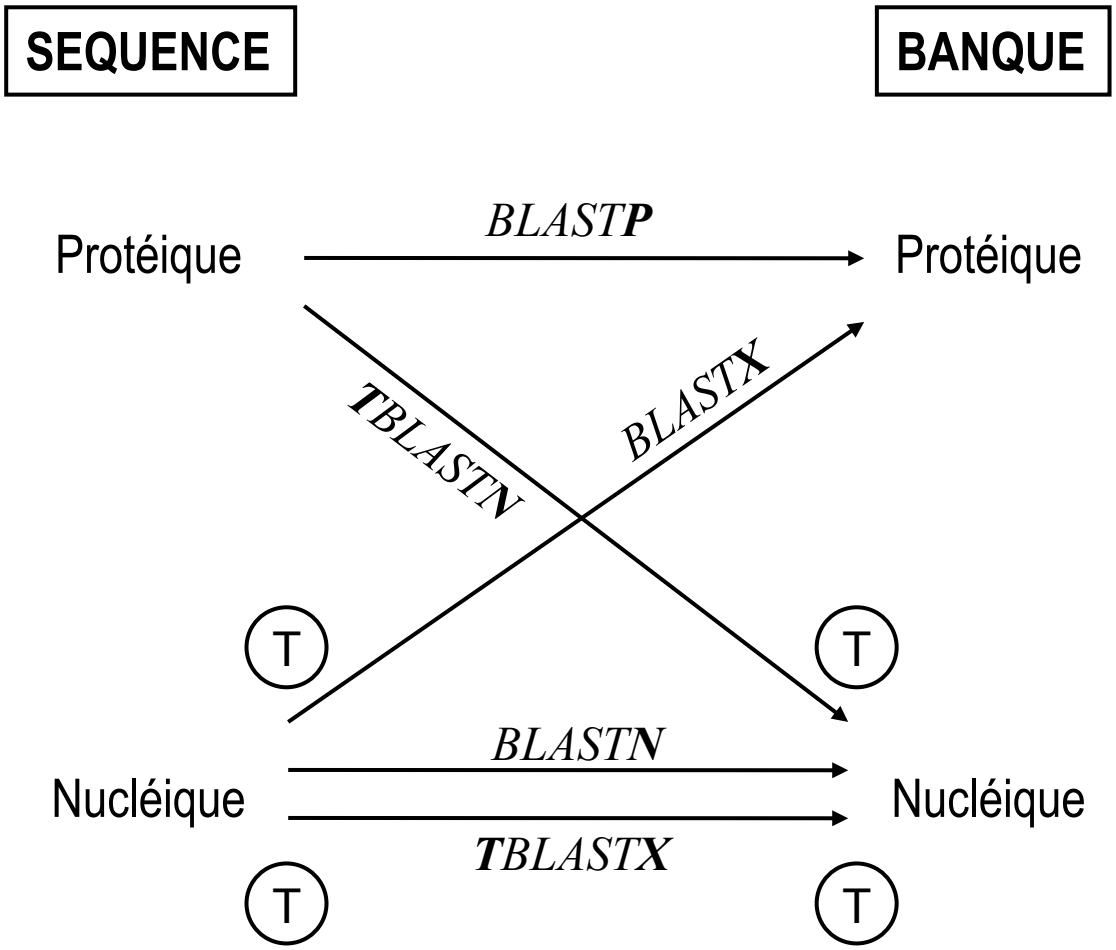
Blast

- Si le score s'écarte de trop (seuil défini par défaut, *zone X figure*) du score le plus haut établi pour l'alignement, blast arrête d'étendre l'alignement. Cette procédure d'extension est répétée à l'autre extrémité de l'alignement.
- On obtient alors un alignement sans gap, appelé **HSP** (High Scoring Pair) qui est conservé si sa e-value est en dessous du seul défini.



Extension de l'alignement vers la droite

Blast


SEQUENCE
BANQUE
Protéique
BLASTP
Protéique
TBLASTN
BLASTX
T
T
BLASTN
Nucléique
Nucléique
TBLASTX
T
T

Blast

- E-value
  - $E = Y Z K e^{-\lambda S}$
    - s = score authentique
    - Y = longueur de la séquence
    - Z = taille de la banque
    - K et λ = constante
- E = Probabilité d'observer au hasard ce score à travers la banque de séquences considérée.

On considère que :

E-Value	Conclusion
< e-100	match exact (même gène, même espèce)
e-100...e-50	gènes quasimment identiques (allèles, mutations, espèces voisines)
e-50...0.1	relation plus lointaines
> 0.1	séquenes en général inintéressantes

Blast

E-value = nombre d'alignements attendus par hasard ayant un score supérieur au score obtenu dans la banque considérée

- Plus la valeur est faible, plus l'alignement est fiable
- Dépend du nombre total de résidus contenus dans la banque
- Ces valeurs ne sont pas comparables entre deux banques

Le score HSP = Σ bonus pour appariement

- Σ malus pour mésappariement
- Σ malus pour ouverture d'un gap
- Σ malus pour extension d'un gap

Blast

Une faible valeur de la E value indique que le résultat n'est pas du au hasard

Sequences producing significant alignments:	Score (bits)	E Value
swissprot:CTRB_HUMAN Chymotrypsinogen B precursor (EC 3.4.21.1).	433	e-121
swissprot:CTR2_CANFA Chymotrypsinogen 2 precursor (EC 3.4.21.1).	386	e-107
swissprot:CTRB_RAT Chymotrypsinogen B precursor (EC 3.4.21.1).	383	e-106
swissprot:CTRB_BOVIN Chymotrypsinogen B (EC 3.4.21.1).	348	4e-96
swissprot:CTRA_BOVIN Chymotrypsinogen A (EC 3.4.21.1).	330	1e-90
swissprot:CTRA_GADMO Chymotrypsin A precursor (EC 3.4.21.1).	286	2e-77

Un score élevé, ou mieux une série de scores élevés, indique une relation (homologie ou famille de gènes)

L'alignement : Travaux Pratiques

Deux fragments du génome d'*Escherichia coli* :

>fragment1 : 1100 bp - from 779751 bp to 780850 bp
>fragment2 : 400 bp - from 2518951 bp to 2519350 bp

Réaliser l'alignement global et local de ces 2 fragments,
en utilisant respectivement

- Stretcher (http://emboss.bioinformatics.nl/cgi-bin/emboss/stretcher) global
- Matcher (http://emboss.bioinformatics.nl/cgi-bin/emboss/matcher) local.

Donnez les résultats de vos 2 alignements. Que pouvez-vous déduire ?

Vous pouvez visualiser les alignements avec le logiciel seaview.

Rmq : pour le logiciel Matcher : vous devrez modifier quelques paramètres dans *"Additional section"* puis *"Number of alternative matches*" : choisir la valeur 5 (le logiciel cherchera les 5 meilleurs alignements possibles).

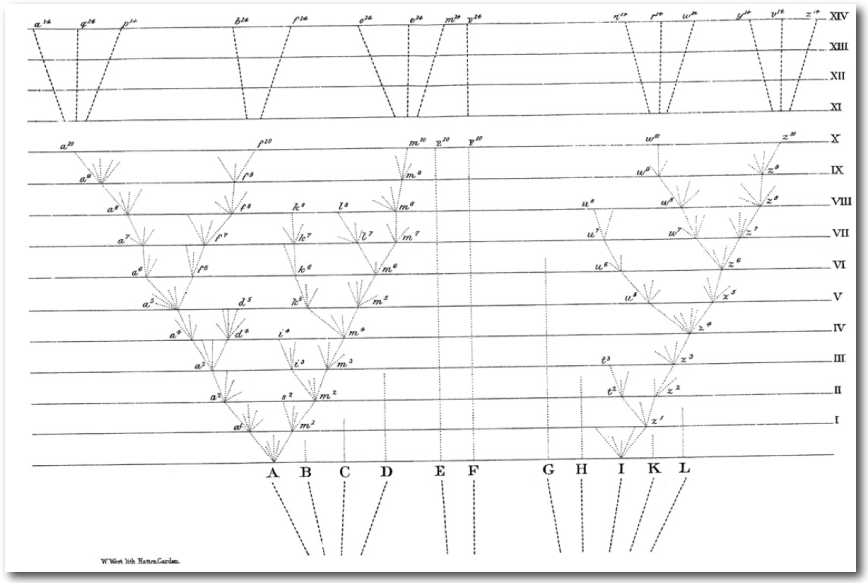
PARTIE I : Phylogénie moléculaire

LES ÉLÉMENTS D'UNE PHYLOGENIE

- Les caractères
- Le jeu de données
- Les banques et formats
- Les alignements
- Les arbres

Les arbres

Les premières représentations



Sens de lecture

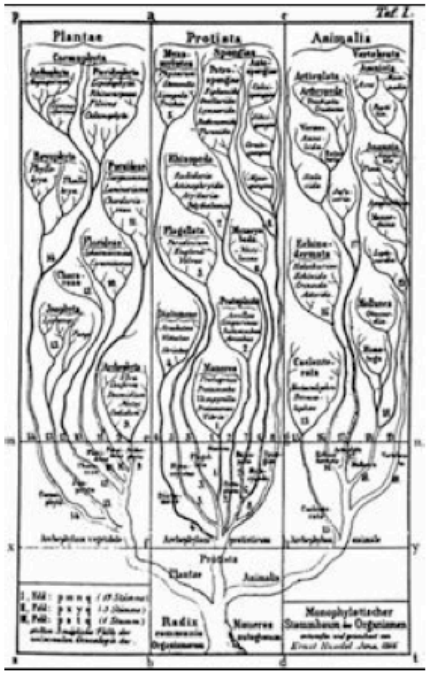
Lamarck, 1809

Darwin, 1859

Les arbres

Volonté de représenter le développement paléontologique des organismes par analogie avec l'ontogénie ou histoire du développement individuel

Haeckel, 1860

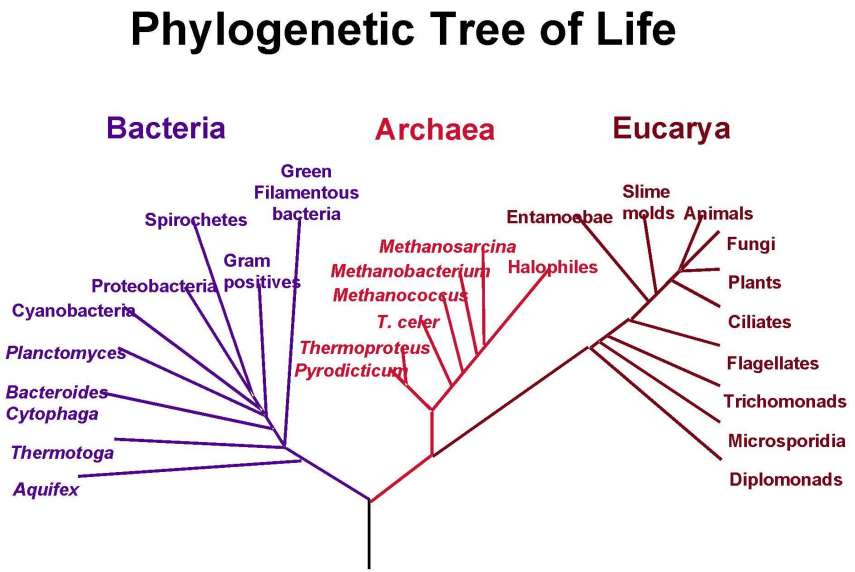


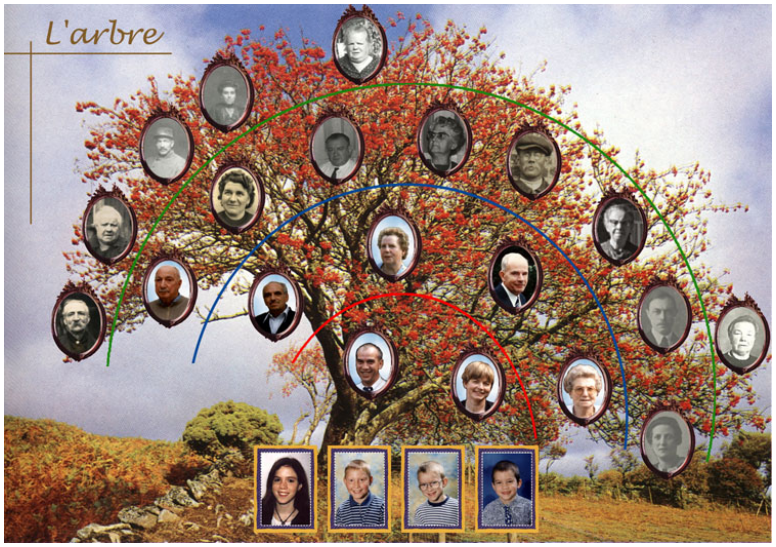
Sens de lecture

Ontogénèse : développement progressif d'un être vivant de sa conception à sa mort
Phylogenèse : caractérise le développement d'une espèce.

Les arbres

Différences entre un arbre phylogénétique et un arbre généalogique.

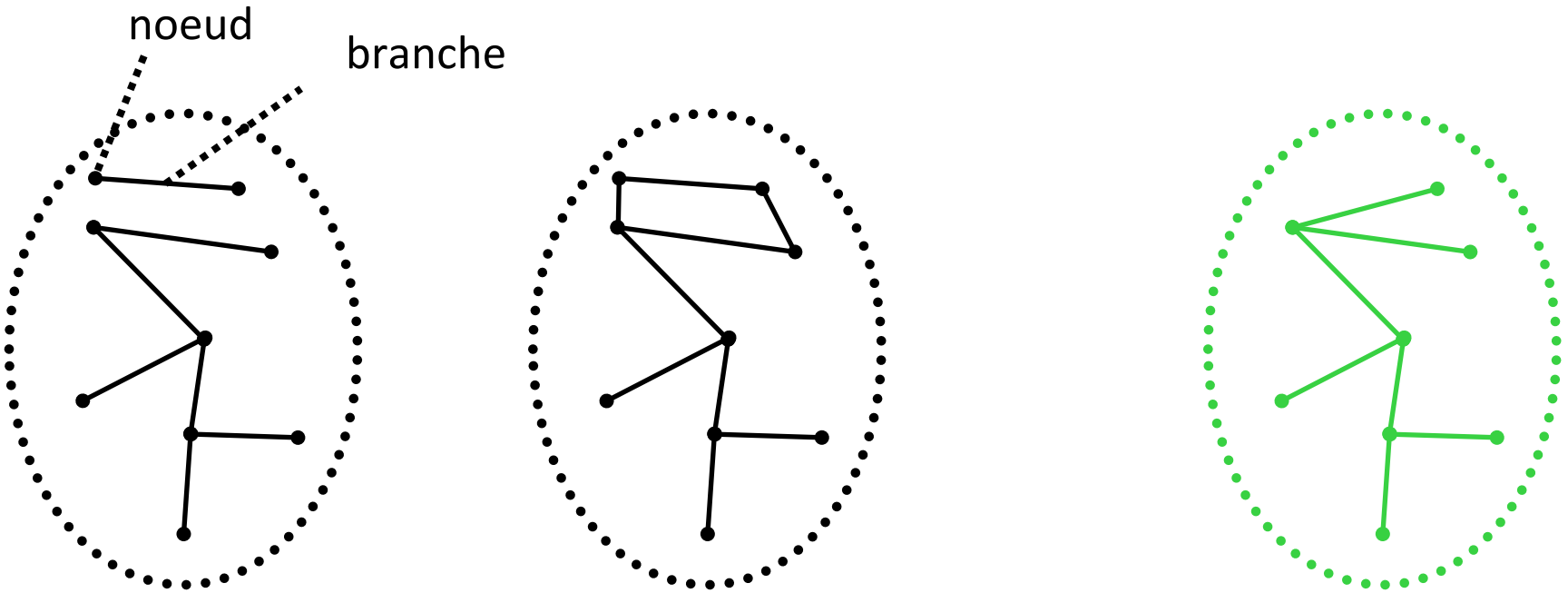




1. il est construit à partir de groupes, d'espèces, et non d'individus,
2. il exprime les relations de parenté et non de descendance ; c'est-à-dire que l'identification des ascendants importe moins que la parenté relative de deux espèces contemporaines,
3. il infère le passé et en ce sens se pose comme hypothèse.

Les arbres

- Les relations évolutives sont représentées en créant un structure arborescente appelée phylogénie ou **arbre** qui illustre les relations entre les espèces (ou les séquences).
- Arbre = réseau connexe non cyclique

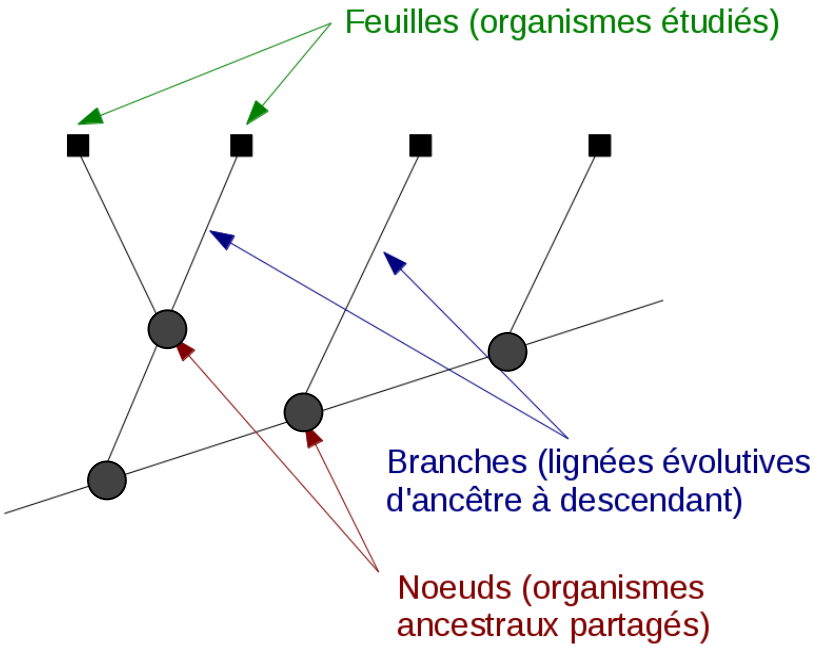


Réseau non connexe non cyclique

Réseau connexe cyclique

Réseau connexe non cyclique

Les arbres



HTU : Hypothetical Taxonomic Unit

OTU : Operational Taxonomic Unit

Branches représentent l'évolution moléculaire

Internes : reliant 2 nœuds

Externes : nœud à feuille

Longueur de branche :

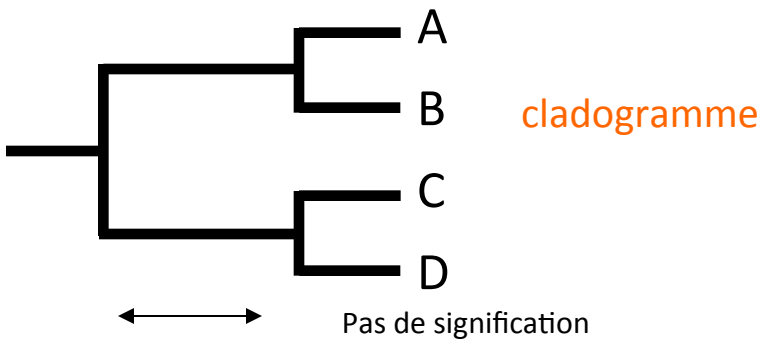
Nombre moyen de remplacements d'un résidu par un autre sur chaque site de la molécule étudiée le long de cette branche (unité : nb de substitutions/site)

La longueur de branche est proportionnelle à l'éloignement en terme d'évolution entre les séquences et leur ancêtre

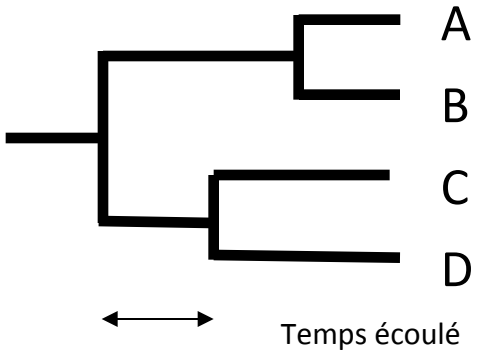
Les arbres

Les méthodes **cladistiques et phénétique** construisent un arbre (dendrogramme)

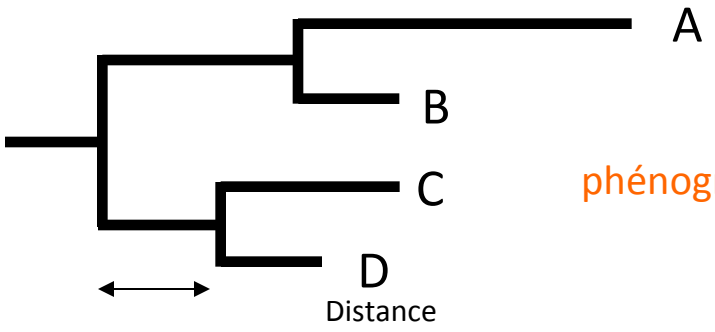
- cladogramme - un dendrogramme exprimant les relations phylogénétiques entre taxa et construit à partir de l'analyse cladistique;
- phénogramme - un dendrogramme obtenu par méthodes de distance où les relations entre taxa expriment des degrés de similitude globale;
- phylogramme - un dendrogramme dont la longueur des branches est proportionnelle au nombre de changements évolutifs.



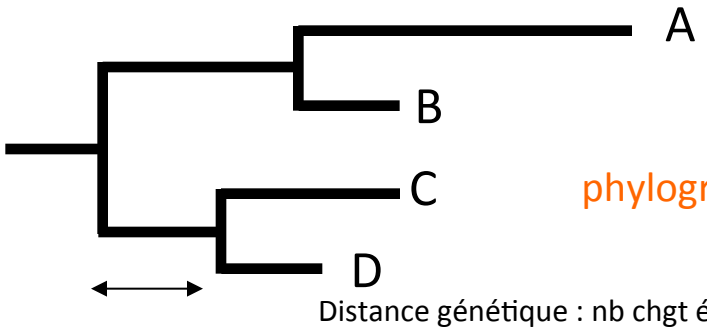
cladogramme



arbre ultramétrique

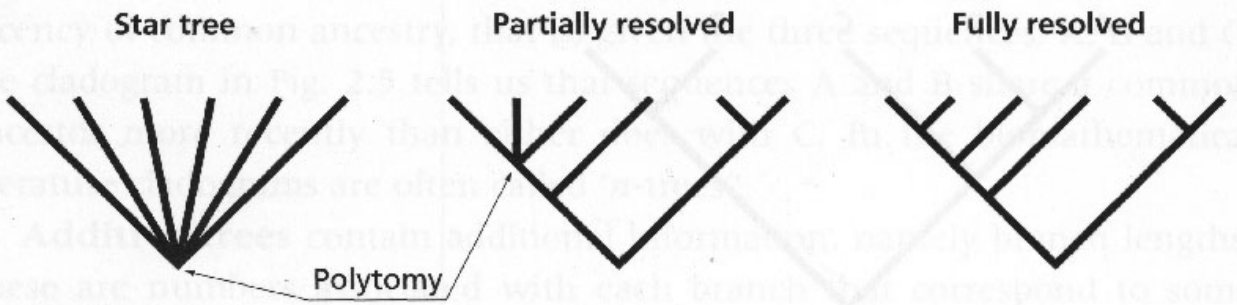


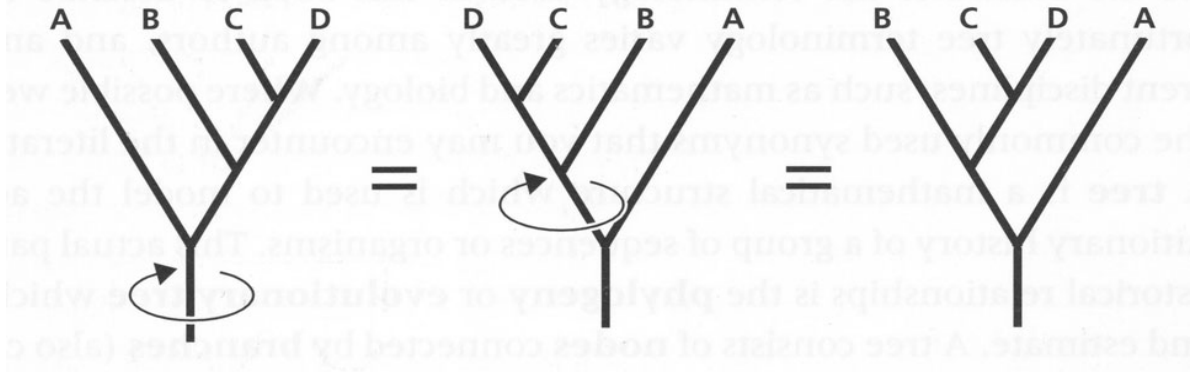
phénogramme



phylogramme

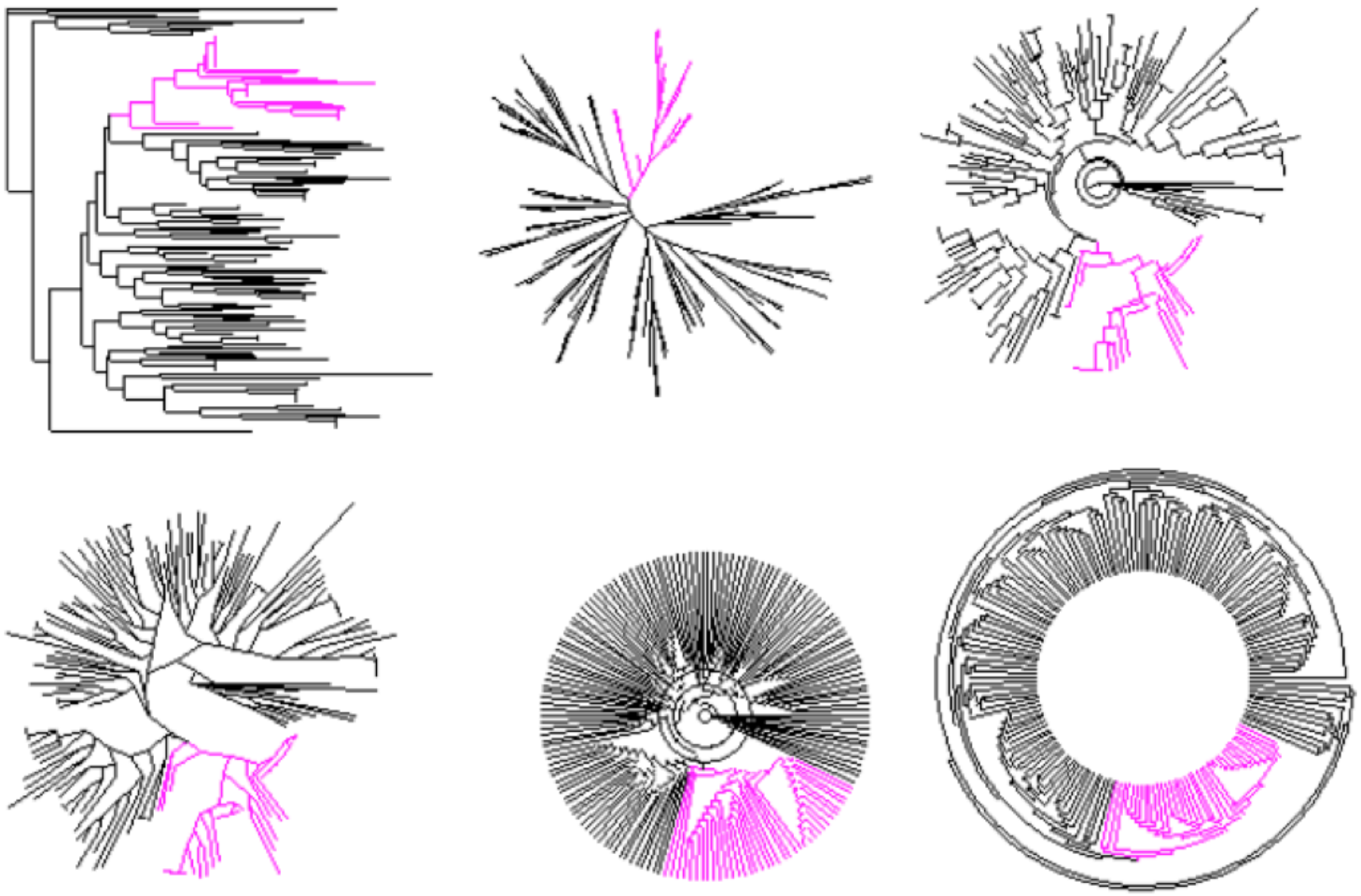
Les arbres





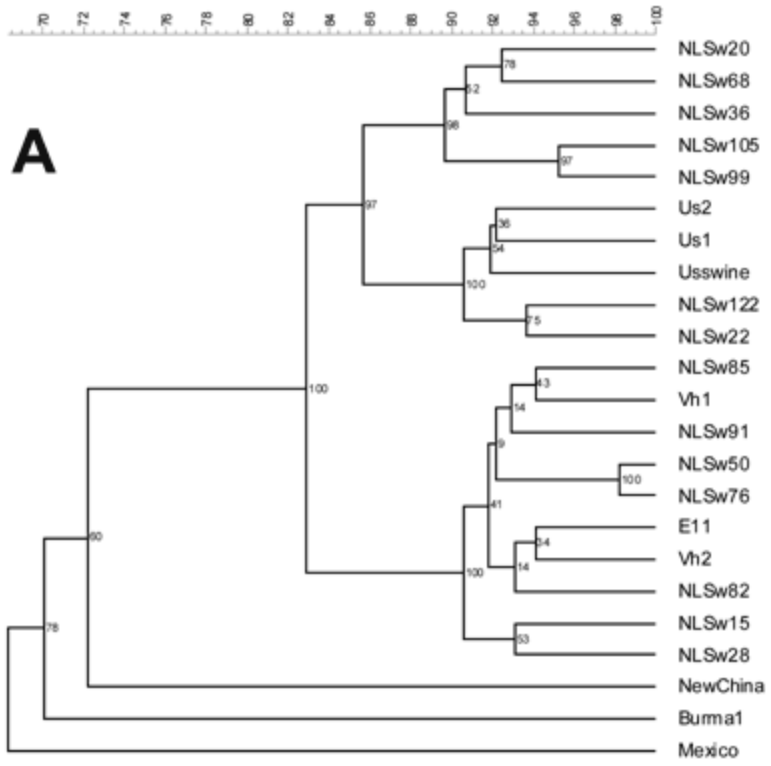
Trois différentes représentations de la même topologie d'arbre

Les arbres



Différentes représentations de la même topologie d'arbre

Les arbres : enracinement



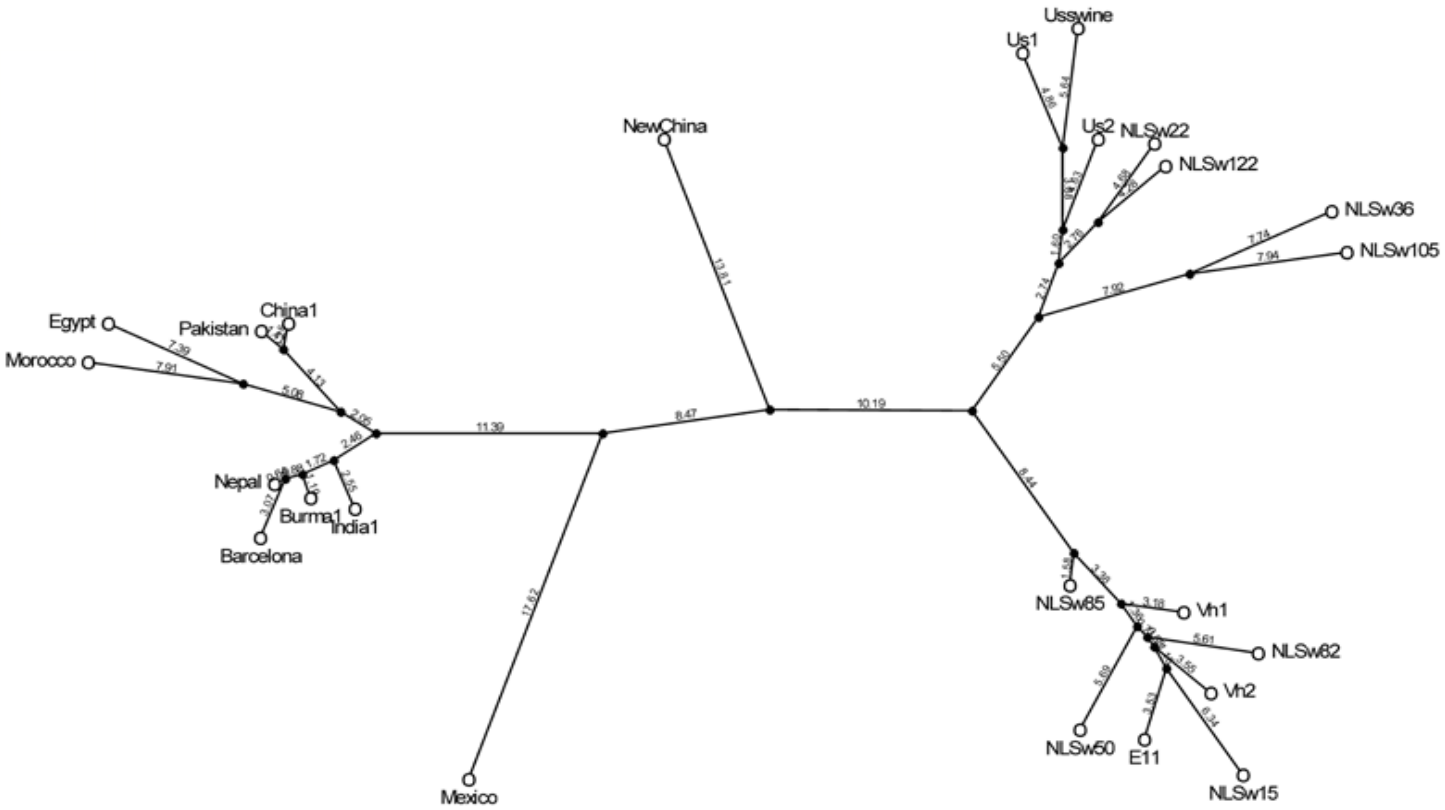
Ancien → Recent

Un arbre raciné à un nœud (la racine) qui représente un événement dans le temps plus ancien que n'importe quel autre nœud dans l'arbre.

Un arbre raciné a une direction(les nœuds peuvent être ordonnés en termes "d'anciens" ou "récents").

Dans un arbre raciné la distance entre deux nœuds est représenté le long de l'axe du temps.

Les arbres : enracinement



Dans un arbre non raciné il n'y a pas de direction.
Nous ne savons pas si un nœud est plus ancien qu'un autre.

La distance le long des branches représente directement la distance entre les nœuds.

Les arbres : enracinement

Une stratégie couramment utilisée pour «raciner un arbre» est d'inclure les espèces « outgroupe » dans l'analyse

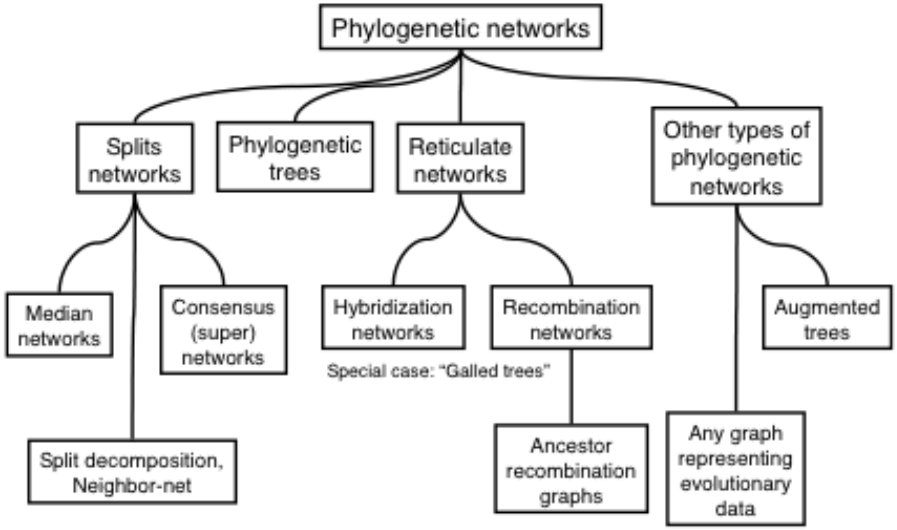
Ces espèces sont connus pour être plus lointaines que les espèces d'intérêt.

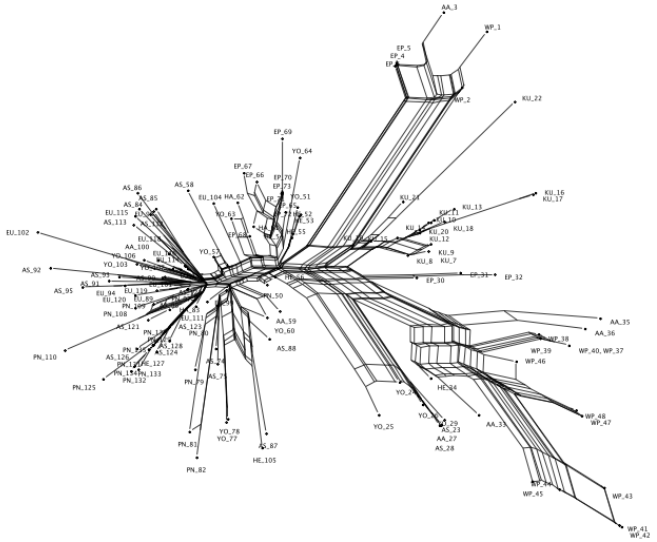
Bien que l'arbre inféré pour toutes les espèces soit sans racines**, la racine est censée être située le long de la branche** qui mène à « l' outgroupe » afin que l'arbre de l'espèce ingroup soit enraciné.

Cette stratégie est appelée enracinement « outgroupe ».

Les arbres

Il existe d'autre représentation : les réseaux phylogénétiques

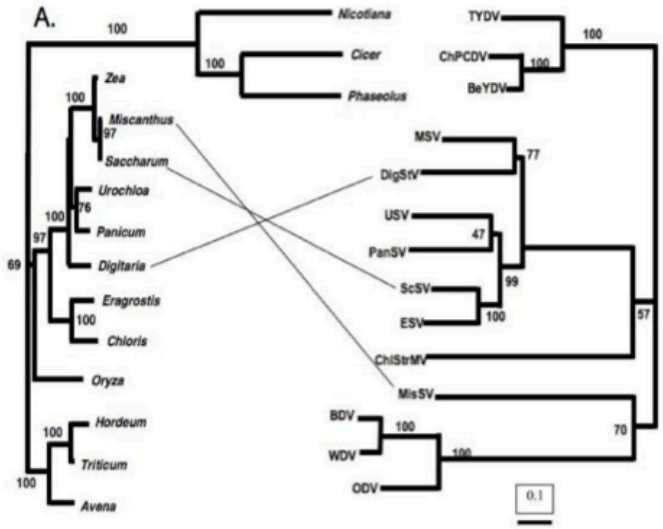


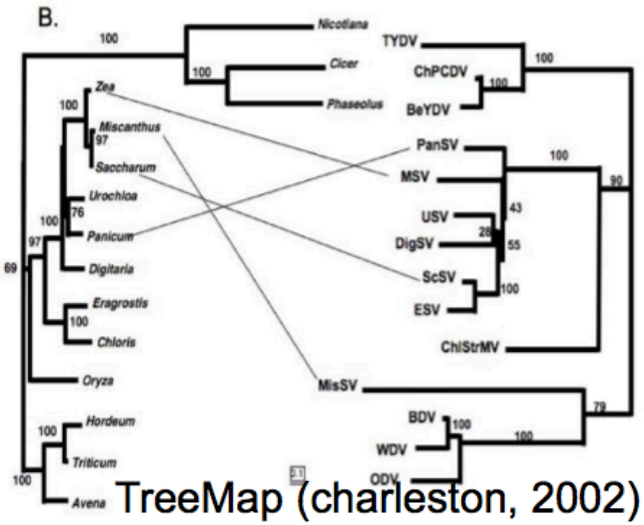


- permettent de rendre comptes des évènements évolutifs : les transferts latéraux, les recombinaisons, les duplications
- de représenter des arbres consensus

Les arbres

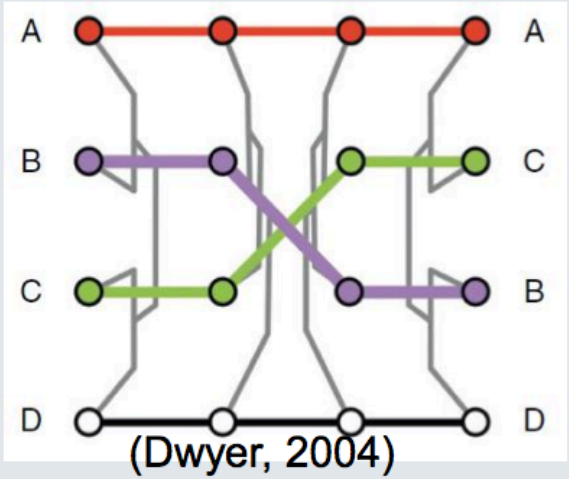
« Tanglegram » hôtes/parasites



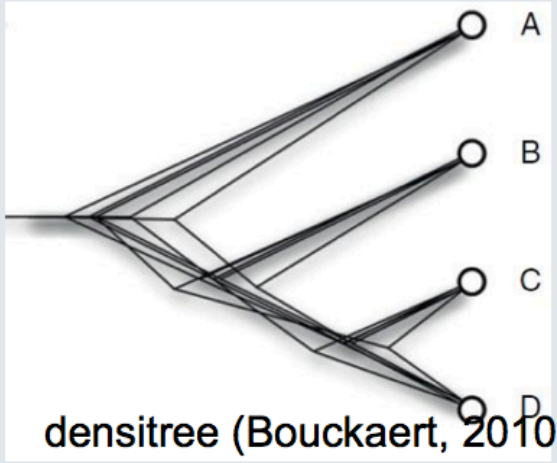


TreeMap (charleston, 2002)

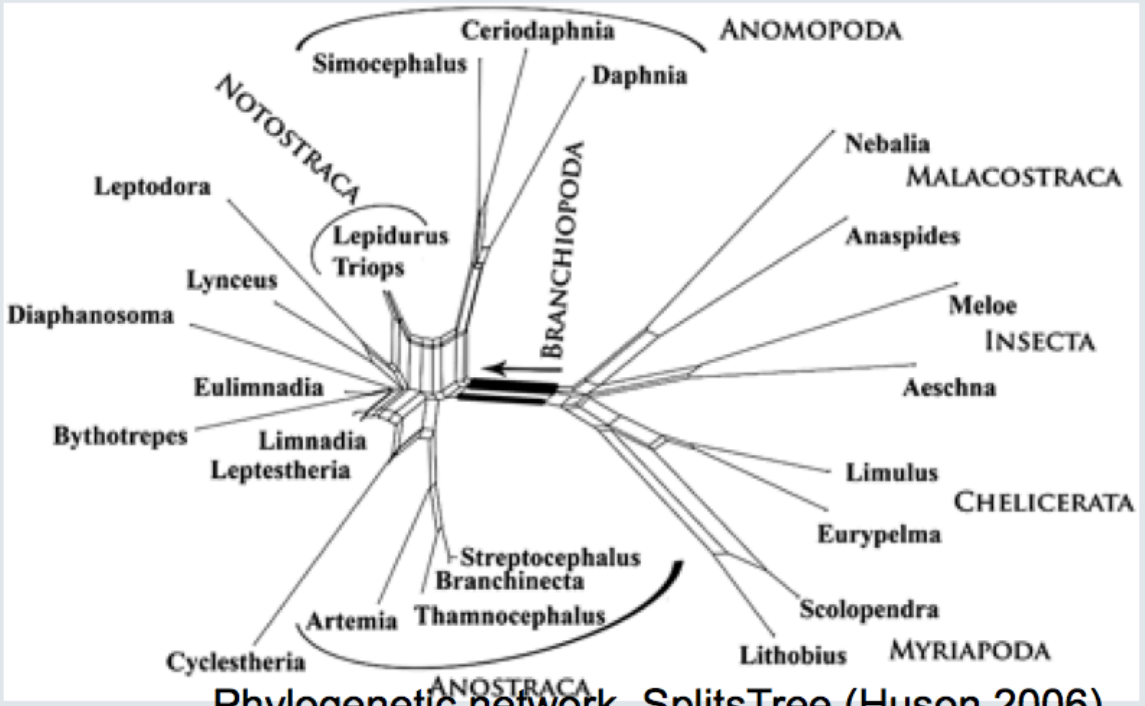
« uncertainty » (tree topology and branch lenght)



(Dwyer, 2004)



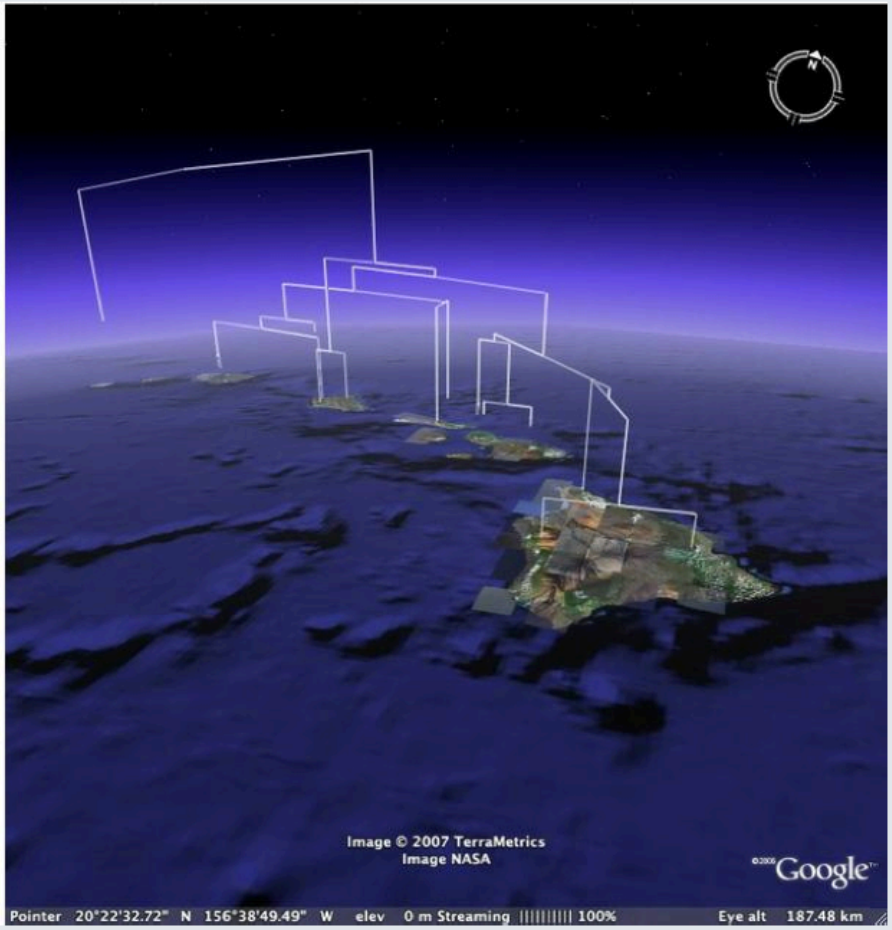
densitree (Bouckaert, 2010

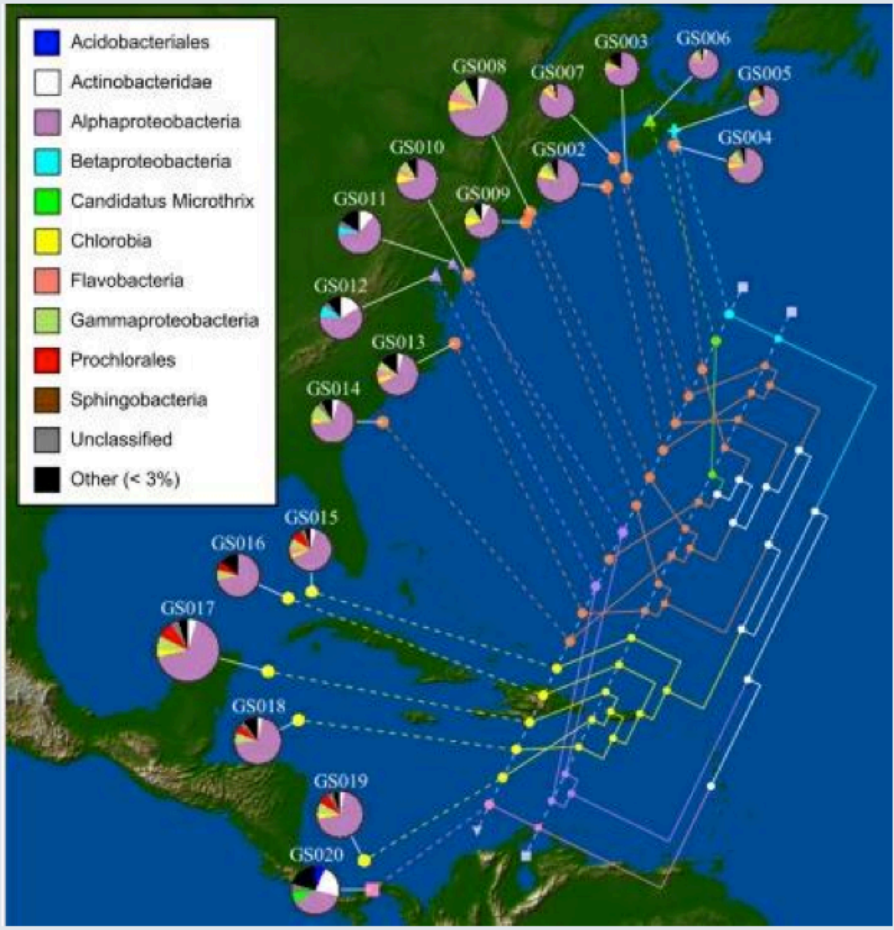


Phylogenetic network, SplitsTree (Huson 2006)

Les arbres

Géophylogénies

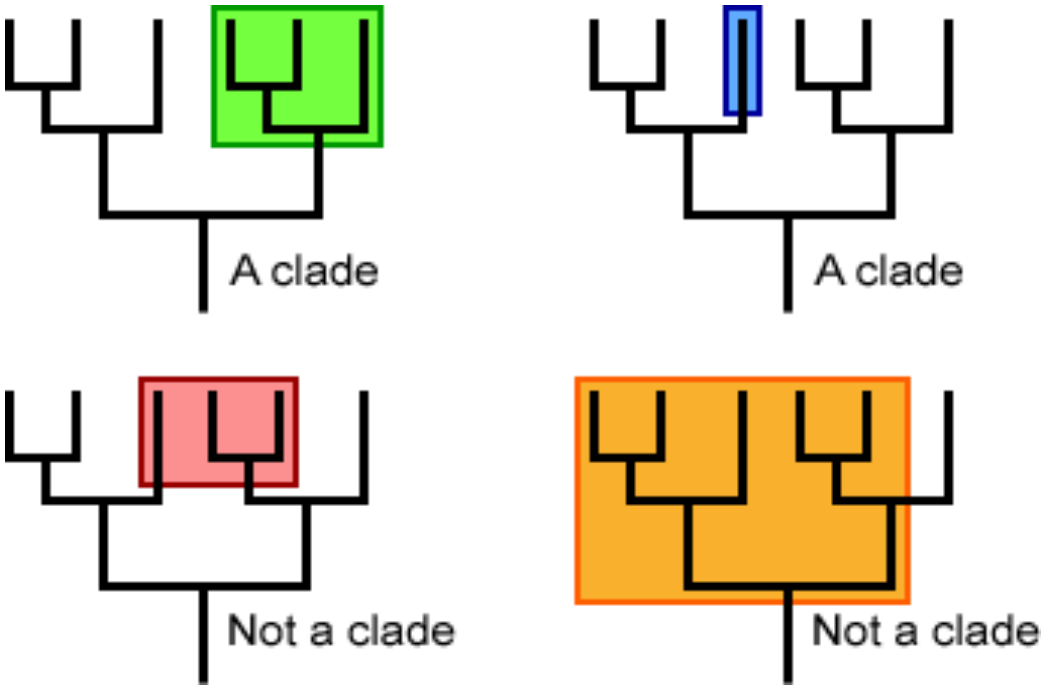




Google Earth GenGIS (Parks 2009)

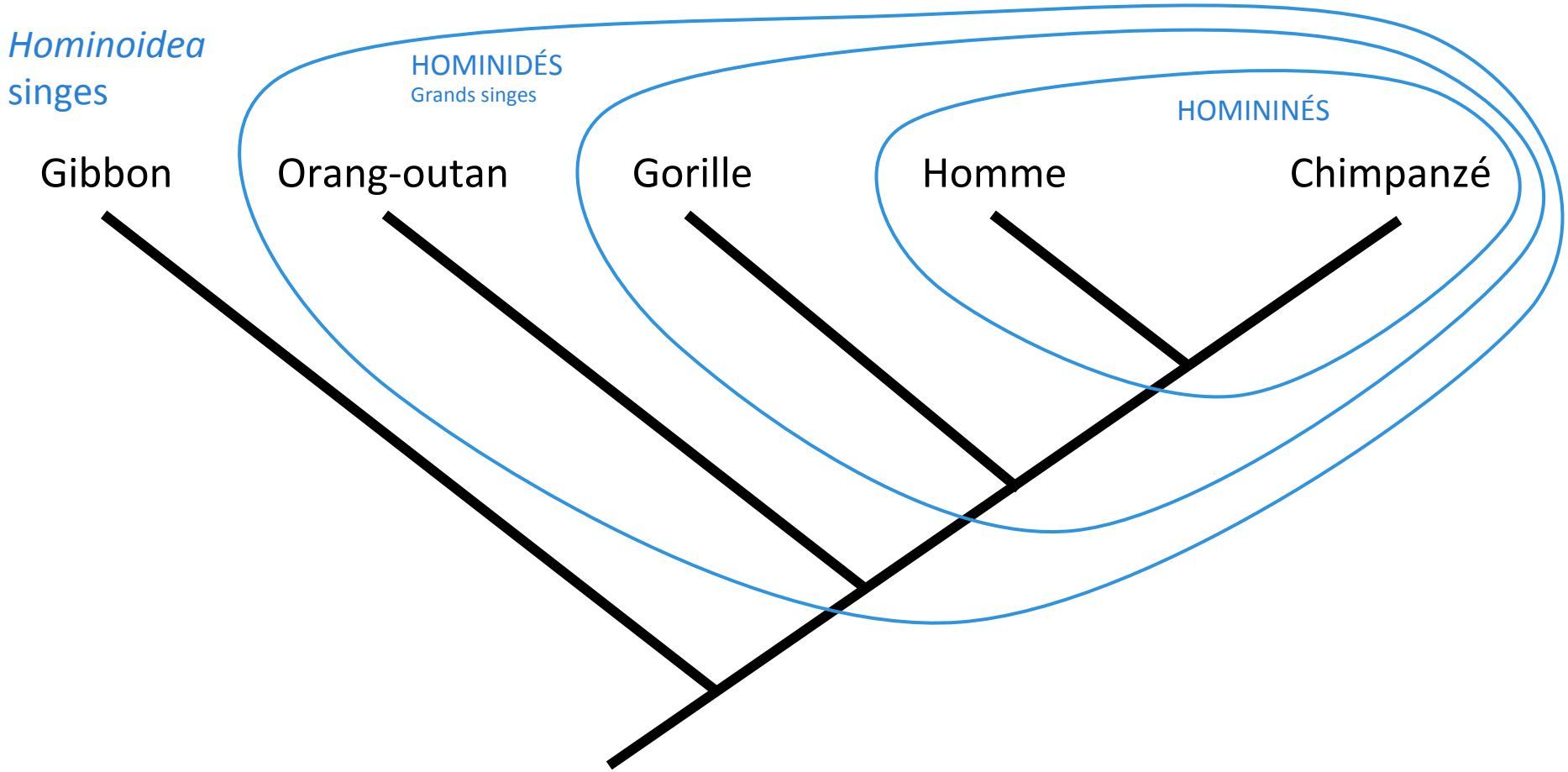
Les arbres : Les clades

- **Clade** : groupe monophylétique d'organismes vivants : c'est-à-dire qui inclut un ancêtre commun et tous ses descendants.



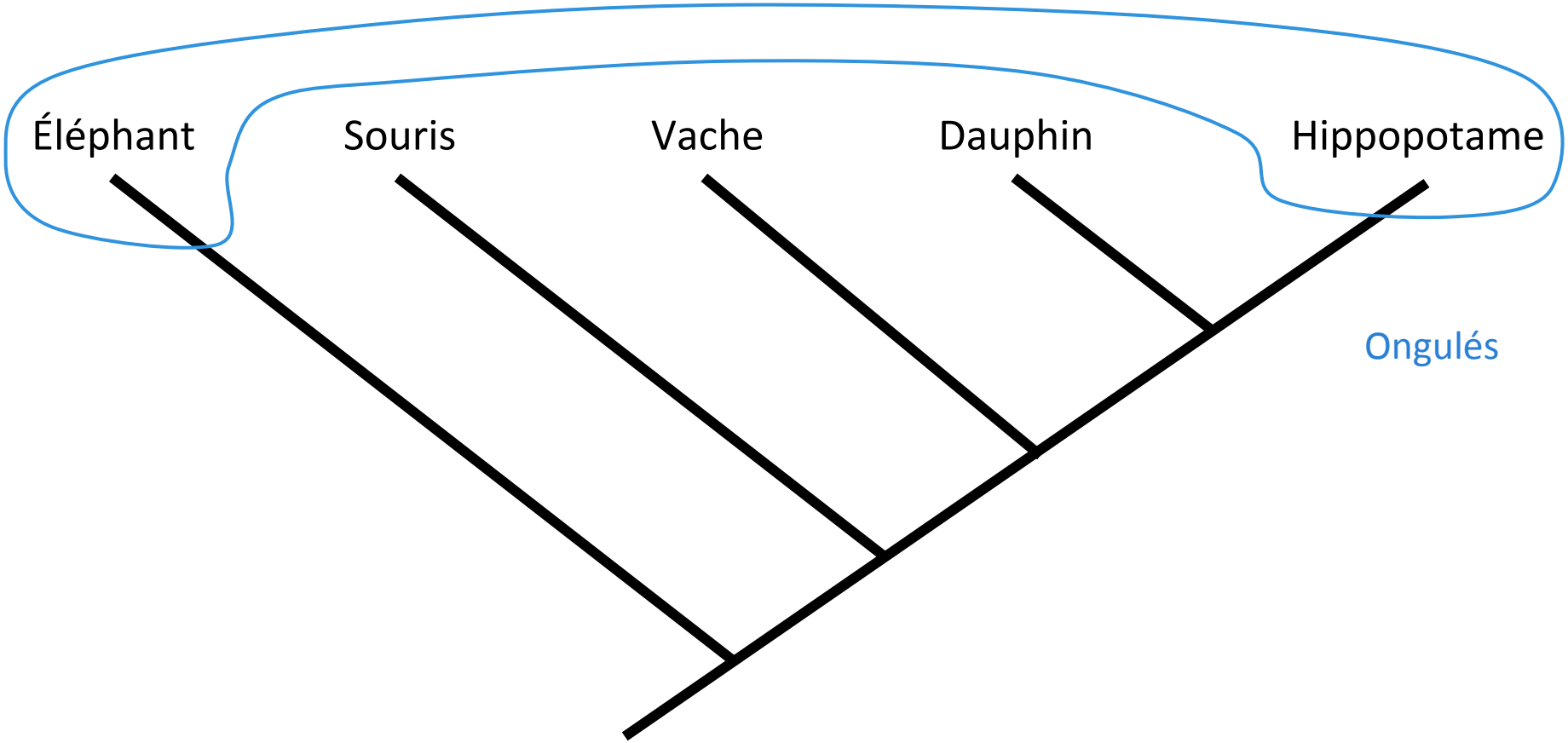
Les arbres : les groupes

Un groupe monophylétique contient l'ancêtre commun et tous ses descendants



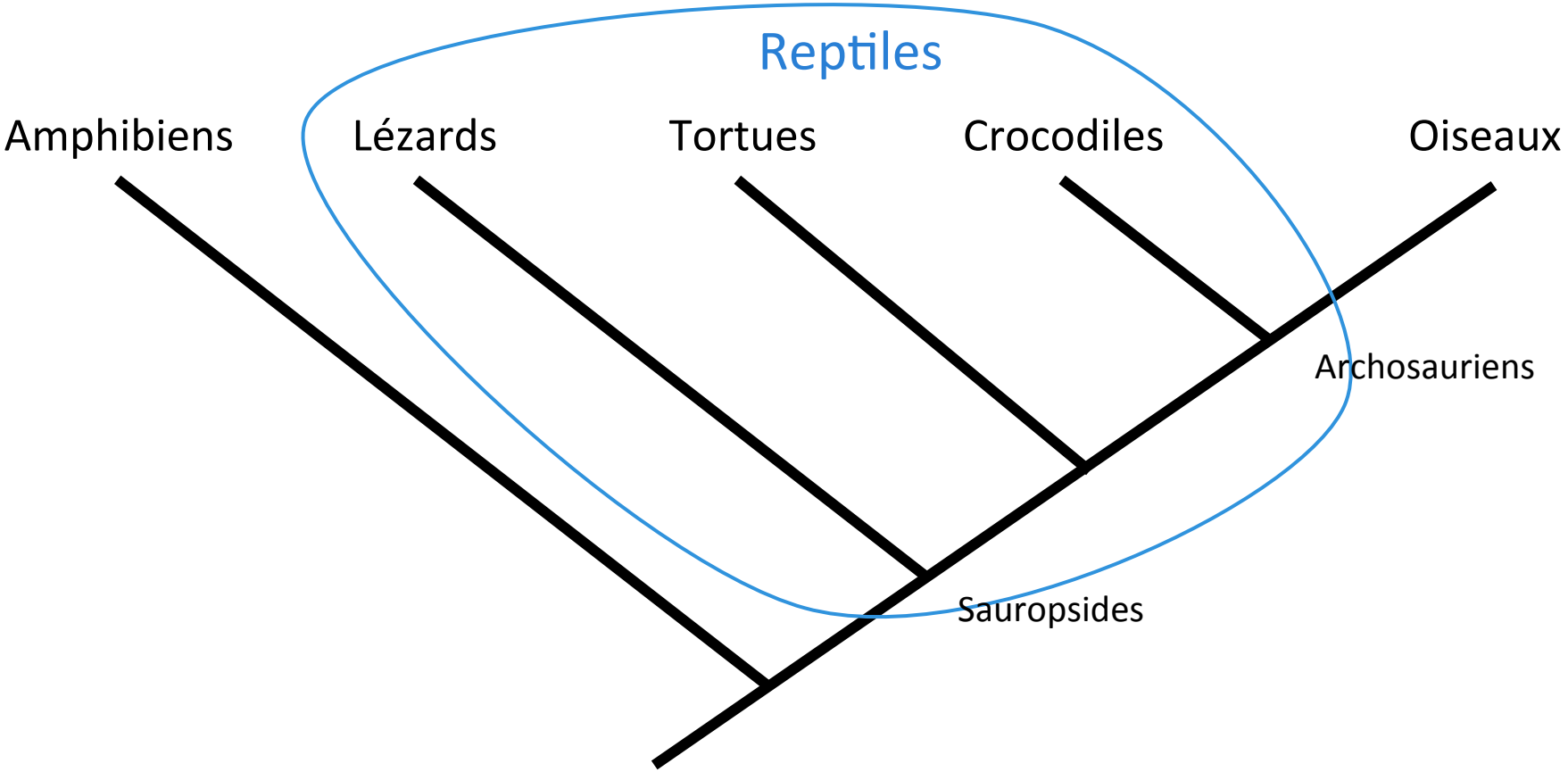
Les arbres : les groupes

Un groupe polyphylétique ne contient pas l'ancêtre commun



Les arbres : les groupes

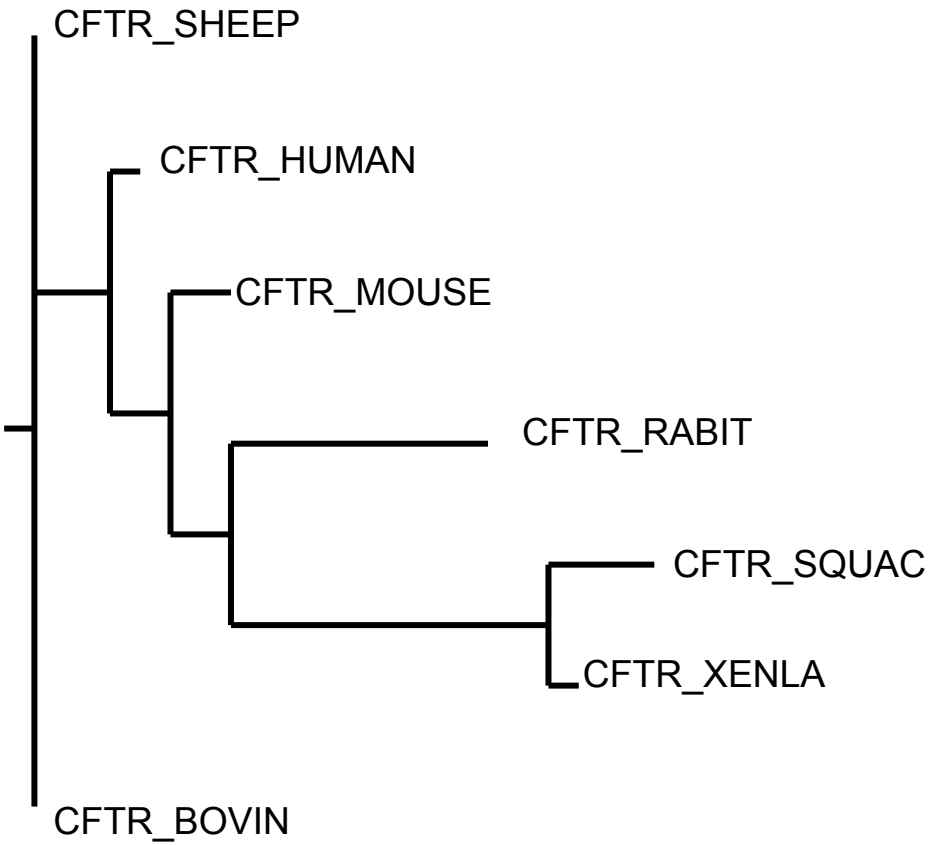
Un groupe paraphylétique contient l'ancêtre commun et seulement certains de ses descendants



Les arbres : les formats

Newick Tree Format

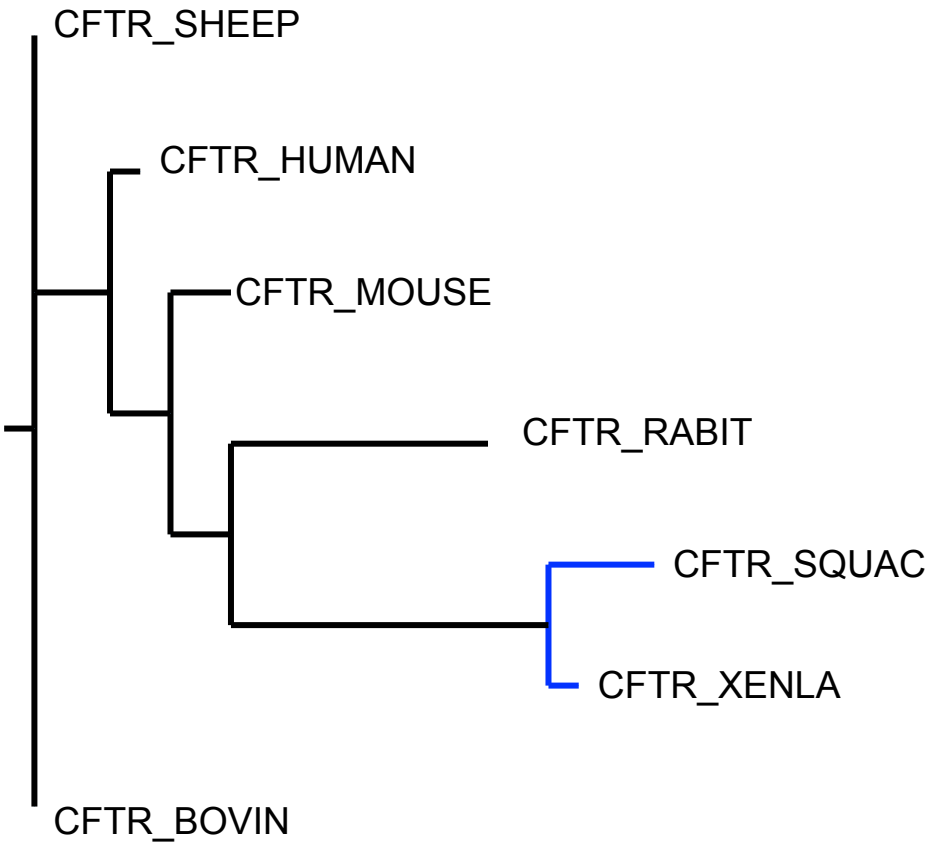
(CFTR_SHEEP:0.01457,(CFTR_HUMAN:0.16153, (CFTR_MOUSE:0.70599, (CFTR_RABIT:2.76042,(CFTR_SQUAC:1.27192,CFTR_XENLA:0.28818) :3.42183):0.77076):0.65873)0.73937, CFTR_BOVIN:0.00953);



Les arbres : les formats

Newick Tree Format

(CFTR_SHEEP:0.01457,(CFTR_HUMAN:0.16153, (CFTR_MOUSE:0.70599, (CFTR_RABIT:2.76042,(CFTR_SQUAC:1.27192,CFTR_XENLA:0.28818) :3.42183):0.77076):0.65873)0.73937, CFTR_BOVIN:0.00953);

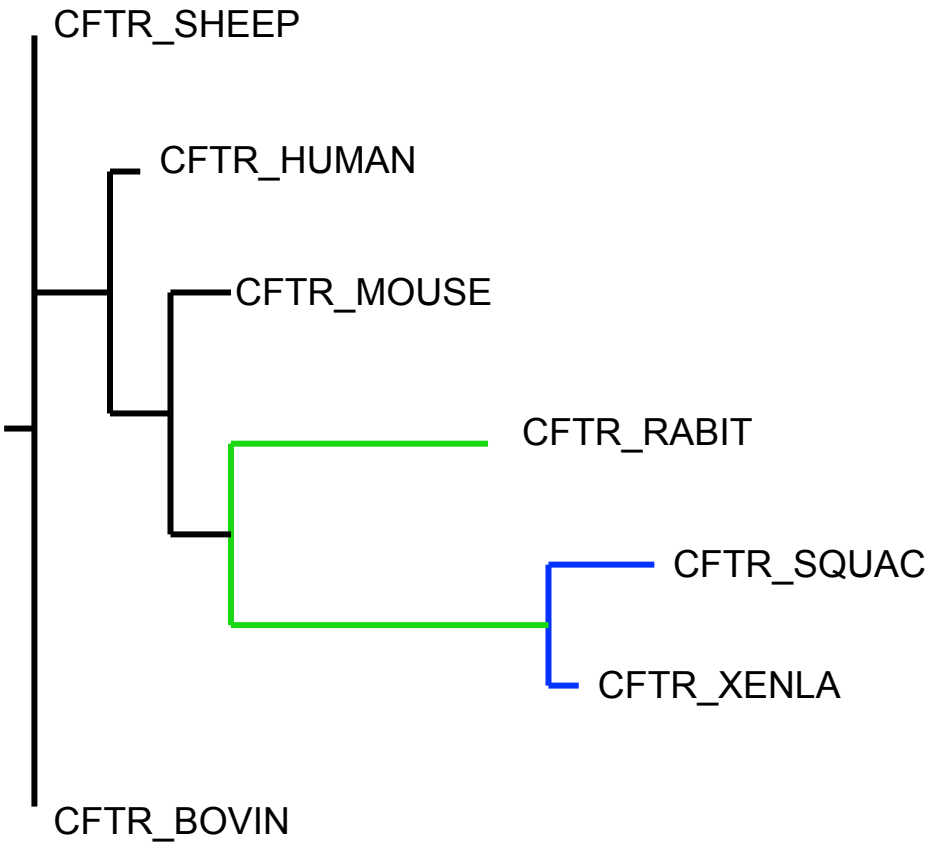


(CFTR_SQUAC:1.27192,
CFTR_XENLA:0.28818)

Les arbres : les formats

Newick Tree Format

(CFTR_SHEEP:0.01457,(CFTR_HUMAN:0.16153, (CFTR_MOUSE:0.70599, (CFTR_RABIT:2.76042,(CFTR_SQUAC:1.27192,CFTR_XENLA:0.28818) :3.42183):0.77076):0.65873)0.73937, CFTR_BOVIN:0.00953);

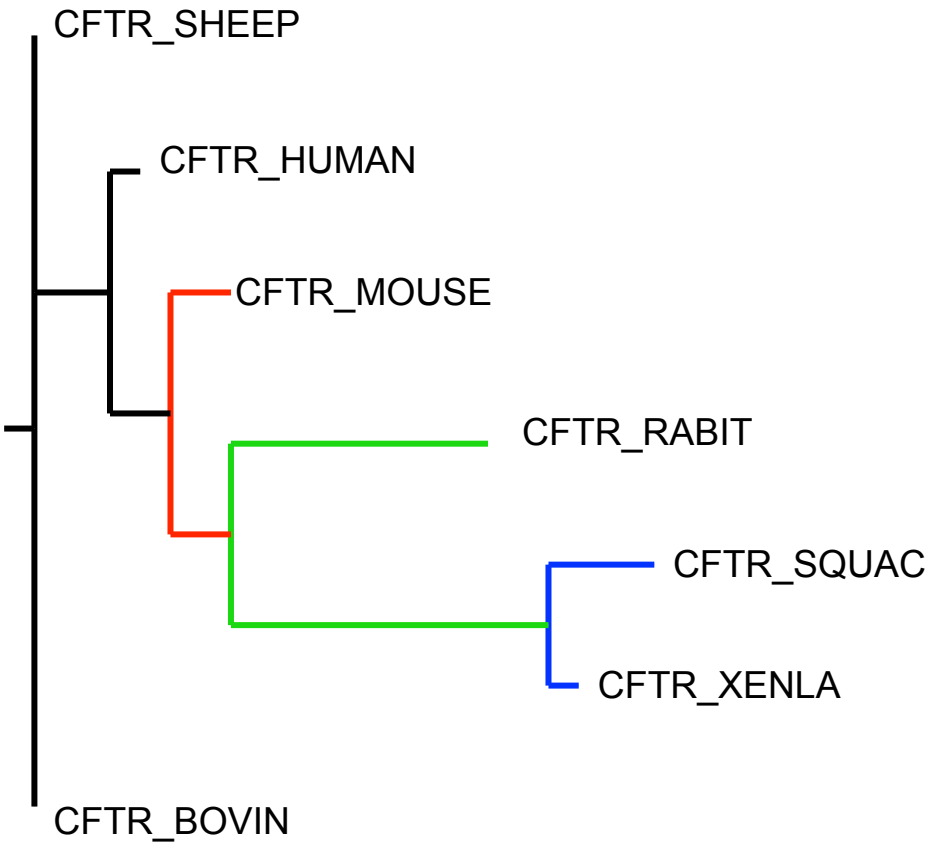


(CFTR_RABIT:2.76042,
(CFTR_SQUAC:1.27192,
CFTR_XENLA:0.28818)
:3.42183)

Les arbres : les formats

Newick Tree Format

(CFTR_SHEEP:0.01457,(CFTR_HUMAN:0.16153, (CFTR_MOUSE:0.70599, (CFTR_RABIT:2.76042,(CFTR_SQUAC:1.27192,CFTR_XENLA:0.28818) :3.42183):0.77076):0.65873)0.73937, CFTR_BOVIN:0.00953);

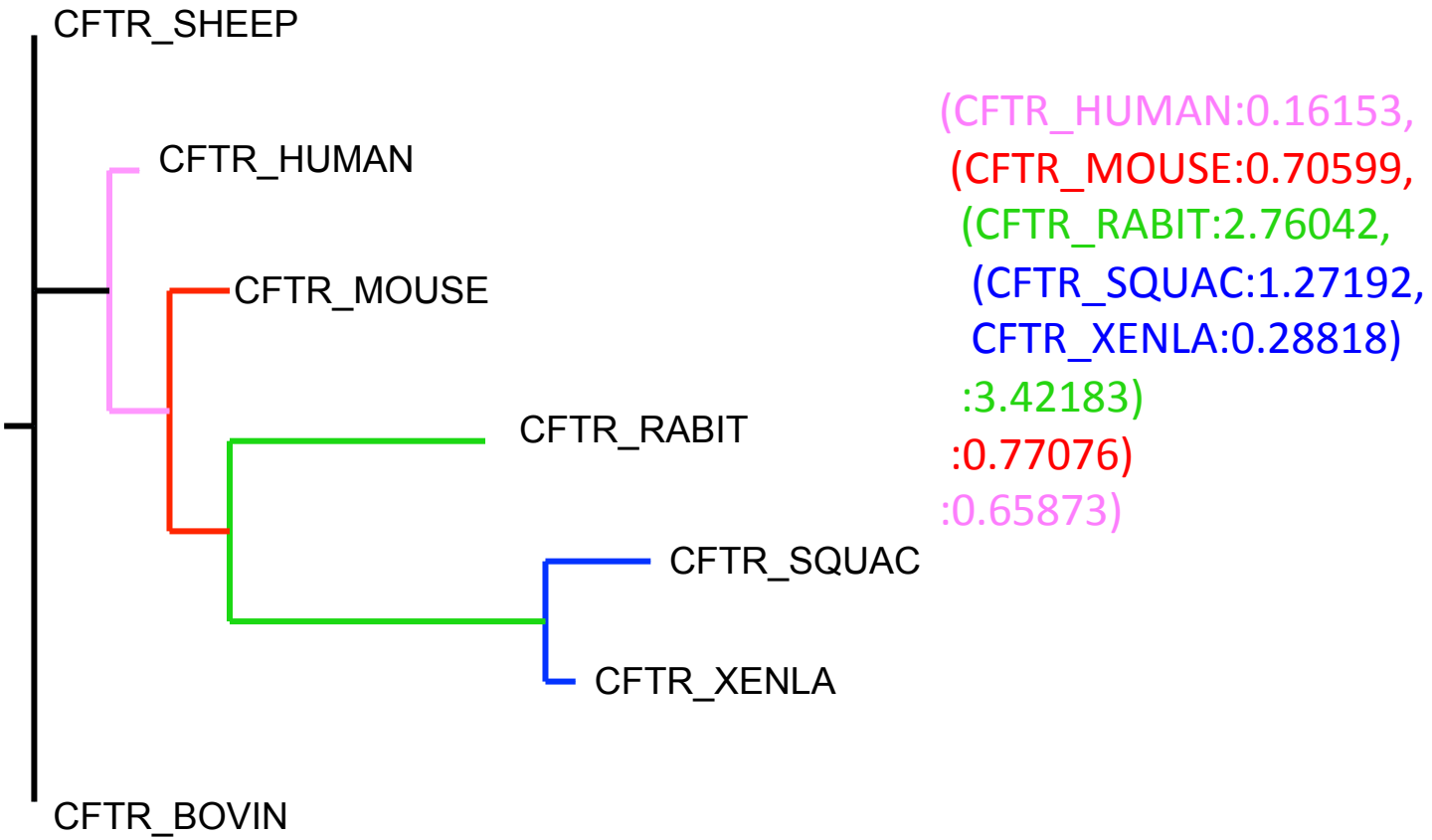


(CFTR_MOUSE:0.70599,
(CFTR_RABIT:2.76042,
(CFTR_SQUAC:1.27192,
CFTR_XENLA:0.28818)
:3.42183)
:0.77076)

Les arbres : les formats

Newick Tree Format

(CFTR_SHEEP:0.01457,(CFTR_HUMAN:0.16153, (CFTR_MOUSE:0.70599, (CFTR_RABIT:2.76042,(CFTR_SQUAC:1.27192,CFTR_XENLA:0.28818) :3.42183):0.77076):0.65873)0.73937, CFTR_BOVIN:0.00953);

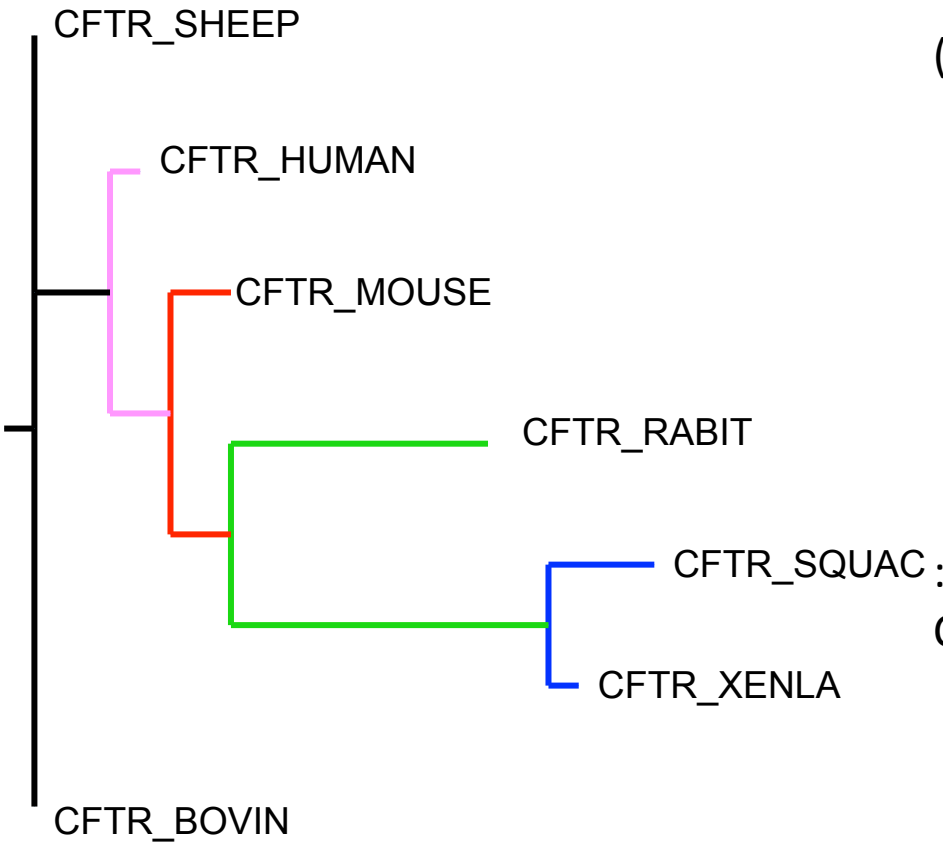


(CFTR_HUMAN:0.16153,
(CFTR_MOUSE:0.70599,
(CFTR_RABIT:2.76042,
(CFTR_SQUAC:1.27192,
CFTR_XENLA:0.28818)
:3.42183)
:0.77076)
:0.65873)

Les arbres : les formats

Newick Tree Format

(CFTR_SHEEP:0.01457,(CFTR_HUMAN:0.16153, (CFTR_MOUSE:0.70599, (CFTR_RABIT:2.76042,(CFTR_SQUAC:1.27192,CFTR_XENLA:0.28818) :3.42183):0.77076):0.65873)0.73937, CFTR_BOVIN:0.00953);

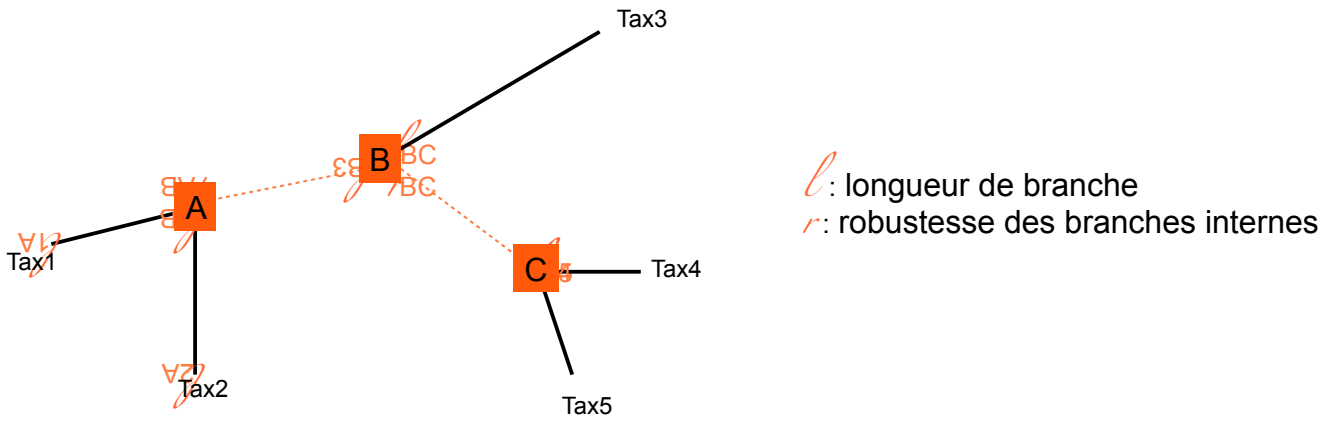


(CFTR_SHEEP:0.01457,
(CFTR_HUMAN:0.16153,
(CFTR_MOUSE:0.70599,
(CFTR_RABIT:2.76042,
(CFTR_SQUAC:1.27192,
CFTR_XENLA:0.28818)
:3.42183)
:0.77076)
:0.65873)
:0.73937,
CFTR_BOVIN:0.00953);

Les arbres : les formats

- Format standard d'écriture des arbres :
  - Les feuilles issues d'un même nœud sont groupées entre parenthèses
  - Les feuilles et groupes de feuilles sont séparés par des virgules
  - On termine toujours l'arbre par un point virgule

Ex : ((Tax1:ℓ_{1A},Tax2:ℓ_{2A}) r_{AB}:ℓ_{AB},Tax3:ℓ_{B3},(Tax4:ℓ_{C4},Tax5:ℓ_{C5})r_{BC} :ℓ_{BC}) ;

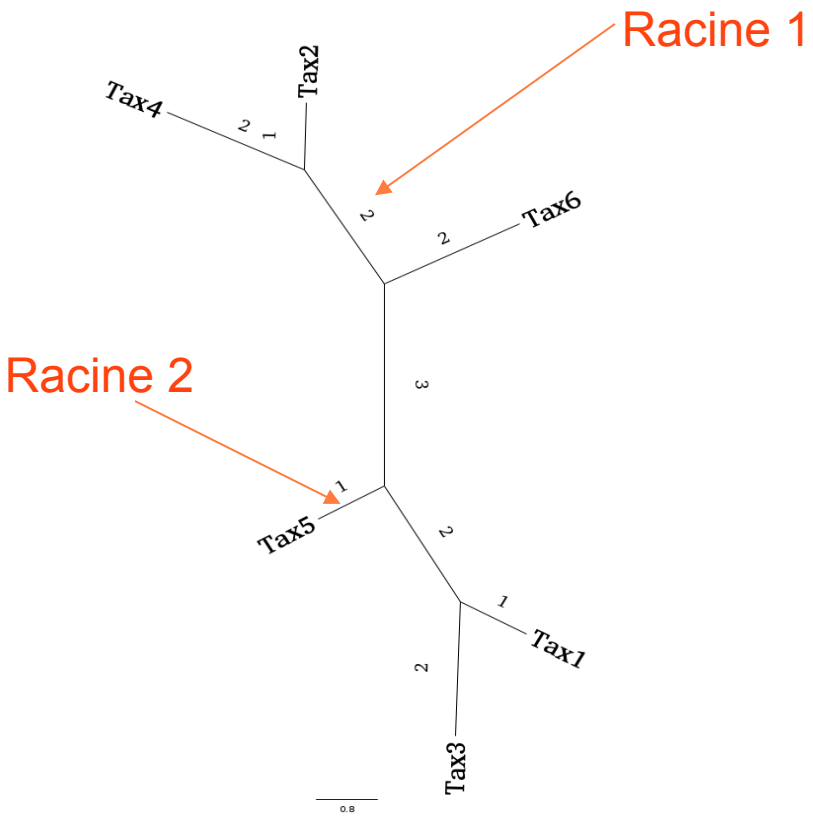


- Dessiner l'arbre suivant :
  - ((Tax1:1,Tax3:2):2,Tax5:1,((Tax4:2,Tax2:1):2,Tax6:2):3);

Les arbres : les formats

- Dessiner l'arbre suivant :

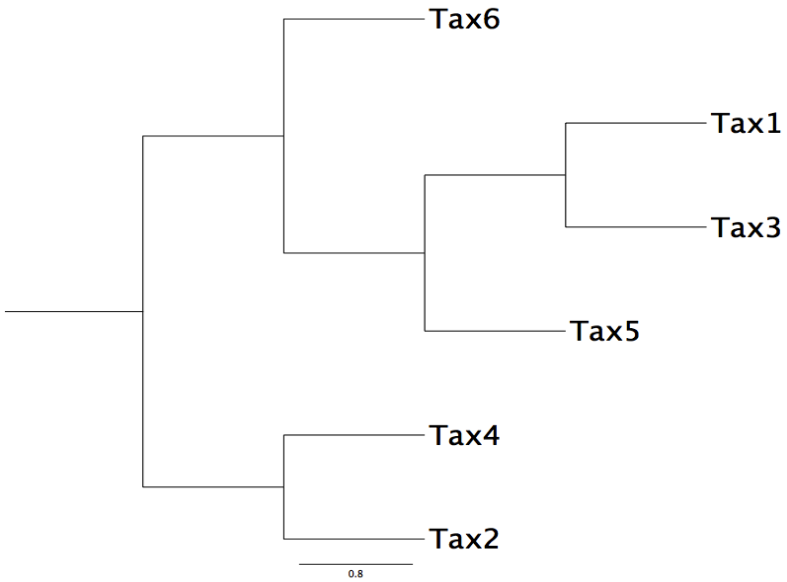
  ((Tax1:1,Tax3:2):2,Tax5:1,((Tax4:2,Tax2:1):2,Tax6:2):3);



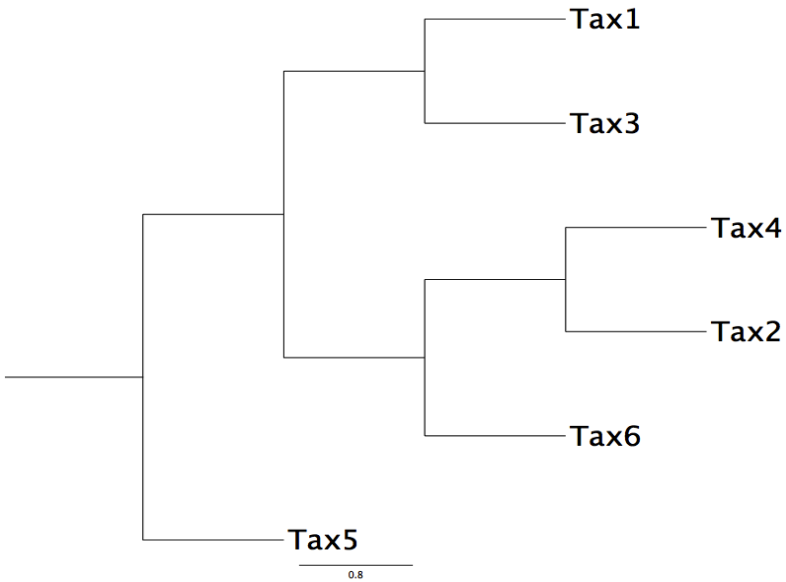
Dessiner les arbres racinés correspondants pour la Racine 1 puis pour la Racine 2

Les arbres : les formats

- Racine1



- Racine2



Les arbres : les formats

Le format NHX (New Hampshire eXtended)

http://www.phylosoft.org/NHX

NHX introduit des balises pour associer divers champs de données à un noeud d'un arbre phylogénétique

- noeuds internes et externes peuvent être étiquetés
- l'ordre des balises n'est pas important
- la longueur de toutes les données basées sur des chaînes de caractères est illimitée (par exemple nom, espèces)

(((ADH2:0.1[&&NHX:S=human:E=1.1.1.1], ADH1:0.11[&&NHX:S=human:E=1.1.1.1]):
0.05[&&NHX:S=Primates:E=1.1.1.1:D=Y: B=100], ADHY:
0.1[&&NHX:S=nematode:E=1.1.1.1],ADHX:0.12[&&NHX:S=insect:E=1.1.1.1]):
0.1[&&NHX:S=Metazoa:E=1.1.1.1:D=N],
(ADH4:0.09[&&NHX:S=yeast:E=1.1.1.1],ADH3:0.13[&&NHX:S=yeast:E=1.1.1.1],
ADH2:0.12[&&NHX:S=yeast:E=1.1.1.1],ADH1:0.11[&&NHX:S=yeast:E=1.1.1.1]):0.1
[&&NHX:S=Fungi])[&&NHX:E=1.1.1.1:D=N];

Les arbres : les formats

Le format Nexus

Suffix .nxs .nex

Paup, MrBayes, Mesquite, MacClade

```
#NEXUS
BEGIN TAXA;
      Dimensions NTax=4;
      TaxLabels fish frog snake
mouse;
END;

BEGIN CHARACTERS;
      Dimensions NChar=20;
      Format DataType=DNA;
      Matrix
      fish   ACATA GAGGG TACCT CTAAG
      frog   ACATA GAGGG TACCT CTAAG
      snake  ACATA GAGGG TACCT CTAAG
      mouse  ACATA GAGGG TACCT CTAAG
END;

BEGIN TREES;
      Tree best=(fish, (frog,
(snake, mouse)));
END;;
```

Les arbres : les formats

Le format PhyloXML

Han, Mira V.; Zmasek, Christian M. (2009).
phyloXML: XML for evolutionary biology and comparative genomics" BMC Bioinformatics (United Kingdom: BioMed Central) 10: 356

PhyloXML est un langage XML pour l'analyse, l'échange et le stockage des arbres phylogénétiques (ou des réseaux) et les données associées. La structure de phyloXML est décrit par XML Schema Definition (XSD) langue.

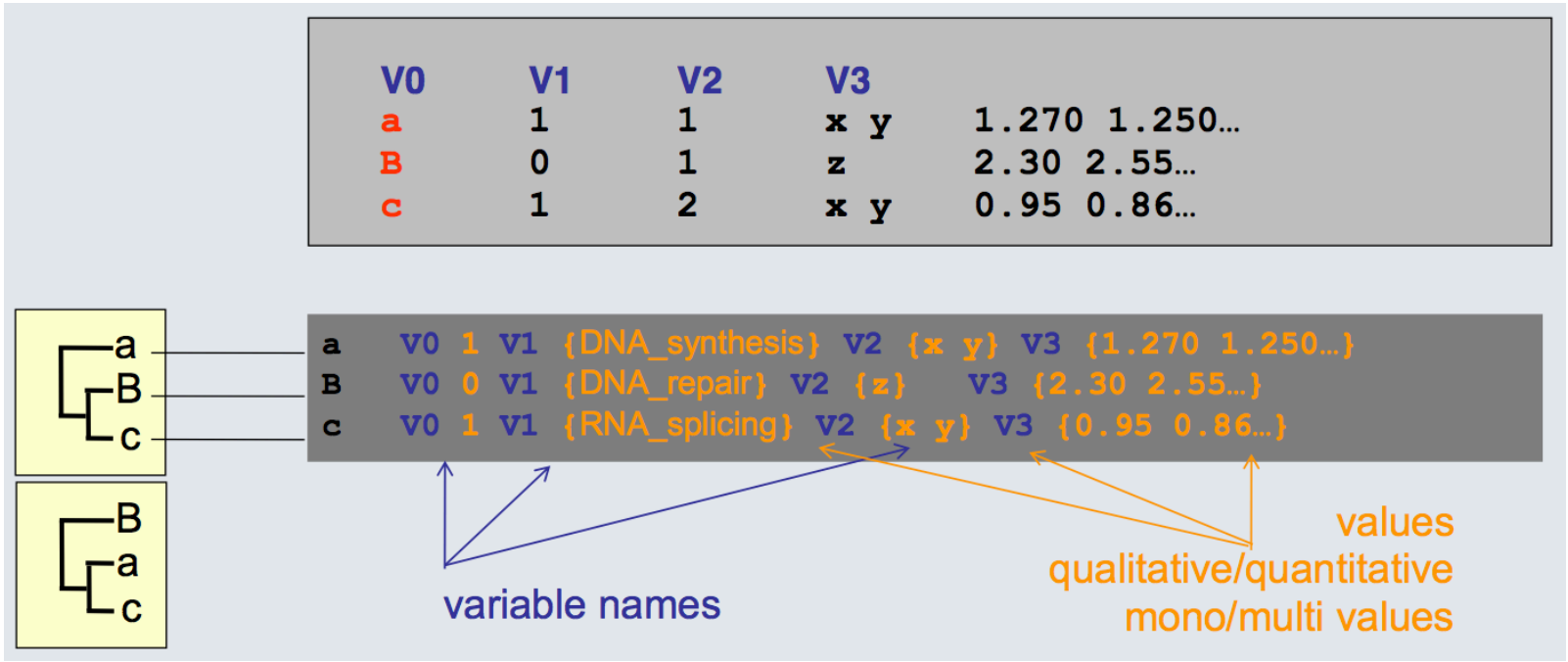
```
<phyloxml xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.phyloxml.org http://
www.phyloxml.org/1.10/phyloxml.xsd" xmlns="http://www.phyloxml.org">
<phylogeny rooted="true">
<name>example from Prof. Joe Felsenstein's book "Inferring Phylogenies"</name> <description>MrBayes based on MAFFT alignment</
description>
<clade>
    <clade branch_length="0.06">
      <confidence type="probability">0.88</confidence>
      <clade branch_length="0.102">
        <name>A</name>
      </clade>
<clade branch_length="0.23"> <name>B</name>
</clade>
</clade>
    <clade branch_length="0.4">
      <name>C</name>
    </clade>
  </clade>
 </phylogeny>
</phyloxml>
```

Les arbres : les formats

Le format TreeDyn

Fichiers d'annotation externes avec le nom des feuilles (unique) pour faire la liaison avec des couples de variables/valeurs.

Cette organisation présente l'avantage d'être indépendante mais totalement compatible avec le standard "Newick".



LES MÉTHODES DE CONSTRUCTION D'ARBRES

Méthodes de construction d'arbres

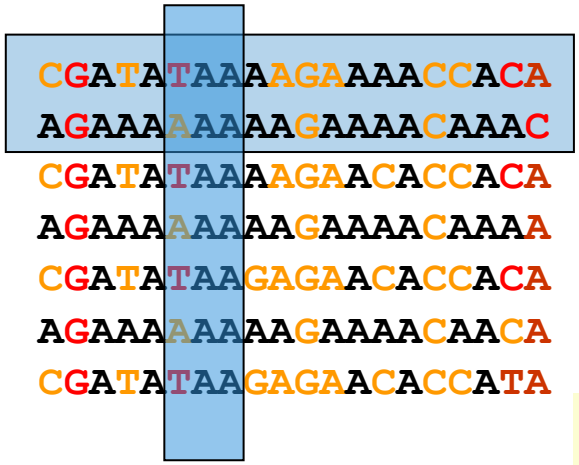
Il existe 3 approches :

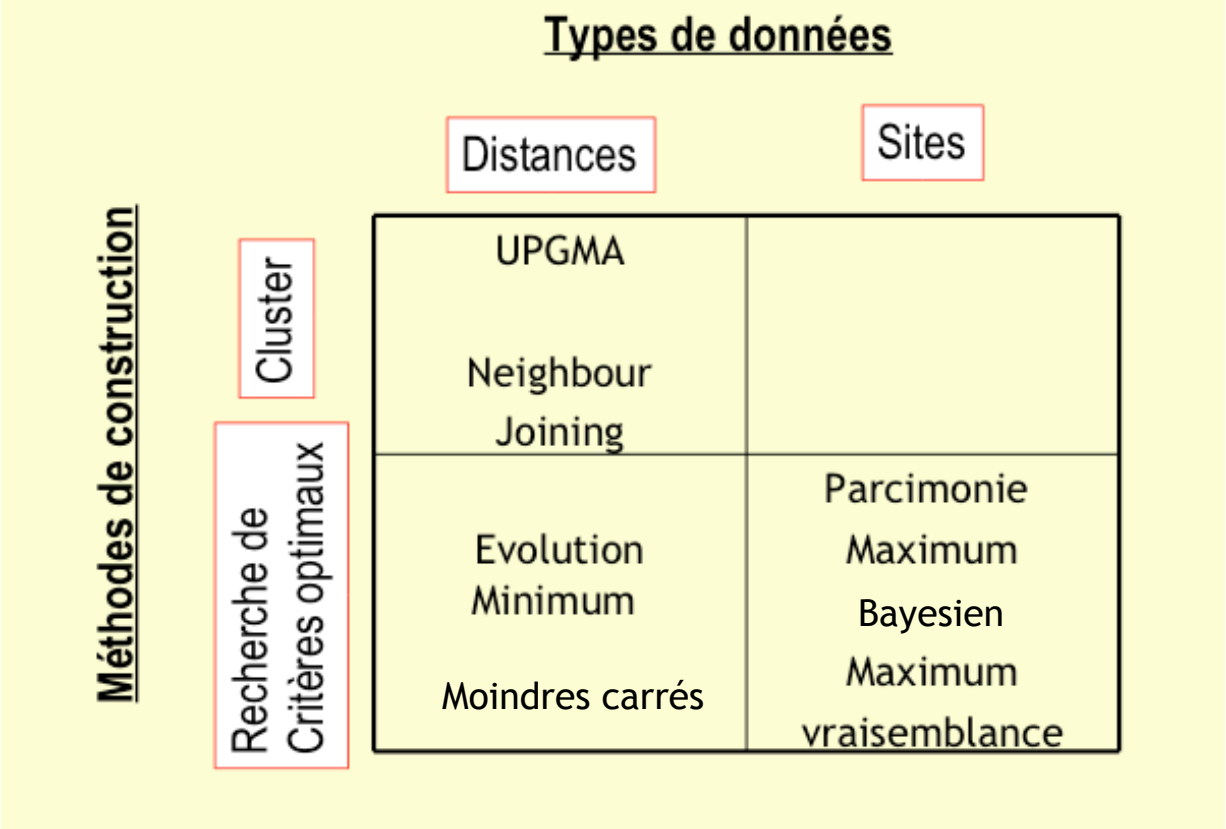
L'approche cladistique cherche en particulier à déterminer les caractères propres à une branche, qui « signent » un apparentement.

L'approche phénétique, une classification basée uniquement sur des mesures de distance entre taxons (évaluées par exemple en comptant les différences de séquences d'ADN) sans chercher à faire une interprétation phylogénétique.

L'approche probabiliste qui construit des arbres phylogénétiques en utilisant des modèles d'évolution des caractères (le plus souvent moléculaires, mais pas obligatoirement).

Méthodes de construction d'arbres





Méthodes de construction d'arbres

Phénétique vs Cladistique

L'approche phénétique (taxonomie numérique) se veut complètement objective. C'est une approche très quantitative dans laquelle tous les traits (qu'ils soient homologues ou non) sont traités également.

Cette méthode se révèle peu pertinente lorsqu'on l'applique aux caractères morphologiques en raison des analogies : convergence évolutives.

Elle s'applique préférentiellement sur des caractères moléculaires où le nombre de caractères pris en compte est important

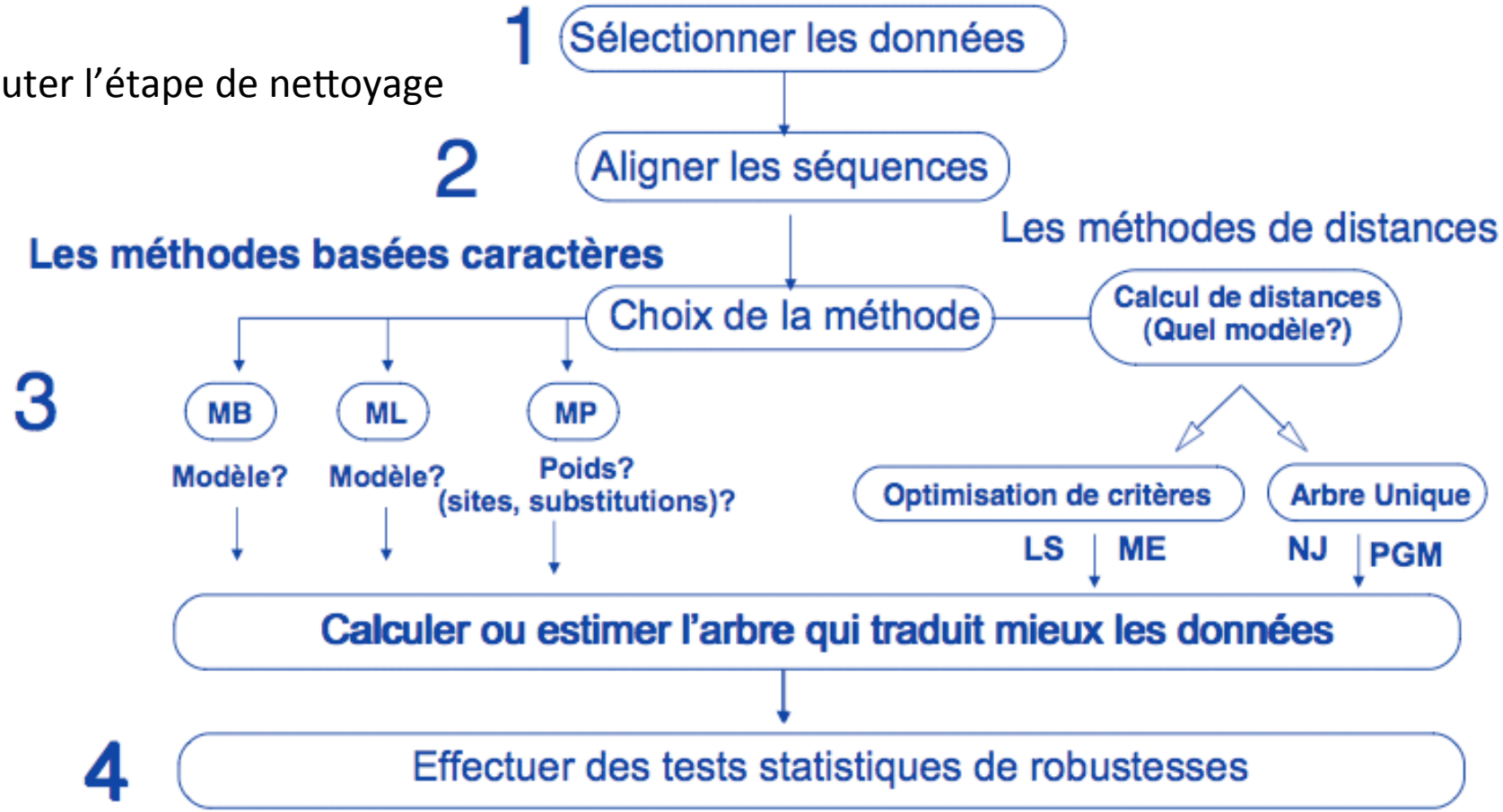
La cladistique hiérarchise les caractères comparés. Ne sont en fait regroupés dans un même taxon que les êtres vivants qui partagent des caractères homologues : partage d'une ascendance commune.

Les homologies sont en fait vues comme des innovations évolutives partagées : synapomorphies

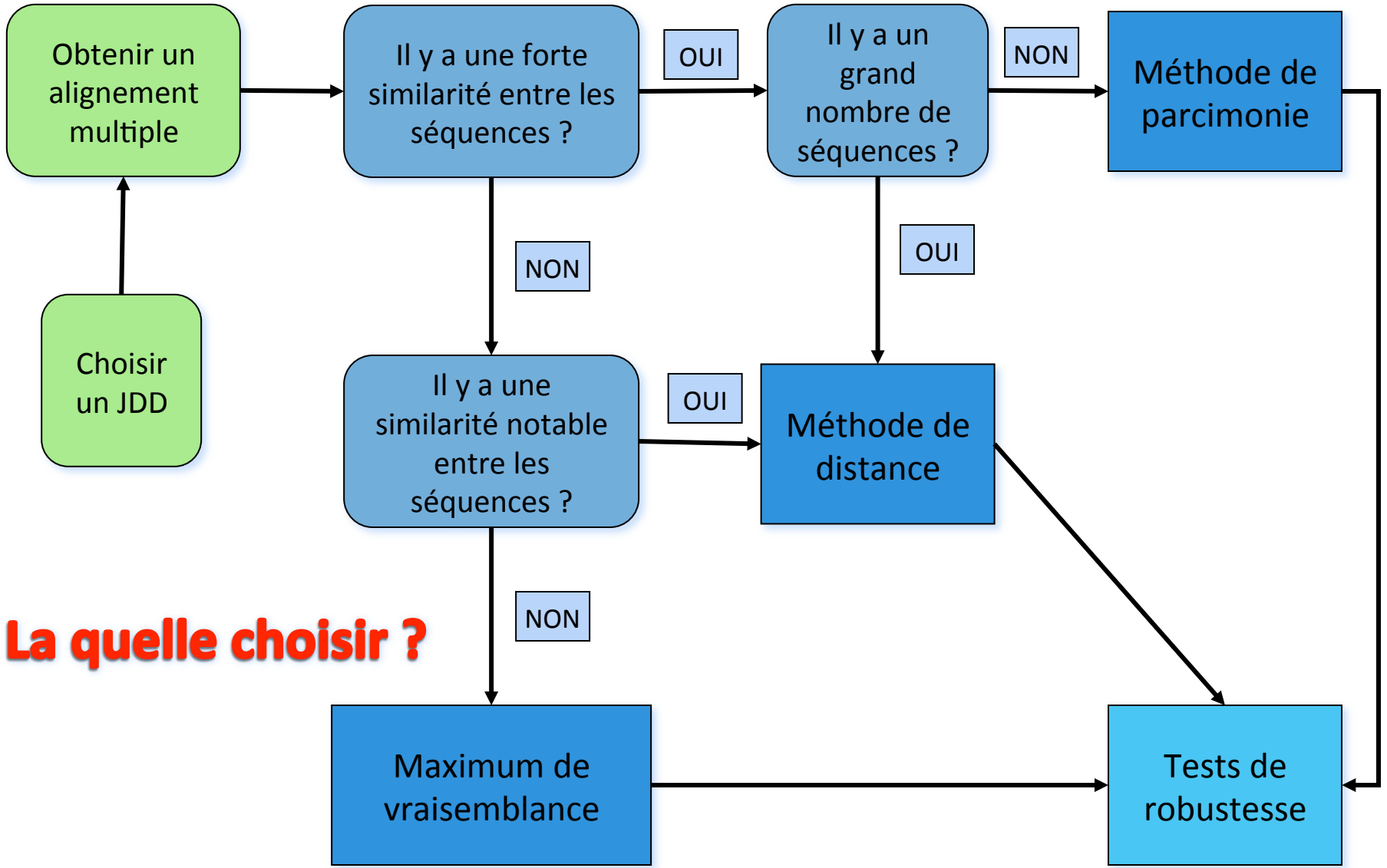
Méthodes de construction d'arbres

Méthodologie

Ajouter l'étape de nettoyage



Méthodes de construction d'arbres



La quelle choisir ?

bayesien