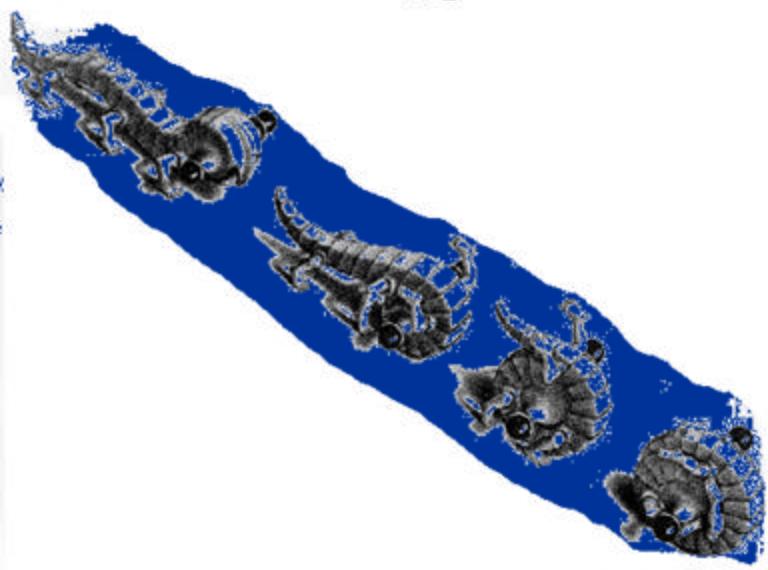
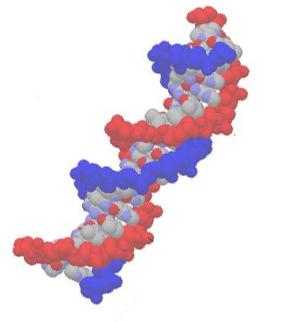


Masters DBM,
Université Joseph Ki-Zerbo



Principes et outils de la Bioinformatique

Octobre 2020

Dr Fidèle TIENDREBEOGO

Dr Ezechiel B. TIBIRI

Chercheur, CNRST, INERA

MOLECULAR SEQUENCES

Alignment Methods



BIOINFORMATICS

ALIGNMENT

*Sequence Evolution Models
Phylogenetic Methods*



PHYLOGENETICS

EVOLUTIONARY TREE (time scale = genetic distance)

Molecular Clock Models



EVOLUTIONARY TREE (time scale = years)

Coalescent Models



POPULATION GENETICS

POPULATION GENETIC PROCESSES (e.g. selection, migration, population dynamics)

Définitions et Rappels

Obtention des données

Alignement

Phylogénie :

Méthodes de distance

Parcimonie

Maximum likelihood

Analyse bayésienne

Horloge moléculaire

Recombinaison

Définitions & Rappels

Biologie et Bioinformatique

- L'étude des êtres vivants (relations et interactions) >> nombreuses données (un grand nombre non encore interprétées!)
- Pour mieux organiser cette manne de données >> nouvelle discipline: la bioinformatique
- Disciplines associées : biologie, informatique, mathématiques (probabilités, statistiques etc.)

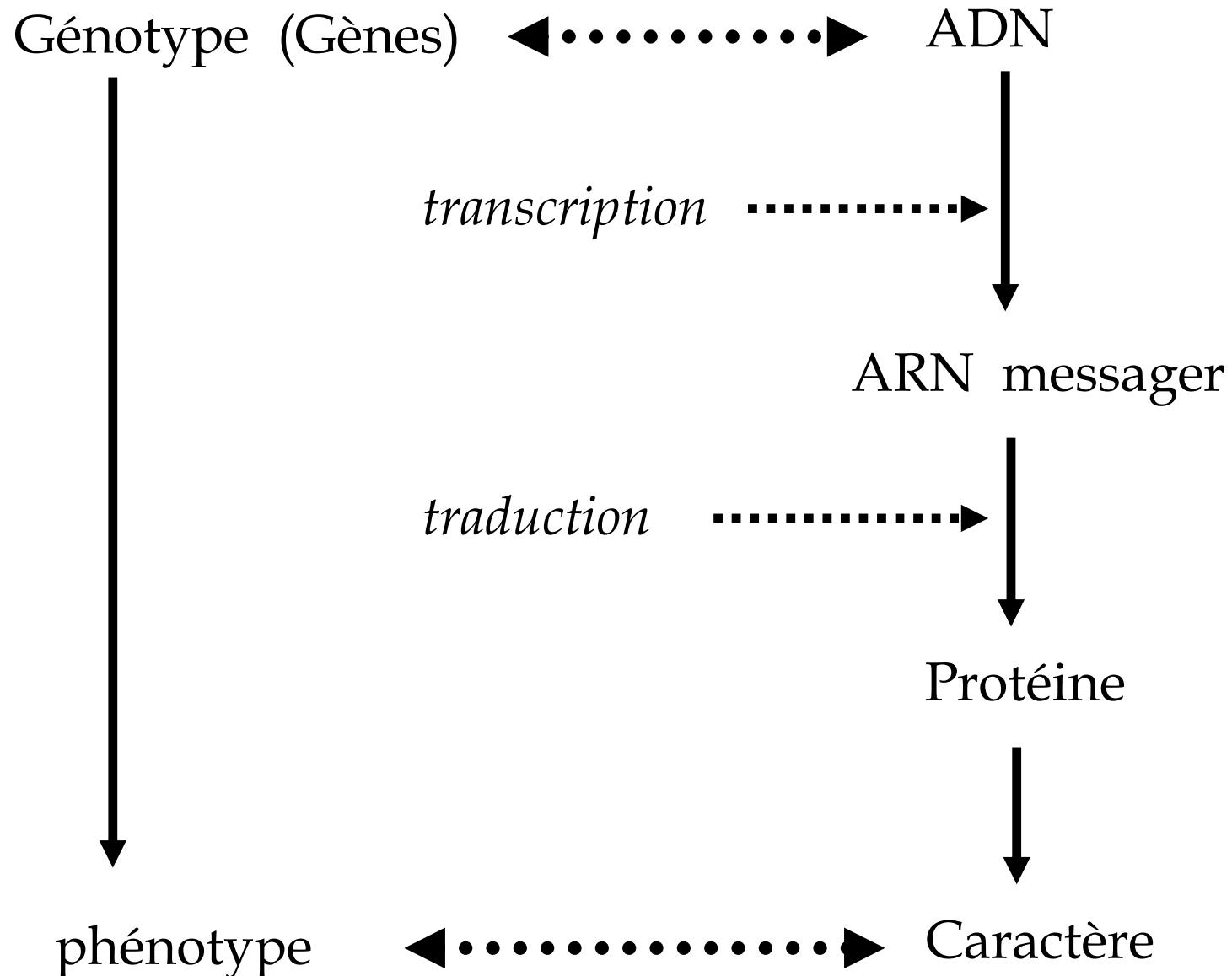
Principaux champs d'intérêt de la bioinformatique

- Collecte, stockage, organisation, gestion des données biologiques généralement sous forme de séquences nucléotidiques/protéiques et ChIP-Seq
- Distribution des données et mise à disposition de la communauté scientifique
- Analyse et interprétation de l'information biologique collectée : recherche de structures, identification de fonctions, classification des espèces, mécanismes d'évolution des êtres vivants, etc.
- Production d'outils (algorithmes et programmes informatiques) pour mieux répondre aux différents besoins

Macromolécules Support de l'information biologique

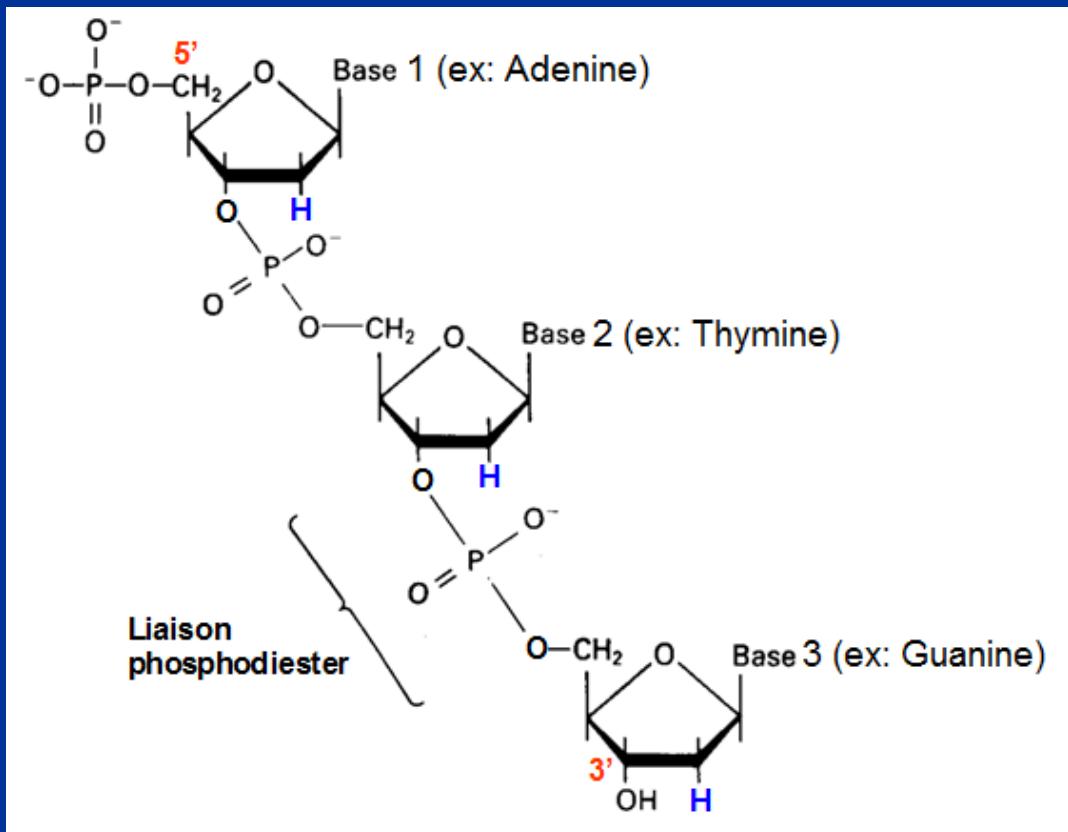
- L'information génétique héréditaire des organismes vivants est portée par deux macromolécules : ADN et ARN
- Cette dernière est fonctionnelle lorsqu'elle est traduite en Protéines
- Niveau ADN ou ARN : information génétique est sous forme de gènes >>> caractères phénotypiques
- L'expression phénotypique des gènes dépend de leur état c-a-d du génotype
- Le passage du génotype au phénotype constitue un ensemble de processus appelés Dogme Central

Le Dogme central



Généralité sur les acides nucléiques : ADN et ARN

- L'ADN est un polymère de désoxyribonucléotides (dNTP) formés par : le désoxyribose triphosphate et une base azoté purique R (Adénine, Guanine) ou pyrimidique Y (Cytosine, Thymine)
- Les dNTPs sont reliés entre eux par des liaisons phosphodiester

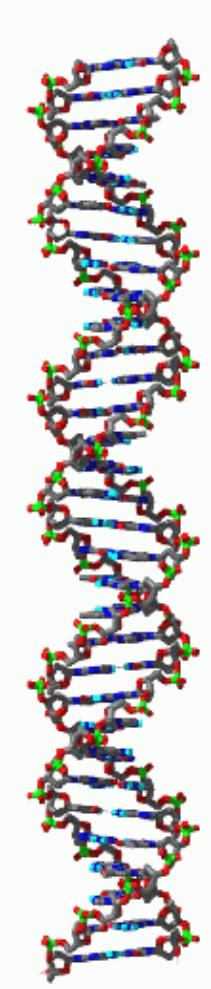


- extrémité 5'-Phosphate du 1^{er} dNTP libre
- extrémité 3'-OH du dernier dNTP libre

5'AGTCCGTAAGTCGGCTT3'

Généralités sur les acides nucléiques : l'ADN

- L'ADN est généralement bicaténaire (double brin) : les deux brins sont complémentaires; règle de complémentarité A-T et G-C
- L'ADN de certains virus est monocaténaire (un brin)
- La succession des nucléotides dans la molécule d'ADN est la séquence de cette molécule ADN
- La longueur de la séquence est exprimée en nombre de nucléotides (ADN monocaténaire) ou en paires de bases, pdb (ADN bicaténaire)
- La perpétuation de l'information génétique est réalisée via la réPLICATION sémi-conservative de l'ADN



Généralités sur les acides nucléiques : l'ARN

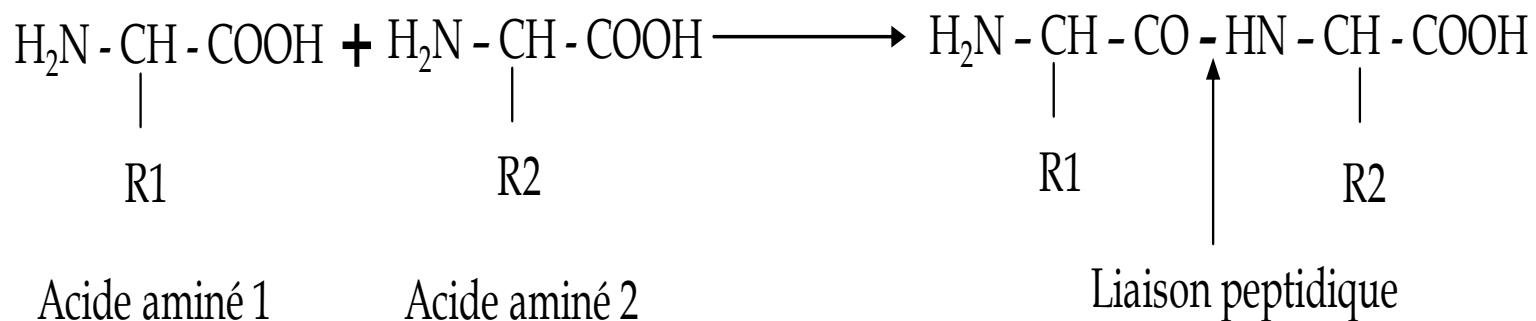
- L'ARN est généralement issu de la transcription de l'ADN; il constitue aussi le matériel génétique des viroïdes et de certains virus
- Il y a trois types d'ARN : messager, ribosomal, transfert; seul le messager est porteur de l'information génétique
- L'ARN est formé sur le même modèle que l'ADN mais avec deux grandes différences : le ribose remplace le désoxyribose / les bases azotés sont : A, U, G, C (complémentarité A-U et G-C)
- L'ARN est monocaténaire mais peut avoir des zones de complémentarité entraînant des structures bicaténaires

Généralités sur protéines

- Les protéines sont des chaînes polypeptidiques formées à partir d'acides aminés
- Les protéines sont issues de la traduction de l'information présente au niveau de l'ARNm; les parties correspondantes au niveau de l'ADN sont dites régions codantes
- Les différentes séquences d'une molécule d'ADN capables de coder pour une protéine sont identifiées sous le nom de cadres ouverts de lecture ou Open Reading Frames, ORF
- La traduction ARNm>>>Protéines se fait par l'intermédiaire du Code génétique qui fait correspondre un codon (triplet de nucléotides) à un acide aminé donné.

Généralités sur protéines

- Les protéines sont représentées par leurs séquences d'acides aminés en utilisant le code à une lettre
- Chaque séquence a une extrémité N-terminal et une extrémité C-terminal



MARKGKKINSNQGQQGKKKSRRPRGRSVEPQ

Extrémité N-terminale

Extrémité C-terminale

Deux champs particuliers d'application bioinformatique

(1) L'évolution moléculaire

- Des erreurs de copies se produisent pendant la réPLICATION ou la transcription de l'ADN
- Ces changements dans la séquence d'ADN ou mutations (insertions, délétions, substitutions de bases) peuvent être transmises à la descendance
- L'accumulation progressive des mutations est à la base du processus de l'évolution des organismes
- L'évolution des organismes a d'abord été appréhendée par l'intermédiaire des caractères phénotypiques
- Séquences, nouveaux outils pour inférer l'histoire évolutives des êtres : origine, mécanismes évolutifs, classification etc.

(2) La modélisation des systèmes biologiques

- Les fonctions biologiques sont régies par des systèmes souvent complexes
- Pour comprendre ces fonctions > nécessité de structures moins complexes appelées modèles : exemple : structure tridimensionnelle des protéines, de l'ADN; mécanisme de la production de l'ATP, etc.
- Le modèle doit cependant être valide c-a-d permettre d'identifier et d'extraire les paramètres pertinents du système biologique
- L'approche bioinformatique consiste à intégrer de façon quantitative les paramètres identifiés, de mettre en œuvre des méthodes de calculs afin de décrire les propriétés du système et d'en prédire les comportements

Gestion et distribution des données biologiques : les bases de données

- 1972, mise sur pied de la première banques de données biologiques: la Protein Information Ressource, PIR
- Mais les véritables banques ou bases de données biologiques ont été établis dans les années 1980s.
- Au début, banques de données nationales mais intégration rapide de plusieurs banques pour une plus grande efficacité et éviter les doublons

- Types de données dans les banques : séquences nucléiques puis protéiques, données de cartographie génique, d'expression de gènes, motifs protéiques etc.
- Fiabilité des données : assez bonne mais erreurs existent notamment sur les toutes premières séquences soumises; exemple : segments de vecteurs de clonage présents dans les séquences de certains organismes

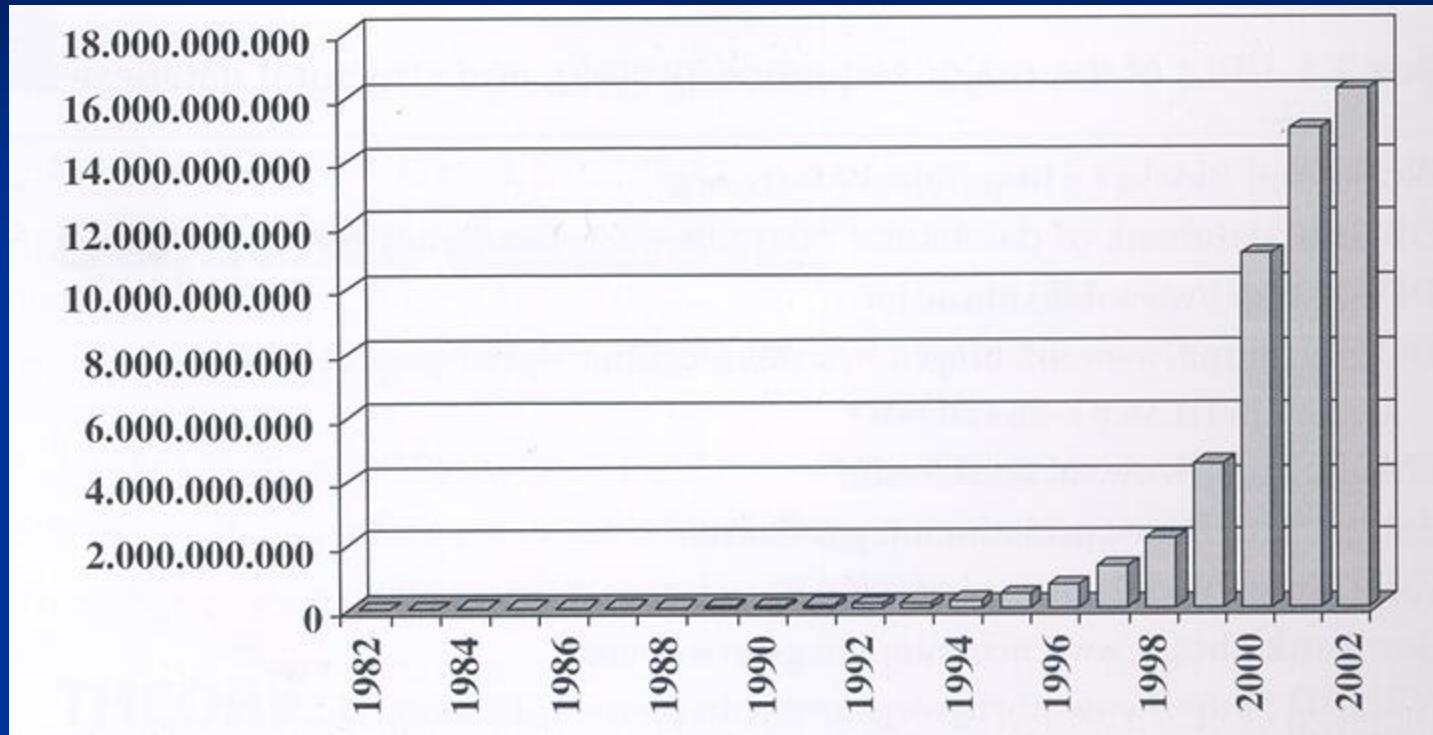
Format de conservation des données biologiques

- Quelque soit le type de banque, les données sont conservées sous forme de fichiers informatiques
- Chaque donnée stockée possède un numéro d'identification propre (Accession number); chaque banque a cependant sa propre méthode d'attribution des numéros d'identification; ex AF192135 pour la séquence complète d'un isolat du HIV1 chez Genebank
- Grâce à l'interface web, les numéros d'identification sont généralement des liens conduisant aux données correspondantes

Quelques formats de fichiers de conservation des données

Programme ou [format]	Type de fichier	Extension
EditSeq (Lasergene)	Sequence nucléique individuelle	*.seq
	Séquence protéique individuelle	*.pro
FASTA	Séquence individuelle	*.fas
[Genbank Flat File]	Séquence individuelle	*.gbk
GCG DNA File	Sequence nucléique	*.gcf
GCG Protein File	Séquence protéique	*.pep
Megalign (Lasergene)	Alignement de séquences	*.meg
CLUSTAL	Alignement de séquences	*.aln
Paup ([paup] ou[nexus])	Alignement de séquences	*.pau ou *.nex
Trevview	Arbre phylogénétique	*.tre, *.dnd, *.phy

Augmentation du nombre de séquences disponibles dans les bases de données (en 2012 +69 billiards)



International Nucleotide Sequence Database Collaboration
Genbank, EMBL, DDJB @ NCBI (NIH)

General nucleic acid sequence databases:

- ✓ EMBL (European Molecular Biology Laboratory) database, maintained at EMBL-EBI (European Bioinformatics Institute, Hinxton, UK)
- ✓ GenBank, maintained at NCBI (National Center for Biotechnology Information, Bethesda, Maryland, USA)
- ✓ DDBJ (DNA Data Bank of Japan), maintained at NIG/CIB (Mishima, Japan)

Obtention des séquences

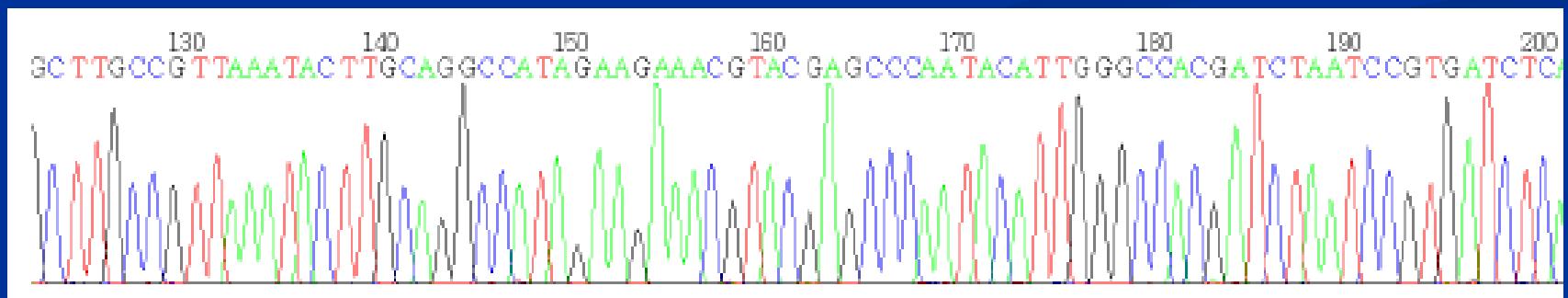
1 - Clonage et séquençage



Extraction
ADN
ou
ARN

Amplification
PCR
ou
RT-PCR

Séquençag
e ADN
ou
ADNc



2 - BLAST

Basic Local Alignment and Search Tool

NCBI Blast - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&CLIENT=web&DA=

Google Alta Vista Exalead

Exalead - Rechercher sur Internet avec le ... Clicanoo | Journal de l'Île de la Réunion, le q... NCBI Blast

NCBI nucleotide-nucleotide BLAST

Nucleotide Protein Translations Retrieve results for an RID

ACCGGTTGGC CCCGCCCCCT TTAATGTGGT CCCCAGGCAC TACTTATGTC
GGCCATTCAT
61 GATGTAGCTT TAAAGGTTAT GTATTAGTGG TGGGCCACTA
TATACTTGCA GGCGAAGTTG
121 TGGCTAGTGC GCAATGTGGG ATCCACTGGT GAATGAGTTT
CCAGACTCGG TGCATGGCT

Search Set subsequence From: To:

Choose database Human genomic plus transcript
Mouse genomic plus transcript
Others (nr etc.): nr

NEW Two new Human and Mouse databases combine genomic plus transcript alignments in a single report. You can also choose from Others to use nr or an existing database.

Now: **BLAST!** or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: All organisms

Choose filter Low complexity Repeats Human Mask for lookup table only Mask lower case

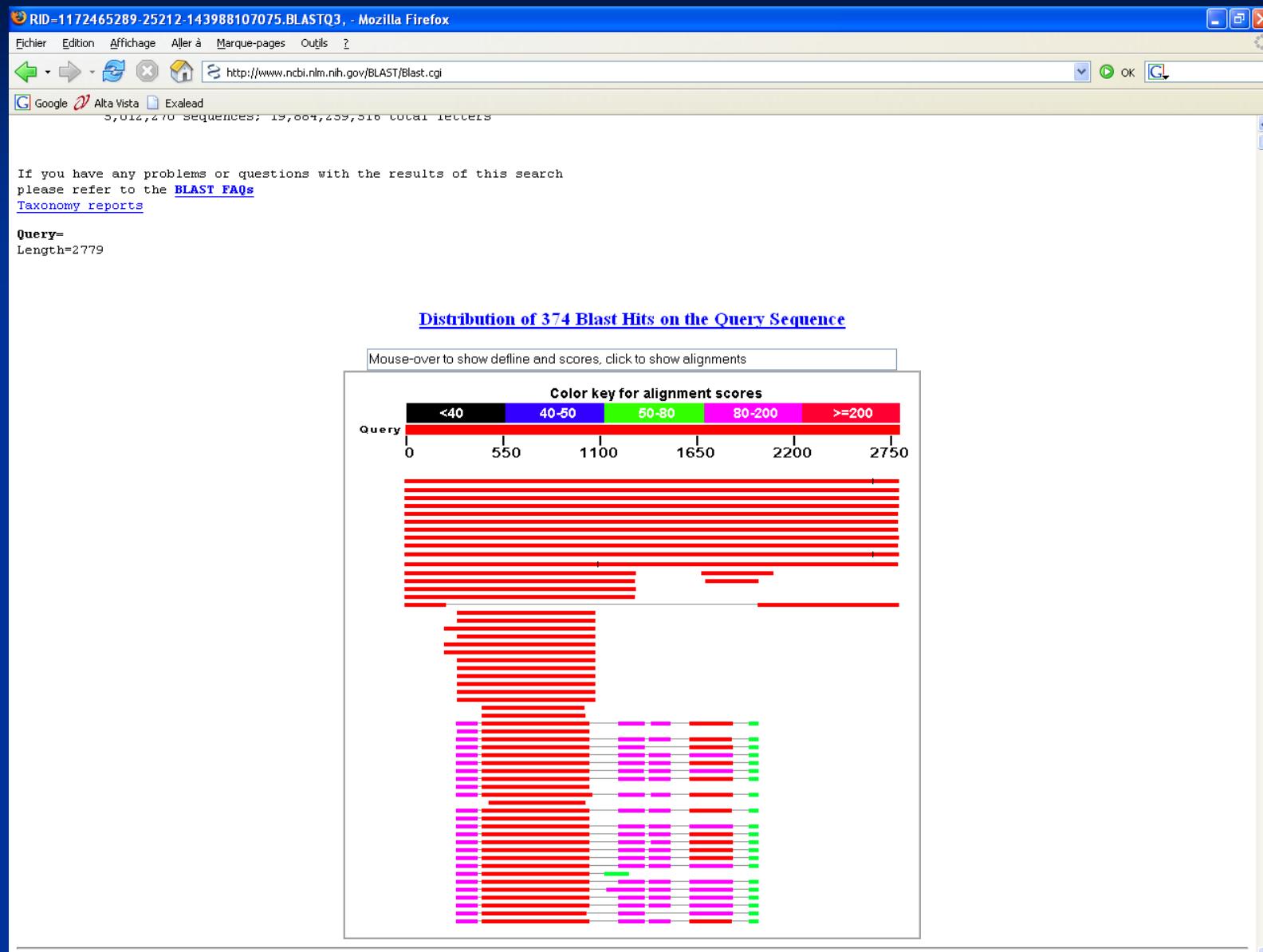
Expect 10

Word Size 11

Other advanced

The screenshot shows the NCBI BLAST search interface. At the top, the browser title bar reads "NCBI Blast - Mozilla Firefox". The menu bar includes "Fichier", "Edition", "Affichage", "Aller à", "Marque-pages", "Outils", and a question mark icon. The address bar shows the URL "http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&CLIENT=web&DA=". Below the address bar is a toolbar with icons for Google, Alta Vista, and Exalead, followed by a search bar containing "Clicanoo | Journal de l'Île de la Réunion, le q...". The main content area has a blue header "nucleotide-nucleotide BLAST" with tabs for "Nucleotide", "Protein", "Translations", and "Retrieve results for an RID". The main search area contains a sequence alignment between a query sequence (ACCGGTTGGC CCCGCCCCCT TTAATGTGGT CCCCAGGCAC TACTTATGTC) and a database sequence (GGCCATTCAT). The alignment is shown with arrows indicating matches. Below the alignment are buttons for "Search", "Set subsequence", and "From:" and "To:" input fields. A "Choose database" dropdown menu is open, showing "Human genomic plus transcript", "Mouse genomic plus transcript", and "Others (nr etc.)" with "nr" selected. A note in a box says "NEW Two new Human and Mouse databases combine genomic plus transcript alignments in a single report. You can also choose from Others to use nr or an existing database." At the bottom, there are buttons for "BLAST!", "Reset query", and "Reset all". A second section at the bottom is titled "Options for advanced blasting" and includes fields for "Limit by entrez query" (with a dropdown for "All organisms"), "Choose filter" (with checkboxes for "Low complexity", "Repeats", "Human", "Mask for lookup table only", and "Mask lower case"), "Expect" (set to 10), "Word Size" (set to 11), and "Other advanced" (with an empty input field).

BLAST



3 - Obtenir tout un groupe de séquences

The screenshot shows the NCBI homepage. At the top, there is a navigation bar with links for PubMed, All Databases, BLAST, OMIM, Books, TaxBrowser, and Structure. Below the navigation bar is a search bar with dropdown menus for 'Search' and 'All Databases', and a 'Go' button. To the left, there is a sidebar with links for SITE MAP, Alphabetical List, Resource Guide, About NCBI, An introduction to NCBI, GenBank, and Sequence. The main content area features a section titled 'What does NCBI do?' which describes the organization's mission. To the right of this section is a 'Hot Spots' sidebar with links to Assembly Archive, Clusters of orthologous groups, Coffee Break, Genes & Disease, and NCBI Handbook. A red curved arrow points from the text below to the 'TaxBrowser' link in the navigation bar.

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence

National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books **TaxBrowser** Structure

Search All Databases for Go

▶ What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes

Hot Spots

- ▶ Assembly Archive
- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook

Aller dans TaxBrowser et taper le nom de l'organisme dont on souhaite les séquences

ex : begomovirus

 NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for as lock

Display levels using filter:

Begomovirus

Taxonomy ID: 10814
Rank: genus
Genetic code: [Translation table 1 \(Standard\)](#)
Other names:
synonym: [Geminivirus subgroup III](#)
synonym: [Subgroup III Geminivirus](#)

[Lineage \(full\)](#)
[Viruses](#); [ssDNA viruses](#); [Geminiviridae](#)

[ICTV homepage](#)

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	2,286	-
Protein	5,960	-
Structure	2	-
Genome Sequences	191	-
Popset	12	3
3D Domains	2	-
PubMed Central	127	36
Gene	1,028	-
Taxonomy	603	1

Séquence nucléotidique

Génome complet

NB : Toutes les séquences complètes ne sont pas enregistrées en tant que tel

Choisir « FASTA » dans le menu « Display » et « File » dans le menu « Send To »

On a alors un fichier FASTA avec toutes les séquences disponibles complètes ou non

→ Trier les séquences par la suite

Alignment

Point de départ d'une analyse de séquence :
un lot de séquences d'acides nucléiques ou de protéines **aligné**

Aligner, c'est rechercher et mettre en concordances mes résidus homologues d'un jeu de séquence

La qualité de l'alignement est cruciale

On peut faire disparaître le signal en alignant « trop »

Chaque colonne d'un alignement est supposée contenir des résidus homologues dérivant d'un ancêtre commun

Parfois les portions incertaines d'un alignement doivent être éliminée

<i>Thermus ruber</i>	UCCGAAAGC-UAAAAGA-CCGAAAG=CUCAA=CUUCGG=GGGU=GCGUUGGA
<i>Th. thermophilus</i>	UCCGAAAGU-GAAAAGA-CCACGG=CUCAA=CCGUGGG=GGGA=GCGUGGGGA
<i>E.coli</i>	UCAGAAAGU-GAAAAC-CCCGGG=CUCAA=CCUGGG=ACAU=GCAUCUGA
<i>Ancyst. nidulans</i>	UCUGUUGU-CAAAGC-GUGGGG=CUCAA=CCUCAU=ACAG=GCAAUUGGA
<i>B.subtilis</i>	UCUGUUGU-GAAAAGC-CCCGGG=CUCAA=CCGGGG=AGGG=UCAUUGGA
<i>Chl. aurantiacus</i>	UGGGGCGU-GAAAAGC-GCCCCG=CUCAA=CGGGGC=GGGG=CGCGCGCA
match	*** → *** * * * * *

Match = identité

Mismatch = mutation

Gap = indel

NB : La plupart des méthodes d'analyse prennent en compte les substitutions mais pas les gaps

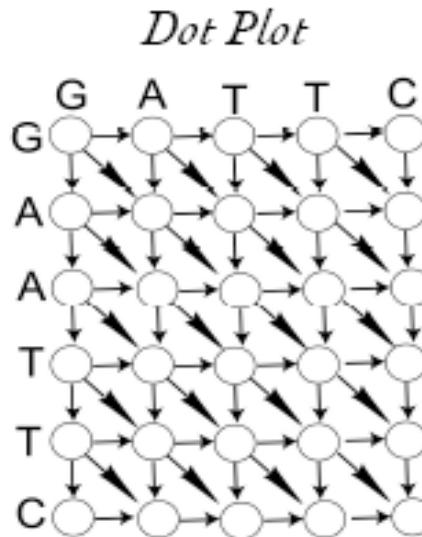
- Alignement plus ou moins facile
- Séquence codante ou pas
 - Utiliser les AA (codons) pour alignement
 - Considérer les types d'AA (taille, polarité, hydrophobicité)
- Séquences plus ou moins divergentes
- Homologie variable selon région
- Alignement atteint par ajout d'événements d'insertion-délétion (*indels*) à l'aide de *gaps* : limités par *pénalités* (sauf aux extrémités)

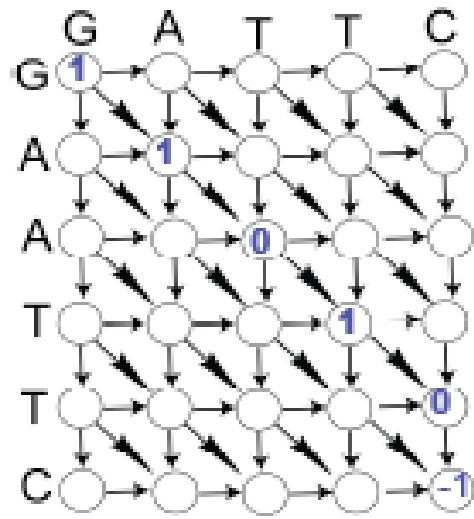
Exemple d'alignement manuel

- But de l'alignement automatique : maximiser le *score* de l'alignement
- Exemple

GATTC
GAATTC

On définit :
Match - +1
Mismatch - 0
Indel - -1



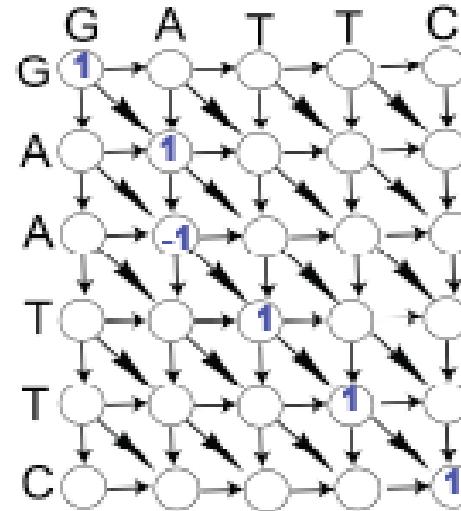


GATTC-

GAATTC

Score = 2

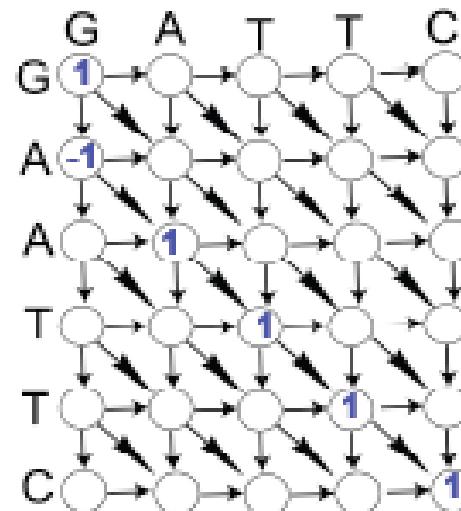
2 alignements optimaux



GA-TTC

GAATTC

Score = 4



G-ATTC

GAATTC

Score = 4

Match +1

Mismatch 0

Gap -1

- En plus de la pénalité d'introduction des gaps (*gap opening penalty*), on définit une pénalité pour l'extension des gaps (*gap extension penalty*), moins élevée (encourage extension, pas des trous partout)
- GOP et GEP peuvent varier le long des séquences, en fonction de la présence de gaps et de caractéristiques biochimiques (e.g. AA hydrophiles)
- On peut aussi pondérer différemment les substitutions (certaines sont plus faciles que d'autres ; e.g. pour AA : matrice BLOSUM 62)

- Problème complexe analytiquement : on ne peut garantir le “meilleur” alignement quand le nombre de séquences augmente (alignement multiple)
- Alignement progressif (e.g. *Clustal*)
 - Calcul d'un arbre-guide (NJ) pour alignement des paires de séquences
 - Aline d'abord les séquences les plus proches et ainsi de suite
 - Rapide mais pas de critère d'optimalité

Format clustal

CLUSTAL W (1.8) multiple sequence alignment

Homo_sapiens	AGUCGAGUC---GCAGAAAC
Pan_paniscus	AGUCGCGUCG--GCAGAAAC
Gorilla_gorilla	AGUCGCGUCG--GCAGAUAC
Pongo_pigmaeus	AGUCGCGUCGAAGCAGA--C
	***** *** ***** *

Homo_sapiens	GCAUGAC-GACCACAUUUU-
Pan_paniscus	GCAUGACGGACCACAUCAU-
Gorilla_gorilla	GCAUCACGGAC-ACAUCAUC
Pongo_pigmaeus	GCAUGACGGACCACAUCAUC
	***** ** *** ***** *

Homo_sapiens	CCUUGCAAAG
Pan_paniscus	CCUUGCAAAG
Gorilla_gorilla	CCUCGCAGAG
Pongo_pigmaeus	CCUUGCAGAG
	*** *** **

Format Phylip

4 50

Homo sapie AGUCGAGUC---GCAGAAACGCAUGAC-GACC
Pan panisc AGUCGCGUCG--GCAGAAACGCAUGACGGACC
Gorilla go AGUCGCGUCG--GCAGAUACGCAUCACGGAC-
Pongo pigm AGUCGCGUCGAAGCAGA--CGCAUGACGGACC

ACAUUUU-CCUUGCAAAG
ACAUCAU-CCUUGCAAAG
ACAUCAUCCCUCGCAGAG
ACAUCAUCCCUUGCAGAG

Format FASTA

```
>Homo_sapiens
AGUCGAGUC---GCAGAAACGCAUGAC-GACCACAUUUU-CCUUGCAAAG
>Pan_panis
AGUCGGCGUCG--GCAGAAACGCAUGACGGACCACAUCAU-CCUUGCAAAG
>Gorilla
AGUCGCGUCG--GCAGAUACGCAUCACGGAC-ACAUCAUCCCUCGCAGAG
>Pongo_pigmy
AGUCGCGUCGAAGCAGA--CGCAUGACGGACCACAUCAUCCUUGCAGAG
```

Format Nexus

```
#NEXUS
[TITLE: Four Anthropoidea]

begin data;
dimensions ntax=4 nchar=50;
format interleave datatype=RNA missing=N gap=-;

matrix
Homo_sapiens          AGUCGAGUC---GCAGAACGCAUGAC-GAC
Pan_paniscus           AGUCGCGUCG--GCAGAACGCAUGACGGAC
Gorilla_gorilla        AGUCGCGUCG--GCAGAUACGCAUCACGGAC
Pongo_pigmaeus         AGUCGCGUCGAAGCAGA--CGCAUGACGGAC

Homo_sapiens           CACAUUUU-CCUUGCAAAG
Pan_paniscus           CACAUCAU-CCUUGCAAAG
Gorilla_gorilla        -ACAUCAUCCCUCGCAGAG
Pongo_pigmaeus         CACAUCAUCCCUCUGCAGAG
;
```

Mesure de la diversité

■ Estimation de la diversité génétique (BIOEDIT, MEGA)

■ % d'identités (entre 2 séquences)

$$\% = (1 - I / N) * 100$$

N : nombre de sites de l'alignement
I : nombre de sites différents

	I/N
CMV1	ATGGGATTGAACGAGAGCGCAGTGACAAACGT CGA ACTCCAGCTGGCTCGTAAAGAAGACC
CMV2 G
CMV3 C

	I/N
CMV1	4/61 93.4 %
CMV2	2/57 96.5%
CMV3	6/61 90.2 %

	I/N
CMV1	1/20 95.0 %
CMV2	8/19 57.9 %

■ Indice de diversité nucléotidique (n séquences)

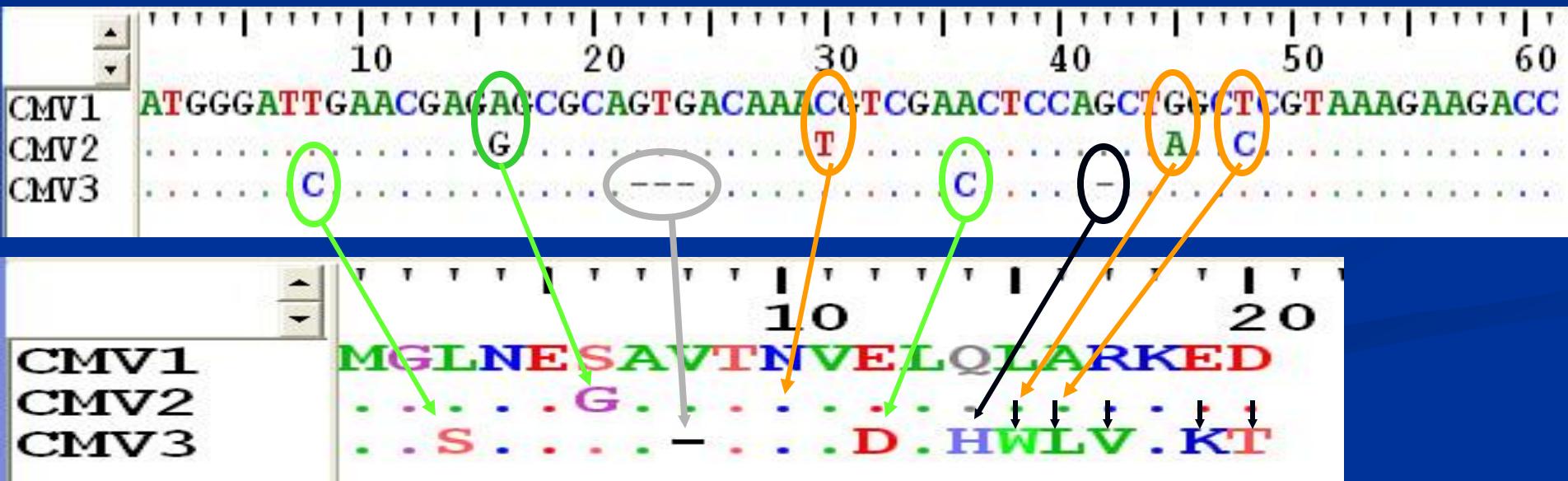
$$\pi = n / (n - 1) \sum X_i X_j \pi_{ij}$$

X_i : fréquence estimée de la i^e séquence

π_{ij} : proportion de nucléotides différents entre les séquence i et j (= 1-%ident.)

MutatioNs

■ Synonyme (silencieuse) ou non synonyme

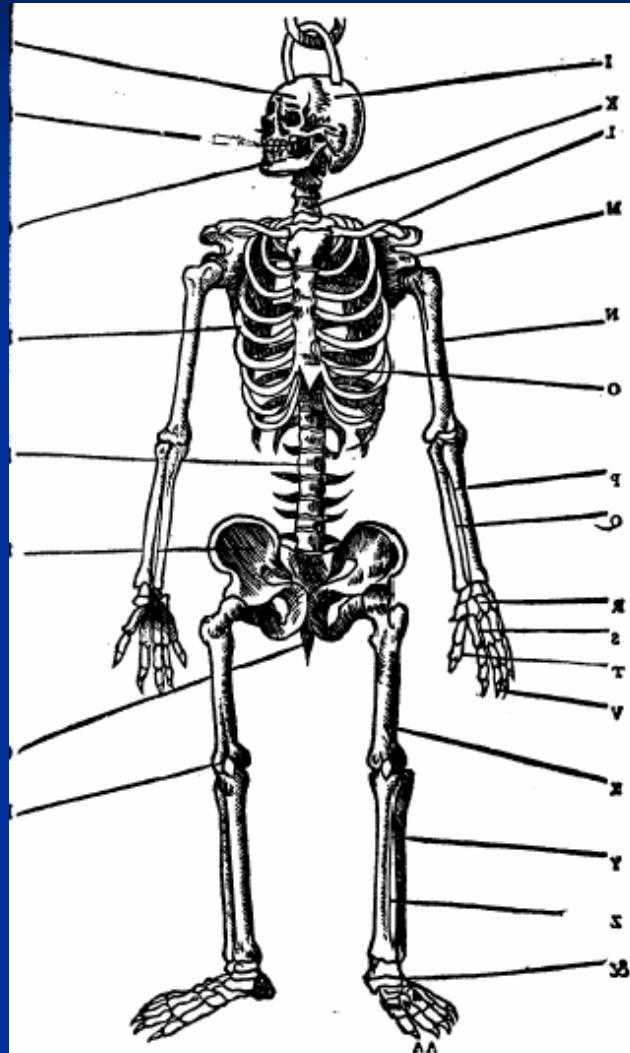


NB : valable dans le cas des séquences codantes

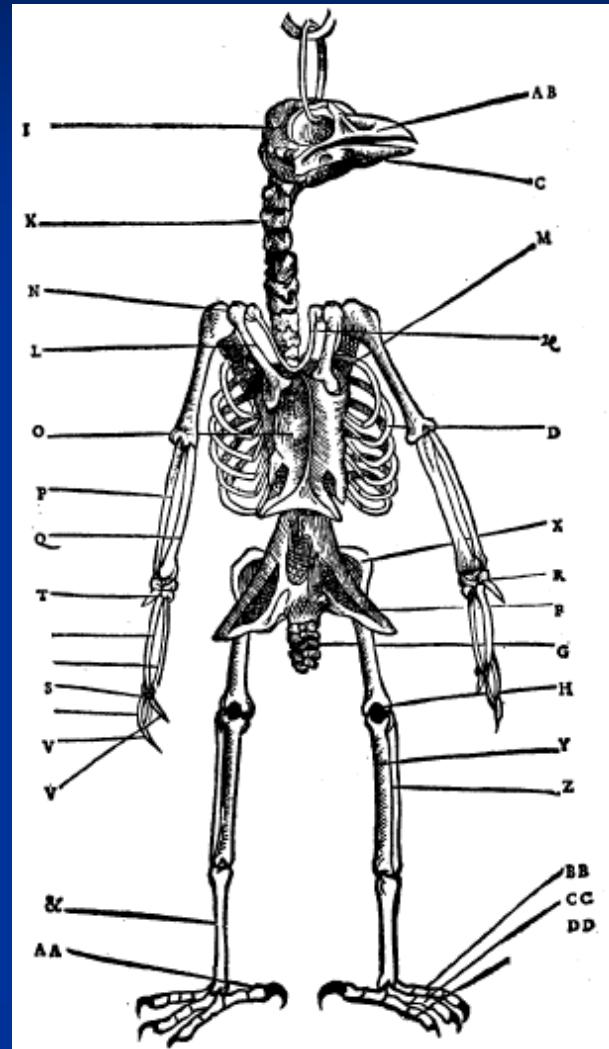
- Taux de mutation indicateur de la diversité génétique
 - Taxonomie
- Accumulation de séquences nucléotidiques diverses permet d'inférer les processus évolutifs sous jacents
 - Phylogénie
 - Mesure de la pression de sélection

Phylogénie

La biologie comparative est ancienne...

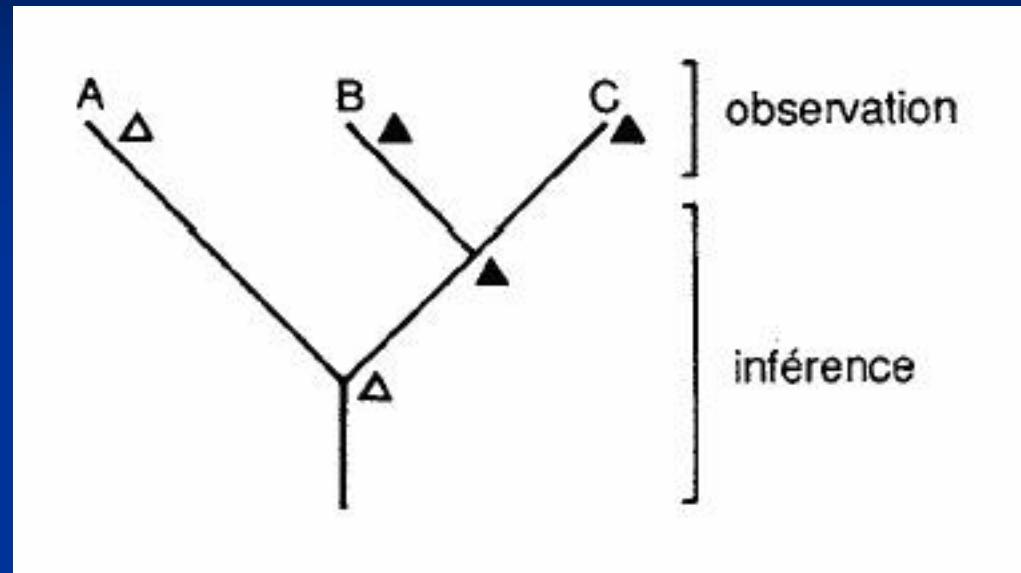


L'ame des os humains a des similitudes avec l'anatomie des os d'un oiseau.
G. Belon (1557).



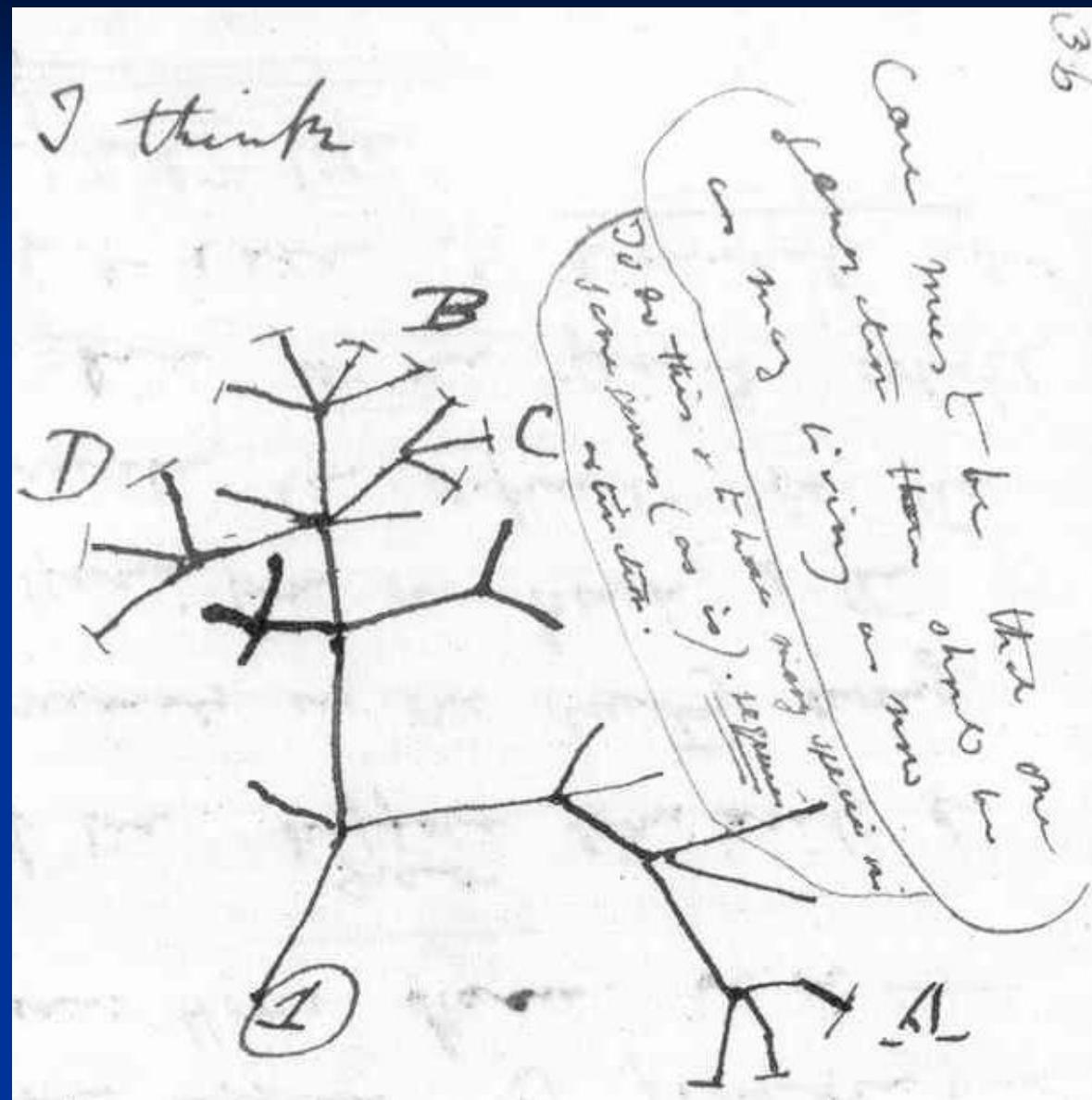
Anatomie des os d'un oiseau mise en comparaison à celle de l'homme,
pour montrer l'affinité des deux.
P. Belon (1557).

L'inférence phylogénétique



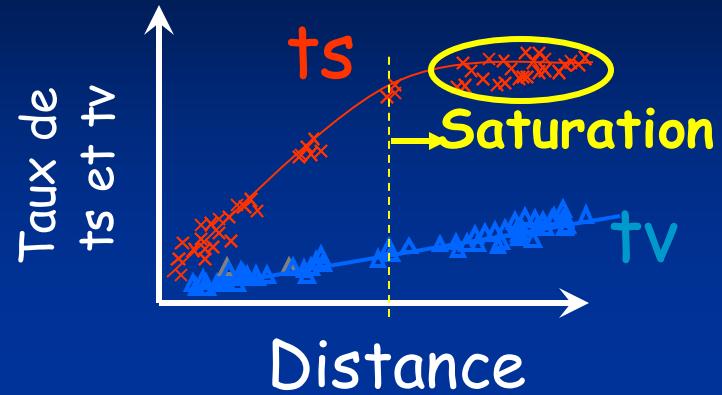
- Reconstruire l'histoire évolutive de taxons ou de caractères
- Analyser les caractères et les vitesses d'évolution

Premier arbre phylogénétique par Darwin



■ Reconstruction phylogénétique

- Alignement des séquences
- Modèle de mutations (! saturation)
- Méthodes
 - Phénétiques (distances) : UPGMA, NJ
 - Cladistique (substitutions minimales) : parcimonie
 - Probabiliste (ML, Bayesian) : Quartet Puzzling, MrBayes
- Hypothèses d'horloge moléculaire

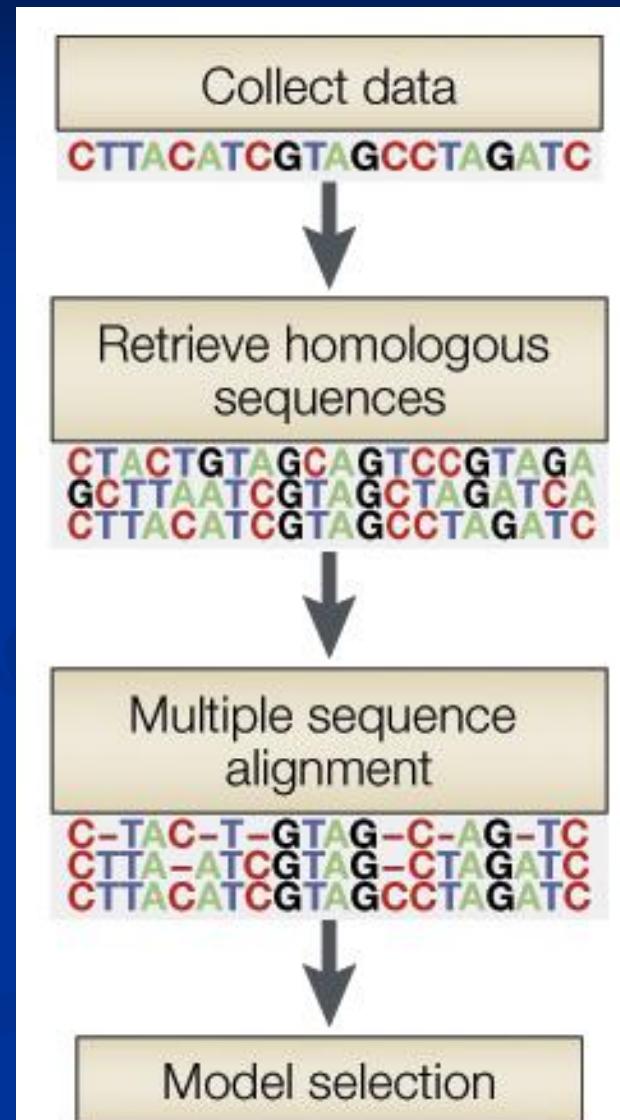


Le processus de reconstruction phylogénique (1)

Récolte des données

Généralement quelques
“outgroup” sont inclus

L'alignement pour que les
nucléotides d'une colonne
soit issus d'un même
ancêtre commun (gaps)



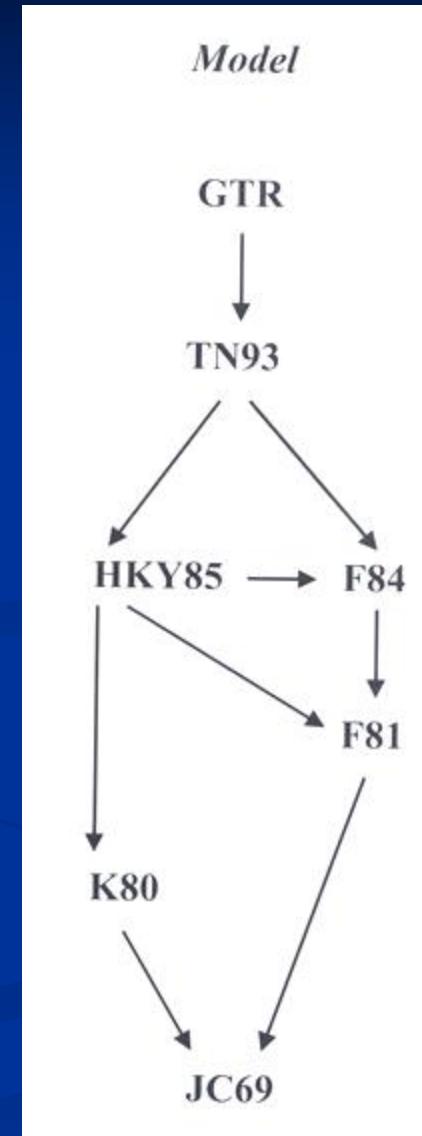
Le processus de reconstruction phylogénique (2)

Un modèle d'évolution des données doit être choisi

Augmenter la complexité du modèle augmente l'explication des données mais augmente aussi la variance des paramètres estimés

Les stratégies de sélection de modèle ont pour but de déterminer le niveau de complexité du modèle adapté à un jeu de données

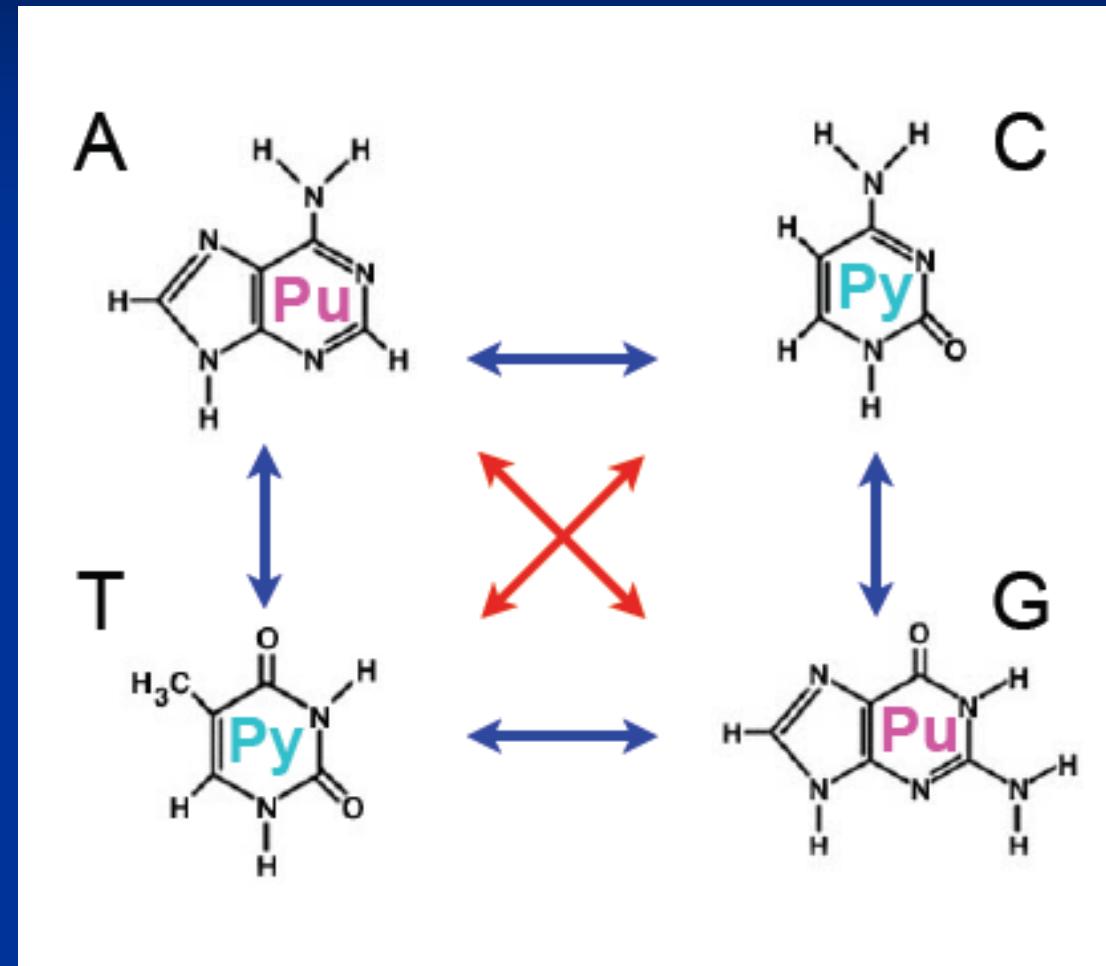
Model Test permet de déterminer aisément le modèle à choisir



Modèles d'évolution

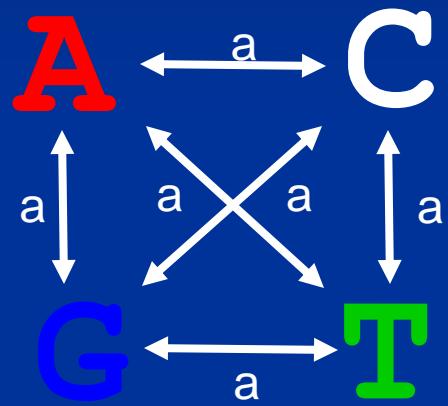
Transversions (tv)

Transitions (ts)

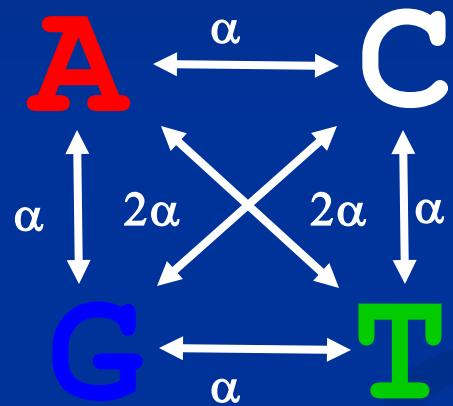


ts > tv

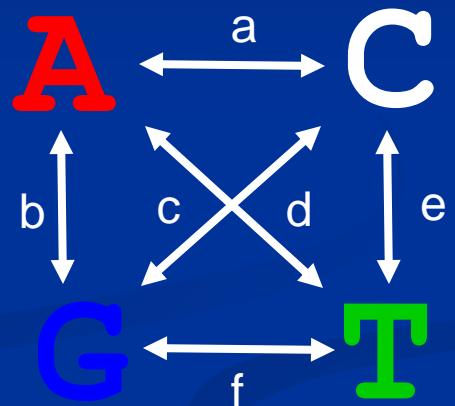
Jukes Cantor



Kimura

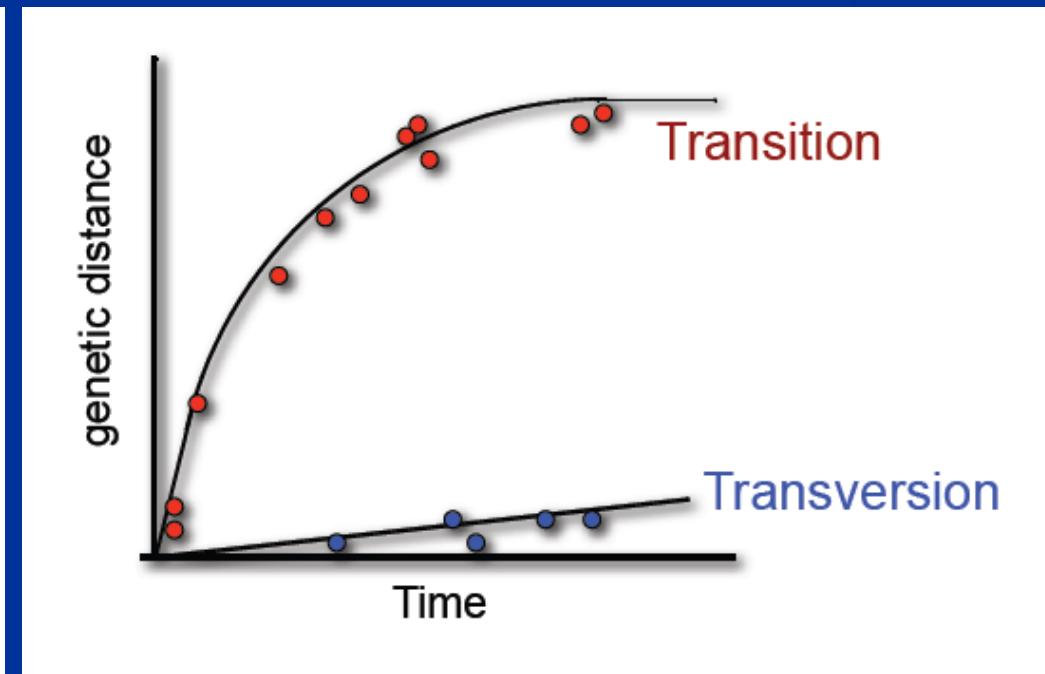
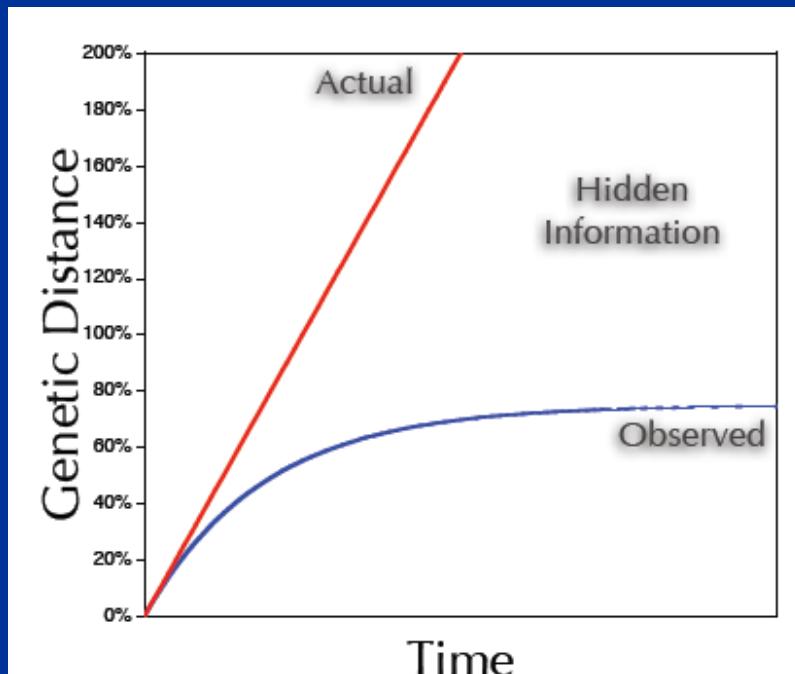


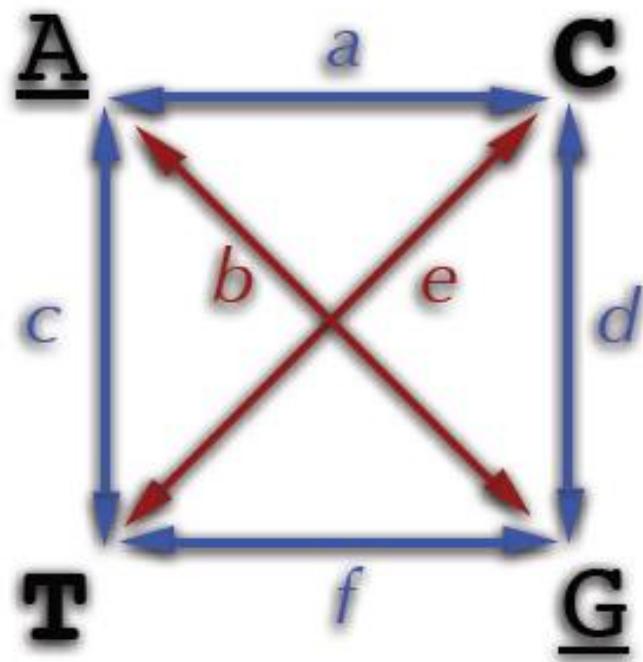
General



Saturation: perte du signal phylogénétique

Lorsque les séquences comparées ont subies un trop grand nombre de substitutions depuis leur divergence, il est impossible de déterminer un arbre phylogénétique





a, b, c, d, e, f = relative rate parameter

$$Q = \begin{pmatrix} . & \mu a \pi_c & \mu b \pi_G & \mu c \pi_T \\ \mu a \pi_A & . & \mu d \pi_G & \mu e \pi_T \\ \mu b \pi_A & \mu d \pi_C & . & \mu f \pi_T \\ \mu c \pi_A & \mu e \pi_C & \mu f \pi_G & . \end{pmatrix}$$

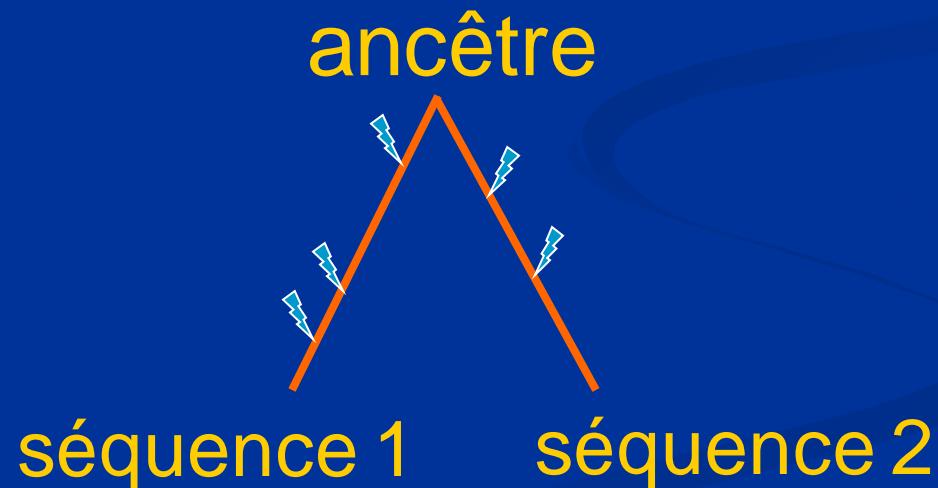
μ = nucleotide substitution rate

π_X = frequency of base X
(usually estimated from the data)

Distance évolutive

Mesure du nombre de substitutions qui ont eu lieu depuis la divergence avec le dernier ancêtre commun

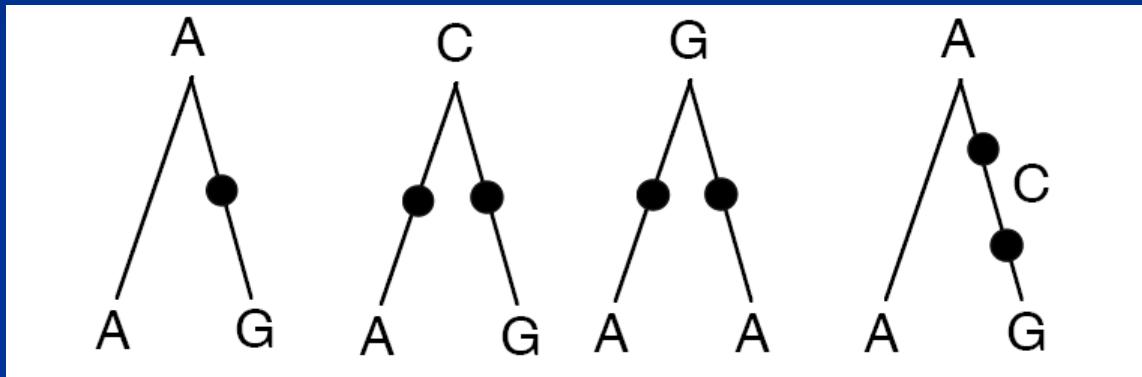
Exprimé en substitutions par sites



Distance évolutive

- Problème des changements multiples

d (dist évol. vraie) > p (% substitutions)



$$d = p + \text{changements cachés}$$

Des Hypothèses de régularité permettent d'estimer d à partir de p

$$d = -\frac{1}{2} \ln[(1 - 2p - q)\sqrt{1 - 2q}]$$

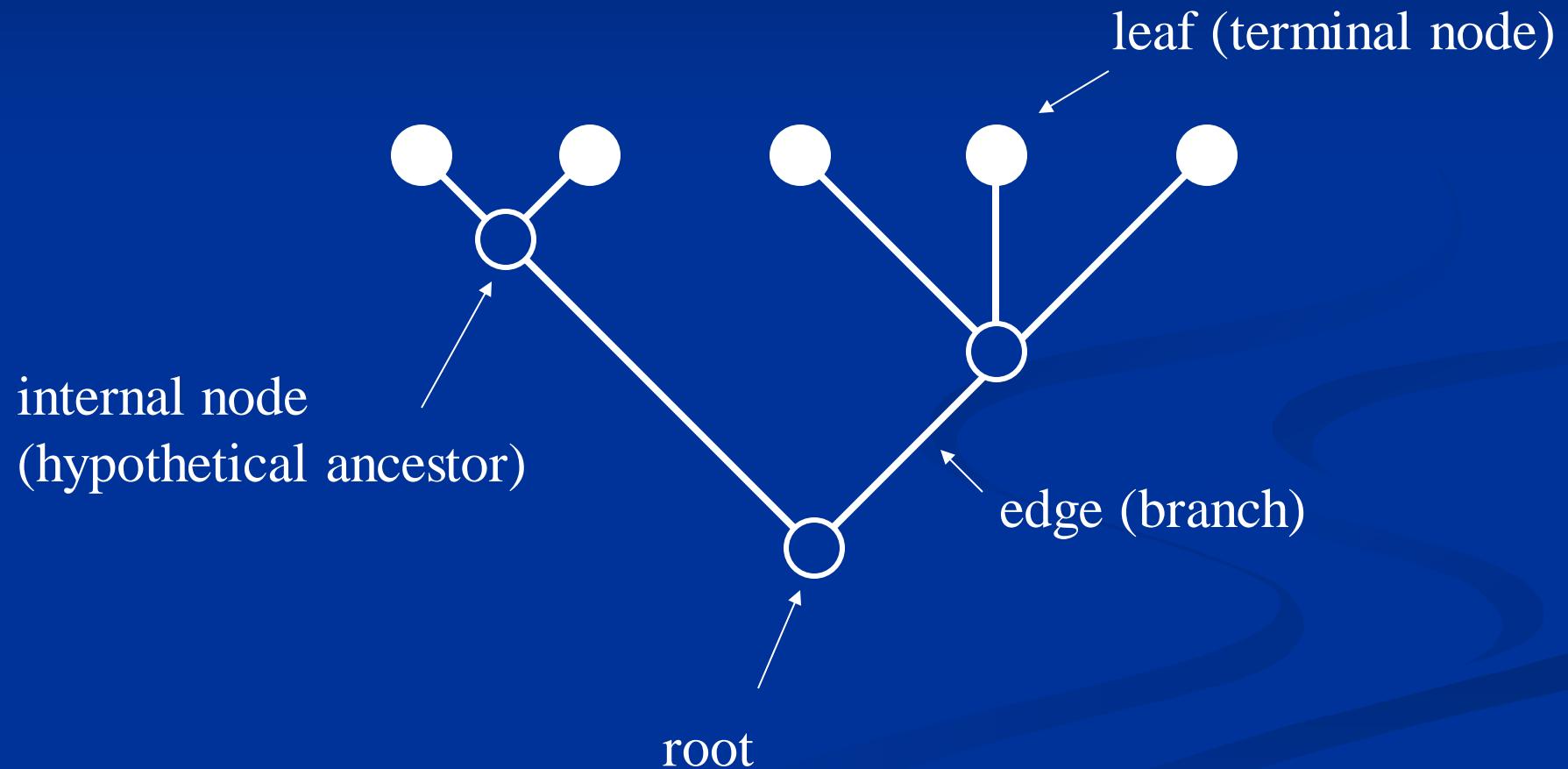
Kimura (1980) J. Mol. Evol. 16:111

d : distance

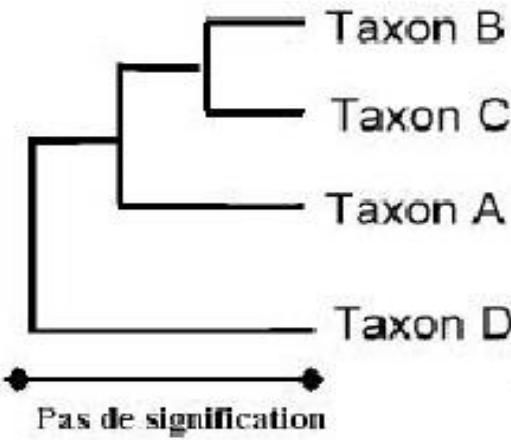
p : taux de transition

q : taux de transversion

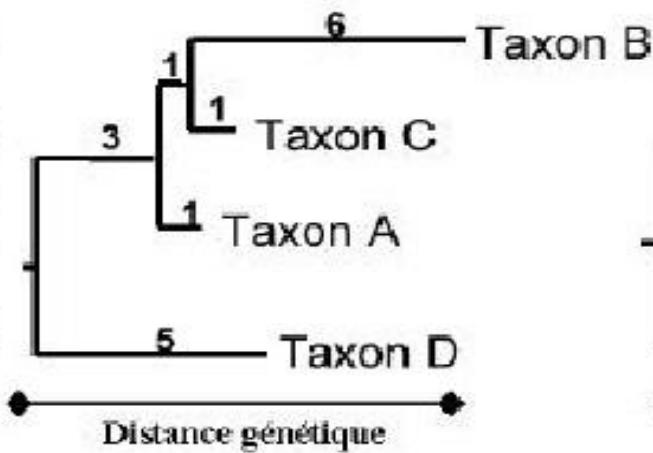
Terminologie



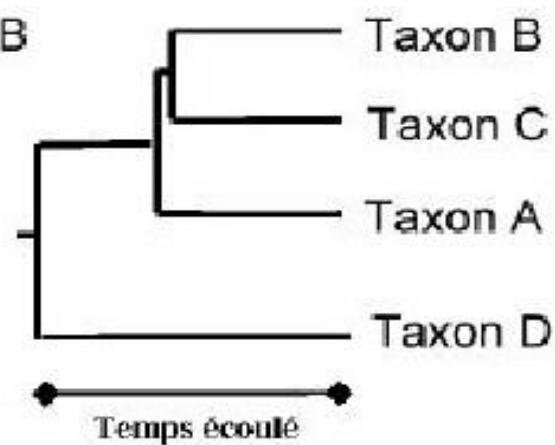
Cladogramme



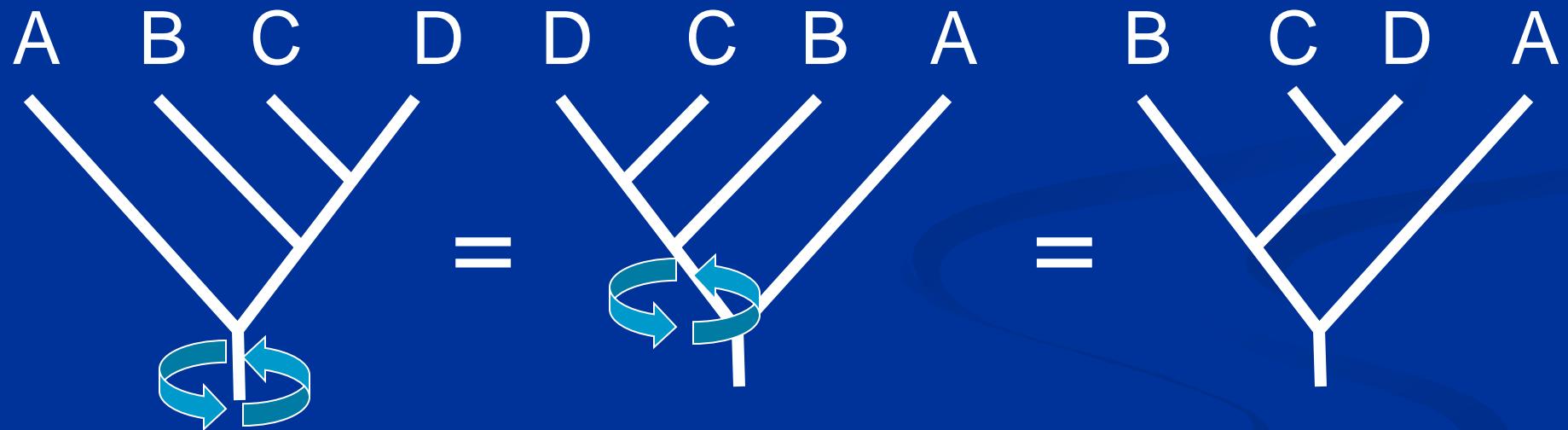
Phylogramme



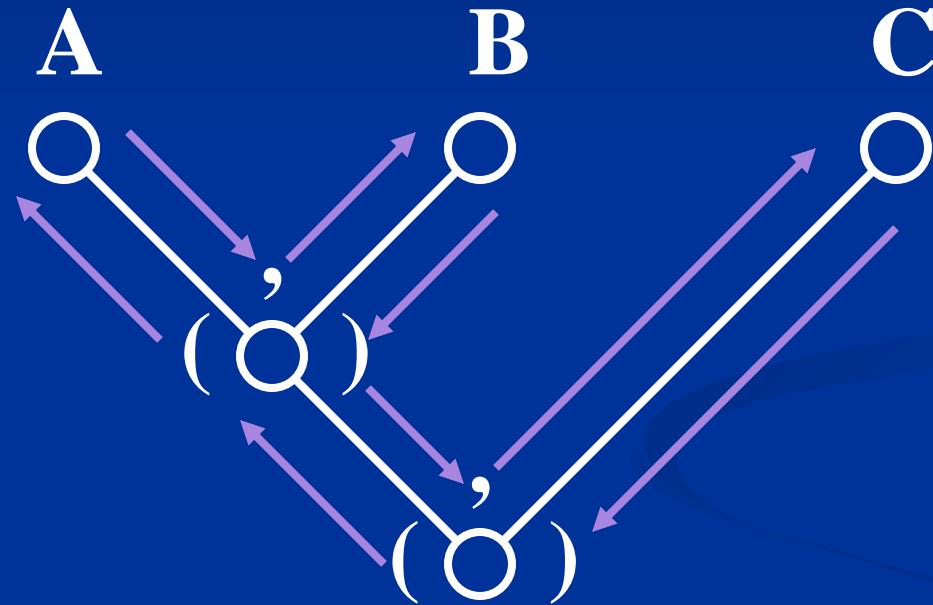
Arbre Ultramétrique



L'ordre des taxons n'est pas important



Description d'un arbre



$((A,B),C)$

Newick format

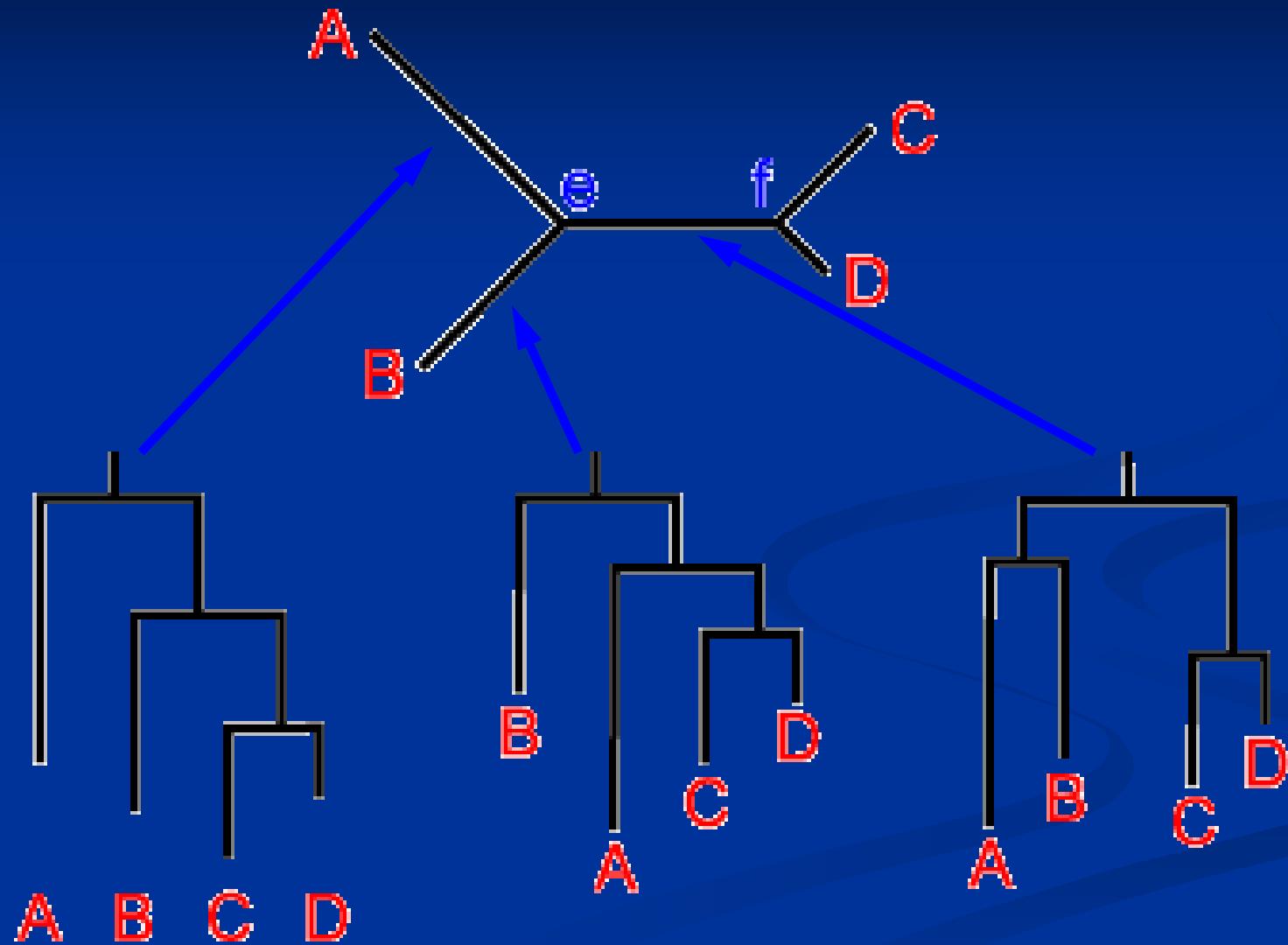
Racine des arbres

La plupart des méthodes phylogénétiques produisent des arbres non enracinés. Détection de différences entre séquences mais pas de moyens d'orienter ces changements relativement au temps

2 moyens d'enraciner un arbre :

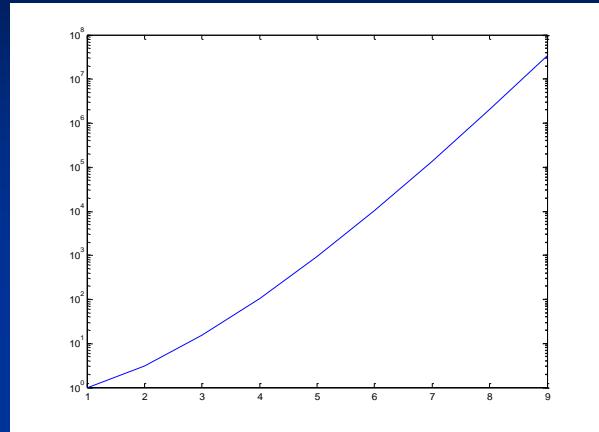
1. Utiliser un outgroup : une séquence connue à priori comme étant extérieure au groupe étudier
2. Faire l'hypothèse de l'horloge moléculaire. Tous les groupes ont évolués à la même vitesse. La racine est à équidistance de tous les feuilles de l'arbre (Mid Point Rooting)

Enraciner un arbre



Nombre d'arbres possibles

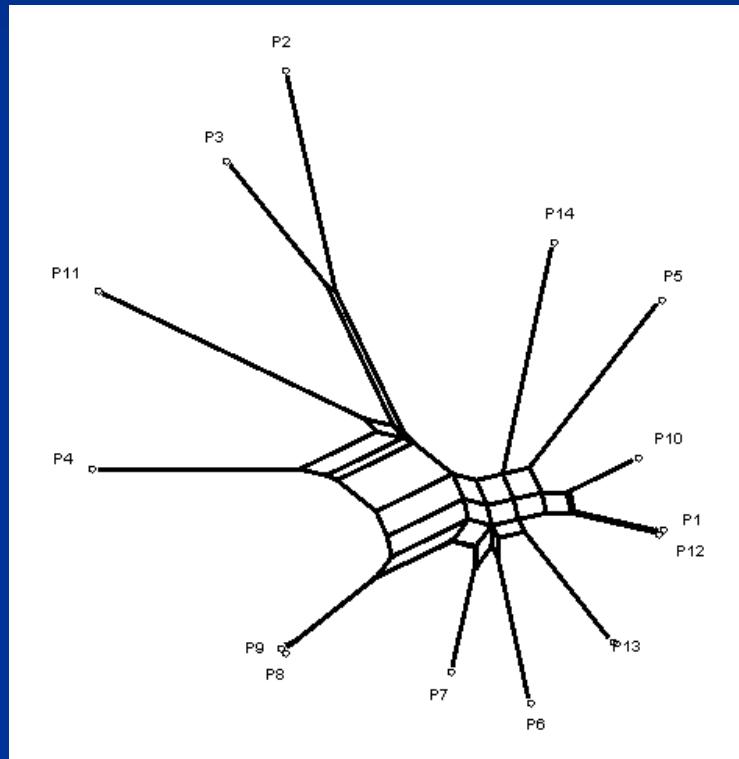
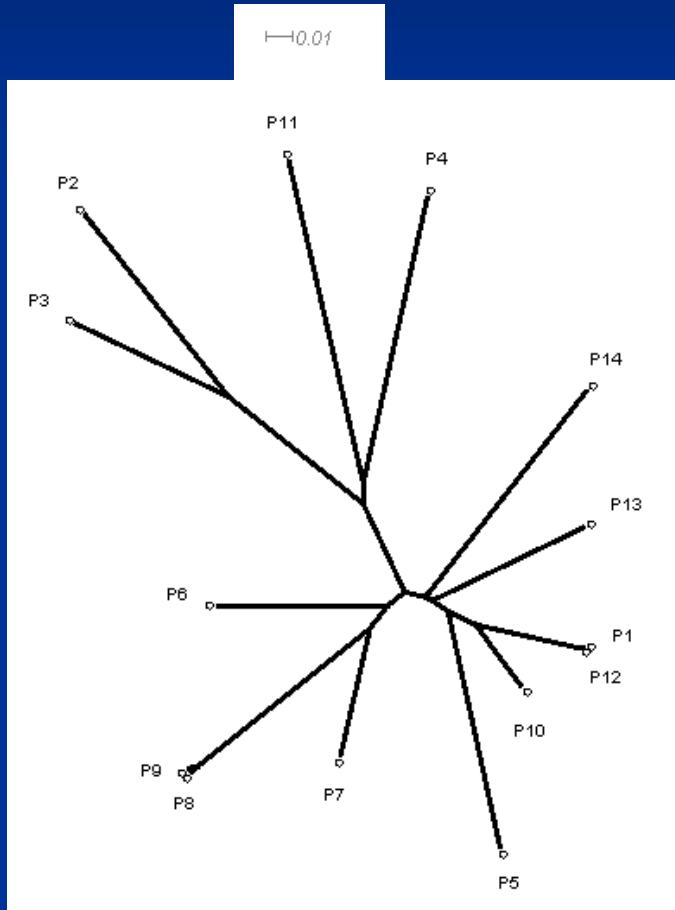
Log # trees



OTU's

Nb de taxons	Enraciné	Non enraciné
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	95 103	945
8	135 135	95 103
9	5 702 202	135 135
10	25 594 344	5 702 202

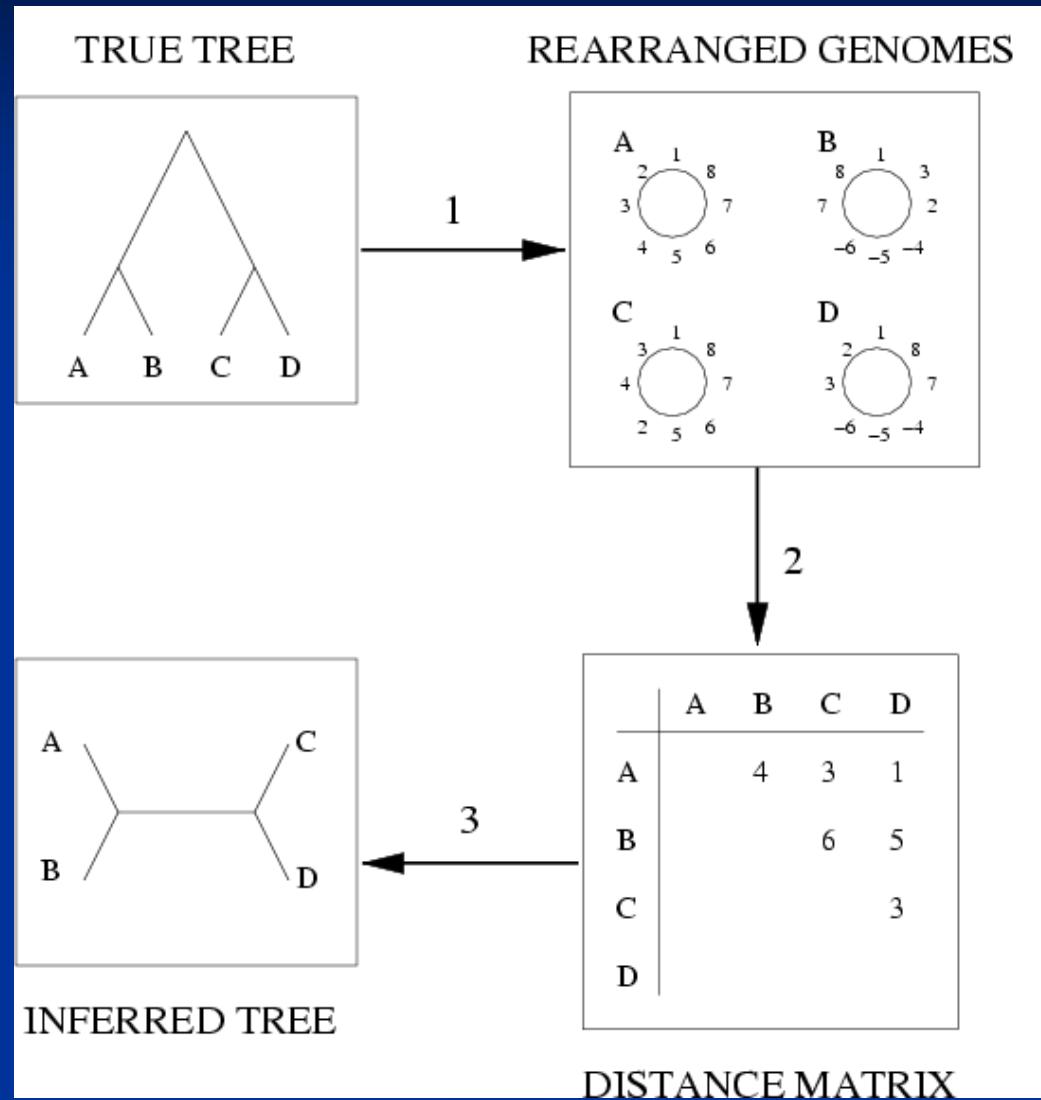
Représentation de plusieurs phylogénie en même temps : network



	Tree Evaluation	Clustering
Character State	Maximum Parsimony Maximum Likelihood Bayesian Inference	
Distance Matrix	Fitch-Margoliash	Neighbour-Joining UPGMA

Neighbour joining

Distance-based methods



Méthode du Neighbor-Joining

Très utilisé

Relativement rapide

Efficace lorsque la divergence entre séquence est faible

La première étape est la conversion des séquences en matrice de distance représentant la distance évolutive entre séquences

Méthode du Neighbor-Joining

Etape 1 : Créer une matrice de distance entre les séquences

Etape 2 : Regrouper les deux taxons les plus proches et le considérer comme un seul

Etape 3 : Recalculer une matrice de distance et recommencer l'opération jusqu'à ce qu'il ne reste plus que 2 taxons

Propriétés du Neighbor-Joining

Méthode rapide, même pour un centaine de séquences

Produit toujours des arbres non enracinés

Détermine le bon arbre si les données correspondent à
un arbre et si elles sont bien estimées

Parcimonie

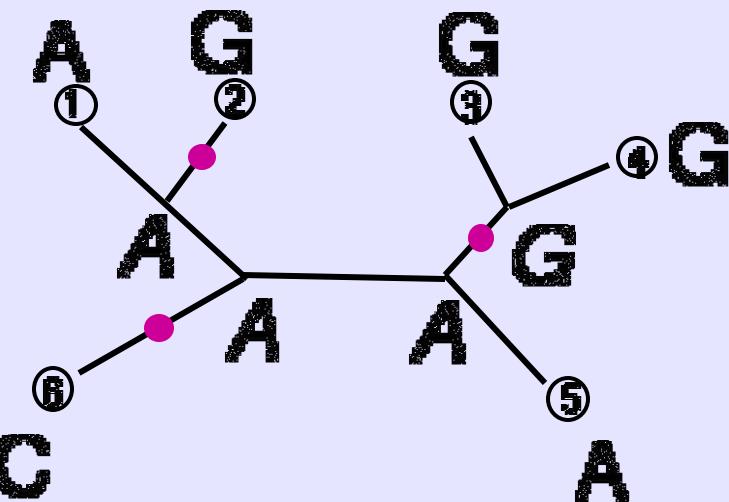
Parcimonie

Les arbres possibles sont évalués par le nombre de changements évolutifs (mutations) produisant les données (score).

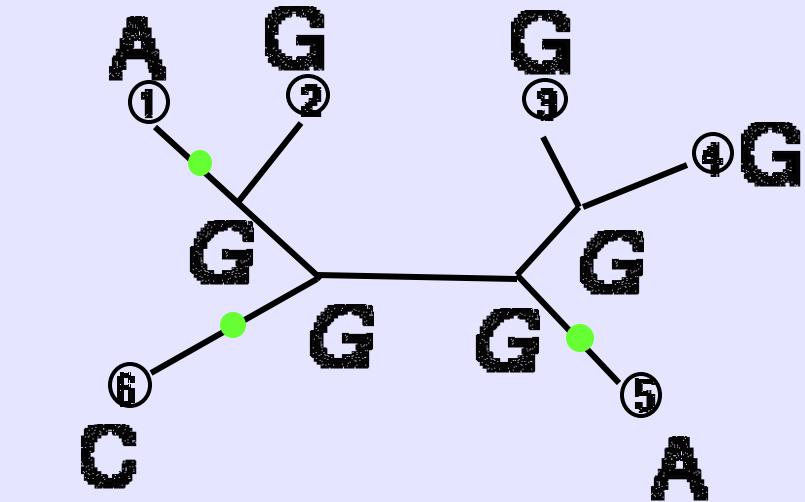
Le **maximum de parcimonie**, est atteint quand les données peuvent être produites avec le minimum de mutations

Etape 1 : pour une topologie d'arbre donnée et pour un site de l'alignement, déterminer quels sont les résidus ancestraux requièrent le nombre le plus faible de changement pour l'ensemble de l'arbre (d)

Exemple:



X : ancestral nucleotide



● : substitution event

*Pour ce site et pour une topologie d'arbre donnée, il faut au moins 3 substitutions pour expliquer la structure de l'arbre.
Plusieurs scénarios sont ici possible.*

Etape 2 : Calculer d pour chaque site de l'alignement.
Ajouter d pour tout l'alignement

→ On obtient L qui est égal à la longueur totale de l'arbre

Etape 3 : Calculer L pour toutes les topologies possibles d'arbre. Ne retenir que l'arbre le plus court

→ Arbre requérant le moins de changement (le plus parcimonieux)

Quelques propriétés de la parcimonie

Plusieurs arbres peuvent avoir le même score de parcimonie

Le nombre d'arbres augmente très rapidement avec le nombre de séquences comparées.

- La recherche de l'arbre le plus parcimonieux doit être réduite à une fraction de tous les arbres possibles (recherche heuristique). Il n'y a pas d'assurance de trouver le meilleur arbre

Maximum likelihood

Maximum likelihood (ML)

En ML, une hypothèse est jugée sur comment elle prédit les données observées. L'arbre qui a la plus grande probabilité de produire les données observées est préféré (Maximum de vraisemblance)

Pour utiliser cette approche, nous devons être capable de calculer la probabilité de données suivant une phylogénie

Il faut un modèle d'évolution de séquence qui décrit la probabilité relative de divers événements

Tous les “chemins” pour obtenir des données sont considérés.

Maximum likelihood

(programmes MEGA, fastDNAmI, PAUP, PROML...)*

Hypothèses

- Les substitutions suivent un modèle probabiliste pour lequel l'expression mathématique, mais pas les paramètres, sont connus
- Les sites évolues indépendamment les uns des autres
- Les probabilités de substitutions ne changent pas avec le temps sur chaque branche, mais peuvent varier d'une branche à l'autre

Maximum likelihood

Etape 1 : Considérons un arbre enraciné, un site donné et un set de longueur de branche. On peu calculer la probabilité que les nucléotides de ce site aient évolué selon cet arbre

Etape 2 : calculer la probabilité pour l'ensemble des séquences

$$P(Sq1, Sq2, Sq3, Sq4) = \prod_{\text{all sites}} P(S1, S2, S3, S4)$$

Maximum likelihood

Etape 3 : calculer les longueurs de branches et la valeur du paramètre q qui donne la valeur de $P(Sq_1, Sq_2, Sq_3, Sq_4)$ la plus forte. C'est la vraisemblance de l'arbre.

Etape 4 : calculer la vraisemblance de tous les arbres possibles.

➤ On choisit l'arbre ayant la plus grande vraisemblance.

Maximum likelihood

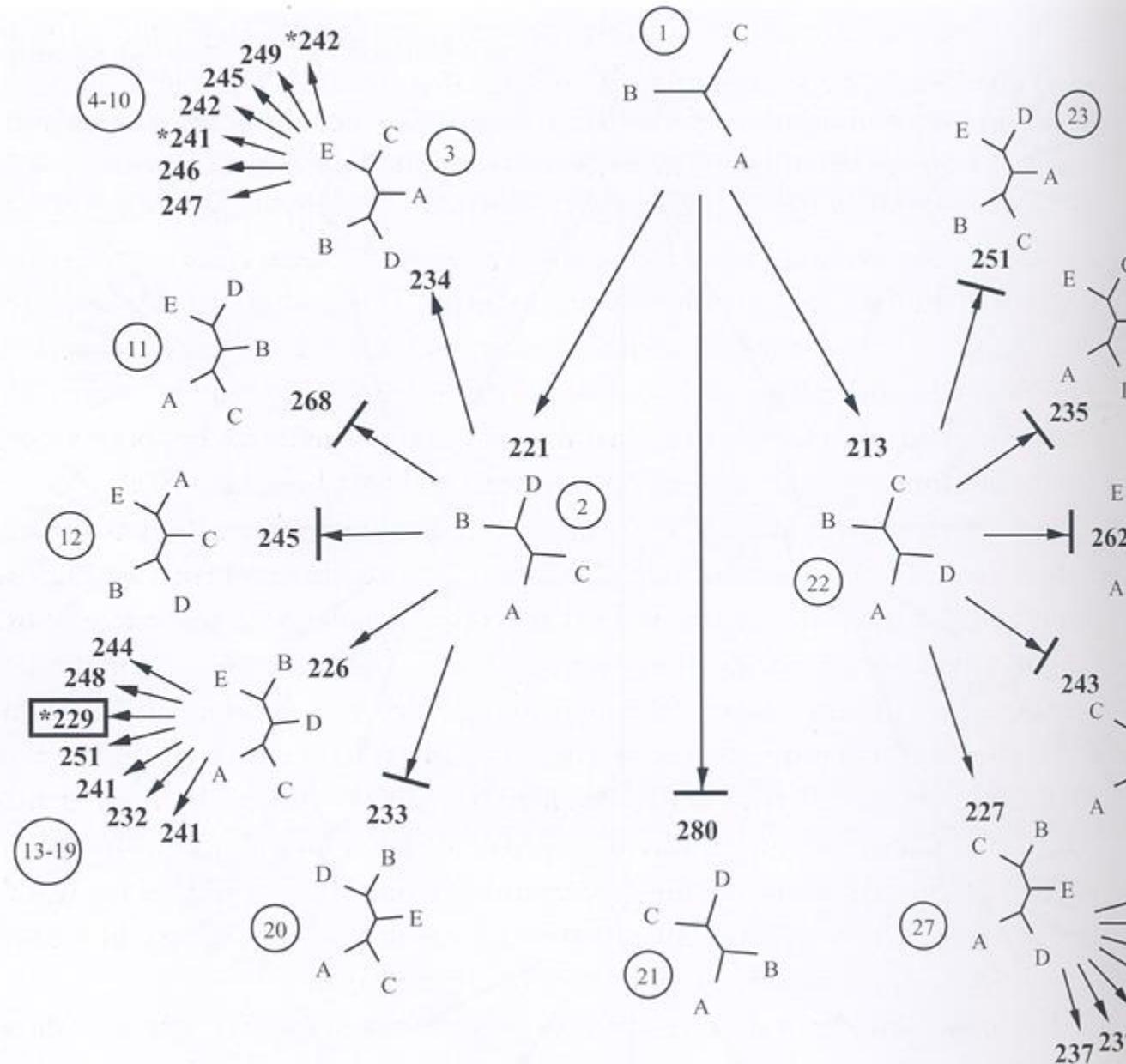
Méthodes la plus intéressante d'un point de vue théorique.

Des simulations ont montré que cette méthode marche mieux que les autres dans la majorité des cas

Méthodes demandant énormément de calcul

Il est impossible d'évaluer tous les arbres (trop nombreux). Une exploration partielle des arbres est effectuée

Méthode du Branch and bound



Analyse bayésienne

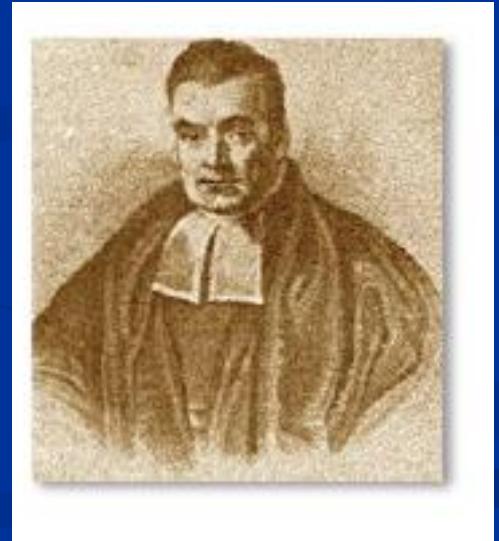
Les méthodes précédentes tentent de déterminer un seul arbre qui décrit au mieux les données

Mais elles ne cherchent pas dans tout l'espace et c'est donc difficile de trouver le meilleur arbre



En utilisant les statistiques bayesiennes, il n'y a pas de recherche du meilleur arbre

L'espace dans lequel un arbre est recherché est limité par des informations à priori (prior information)



$$P(A|B)P(B) = P(A \cap B)$$

But : déterminer la probabilité *a posteriori* suivant un alignement donné

$$\Pr(\tau | X) \propto \iint_{v,\theta} \Pr(X | \tau, v, \theta) \cdot \Pr_{prior}(v, \theta) dv d\theta$$

likelihood de l'arbre et des paramètres

Probabilité à priori des valeur du paramètre

τ : topologie de l'arbre

X: alignement de séquence

v: longueur des branches

θ : paramètres du modèle de substitution (e.g., transit/transv ratio)

Le calcul de $\Pr(t|X)$ est généralement impossible

Une méthode appelée

Metropolis-coupled Markov chain Monte Carlo (MC³) est utilisée pour générer un échantillon de la distribution des arbres a posteriori

(exemple: génération d'un échantillon aléatoire de 10,000 arbres)

Résultats :

On retient l'arbre ayant la plus grande probabilité (celui qui est trouvé le plus fréquemment dans l'échantillonnage)

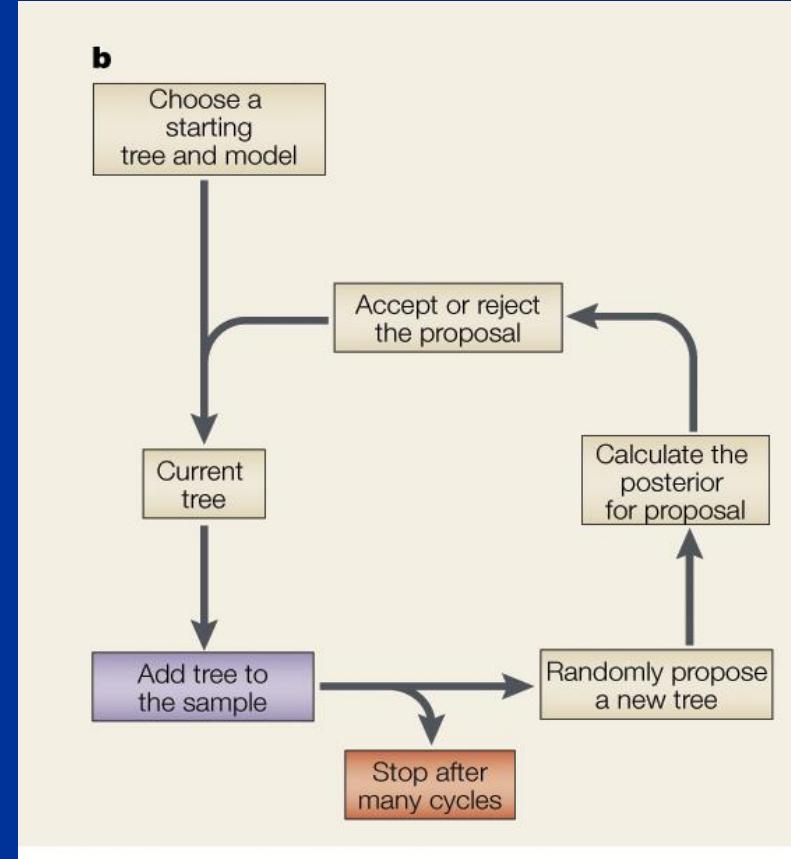
On calcul la probabilité a posteriori de chaque partition de l'arbre : fraction d'arbre contenant cette partition

Markov chain Monte Carlo

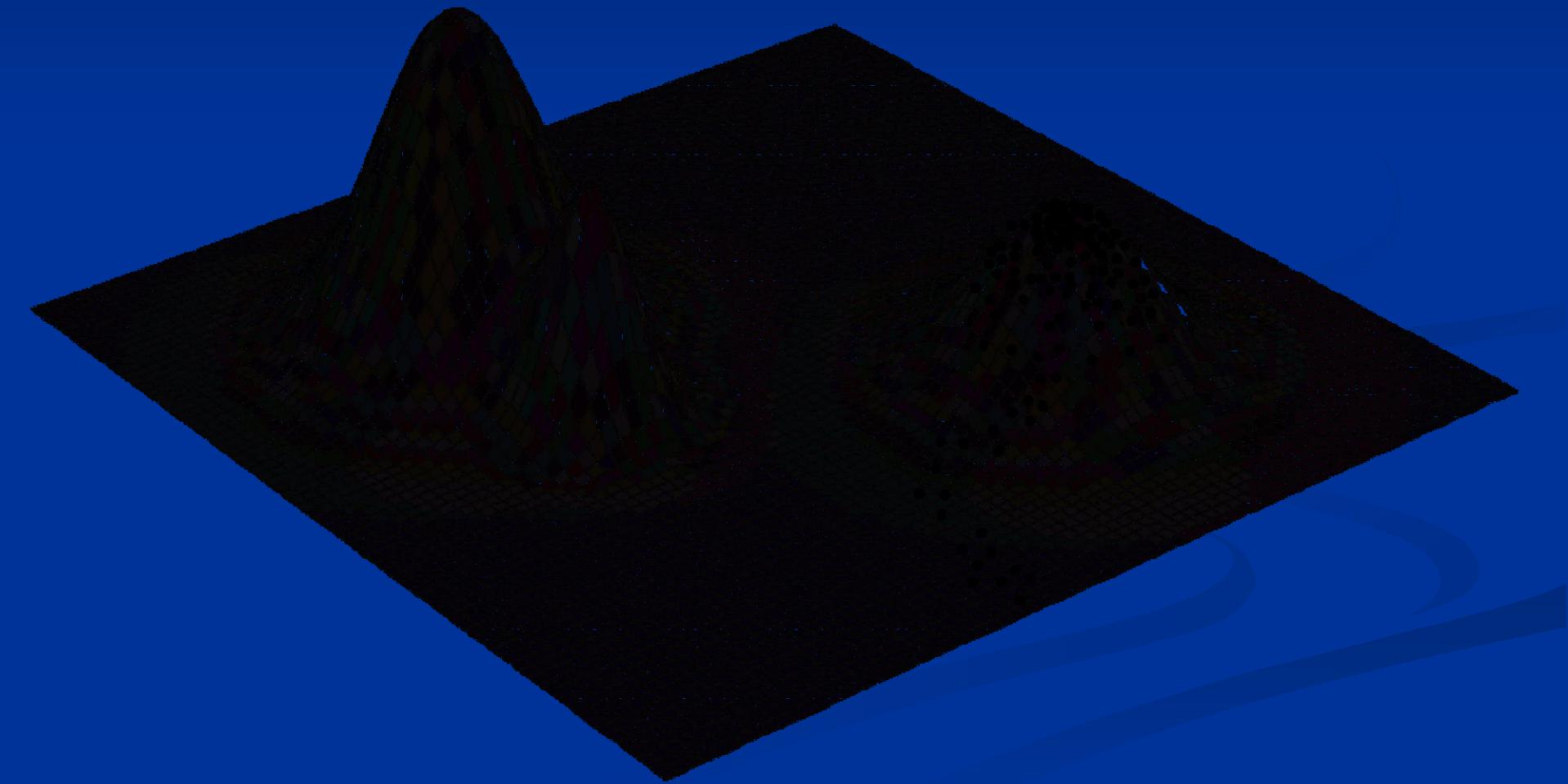
La méthode du Markov chain Monte Carlo (MCMC) est similaire aux algorithmes de recherche d'arbre

A partir d'un arbre initial, un nouvel arbre est proposé de manière aléatoire.

L'algorithme MCMC spécifie aussi les règles d'acceptation et de rejet d'un arbre

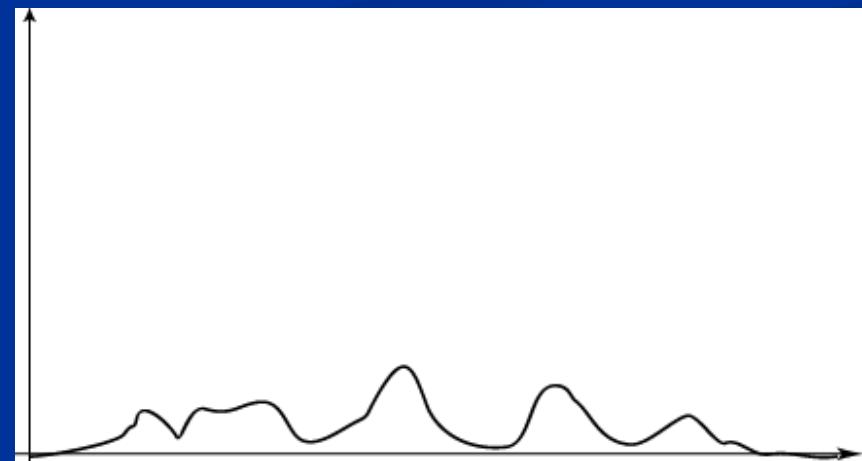
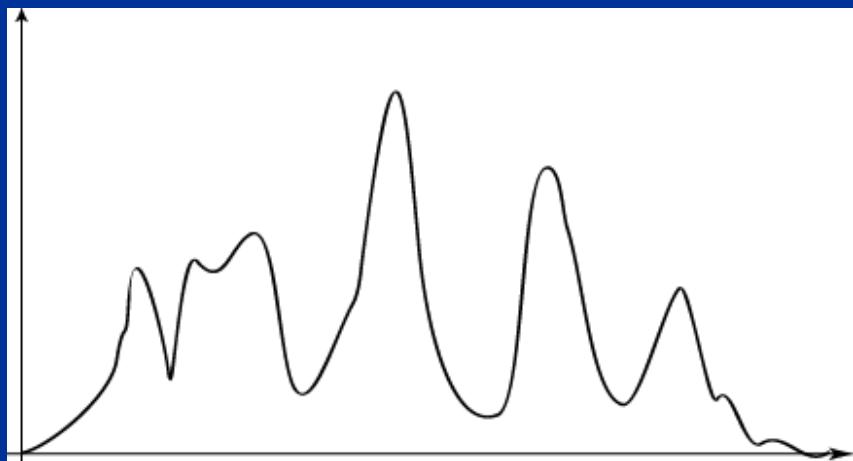


Markov chain Monte Carlo



Markov chain Monte Carlo

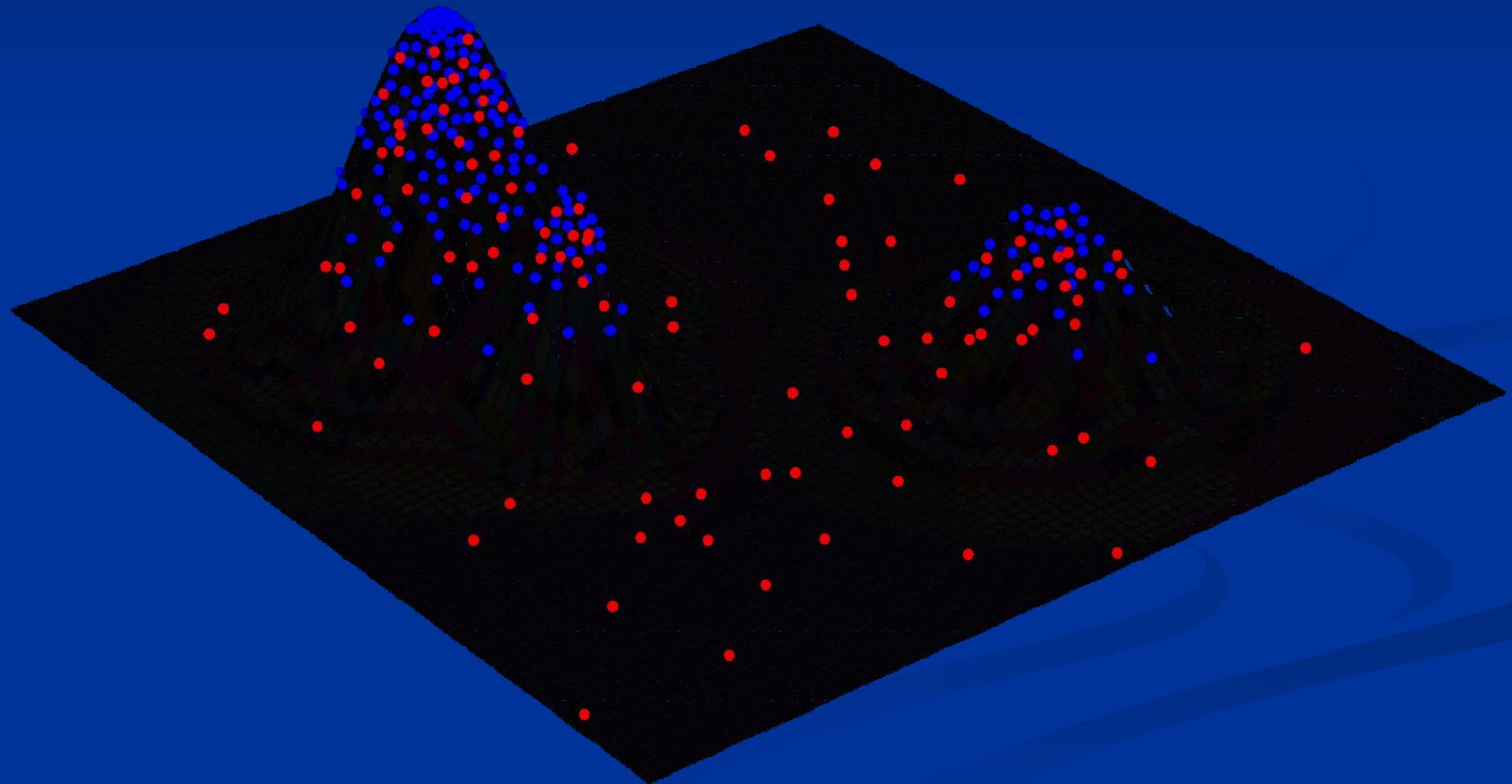
Metropolis Coupled Markov Chain Monte Carlo



Markov chain Monte Carlo



Markov chain Monte Carlo



Les approches Bayesienne permettent l'utilisation de modèle d'évolution complexe

Estimation du temps de divergence

Recherche de résidus important pour la sélection naturelle

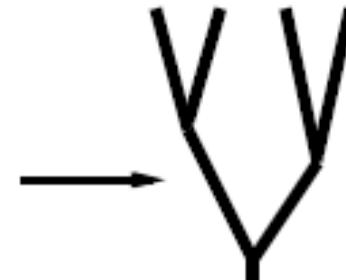
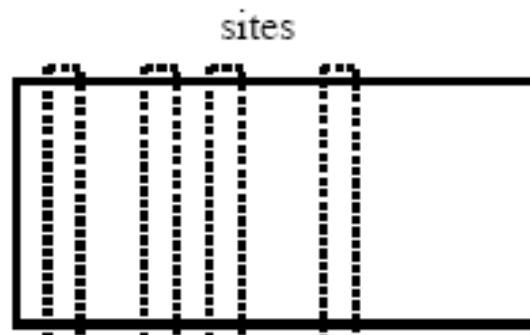
Détection de point de recombinaison

...

Bootstrap

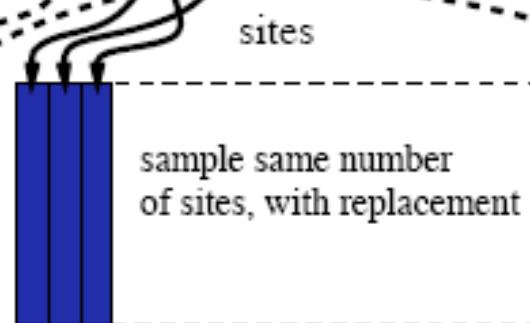
Original
Data

sequences



Bootstrap
sample
#1

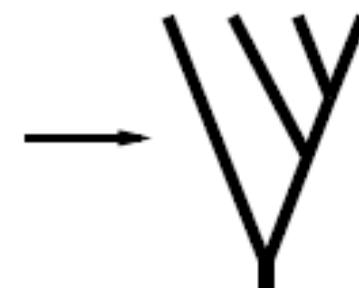
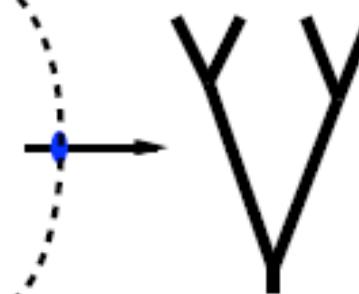
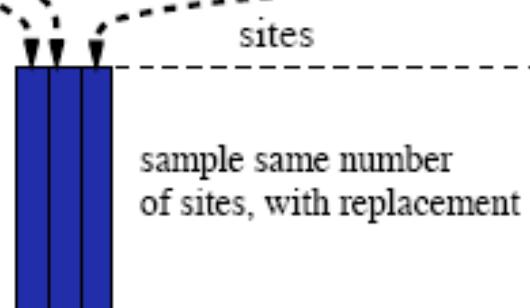
sequences



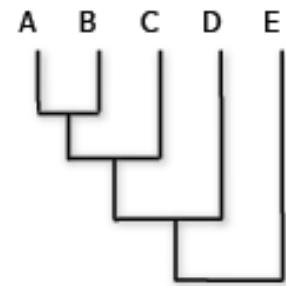
Estimate of the tree

Bootstrap
sample
#2

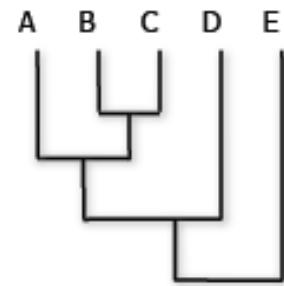
sequences



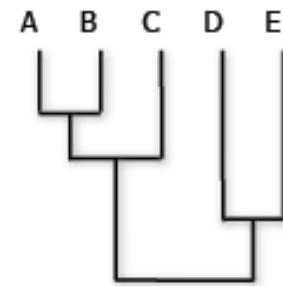
(and so on)



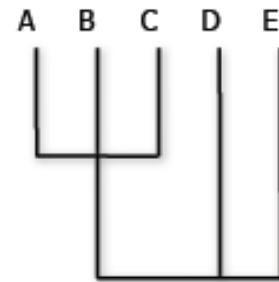
Tree 1



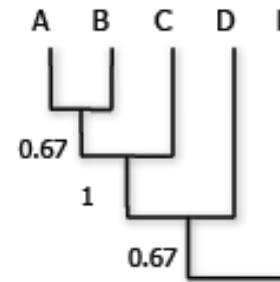
Tree 2



Tree 3



Strict
consensus tree



Majority-rule
consensus tree

Bootstrap

Les branches internes supportées par plus de 70 % (parfois 90 %) des arbres sont considérées comme statistiquement significatives

Cette procédure ne permet pas de dire si la méthode de reconstruction phylogénétique était bonne ou non. Un mauvais arbre peu avoir ces branches avec de fortes valeurs de bootstrap

Conclusion

Comparaison des méthodes

Table 1 | **Comparison of methods**

Method	Advantages	Disadvantages	Software
Neighbour joining	Fast	Information is lost in compressing sequences into distances; reliable estimates of pairwise distances can be hard to obtain for divergent sequences	PAUP* MEGA PHYLIP
Parsimony	Fast enough for the analysis of hundreds of sequences; robust if branches are short (closely related sequences or dense sampling)	Can perform poorly if there is substantial variation in branch lengths	PAUP* NONA MEGA PHYLIP
Minimum evolution	Uses models to correct for unseen changes	Distance corrections can break down when distances are large	PAUP* MEGA PHYLIP
Maximum likelihood	The likelihood fully captures what the data tell us about the phylogeny under a given model	Can be prohibitively slow (depending on the thoroughness of the search and access to computational resources)	PAUP* PAML PHYLIP
Bayesian	Has a strong connection to the maximum likelihood method; might be a faster way to assess support for trees than maximum likelihood bootstrapping	The prior distributions for parameters must be specified; it can be difficult to determine whether the Markov chain Monte Carlo (MCMC) approximation has run for long enough	MrBayes BAMBE

For a more complete list of software implementations, see online link to Phylogeny Programs.

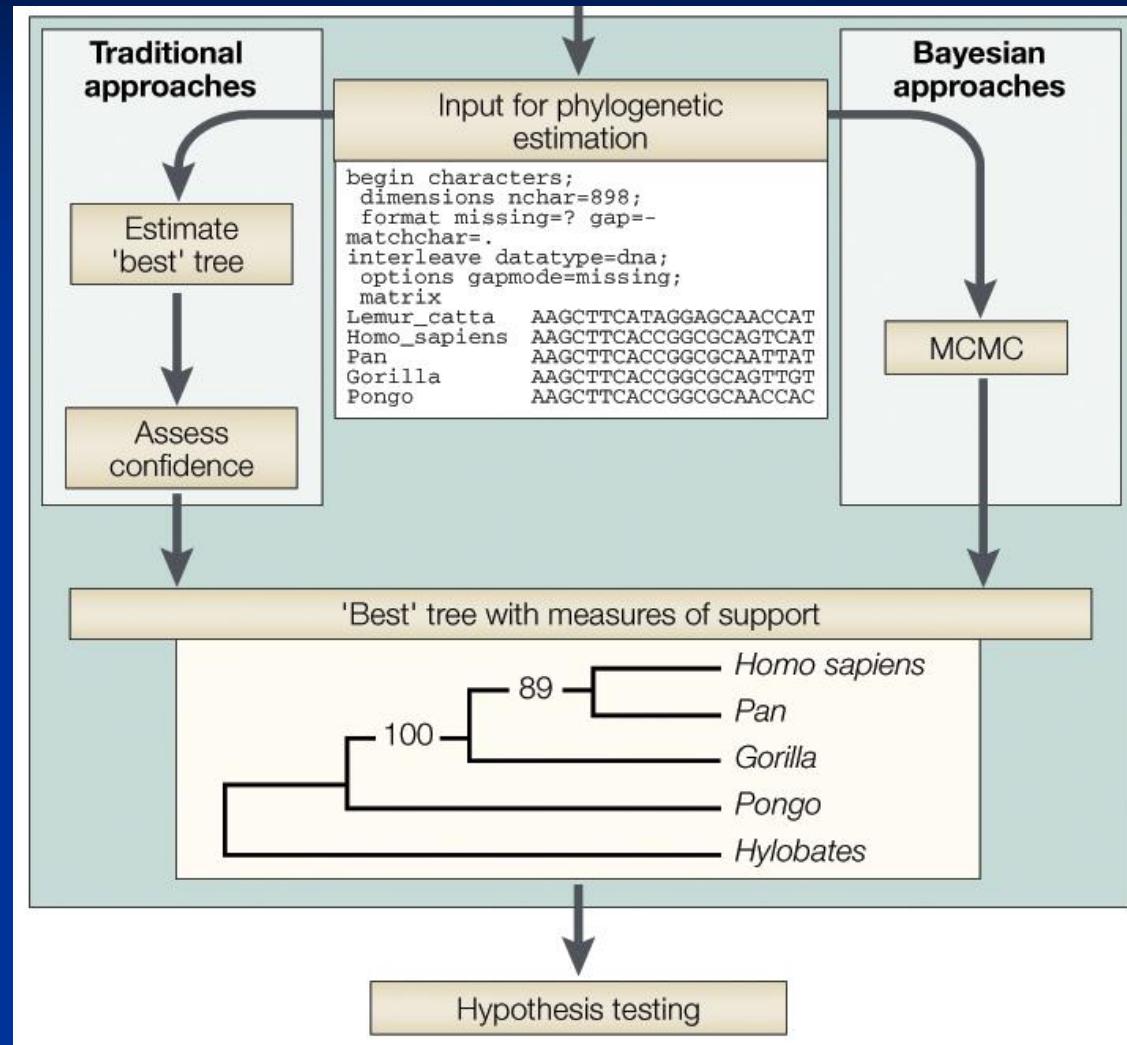
Bayesian vs ML

Maximum likelihood

Recherche de l'arbre maximisant les chances de voir les données par rapport à l'arbre ($P(\text{Data} | \text{Tree})$)

Bayesian inference

Recherche de l'arbre maximisant les chances de voir l'arbre par rapport aux données ($P(\text{Tree} | \text{Data})$)



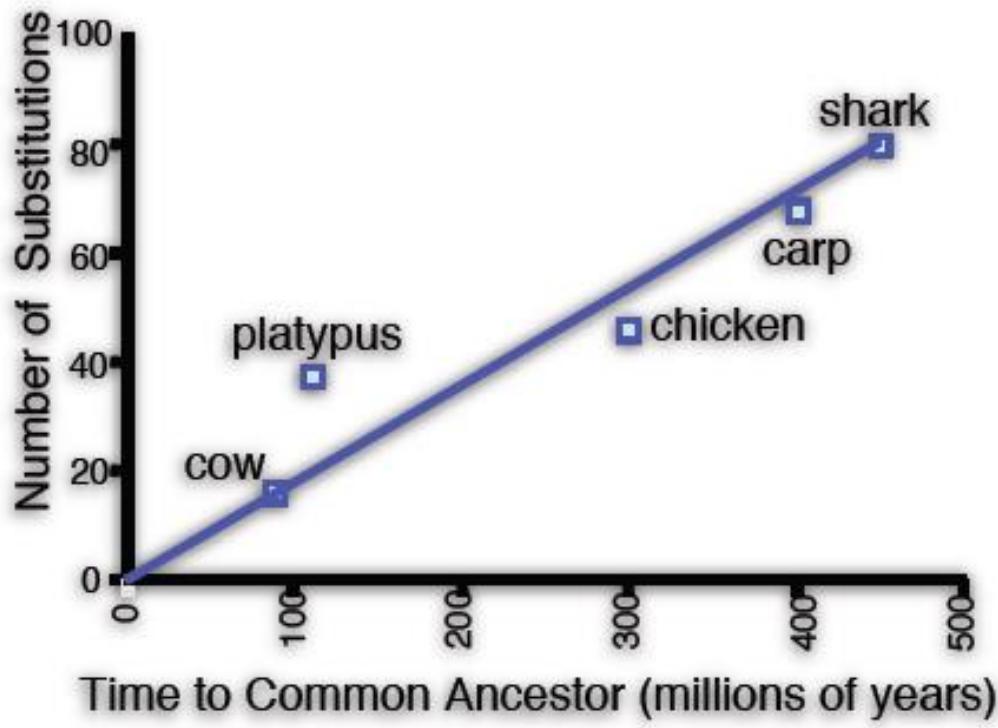
Temps d'analyse

TABLE 1. Average run times for various methods. The computing times were measured on a 1.8-GHz (1 Go RAM) PC with Linux. For PHYML, the number in parentheses is the average number of refinement stages.

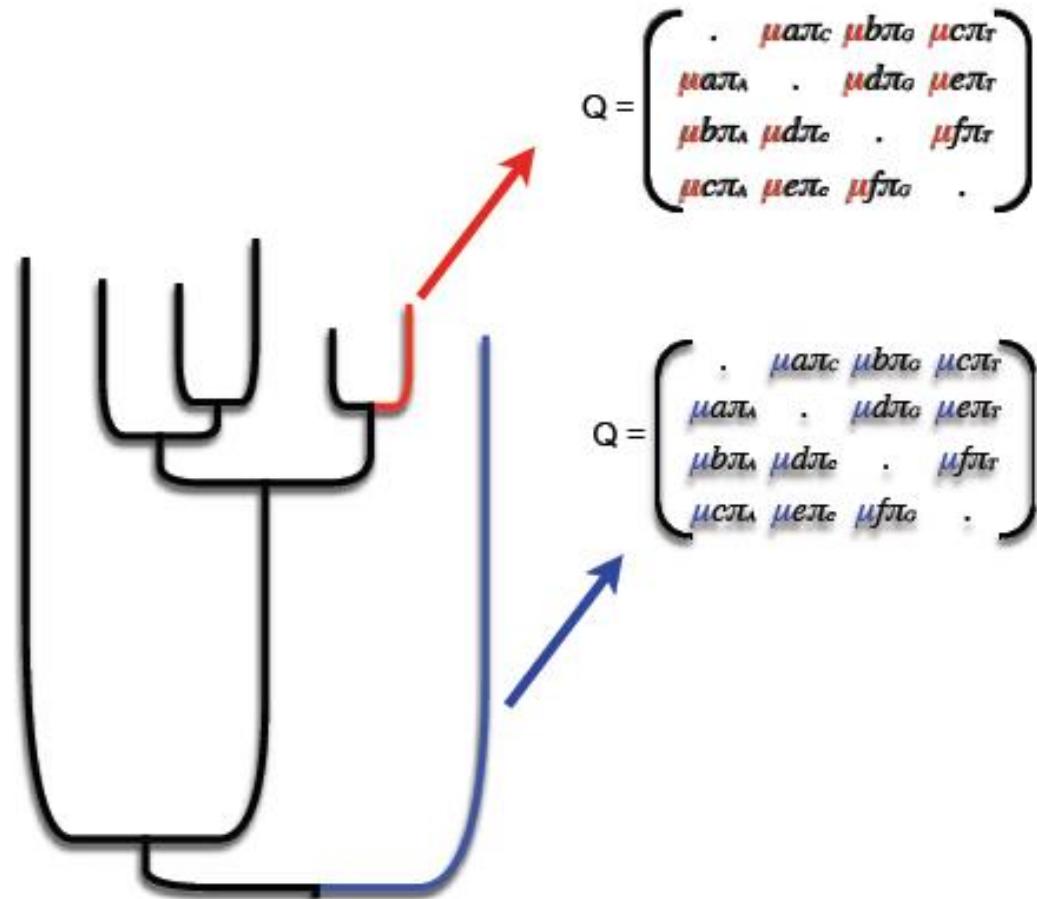
Method	Simulations		Real data	
	40 taxa (500 bp)	100 taxa (500 bp)	218 taxa (4,182 bp)	500 taxa (1,428 bp)
DNADIST+ NJ/BIONJ	0.3 sec	2.3 sec	50 sec	2 min, 19 sec
DNADIST+ Weighbor	1.5 sec	22 sec	4 min, 52 sec	58 min, 40 sec
DNAPARS	0.5 sec	6 sec	4 min, 4 sec	13 min, 12 sec
PAUP*	3 min, 21 sec	1 hr, 4 min		
PAUP* + NJ	1 min, 10 sec	22 min	10 hr, 50 min	
MrBayes	2 min, 6 sec	32 min, 37 sec		
fastDNAml	1 min, 13 sec	26 min, 31 sec		
NJML	15 sec	6 min, 4 sec		
MetaPIGA	21 sec	3 min, 27 sec	4 hr, 45 min	9 hr, 4 min
MetaPIGA+ NJ	6 sec	23 sec	1 hr, 40 min	3 hr
PHYML	2.7 sec (6.4)	12 sec (8.3)	8 min, 13 sec (15)	11 min, 59 sec (13)

Horloge moléculaire

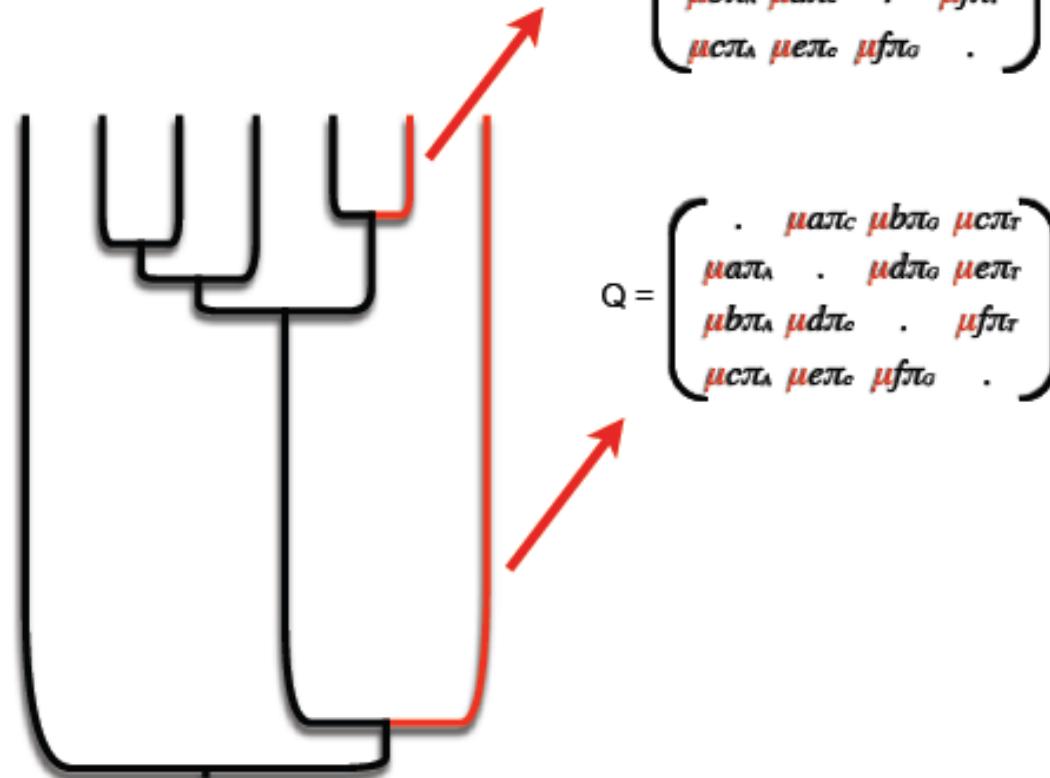
Molecular Clock for α -globin

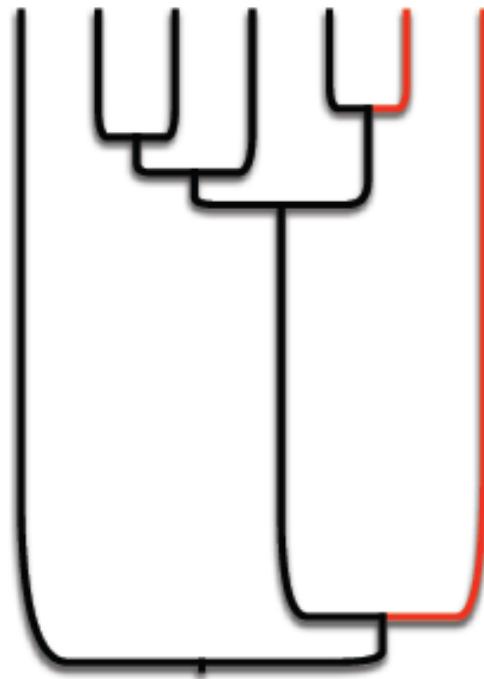
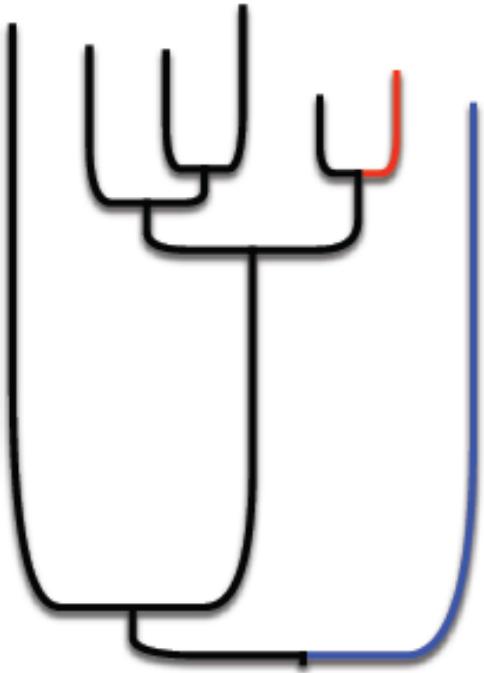


Unrooted model, or ‘non-clock model’



Strict molecular clock model

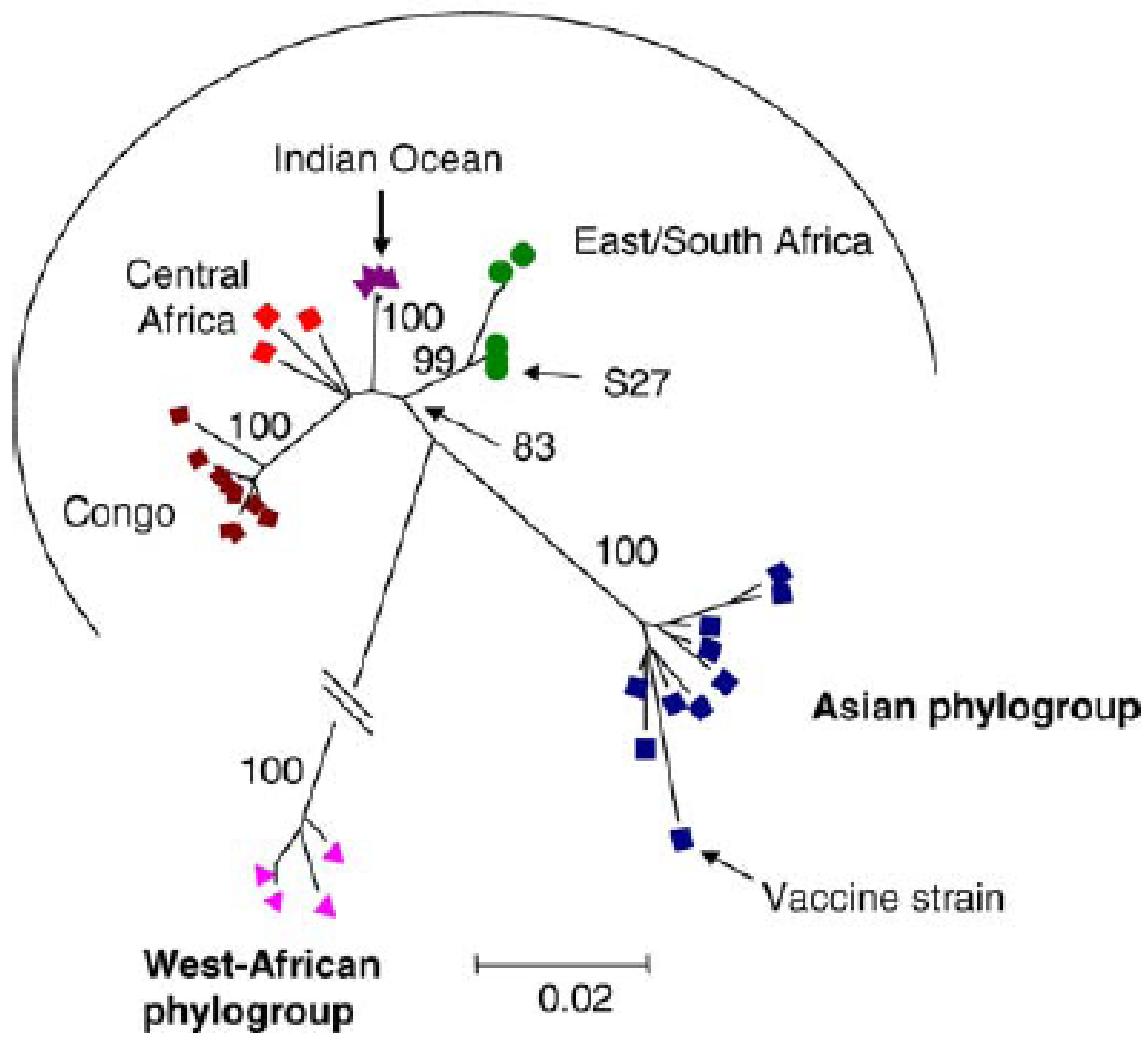




$$2(\ln L_c - \ln L_{nc}) \sim \chi^2_{2n-2}$$

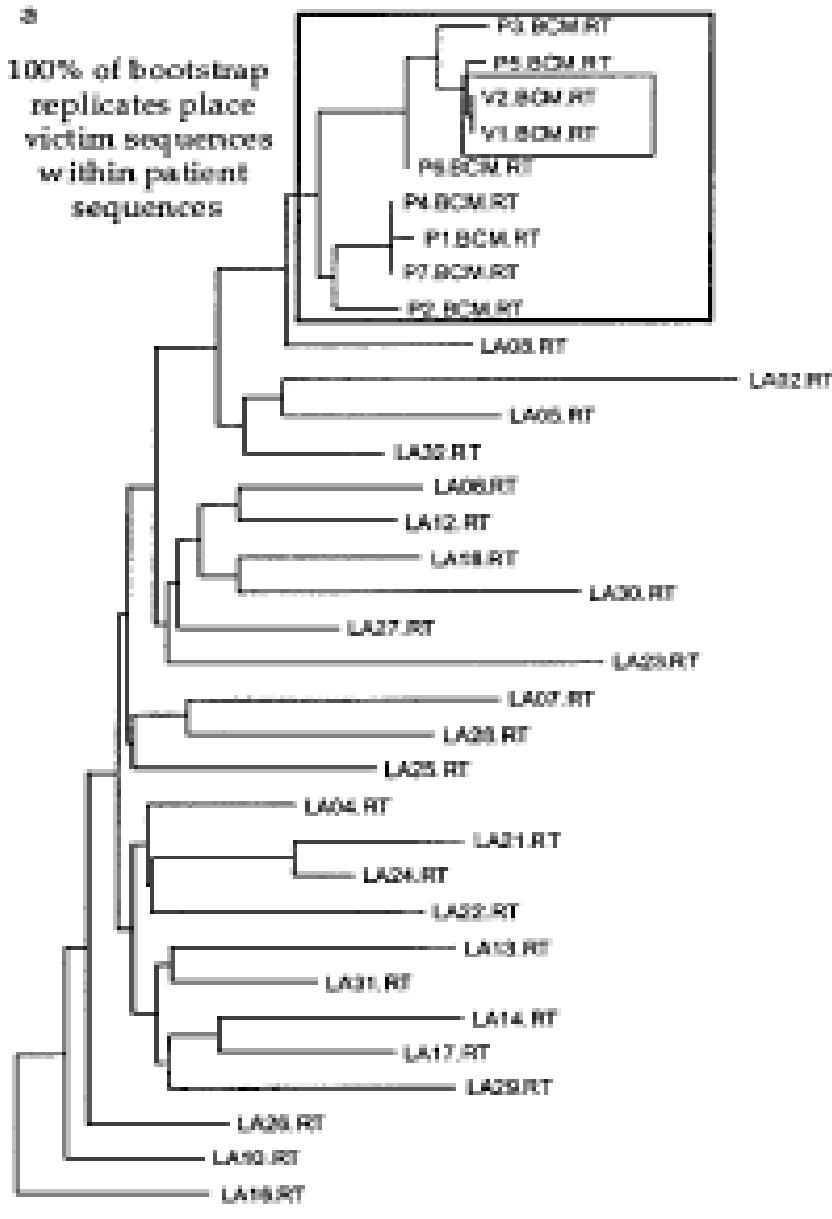
Exemples

Relation entre les isolats de Chikungunya de différentes origines



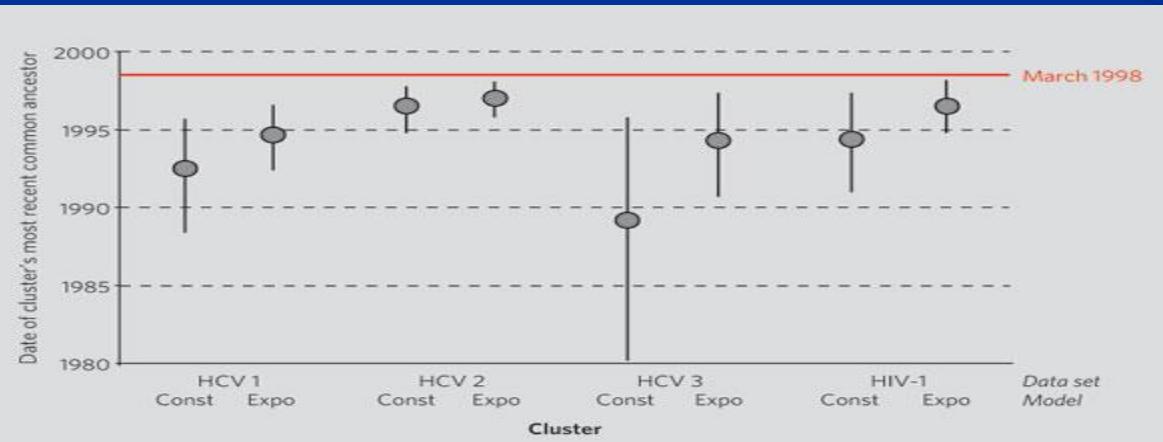
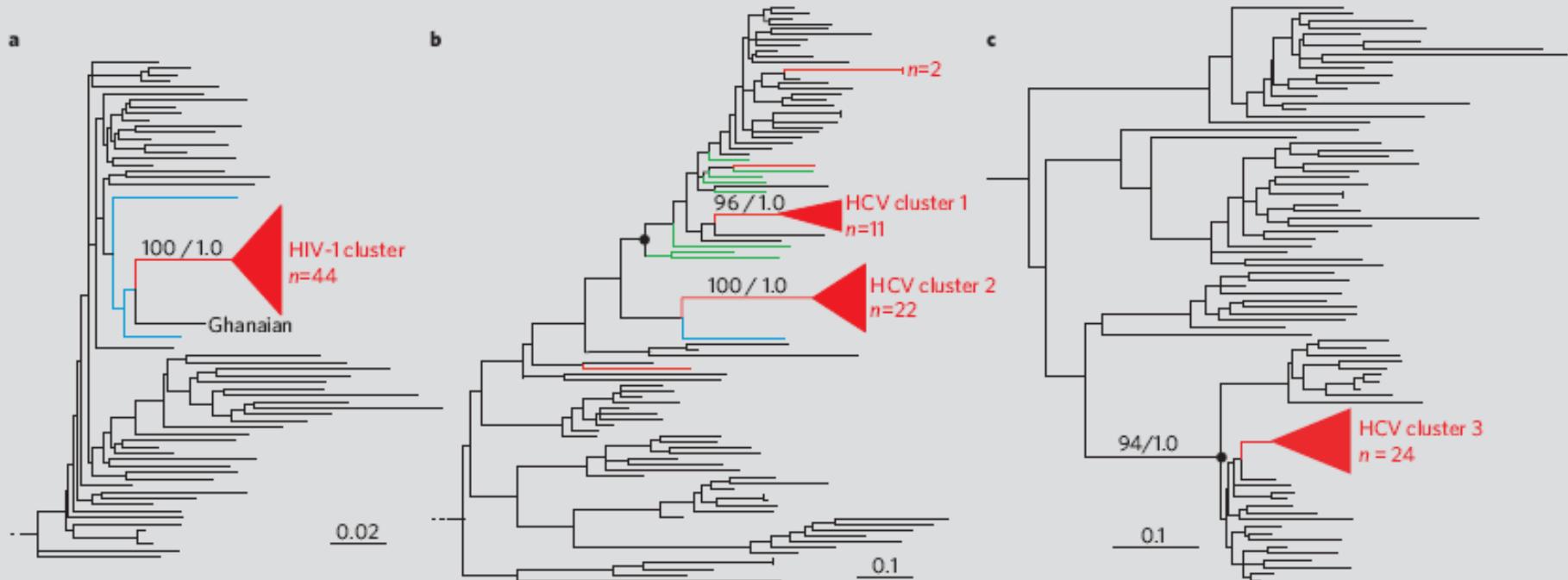
8

100% of bootstrap replicates place victim sequences within patient sequences



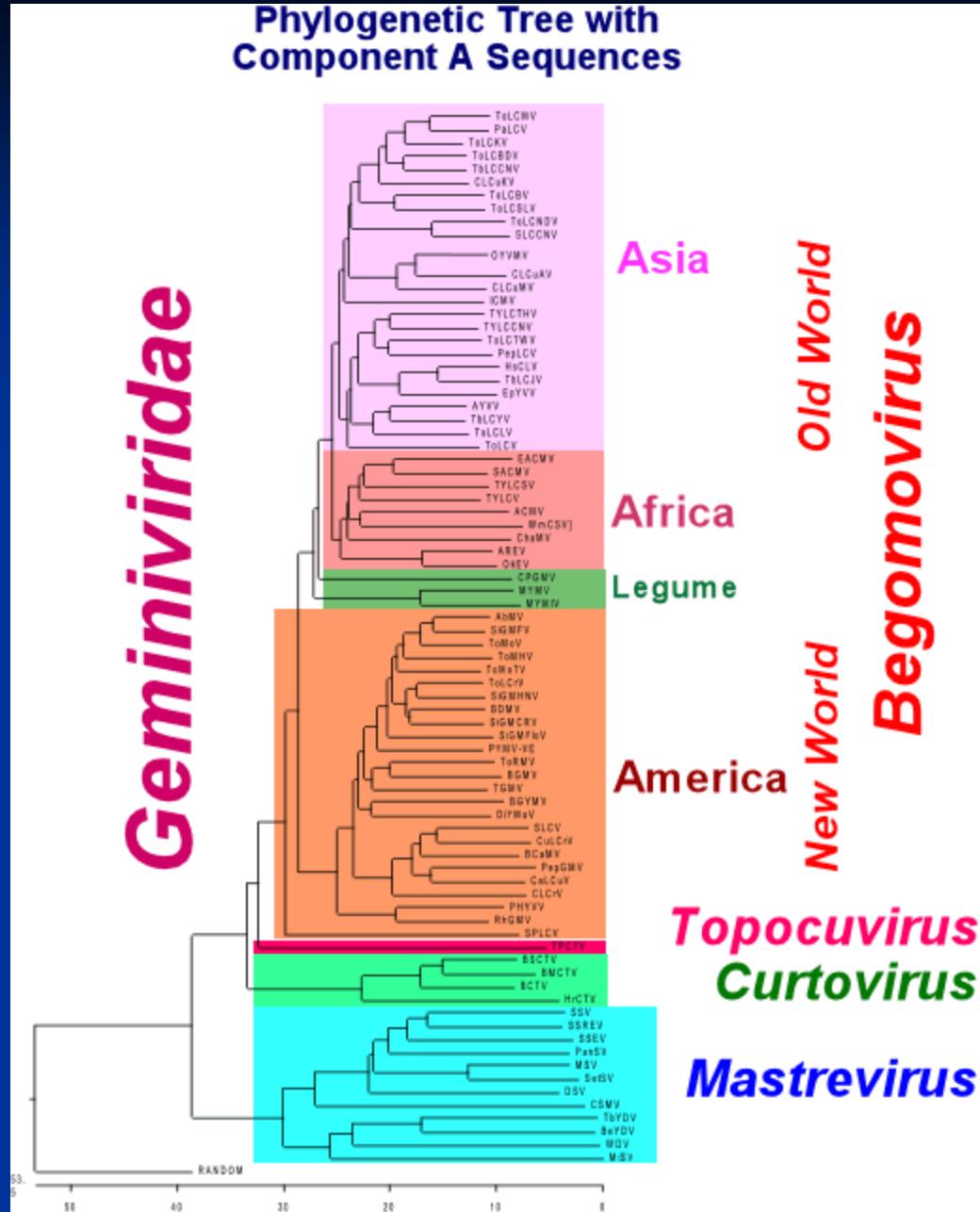
Preuve de la transmission
par un médecin du virus du
SIDA de ses patients à son
ex-femme...

Epidémie de HIV-1 et HCV en Libye : 5 infirmières bulgares et un médecin Palestinien accusé de transmission volontaire du virus



De Oliveira et al,
Nature 2006

Classification des Begomovirus

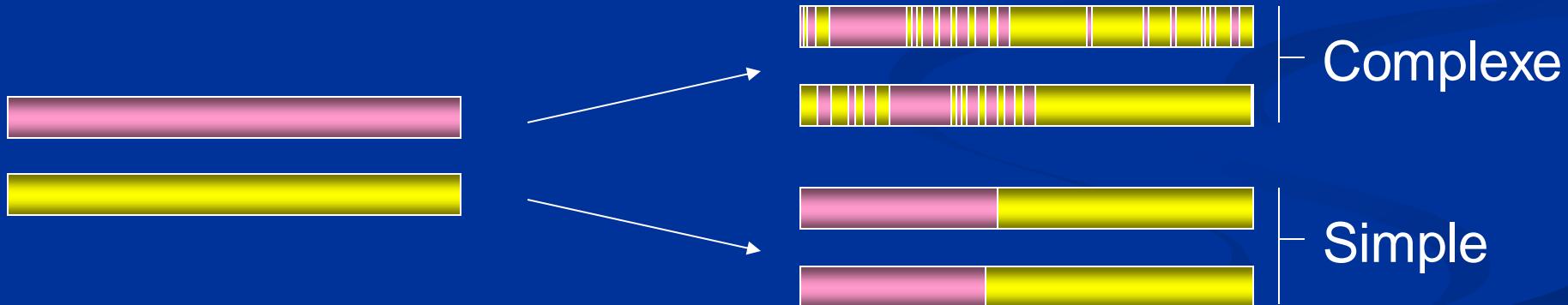


Recombinaison

Homologue ou non homologue

Homologue : séquences proches (même gène...)

Non Homologue : séquences différentes



Presque tout les génomes recombinent
Pourquoi est ce important ?

C'est important pour la phylogénie

La recombinaison permet à différentes régions génomiques d'avoir des histoires évolutives différentes – i.e. un arbre phylogénétique ne peut décrire à lui seul l'histoire de séquence recombinante

(solution : network...)

→ Complique la phylogéographie, l'horloge moléculaire...

Causé par une réparation de rupture d'ADN double brin ?

Saut de répliqueuse ?

Un moyen efficace d'aquérir les mutations bénéfiques :

La recombinaison permet un taux de mutation plus important et une exploration plus aggressive de la diversité.

La recombinaison permet des sauts de pics de fitness

Estimation du taux de recombinaison de la population :
Lamarck, LDHat, Sites...

Détection de la recombinaison au seins d'un alignement
de séquences :
Simplot, VisRD, RDP...

Déetecter au sein d'un alignement de séquences la présence de recombinaison

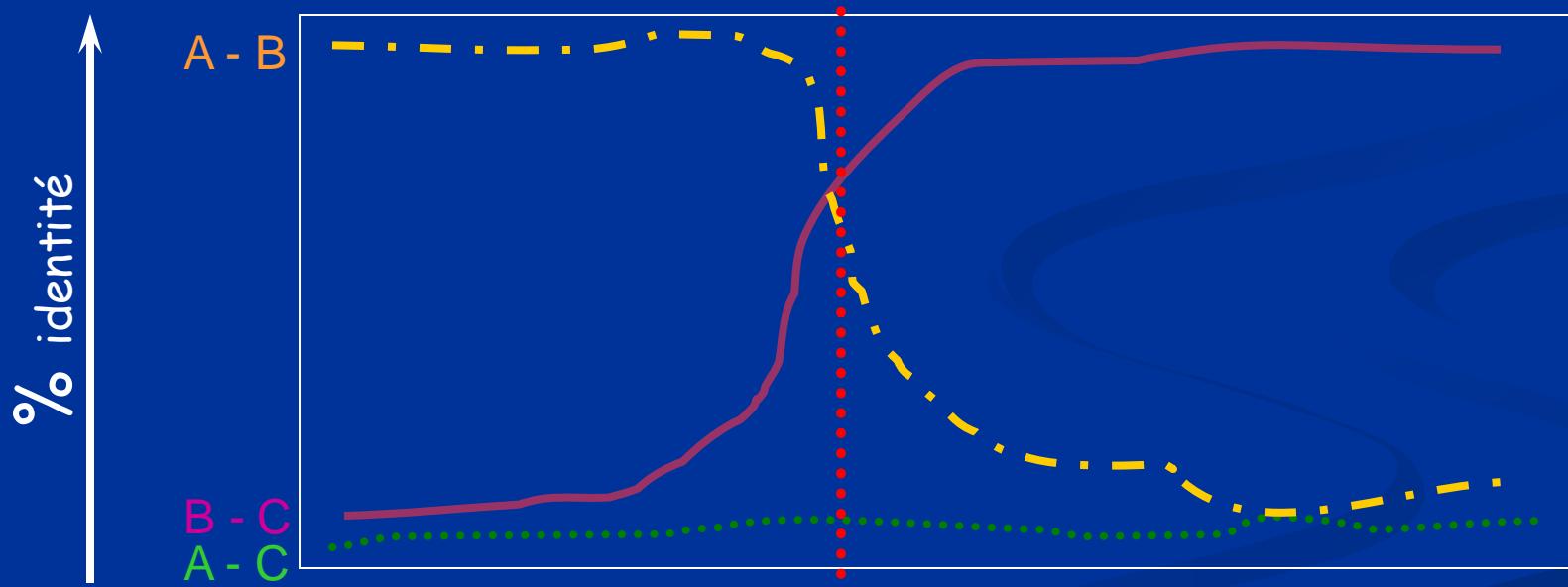
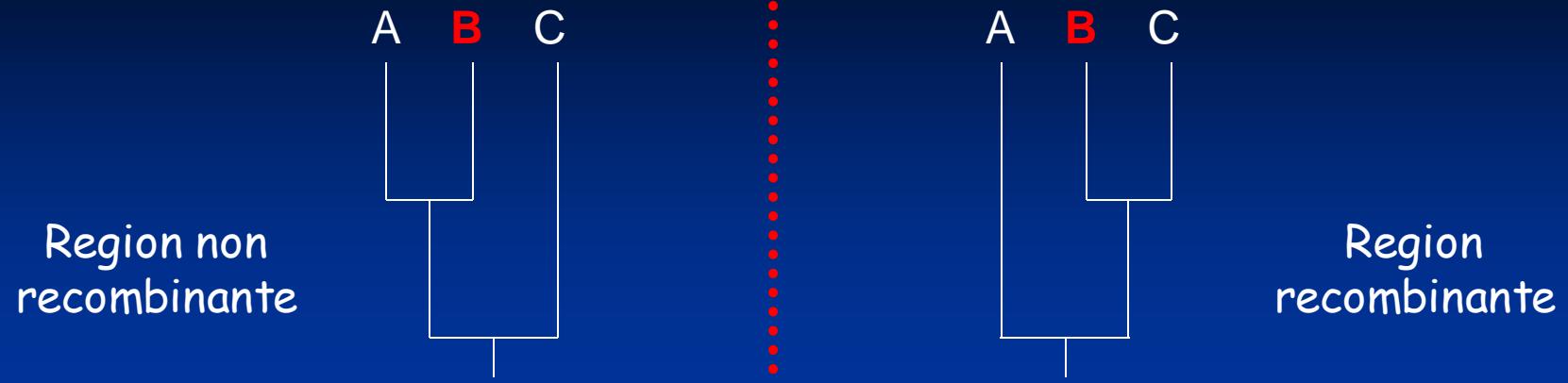
- Recombinaison « vraie »
- Pseudorecombinaison

Conditions d'emploi :

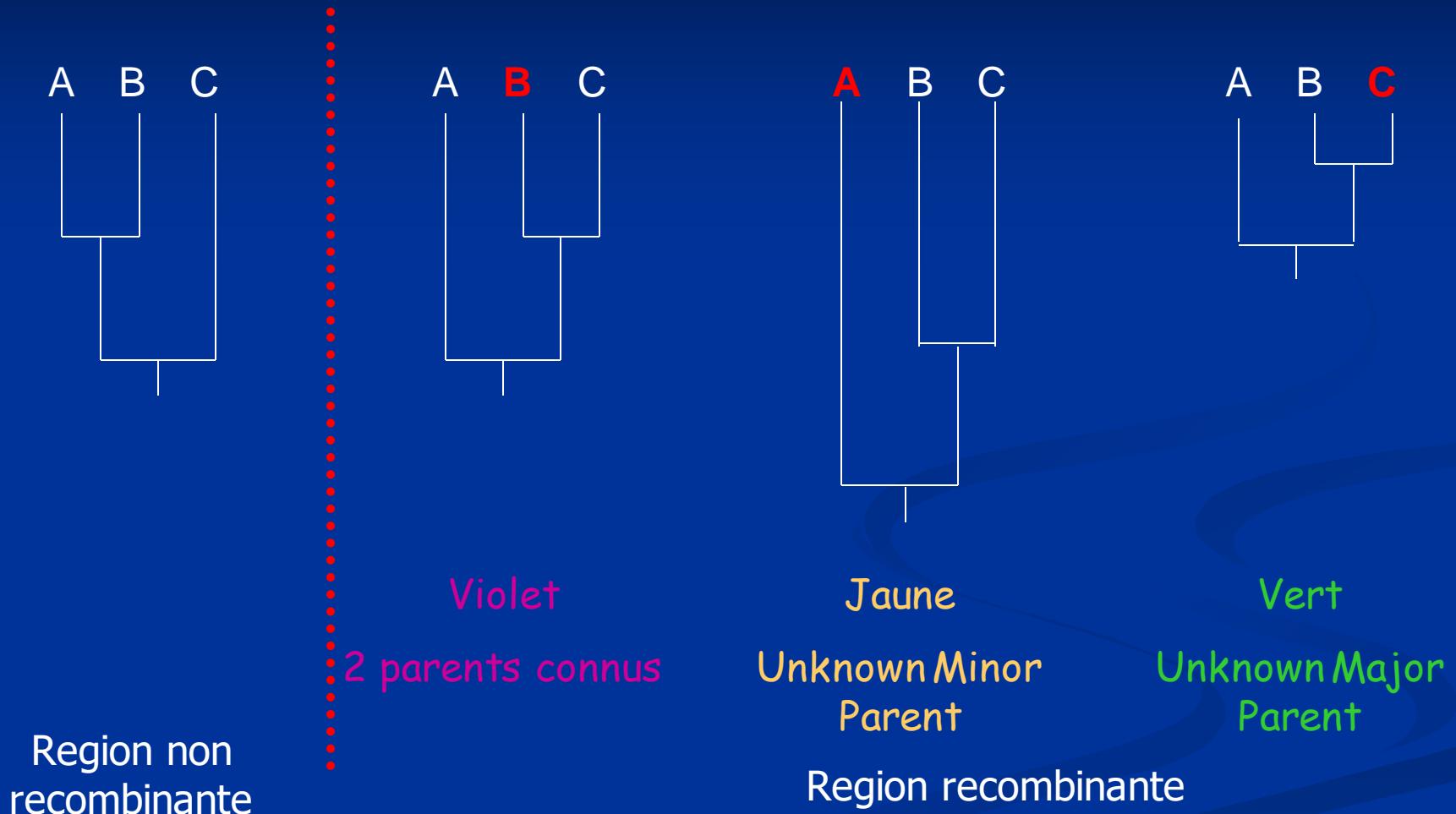
Variabilité génétique

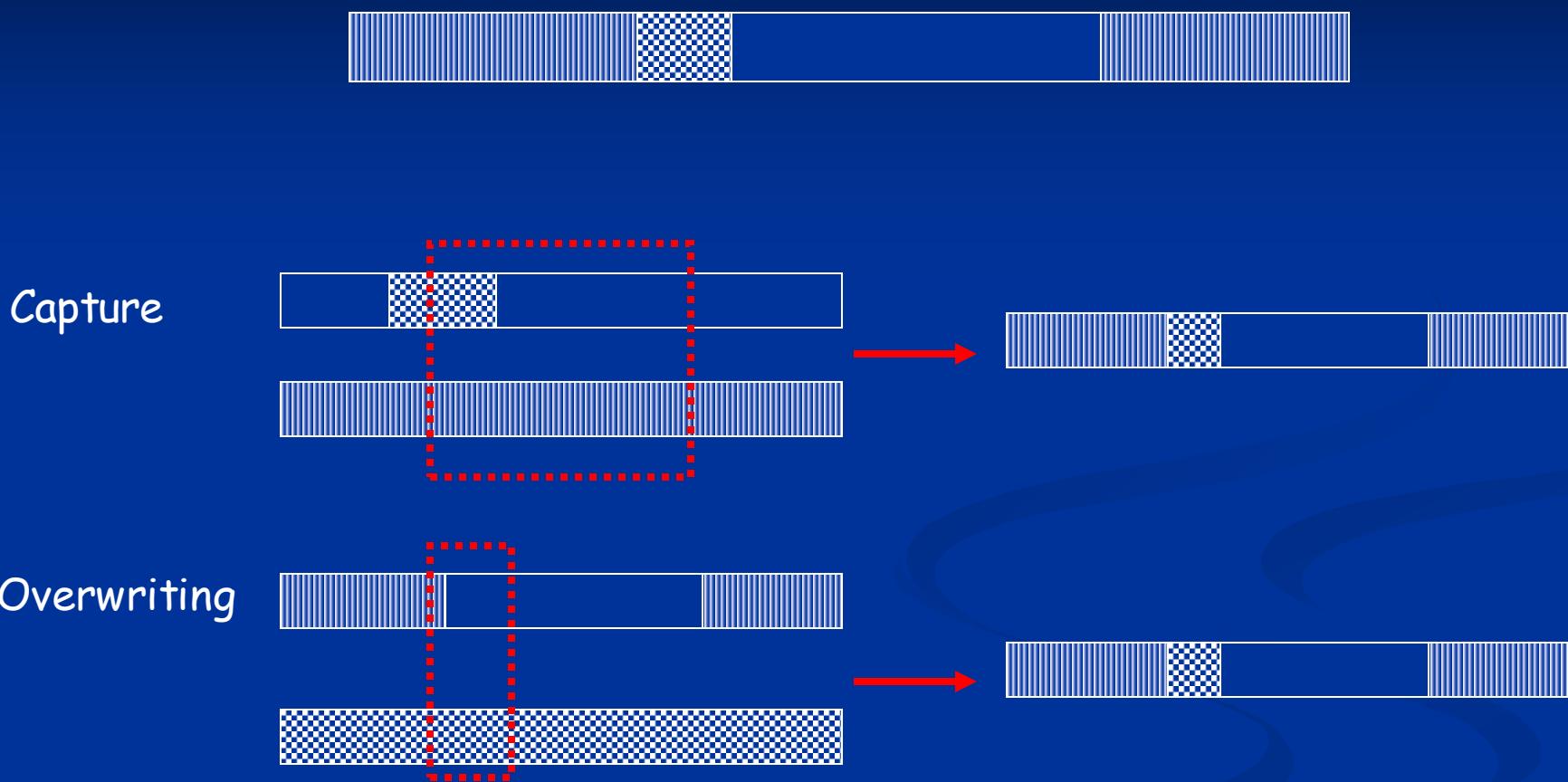
Divergence après recombinaison réduite

Analyser que des séquences présentant une identité > 60%



Trois types de recombinaisons





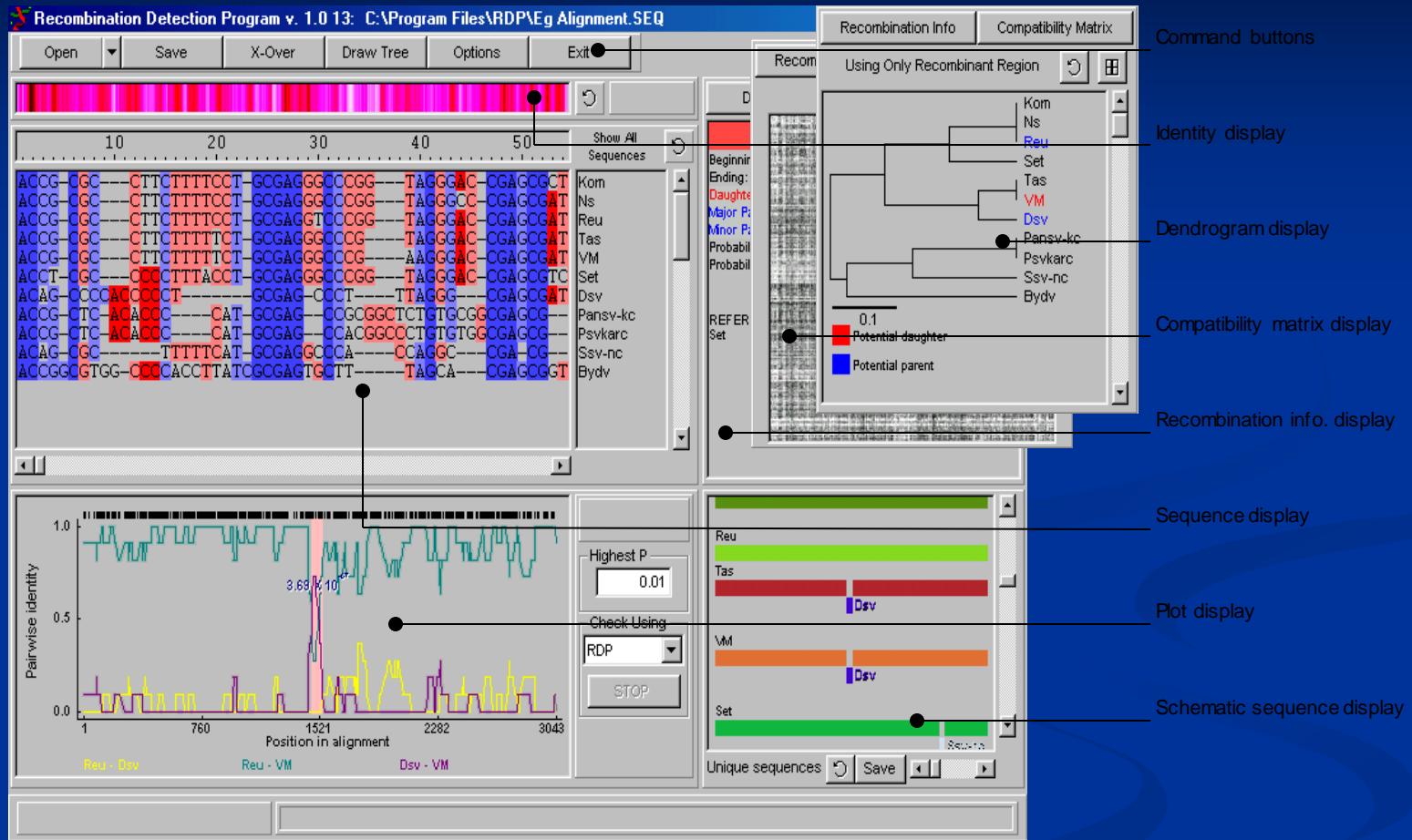


Figure 1. The main components of the RDP interface.

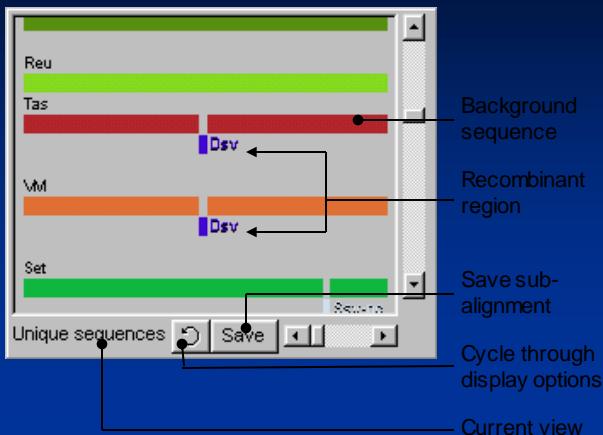


Figure 2. The schematic sequence display

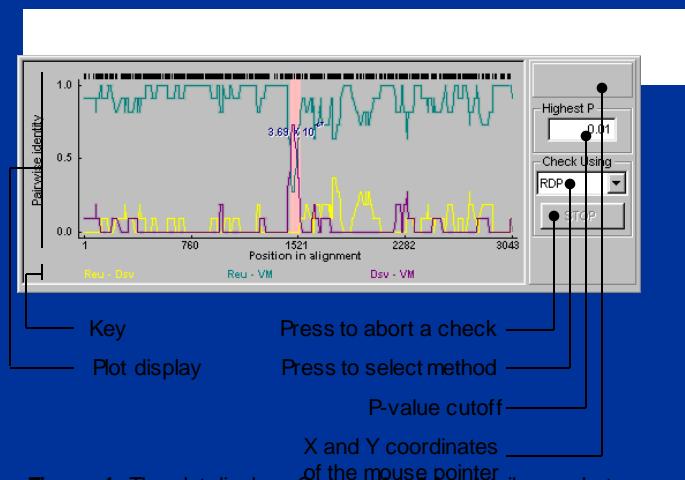


Figure 4. The plot display. See section 8 for details on what is plotted.

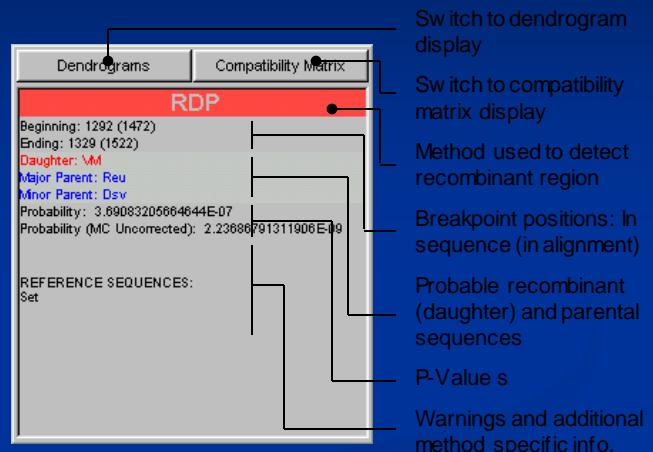


Figure 3. The recombination information display. Breakpoint positions are specific for the daughter (or potentially recombinant) sequence – the breakpoint positions in the alignment are given in parentheses. The “major parent” is a sequence closely related to that from which the greater part of the daughter’s sequence may have been derived. The “minor parent” is a sequence closely related to that from which sequences in the proposed recombinant region may have been derived. P-values that are displayed are either multiple comparison (MC) corrected or uncorrected.

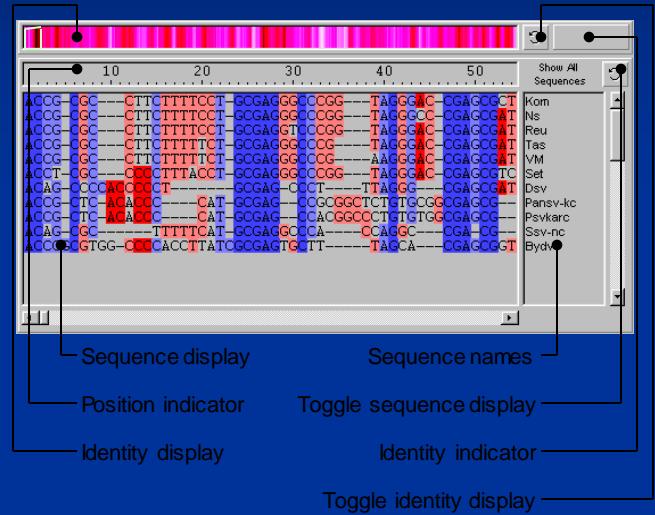


Figure 5. The sequence display.

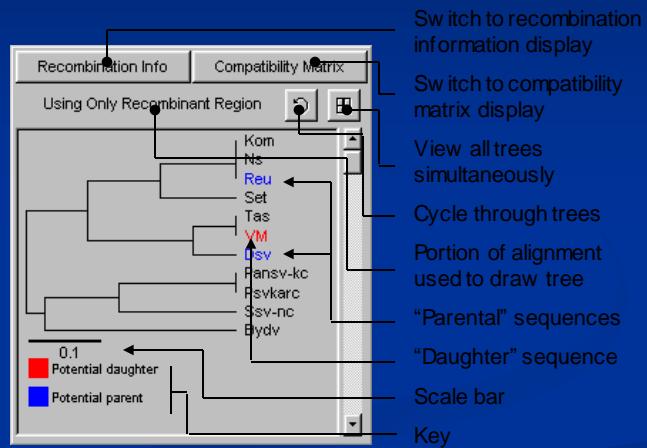


Figure 6. The dendrogram display.

RDP (Martin et Rybicki, 2000)

GENECONV (Padidam et al., 1999)

Bootscan (Salminen et al., 1995)

MaxChi (Maynard Smith, 1992)

Chimaera (Maynard Smith, 1992)

SiScan Gibbs et al., 2000)

LARD

Reticulate

Topal

Phylpro

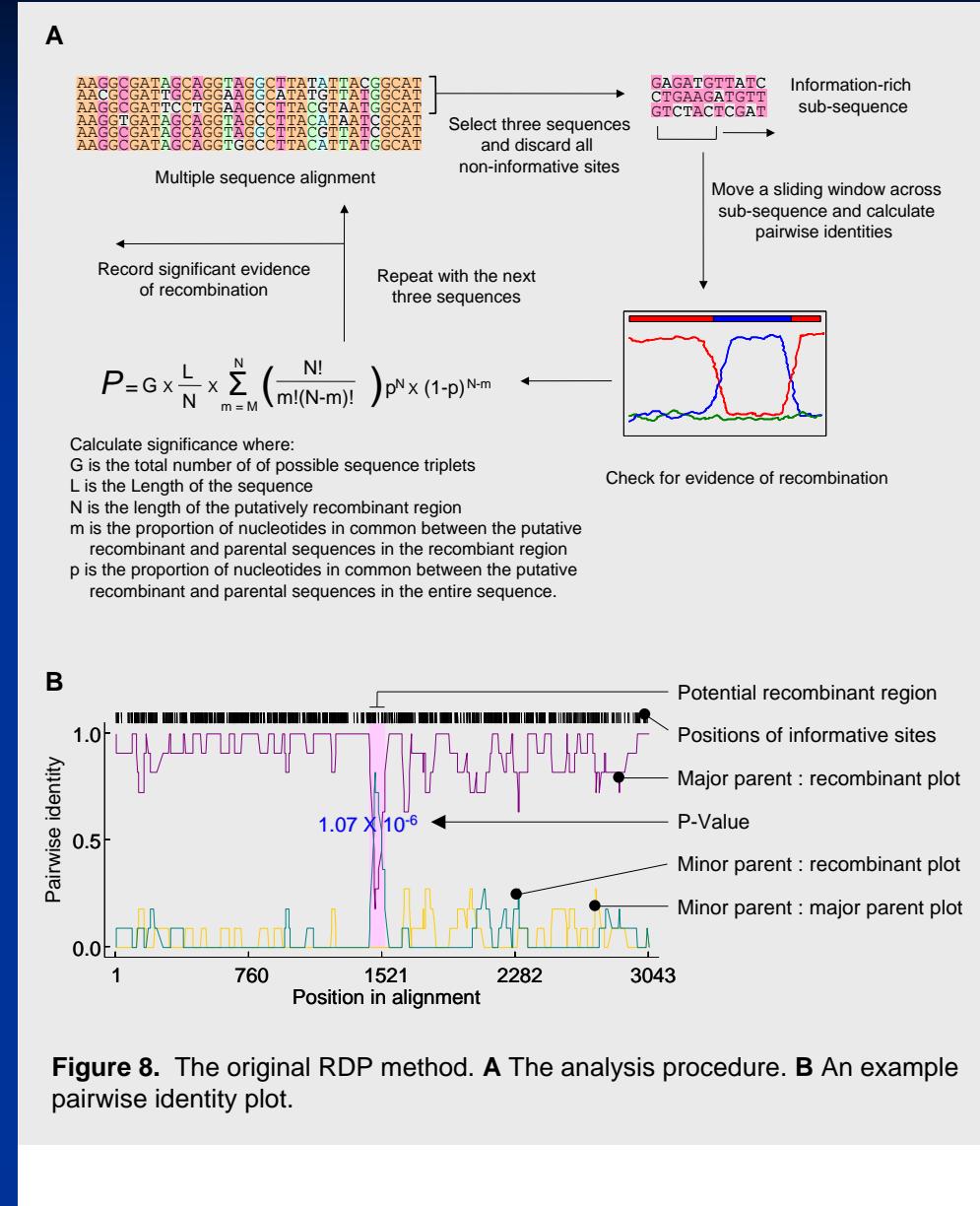
RDP

Analyse de l'alignement
en examinant chaque
triplet de séquence en 3
étapes

Élimination des sites non
informatif

Fenêtre est bougé le
long de l'alignement et
un pourcentage d'identité
est calculé pour les trois
triplet

Calcul d'une p-Value



GENECONV

Analyse de l'alignement en examinant chaque paire de séquence en 3 étapes

Élimination des sites monomorphe

Recherche de zone à forte similarité

Calcul d'une p-Value

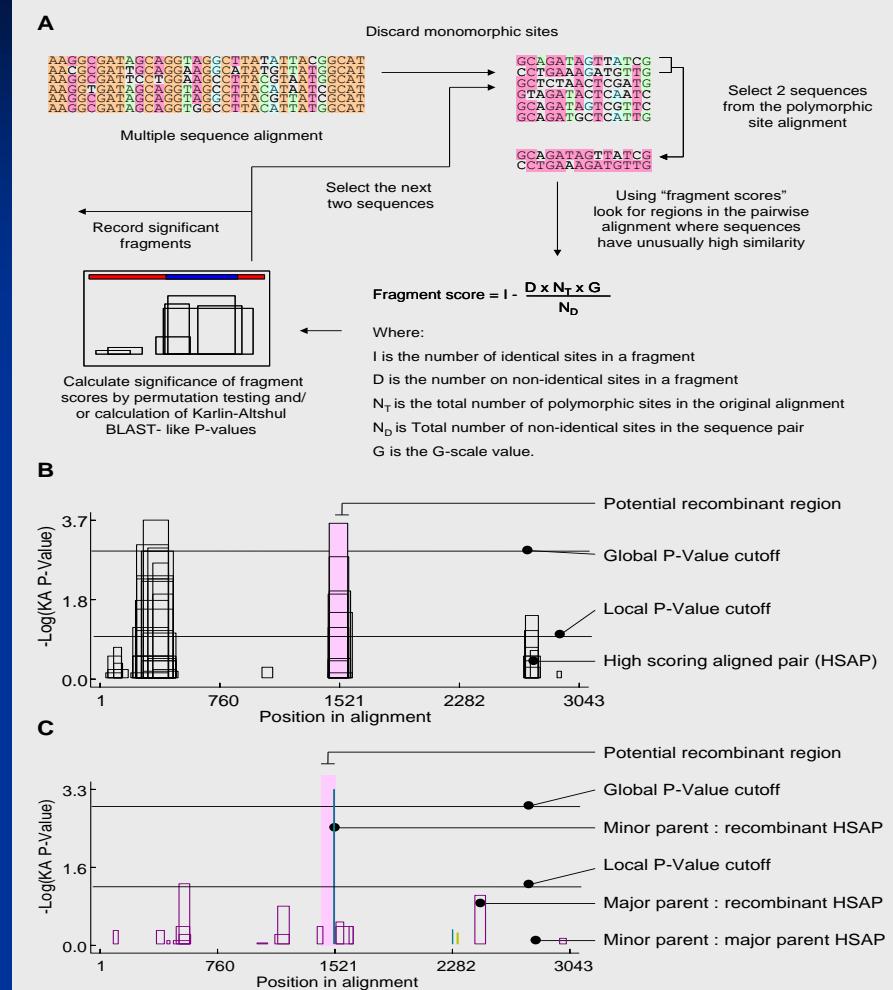


Figure 9. The GENECONV method. **A** The analysis procedure. **B** An example plot of high scoring aligned pairs (HSAPs or fragments). **C** An example plot in which GENECONV is used to check the RDP derived result in Fig 8 B.

MaxChi

Identifie les points de recombinaison. Analyse des séquences par paire

Elimination des sites monomorphes

Une fenêtre est bougée le long de la paire des séquences

Une valeur de chi² est calculée sur le nombre de site polymorphe entre les deux séquences pour chaque côté de la fenêtre

Une forte valeur indique un point de recombinaison

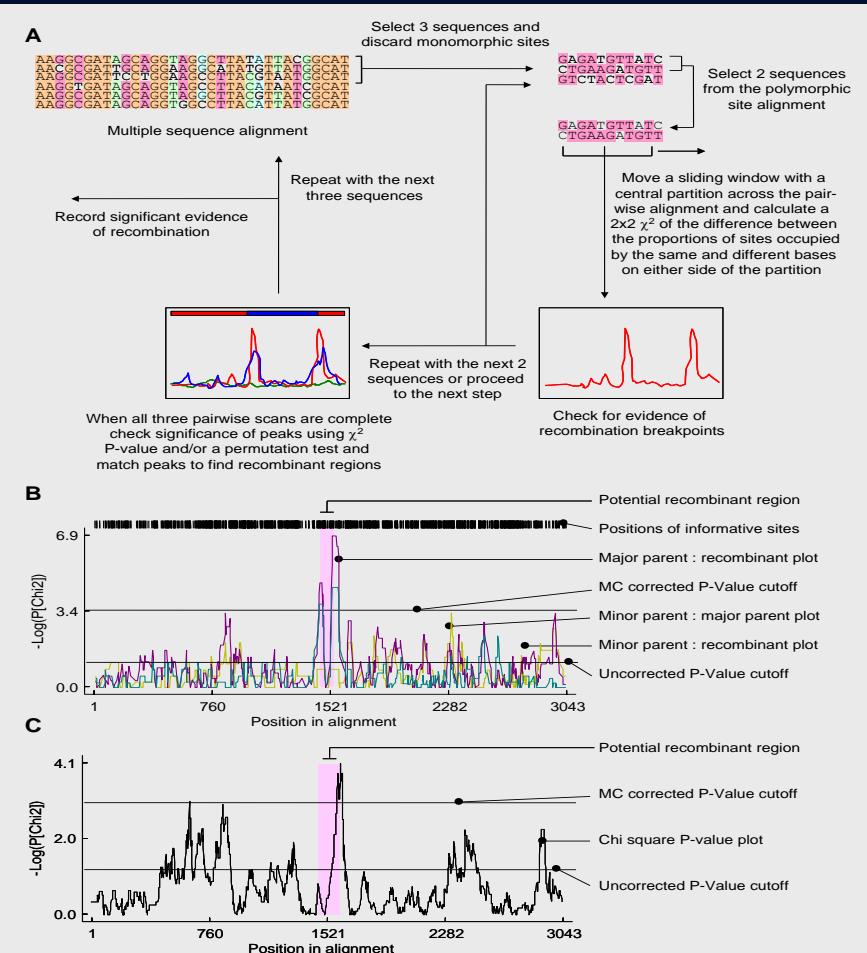


Figure 11. The MaxChi method. **A** The analysis procedure when the MaxChi “scan triplets” setting is used. When the “scan entire dataset simultaneously” setting is used the analysis procedure is the same except that there is only one analysis cycle with the polymorphic site alignment being produced from the entire alignment (instead of it being produced from the currently selected triplet) **B** An example of Chi squared P-value plots used to confirm the RDP derived result in Fig 8 B. **C** An example plot in which MaxChi is used to check the GENECOVN derived result in Fig 9 B.

Après ce type d'analyse, on a déterminer une zone de recombinaison et les 3 séquences impliquées (2 parents et une séquence fille) mais on ne sait pas qui est qui.

Pour déterminer qui est la séquences recombinante :

Construction d'un arbre Neighbour-Joining à partir de l'alignement complet

Pour chaque séquence du triplet, détermination du nombre de changement nucléotidique entre l'ancêtre commun le plus proche (défini par parcimonie) et les séquences actuelles

La branche avec la plus de changement entre la région recombinante relativement au reste est défini comme étant le recombinant

Heureusement, tout ça est automatique...

Conclusion

Conclusions

La reconstruction phylogénétique est devenu courante lors de l'analyse de nouvelles séquences

Les outils d'analyses de séquences sont en plein boum : parfois difficile à suivre, mais très prometteur

Références

The phylogenetic handbook: a practical approach to DNA and Protein phylogeny.

Marco Salemi, Anne-Mieke Vandamme

La reconstruction phylogénétique,

Pierre Darlu, Pascal Tassy
(http://sfs.snv.jussieu.fr/pdf/Darlu_Tassy_online.pdf)

Molecular Evolution and Phylogenetics

Masatoshi Nei, Sudhir Kumar

The phylogenetic handbook: a practical approach to DNA and Protein phylogeny, Marco Salemi, Anne-Mieke Vandamme

La reconstruction phylogénétique, Pierre Darlu, Pascal Tassy (http://sfs.snv.jussieu.fr/pdf/Darlu_Tassy_online.pdf)

Molecular Evolution and Phylogenetics, Masatoshi Nei, Sudhir Kumar

A suivre...

