Bioinformatique M1: Lecture 7

P. Derreumaux

PHYLOGENIE

L'homologie : point de départ de la phylogénie moléculaire

- homologie → ancêtre commun
- changements mineurs : substitutions, insertions ou délétions de courte taille.
- remaniements complexes : fusion ou fission de gènes, ... ⇒ relations d'homologie ne concernent que des segments du gène.

On parle de domaines protéiques homologue et non de protéines homologue

Pour la phylogénie moléculaire, on utilise des gènes orthologues pour reconstituer l'histoire des espèces.

Conditions préalables requises pour la construction de bons arbres

- Disposer du plus grand nombre de gènes homologues possible
- Procéder à un alignement multiple fiable
- Eliminer les régions ambigues, les positions non informatives
- Faire appel à un groupe extérieur (pour enraciner un arbre
- Evaluer l'arbre statistiquement.

Limites de la phylogénie moléculaire

- différents arbres selon l'algorithme
  ↳ arbre réel

Use of DNA or protein sequences for phylogeny.

- DNA or RNA for non-coding sequence evolution (e.g miRNA, binding sites of transcription factors).
- DNA also to study the molecular evolution of protein-coding genes

From Alignments, estimation of:

K_S: number of synonymous substitutions (or silent: no amino-acid change)

K_a: number of non-synonymous substitutions

if $K_a << K_S$, the DNA sequence is under negative (purifying) selection

if $K_a >> K_S$, positive selection

- Protein sequences to be used when preferred (20 AA vs. 4 bases)

Four stages of phylogenetic analysis

Molecular phylogenetic analysis may be described in four stages:

[1] Selection of sequences for analysis

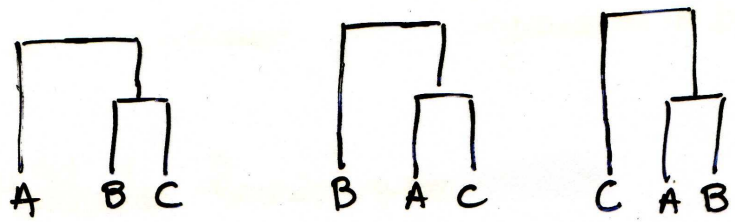
[2] Multiple sequence alignment

[3] Tree building

[4] Tree evaluation

Possible ways of drawing a tree

- Only 3 trees possible for 3 species



- For 4 species → 15 different topologies

- number of possible trees for n OTU's

→ rooted trees = $\frac{(2n-3)!}{2^{n-2}(n-2)!}$

→ unrooted trees = $\frac{(2n-5)!}{2^{n-3}(n-3)!}$

ex 20 OTU → $8\ 10^{21}$ rooted trees
$2\ 10^{20}$ unrooted trees

Problem is a NP (non-polynomial) problem
∝ exponential in character

→ Exhaustive enumeration impossible

Méthodes de construction des arbres

- La Technique de 'Séparation et Evaluation'

- Méthodes heuristiques
  - fondées sur les distances (phénétiques)
    ex UPGMA et 'Neighbor-Joining'
  - fondées sur les séquences et caractères (cladistiques)
    ex Méthode de parcimonie
    Méthode de compatibilité
    Méthode de 'maximum likelihood'

Stratégie "branch and bound" *

(Séparation et évaluation)

"Branch and bound" est une statégie exacte permettant de trouver l'arbre de phylogénie maximal pour 20 espèces ou plus.

Méthode:

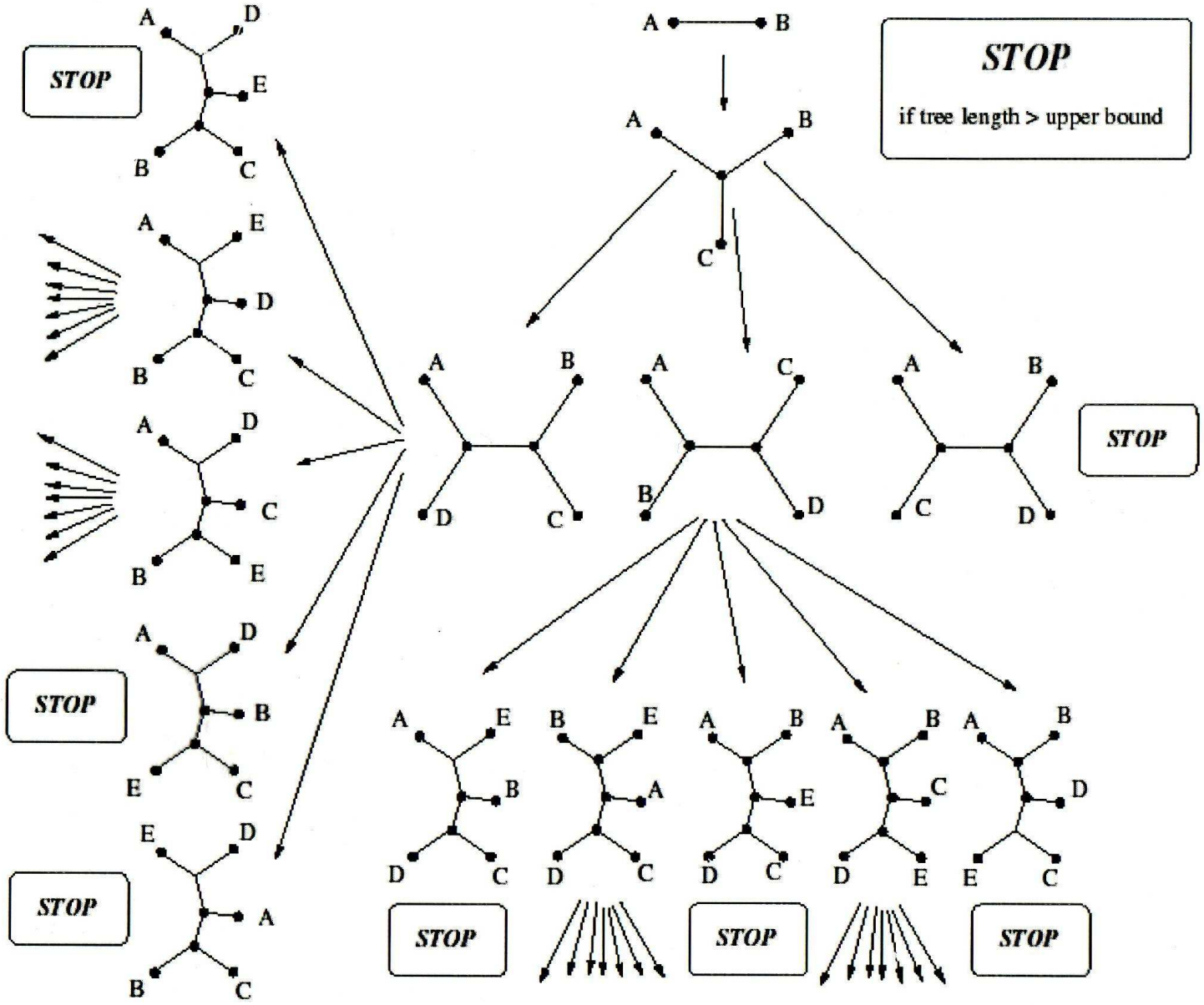
1) Obtenir une borne supérieure du nombre de mutations (par Neighbor Joingning, par exemple) ou de la longueur de l'arbre

2) Construire toutes les topologies d'arbres en ajoutant les espèces une à une

3) Si, pour une topologie donnée, le nombre de mutations est plus grand que la borne supérieure, alors arrêter d'ajouter des espèces à cette topologie

*Hendy, M.D. et Peeny, D., Branch an bound algorithms to determine minimal evolutionary trees, Mathematical Biosciences, 60, pp.133-142, 1982.

Stratégie “branch and bound”



Méthodes de distances (1)

Les méthodes de distances sont basées sur la **similitude globale**.

⇒ On compare les états de caractères entre les organismes pris deux à deux et on **compte le nombre de différences**.

Exemple :

	1	T	T	A	T	T	A	A
	2	A	A	T	T	T	A	A
Séquences	3	A	A	A	A	A	T	A
	4	A	A	A	A	A	A	T

(2) A partir des nombres de différences, on établit une **matrice de distances**

Matrice de distances

	1	2	3	4
1	-	3	5	5
2	3	-	4	4
3	5	4	-	2
4	5	4	2	-

(3) Ces distances vont permettre **l'élaboration d'un arbre** dont la longueur des branches va représenter un nombre de différences.

⇒ Il existe de nombreuses méthodes pour reconstruire un arbre à partir d'une matrice de distances

⇒ Exemple avec l'UPGMA (Unweighted pair group method with arithmetic means)

METHODE DE L'UPGMA

⇒ On cherche les taxons i et j pour lesquels la distance ij est la + petite
⇒ On définit la profondeur du point de branchement entre i et j comme étant dij/2
⇒ Si i et j étaient les 2 derniers ensembles, l'arbre est complet
⇒ On recalcule la matrice en prenant comme distance U à chaque autre taxon k, la moyenne des distances dki et djk (dij+djk)/2
⇒ Retour à l'étape 1

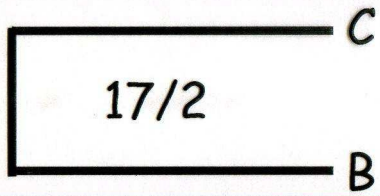
Méthode de distances (3)

Matrice de distances

	A	B	C	D	E
A	-	52	64	65	73
B		-	17	88	97
C			-	94	99
D				-	47
E					-

(1) On cherche les taxons i et j pour lesquels la distance ij est la + petite

(2) On place C et B dans l'arbre avec comme longueur de branche 17/2



(3) On réduit la matrice

A(BC) =(52+64)/2=58

D(BC) = (88+94) /2=91

E(BC) = (97+99)/2 = 98

Matrice de distances

	A	BC	D	E
A	-	58	65	73
BC		-	91	98
D			-	47
E				-

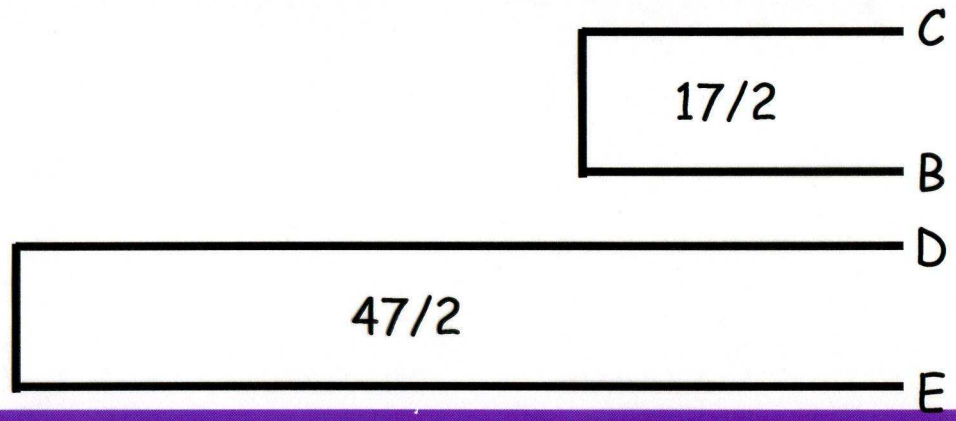
Méthode de distances (3)

Matrice de distances

	A	BC	D	E
A	-	58	65	73
BC		-	91	98
D			-	47
E				-

(1) On cherche les taxons i et j pour lesquels la distance ij est la + petite

(2) On place D et E dans l'arbre avec comme longueur de branche 47/2



(3) On réduit la matrice

(A)(DE) = (65+73)/2 = 69

(BC)(DE) = (91+98)/2 = 94.5

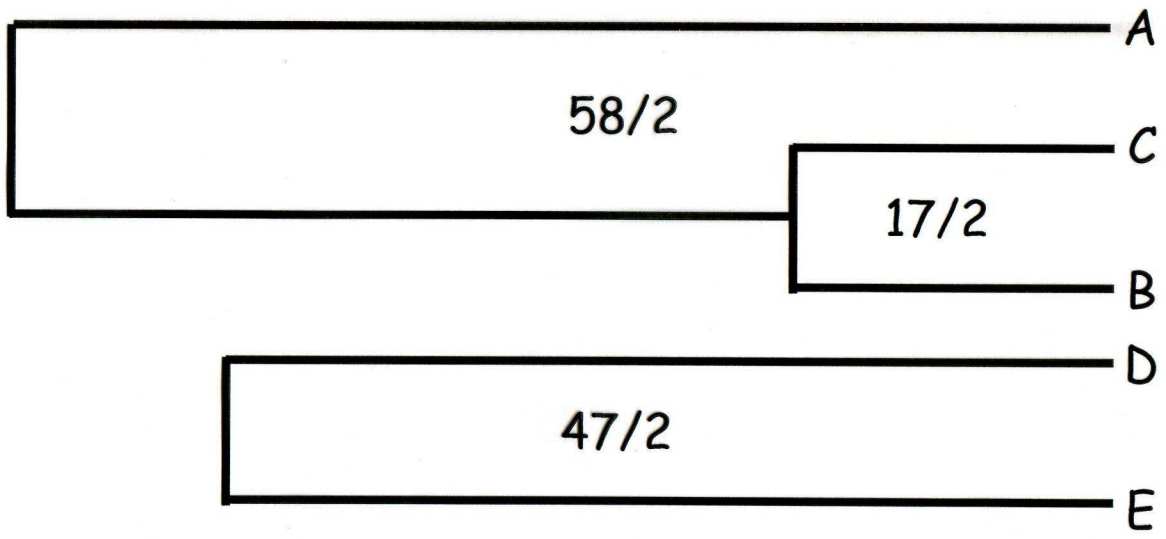
Méthode de distances (3)

Matrice de distances

	A	BC	DE
A	-	58	69
BC		-	94.5
DE			-

(1) On cherche les taxons i et j pour lesquels la distance ij est la + petite

(2) On place A et BC dans l'arbre avec comme longueur de branche 58/2



(3) On réduit la matrice

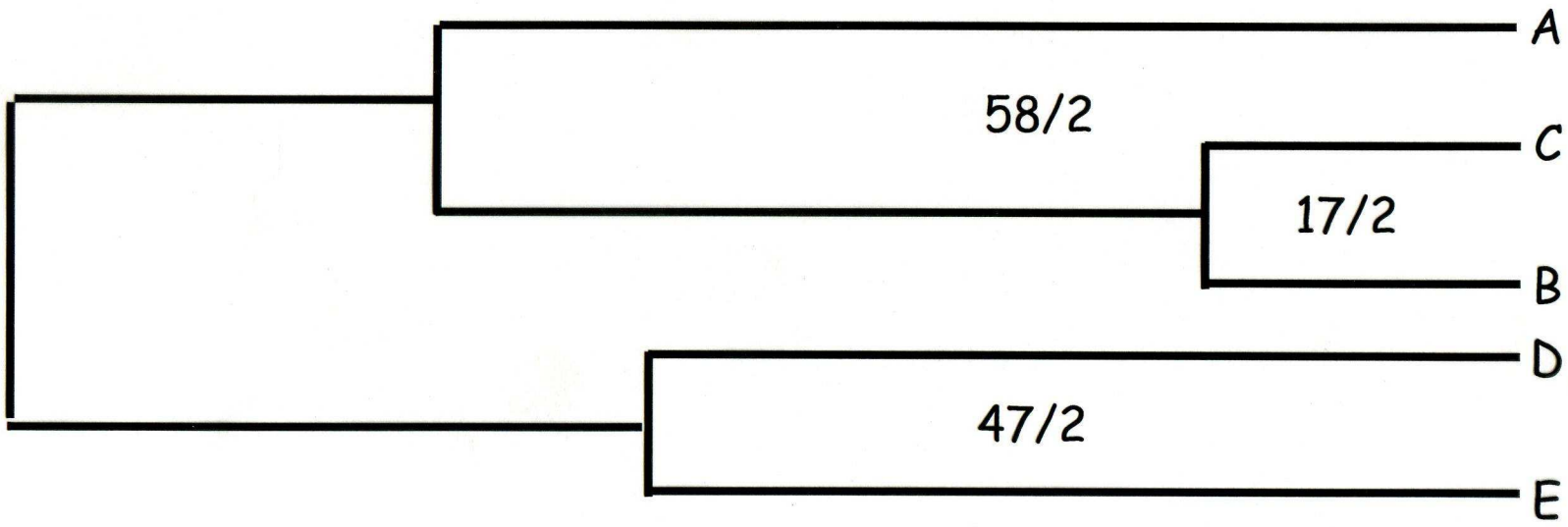
(A(BC))(DE) =(69+94.5)/2=81.75

Méthode de distances (3)

Matrice de distances

	DE
ABC	81.75

On place ABC et DE dans l'arbre avec comme longueur de branche 81.75/2



Méthode de distances (4)

Attention : dans la plupart des cas l'UPGMA n'est pas une bonne méthode de reconstruction

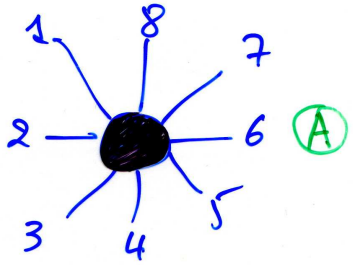
=> On préfère utiliser d'autres méthodes de distances comme le Neighbor-Joining

Neighbor-Joining.

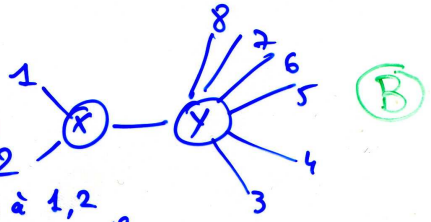
- Estimation de l'arbre en minimisant la distance évolutive.
- Principe : recherche séquentiellement les voisins minimisant la longueur totale de l'arbre.

N-3 fois

- Arbre en étoile



- Séparation 1 paire d'individus (OTU 1 et 2 sont regroupés)



X : noeud commun à 1, 2
Y : noeud — 3, ... 8

→ estimation longueur branche

- Séparation des $N(N-1)/2$ paires différentes

→ paire d'individus qui donne longueur minimale et réduite à un individu

Pour l'arbre donné en (B), la somme (S_{12}) de toutes les longueurs de branche

$$S_{12} = \frac{1}{2(N-2)} \sum_{k=3}^{N} (d_{1k} + d_{2k}) + \frac{1}{2} d_{12} + \frac{1}{N-2} \sum_{3 \le i < j \le N} d_{ij}$$

Méthode cladistique : Méthode de parcimonie

Principe : Recherche de l'arbre dont la topologie demande le nombre le plus faible de changements evolutifs (substitutions)

Calcul du nombre minimum de substitutions : algorithme de Fitch (1971).

1. Reconstitution exhaustive des états ancestraux possibles
2. Choix des états ancestraux les plus parcimonieux
3. Calcul du nombre de substitutions

Unweighted versus weighted parsimony

- In unweighted parsimony all substitutions between nucleotides or amino acids are expected to occur in equal frequencies in all directions
- This is biologically unrealistic – e.g. transitions occur more frequently than transversions & amino acid substitutions vary greatly for different residue pairs
- Weighted parsimony attempts to address this by giving different weights to different types of substitutions when counting changes on topologies
- For example, 1st 2nd & 3rd codon positions vary at different rates so a weighting scheme such as w1 = 3, w2 = 5 and w3 =1 that reflects the relative rate of change can be employed
- Similarly different weights can be used for relatively more common transitions (w=1) and more rare transversions (w=2)
- Much effort has been put into developing weighted parsimony methods for amino acid sequences that reflect the genetic code and the biochemical properties of the different residues (unclear how much these improve the results)

Advantages and disadvantages of parsimony

Advantages:

1 – based on a logically coherent and biologically plausible model of evolution
2 – free from assumptions used in distance estimations
3 – better than distance methods when extent of sequence divergence is low (≤ 10%), rate of substitution is constant, number of residues is large
4 – very useful for certain types of molecular data e.g. insertions and deletions
5 – provides several ways to evaluate the support for the topologies produced e.g. measures of homoplasy, ranked list of trees based on length

Disadvantages:

1 – gives incorrect topologies when backward substitutions are present (common with nucleotides) and when the number of sites is fairly small
2 – gives incorrect topologies when rate of substitution varies substantially across lineages
3 – long branch attraction – long branches (and short branches) tend to group together on reconstructed tree → problems for all methods.
4 – difficult to treat the results in a statistical framework

Méthode cladistique : Méthode de compatibilité

principe : le critère d'optimalité des arbres n'est plus le plus petit nombre de pas mais le plus grand nombre de sites compatible.

Méthode cladistique : Méthode de vraisemblance maximum

principe :
- utilise un modèle explicite d'évolution.
- P_{ij}^t : probabilité qu'un site initialement avec le nucléotide i change en un nucléotide j au cours du temps t

$$P_{ij}^t = \delta_{ij} e^{-\mu t} + (1 - e^{-\mu t}) g_j.$$

$\delta_{ij} = 1$ si $i = j$ et $\delta_{ij} = 0$ sinon

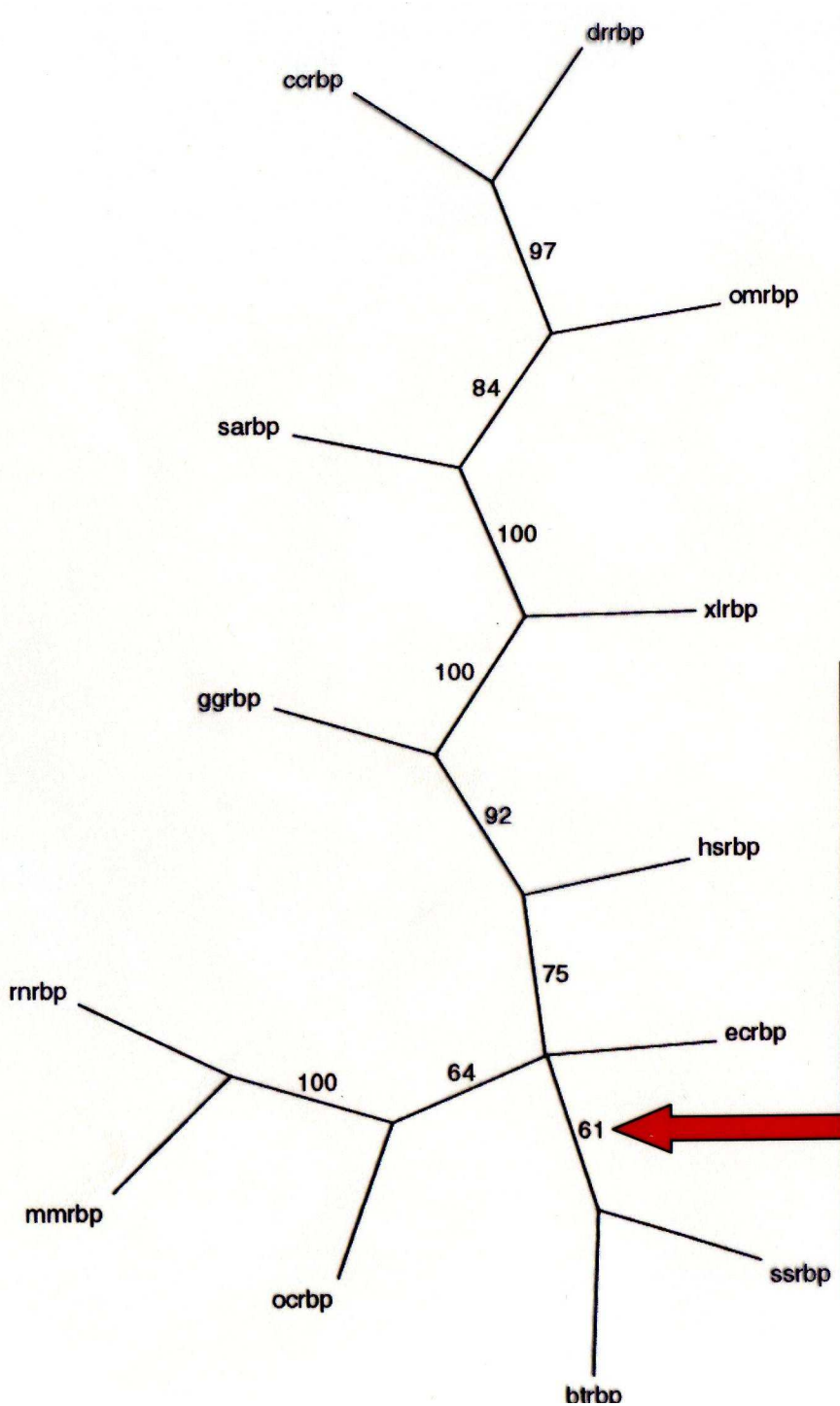
g_j : fréquence du nucléotide j.

- méthode consiste à évaluer pour chacune des configurations explorées la probabilité globale résultant de la sommation des probabilités pondérées calculées sur chacun des sites.
- La méthode retient le ou les arbre avec P_{max}.

Evaluating trees: bootstrapping

Bootstrapping is a commonly used approach to measuring the robustness of a tree topology. Given a branching order, how consistently does an algorithm find that branching order in a randomly permuted version of the original data set?

To bootstrap, make an artificial dataset obtained by randomly sampling columns from your multiple sequence alignment. Make the dataset the same size as the original. Do 100 (to 1,000) bootstrap replicates. Observe the percent of cases in which the assignment of clades in the original tree is supported by the bootstrap replicates. >70% is considered significant.



Resampling method
Bootstrap Resample characters
Jackknife with 50 % deletion Emulate "Jac" resampling
Number of replicates: 100 Random number seed: 808890985
Type of search
Full heuristic "Fast" stepwise-addition
Branch-and-bound Neighbor-joining/UPGMA (distance only)
Consensus tree options
Retain groups with frequency > 50 %
Include groups compatible with 50% majority-rule consensus
Show table of partition frequencies
Don't show groups with bootstrap proportions ≤ 5 %
Character-weight handling... Save trees to file
Defaults... Cancel Continue

In 61% of the bootstrap resamplings, ssrbp and btrbp (pig and cow RBP) formed a distinct clade. In 39% of the cases, another protein joined the clade (e.g. ecrbp), or one of these two sequences joined another clade.

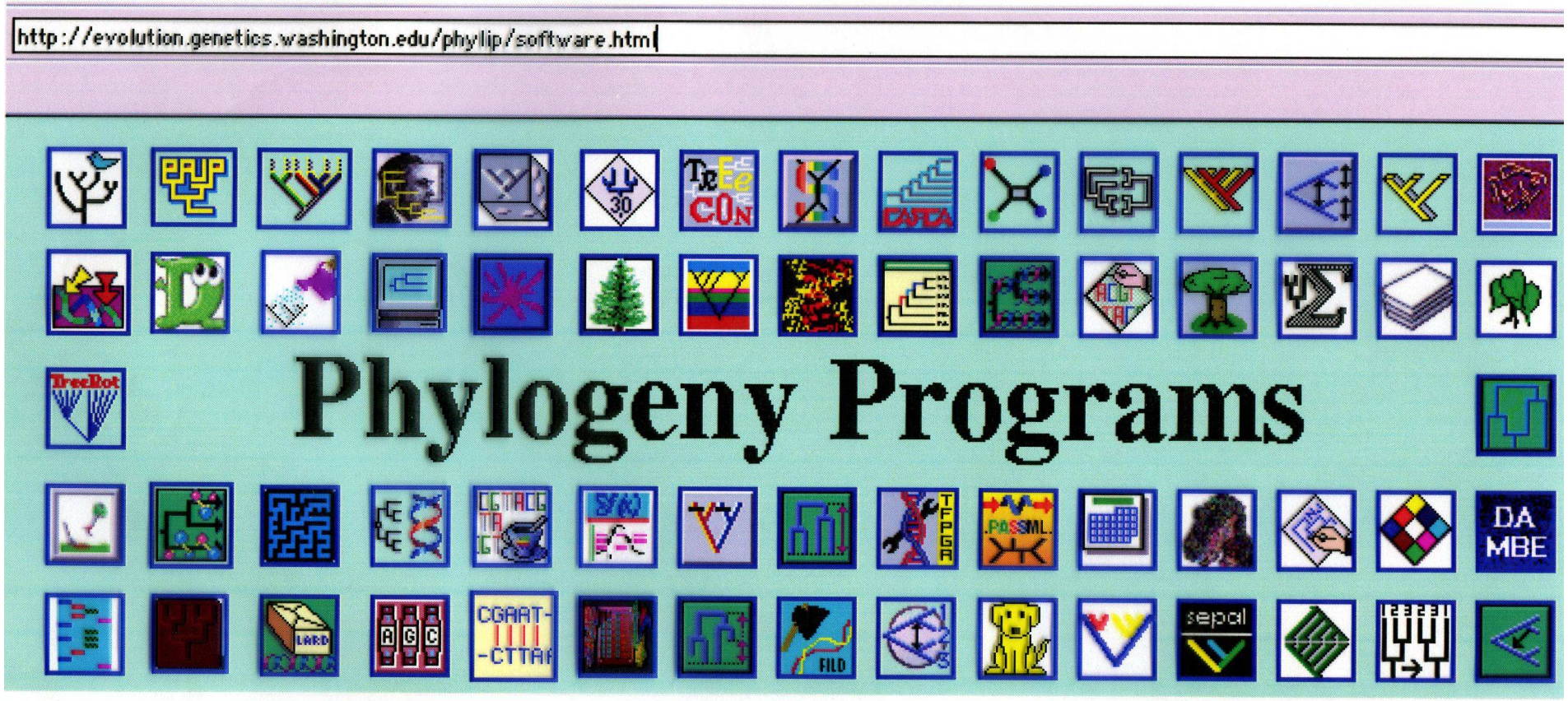
Which Method to Choose?

- depends upon the sequences that are being compared
  - strong sequence similarity:
    - maximum parsimony
  - clearly recognizable sequence similarity
    - distance methods
  - All others:
    - maximum likelihood

Which Method to Choose?

- Best to choose at least two approaches
- Compare the results – if they are similar, you can have more confidence

http://evolution.genetics.washington.edu/phylip/software.html



This site lists 200 phylogeny packages. Perhaps the best-known programs are PAUP (David Swofford and colleagues) and PHYLIP (Joe Felsenstein).