M2 - BIM - PHYL

Maximum likelihood

Algorithmes de reconstruction des arbres phylogénétiques

Alessandra Carbone
Université Pierre et Marie Curie

Nucleotide/amino-acid sequences

Gly Ala Ile Leu Asp Arg

-GGAGCCATATTAGATAGA-

-GGAGCAATTTTTGATAGA-

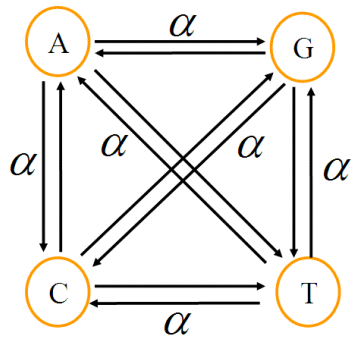
Gly Ala Ile Phe Asp Arg

3 different DNA positions but only one different amino acid position: 2 of the nucleotide substitutions are therefore synonymous and one is non-synonymous.

DNA yields more phylogenetic information than proteins. **The nucleotide sequences of a pair of homologous genes have a higher information content than the amino acid sequences of the corresponding proteins**.

This is because mutations that result in synonymous changes alter the DNA sequence but do not affect the amino acid sequence. (On the other hand, amino-acid sequences are more efficiently and more precisely aligned.)

Jukes-Cantor model : all changes occur at equal probabilities



α = the rate of substitution

The nucleotide substitution process of DNA sequences specifies the relative rates of change of each nucleotide along the sequence.

	A	C	G	T
A	-3/4	1/4	1/4	1/4
C	1/4	-3/4	1/4	1/4
G	1/4	1/4	-3/4	1/4
T	1/4	1/4	1/4	-3/4

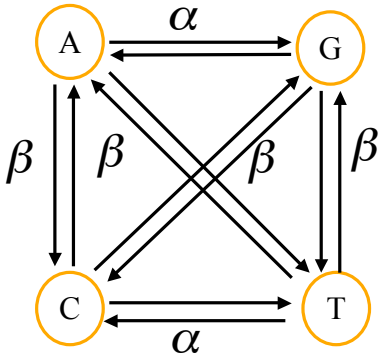
For the Jukes-Cantor model, the expected substitution rate per site is 3/4, and the number of expected substitutions in time t is 3/4t.

Kimura model : A more realistic simulation represents different probabilities for transitions than for transversions

While the probability of a transversion is twice as high as for a transition, we often see that the ratio

transitions/transversions ≠ ½!!

The ratio is often much higher!
-> higher mutation rate for transitions



α -> **transitions**
β -> **transversions**

Tree construction: how to proceed?

1. Consider the set of sequences to analyse ;
2. Align "properly" these sequences ;
3. Apply phylogenetic making tree methods ;
4. Evaluate statistically the phylogenetic tree so obtained.

Methodology :

1- Multiple alignment;
2- Bootstrapping;
3- Consensus tree construction and evaluation.

Alignment is essential preliminary to tree construction

```
GACGACCATAGACCAGCATAG
GACTACCATAGA-CTGCAAAG
***  ********  * ***  **

GACGACCATAGACCAGCATAG
GACTACCATAGACT-GCAAAG
***  *********  ***  **
```

Two possible positions for the indel

If errors in indel placement are made in a multiple alignment then the tree reconstructed by phylogenetic analysis is unlikely to be correct.

Steps in Multiple Sequence Alignments

A common strategy of several popular multiple sequence alignment algorithms is to:

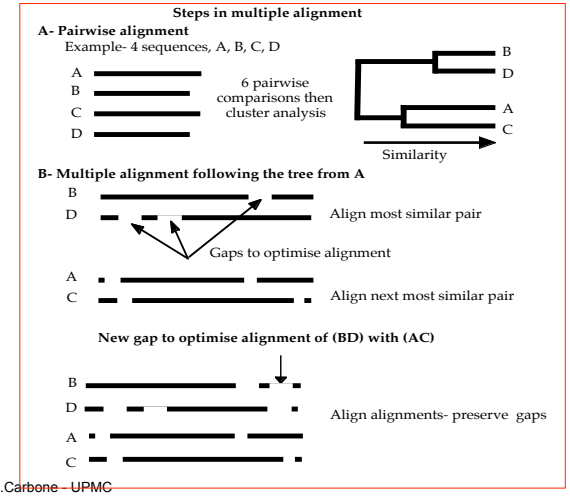
1- generate a pairwise distance matrix based on all possible pairwise alignments between the sequences being considered;

2- use a statistically based approach to construct an initial tree;

3- realign the sequences progressively in order of their relatedness according to the inferred tree;

4- construct a new tree from the pairwise distances obtained in the new multiple alignment;

5- repeat the process if the new tree is not the same as the previous one.



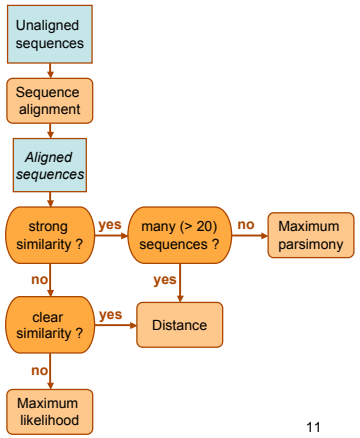
Phylogenetic tree construction methods

- A phylogenetic tree is characterised by its topology (form) and its length (sum of its branch lengths) ;
- Each node of a tree is an estimation of the ancestor of the elements included in this node;
- There are 3 main classes of phylogenetic methods for constructing phylogenies from sequence data :

- Distance based methods (eg. Neighbour Joining (NJ)): find a tree such that branch lengths of paths between sequences (species) fit a matrix of pairwise distances between sequences.
- Maximum Parsimony : find a phylogenetic tree that explains the data, with as few evolutionary changes as possible.
- **Maximum likelihood** : find a tree that maximizes the probability of the genetic data given the tree.

Phylogenetic inference from sequence comparison

Alternative approaches
- Maximum parsimony
- Distance
- Maximum likelihood



Source: Mount (2000)

Tree evaluation: bootstrapping

- sampling technique for estimating the statistical error in situations where the underlying sampling distribution is unknown
- evaluating the reliability of the inferred tree - or better the reliability of specific branches [are members of a clade always members of that clade?]

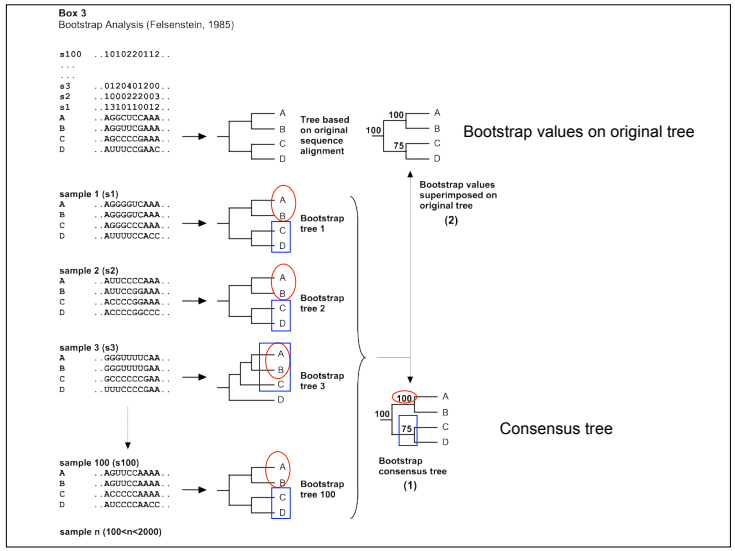
How to proceed:
- From the original alignment, columns in the sequence alignment are chosen at random 'sampling with replacement'
- a new alignment is constructed with the same size as the original one
- a tree is constructed

This process is repeated 100 of times [can take a long time]

Show bootstrap values on phylogenetic trees

- majority-rule consensus tree
- map bootstrap values on the original tree

! Bootstrapping doesn't really assess the accuracy of a tree, it only indicates the **consistency of the data**



Maximum likelihood: the concept of likelihood

Problem: given some data D, a decision must be made about an adequate explanation of the data

Formulation of a specific model and of a hypothesis

In phylogeny, part of the model considers that sequences are evolving according to a tree.

that is based on:
- a tree structure
- branch lengths
- parameters of the model of sequence evolution
- ….

Example: a coin tossing

Coin flipped n=100 times. Observed data: h=21 heads and t=79 tails
D=(21,79)

Model:
- head appears with probability θ when the coin is flipped
- the outcomes of the experiments are independent
- θ does not change during the experiment
- the experiment has only two possible outcomes

Since both head and tails have been obtained, $0< \theta < 1$ and, the probability to obtain exactly h times head is

$$L(\theta)= \Pr[H=h]=C^n_h\ \theta^h\ (1-\theta)^{n-h}$$

This probability can be seen as a function of θ, where n and h are given and it defines the likelihood function L.

Some hypothesis (i.e. values of θ) generate the observed data with higher probability than others. The maximum is obtained for $\theta=21/100$.

To compute easily L(θ), first compute the logarithm of L(θ) which results in:

$$\log[L(\theta)] = \log\left(C^{n}_{h}\right) + h\log(\theta) + (n-h)\log(1-\theta)$$

The problem is now to find the value of θ maximizing the function.
By applying calculus we compute the derivative and solve the equation where the derivative equals 0:

$$\frac{\partial \log [L(\theta)]}{\partial \theta} = \frac{h}{\theta} + \frac{n-h}{1-\theta} = 0$$

This gives θ=h/n. This is called the maximum-likelihood estimate (MLE) of the probability of seeing a head in a single coin toss.

In evolution, point mutations are considered chance events, just like tossing a coin. Therefore, the probability of finding a mutation along one branch in a phylogenetic tree can be calculated by using the same maximum likelihood framework.

The main idea behind phylogeny inference with maximum likelihood is to determine the tree topology, branch lengths, and parameters of the evolutionary model that maximize the probability of observing the sequences at hand.

The likelihood function is the conditional probability of the data (i.e. sequences) given a hypothesis (i.e. a model of substitution with a set of parameters θ and a tree τ)

$$L(\tau, \theta) = \text{Prob}(\text{Data} \mid \tau, \theta)$$
$$= \text{Prob}(\text{aligned sequences} \mid \text{tree, model of evolution})$$

The MLE of τ and θ are those making the likelihood function **as large as possible**.

Warning:

The likelihood function is not a « probability ». It is the probability of the observed event, not of the unknown parameters. The parameters do not depend on chance.

The probability of getting the observed data has nothing to do with the probability that the underlying model is correct.

The simpler tree: two nodes and one branch between them.
Estimation of branch length that produces the data with maximum likelihood.

Sequences evolve with the Jukes-Cantor model:
each position evolves independently and with the same evolutionary rate.

The number of expected substitutions in time t is d=3/4t.
The alignment has length n for the two sequences:
$S_i = s_i(1)s_i(2)\ldots s_i(n)$ where $s_i(j)$ is the nucleotide of sequence i at position j.

$$\log(L(d)) = C + l_0 \log[P_{xx}(d)] + l_1 \log[P_{xy}(d)]$$

where l_0 is the number of identical pairs of nucleotides
l_1 is the number of different pairs
$l=l_0+l_1$
$D=(l_0, l_1)$

which is maximized at $d = -3/4 \log\left[1 - 4/3 \frac{l_1}{l_0 + l_1}\right]$

$$\log (L(d)) = C + l_0 \log[P_{xx}(d)] + l_1 \log[P_{xy}(d)]$$

is maximized at $d = -3/4 \log \left[1 - 4/3 \frac{l_1}{l_0 + l_1} \right]$

The maximum-likelihood tree relating the sequences S_1 and S_2 is a straightline of length d, with the sequences at its end-points.

This example was completely computable because :
- JC is the simplest model of sequence evolution
- the tree has a unique topology

Maximum likelihood for tree identification :
the complex case

According to this method:

- the bases (nucleotides or amino acids) of all sequences at each site are considered separately (as independent),
- the log-likelihood of having these bases are computed for a given topology by using the same evolutionary model. This log-likelihood is added for all sites, and the sum of the log-likelihood is maximized to estimate the branch length of the tree.
-A rate-specific factor of evolution can be assumed for all sites

This procedure is repeated for all possible topologies, and the topology that shows the highest likelihood is chosen as the final tree.

Note that :
1. ML estimates the branch lengths of the final tree ;
2. ML methods are usually consistent ;
3. Drawback : they need long computation time to construct a tree.

In theory, each site could be assigned :

- its own model of sequence evolution, and
- its own set of branch length

The goal to reconstruct a tree from an alignment becomes almost computationally intractable and several simplifications are needed.

Some observations coming from biology:

• Kimura model vs JC. Since transitions (exchanging purine for a purine and pyrimidine for a pyrimidine) are observed roughly 3 times as often as transversions (exchanging a purine for a pyrimidine or vice versa); it can be reasonably argued that a greater likelihood exists that the sequence with C and T are more closely related to each other than they are to the sequence with G.

• Calculation of probabilities is complicated by the fact that the sequence of the common ancestor to the sequences considered is unknown.

• Multiple substitutions may have occurred at one or more sites and all sites are not necessarily independent or equivalent.

Still, objective criteria can be applied to calculate the probability for every site and for every possible tree that describes the relationships of the sequences in a multiple alignment.

The maximum likelihood is computed, analytically if possible, with the following formula

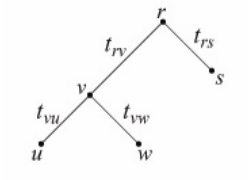
$L(\tau, M, \rho) = Pr[D, \tau, M, \rho] = \prod_{j=1...l} Pr[D_j, \tau, M, \rho]$

where D is the data
τ is the tree
M is the evolutionary model
ρ is the vector of rates

Computing the probability of an alignment for a fixed tree

Consider a tree τ with its branch lengths (i.e. number of substitutions),
M the model of sequence evolution
ρ the site-specific rate vector ρ_j=1 for each site j

Goal: to compute the probability of observing one of the 4^n possible patterns in an alignment of n sequences



To compute Pr[Dj, τ, M, **1**] for a specific site j, where Dj=(u,w,s) are the nucleotides observed, it is necessary to know the ancestral states v and r. The conditional probability of the data given v and r will be:

$$P_{r\to s}(t_{rs}) \cdot P_{r\to v}(t_{rv}) \cdot P_{v\to u}(t_{vu}) \cdot P_{v\to w}(t_{vw})$$

In almost any realistic situations the ancestral sequences are not available and therefore one sums over all possible combinations of ancestral states of nucleotides:

$L(D, \tau, M, \mathbf{1}) = Pr[D, \tau, M, \mathbf{1}] =$

$$\sum_r \sum_v P(r) \cdot P_{r\to s}(t_{rs}) \cdot P_{r\to v}(t_{rv}) \cdot P_{v\to u}(t_{vu}) \cdot P_{v\to w}(t_{vw})$$

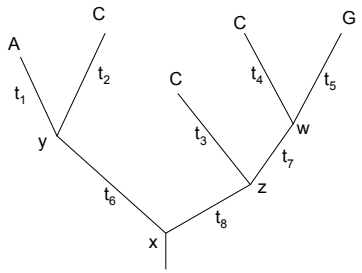
where P(r) is the probability for nucleotide r.

When several positions are taken into account: we shall repeat the previous computation for each of the positions separately, and multiply the results (note that by assumptions, the positions are pairwise independent). The equation is:

$$\begin{aligned} L &= P(M|T) = \prod_{\text{character j}} P(M_j|T) \\ &= \prod_{\text{character j}} \left\{ \sum_{\text{reconstruction R}} P(M_j, R|T) \right\} \\ &= \prod_{\text{character j}} \{ \sum_{\text{reconstruction R}} \left[P(root) \cdot \prod_{\text{edge u}\to\text{v}} P_{u\to v}(t_{uv}) \right] \} \end{aligned}$$

Remark: for the formulation, the tree is assumed to be rooted.
If the model is reversible, that is $P(x)P_{x\to y}(t) = P(y)P_{y\to x}(t)$, then the tree is unrooted and a root can be arbitrarily chosen, with no change in the tree likelihood.

Example :



P(A,C,C,C,G,x,y,z,w|T)=
P(x) $P(y|x,t_6)$ $P(A|y,t_1)$ $P(C|y,t_2)$
$P(z|x,t_8)$ $P(C|z,t_3)$
$P(w|z,t_7)$ $P(C|w,t_4)$ $P(G|w,t_5)$

L'expression est pleine de termes: 4^4=256 (4 nœuds internes qui peuvent avoir associés 4 possibles lettres)
n=10 , 262.144 termes
n=20, 274.877.906.944 termes

The sum in the formula can be efficiently assessed by evaluating the likelihoods moving from the endnodes of the tree to the root (Felsenstein, 1981).

Felsenstein's pruning algorithm :

We denote $C_j(X,v)$ = P(subtree with root v | v_j = X)

where $C_j(X,v)$ is the likelihood conditioned to the subtree with root v, that is the probability to observe a subtree at root v with v labelled by X (i.e. X takes a value among the four A,T,C,G) at position j in the alignment.

Initialisation:

For each leaf v labelled X: $C_j(X,v) = \begin{cases} 1 & \text{si } v_j=X \\ 0 & \text{sinon} \end{cases}$

Recursion:

Cross the tree in postfix order ; for each internal node v with children u and w, compute for each possible value X (= value at position j):

$$C_j(X,v) = [\sum_y C_j(y,u) \cdot P_{x \to y}(t_{vu})\,] \cdot [\sum_y C_j(y,w) \cdot P_{x \to y}(t_{vw})\,]$$

The final solution is $L = \prod_{j=1}^{m} [\sum_x C_j(X,racine) \cdot P(x)]$

Complexity: for n sequences, m positions, and k possible values (k=4 for DNA sequences), the algorithm is $O(m \cdot k^2)$ on O(n) nodes. Therefore, the time of computation is $O(n \cdot m \cdot k^2)$.

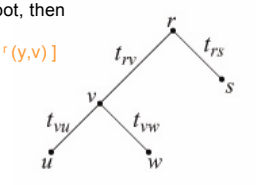
Optimal branch-length calculation for a given tree topology

The previous formula demands the branch lengths of the tree to be known.
In practice this is not usually the case.

The tree branching length is computed by maximizing the log-likelihood function.
Suppose to know all lengths except t_{rv}. If r is the root, then

$$\log L = \sum_{j=1}^{m} \log [\sum_{x,y} P(x) \cdot C_j^{v}(x,r) \cdot P_{x \to y}(t_{rv}) \cdot C_j^{r}(y,v)\,]$$

$C_j^{u}(X,y)$ in a tree where u is the root



We have to maximize log(L) wrt t_{rv}. To do this, people use numerical methods like the Newton's methods or other numerical methods (EM) and often the result depends on the numerical method.

Remark: we can try to optimize branch length one after the other. This approach works fine. After some steps in the tree, typically the likelihood converges, and the result is a phylogenetic tree close to the optimal.

Finding a maximum likelihood tree

Search space is impossibly large when n grows

$(2n - 3)!!$ possible rooted trees

$(2n - 5)!!$ possible unrooted trees

When computing the maximum likelihood tree, the model parameters and branch lengths have to be computed for each tree, and then the tree that yields the highest likelihood has to be selected.

Reasonable trees are suggested by heurystics.

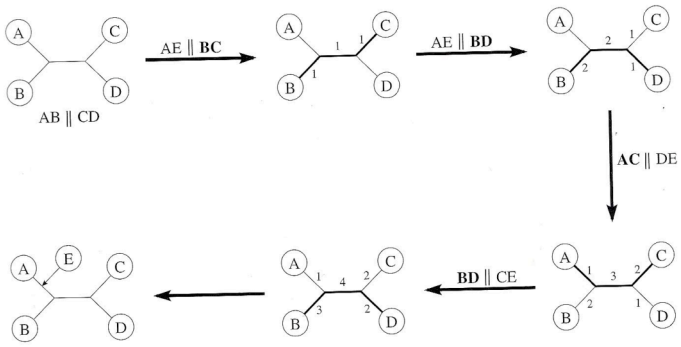
We present the quartet-puzzling algorithm.

Given a set of n aligned nucleotide sequences, any group of 4 of them is called a quartet. The quartet-puzzling algorithm analyzes all possible quartets in a dataset, taking advantage of the fact that for a quartet, just three unrooted tree topologies are possible.

The algorithm is a three step procedure:

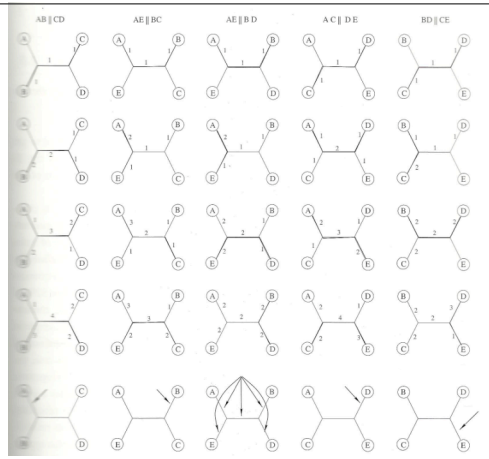
-the maximum-likelihood step - it computes for each of the C^n_4 possible quartets the maximum-likelihood values L_1, L_2, L_3 for the three possible 4-sequence trees. 3 C^n_4 likelihoods result from this step.

-the quartet-puzzle step - to compute an intermediate tree by inserting sequences sequentially in an already reconstructed subtree.



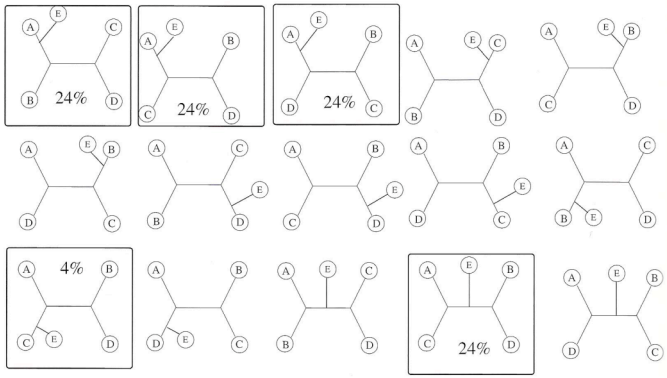
The original quartet is tested against all possible quartets defined by 3 letters (out of the 4: A,B,C,D) and E, and penalties are added to the original branches.

Sequence E is inserted at the branch with minimal penalty.



All intermediate trees for a small example of 5 sequences.

For large data, it is not feasible to compute all intermediate trees and the algorithm is repeated at least a 1000 times for various input orders of sequences to avoid reconstruction artifacts due to the ordering of the sequences and to get a representative collection of trees.



Frequence of appearance for tree-topologies occurring in the case of 5-sequences.

-majority rule consensus : computed from the resulting intermediate trees.

This step provides information about the number of times a particular grouping occurred in the intermediate trees. This reliability value measures (in %) how frequently a group of sequences occurs among all intermediate trees.

All groups that occur in >50% of the collection of intermediate trees are represented in the majority-rule consensus tree. This tree is not necessarily the maximum likelihood tree.

Cluster	T-1	T-2	T-3	T-4	T-5	Total frequency
(A, B)	0	0	0	0	0	0
(A, C)	0	0	0	0	24	24
(A, D)	0	0	0	0	0	0
(A, E)	24	24	24	0	0	72
(B, C)	0	0	24	0	0	24
(B, D)	0	24	0	4	24	52
(B, E)	0	0	0	0	0	0
(C, D)	24	0	0	0	0	24
(C, E)	0	0	0	4	0	4
(D, E)	0	0	0	0	0	0

Outcome of the computation of the majority-rule consensus tree for the five sequence example (in %)

Problèmes avec la reconstruction phylogénétique

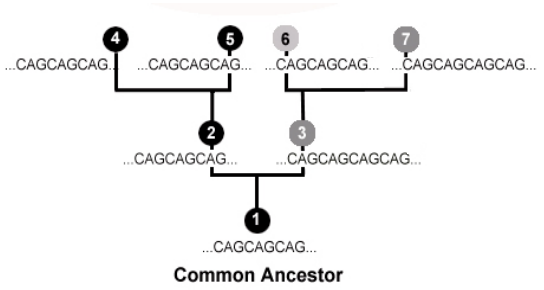
- Important: une seule reconstruction phylogénétique fourni très souvent une image incomplète de la réalité.
- Quand plusieurs méthodes (parcimonie, basés sur la distance, probabilistes) donnent le même résultat, alors il est plus probable d'avoir une réponse correcte.

1. Homoplasie

- Given:
  - 1: ...CAGCAGCAG...
  - 2: ...CAGCAGCAG...
  - 3: ...CAGCAGCAGCAG...
  - 4: ...CAGCAGCAG...
  - 5: ...CAGCAGCAG...
  - 6: ...CAGCAGCAG...
  - 7: ...CAGCAGCAGCAG...
- Les séquences 1, 2, 4, 5 et 6 semblent avoir évolues à partir d'un ancêtre commun, avec une insertion qui amène à la présence de 3 et 7

Homoplasie

- Mais si l'arbre vrai était celui-ci ?

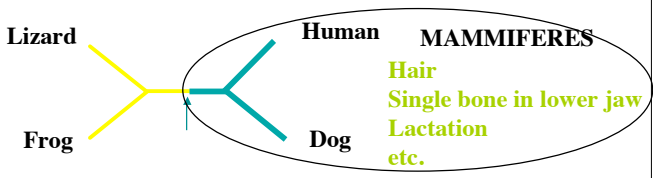


Homoplasie

- 6 a évolué séparément par rapport à 4 et 5, mais la parsimonie grouperait 4, 5 et 6 ensemble comme ayant évolués d'un ancêtre commun.
- Homoplasie: évolution indépendante (ou parallèle) d'un même/similaire caractère.
- Les résultats de parsimonie **minimisent** l'homoplasie, de telle façon que si l'homoplasie est fréquente, la parsimonie peut donner des résultats très erronés.

1. Caractères contradictoires: les queues

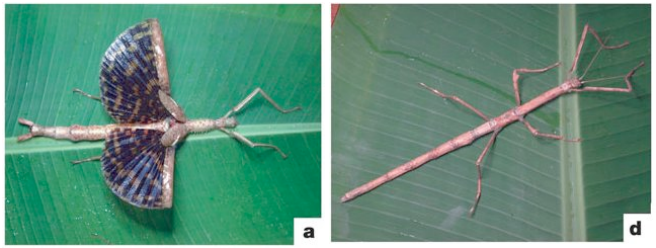
Un arbre phylogénétique a une plus forte probabilité d'être correcte quand il est soutenu par plusieurs caractères, comme vu dans l'exemple:



- **Note: dans ce cas, les queues sont homoplastiques**

2. Combien de fois l'évolution a re-inventé les ailes?

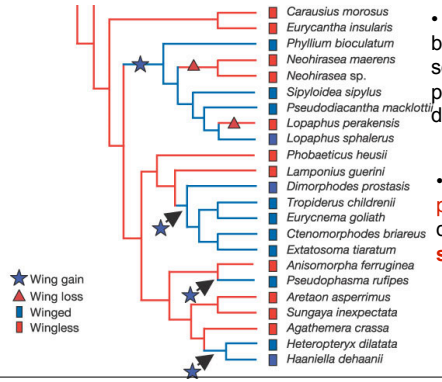
- Whiting, et. al. (2003) a regardé aux insectes avec et sans ailes.



Réinventer les ailes

- Etudes précédents ont montré des transitions "avec → sans" ailes
- Les transitions "sans → avec" ailes sont beaucoup plus compliquées (elles demandent le développement de plusieurs chemins biochimiques nouveaux)
- La reconstruction de plusieurs arbres phylogénétiques a amené à établir toujours une re-évolution des ailes.

Arbre phylogénétique le plus parcimonieux des insectes avec et sans ailes

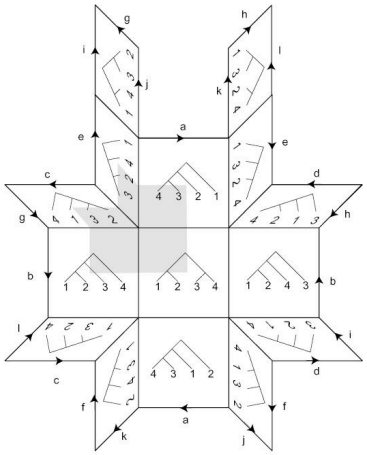


- arbre phylogénétique basé sur les **deux :** séquences d'ADN et présence/absence d'ailes
- la reconstruction la plus parcimonieuse donne un **ancêtre sans ailes**

Pourquoi les insectes sans ailes volent encore?

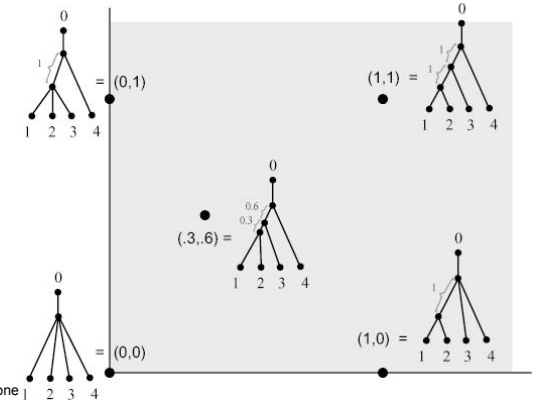
Car les reconstructions phylogénétiques les plus parsimonieuses demandent une re-invention des ailes, il est possible que les chemins de développement des ailes soient conservés dans les insectes sans ailes.

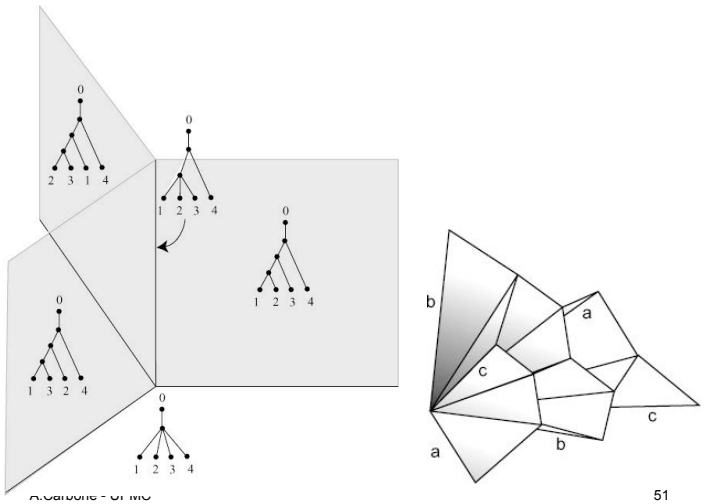
Géométrie de l'espace des arbres phylogéniques

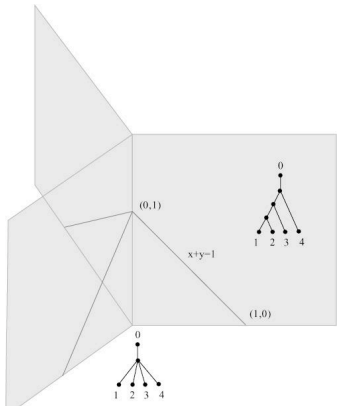


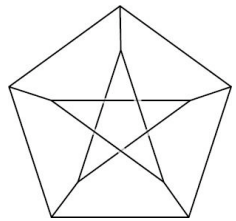
Arbres de 4 feuilles
et leur transformation

Arbres métriques: avec noeuds étiquettés









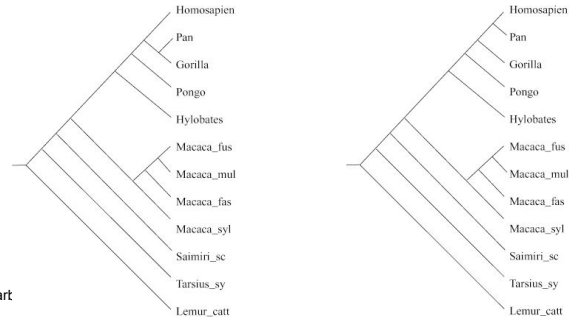
Graphe de Peterson

Propriétés de cet espace

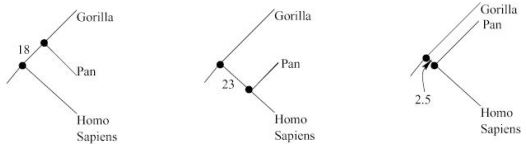
- Il s'agit d'un espace CAT(0), à curvature non-positive.
- Il y a toujours une géodesique entre chaque paire de points.
- Dans un espace CAT(0) les **centroides** existent pour chaque ensemble fini de points, et la fonction centroide est convexe.

Exemple

```
'Lemur_catta'      AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCCA...
'Tarsius_syrichta'AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCTCCC...
'Saimiri_sciureus'AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGTCTA...
'Macaca_sylvanus' AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTTCCA...
'Macaca_fascicul.'AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTTCCA...
'Macaca_mulatta'   AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTTCCA...
'Macaca_fuscata'   AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTTCCA...
'Hylobates'        AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCC...
'Pongo'            AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCC...
'Gorilla'          AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCAT...
'Pan'              AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCAT...
'Homo_sapiens'     AAGCTTCACCGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCAT...
```



Centroide de deux arbres phylogénétiques



Références

[1] L.Billera, S.Holmes, K.Vogtmann, Geometry of the space of phylogenetic trees, *Advances in Applied Mathematics*, 27:733-767, 2001.

[2] H. L. Bodlaender, M. R. Fellows, and T. J. Warnow. Two strikes against perfect phylogeny. In *Proc. 19th*. Springer, 1992.

[3] C. Bron and J. Kerbosch. Finding all cliques of an undirected graph. *Communications of the Association for Computing Machinery*, 16:575–577, 1973. Algorithm 457.

[4] W. H. E. Day. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49:461–467, 1986.

[5] J. Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22:240–249, 1973.

[6] J. Felsenstein. Phylogenies from molecular sequences: inference and reliability. *Annuals Rev. Genetics*, 22:521–565, 1988.

[7] J. Felsenstein. *Inferring Phylogenies*. ASUW Publishing, Seattle, WA, 1998.

[8] W. M. Fitch. Toward defining course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, 20:406–416, 1971.

[9] L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3:43–49, 1982.

[10] Jr. G.F. Estabrook, C.S. Johnson and F.R. McMorris. An algebraic analysis of cladistic characters. *Discrete Mathematics*, 16:141–147, 1976.

[11] D. Gusfield. The steiner tree problem in phylogeny. Technical Report 334, Yale University, Computer Science Dep., 1984.

[12] Dan. Gusfield. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, 1997.

[13] M. D. Hendy and D. Penny. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, 60:133–142, 1982.

[14] T. H. Jukes and C. Cantor. *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[15] S. Kannan and T. Warnow. Inferring evolutionary history from dna sequences. *SIAM J. Comput.*, 23:713–737, 1994.

[16] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Molecular Evolution*, 16:111–120, 1980.

[17] W. H. Li. *Molecular Evolution*, chapter 5, pages 105–112. Sinauer Associates, Inc., Publishers, Sunderland, Massachusetts, 1997.

[18] T. Yano M. Hasegawa, H. Kishino. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Molecular Evolution*, 22:160–174, 1985.

[19] C. D. Michener and R. R. Sokal. A quantitative approach to a problem in classification. *Evolution*, 11:130–162, 1957.

[20] A. Krogh G. Mitchison R. Durbin, S. Eddy. *Biological Sequence Analysis*, chapter 7, pages 160–191. Cambridge University Press, Cambridge, United Kingdom, 1998.

[21] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[22] D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics*, 28:35–42, 1975.

[23] M. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *J. of Classification*, 9:91–116, 1992.

[24] A. Wilson and R. Cann. The recent african genesis of humans. *Scientific American*, April, 1992.

[25] E. O. Wilson. A consistency test for phylogenies based on contemporaneous species. *Systematic Zoology*, 14:214–220, 1965.