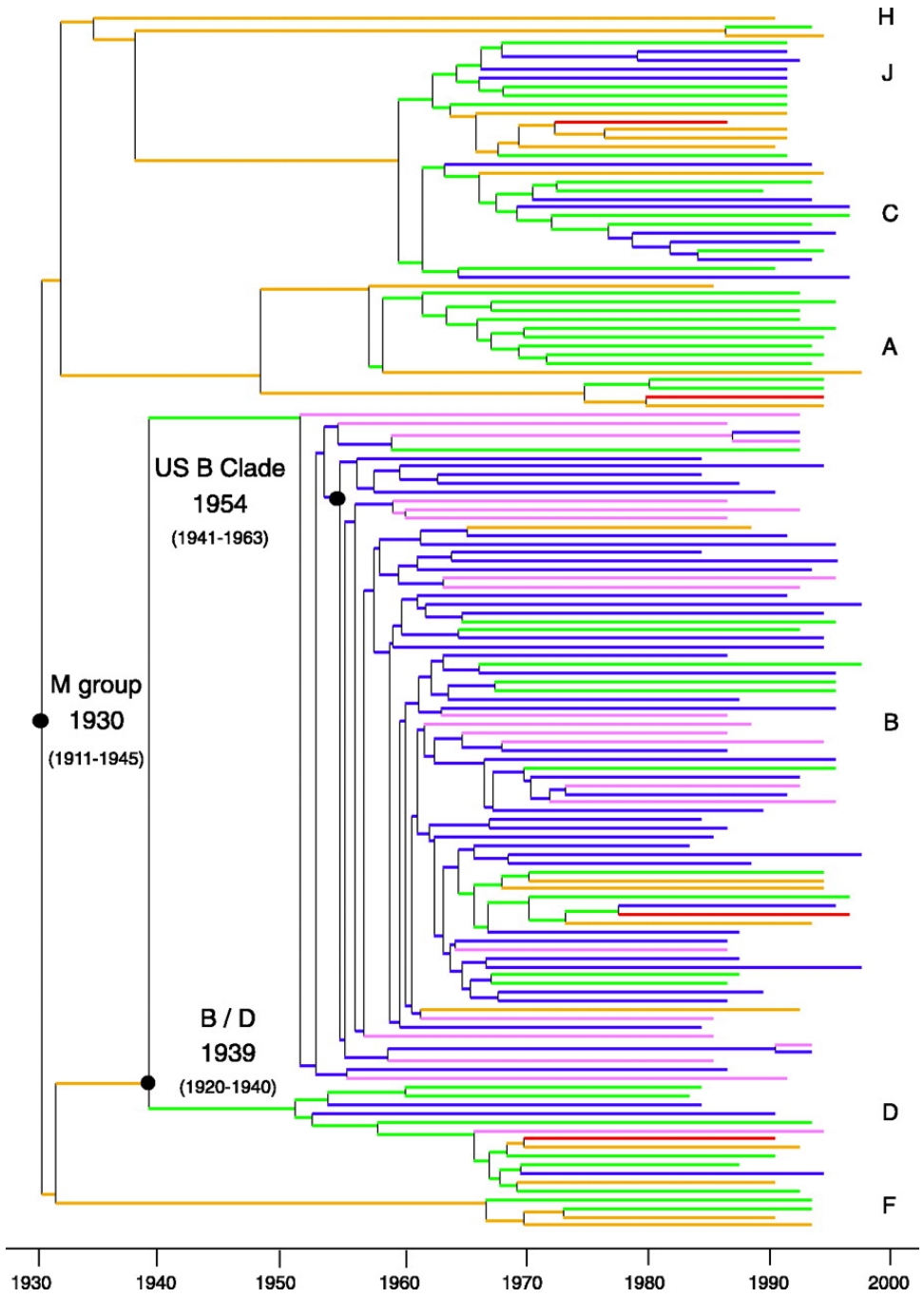
Web References for HIV

The HIV/SIV jump problem, when did it occur?

Quest for the Origin of AIDS

Did Modern Medicine Spread an Epidemic?

The river without a paddle


H
J
C
A
US B Clade
1954
(1941-1963)
M group
1930
(1911-1945)
B
B / D
1939
(1920-1940)
D
F
1930
1940
1950
1960
1970
1980
1990
2000

Souvent ces modeles utilisent la notion d'horloge moleculaire. Ayant un processus homogene dans le temps, le nombre de mutation sera proportionnel au temps passé.

Molecular Clock : Horloge Moléculaire

Phylogénies

L'idée de la representation des relations d'apparente par des arbres est vieille de plusieurs milliers d'annees , du temps des Grecs. Les linguistes aussi utilisent des arbres pour elaborer l'histoire des langues et leurs relations.

Leurs arbres ne presentent pas la meme symmetrie entre feuilles soeurs.

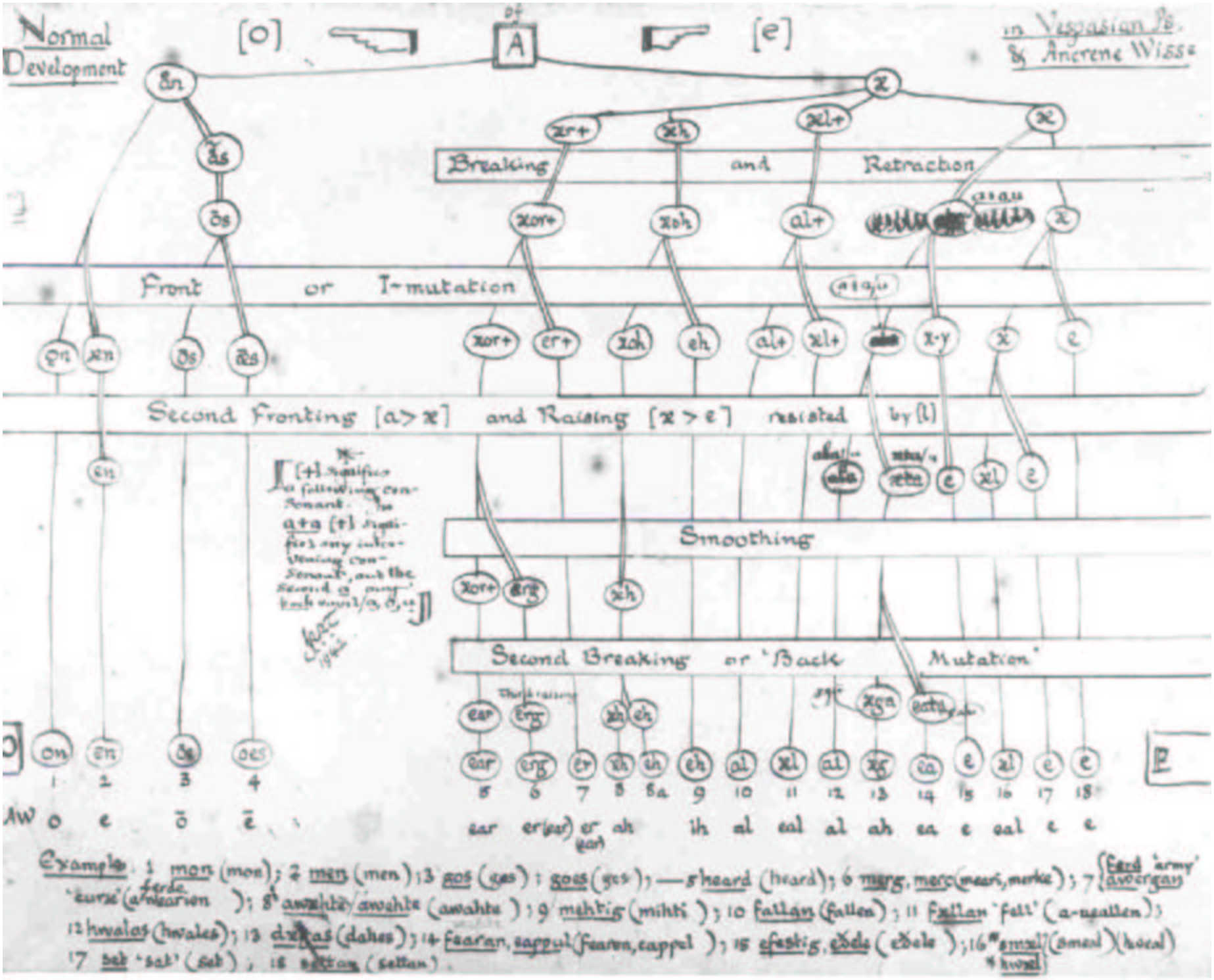


Chart of vowel-changes, 1942

Voisins sur l'arbre partagent le meme ancetre. Les caracteres derives

avec les memes ancetres sont dits homologues. (quelquefois appeles IBD)

Une distinction importante entre similarité et homologie.

Les soeurs definissent un ancetre commun ou groupe monophyletique. C'est souvent le but d'etudes de phylogenies que de trouver ces groupes monophyletiques.

Pendant 200 ans les biologists ont construit leurs arbres avec des caracteres morphologiques. [1] data.

L'explosion des données genetiques disponibles par les methodes de sequencage rendent la construction d'arbres tres faciles, et tout le monde s'en sert.

Voici quelques questions interessantes a aborder:

[1] data about presence or absence of wings, sepals, hair, nodules,...

- Les modeles parametriques sont elles preferables au modeles non parametriques.

- Quels sont les meilleurs codages des données?

- Est-ce qu'il ne faut pas changer le poids de certains caracteres qui ont des conflits avec l'alignement ou l'arbre?

- Quelles methodes utilisees pour la validation d'arbres?

- Quelle parametrization, est-ce que les methodes sont consistents, identifiable, robuste?

- Comment combiner des donnees de genes differents en un seul arbre?

- Comment incorporer l'information a priori?
  Le dilemme Bayesien?

Les programmes, logiciels, resources seront disponibles a:
`http://www-stat.stanford.edu/~susan/courses/BIMM/BIMM.html`

Sur quelles données nous basons nous pour construire l'arbre?

Données moleculaires, acides aminés ou nucleotides.

La premiere etape est la construction d'un alignement des sequences.

Quatre ecoles principales de construction/estimation

- Maximum de vraisemblance.
- Méthodes basés sur les distance.
- Méthodes basés sur la parcimonie methods.
- Méthodes Bayesiennes.

Données Alignées

```
      21   383
VVi       M-SGTAGQVICCKAAVAWEAGKPVIEEVEVAPPQAMEVRLKILYTSLCH
Zma1      M--ATAGKVIKCKAAVAWEAGKPSIEEVEVAPPQAMEVRVKILFTSLCH
Zma2      M--ATAGKVIKCRAAVTWEAGKPSIEEVEVAPPQAMEVRIKILYTALCH
Hvu1      M--ATAGKVIKCKAAVAWEAGKPTMEEVEVAPPQAMEVRVKILFTSLCH
Hvu2      M--ATAGKVIKCKAAVAWEAGKPSMEEVEDAPPQAMEVRDKILYTALCH
Hvu3      M--ATAGKVIKCKAAVAWEAGKPSIEEVEVAPPQAMEVRVKILYTALCH
Tae       M--ATAGKVIECKAAVAWEAGKPSIEEVEVAPPHAMEVRVKILYTALCH
Osa1      M--ATAGKVIKCKAAVAWEAGKPSIEEVEVA--KEMEVRVKILFTSLCH
Osa2      M--AT-GKVIKCKAAVAWEAGEASIEEVEVAPPQRMEVRVKILYTALCH
Ath       M-S-TTGQIIRCKAAVAWEAGKPVIEEVEVAPPQKHEVRIKILFTSLCH
Psa       M-SNTVGQIIKCRAAVAWEAGKPVIEEVEVAPPQAGEVRLKILFTSLCH
Fan       M-SSTEGKVICCRAAVAWEAGKPVIEEVEVAPPHPNVVRVKILYTSLCH
Tre       M-SNTAGQVIKCRAAVAWEAGKPVIEEVEVAPPQAGEVRLKILFTSLCH
Stu       M-STTVGQVIRCKAAVAWEAGKPVMEEVDVAPPQKMEVRLKILYTSLCH
Pgl       M-A-TAGKVIKCKAAVAWEAGKPSIEEVEVAPPQAMEVRVKILYTSLCH
Phy       MSSNTAGQVIRCKAAVAWEAGKPVIEEVEVAPPQKMEVRLKILFTSLCH
Pde       M-SSTVGKVIRCKAAVAWEAAKPSIEEVEVAPPQANEVRLRILFTSLCH
Pta       MASSTAGQVIKCKAAVAWAAGEPKIEEVEVAPPQAMEVRVKIHYTALCH
```

```
Fra      M-SSTEGKVICCRAAVAWEAGKPVIEEVEVAPPQANVVRVKILYTSLCH
Mal      M-SNTAGQVIRCRAAVAWEAGKPVIEEVEVAPPQANEVRIKILFTSLCH
Lyc      M-STTVGQVIRCKAAVAWEAGKPVMEEVDVAPPQKMEVRLKILYTSLCH
```

Transformations

1. Pour le maximum vraisemblance, toute la matrice des sequences est employee. Meme les colonnes qui sont conservees: celles ci servent la contruction des frequences de chaque character et de l'estimation du taux de mutations.

2. La methode de parcimonie n'emploient que les caracteres informatives:

   Les sites sont dites informatives si elles permettent de distinguer entre des arbres possibles. Les sites qui montrent pas de differences (completement conservees) ou seulement une difference pour une espece ou taxon ne sont pas informatives.

3. Les methodes basees sur les distances ont une strategie intermiediaire. Une premiere etape consiste dans le calcul des

parametres de taux de mutation et du nombre de changements de types differents. Apres cela seules les distances calculees sont utilisees.

Arbre de gènes, arbre des especes

L'information contenue dans un gène ne produira de l'information que sur l'histoire de ce gène, l'arbre en question sera apelé 'gene-tree'.

La combinaison des arbres gènes en arbre espèce est un probleme ouvert. On utilise en general des méthodes du consensus.

Le modele de substitution

Pour commencer considerons le cas ou les donnees sont des nucleotides ADN: **purines ('A', 'G')** et **pyrimidines ('T', 'C')**. Plusieurs types de modeles de substitution , le plus simple s'appelle Jukes-Cantor: toute mutation a la meme probabilite.

Que ce soit une (**transversion**), for instance from purines to pyramidines within each type, (**transition**), for instance from purines to purines. The rate matrix Q is of the form:

$$Q = \begin{array}{c|cccc} & A & T & C & G \\ A & -3\alpha & \alpha & \alpha & \alpha \\ T & \alpha & -3\alpha & \alpha & \alpha \\ C & \alpha & \alpha & -3\alpha & \alpha \\ G & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

Le modele a deux parametres s'appelle le modele de Kimura

$$Q = \begin{array}{c|cccc} & A & T & C & G \\ A & -\alpha-2\beta & \beta & \beta & \alpha \\ T & \beta & -\alpha-2\beta & \alpha & \beta \\ C & \beta & \alpha & -\alpha-2\beta & \beta \\ G & \alpha & \beta & \beta & -\alpha-2\beta \end{array}$$

Le modele le plus complet (GTR) est de la forme

$$Q = \begin{array}{c|cccc} & A & T & C & G \\ A & - & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} \\ T & \alpha_{2,1} & - & \alpha_{2,3} & \alpha_{2,4} \\ C & \alpha_{3,1} & \alpha_{3,2} & - & \alpha_{3,4} \\ G & \alpha_{4,1} & \alpha_{4,2} & \alpha_{4,3} & - \end{array}$$

La matrice de substitution donne la probabilite de changement d'un nucleotide pendant le temps t par une chaine de Markov a generateur infitessimal Q.

Dans le cas des aa on obtiendrait une matrice (20×20 instead of 4×4), mais tous les calculs sont semblables.

Distance based methods

Variants of hierarchical cluster analysis.

The aim is to reconstruct the distances as computed between the two sequences of the two species x and y by distances along the edges of the tree forming a path between x and y.

First a distance matrix is constructed between the N units in some way. These distances d_{xy} are supposed to estimate the unknown 'true evolutionary' distances between x and y as they would be measured along the unknown true tree $\mathcal{T}$.

For the Jukes-Cantor model which assumes equal rates of substitution between all base pairs provides the estimate of distances between sequences x and y as:

$$d_{xy} = -\frac{3}{4}\log(1 - \frac{4}{3}(1 - (\frac{\#AA}{k} + \frac{\#CC}{k} + \frac{\#GG}{k} + \frac{\#TT}{k})))$$

where k denotes the number of characters (columns) in the data matrix, and $\#AA$ denotes the number of times there is an A in x matched with an A in y.

Once the distances are decided upon, the parametric model is left behind and a clustering technique such as hierarchical clustering with average groups is used to find the tree from the distances.

Remarks:

If we knew the true evolutionary distances between species, we could build an additive tree that reproduced the distances along the tree in a unique way.

The existence of an additive tree reproducing the distances faithfully is not always ensured, a sufficient condition for this to be possible is called the **four point condition**(for all quadruples):
$d_{AB} + d_{CD} \leq max(d_{AC} + d_{BD}, d_{AD} + d_{BC})$.
This means that one of the two sums is minimum and the other two are equal. Notice that this is not the same as the ultrametric property which says that for any three points: A, B, C:

$$d_{AC} \leq max(d_{AB}, d_{BC})$$

If the distances obey the ultrametric property the distances can be fit to a binary tree with leaves equally distant from the root.

Unfortunately distances computed from real data never obey this property.

Additivity is destroyed by:

- Homoplasy (reversal, parallelism and convergence) which is caused by superimposed changes.
- An uneven distribution of change rates.
- Measurement error.
- **Paralogous** sequences.

We concentrate on distances that are computed from substitution models such as Jukes and Cantor's one-parameter model, Kimura's two-parameter model, or even the complex 12-parameter model for the substitution matrices. These models provide estimates of

differences between sequences computed from the frequencies of various changes in the sequences.

Parsimony method

Nonparametric procedures. Farris (1983), has a justification for parsimony : "minimizes requirements of ad hoc hypotheses of homoplasy".

Analogy is made between homoplasies and residuals, (part of the data that the tree does not explain), minimizing homoplasies is akin to minimizing residuals in regression.

Roughly this method can be seen as based on the assumption that "evolution is parsimonious" which means that there should be no more evolutionary steps than necessary.

Thus the best trees are the ones that minimize the number of changes between ancestors and descendants. Under independence of each of the characters, this has a clear combinatorial translation.

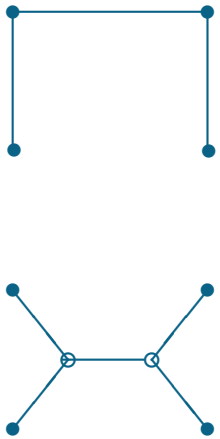
The parsimony tree as a combinatorial problem

Unrooted parsimony trees.

Recall that the Hamming distance between two units is the number of changes needed to bring one to the other. This assumes that all changes in a categorical character are counted as one step.

$$\mathrm{d_H}(AACTGGG, AACTGGC) = \mathrm{d_H}(AACTGGG, AACTGGA) = 1$$

Here, given N points in a metric space, the Steiner problem is that of finding the shortest tree connecting the N points where one is allowed to add extra vertices. Thus, with 4 points arranged at the vertices of a unit square, one would add a fifth point in the center to form the Steiner tree.



The minimum spanning tree andthe Steiner tree of the 4 vertices of a rectang

Although statisticians are not familiar with minimal Steiner trees, they may have encountered minimal spanning trees as used by Friedman and Rafsky (1985). The relation between the two is well explained in Gardner's wonderful chapter on Steiner trees (Chapter 22, Gardner (1997)). He explains how minimal spanning trees are good "starting points" since in the plane for instance they can only be 13% longer than Steiner trees.

As a combinatorial problem, the maximum parsimony tree is the problem of finding the Steiner points or Steiner tree for Hamming

distance between the units, under the constraint that the tree be binary.

The problem of finding a minimal Steiner tree is that of finding the Steiner points (representing ancestors) that minimize the complete length of the tree. Steiner points are points that are added to a graph so that its minimal spanning tree becomes shorter.

Computation issues

The minimal Steiner tree problem is NP-hard, meaning that no algorithm is known that will compute an optimal tree in polynomial time in the number of species N.

Much work has been done to implement good heuristic algorithms for finding approximately optimum trees. Swofford's `PAUP`, Felsenstein's `Phylip`, and Goloboff's `NONA` all contain clever use of branch and bound techniques and branch swapping to find acceptable answers.

N=500 can now be done routinely.

Parsimony as a statistical procedure

Felsenstein (1983) lists parsimony in a section entitled a section on parsimony as "non-statistical approaches". Farris says (1983) says the "statistical approach to phylogenetic inference was wrong from the start, for it rests on the idea that to study phylogeny at all one must first know *in great detail* how evolution has proceeded". Both these authors identify statistics with parametric modeling.

Many data-analytic procedures such as correspondence analysis, projection pursuit, neural nets, classification and regression trees (CART) and minimal spanning trees have proved that complex situations can be satisfactorily understood by heuristic procedures before any theoretical framework supposing a probabilistic model justifies their properties (Diaconis and Efron (1984)).

Example of data set where the tree was known

T7 data experimentally generated phylogeny, Hillis et al. (1992) In `phylip` form (and only 'informative' sites):

```
    9  21
R           C C G C C G G C C G G C C A G C G G G G T
J           C C C C G T A C C G G T C A A C G G G G T
K           T C C C G C A C C G A T C A A T G G G G G
L           T C C C G C A C C G A T C A A T G G G G G
M           C T C C G T A C C G G T C A A C G G G G T
N           C C T T A C G T T A G C T G G C A A A A T
O           C T C C G C G C T G G C C G G C A G A A T
P           C C C C A C G C T G G C C G G C A G A A T
Q           C C T T A C G T T A G C T G G C A A A A T
```

If the data set is put into a file called `infile` it will automatically be processed by any `phylip` program that is called. Otherwise if there is no current infile, `phylip` will ask for a file name, then there is a

dialogue menu that allows the user to specify all the options.

Maximum Parsimony Tree

This is part the output from the `phylip` command `dnapars`:

```
One most parsimonious tree found:
                +--------0
     +----------6
     !          !  +-----P
     !          +--7
     !             !  +--Q
     !             +--8
  +--5                +--N
  !  !
  !  !                +--L
  !  !          +-----3
  !  !          !     +--K
--1  +----------2
  !             !     +--M
  !             +-----4
  !                   +--J
  !
  +----------------------R
  remember: this is an unrooted tree!
```

```
requires a total of      25.000
 steps in each site:
      0   1   2   3   4   5
   *-----------------------------
  0!       1   2   2   1   2
 10!   1   1   1   1   1   1
 20!   1   1
```

Output: the Newick notation

The output file called `treefile` contains the following line (the tree in parentheses format):

```
(((O,(P,(Q,N))),((L,K),(M,J))),R);
```

Rooting the Tree

At least one of the taxonomic units has a special function. For a statistician it would be seen as a simple outlier: the biologists voluntarily include what they call an **outgroup** to locate the root of the tree. The root is situated by creating an unrooted tree and the edge that joins the outgroup to the other species will be the support for the root. This is a clever use of prior information that simplifies the problem considerably, (by a factor of $(2\mathrm{N}-3)$). What is less obvious to the outsider is why, once the root's position is decided upon, the biologists keep the outgroup in the data set - it seems to distort the image of the closer group (called the **ingroup**), in fact outgroups also provide information on the root's characters, and so on the ancestral states of the character. This seems to be a security check, if in fact the outgroups become misplaced or lost in the tree, then there are signs of trouble. Many methods have trouble as soon

as 2 very different outgroups are present (this is named the **long branch attraction problem**), just as in regression two opposite outliers can completely redefine the regression line.

Homoplasy

A character change may become invisible through time, because there has been a **reversal** or **back-substitution** for instance:

$$A \longrightarrow G \longrightarrow A.$$

There are also changes of exactly the same type that appear in different parts (clades) of the tree, giving a false impression of similarity. This is called **parallelism**.

Another variant is substitutions that occur in different clades but have the same results:

$$\left.\begin{array}{r} A \longrightarrow G \longrightarrow A \\ A \longrightarrow C \longrightarrow T \longrightarrow A \end{array}\right\} \textbf{convergent substitutions.}$$

The effect on the resulting measurements of differences between units are the same: there is an error; units appear to be more similar than

they would be if the complete history were known. Collectively these are called **homoplasy**.

Parametric models that take homoplasy into account are the motivation for the 'modified evolutionary distance' computations. Whether they include 1 or 12 parameters they try to retrieve some of the variability lost through homoplasy. Some authors feel that this possibility of error-correction in parametric methods is so essential that it justifies using such models even when they have not been proved to fit the actual phenomenon.

Parsimony methods are sometimes limited to shorter stretches of time to limit the homoplasy; 'long branches' are undesirable in parsimony methods.

Maximum vraisemblance trees

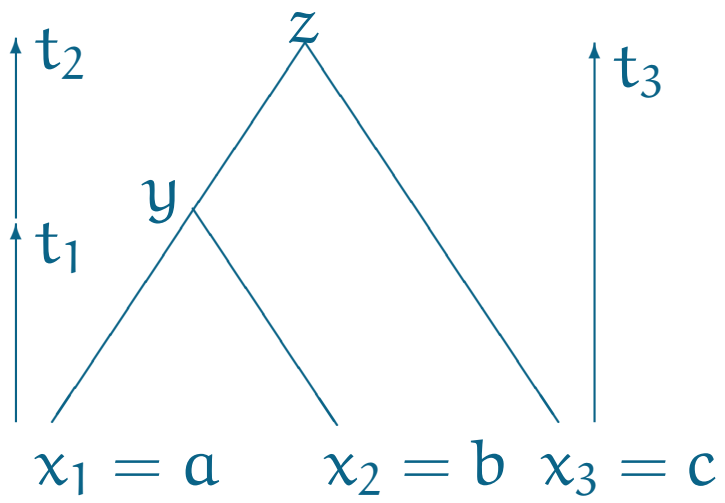
For a statistician this is the easiest of the methods to understand. A parametric model (θ, T) is postulated, θ is a η-dimensional vector that we explain below and $\mathcal{T}$ is the tree's topology. Under this model the vraisemblance for each possible tree $\mathcal{T}$ is separately computed for each character, the independence of characters then allows the total vraisemblance of the tree for all data to be computed by taking the product.

The first part of the vector of parameters θ comes from the Markovian substitution model as explained before.

The number of other parameters that have to be specified depends on the complexity of the model. If a molecular clock [2] is postulated, speciation times $\{t_1, t_2, ...t_{N-2}\}$ (splitting events) are the other

[2] branch lengths in evolutionary change depend linearly on time

parameters. Otherwise both the branch lengths $\{v_1, v_2, ...v_{N-2}\}$ and the different rates along those branches have to be parametrized.



The substitution parameters are estimated from the data. A complete model including distributions of separation events is postulated and the vraisemblance can be computed for each possible tree by computing the vraisemblance of the tree for each site $X_{.j}$:

$$f(X_{.j}|\theta_1, \theta_2, \ldots, \theta_\eta, \mathcal{T}).$$

This actually requires computing the vraisemblance of all the

subtrees, so the method is recursive.

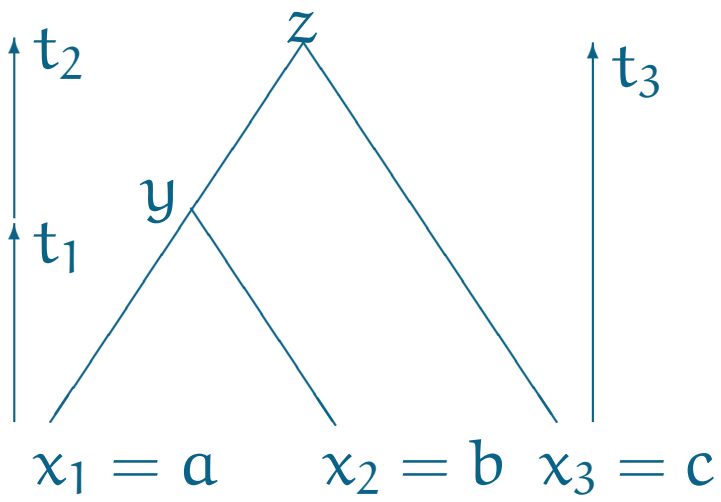
$$\mathcal{L}(\theta_1, \theta_2, \ldots, \theta_\eta | X_{.1}, X_{.2}, \ldots, X_{.k}, \mathcal{T}) = \prod_{j=1}^{k} f(X_{.j} | \theta, \mathcal{T})$$

The essential assumptions:

1. Each site in the sequence evolves independently.

2. Different lineages evolve independently.

3. Each site undergoes substitution at an expected rate (can be extended to a series of rates with a given distribution).



Vraisemblance:P(data|Tree,t's,ancestors,mutation rates). Based on the probabilities computed given the tree and for potential ancestors ($t_3 = t_1 + t_2$)

$$\begin{aligned} P(a, b, c, y, z|T, t) &= P(a|y, t_1)P(b|y, t_1)P(c|z, t_3)P(y|z, t_2)P(z) \\ P(a, b, c, |T, t) &= \sum_z \pi_z P_{zc}(t_3) \sum_y P_{zy}(t_2)P_{ya}(t_1)P_{yb}(t_1) \end{aligned}$$

This is a function of t_1, t_2 whose values are estimated as the maximum for a given tree topology, then for the ml estimate is made for each T. The T with the maximum value is the maximum

vraisemblance estimate.
We can consider the vraisemblance computation, one character at a time. Starting from the root, or starting from the leaves, Felsenstein's transversal method starts from the leaves, we abbreviate the character we are interested from x_{ij} to x_i. For two leaves with the residue a at their common ancestor, (the root here) :

$$P(x^1, x^2, a|\mathcal{T}, \theta_1 = t_1, \theta_2 = t_2) = \pi_a P(x^1|a, \theta_1) P(x^2|a, \theta_2)$$

The root is an unknown nuisance parameter that we integrate out:

$$P(x^1, x^2|\mathcal{T}, \theta_1 = t_1, \theta_2 = t_2) = \sum_a \pi_a P(x^1|a, \theta_1) P(x^2|a, \theta_2)$$

Call $m[i]$ the direct parent of i, and $P(L_i|a)$ denote the probability of all nodes below i given that the node i is a. We number the inner nodes from (n+1) to (2n-2), these ancestral nodes are all unknown, so we have to sum the probabilities of all their possible assignments to

compute the complete vraisemblance of the tree, given its edge lengths $(\theta_1, \theta_2, \ldots, \theta_{2n-2})$.

The algorithm is similar to the forward algorithm in HMM (see chapter 3), we are going to sum over possible paths, working upwards from the leaves.

Compute $P(L_j|e), P(L_k|f)$ for all e and f at daughter nodes j, k of i

$$P(L_i|a) = \sum_{b,c} P(b|a, t_j) * P(L_j|b) * P(c|a, t_k) * P(L_k|c)$$

We can write down the complete probability as a sum. We denote the alphabet of possible residuals $\mathcal{A}$,

$$P(x^1, x^2, \ldots, x^{(2n-2)}|\mathcal{T}, \theta)$$

$$= \sum_{(a^{n+1},\ldots,a^{2n-1})\in\mathcal{A}^{n-2}} \pi_{a^{2n-1}} \prod_{n+1}^{2n-2} P(a^i|a^{m[i]}, \theta_i) \prod_{1}^{n} P(x^i|a^{m[i]}, \theta_i)$$

the computational algorithm evaluates $P(L_i|a)$ for the children j and k such that $m[j] = m[k] = i$, we compute $P(L_j|b)$ and $P(L_k|c)$ for all possible b and c.

These instructions allow us to compute the vraisemblance of any tree, given its branching order (sometimes called topology) and its branch lengths. For the maximum vraisemblance computation, we need to compute the tree that maximises the vraisemblance, first for a given branching order, find the branch lengths that maximise the vraisemblance. This can be done by taking the derivative $\frac{\partial P(x^j|x^{m[j]},\theta_k)}{\partial \theta_j}$ in order to use the conjugate gradient method for optimising the edge lengths, or we can take an EM approach as Felsenstein, 1981 suggests and implemented in his `phylip` program.

Finding the vraisemblance of one tree is an NP complete problem[3], thus as we need to look at all the topologies, of which there are exponentially many; we see the exact computation becomes quickly

[3] There is no known polynomial time algorithm that finds the tree with maximum vraisemblance.

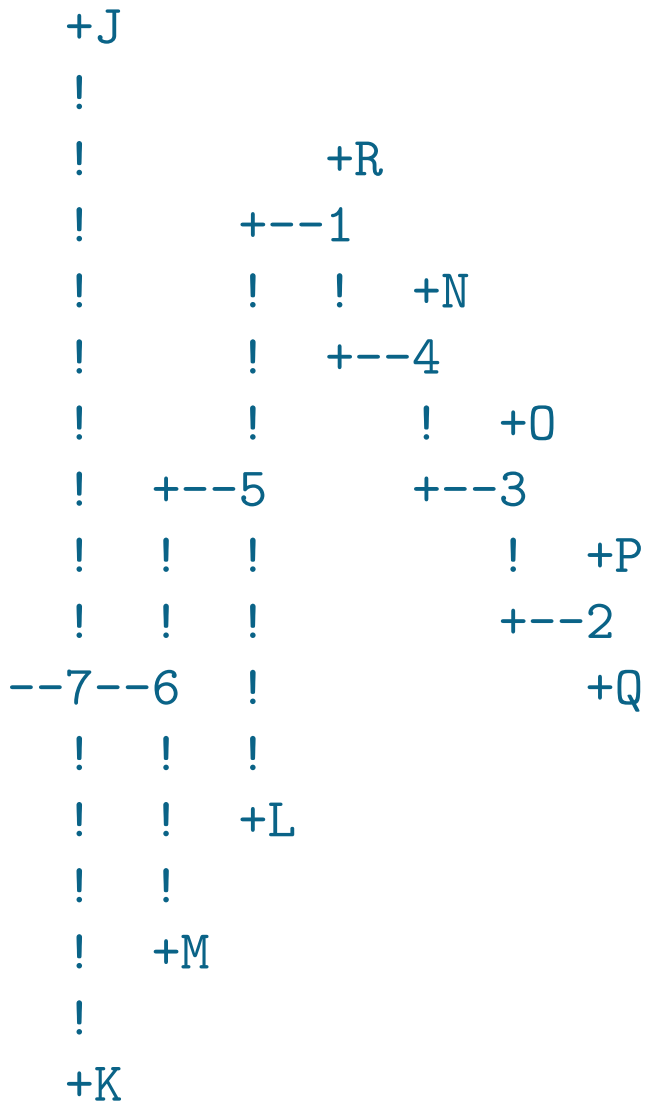
intractable as the number of leaves increases.

Most of the implementations actually use randomised optimisation procedures, where one is not ensured to obtain the optimal solution, there are several approaches, such as:
simulated annealing (Barker's LVB)
genetic algorithm.
`phyml` algorithm.

Many use random initial starting points, so going through the procedure several times can give an idea of whether one has found a stable optimum.

Maximum vraisemblance trees: Output from `phylip` program `dnaml`:

```
Nucleic acid sequence Max. Vraisemblance, vers. 3.572c
Empirical Base Frequencies:
   A       0.27778    G 0.22685
   C       0.22325  T(U)0.27212
Transition/transversion ratio =   2.000000
(Transition/transversion parameter =   1.519971)
```




```
Ln Vraisemblance =  -344.10331
Examined   95 trees
 Between    And     Length  Approx.Conf.Limits
 -------    ---     ------  ------- ----------
   7      J        0.00006 (  zero, infinity)
```

```
 7          6     0.00003 (  zero, infinity)
 6          5     0.00006 (  zero, infinity)
 5          1     0.00936 (  zero,  0.02236) **
 1      R         0.00466 (  zero,  0.01384) **
 1          4     0.00469 (  zero,  0.01389) **
 4      N         0.00462 (  zero,  0.01369) **
 4          3     0.00003 (  zero, infinity)
 3      O         0.00462 (  zero,  0.01369) **
 3          2     0.00003 (  zero, infinity)
 2      P         0.00462 (  zero,  0.01369) **
 2      Q         0.00003 (  zero, infinity)
 5      L         0.00006 (  zero, infinity)
 6      M         0.00003 (  zero, infinity)
 7      K         0.00003 (  zero, infinity)
 *  = significantly positive, P < 0.05
** = significantly positive, P < 0.01
```

Parametric bootstrap generation of sequences

Suppose we had the `treefile` from a previous `phylip` output, the generation of sequences is done using `Seq-gen` (Rambaut and Grassly, 1997) by :

```
seq-gen -mHKY -t3.0 -l27 -n100 < treefile > example.T7
```

For which the output looks like:

```
Sequence Generator - seq-gen, Version 1.04
(c) Copyright, 1996 Andrew Rambaut and Nick Grassly
Department of Zoology, University of Oxford
South Parks Road, Oxford OX1 3PS, U.K.
Simulating 11 taxa, 27 bases
  for 1 tree(s) with 100 dataset(s) per tree
Branch lengths assumed to be number of substitutions
per site
Rate homogeneity of sites.
Model=HKY
```

```
  transition/transversion ratio = 3 (kappa=6)
  frequencies = A:0.25 C:0.25 G:0.25 T:0.25
0%|____________________|100%
  [....................]
Time taken: 0.12 seconds
```

The data file example.T7 generated looks like this:

```
 11 27
Pfa4       CCGACCTCCAAGATTCGCTATGACAAT
Pvi10      CCGACCTCCAAGATTCGCTATGACAAT
Pcy9       CCGACCTCCAAGATTCGCTATGACAAT
Pkn8       CCGACCTCCAAGATTCGCTATGACAAT
Pfr7       CCGACCTCCAAGATT.........etc

 11 27
Pfa4       ATGGTAGCGGATAACTGACTTCATCGA
Pvi10      ATGGTAGCGGATAACTGACTTCATCGA
Pcy9       ATGGTAGCGGATAACTGACTTCATCGA
```

```
Pkn8        ATGGTAGCGGATAACTGACTTCATCGA
Pfr7        ATGGTAGCGGATAACTGACTTCATCGA
Pma3        ATGGTAGCGGATAA.........etc
```

This file example. T7 was then submitted to the `phylip` program `dnapars` with the option multiple data sets indicating that there were 100 data sets to analyze, the first part of the output from this looked

((R,((((((M,K),L),N),Q),(J,P))),O)[0.0100];
((R,((((((M,K),L),N),(J,Q)),P)),O)[0.0100];
((R,((((((M,K),L),(J,N)),Q),P)),O)[0.0100];
((R,((((((M,K),(J,L)),N),Q),P)),O)[0.0100];
((R,((((((M,(J,K)),L),N),Q),P)),O)[0.0100];
((((((((J,M),(R,K)),L),N),Q),P),O)[0.0100];
((((((((J,(R,M)),K),L),N),Q),P),O)[0.0100];
((((((((R,J),M),K),L),N),Q),P),O)[0.0100];
((R,((((((J,M),K),L),N),Q),P)),O)[0.0100];
((((((((R,(J,M)),K),L),N),Q),P),O)[0.0100];
((R,J),((((((M,K),L),N),Q),P)),O)[0.0100];
((J,(R,((((((M,K),L),N),Q),P))),O)[0.0100];
((R,(J,((((((M,K),L),N),Q),P))),O)[0.0100];
((R,((J,((((((M,K),L),N),Q)),P)),O)[0.0100];
((R,(((J,((((M,K),L),N)),Q),P)),O)[0.0100];

like this: ((R,((((J,((M,K),L)),N),Q),P)),O)[0.0100];
((R,((((((J,(M,K)),L),N),Q),P)),O)[0.0100];
(((J,(R,M)),((((K,L),N),Q),P)),O)[0.0100];
((((R,J),M),((((K,L),N),Q),P)),O)[0.0100];

More believable models of Evolution:

The vraisemblance was computed as:

$$\mathcal{L}(\theta_1, \theta_2, \ldots, \theta_\eta | x_{.1}, x_{.2}, \ldots, x_{.k}, \mathcal{T}) = \prod_{j=1}^{k} f(x_{.j} | \theta, \mathcal{T})$$

Variation of rates of substitution among sites.

Variable sites models for the rates considers the sites to have different rates. The new vraisemblance takes the different rates into account:

$$P(x | T, t, r_K) = \prod_{k=1}^{K} P(x_k | T, r_k t)$$

We do not have enough information about the sites to know what these rates should be, so we integrate out the variation by integrating

out over all values of r using a prior for the rates. Yang proposes to use a gamma $g(r, \alpha, \alpha)$ prior which has mean 1 and variance $1/\alpha$ for the rates.

The vraisemblance now becomes:

$$P(x|T, t, \alpha) = \prod_{k=1}^{K} \int_0^\infty P(x_k|T, rt) g(r, \alpha, \alpha) dr$$

For each T, this is maximised with respect to t and α.

Actually better by far to use α from other data.

In practice a discrete sum approximation is sufficient.

Similar approach is to use a hidden Markov model for the states (Felsenstein and Churchill)

$$P(x|T, t, \alpha_s) = \prod_{k=1}^{K} \sum_{k=1}^{m} a_{kl} P(x_k|T, r_l) g(r, \alpha, \alpha)$$

Different areas can thus be defined:

- Surface sites of proteins may be exposed to more substitutions.
- Loops with exposed sites.
- Beta sheets have an alternance of buried and exposed sites.

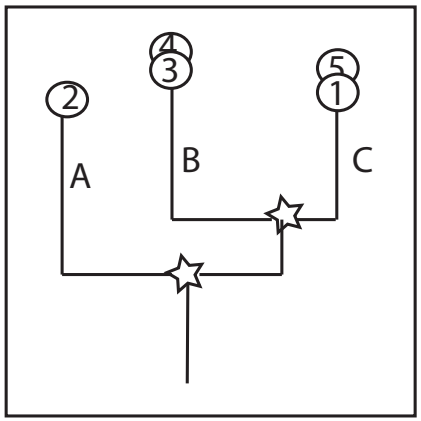
There are statistical questions that come up with trees, such as how to construct confidence regions for trees.

How sure are we of the answers ?

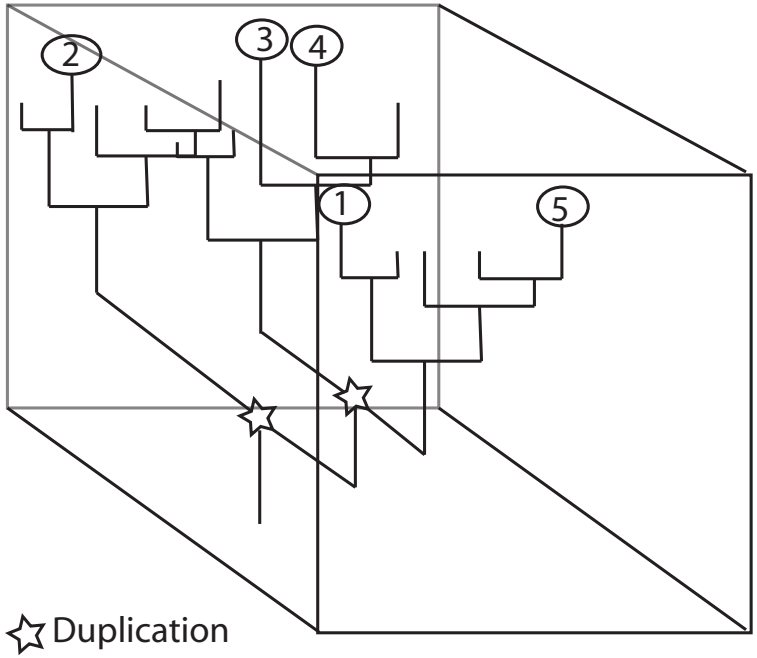
Phylogenetic Trees and Variability. There are statistical questions that come up with trees, such as how to construct confidence regions for trees.

- Aggregating/Combining trees, (Arrow's paradox)
- Stability of sets of trees, (Robustness and inference)
- Comparisons of sets of trees of several kinds.
- Explanation of one set of trees by another : regression in treespace.
- Combining trees with other data.
- Confidence Statements for trees.

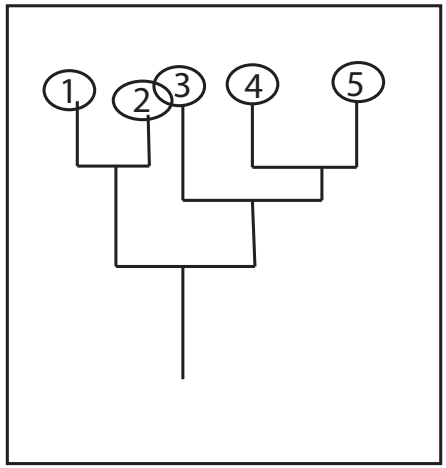
Is the tree the right representation?



Gene relationship



☆ Duplication



Speciation (taxon dimension)

Gene duplication events and speciation events. paralogs are from duplication events orthologs are without duplication events

Codon substitution models, tests and consequences:

Consensus of Trees, Bootstrap Values

M. Singh - On Phylogeny(psfile)
http://theory.lcs.mit.edu/~mona/18.417/spring98/lecture-22.ps

Systematics and Phylogenetic Inference
http://www.nyu.edu/projects/fitch/courses/evolution/html/systematics.l

An algorithmic approach to the ML tree
http://www.math.tau.ac.il/~rshamir/algmb/scribe00/html/lec08/node24