Molecular Evolution: Phylogenetic Tree Building

Lecture 1: IMS Workshop, Singapore, Susan Holmes

Bio-X and Statistics, Stanford University

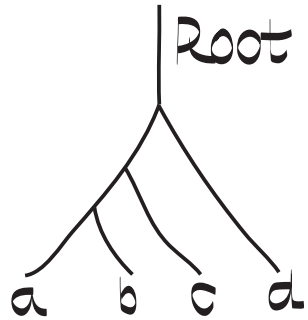
NSF grant #0241246 and NIH-R01GM086884-2

Background foundations of Phylogeny

1. What is a Tree?
2. Gene Tree.
3. Model for Molecular Evolution.
4. Mutation Rates and Edge Lengths.
5. Examples of estimation methods for trees: parsimony.
6. ML estimation.
7. Parametric Bootstrap for ML.
8. Bayesian Approach.
9. Distance based tree building.
10. Hierarchical Clustering Trees.

Phylogenetic Trees

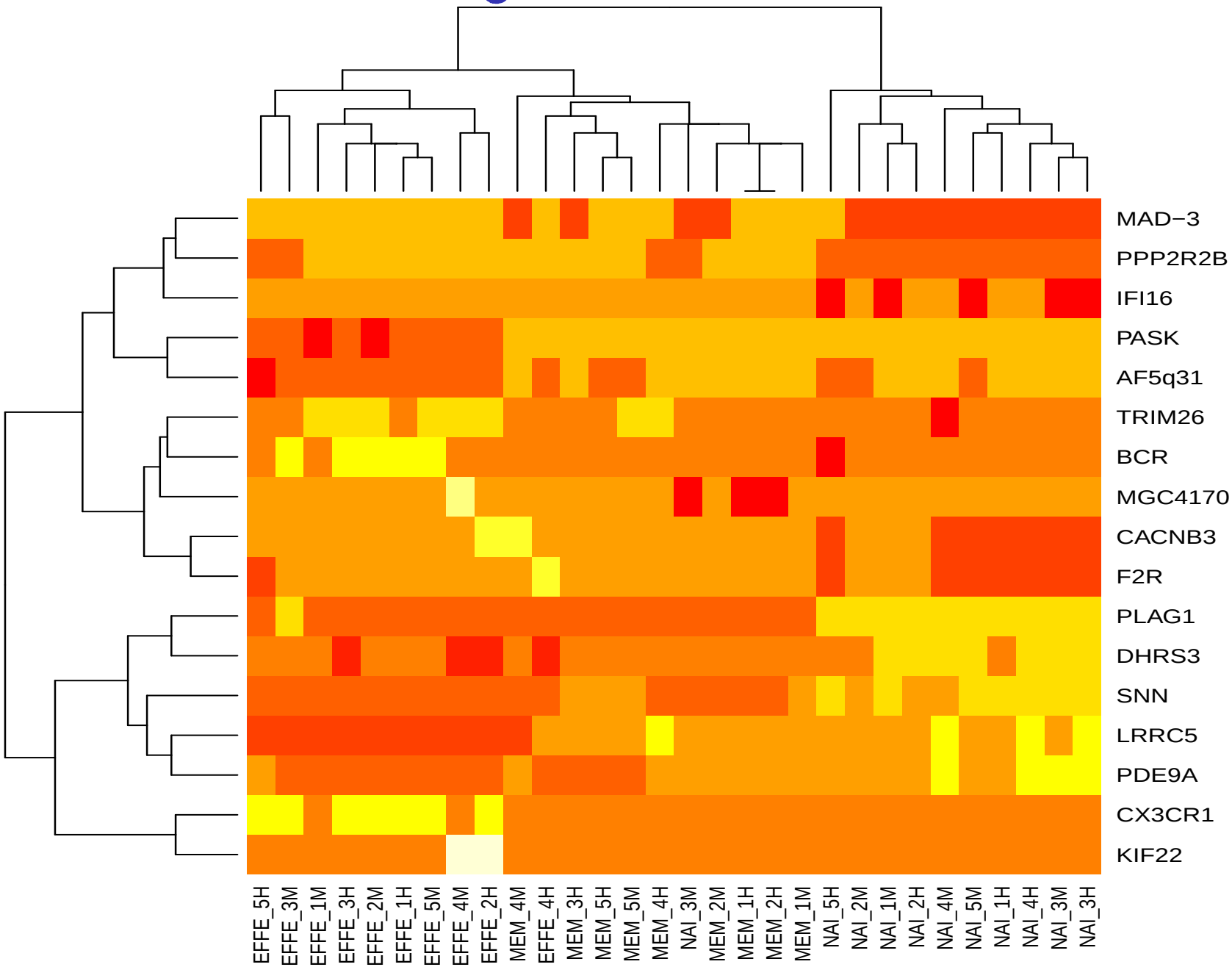
```
   11   1620
Pre1        GTACTTGTTA GGCCTTATAA GAAAAAAGT- TATTAACTTA AGGAATTATA
Pme2        GTATCTGTTA AGCCTTATAA AAAGATAGT- T-TAAATTAA AGGAATTATA
Pma3        GTATTTGTTA AGCCTTATAA GAGAAAAGTA TATTAACTTA AGGA-TTATA
Pfa4        GTATTTGTTA GGCCTTATAA GAAAAAAGT- TATTAACTTA AGGAATTATA
Pbe5        GTATTTGTTA AGCCTTATAA GAAAAA--T- TTTTAATTAA AGGAATTATA
Plo6        GTATTTGTTA AGCCTTATAA GAAAAAAGT- TACTAACTAA AGGAATTATA
Pfr7        GTACTTGTTA AGCCTTATAA GAAAGAAGT- TATTAACTTA AGGAATTATA
Pkn8        GTACTTGTTA AGCCTTATAA GAAAAGAGT- TATTAACTTA AGGAATTATA
Pcy9        GTACTCGTTA AGCCTTTTAA GAAAAAAGT- TATTAACTTA AGGAATTATA
Pvi10       GTACTTGTTA AGCCTTTTAA GAAAAAAGT- TATTAACTTA AGGAATTATA
Pga11       GTATTTGTTA AGCCTTATAA GAAAAAAGT- TATTAATTTA AGGAATTATA

            ACAAAGAAGT AACACGTAAT AA--ATTTAT TTTATTT--- -AGTGTGTAT
            ACAAAGAAGT AACACGTAAT AA--ATTATA TTTATTA--- -AGTGTGTAT
            ACAAAGAAGT AACACATAAT AAA-TTTCGA -ATATTT--- -AGTGTGTAT
            ACAAAGAAGT AACACGTAAT AA--ATTTAT TTTATTT--- -AGTGTGTAT
            ACAAAGAAGT AACACATAAT AT--ATTTAC TATATTT--- -AGTGTGTAT
            ACAAAGAAGC AACACATAAT AAAGCTGCGT CTTATTT--- -AGTGTGTAT
            ACAAAGAAGT AACACGTGAA ATGGATTAAC TCCATTTTTT TAGTGTGTAT
            ACAAAGAAGT AACACGTAAT --GGATTCT- TCCATTTT-- TAGTGTGTAT
            ACAAAGAAGT AACACGTAAT --GGATCCG- TCCATTTT-- TAGTGTGTAT
            ACAAAGAAGC GACACGTAAT --GGATCCG- TCCATTTT-- TAGTGTGTAT
            ACAAAGAAGC AACACATAAT AAAACTTTGT TTTATTT--- -AGTGTGTAT
```



Estimated in different ways from DNA/AA data:

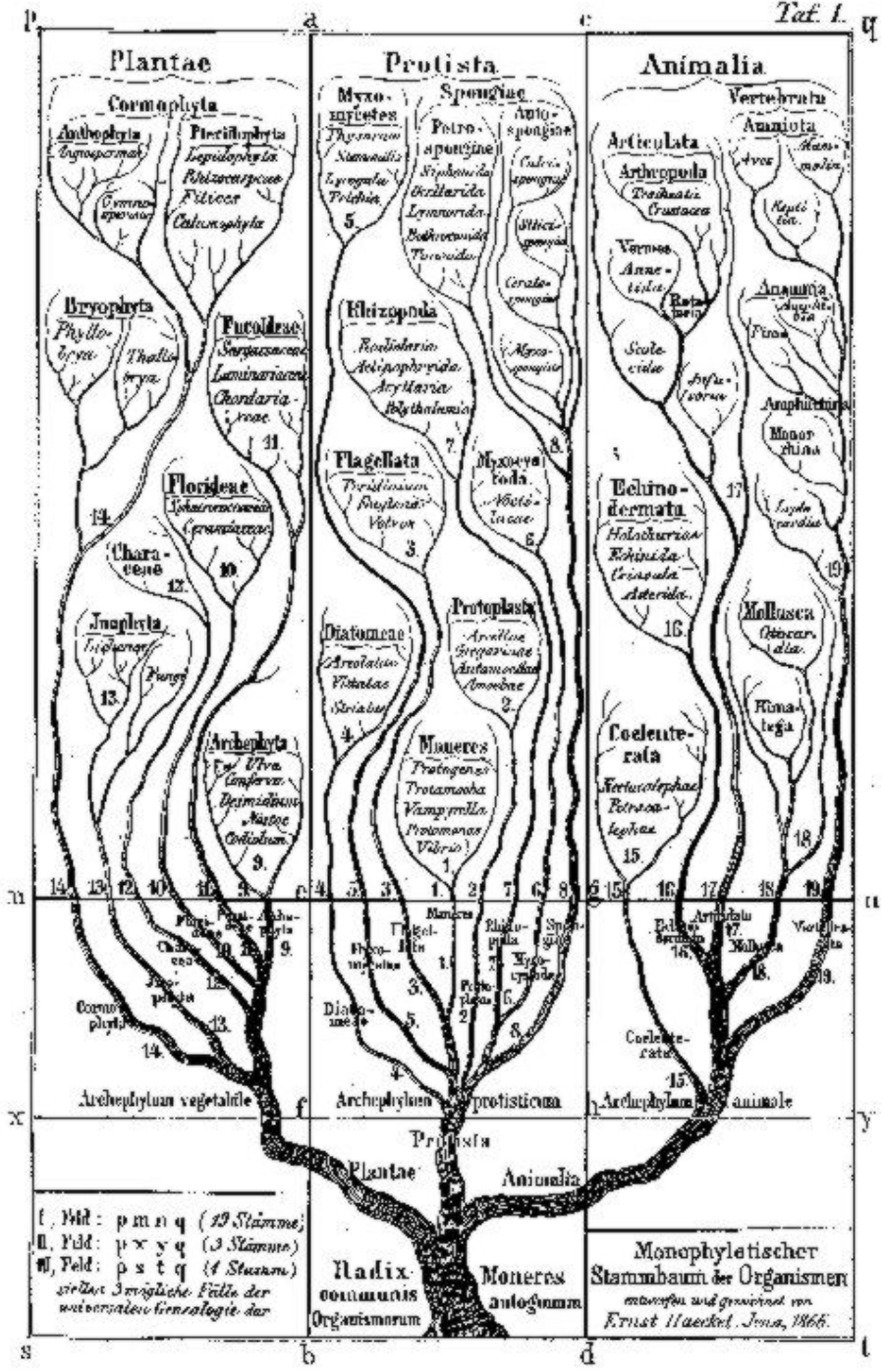
- Distance based methods: NeighborJoining, UPGMA,..
- Parsimony: Steiner tree problem.
- ML estimation, Phyml, FastML,RaxML,
- Bayesian estimation, Mr Bayes by MCMC.

Hierarchical Clustering Trees

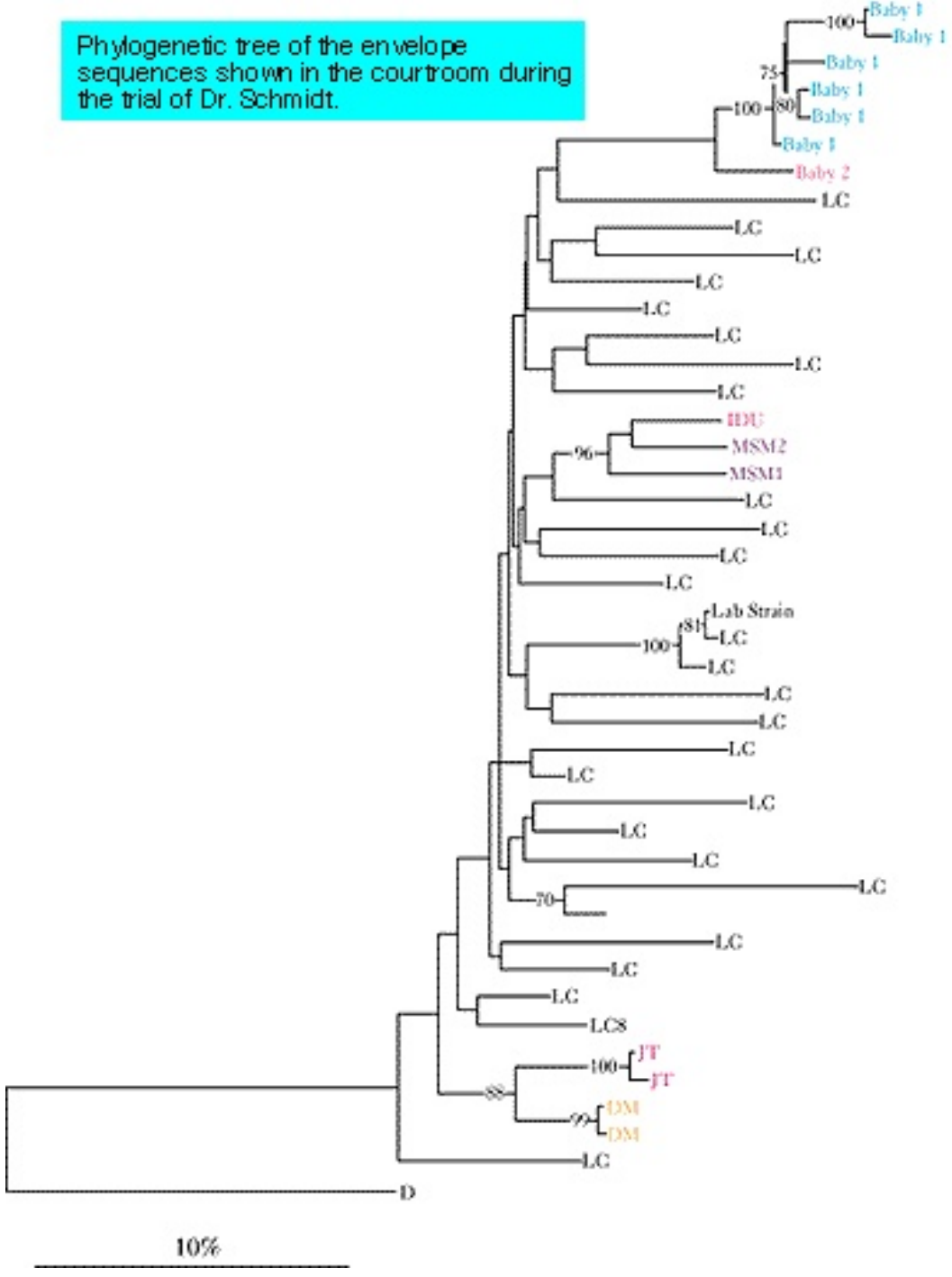


An introduction to Phylogeny

Representation of biological families by trees predates Darwin's theory of evolution, although the latter gave such representations a true explanatory justification. For biologists, at each branch of the tree are situated separation events that split orders or families or genera or species. An early example is the classification made by Haeckel, 1870.

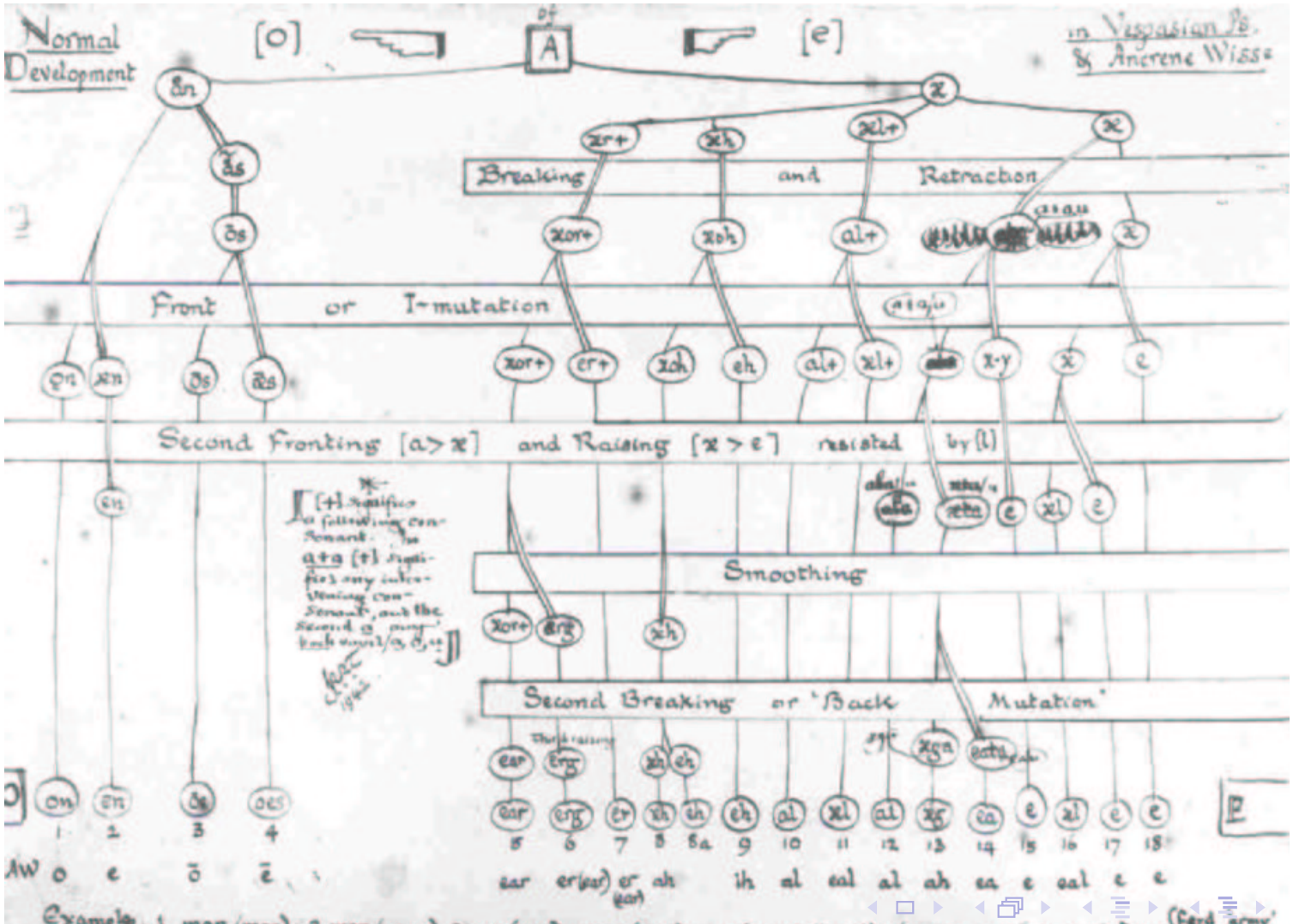

Taf. I.
Plantae
Protista
Animalia
Cormophyta
Thallophyta
Bryophyta
Fucoideae
Florideae
Characeae
Archephyta
Rhizopoda
Flagellata
Diatomeae
Protoplasta
Moneres
Vertebrata
Articulata
Arthropoda
Echinodermata
Coelenterata
Mollusca
Archephylum vegetabile
Archephylum protisticum
Archephylum animale
Radix communis Organismorum
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von
Ernst Haeckel. Jena, 1866.

Phylogenetic tree of the envelope sequences shown in the courtroom during the trial of Dr. Schmidt.



Less symmetrical Phylogenies

Linguistics use trees to map out the history of language. Linguists use trees, but they have an ancient form and a novel form. So their trees do not have symmetry between siblings.



Remarks

- Neighbors on the tree share the same ancestor.
- Characters that are derived from this common ancestry are called homologous.
- Many geneticists doing population studies replace the term homology by identity by descent (IBD).
- The distinction between homology and similarity is a subtle one.
- In particular, sisters in the tree defined by a common ancestor are called clades or monophyletic groups, they have more than just similarities in common.
- Finding monophyletic groups is one of the goals of phylogenetic studies.
- Data used to be presence or absence of wings, sepals, hair, nodules, blood groups,
- Replaced by DNA from Sanger sequencing, microarray chips and pyrosequencing. More recently the explosion of

Motivation for simple Models: HIV

HIV is very fast mutating virus. So much so that they have a rate within its host is 10 years, which is a way the virus protects itself by evolving very fast.
Much of what we are going to study doesn't happen at that scale of time. In the study of trees, we'd like to go back to find the ancestors, who are the closest relationship.
Part of what you're looking at is who are your closest siblings in the phylogenetic trees. All these methods are used at different levels and so the object we'll look at may be different species, but you could also have different populations and study human populations, that's a mainstay now in anthropology.
You can also have different genes. We give different names to these leaves we call them operational taxonomic units. These methods work where ever these leaves are. They're going to be the object of study. In statistics, the unknown tree is the parameter and we want to try to estimate it.

A gene tree

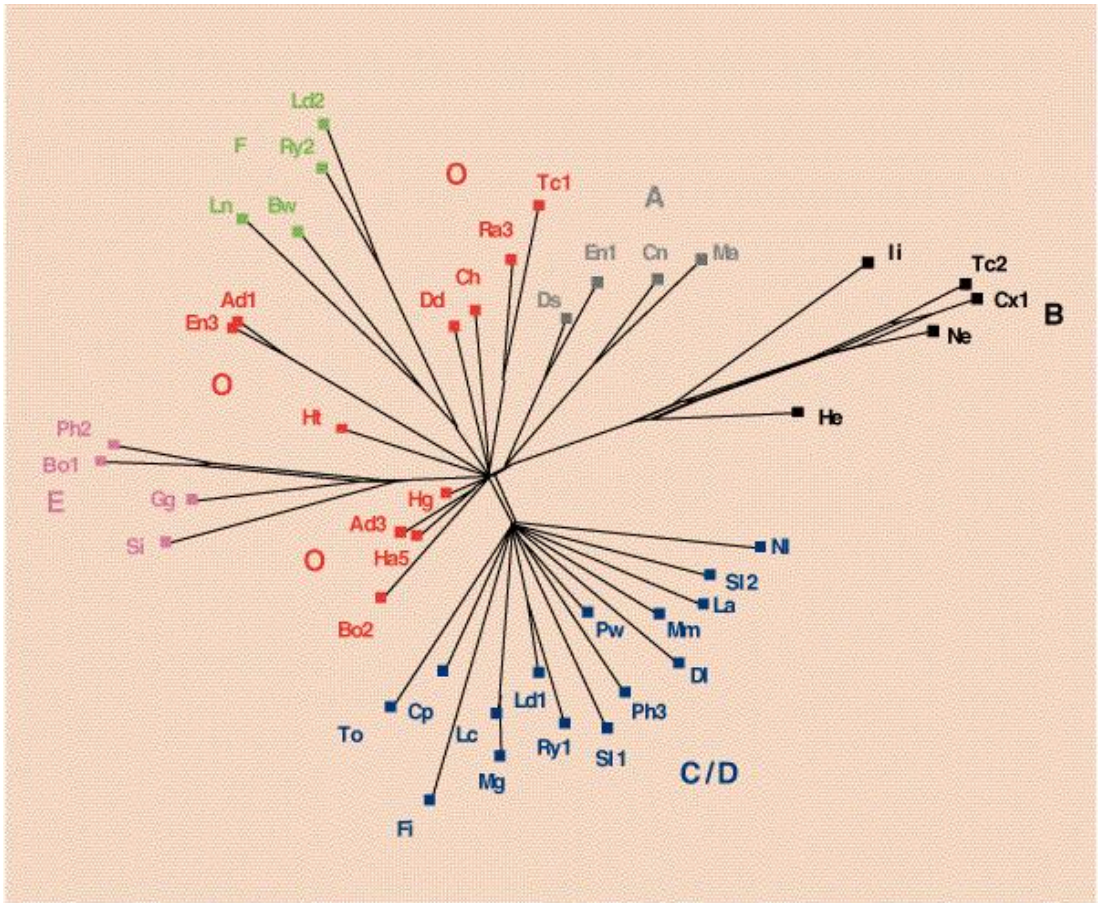
A gene sequence might be about 2000 base pairs long. One of the problems that has occurred in the last 20 years is that biologists believe that the way evolution works is that there would only be one species tree.

Different genes have different histories, so you get different gene trees. Putting them together is also a statistical problem: trying to find out what the average of the different genes are.

We're going to study the evolutionary process as one of our models for trying to understand what happens over time and how these mutations occur. What we see with the data is some columns with changes.

We're going to try to make a model for how these substitutions occur and use that model in various ways to try to make up the tree. The models we use are all Markovian. If you write them in discrete time, we have probability of a change occurring as the transition probability.

Copying Model not only for DNA



Chaucer

Continuous time Markov chains

Memoryless Property $P(Y(u+t)=j|Y(t)=i)$ doesn't depend on time before t

Time homogeneity $P(Y(h+t)=j|Y(t)=i)$ doesn't depend on t, only depends on h, time between the events.

Instantaneous transition rate

$$P_{ij}(h) = q_{ij}h + o(h), j \neq i.$$

$$P_{ii}(h) = 1 - q_i(h) + o(h), \qquad q_i = \sum_{j \neq i} q_{ij}$$

q_{ij} is known as the instantaneous transition rate.

Times between changes are exponential

$$
\begin{aligned}
P(T \geq t+h) &= P(T \geq t)P(T \geq t+h | T \geq t) \\
P(T \geq t+h) &= P(T \geq t)P(T \geq h) \\
&= P(T \geq t)(1 - q_i h + \ldots) \\
\frac{P(t \geq t+h) - P(T \geq t)}{h} &= -q_j P(T \geq t) \\
\frac{dP(T \geq t)}{dt} &= -q_i P(T \geq t) \\
P(T \geq 0) &= 1
\end{aligned}
$$

gives solution

$$
\begin{aligned}
P(T \geq t) &= e^{-q_i t} \\
P(T \leq t) &= 1 - e^{-q_i t} \\
f(t) &= q_i e^{-q_i t} \sim \mathrm{Exp}(q_i)
\end{aligned}
$$

Derivative of P

$$\frac{P_{ij}(t+h) - P_{ij}(t)}{h} = -q_j P_{ij}(t) + \sum_{k \neq j} q_{kj} P_{ik}(t)$$

as $h \longrightarrow 0$,

$$\frac{dP_{ij}(t)}{dt} = -q_j P_{ij}(t) + \sum_{k \neq j} q_{kj} P_{ik}(t)$$

The simplest possible model we'll study, the mutations are all equally likely. This model, called a Jukes-Cantor model is a one parameter model. We suppose that every transition is reversible and that the probability is that they're all equal.
Particular case of Jukes-Cantor: $q_j = 3\alpha$ and $q_{ij} = \alpha, i \neq j$.

$$\begin{aligned}
\frac{dP_{ij}(t)}{dt} &= -3\alpha P_{ij}(t) + \alpha \sum_{k \neq j} P_{ik}(t) \\
&= -3\alpha P_{ij}(t) + \alpha(1 - P_{ij}(t)) \\
&= \alpha - 4\alpha P_{ij}(t) \\
P_{ii}(0) &= 1 \text{ and } P_{ij}(0) = 0
\end{aligned}$$

gives solutions

$$\begin{aligned}
P_{ii}(t) &= \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \\
P_{ij}(t) &= \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}
\end{aligned}$$

The rate matrix Q is of the form:

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

The Kimura two parameter model is:

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

The 12 parameter model is of the form

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

The substitution matrix gives the probability of the change of a nucleotide during a time t as the continuous Markov chain with infinitesimal generator Q.
In the case of the amino acids we would have bigger matrices (20×20 instead of 4×4), but most of the other computations carry through.
The best reference about these subjects are the books by W. H Li and WH Li and D. Graur. See also Page and E. Holmes on Molecular Evolution: A phylogenetic approach.

Estimating the rates

- Call λ the amino acid replacement rate per year,

$$\lambda = \frac{K}{2t} = \frac{\#\text{substit.}}{2 \times \text{divergence time}}$$

- Probability that a site stays unchanged through t intervals is $(1-\lambda)^{2t}$
- The probability D_t of one or more replacements occurring in t units of time is

$$1 - (1-\lambda)^{2t}$$

-

$$\begin{aligned} 1 - D_t &= (1-\lambda)^{2t} \\ \log(1-D_t) &= 2t \log(1-\lambda) \\ \log(1-D_t) &= \frac{K}{\lambda} \log(1-\lambda) \simeq -K \end{aligned}$$

Expected proportion of differences between sequences at time t.

Example : β globin molecule in primates

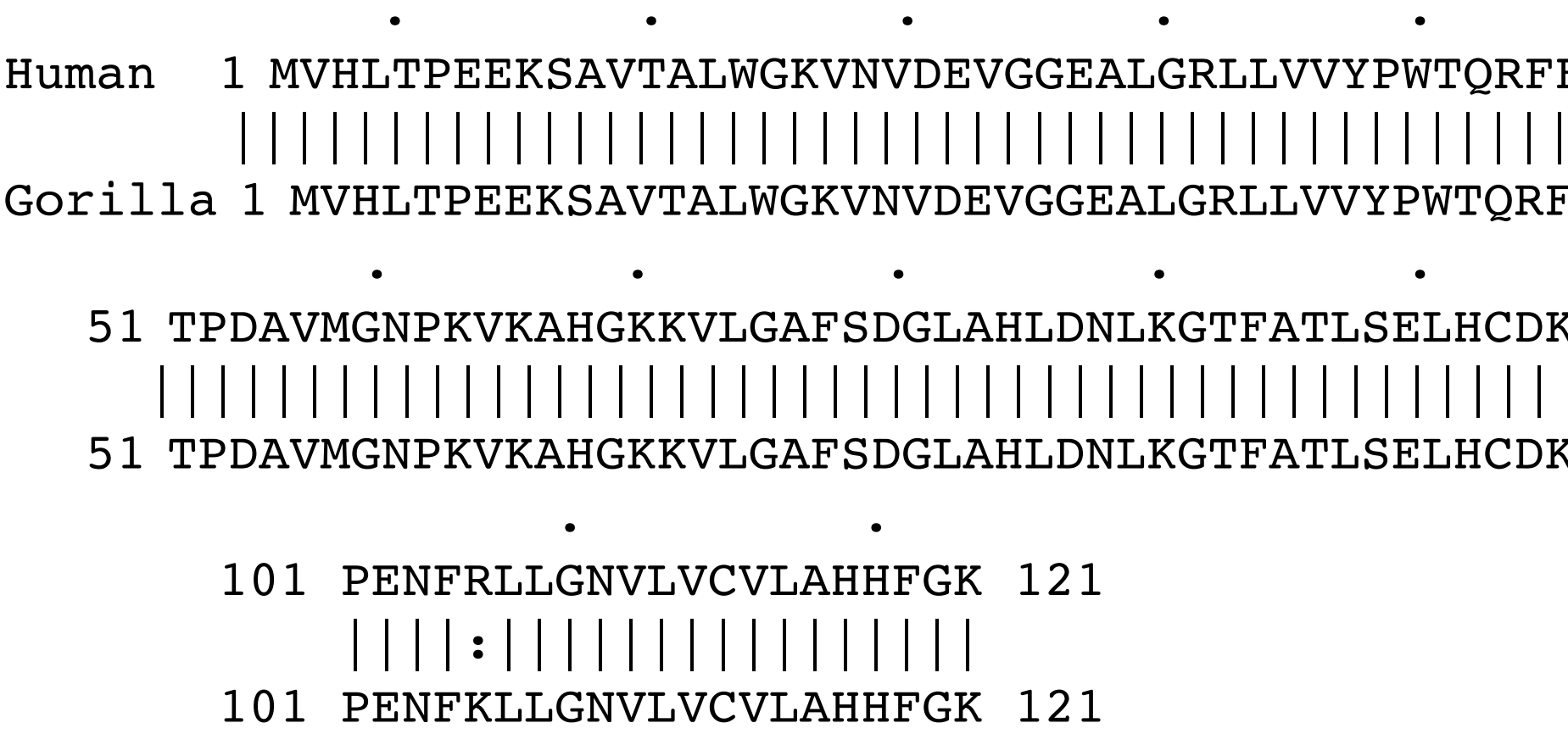
contains 146 amino acids, the estimates of the number of differences are:

Time of div. (millions of years)	Average # of amino acid changes	average $\hat{D}$ differ.	-log$(1-\hat{D})$
85	25.5	25.5/146	.192
60	24	24/146	.180
42	6.25	6.25/146	.044
40	6.0	6.0/146	.042
30	2.5	2.5/146	.018
15	1.5	1.5/146	.007

The slope is around $a = .002$, and the evolution rate is half of this, so: 10^{-3} per million years or 10^{-9} per year.

```
Human    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Gorilla  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Rabbit   MVHLSSEEKSAVTALWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDLSSANAVMNNPK
Cow      M..LTAEEKAAVTAFWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPK
Goat     M..LTAEEKAAVTGFWGKVKVDEVGAEALGRLLVVYPWTQRFFEHFGDLSSADAVMNNAK
Mouse    MVHLTDAEKAAVSCLWGKVNSDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNAK
Chicken  MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM
Carp     MVEWTDAERSAIIGLWGKLNPDELGPQALARCLIVYPWTQRYFASFGNLSSPAAIMGNPK

         61                                                        120
Human    VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Gorilla  VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG
Rabbit   VKAHGKKVLAAFSEGLSHLDNLKGTFAKLSELHCDKLHVDPENFRLLGNVLVIVLSHHFG
Cow      VKAHGKKVLDSFSNGMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARNFG
Goat     VKAHGKKVLDSFSNGMKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVVVLARHHG
Mouse    VKAHGKKVITAFNDGLNHLDSLKGTFASLSELHCDKLHVDPENFRLLGNMIVIVLGHHLG
Chicken  VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS
Carp     VAAHGRTVMGGLERAIKNMDNIKATYAPLSVMHSEKLHVDPDNFRLLADCITVCAAMKFG

          121                         148
Human    .KEFTPPVQAAYQKVVAGVANALAHKYH
Gorilla  .K..........................
Rabbit   .KEFTPQVQAAYQKVVAGVANALAHKYH
Cow      .KEFTPVLQADFQKVVAGVANALAHRYH
Goat     .SEFTPLLQAEFQKVVAGVANALAHRYH
Mouse    .KDFTPAAQAAFQKVVAGVATALAHKYH
Chicken  .KDFTPECQAAWQKLVRVVAHALARKYH
Carp     PSGFSPNVQEAWQKFLSVVVSALCRQYH
```

Human beta-globin vs. Gorilla beta-globin
Percent Similarity: 100
Percent Identity: 99



We're going to separate out two problems, which in today's age of computing, should be mixed together: alignment and trees.
I'm going to suppose we have sequences either of amino acids or nucleotides which we have aligned. This is an example data set I did in my first phylogeny paper I wrote was with Brad Efron in which we analyzed malaria data. These are malaria sequences from 11 different species of malaria. Two of the species of malaria are human malaria. The others are from different animals. The question in trying to find out information from the families has a lot of influence on designing vaccines.

Malaria Data

```
    11   1620
Pre1         GTACTTGTTA GGCCTTATAA GAAAAAAGT- TATTAACTTA AGGAATTATA
Pme2         GTATCTGTTA AGCCTTATAA AAAGATAGT- T-TAAATTAA AGGAATTATA
Pma3         GTATTTGTTA AGCCTTATAA GAGAAAAGTA TATTAACTTA AGGA-TTATA
Pfa4         GTATTTGTTA GGCCTTATAA GAAAAAAGT- TATTAACTTA AGGAATTATA
Pbe5         GTATTTGTTA AGCCTTATAA GAAAAA--T- TTTTAATTAA AGGAATTATA
Plo6         GTATTTGTTA AGCCTTATAA GAAAAAAGT- TACTAACTAA AGGAATTATA
Pfr7         GTACTTGTTA AGCCTTATAA GAAAGAAGT- TATTAACTTA AGGAATTATA
Pkn8         GTACTTGTTA AGCCTTATAA GAAAAGAGT- TATTAACTTA AGGAATTATA
Pcy9         GTACTCGTTA AGCCTTTTAA GAAAAAAGT- TATTAACTTA AGGAATTATA
Pvi10        GTACTTGTTA AGCCTTTTAA GAAAAAAGT- TATTAACTTA AGGAATTATA
Pga11        GTATTTGTTA AGCCTTATAA GAAAAAAGT- TATTAATTTA AGGAATTATA

             ACAAAGAAGT AACACGTAAT AA--ATTTAT TTTATTT--- -AGTGTGTAT
             ACAAAGAAGT AACACGTAAT AA--ATTATA TTTATTA--- -AGTGTGTAT
             ACAAAGAAGT AACACATAAT AAA-TTTCGA -ATATTT--- -AGTGTGTAT
             ACAAAGAAGT AACACGTAAT AA--ATTTAT TTTATTT--- -AGTGTGTAT
             ACAAAGAAGT AACACATAAT AT--ATTTAC TATATTT--- -AGTGTGTAT
             ACAAAGAAGC AACACATAAT AAAGCTGCGT CTTATTT--- -AGTGTGTAT
             ACAAAGAAGT AACACGTGAA ATGGATTAAC TCCATTTTTT TAGTGTGTAT
             ACAAAGAAGT AACACGTAAT --GGATTCT- TCCATTTT-- TAGTGTGTAT
             ACAAAGAAGT AACACGTAAT --GGATCCG- TCCATTTT-- TAGTGTGTAT
             ACAAAGAAGC GACACGTAAT --GGATCCG- TCCATTTT-- TAGTGTGTAT
             ACAAAGAAGC AACACATAAT AAAACTTTGT TTTATTT--- -AGTGTGTAT
```

Transitions and Transversions

The probability of changing from a purine to a pyrimidine is called a transversion. If you think about coding sequences, the amino acids you don't code the amino acid if you have a transition. We make the two parameter model is the most used in the study of evolution. We don't have discrete time, that's just a simplification.

Model 0:Jukes Cantor

This model is not a completely realistic model.
All mutations, transversions and translations are equally likely.
The probability of it not changing is $1 - 3\alpha$. This is discrete time markov chain matrix.
You can look at it stationary distribution because you have a perfect symmetry, the left eigenvector is $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$.
This stationary distribution of $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$.
If for a long time you have sequences evolving over time and you're lost track of time and you pull a nucleotide at random it has equal probability of being any of those.

Transitions and Transversions

The probability of changing from a purine to a pyrimidine is called a transversion. If you think about coding sequences, the amino acids you don't code the amino acid if you have a transition. We make the two parameter model is the most used in the study of evolution. We don't have discrete time, that's just a simplification.

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 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

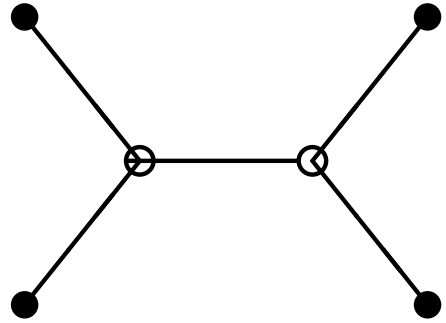
Nonparametric procedures. Farris (1983), has a justification for parsimony : ``minimizes requirements of ad hoc hypotheses of homoplasy".
Analogy is made between homoplasies and residuals, (part of the data that the tree does not explain), minimizing homoplasies is akin to minimizing residuals in regression.
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.

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

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 and the Steiner tree of the 4 vertices of a rectangle.
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.
#species=1500 can now be done routinely.

Parsimony as a statistical procedure

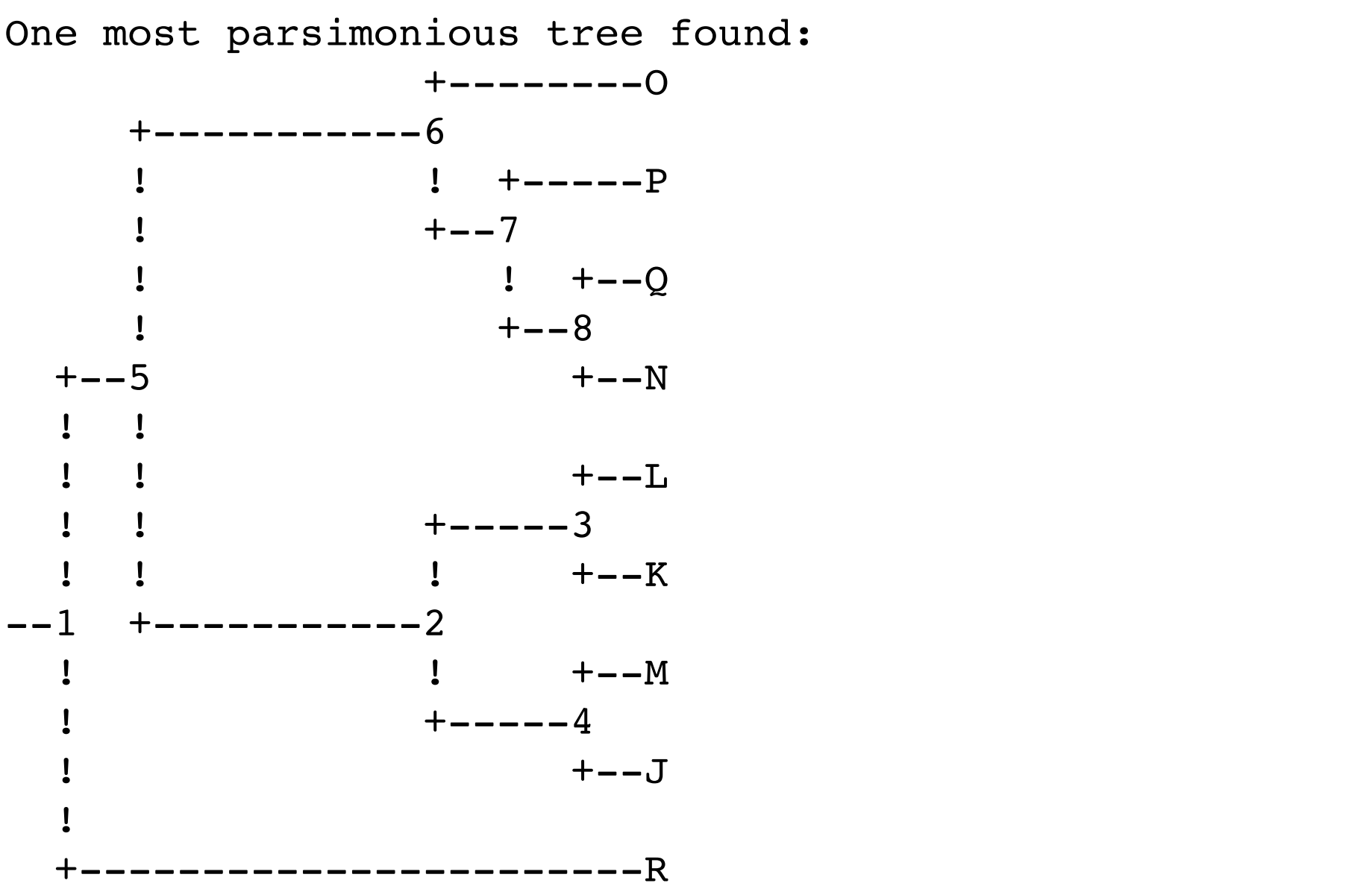
Felsenstein (1983) lists parsimony in a section entitled a section on parsimony as ``non-statistical approaches". Farris says (1983) says the ``statistical approach to phylogenetic inference was wrong from the start, for it rests on the idea that to study phylogeny at all one must first know in great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

In fact parsimony is just a nonparametric method of estimating the tree parameter.

Simple Example

T7 data experimentally generated phylogeny, Hillis et al. (1992) for which the parsimony program will be seen to produce the correct answer. Here is the part of the data set (in `phylip` form) composed of the 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
```


One most parsimonious tree found:
O
6
P
7
Q
8
N
5
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    6    7    8    9
   *-------------------------------------------------
  0!        1    2    2    1    2    2    1    1    1
 10!   1    1    1    1    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 $(2N - 3)$). 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:

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

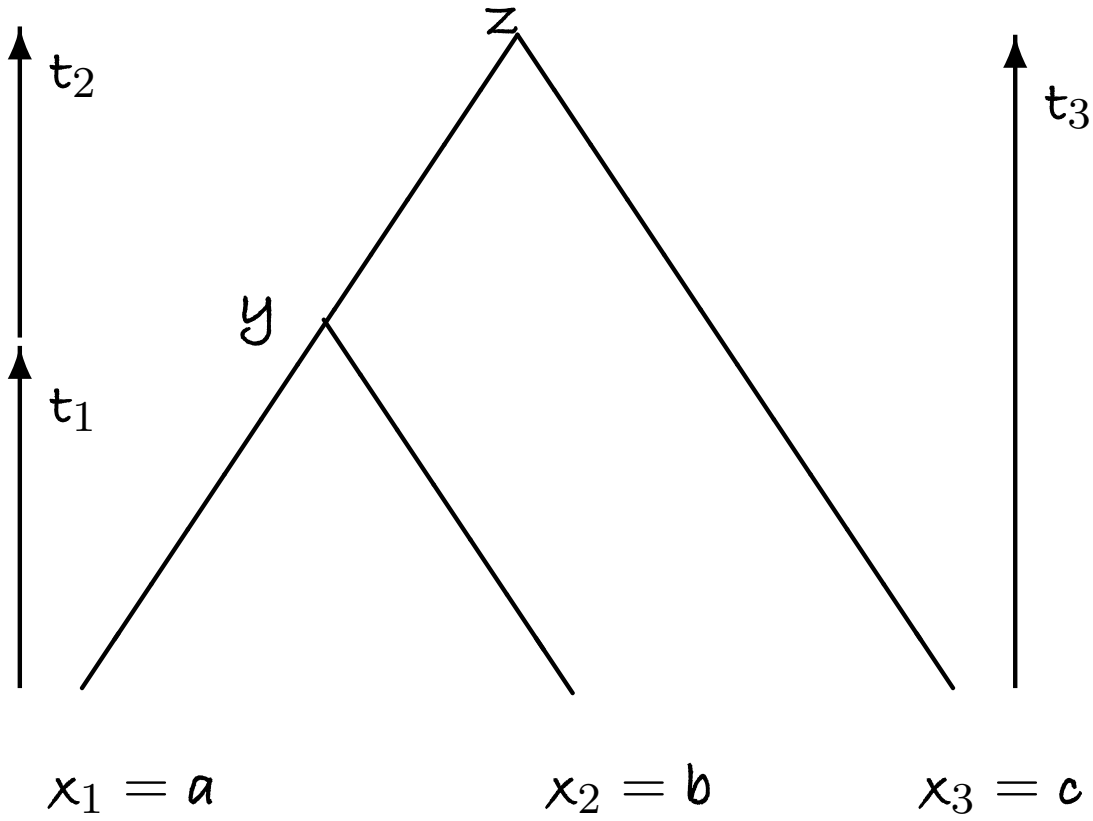
The effect on the resulting measurements of differences between units are the same: there is an error; units appear to be more similar than they would be if the complete history were known. Collectively these are called homoplasy.

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 ho-
moplasy; `long branches' are undesirable in parsimony methods.

Maximum likelihood 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 T is the tree's topology. Under this model the likelihood for each possible tree T is separately computed for each character, the independence of characters then allows the total likelihood 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 [1] is postulated, speciation times $\{t_1, t_2, \ldots t_{N-2}\}$ (splitting events) are the other parameters. Otherwise both the branch lengths $\{v_1, v_2, \ldots v_{N-2}\}$ and the different rates along those branches have to be parametrized.

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



The substitution parameters are estimated from the data. A complete model including distributions of separation events is postulated and the likelihood can be computed for each possible tree by computing the likelihood of the tree for each site x_j:

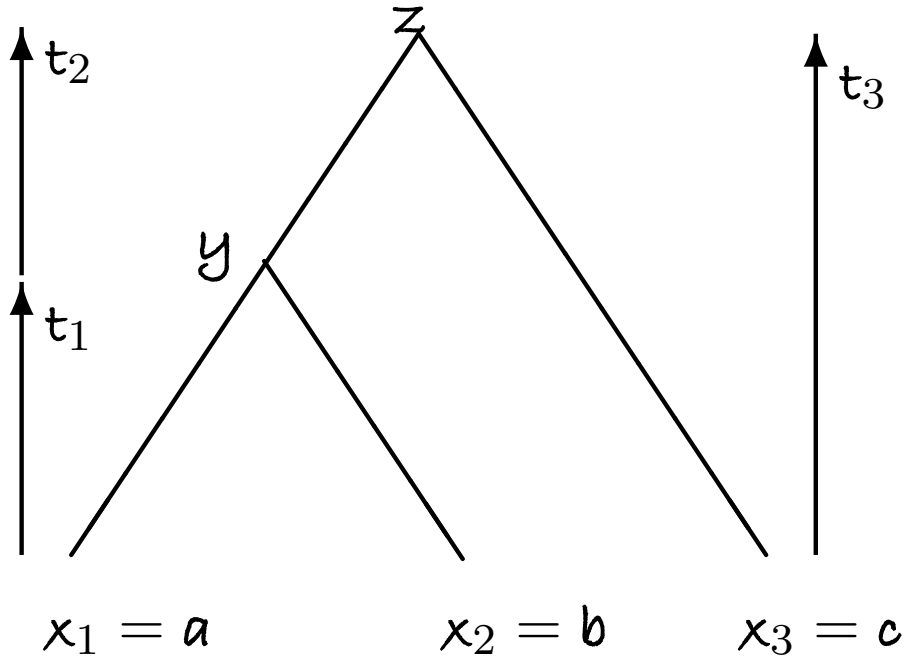
$$f(x_j|\theta_1, \theta_2, \ldots, \theta_\eta, \mathcal{T}).$$

This actually requires computing the likelihood of all the subtrees, so the method is recursive.

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

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).



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

$$P(a, b, c, y, z|T, t) = P(a|y, t_1)P(b|y, t_1)P(c|z, t_3)P(y|z, t_2)P(z)$$

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

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 likelihood estimate.
We can consider the likelihood 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):

$$\mathbb{P}(x_1, x_2, a|\mathcal{T}, \theta_1 = t_1, \theta_2 = t_2) = \pi_a \mathbb{P}(x_1|a, \theta_1)\mathbb{P}(x_2|a, \theta_2)$$

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

$$\mathbb{P}(x_1, x_2|\mathcal{T}, \theta_1 = t_1, \theta_2 = t_2) = \sum_a \pi_a \mathbb{P}(x_1|a, \theta_1)\mathbb{P}(x_2|a, \theta_2)$$

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 likelihood 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.
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 $\mathcal{A}$,

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

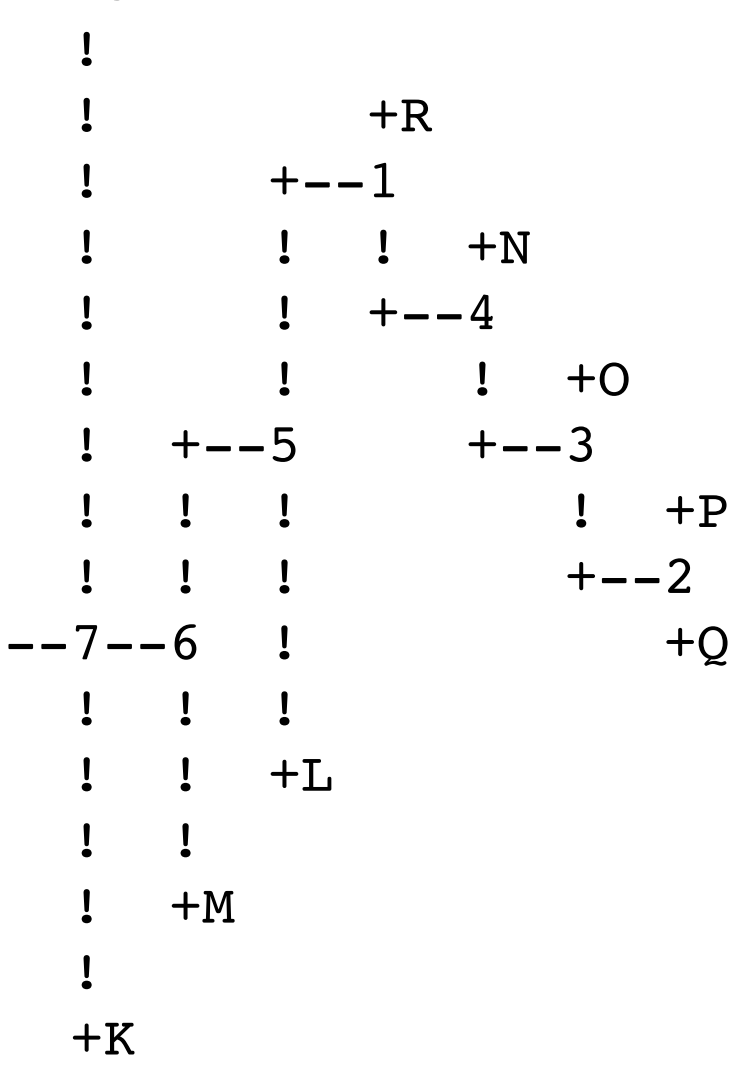
the computational algorithm evaluates $\mathbb{P}(L_i|a)$ for the children j and k such that $m[j] = m[k] = i$, we compute $\mathbb{P}(L_j|b)$ and $\mathbb{P}(L_k|c)$ for all possible b and c.
These instructions allow us to compute the likelihood of any tree, given its branching order (sometimes called topology) and its branch lengths.
For the maximum likelihood computation, we need to compute the tree that maximizes the likelihood, first for a given branching order, find the branch lengths that maximize the likelihood. This can be done by taking the derivative $\frac{\partial \mathbb{P}(x^j | x^{m[j]}, \theta_k)}{\partial \theta_j}$ 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.

Complexity: Hard

Finding the likelihood of one tree is an NP complete problem
Remark :There is no known polynomial time algorithm that finds the tree with maximum likelihood.
Thus as we need to look at all the topologies, of which there are exponentially many; we see the exact computation becomes quickly intractable as the number of leaves increases.
Nice implementations:
phylip, RaXML, FastML, PhyML, (see wikipedia)...
From R: phangorn, phyml.

Maximum likelihood trees: Output from phylip program dnaml:

```
Nucleic acid sequence Max. Likelihood, 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)
```



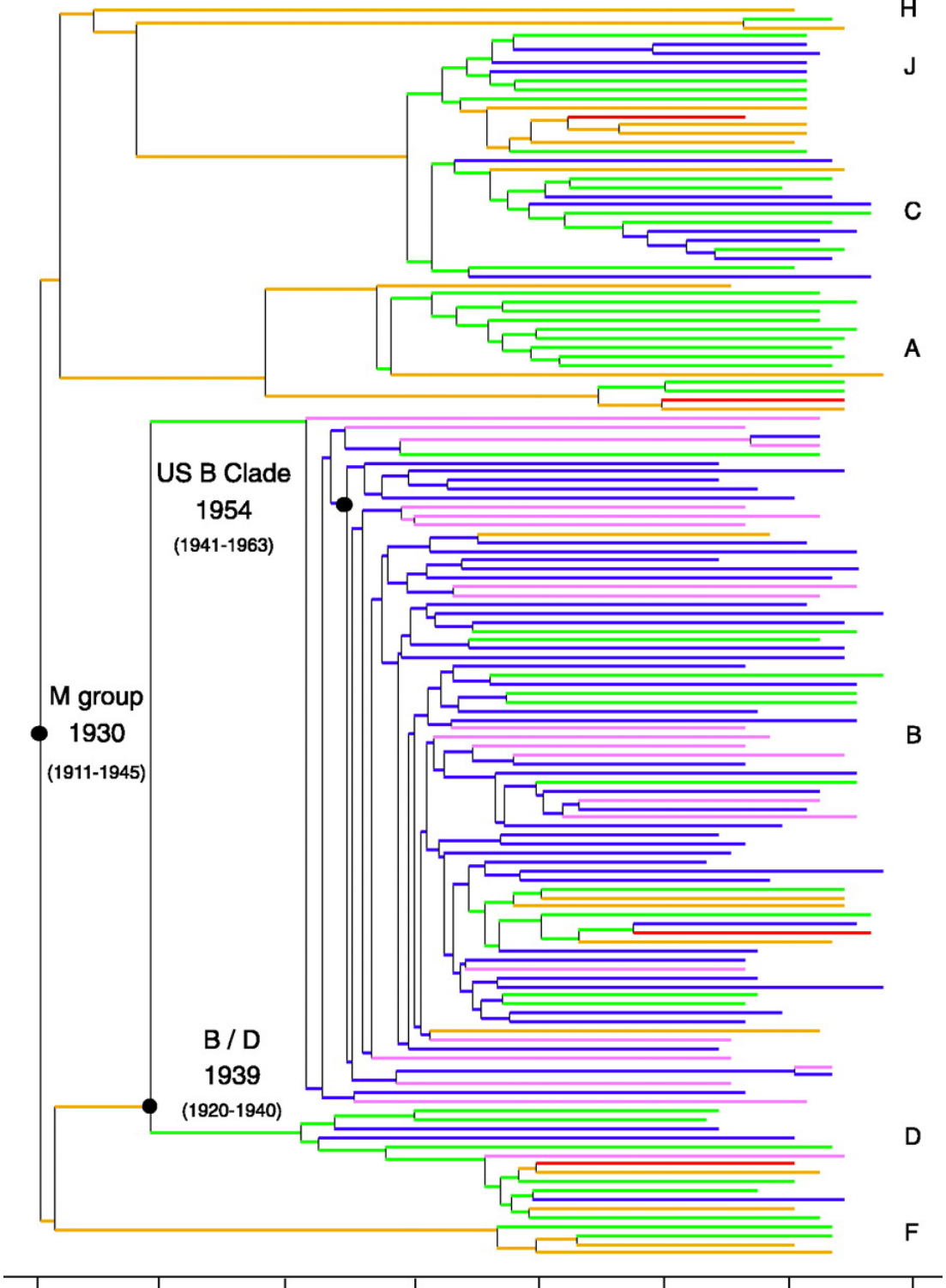
Ln Likelihood = -344.10331
Examined 95 trees

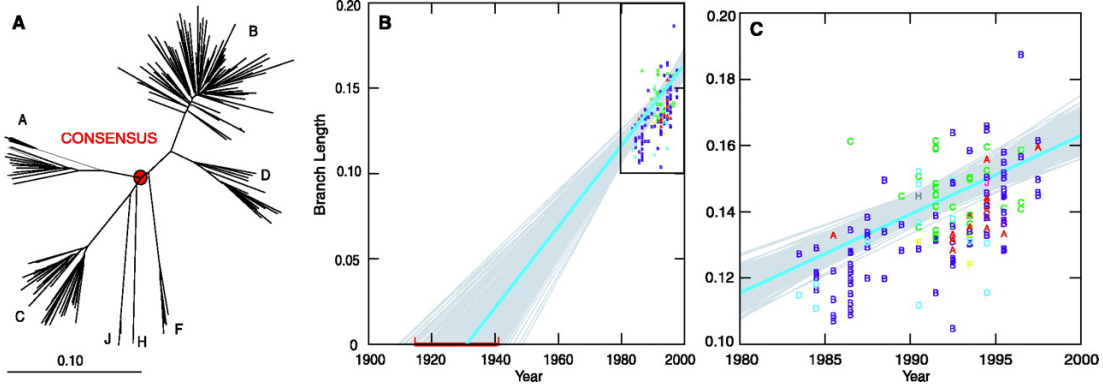
Between	And	Length	Approx.Conf.Limits
7	J	0.00006	(zero, infinity)

```
 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
```

ML Estimate Application: Origins of HIV

The article by Korber et al. provides an estimate of a most recent ancestor. When you see two sequences, how much time went by until the most recent common ancestor.
The English author, Hooper, hypothesis that HIV was spread by dispensaries who were giving the polio vaccination in East Africa. They were supposed to be responsible for diffusing AIDS because the vaccination was grown in monkey tissue. The idea was to try to disprove this occurred at the time of the vaccination program in 1957 and this study was trying to make a confidence interval of the time of the most recent ancestor using as many sequences as they had to make up the whole tree.
One of the reasons this data seemed interested is that this data is freely available on Los Alamos National Laboratories.


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



The ideas is that of the models we are using for molecular evolution, they have this molecular clock.
You have a homogenous process, the number of mutations with be proportionate to time.
There hasn't been much progress in disproving or in proving this molecular clock hypothesis, so the the way it's justified is the average the amount of mutation that occurs over time.

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.T
```

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
R        CCGACCTCCAAGATTCGCTATGACAAT
P      CCGACCTCCAAGATTCGCTATGACAAT
Q        CCGACCTCCAAGATTCGCTATGACAAT
L        CCGACCTCCAAGATTCGCTATGACAAT
M        CCGACCTCCAAGATT.........etc
..
 11  27
R        ATGGTAGCGGATAACTGACTTCATCGA
P      ATGGTAGCGGATAACTGACTTCATCGA
Q        ATGGTAGCGGATAACTGACTTCATCGA
L        ATGGTAGCGGATAACTGACTTCATCGA
M        ATGGTAGCGGATAACTGACTTCATCGA
......         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 like this:

```
((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];
((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];
((M,((R,J),((((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];
```

Notice at the end of each tree is associated a weight.

Molecular Clock

Says that the probability of changes along the edges of the tree are proportional to edgelengths:

Molecular Clock

Says that the probability of changes along the edges of the tree are proportional to edgelengths:

More believable models of Evolution:

The likelihood 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 likelihood 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 likelihood 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.

Full Bayesian Method

- Prior distribution on all tree branching patterns.
- Gamma dsitribution for the rates.
- Compute posterior distribution using MCMC.

Implementations:`MrBayes, Beast`

Open Questions:

- Prior probability model for trees , open question. Uniform distribution on all trees poses big problem:
  $2n - 3!!$ different binary rotted semi-labeled trees with n leaves. With 10, you have more than a million trees.
- How long to run the MCMC? (Diaconis and Holmes, EJP cannot touch the real case)
  Negative results by Mossel and Vigoda on problems with mixtures.
- Using the output from MCMC runs ...we will talk about this.

Distance Based Methods

In phylogenetics, neighbor joining is very similar to the algorithms used for hierarchical clustering.
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 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.

Iterative Agglomeration: Bottom up heuristic

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:

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

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

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.
This can be destroyed by:

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

Hierarchical clustering trees

Built from distances or dissimilarities between the rows of the data matrix [7].
Common examples include computations of dissimilarities in gene expression or in occurrence of words in texts or webpages.
The resulting hierarchical clustering tree has the advantage over simple partitioning methods that one can look at the output in order to make an informed decision as to the relevant number of clusters for a particular data set.
Microarray studies have popularized the use of a double hierarchical clustering or bi-clustering trees where both the rows and columns of the data are clustered. This is the most popular method for visualizing both relations between genes and patient groups in gene expression studies [1, 5].
Many implementations are available; the illustration in Figure in the introduction was made with `heatmap` function in R [9].

References

[1] D.B. Carr, R. Somogyi, and G. Michaels. Templates for looking at gene expression clustering. Statistical Computing & Statistical Graphics Newsletter, 7:20--29, 1997.

[2] J. Chakerian and S. Holmes. Computational methods for evaluating phylogenetic trees, 2010. arXiv.

[3] J. Chakerian and S. Holmes. distory:Distances between trees, 2010.

[4] P. W. Diaconis and S. P. Holmes. Matchings and phylogenetic trees. Proc. Natl. Acad. Sci. USA, 95(25):14600--14602 (electronic), 1998.

[5] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America, 95(25):14863, 1998.

[6] J. Felsenstein. Inferring Phylogenies. Sinauer, Boston, 2004.

[7] J Hartigan. Representation of similarity matrices by trees. Journal of the American Statistical Association, Jan 1967.

[8] S. Holmes. Bootstrapping phylogenetic trees: theory and methods. Statist. Sci., 18(2):241--255, 2003. Silver anniversary of the bootstrap.

[9] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3):299--314, 1996.

[10] E. Mossel and E. Vigoda. Phylogenetic mcmc algorithms are misleading on mixtures of trees. Science, 309(5744):2207--9, Sep 2005.

[11] E. Paradis. Ape (analysis of phylogenetics and evolution) v1.8-2, 2006. http://cran.r-project.org/doc/packages/ape.pdf.