

MA3259 Lecture 9

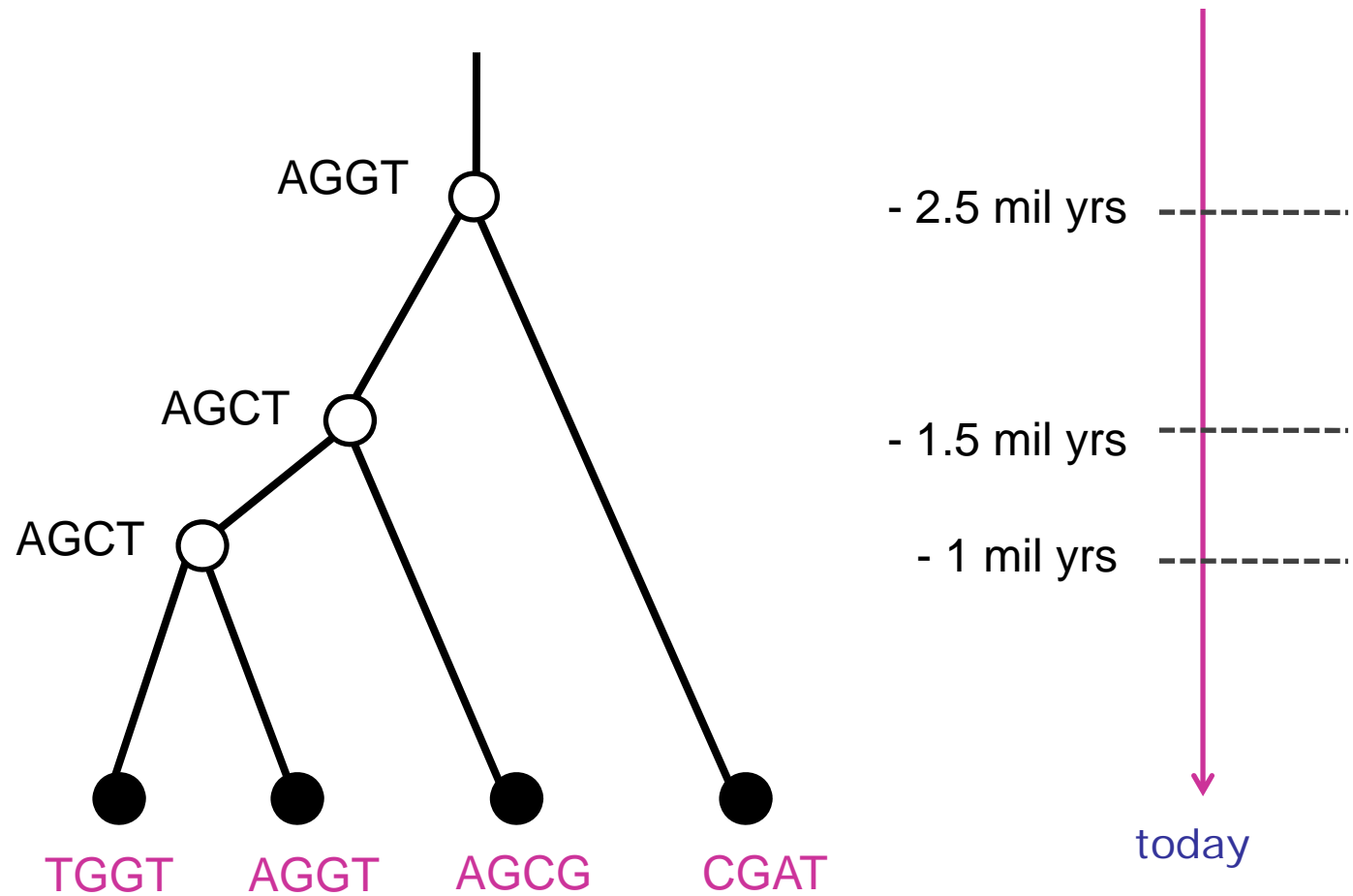
**Molecular Phylogenetic Analysis:
Part 3**

LX Zhang

Department of Mathematics
National University of Singapore

matzlx@nus.edu.sg

1. Computing Sequence Distance

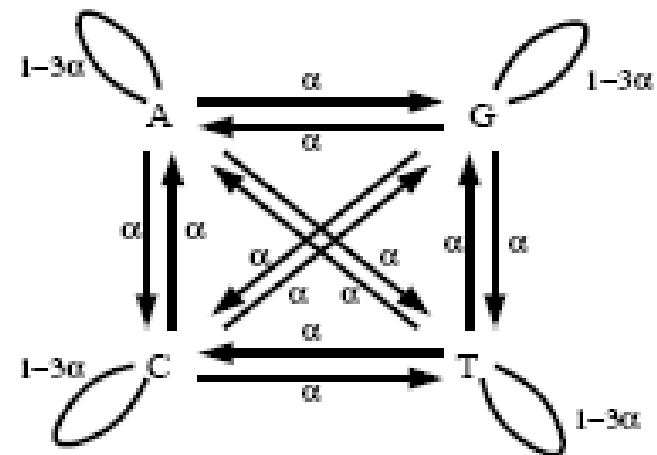


Jukes-Cantor Model for DNA sequences

- The Jukes-Cantor model is to attempt a correction for unobserved changes that are overlain or reversed by others.
- The nucleotide of each site (or position) at the root is random, having uniform distribution.
- The sites evolve independently and identically (i.i.d.)
- If the site changes its state on an edge e , it changes with equal probability to the other states. That is, for any different nucleotides x and y ,

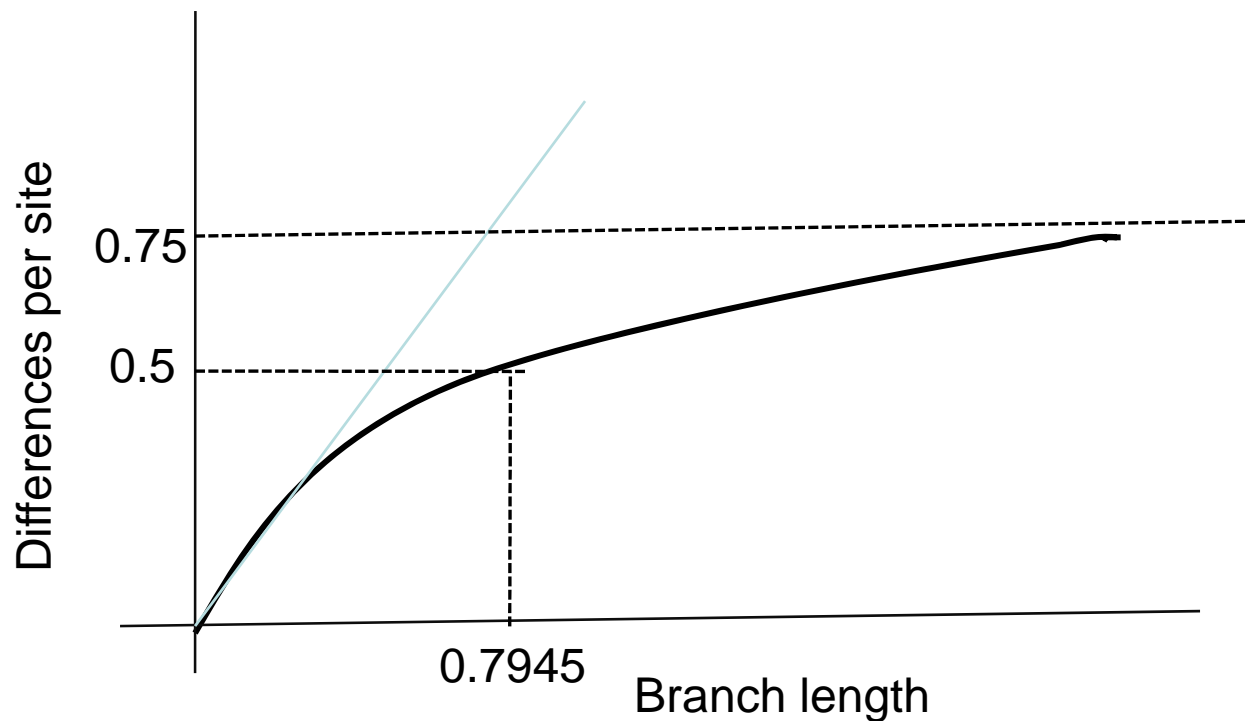
$$\Pr[x \rightarrow y] = \alpha(e), \quad x \neq y,$$

- Under the model, the distance $d(S, T)$ is defined as the expected number of substitutions that occurs at each sites in the evolution from the least common ancestor to S and T .



Under the Jukes-Cantor model, the distance $d(S, T)$ between S and T becomes

$$d(S, T) = -\frac{3}{4} \ln \left(1 - \frac{4}{3} \frac{\text{The number of mismatches}}{\text{The length of alignment}} \right).$$



2. Maximum Likelihood Method

- Given a set of k sequences of n characters each (Data)
- Assume an evolution model M (e.g. Jukes-Cantor):
 - A prior distribution of nucleotides at each site at root
 - The sites evolve independently and identically (i.i.d.)
 - The probability $p(x \rightarrow y / t)$ that a letter x is replaced by another y on a branch of length t for any x and y .
- Find a tree H of k leaves labeled with the that maximizes the conditional probability

$$L = \Pr[\text{Data} \mid H, M],$$

called the **likelihood**, under the model M .

- Since sites evolve i.i.d., using

$$L = \prod_{1 \leq i \leq n} \Pr[\text{Data}^{(i)} \mid H, M],$$

Likelihood calculation

Assume that at the i -th site, the sequences have letters T, A, A , respectively and the following tree H :

-- The root X may take one of the 4 possibilities:

A, G, C, T

-- The internal node Y has also 4 possible state A, G, C, T .

-- For each pair of specific states, say

$X=A, Y=G,$

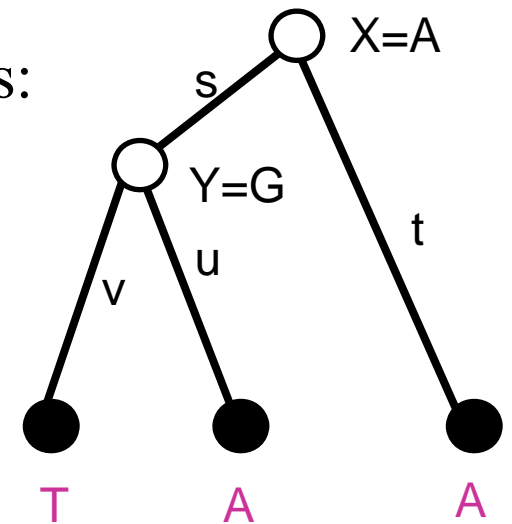
the evolution has the following probability:

$$p[X=A] p(A \rightarrow A | t) p(A \rightarrow G | s) p(G \rightarrow T | v) p(G \rightarrow A | u).$$

-- Considering all the possibilities, the probability is

$$\Pr[\text{Data}^{(i)} | H, M] = \sum_{S \in \Delta} \sum_{Q \in \Delta} p[X = S] p(S \rightarrow A | t) p(S \rightarrow Q | s) p(Q \rightarrow T | v) p(Q \rightarrow A | u)$$

where Δ is the alphabet containing A, G, C, T .



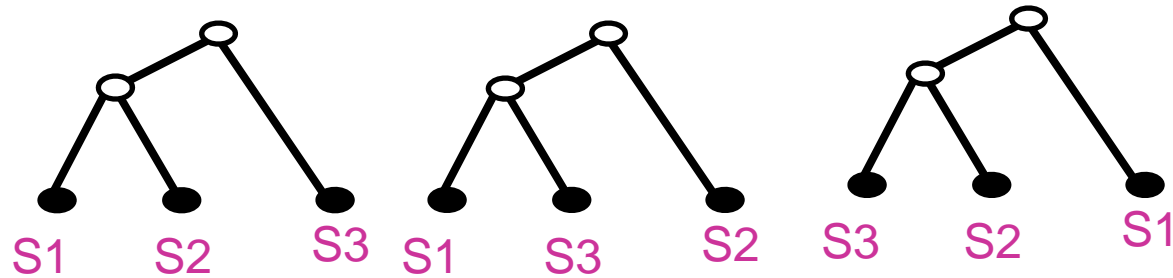
For example, for three sequences

S1: TGGT

S2: AGGT

S3: AGCG

We need to consider the following three different tree topologies:



For each tree, say the left one, the likelihood

$$\Pr[\text{Data} \mid H, M] = P_1 \times P_2 \times P_3 \times P_4$$

$$P_1 = \sum_{S \in \Delta} \sum_{Q \in \Delta} p[X = S] p(S \rightarrow A \mid t) p(S \rightarrow Q \mid s) p(Q \rightarrow T \mid v) p(Q \rightarrow A \mid u)$$

$$P_2 = \sum_{S \in \Delta} \sum_{Q \in \Delta} p[X = S] p(S \rightarrow G \mid t) p(S \rightarrow Q \mid s) p(Q \rightarrow G \mid v) p(Q \rightarrow G \mid u)$$

$$P_3 = \sum_{S \in \Delta} \sum_{Q \in \Delta} p[X = S] p(S \rightarrow C \mid t) p(S \rightarrow Q \mid s) p(Q \rightarrow G \mid v) p(Q \rightarrow G \mid u)$$

$$P_4 = \sum_{S \in \Delta} \sum_{Q \in \Delta} p[X = S] p(S \rightarrow G \mid t) p(S \rightarrow Q \mid s) p(Q \rightarrow T \mid v) p(Q \rightarrow T \mid u)$$

For each tree H , we need to choose 4 branch lengths to maximize the likelihood $\Pr[\text{Data} \mid H, M]$, containing $16 \times 16 \times 16 \times 16$ terms, each of which is a product of 5 probabilities.

So, the maximum likelihood is extremely time-consuming.

- Preferred by some systematists, but even harder than MP in practice. In practice, most systematic biologists use ML on small datasets,
- Theoretically, it is NP-hard.
- The main challenge here is to make it possible to obtain good solutions to ML in reasonable time periods on large datasets.

3. Consistence among trees and distance between trees

It is often to have two or more trees for the same group, often from different types of data or from different methods.

- How to put all the trees together to get one overall estimate of trees?

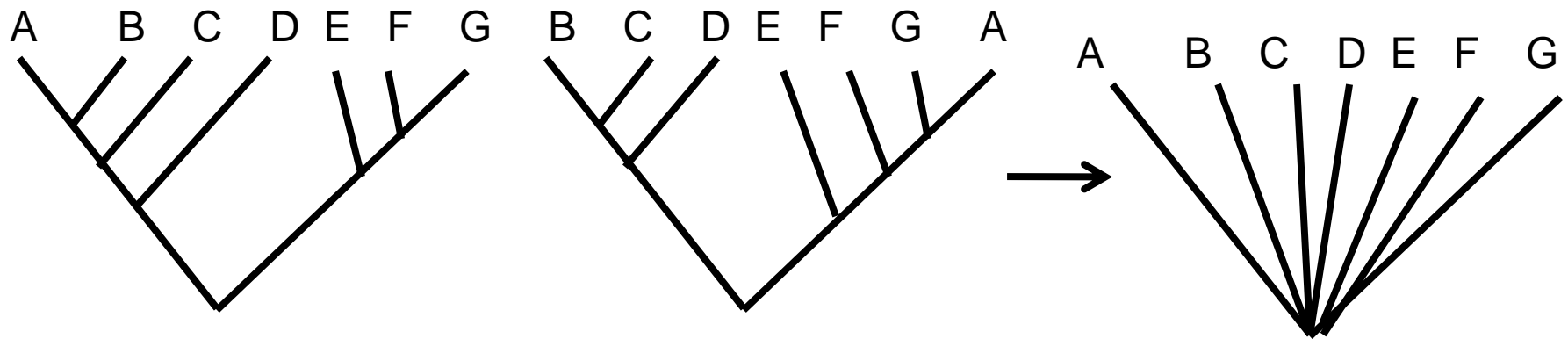
- How to measure the extend of difference between trees?

3.1. Computing consensus trees

Strict Consensus

- Each node presents a subset of leaf labels.
Given a set of trees, Strict consensus defines a tree that contains exactly all the subsets that are on all given rooted trees.





These trees differ only in the placement of A and are hence quite similar. Since the only common subset is the whole set of labels, their strict consensus is completely unresolved.

This shows that limitation of this consensus method.

- Remark:** (1) A consensus tree is not a phylogenetic tree
 (2) Consensus tree can also be defined for a set of unrooted trees.

Majority-Rule Consensus

-- Each node presents a subset of leaf labels.

Given a set of trees, under the majority-rule, the consensus tree contains all the subsets that occur in at least half of the given trees.

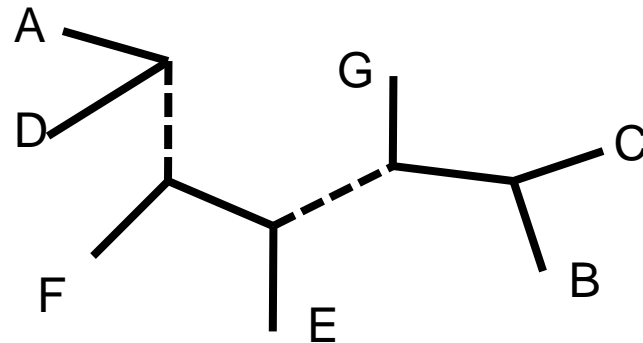
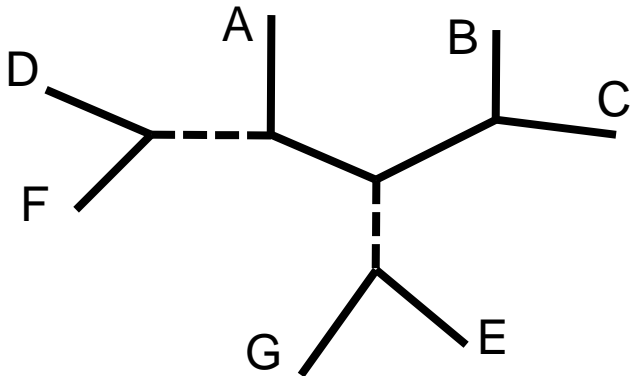
Theorem: The majority-rule consensus tree always exists.



3.2. Distance between trees

Partition metric or Robinson-Foulds distance for unrooted trees

- Each branch in a unrooted tree gives a partition of the set of leaf labels. So, a unrooted tree is uniquely defined by the partitions induced by its branches.
- The Robinson-Foulds distance between two trees is equal to the number of partitions that found in a tree but not in the other. The Robinson-Foulds distance is 4 for the following two trees.



4. Phylogenetic tree applications: Ancient Human Migration

- Two Hypotheses:

Multiregional hypothesis. Our immediate predecessors, now extinct, had wandered out of Africa as early as 2 million years ago. They had evolved into modern human beings, *Homo sapiens*, simultaneously in Africa, Europe and Asia.

'Out of Africa' hypothesis. Modern human beings, *Homo sapiens*, originated in Africa and had wandered out of Africa within the past 100,000 to 200,000 years.

- Use DNA variations in population to answer

Where did we come from?

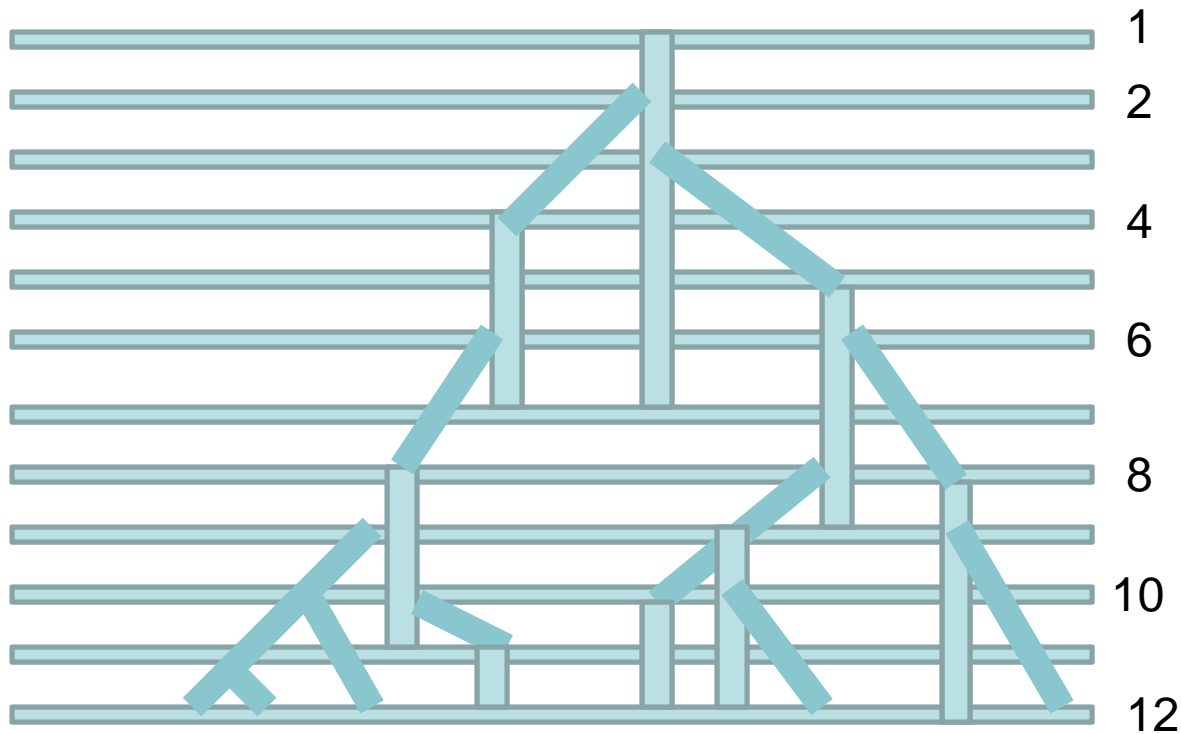
How did we get here?

The latest DNA studies indicate

- All modern human beings share a common female ancestor who lived 140,000 years ago.
- All humans share a common male ancestor who lived in Africa about 60,000 years ago.
- These were not the only humans who lived in these eras.
- Modern humans arose in sub-Saharan Africa and began migrating, starting about 65,000 years ago, to populate first southern Asia, China, Java, and later Europe.

Universal Maternal Ancestor

- All human beings must have an ultimate common female ancestor
- But she did not necessarily live along or in a small population.

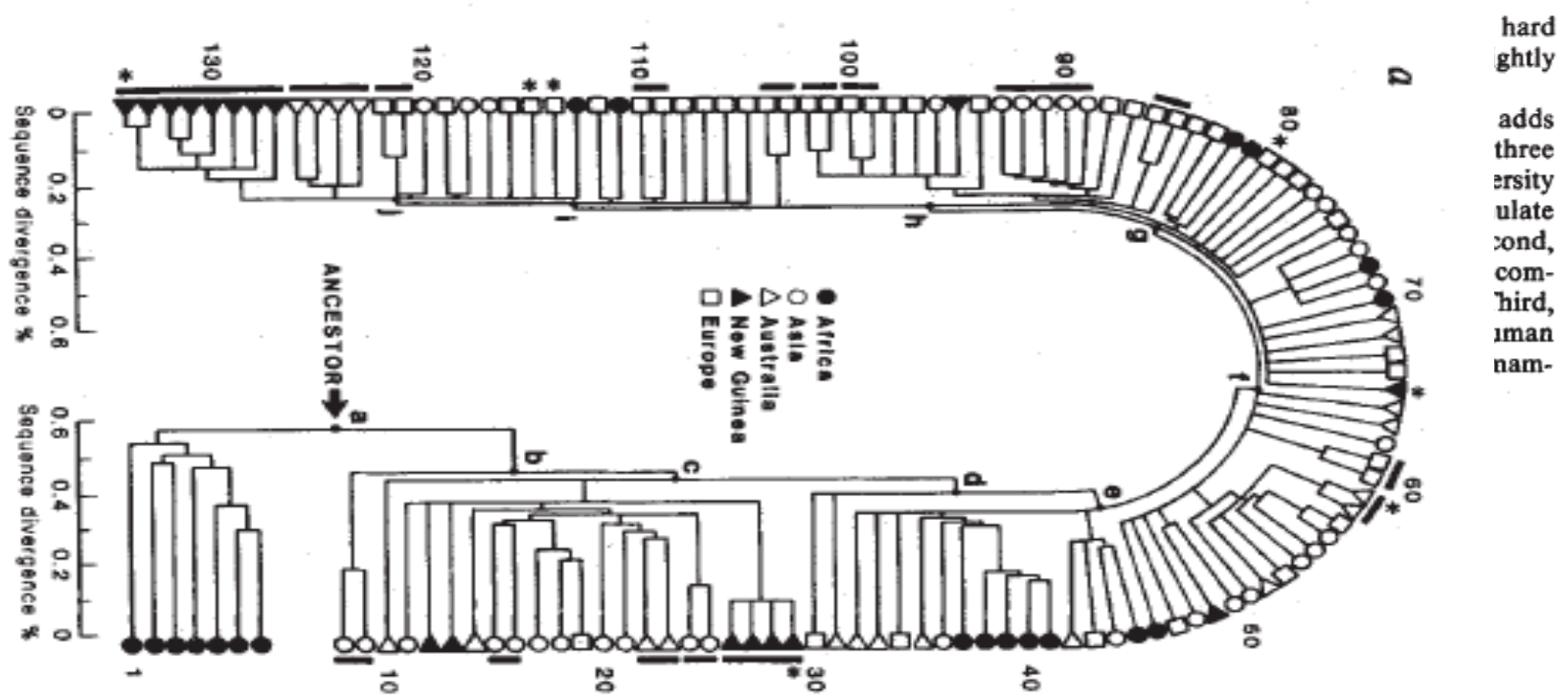


Mitochondrial DNA and human evolution

Rebecca L. Cann*, Mark Stoneking & Allan C. Wilson

Department of Biochemistry, University of California, Berkeley, California 94720, USA

Mitochondrial DNAs from 147 people, drawn from five geographic populations have been analysed by restriction mapping. All these mitochondrial DNAs stem from one woman who is postulated to have lived about 200,000 years ago, probably in Africa. All the populations examined except the African population have multiple origins, implying that each area was colonised repeatedly.



Mitochondrial genome variation and the origin of modern humans

Max Ingman*, Henrik Kaessmann†, Svante Pääbo† & Ulf Gyllensten*

* *Department of Genetics and Pathology, Section of Medical Genetics, Rudbeck Laboratory, University of Uppsala, S-751 85 Uppsala, Sweden*

† *Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany*

The analysis of mitochondrial DNA (mtDNA) has been a potent tool in our understanding of human evolution, owing to characteristics such as high copy number, apparent lack of recombination¹, high substitution rate² and maternal mode of inheritance³. However, almost all studies of human evolution based on mtDNA sequencing have been confined to the control region, which constitutes less than 7% of the mitochondrial

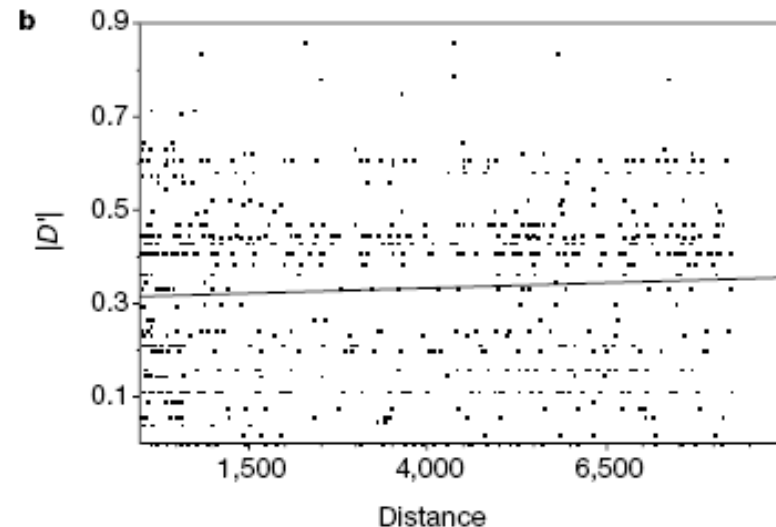


Figure 1 The relationship between linkage disequilibrium, measured by $|D'|$ versus distance between nucleotide sites for all 53 complete human mtDNA genomes. Values of ± 1.0 have been removed. **a**, Individuals of African descent ($n = 1,719$ comparisons), $R^2 = 0.001$; **b**, only non-African individuals ($n = 741$ comparisons), $R^2 = 0.005$.

The complete mitochondrial sequences of 53 people of diverse geographical, racial and linguistic backgrounds were selected. Each sequence has 16500 base pairs. 657 sites were used in the analysis. The mitochondrial DNA is chosen due to the following characteristics:

2. Analysis Method

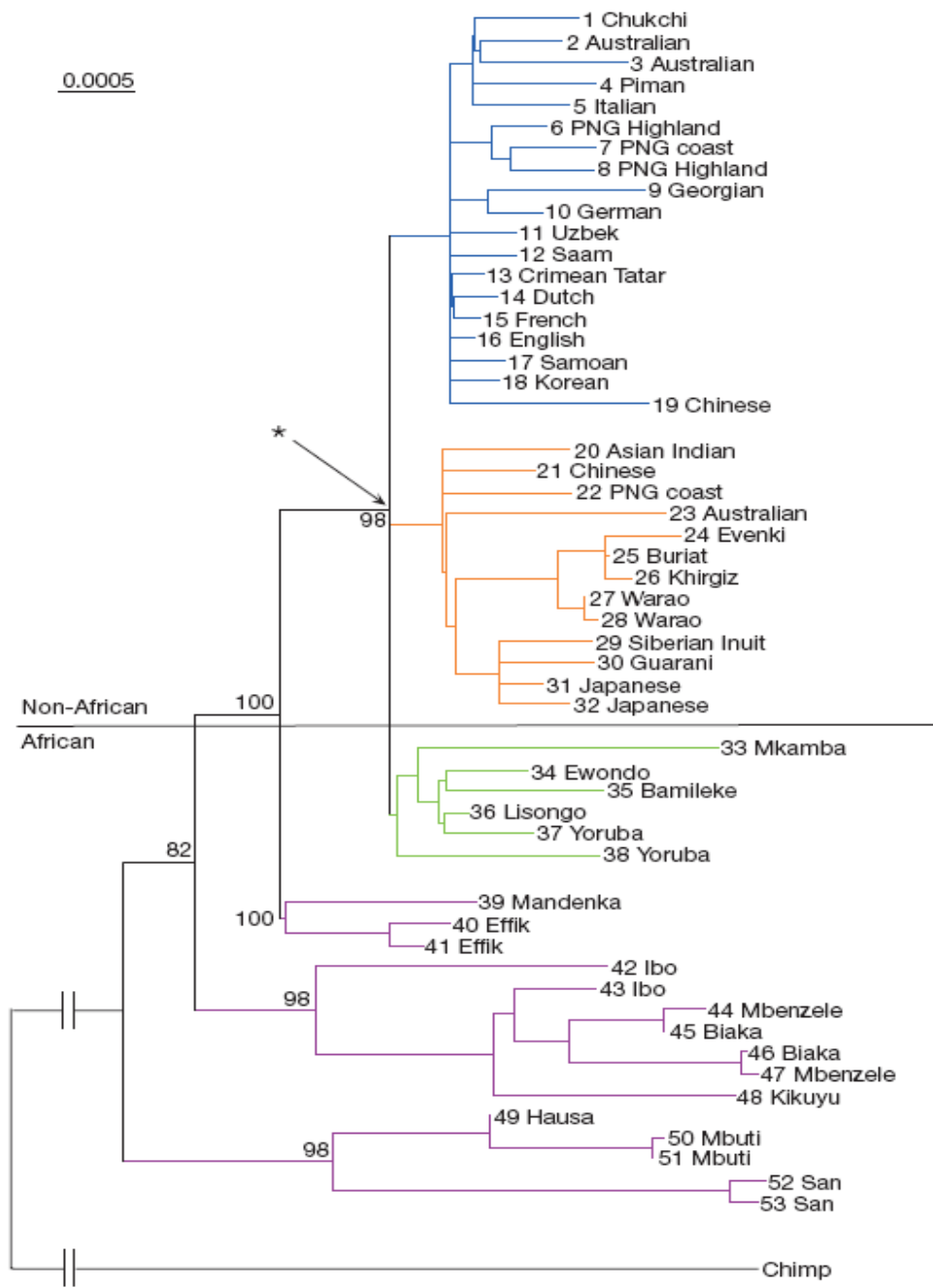
The pairwise distances between the sequences were calculated, vary from 6.0×10^{-5} per site to 6.8×10^{-3} . The average distance is 3.8×10^{-3} per site.

Neighbor-Joining Method was used.

3. Conclusion

The evolutionary tree constructed by Wilson *et al.* and Ingman *et al.* showed a trunk splitting into two major branches. One branch consisted only of Africans; the other contained peoples from everywhere else and some modern Africans.

All of the mtDNA, even samples from regions of the **world** far away from Africa, were highly similar. This suggested that our species is relatively young. But the African samples had the most mutations, thus implying that the African lineage is the oldest and that all modern humans trace their roots back to Africa. They further estimated that modern man emerged from Africa 200,000 years ago with racial differences arising only $52,000 \pm 28,000$ years ago.



4. Phylogenetic tree applications: Tracing HIV transmission

AIDS (acquired immune deficiency syndrome) is the notorious infectious disease now. It is caused by the retrovirus HIV (human immunodeficiency virus), which exists in two forms, HIV1 and HIV2. It was unknown until about 1981.

More than 30 millions are infected and more than 16 millions had already died.

Where did HIV originally come from?

- HIV1 came from chimpanzees (Gao *et al.*, 1999).
- HIV2 came from a type of monkey (Hahn *et al* 2000).
- Multiple introductions of HIV-1 and HIV-2 into Human last century (Korber *et al*, 2000).

All these conclusions are drawn by phylogenetic analysis. The rationale for using phylogenetic analysis is that HIV evolves rapidly. This is a bad property from the perspective of curing an infection. However, its rapid evolution enables fine-scale phylogenetic analysis.

Starting a single virus in a person, the infection will blossom in a tree structure; the viral lines ever expand as the infection continues. Two viruses from one person may have different sequences, but sequences of virus from one person are more similar to each other than they are to sequences of viruses from other person.

- **Florida dentist case.** A Florida dentist was suspected to transmit HIV to his patients (10 patients were eventually discovered to be HIV+). But the dentist was not the unique source of infection. A phylogenetic analysis was conducted to answer whether the dentist was the source of the infection (Ou *et al.* 1992). The output phylogeny includes a cluster of all the HIV+ patients but two and this dentist. This analysis suggested that the dentist was likely the source of infection for the eight patients and the other two had other risk factors.
- **Louisiana physician case** (State of Louisiana Criminal Docket #96CR733913). In December 1994, a woman in Lafayette was diagnosed with HIV and Hepatitis C. She suspected that her **former** lover, a physician, injected her with blood containing HIV in August 1994. During the period of time, he had been giving her vitamin B injections. The reason for this is that she had been negative for both HIV and Hepatitis C viruses at the time of blood donation in April 1990.

Records were discovered in the physician's clinic indicating that blood had been drawn from two patients during the week in question; one patient was previously known to be HIV+ and the other positive for hepatitis C.