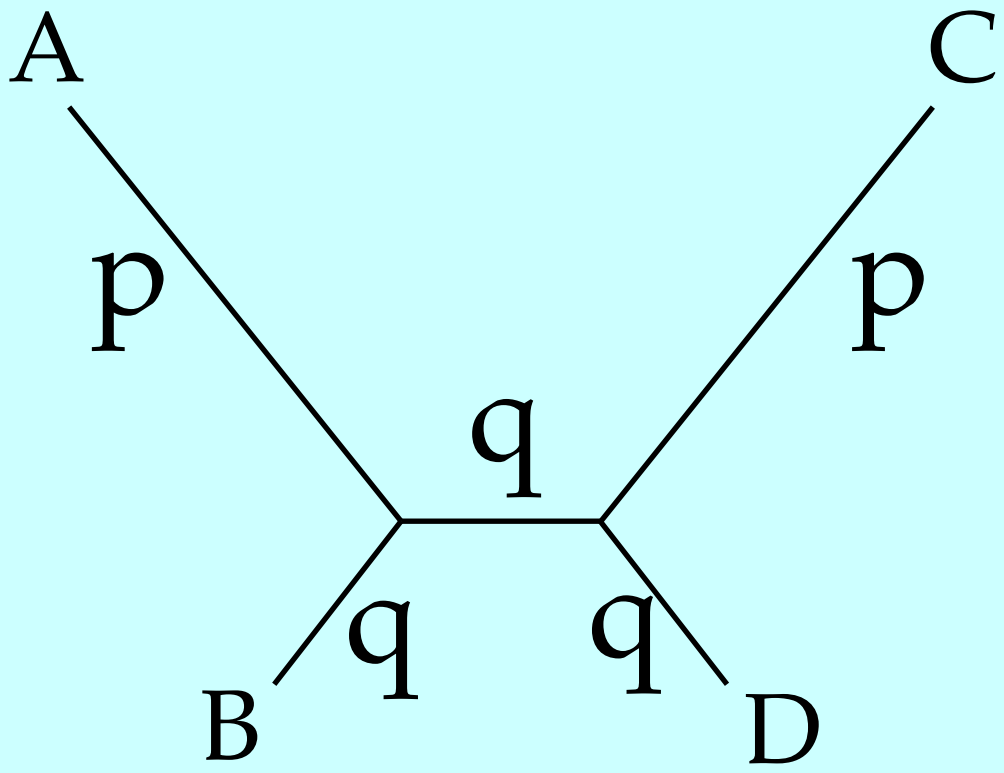
# Week 4: Consistency; History / philosophy, distance methods

Genome 570

January, 2016

## The counterexample




The long branches have probabilities of change $p$, the short branches have probabilities of change $q$. This is the canonical case of “long branch attraction”.

## Pattern probabilities

If tip pattern is 1100, and internal nodes are 1, 1 the probability is

$$\frac{1}{2}(1-p)(1-q)(1-q)pq$$

and in general summing over all four possibilities:

$$\begin{aligned} P_{1100} &= \tfrac{1}{2}\left((1-p)(1-q)^2pq + (1-p)^2(1-q)^2q \right. \\ &\quad \left. + p^2q^3 + pq(1-p)(1-q)^2\right) \end{aligned}$$

# Pattern probabilities

$$
\begin{aligned}
P_{xxyy} &= (1-p)(1-q)[q(1-q)(1-p) + q(1-q)p] \\
&\quad + pq[(1-q)^2(1-p) + q^2p]
\end{aligned}
$$

$$
\begin{aligned}
P_{xyxy} &= (1-p)q[q(1-q)p + q(1-q)(1-p)] \\
&\quad + p(1-q)[p(1-q)^2 + (1-p)q^2]
\end{aligned}
$$

$$
\begin{aligned}
P_{xyyx} &= (1-p)q[(1-p)q^2 + p(1-q)^2] \\
&\quad + p(1-q)[q(1-q)p + q(1-q)(1-p)]
\end{aligned}
$$

## Taking differences

$$P_{xyxy} - P_{xyyx} = (1-2q)\left[q^2(1-p)^2 + (1-q)^2p^2\right]$$

Which is always positive as long as $q < 1/2$ and either $p$ or $q$ is positive. Thus $P_{xyxy} > P_{xyyx}$ so we don't need to concern ourselves with $P_{xyyx}$.

To have $P_{xxyy}$ be the largest of the three, we only need to know that

$$P_{xxyy} - P_{xyxy} > 0$$

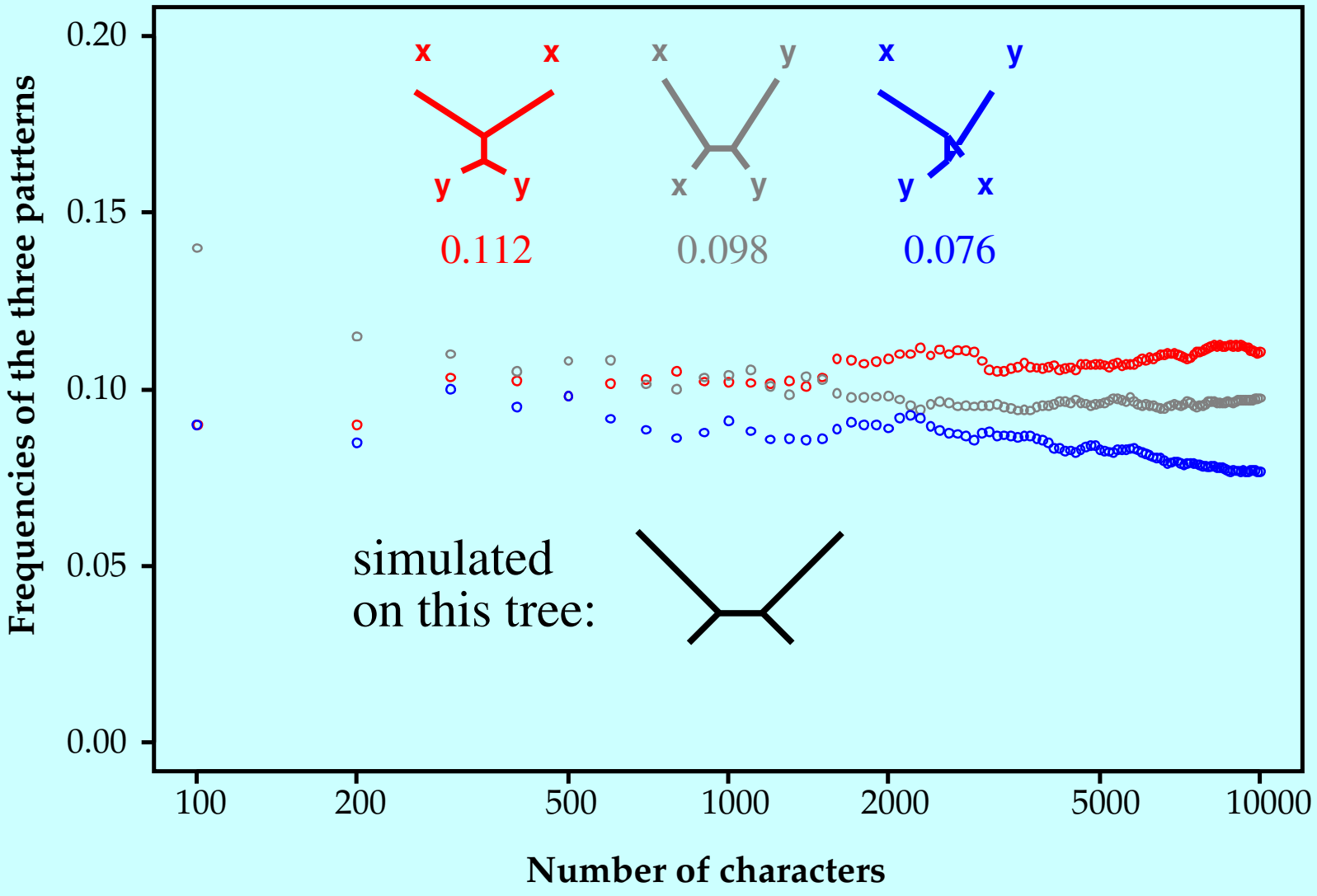
and after a struggle that turns out to require

$$(1-2q)\left[q(1-q) - p^2\right] > 0$$

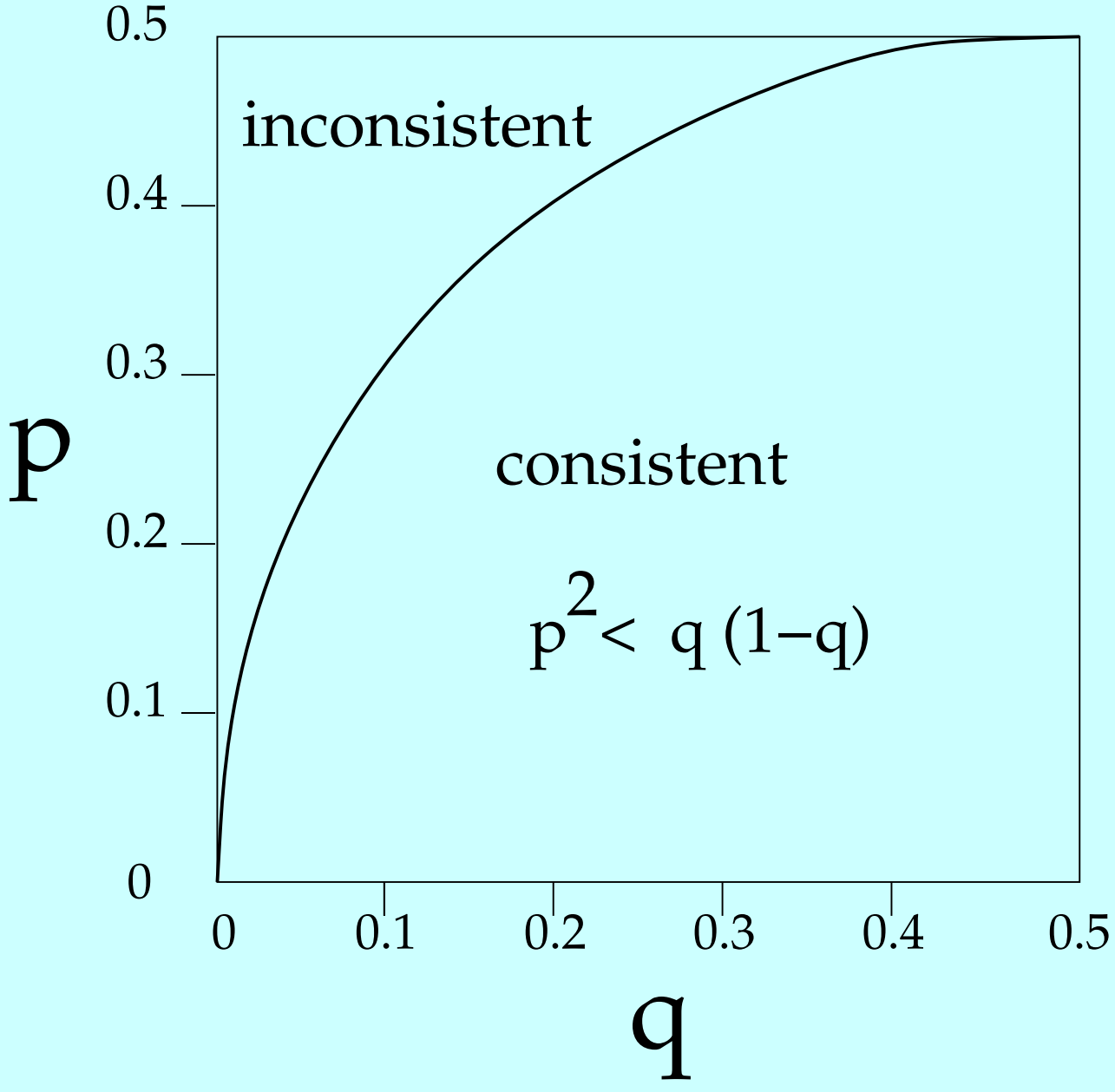
which (provided $q < 1/2$) is true if and only if

$$q(1-q) > p^2$$

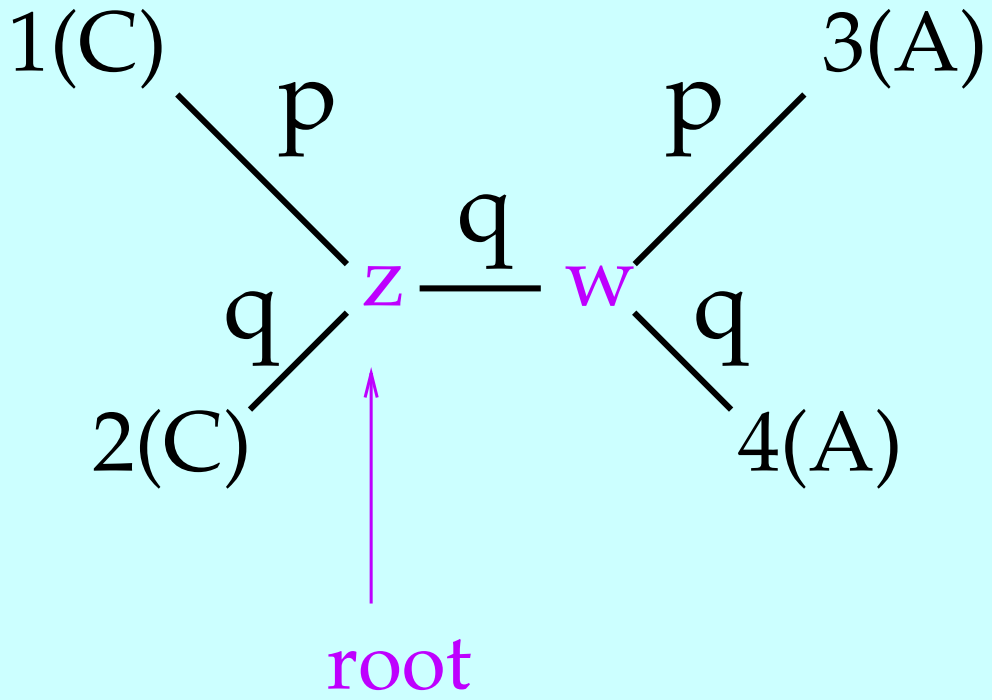
# Pattern frequencies win out



## Conditions for inconsistency



# Example for patterns with DNA



# Calculating a pattern frequency

$$\begin{aligned}\text{Prob}\,[\mathsf{CCAA}] \;=\; & \tfrac{1}{18}\,(1-\mathsf{p})(1-\mathsf{q})^2\mathsf{pq} \;+\; \tfrac{1}{27}\,\mathsf{pq}^2(1-\mathsf{p})(1-\mathsf{q}) \\ & +\; \tfrac{1}{162}\,\mathsf{p}^2\mathsf{q}^2(1-\mathsf{q}) \;+\; \tfrac{7}{972}\,\mathsf{p}^2\mathsf{q}^3 \\ & +\; \tfrac{1}{12}\,(1-\mathsf{p})^2(1-\mathsf{q})^2\mathsf{q}.\end{aligned}$$

# Pattern frequencies

$$\begin{aligned}
\text{Prob [xxyy]} &= (1-p)^2q(1-q)^2 + \tfrac{2}{3}\,p(1-p)q(1-q)^2 \\
&\quad + \tfrac{4}{9}\,p(1-p)q^2(1-q) + \tfrac{2}{27}\,p^2q^2(1-q) \\
&\quad + \tfrac{7}{81}\,p^2q^3
\end{aligned}$$

$$\begin{aligned}
\text{Prob [xyxy]} &= \tfrac{1}{3}\,(1-p)^2q^2(1-q) + \tfrac{2}{9}\,p(1-p)q^2(1-q) \\
&\quad + \tfrac{4}{27}\,p(1-p)q^3 + \tfrac{1}{3}\,p^2(1-q)^3 \\
&\quad + \tfrac{2}{9}\,p^2q^2(1-q) + \tfrac{2}{81}\,p^2q^3
\end{aligned}$$

$$\begin{aligned}
\text{Prob [xyyx]} &= \tfrac{1}{81}\,(1-p)^2q^3 + \tfrac{2}{3}\,p(1-p)q(1-q)^2 \\
&\quad + \tfrac{4}{9}\,p(1-p)q^3 + \tfrac{1}{9}\,p^2q(1-q)^2 \\
&\quad + \tfrac{6}{27}\,p^2q^2(1-q) + \tfrac{2}{81}\,p^2q^3.
\end{aligned}$$

# Conditions for inconsistency with DNA

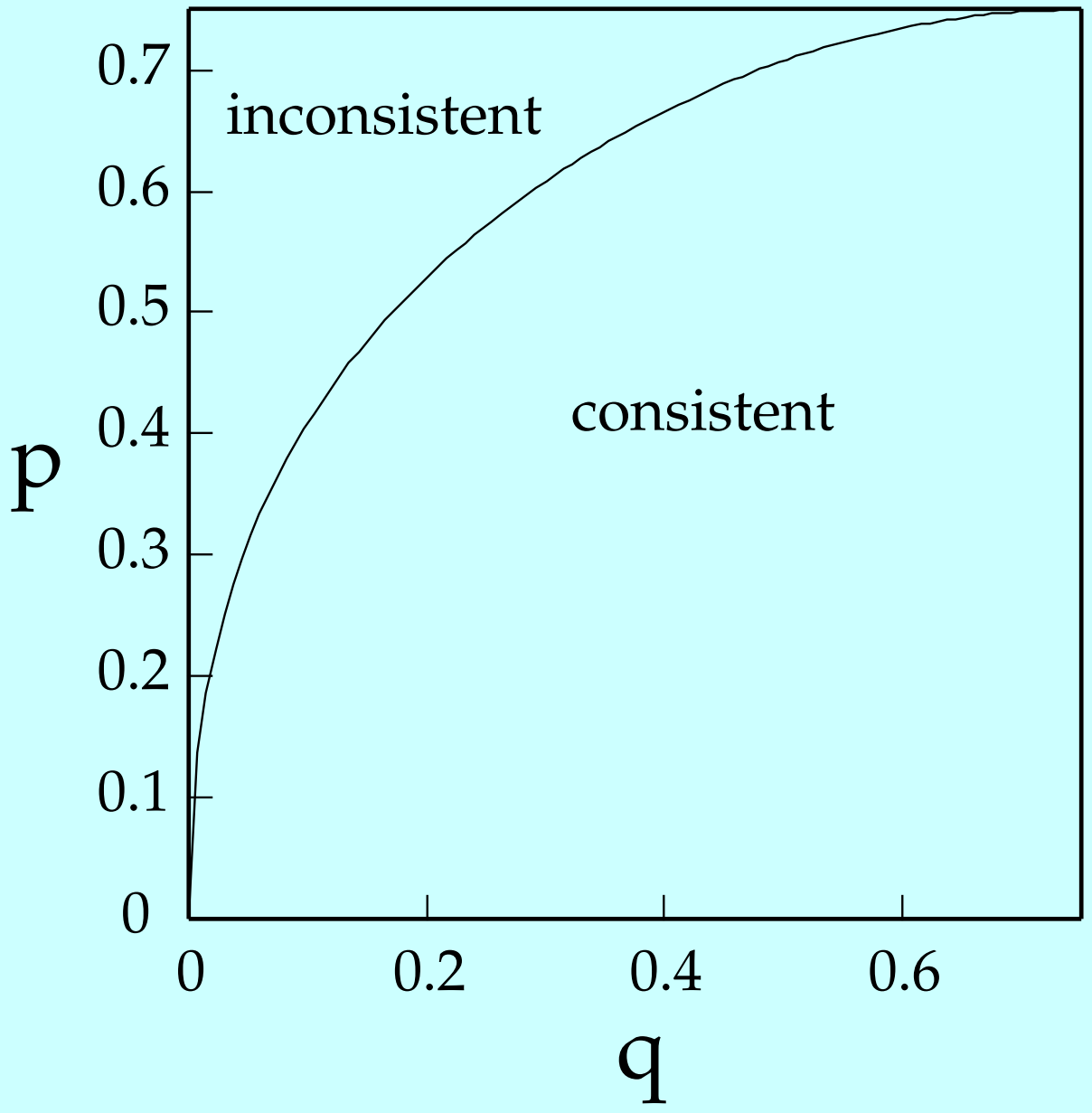
$$p < \frac{-18\,q + 24\,q^2 + \sqrt{243\,q - 567\,q^2 + 648\,q^3 - 288\,q^4}}{9 - 24\,q + 32\,q^2}$$

(This will *not* be on the test).

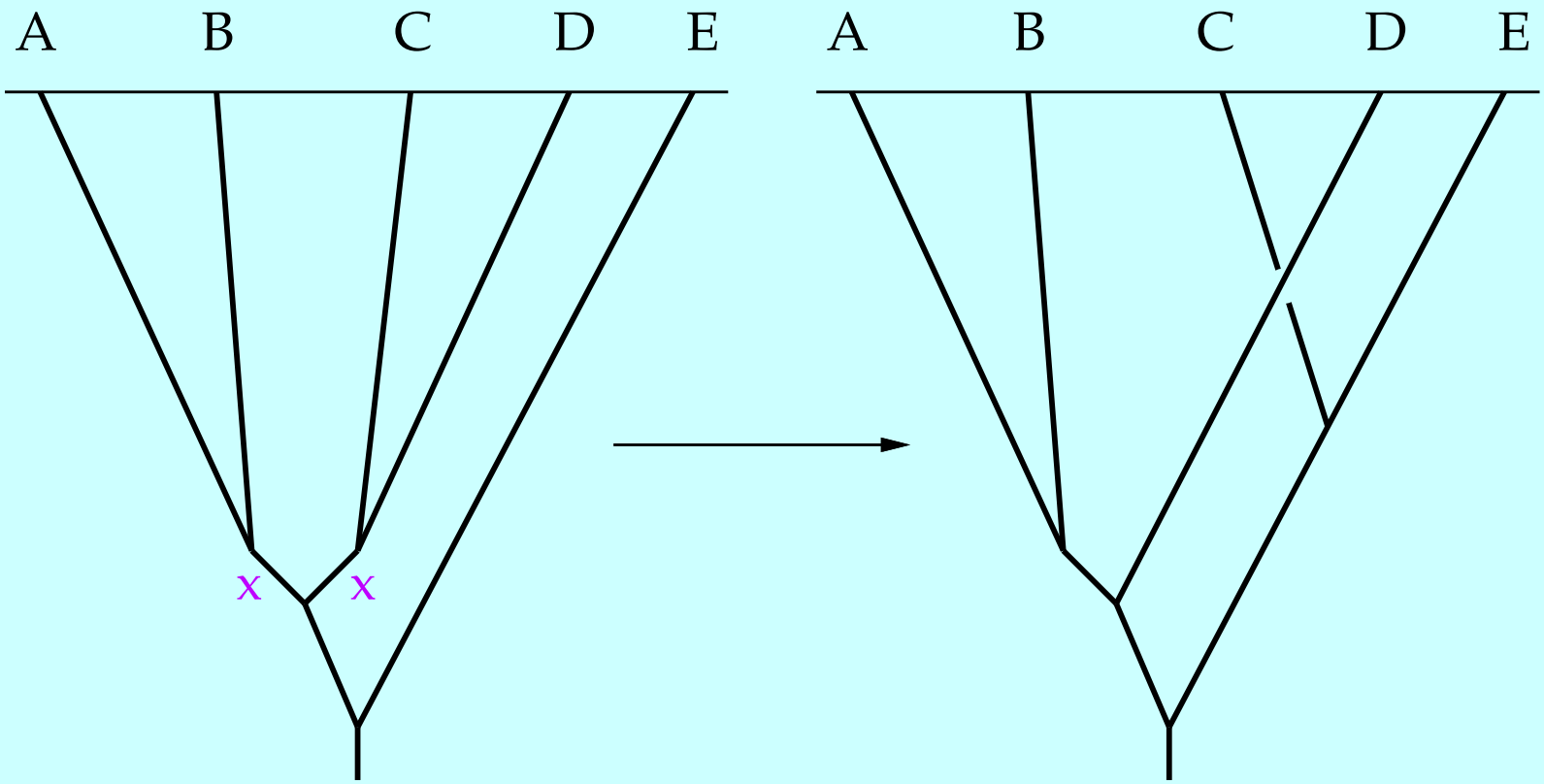
For small $p$ and $q$ this is approximately

$$\frac{1}{3}\,p^2 < q$$

# Conditions for inconsistency with DNA

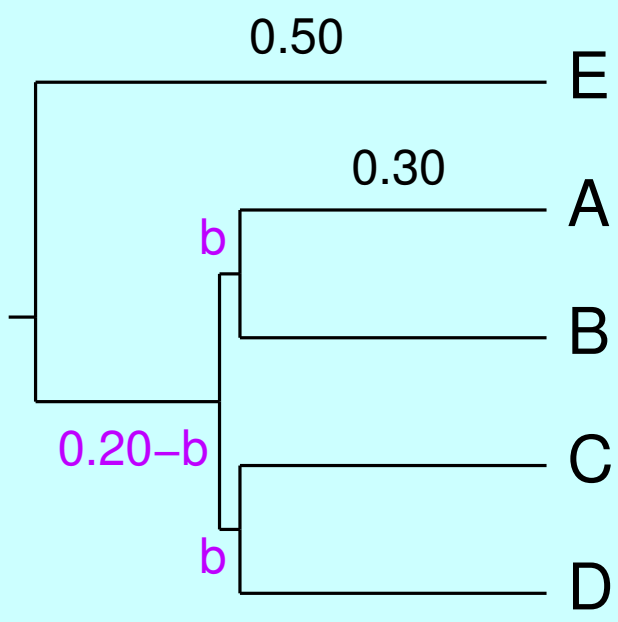


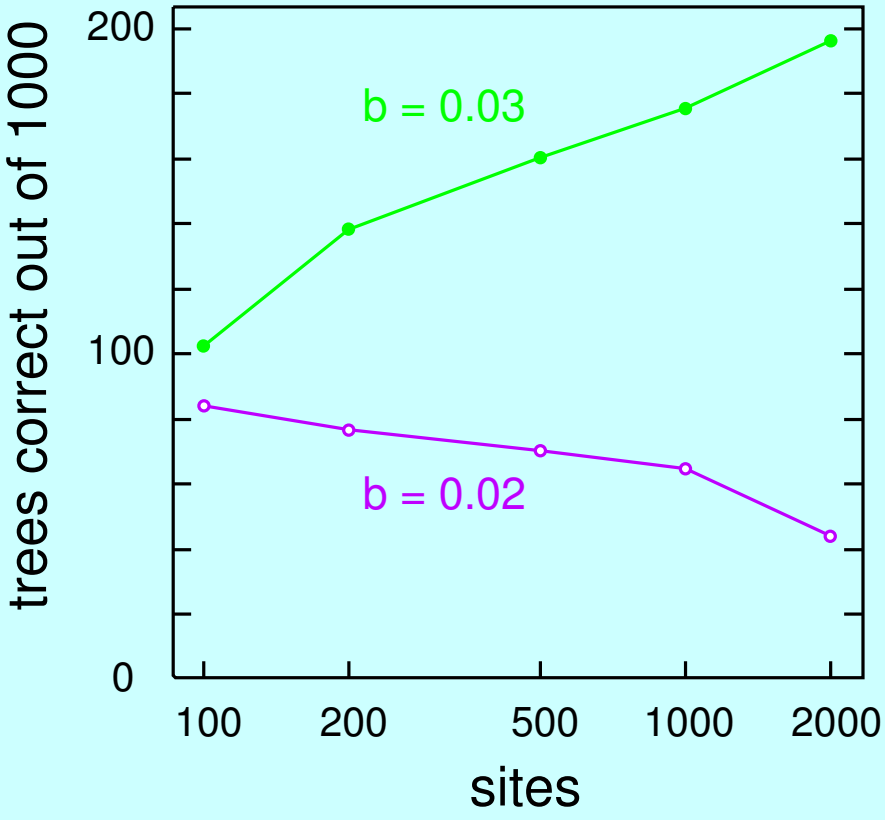
# Inconsistency with a clock




Cases like this were discovered by Michael Hendy and David Penny in 1989.

# Example showing inconsistency with a clock



# History and philosophy

Issues:

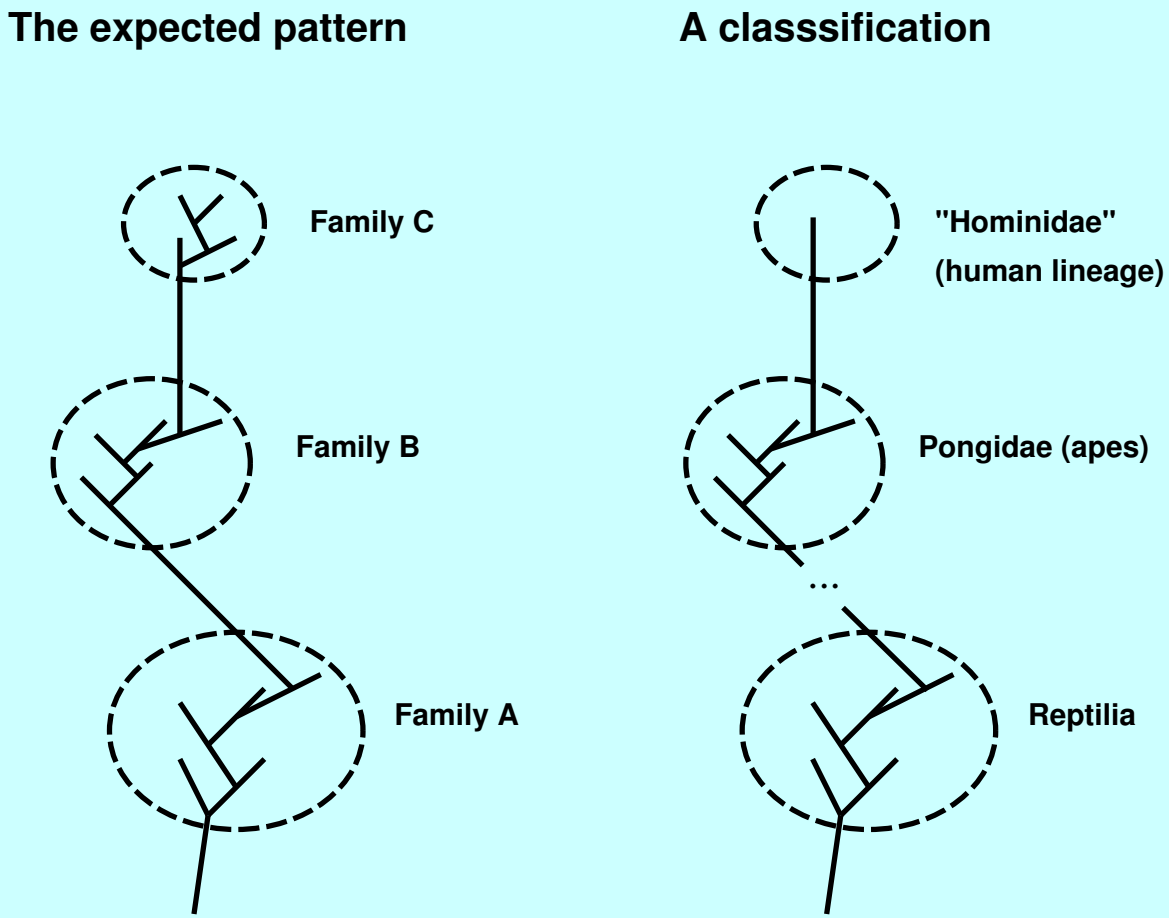
- How did work on numerical methods for phylogenies get started?
- What is the logical basis of inferring phylogenies?
- How does all that relate to classification?

# Ernst Mayr and George Gaylord Simpson

Ernst Mayr (1905-2005) George Gaylord Simpson (1902-1984)

- Major figures in the completion of the “modern synthesis” or “Neodarwinian synthesis” in the 1940s.
- Leaders of the “evolutionary systematics” approach to taxonomic classification, dominant until the 1970s.

# Evolutionary-systematic classification




A pattern of grades with very unequal rates of overall evolution is implicit in the use of paraphyletic groups in Mayr and Simpson's practice.

# A horse tree drawn by Simpson

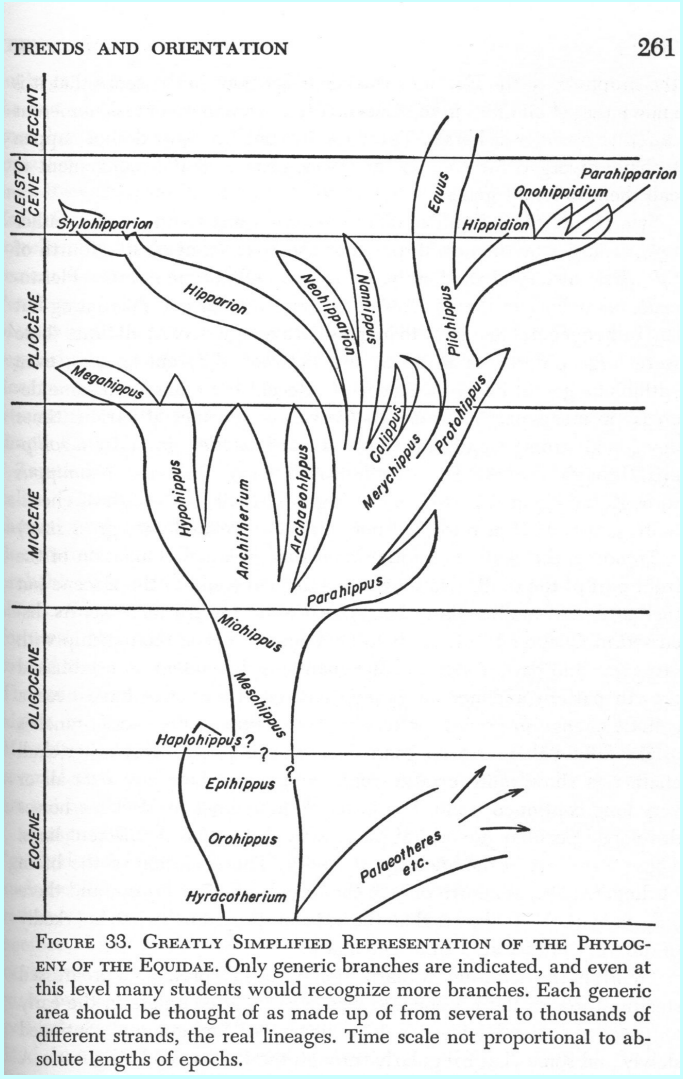
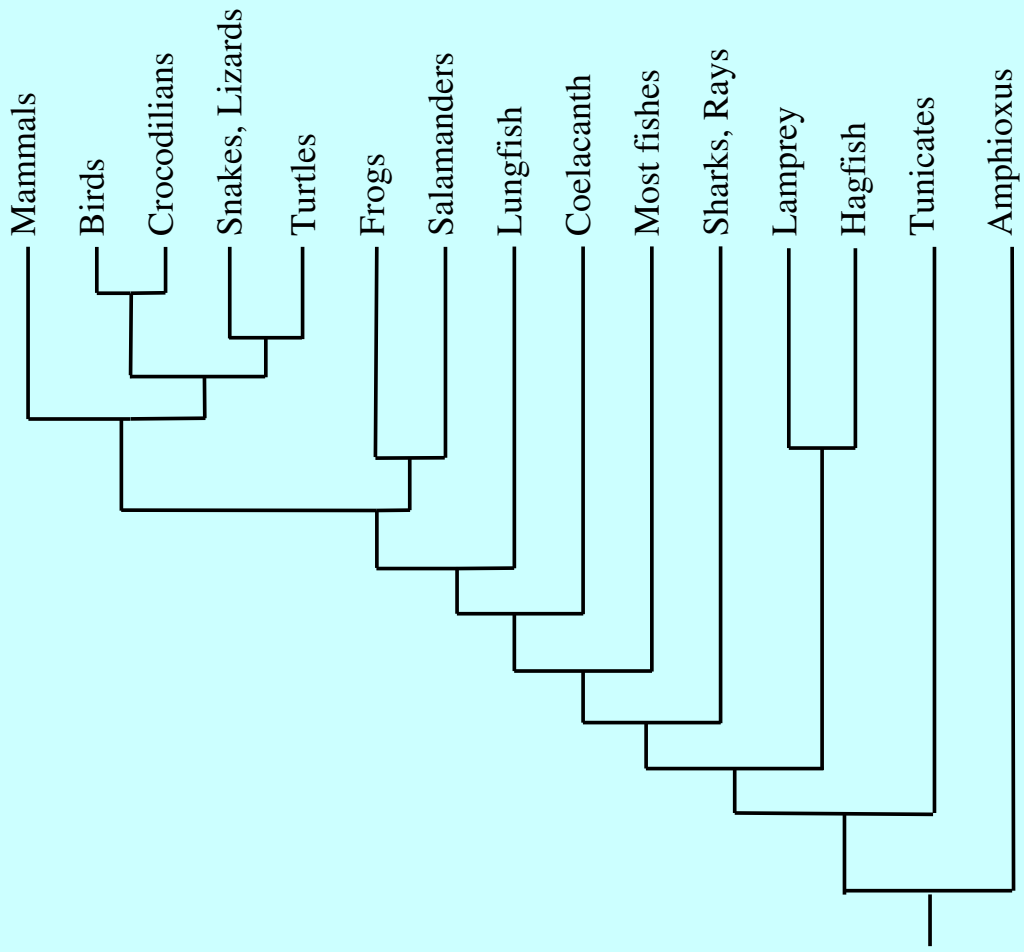



FIGURE 33. GREATLY SIMPLIFIED REPRESENTATION OF THE PHYLOGENY OF THE EQUIDAE. Only generic branches are indicated, and even at this level many students would recognize more branches. Each generic area should be thought of as made up of from several to thousands of different strands, the real lineages. Time scale not proportional to absolute lengths of epochs.

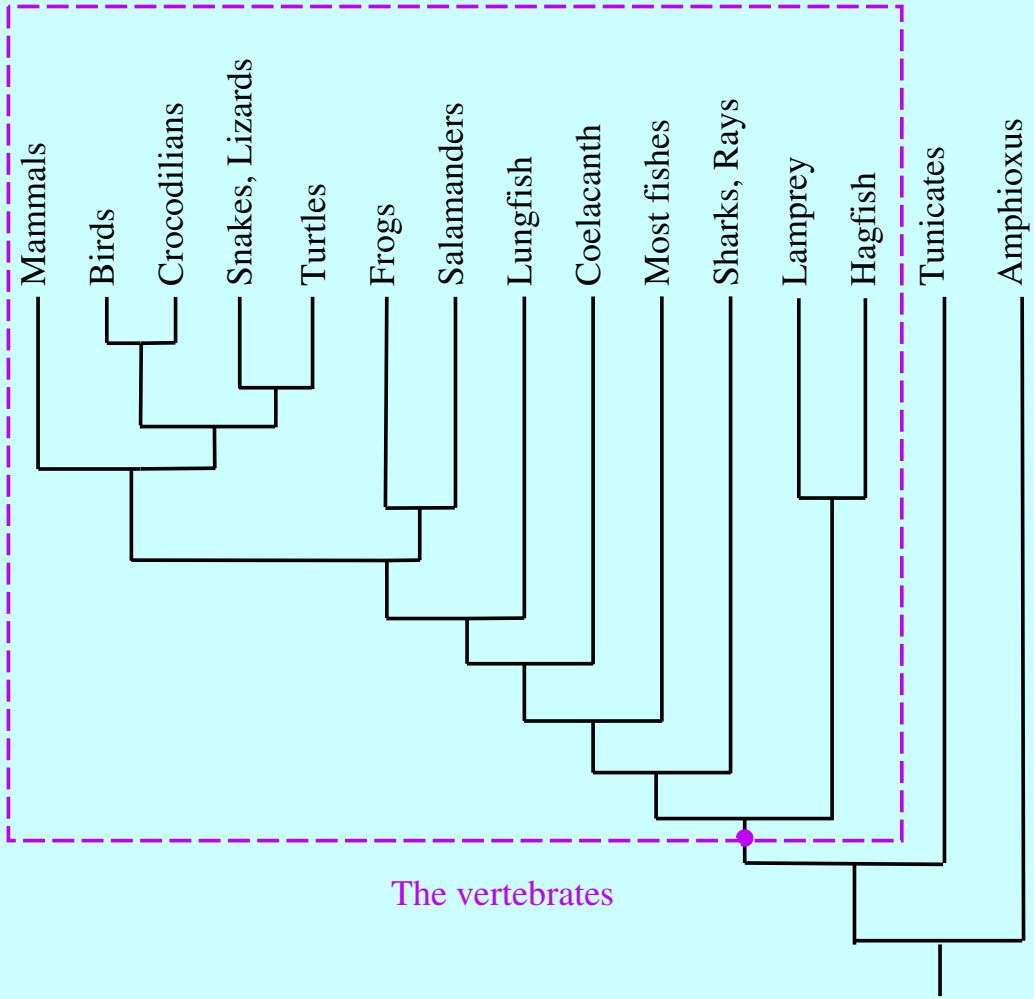
Note the “fat” branches which are somewhat ambiguous. As they emphasize classification over phylogeny, are the trees starting to dissolve?

# A phylogeny of the living chordates

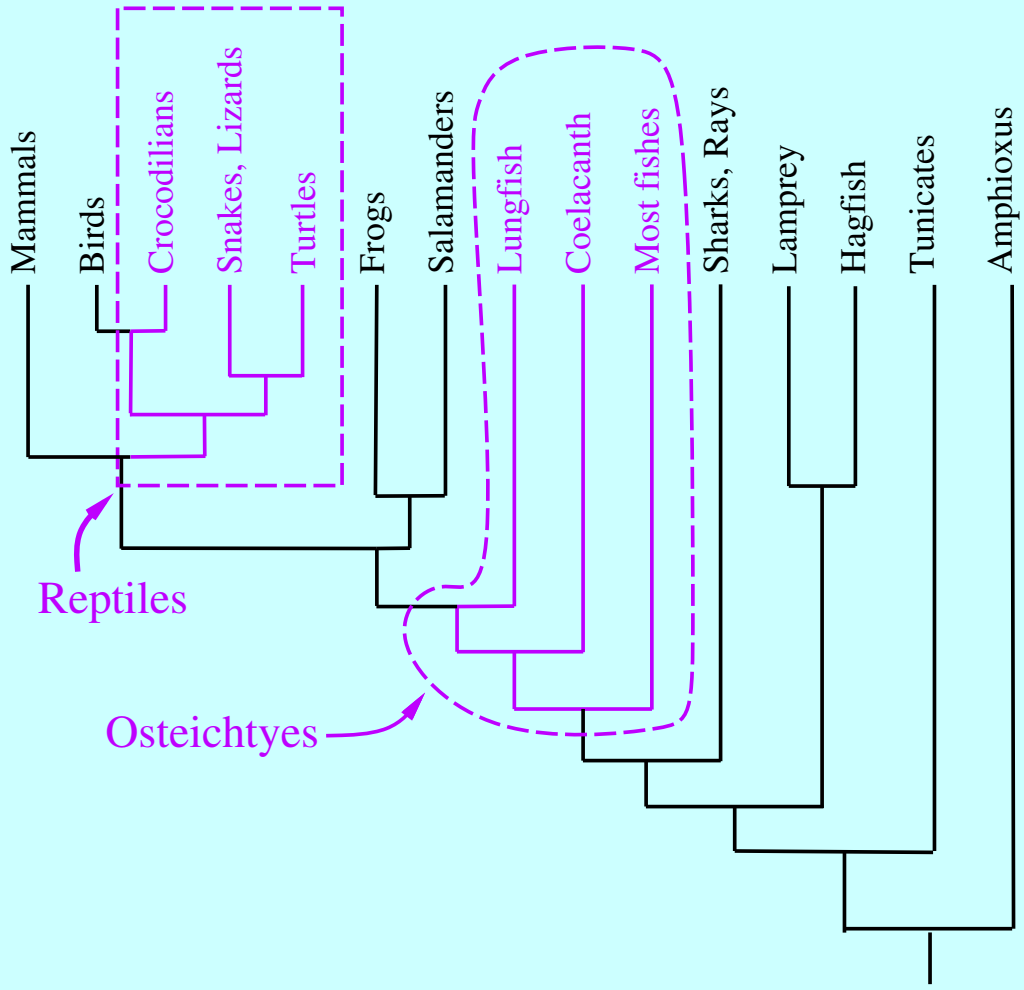



... as an example of groups that are in the traditional classification system but may or may not be monophyletic.

# Vertebrates are a monophyletic group



# Reptiles and fishes are paraphyletic groups




... since their most recent common ancestor is ancestral to other forms too (such as us).

## Positions on classification as of about 1960

- Evolutionary systematics. George Gaylord Simpson and Ernst Mayr led a movement that allowed non-monophyletic (paraphyletic) groups such as reptiles, on the assumption that groups could be separated by real differences of rates of evolution (sometimes "grades" rather than "clades").
- Phylogenetic systematics. Willi Hennig advocated purely monophyletic classification.
- Phenetics. Sokal and Sneath advocated making a classification without reference to evolution, using numerical clustering methods

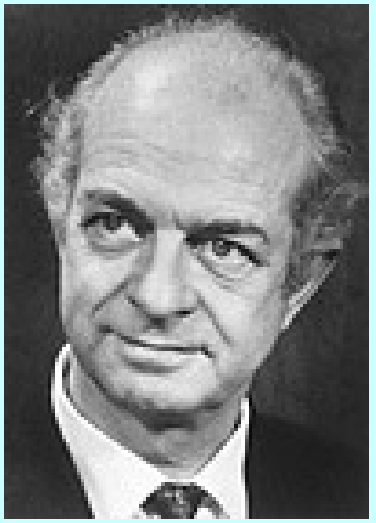
# Technological change post World War II

- Former physicists found molecular biology (first protein sequence, 1951)
- Former codebreakers and atomic bomb builders build the early computers (first stored-program digital computer, 1949)
- Most U.S. universities got their first computer about 1957.
- First sequences of same gene in multiple species in late 1950s.

# Molecular evolution gets off the ground

Zuckerkandl and Pauling in 1962 discussed using trees to infer ancestral sequences, and named this “chemical paleogenetics". They were about 30 years ahead of their time.

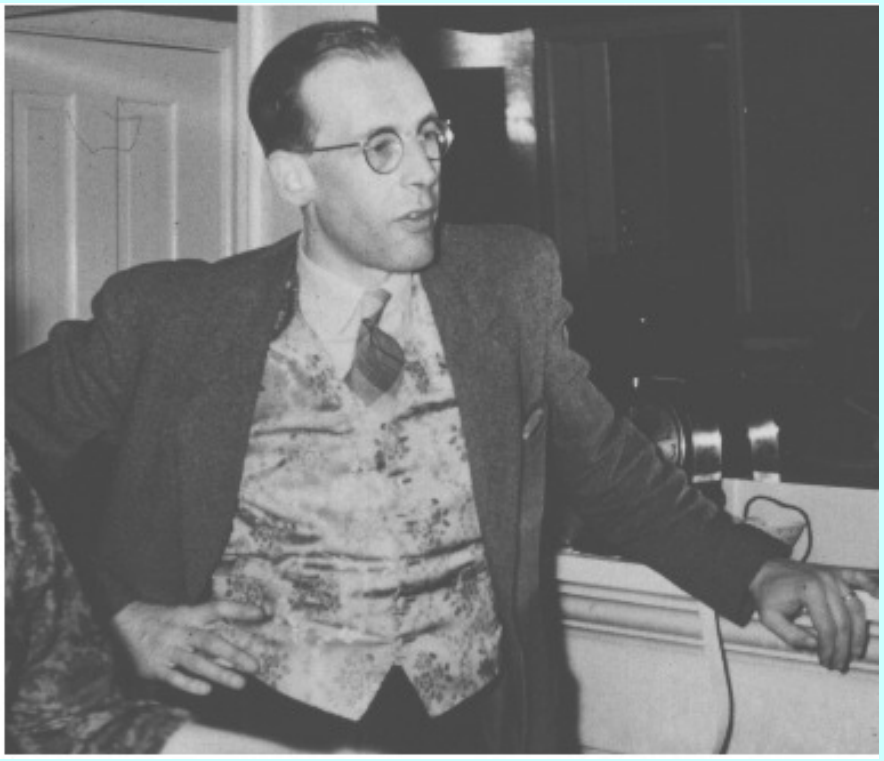
Emile Zuckerkandl in 1986

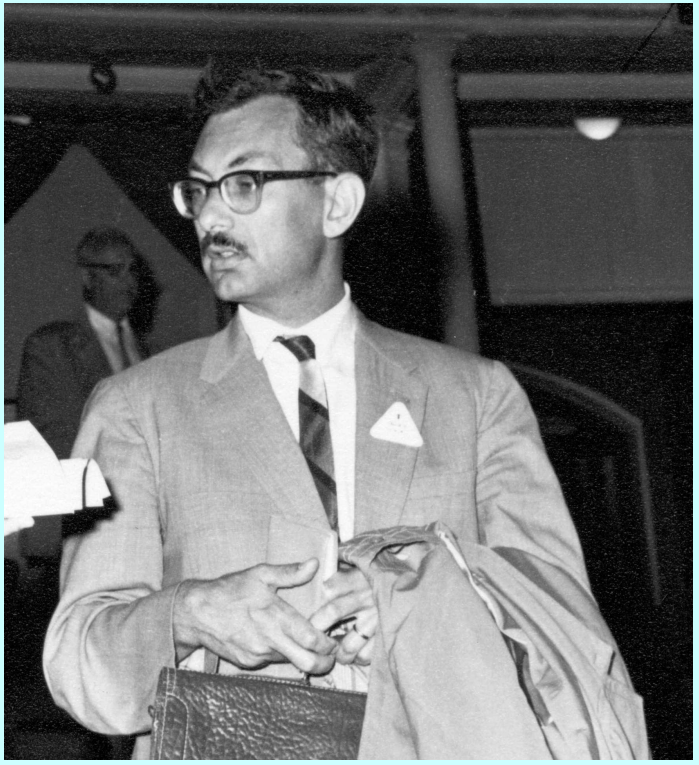


Linus in 1962, *from Nobel Peace Prize web page*

(But then it isn’t fair to anyone to compare them unfavorably to Linus Pauling).

# Peter Sneath in 1962 and Robert Sokal in 1964





... as they were advocating phenetic classification. Sokal did pioneering investigations of parsimony methods – intending to show that they wouldn't work well!

# The first numerical phylogeny, by Sokal and Michener 195

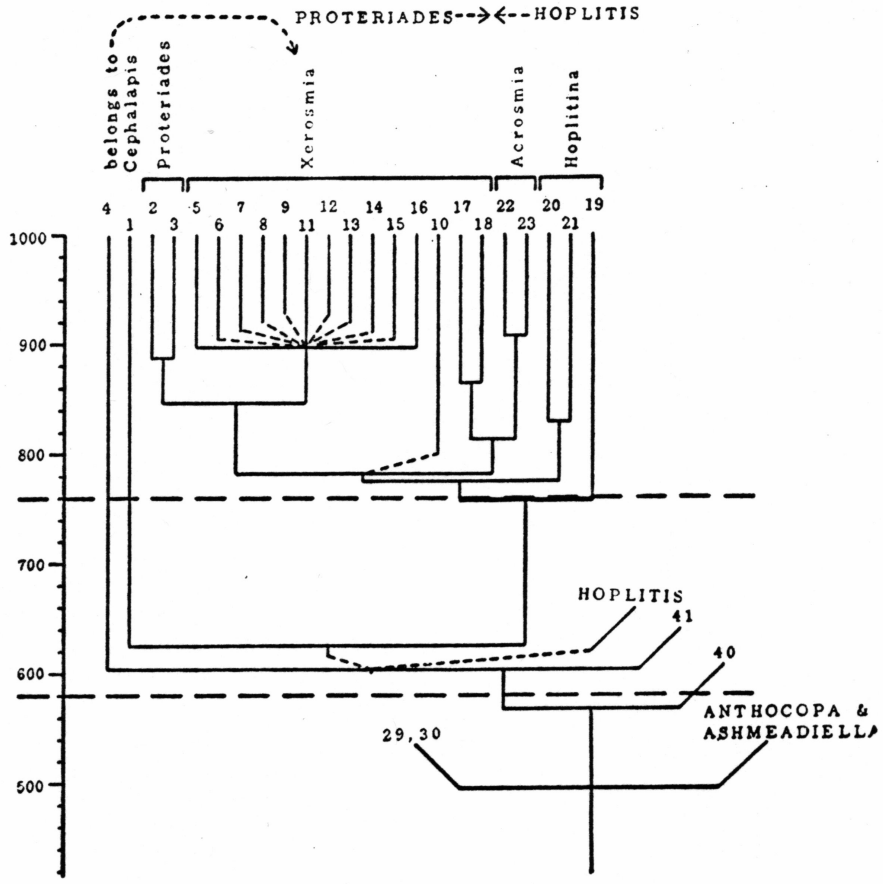
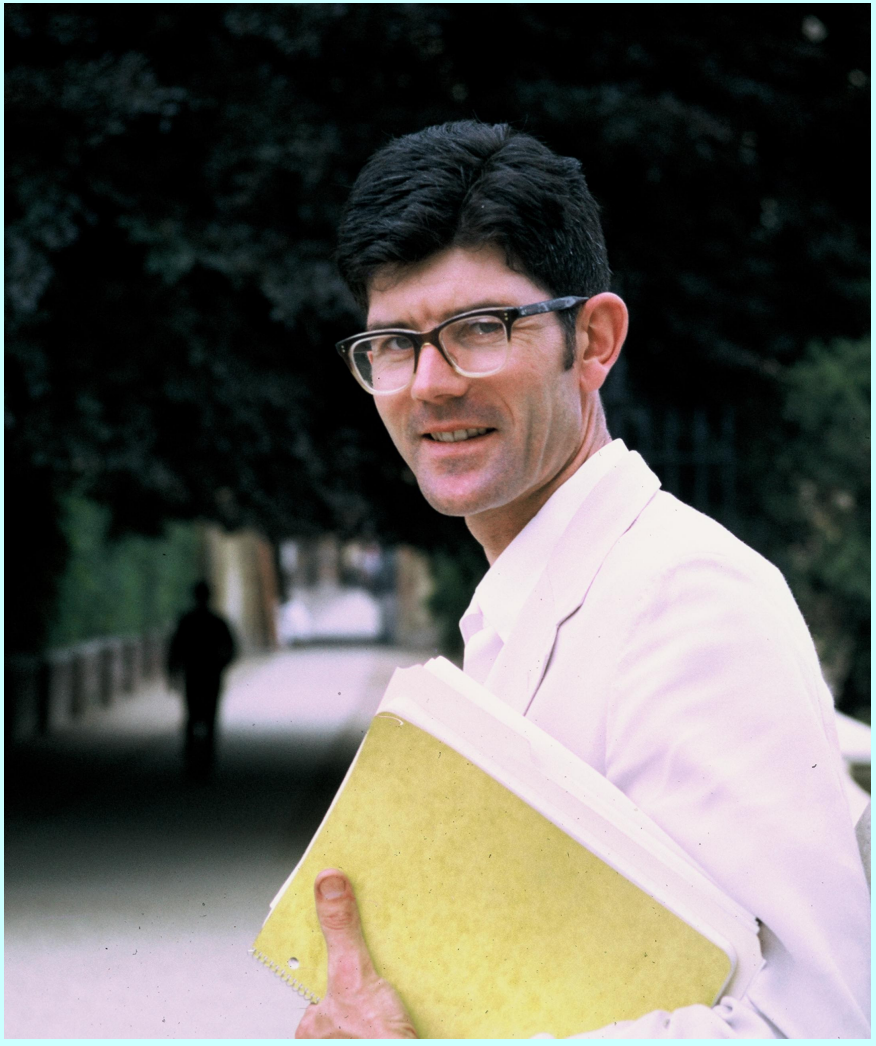



FIG. 5. Diagram of relationships for the genus *Proteriades* obtained by the weighted variable group method.

A tree of bees (Michener is the world's greatest bee systematist). Michener was the one who wanted to interpret this as a phylogeny. It was inferred by a clustering method (not, as misstated in my book, by UPGMA).

# Cavalli-Sforza and Edwards, 1963; Edwards, 1970



The picture on the left was taken by the famous population geneticist Motoo Kimura when he and his family visited Pavia and were taken on an excursion to a vineyard.

# The first phylogeny by parsimony

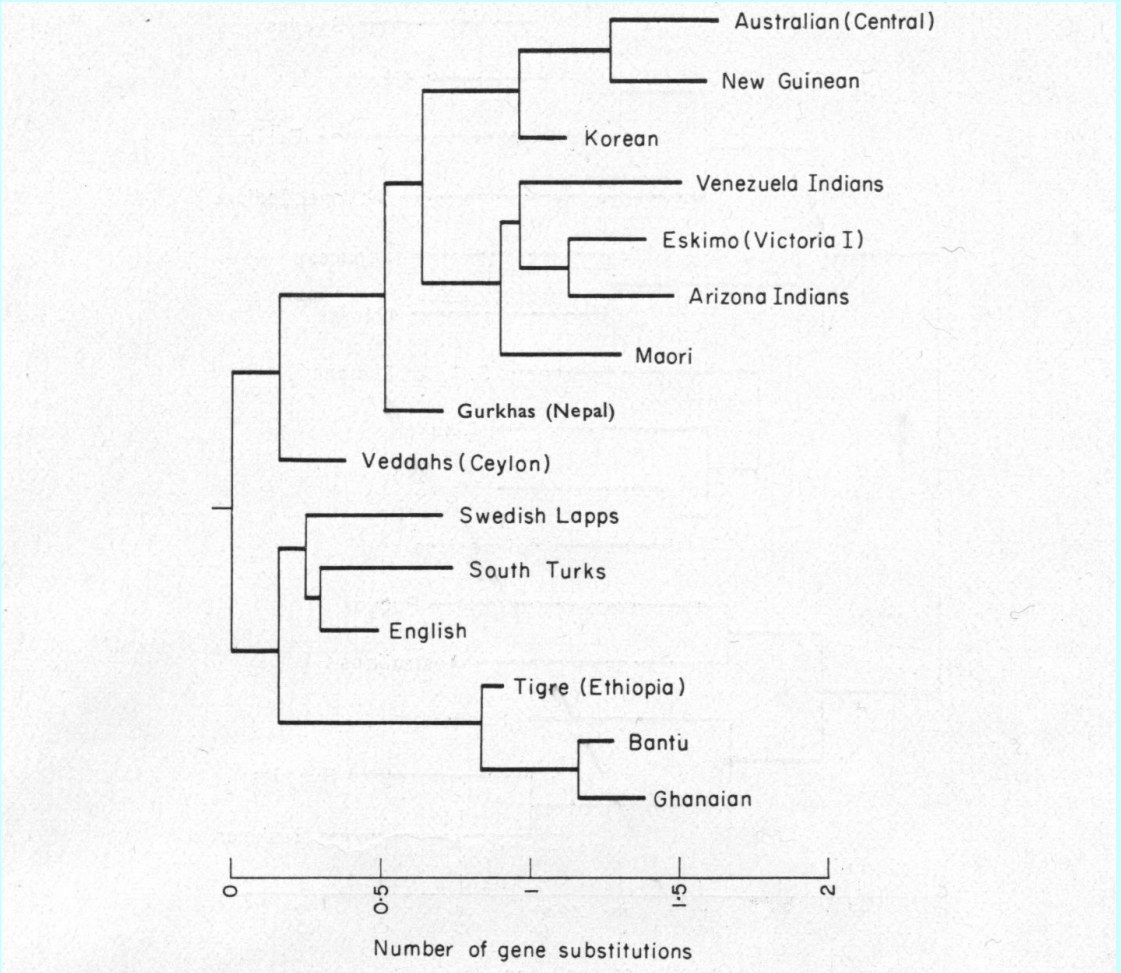



FIG. 5. Tree computed from blood group gene frequencies.

Gene frequencies of blood group polymorphisms. This is by minimum length in a space of gene frequencies.

# That tree drawn out on a map

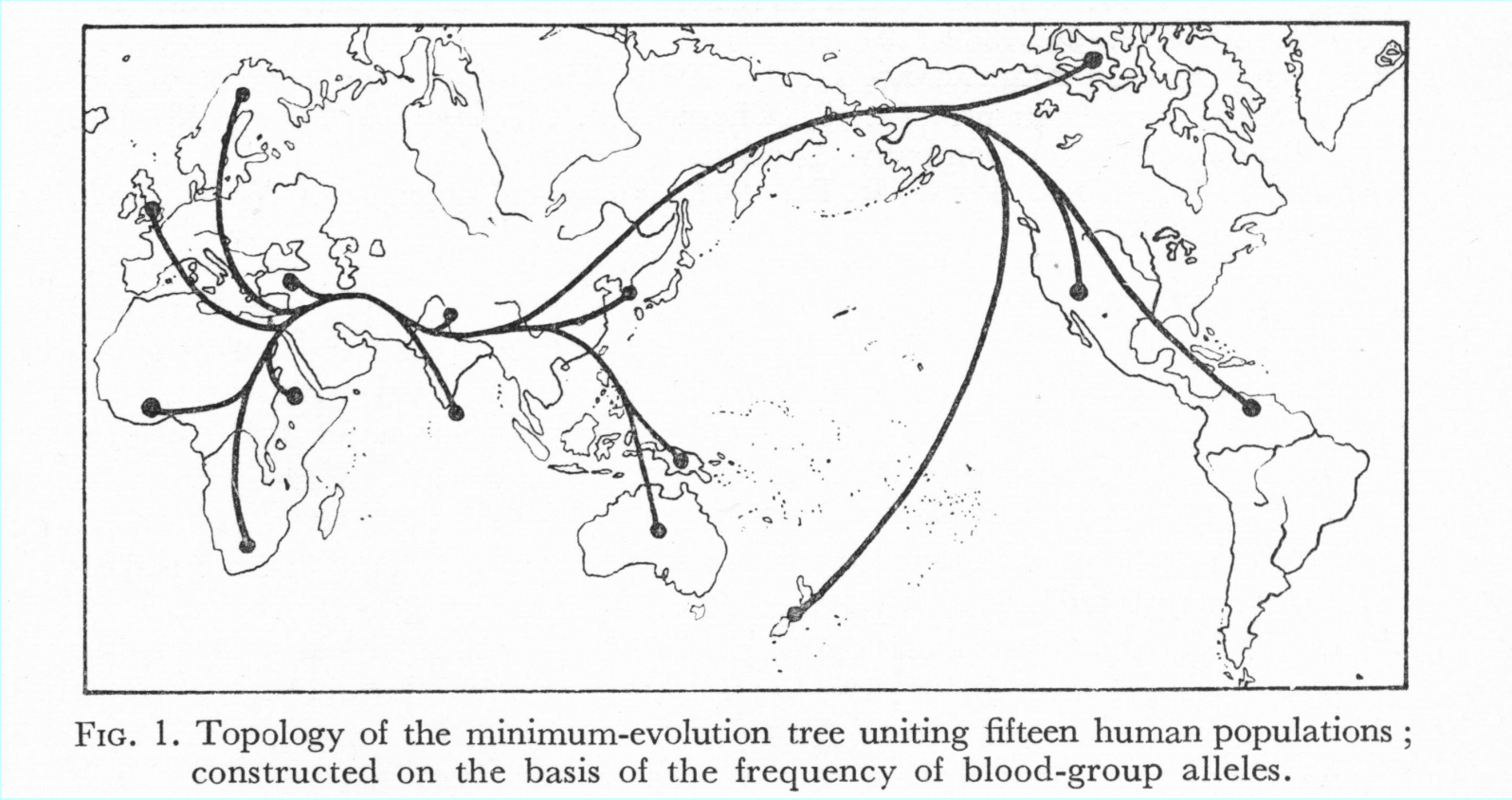
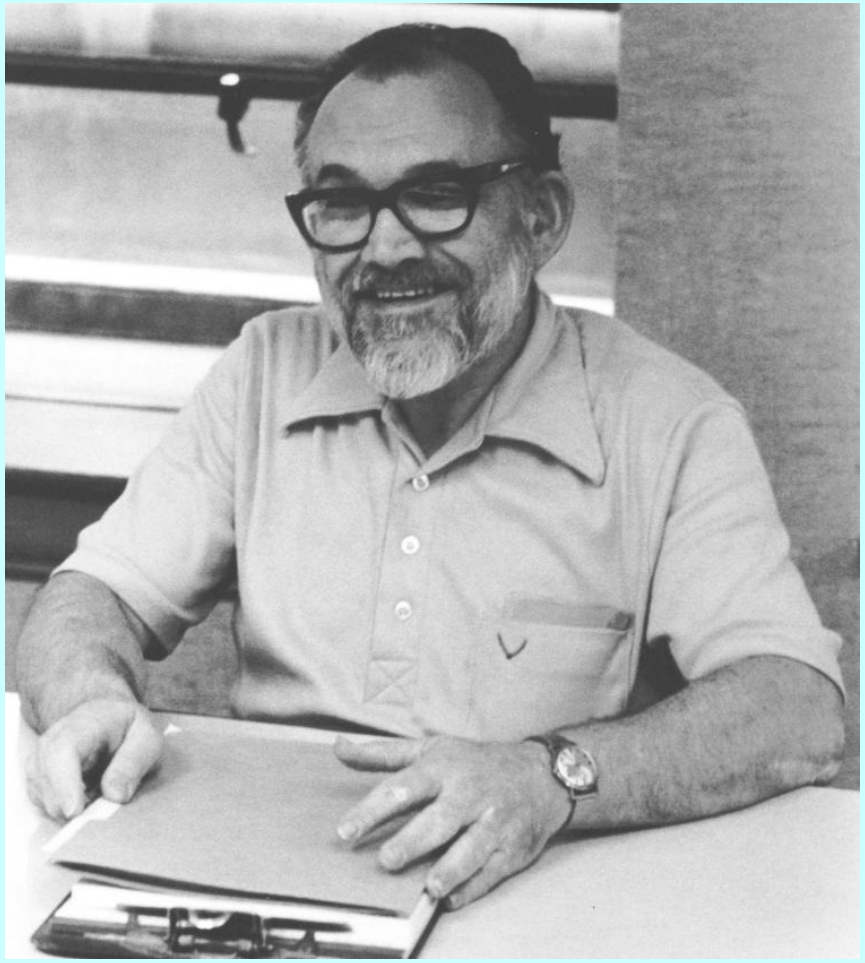
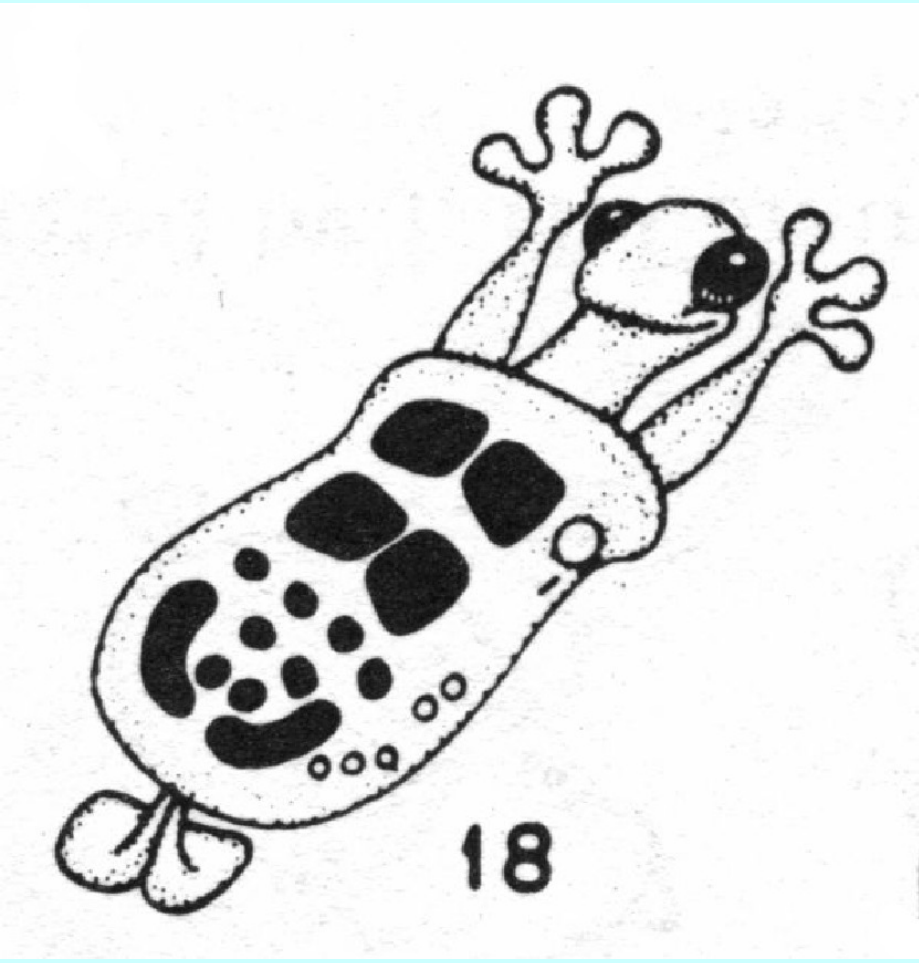


FIG. 1. Topology of the minimum-evolution tree uniting fifteen human populations; constructed on the basis of the frequency of blood-group alleles.

Note that the lineages from the Northwest Coast going down to Polynesia are dependent on where on the map the splits are placed and that is somewhat arbitrary.

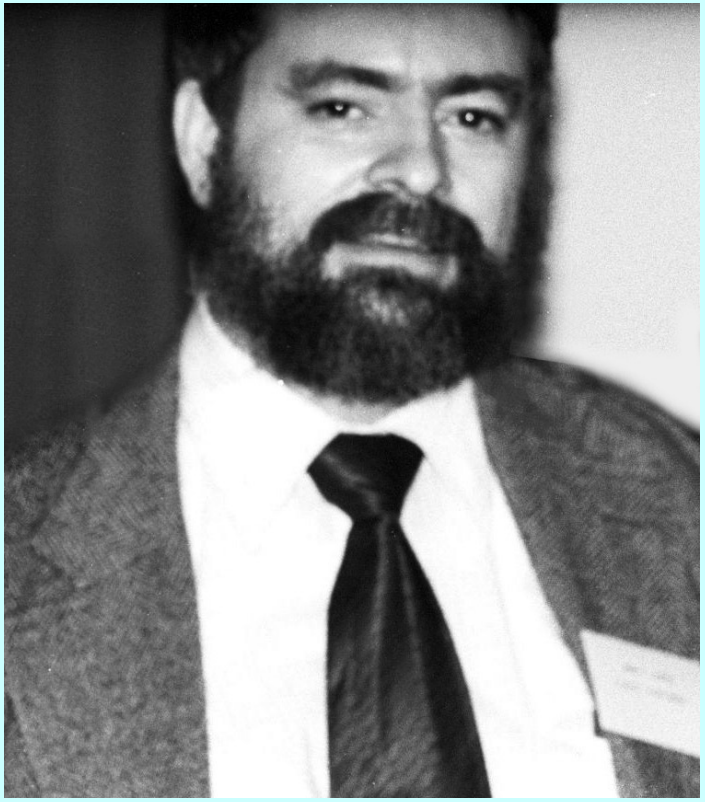
# Camin in the 1970s, one of the Caminalcules

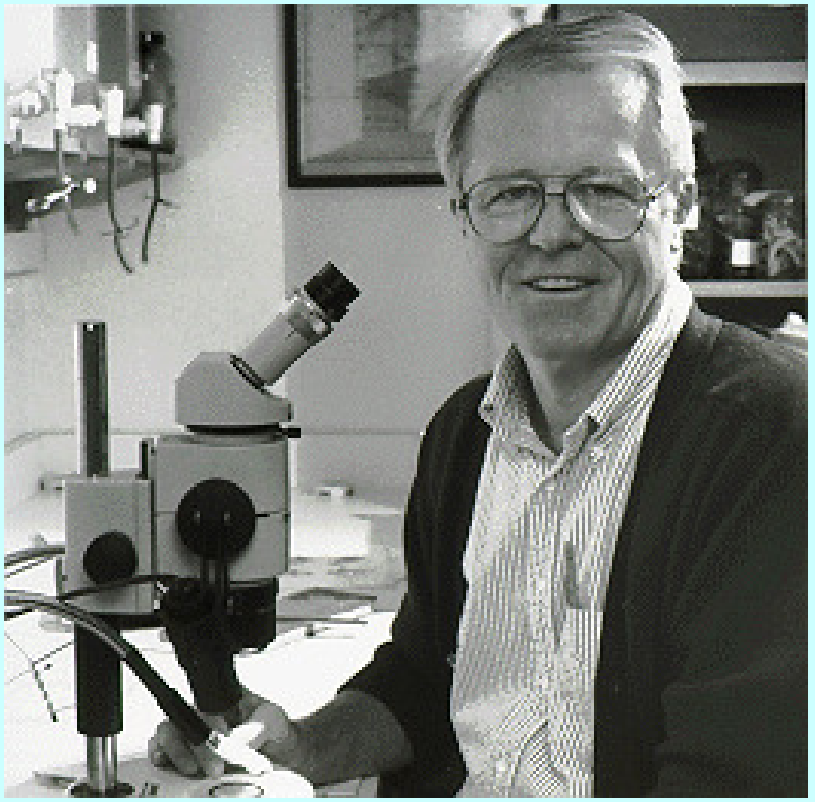





Camin noticed (in 1965) that students who did the best job recovering the true “phylogeny” of the Caminalcules made the reconstruction which required the fewest changes of state.

# J. S. Farris and Arnold Kluge in the 1980s



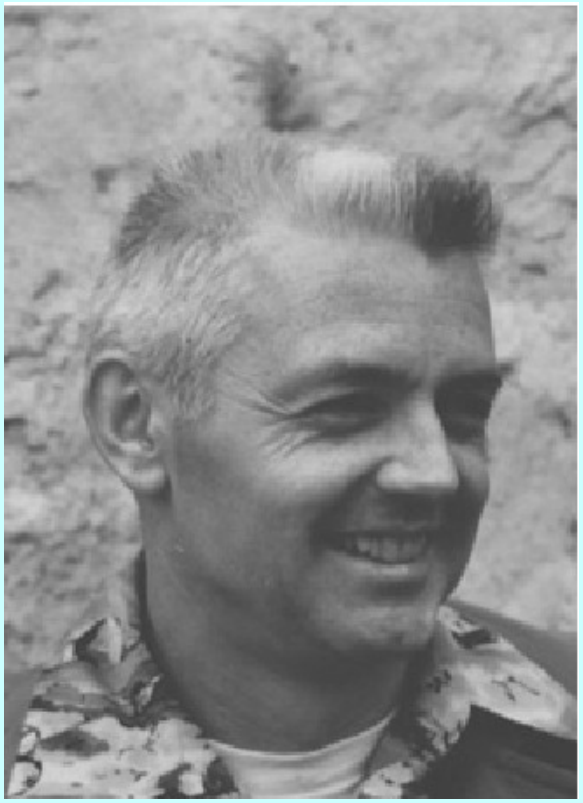


Further developments of parsimony methods, starting in 1969, advocacy of them and of Hennig's approach to classification during the 1970s and on. Central in the rise (in 1980) of the Willi Hennig Society.

# Margaret Dayhoff, 1966

- A major pioneer of molecular databases (starting in 1965)
- (With Richard Eck) made the first numerical phylogenies using molecular data
- Presented trees organized by gene families in the *Atlas of Protein Sequences* (later the PIR database) in 1966.
- Compiled the first empirical substitution rate matrices for amino acids, intended to form the basis of a probabilistic model of protein evolution.

# Walter Fitch, 1975



Walter Fitch (1929-2011) :

- The first major distance matrix method (1967)
- Developed algorithm (1971) that counts changes of state in DNA parsimony.
- Introduced the terms and concepts of orthology and paralogy.
- Co-founded the journal *Molecular Biology and Evolution* and the society SMBE.

# Fitch and Margoliash's 1967 distance tree

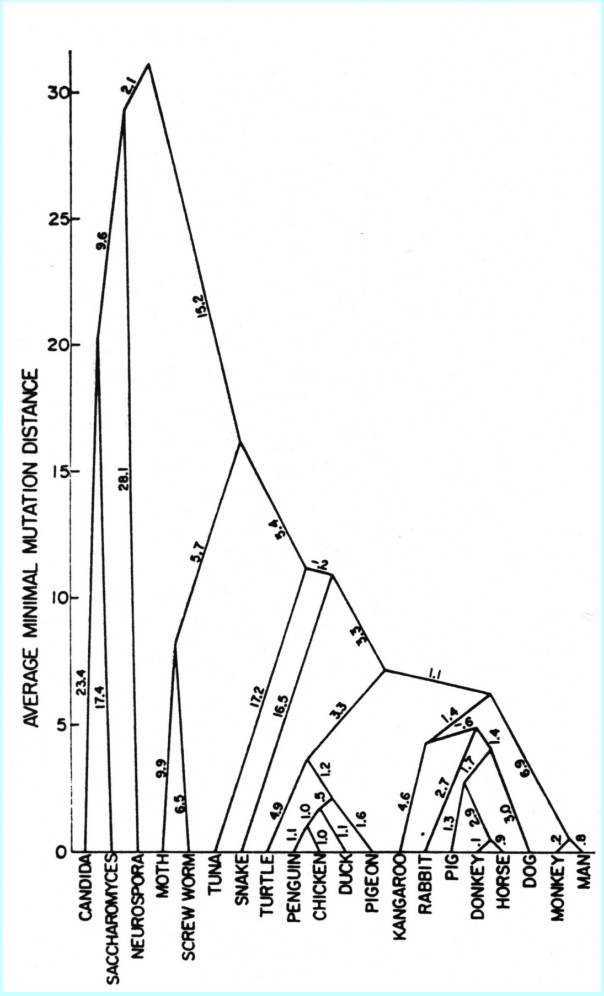
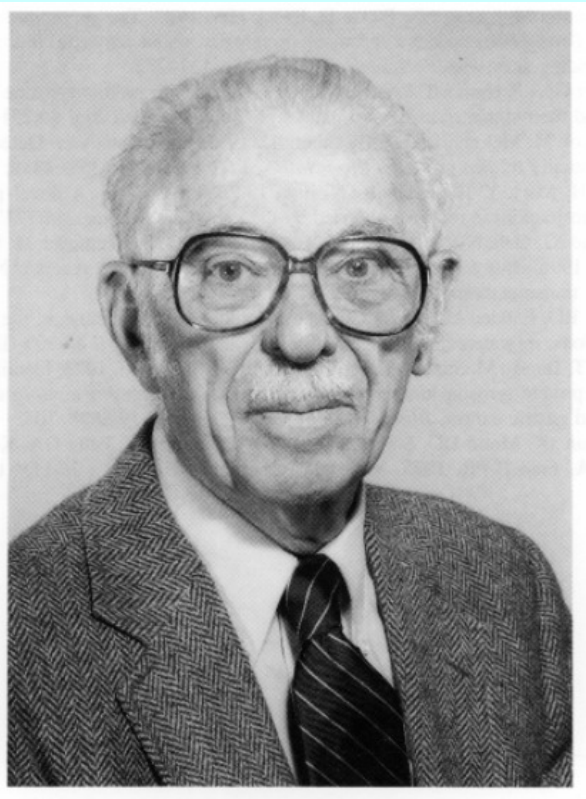
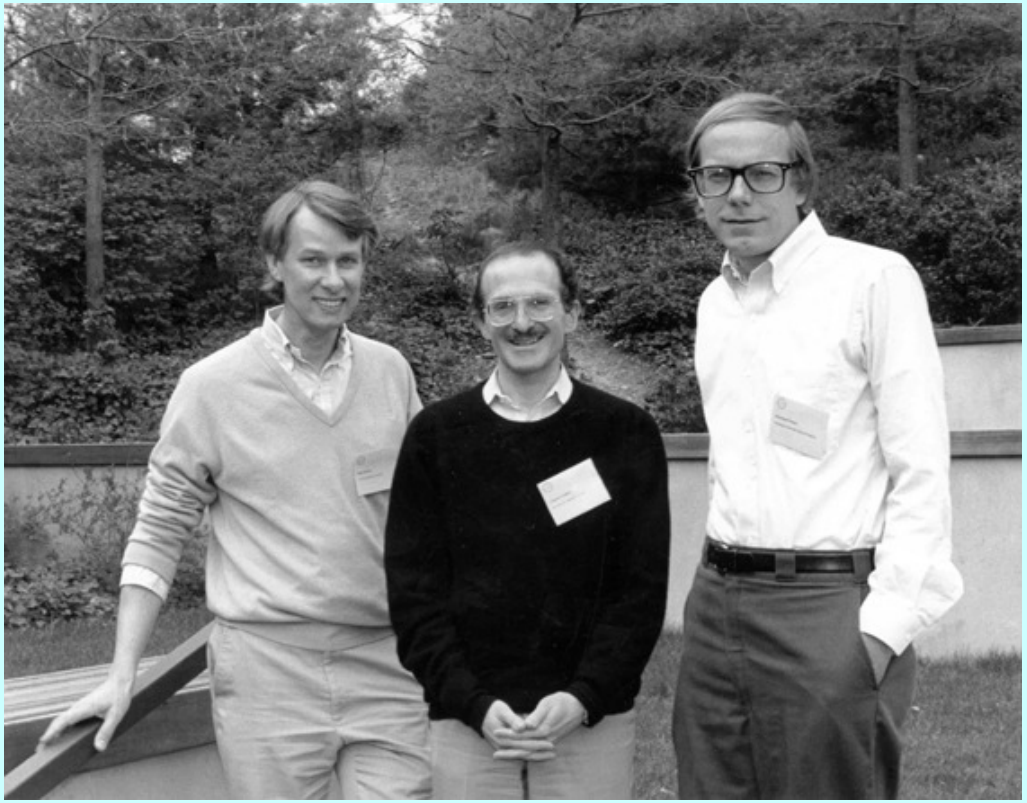



Fig. 3 (right above). A gene phylogeny as reconstructed from observable mutations in several heme-containing globins. See Fig. 2 for details. The percent "standard deviation" (7) for this tree is 1.33.

This is for globin sequences. It is the first distance matrix method published, if you don't count clustering methods.

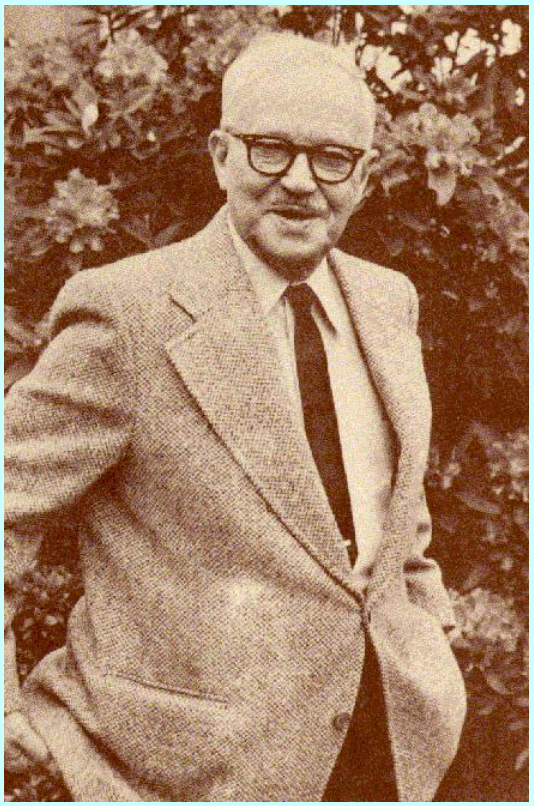
# Thomas Jukes and Charles Cantor, middle, in the 1990s





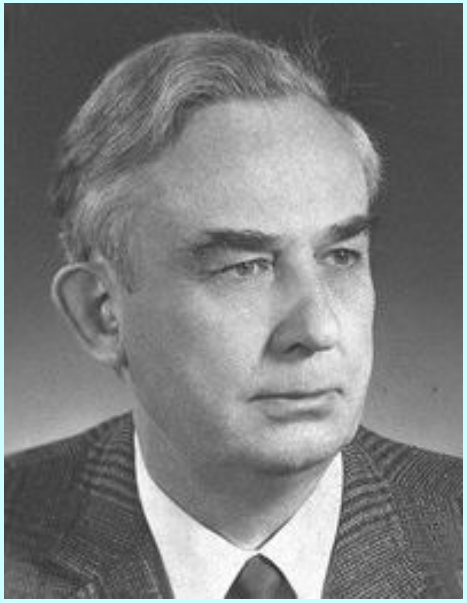
But who are the two guys standing next to Charlie Cantor? (*Hint:* one has a Nobel Prize, the other is a member of my own department). Cantor later made important technical discoveries in genomics. Jukes was a nutritional biochemist who was the primary person responsible for insisting that pregnant women get folic acid in their diet.

# Jerzy Neyman: likelihood on molecular sequences

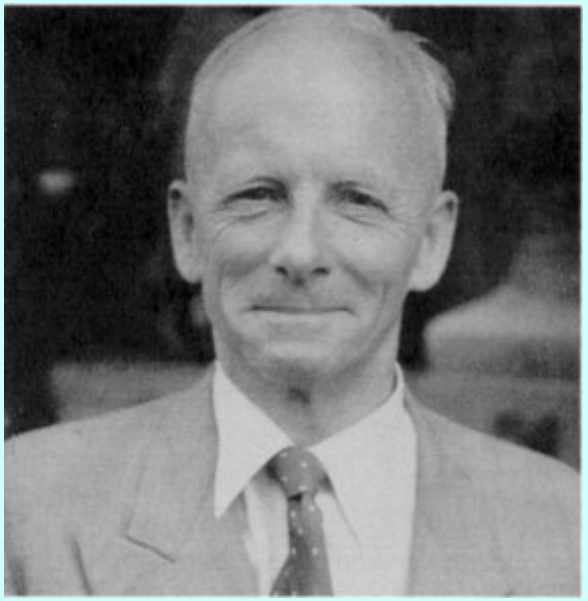


A major figure in mathematical statistics (he invented confidence intervals). Statisticians don't know that he also once worked on maximum likelihood inference of phylogenies from protein sequences (usually he was known as a pointed critic of likelihood).

# Willi Hennig (in 1972) and Walter Zimmermann (in 1959)



Willi Hennig (1913-1976)



Walter Zimmermann (1890-1980)

Zimmermann pioneered the method later advocated and made important by Hennig. Hennig was the major advocate of a purely monophyletic classification system.

## Hennig and conflict of characters

W. Hennig (1966) says that in the case of homoplasy,

> “it becomes necessary to recheck the interpretation of [the] characters"

He also says (1966, p. 121)

> “the more certainly characters interpreted as apomorphous (not characters in general) are present in a number of different species, the better founded is the assumption that these species form a morphological group."

Farris, Kluge, and Eckardt (1970) argue that this should be translated as:

> “the more characters certainly interpretable as apomorphous ..."

# Did Hennig invent parsimony?

Farris (1983, p. 8):

> I shall use the term in the sense I have already mentioned: most parsimonious genealogical hypotheses are those that minimize requirements for ad hoc hypotheses of homoplasy. If minimizing ad hoc hypotheses is not the only connotation of “parsimony" in general useage, it is scarcely novel. Both Hennig (1966) and Wiley (1975) have advanced ideas closely related to my useage. Hennig defends phylogenetic analysis on the grounds of his auxiliary principle, which states that homology should be presumed in the absence of evidence to the contrary. This amounts to the precept that homoplasy ought not be postulated beyond necessity, that is to say parsimony.

## Hennig's auxiliary principle

Hennig discusses the case in which "only one character can certainly or with reasonable probability be interpreted as apomorphous." Hennig (1966, p. 121):

> In such cases it is impossible to decide whether the common character is indeed synapomorphous or is to be interpreted as parallelism, homoiology, or even as convergence. I have therefore called it an "auxiliary principle" that the presence of apomorphous characters in different species "is always reason for suspecting kinship [i.e. that the species belong to a monophyletic group], and that their origin by convergence should not be assumed a priori" (Hennig 1953). This was based on the conviction that "phylogenetic systematics would lose all the ground on which it stands" if the presence of apomorphous characters in different species were considered first of all as convergences (or parallelisms), with proof to the contrary required in each case.

This is usually considered to be a statement of parsimony. Is it?

# Farris and Kluge on Hennig and parsimony

> Unfortunately, AIV is not sufficiently detailed to allow us to select a unique criterion for choosing a most preferable tree. We know that trees on which the monophyletic groups share many steps are preferable to trees on which this is not so. But AIV deals only with single monophyletic groups and does not tell us how to evaluate a tree consisting of several monophyletic groups. One widely used criterion – parsimony – could be used to select trees. This would be in accord with AIV, since on a most parsimonious tree OTUs [tips] that share many states (this is *not* the same as the OTUs' being *described* by many of the same *states*) are generally placed together. We might argue that the parsimony criterion selects a tree most in accord with AIV by "averaging" in some sense the preferability of all the monophyletic groups of the tree. Other criteria, however, may also agree with AIV.
>
> Farris, Eckardt and Kluge, 1970

# Philosophical frameworks: hypothetico-deductive

Gaffney (1979, pp. 98-99)

> In any case, in a hypothetico-deductive system, parsimony is not merely a methodological convention, it is a direct corollary of the falsification criterion for hypotheses (Popper, 1968a, pp. 144-145). When we accept the hypothetico-deductive system as a basis for phylogeny reconstruction, we try to test a series of phylogenetic hypotheses in the manner indicated above. If all three of the three possible three-taxon statements are falsified at least once, the least-rejected hypothesis remains as the preferred one, not because of an arbitrary methodological rule, but because it best meets our criterion of testability. In order to accept an hypothesis that has been successfully falsified one or more times, we must adopt an *ad hoc* hypothesis for each falsification .... Therefore, in a system that seeks to maximize vulnerability to criticism, the addition of *ad hoc* hypotheses must be kept to a minimum to meet this criterion.

## more from Gaffney

Gaffney (1979)

> “the use of derived character distributions as articulated by Hennig (1966) appears to fit the hypothetico-deductive model best."

Gaffney (1979, p. 98)

> “it seems to me that parsimony, or Ockham’s razor, is equivalent to ‘logic’ or ‘reason’ because any method that does not follow the above principle would be incompatible with any kind of predictive or causal system."

## Hypothetico-deductivists on falsification

Eldredge and Cracraft (1980, p. 69) are careful to point out that

> “Falsified" implies that the hypotheses are proven false, but this is not the meaning we (or other phylogenetic systematists) wish to convey. It may be that the preferred hypothesis will itself be “rejected" by some synapomorphies.

Wiley (1981, p. 111):

> In other words, we have no external criterion to say that a particular conflicting character is actually an invalid test. Therefore, saying that it is an invalid test simply because it is unparsimonious is a statement that is, itself, an ad hoc statement. With no external criterion, we are forced to use parsimony to minimize the total number of ad hoc hypotheses (Popper, 1968a: 145). The result is that the most parsimonious of the various alternates is the most highly corroborated and therefore preferred over the less parsimonious alternates.

## Farris on hypothetico-deductivism

Farris (1983, p. 8):

> Wiley [(1975)] discusses parsimony in a Popperian context, characterizing most parsimonious genealogies as those that are least falsified on available evidence. In his treatment, contradictory character distributions provide putative falsifiers of genealogies. As I shall discuss below, any such falsifier engenders a requirement for an ad hoc hypothesis of homoplasy to defend the genealogy. Wiley's concept is then equivalent to mine.

# Philosophical frameworks: Logical-parsimony

Beatty and Fink (1979):

> We can account for the necessity of parsimony (or some such consideration) because evidence considerations alone are not sufficient. But we have no philosophical or logical argument with which to justify the use of parsimony considerations – a not surprising result, since this issue has remained a philosophical dilemma for hundreds of years.

This moves away from considering parsimony as justified by hypothetico-deductive frameworks and invokes it as its own justification.

# Kluge and Wolf on logical parsimony

Kluge and Wolf (1993, p. 196):

> Finally, we might imagine that some of the popularity of the aforementioned methodological strategies and resampling techniques, and assumption of independence in the context of taxonomic congruence and the cardinal rule of Brooks and McLennan (1991), derives from the belief that phylogenetic inference is hypothetico-deductive (e.g. Nelson and Platnick, 1984: 143-144), or at least that it should be. Even the uses to which some might put cladograms, such as “testing" adaptation (Coddington, 1988), are presented as hypothetico-deductive. But this ignores an alternative, that cladistics, and its uses, may be an abductive enterprise (Sober, 1988). We suggest that the limits of phylogenetic systematics will be clarified considerably when cladists understand how their knowledge claims are made (Rieppel, 1988; Panchen, 1992).

Again, moving away from hypothetico-deductivism.

## Elliot Sober on falsification

Sober (1988, p. 126):

> Popper's philosophy of science is very little help here, because he has little to say about *weak* falsification. Popper, after all is a hypothetico-*deductivist*. For him, observational claims are deductive consequences of the hypothesis under test .... Deductivism excludes the possibility of probabilistic testing. A theory that assigns probabilities to various possible observational outcomes cannot be strongly falsified by the occurrence of any of them. This, I suggest, is the situation we confront in testing phylogenetic hypotheses. (AB)C is logically consistent with all possible character distributions (polarized or not), and the same is true of A(BC). [Emphasis in the original]

Sober sees Popper's hypothetico-deductive framework as ill-suited to inference of phylogenies.

# Philosophical foundations: Logical probability?

Popper's corroboration formula

$$C(h, e, b) = \frac{p(e, hb) - p(e, b)}{p(e, hb) - p(eh, b) + p(e, b)}$$

where

| | | |
|---|---|---|
| b | = | background knowledge |
| h | = | hypothesis |
| e | = | evidence (= d, data?) |

$$C(h, e, b) = \frac{\mathrm{Prob}\,(d|h) - \mathrm{Prob}\,(d)}{\mathrm{Prob}\,(d|h) - \mathrm{Prob}\,(d\&h) + \mathrm{Prob}\,(d)}$$

This is used as a justification for parsimony, not actually as a statistical method in spite of people like Kluge calling it "logical probability". Issue: how to compute $\mathrm{Prob}(d)$ and $\mathrm{Prob}(d\&h)$ if you aren't a Bayesian and can't weight by a prior on hypotheses?

## Criticisms of statistical inference

Farris (1983, p.17):

> 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.

Kluge (1997a)

> “As an aside, the fact that the study of phylogeny is concerned with the discovery of historical singularities means that calculus probability and standard (Neyman-Pearson) statistics *cannot* apply to that historical science ...."

If, after tossing a coin multiple times, you lose the coin, does he think you can’t then analyze those data?

## Positions on classification nowadays

- Phylogenetic systematics. Willi Hennig advocated purely monophyletic classification. Now the (strongly) dominant approach.
- Evolutionary systematics. Has almost faded away. Its adherents were reluctant to make it algorithmic.
- Phenetics. Although Sokal and Sneath strongly influenced the field of numerical clustering, their approach to biological classification has few adherents.
- IDMVM One person (me) takes the view that It Doesn't Matter Very Much, as we use the phylogeny, and, given that, we never use the classification system. This is widely regarded as a marginal crackpot view ["A bizarre thumb in the eye to systematists" – Michael Sanderson].