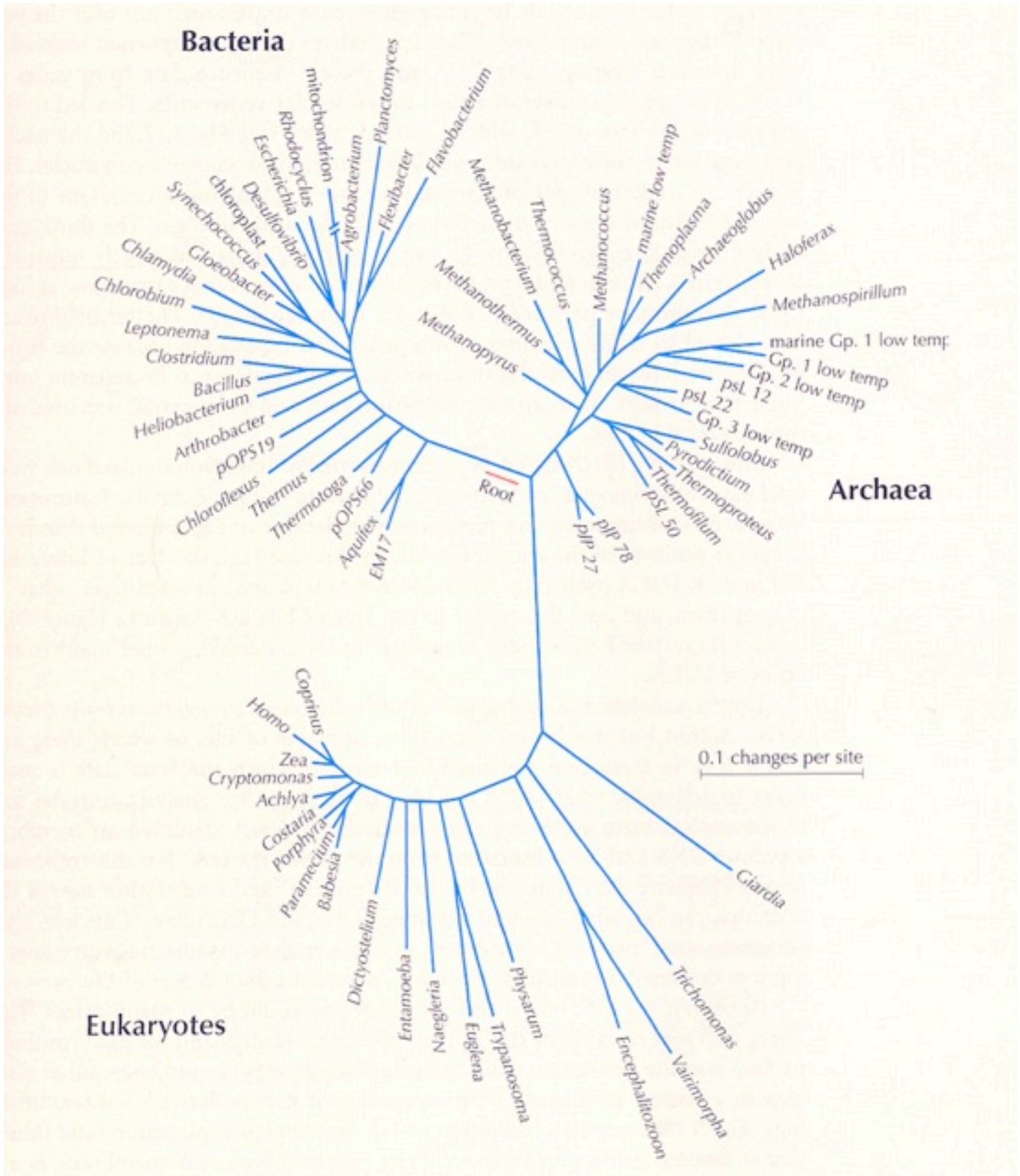
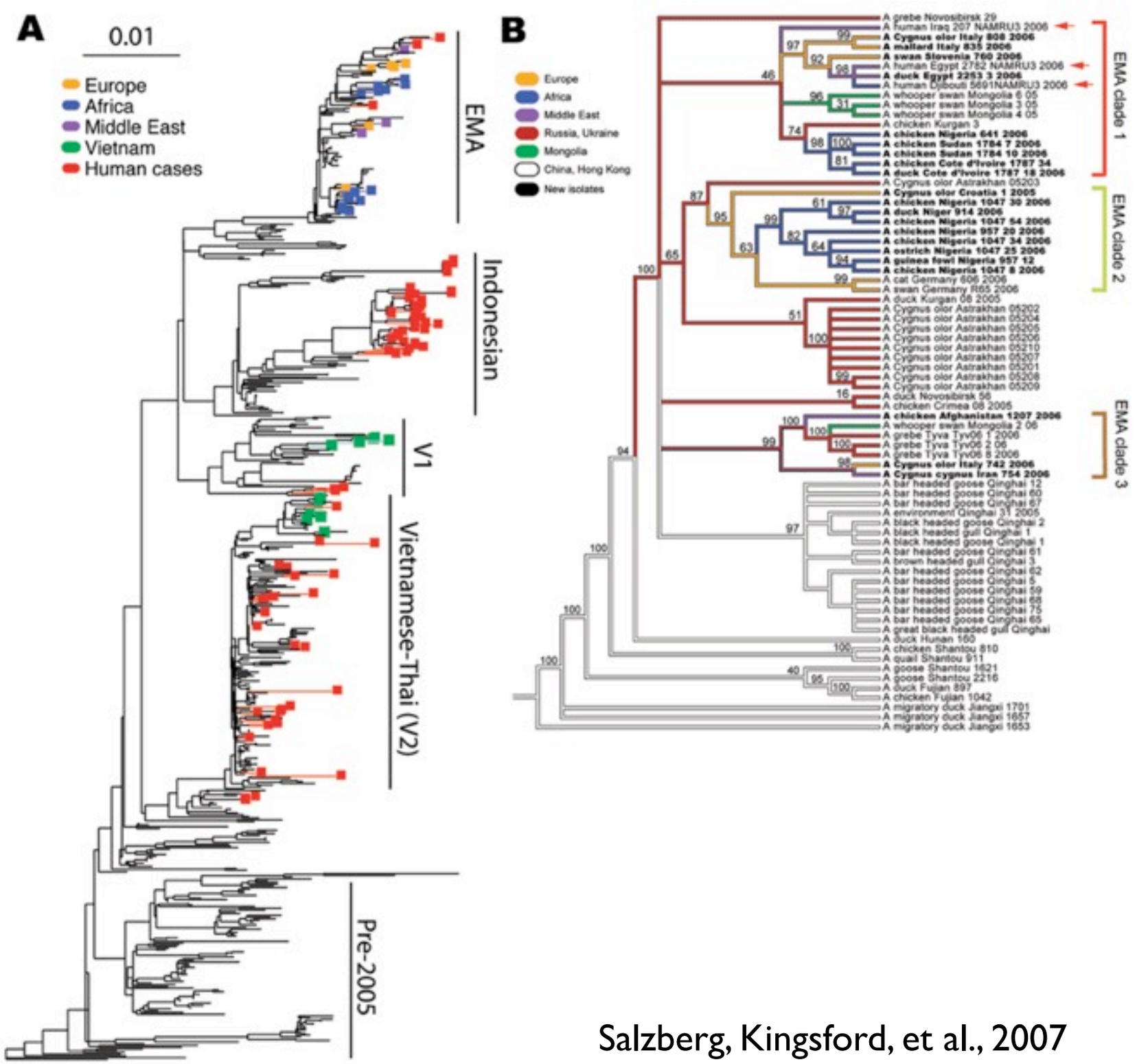
Phylogeny

CMSC 858s

Tree of Life / LUA

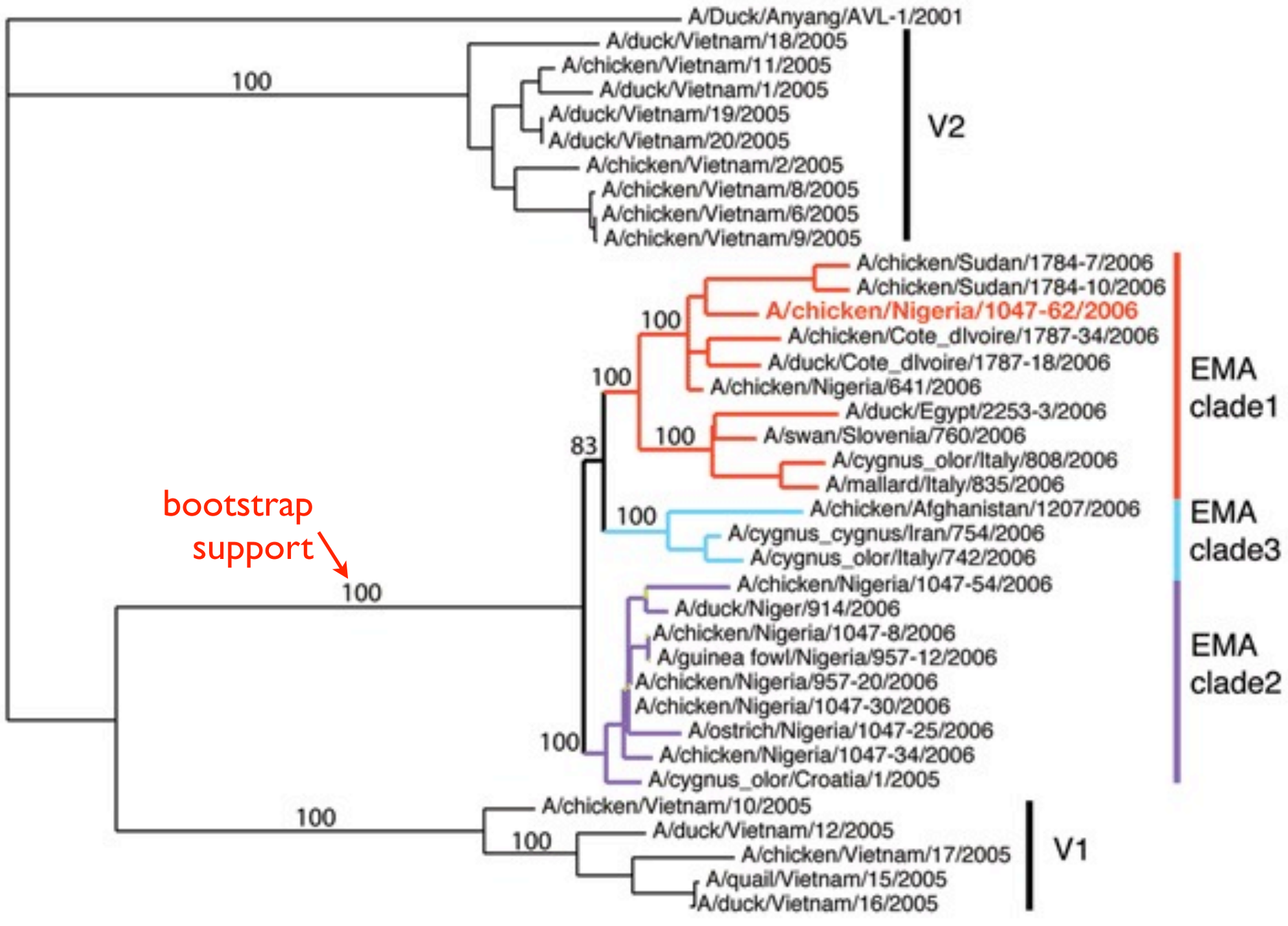


H5N1 Influenza Strains



Salzberg, Kingsford, et al., 2007

H5N1 Influenza Strains



Questions Addressable by Phylogeny

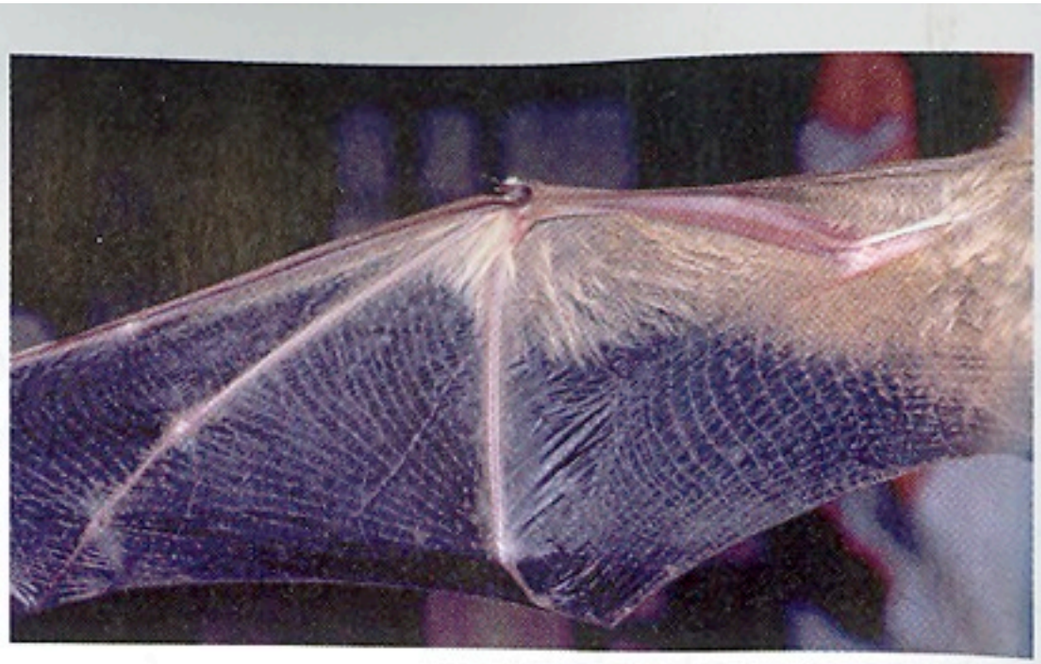
- How many times has a feature arisen? been lost?
- How is a disease evolving to avoid immune system?
- What is the sequence of ancestral proteins?
- What are the most similar species?
- What is the rate of speciation?
- Is there a correlation between gain/loss of traits and environment? with geographical events?
- Which features are ancestral to a clade, which are derived?
- What structures are homologous, which are analogous?

Study Design Considerations

- **Taxon sampling:**
  - how many individuals are used to represent a species?
  - how is the "outgroup" chosen?
  - Can individuals be collected or cultured?

- **Marker selection: Sequence features should:**
  - be Representative of evolutionary history (unrecombined) [mitochondrial & chloroplast dna]
  - have a single copy
  - be able to be amplified with PCR
  - able to be sequenced
  - change enough to distinguish species, similar enough to perform MSA

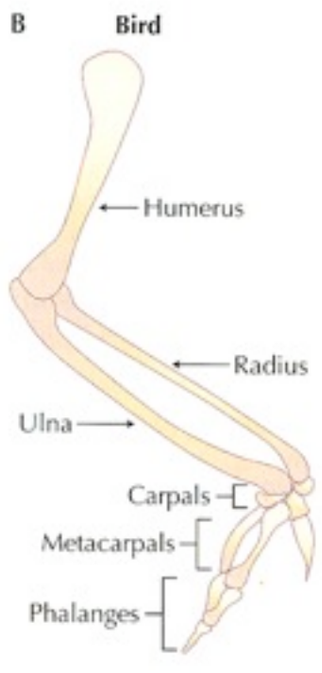
Convergent Evolution

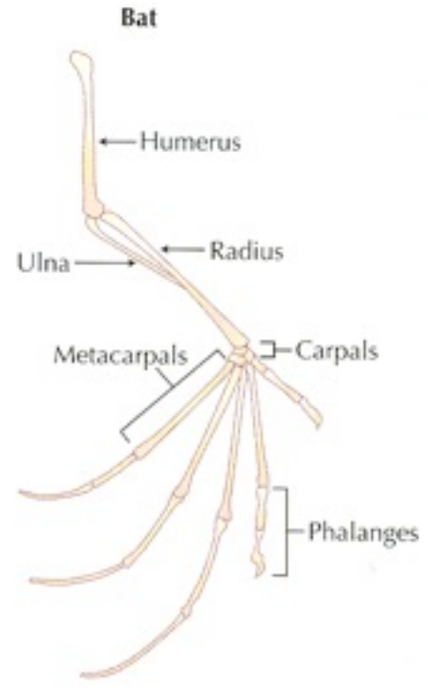


Bird & bat wings arose independently (analogous)

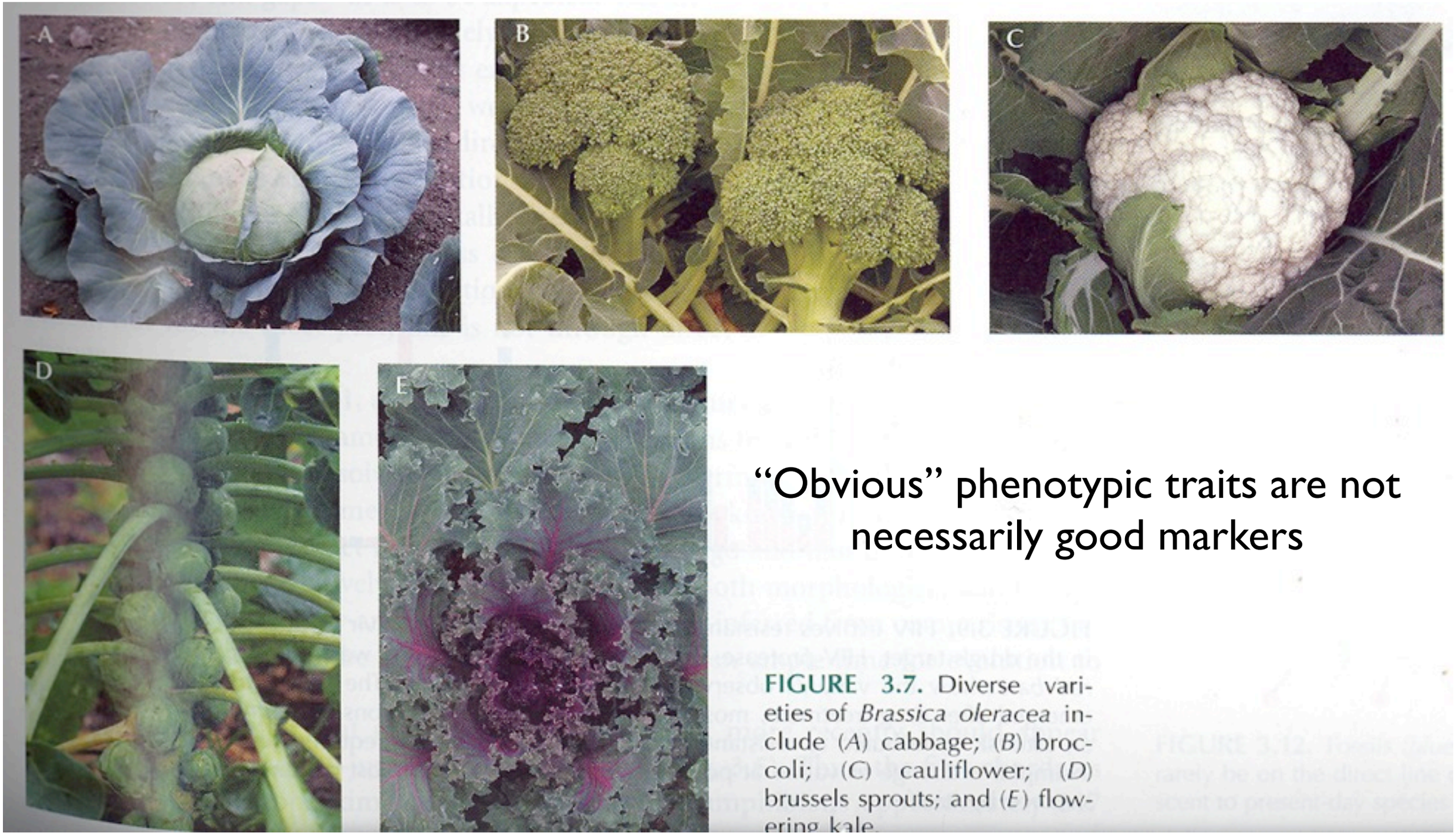
"Has wings" is thus a "bad" trait for phylogenetic inference

Bone structure has common ancestor (homologous)





"Divergent" Evolution



"Obvious" phenotypic traits are not necessarily good markers

FIGURE 3.7. Diverse varieties of *Brassica oleracea* include (*A*) cabbage; (*B*) broccoli; (*C*) cauliflower; (*D*) brussels sprouts; and (*E*) flowering kale.

Searching for Trees

- Distance-based methods:
  - Sequences -> Distance Matrix -> Tree
  - Neighbor joining, UPGMA

- Maximum Likelihood:
  - Sequences + Model -> Tree + parameters

- Bayesian MCMC:
  - Markov Chain Monte Carlo: random sampling of trees by random walk

Additivity

- A distance matrix is additive if a tree can be constructed such that $d_T(i,j)$ = path length from i to j = $M(i,j)$.

- Such a tree faithfully represents all the distances.

- 4-point condition: A metric space is additive if given any 4 points, we can label them so that

$$M_{ij} + M_{kl} = M_{ik} + M_{jl} \geq M_{il} + M_{jk}$$

UPGMA

- Find two most similar taxa (ie. such that M_{ij} is smallest)
- Merge into new "OTU" (operational taxonomic unit)
  - distance from k to to new OTU = average distance from k to each of OTUs members
- Repeat.
- Even if there is perfect tree, it may not find it.

Neighbor Joining

Choose x, y to merge that minimize:

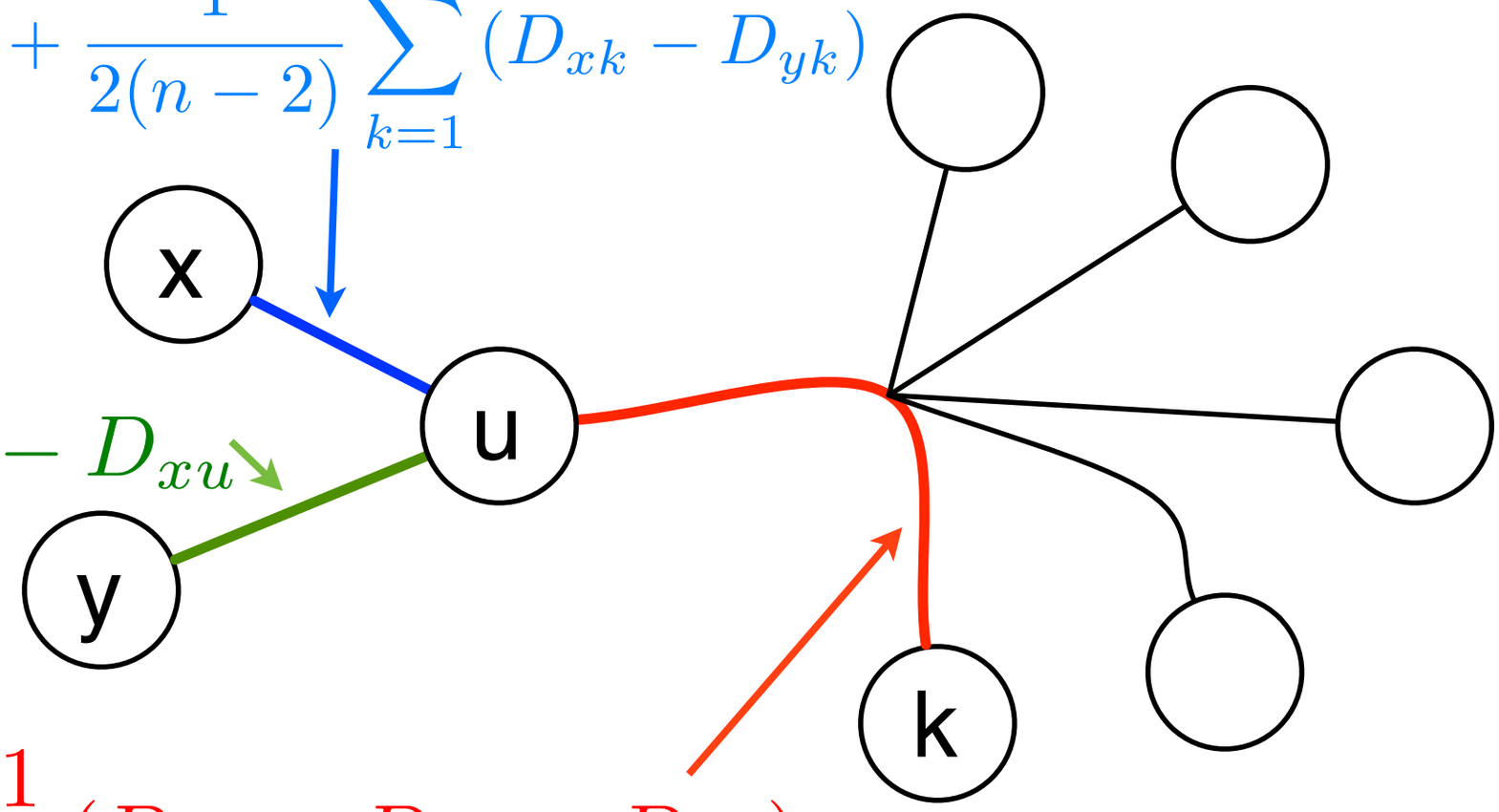
$$Q(x,y) := (n-2)D_{xy} - \left(\sum_{k=1}^{n} D_{xk} + \sum_{k=1}^{n} D_{yk}\right)$$

Update lengths:

$$D_{xu} := \frac{1}{2}D_{xy} + \frac{1}{2(n-2)}\sum_{k=1}^{n}(D_{xk} - D_{yk})$$

$$D_{yu} := D_{xy} - D_{xu}$$

$$D_{uk} := \frac{1}{2}\left(D_{xk} + D_{yk} - D_{xy}\right)$$



Maximum Parsimony

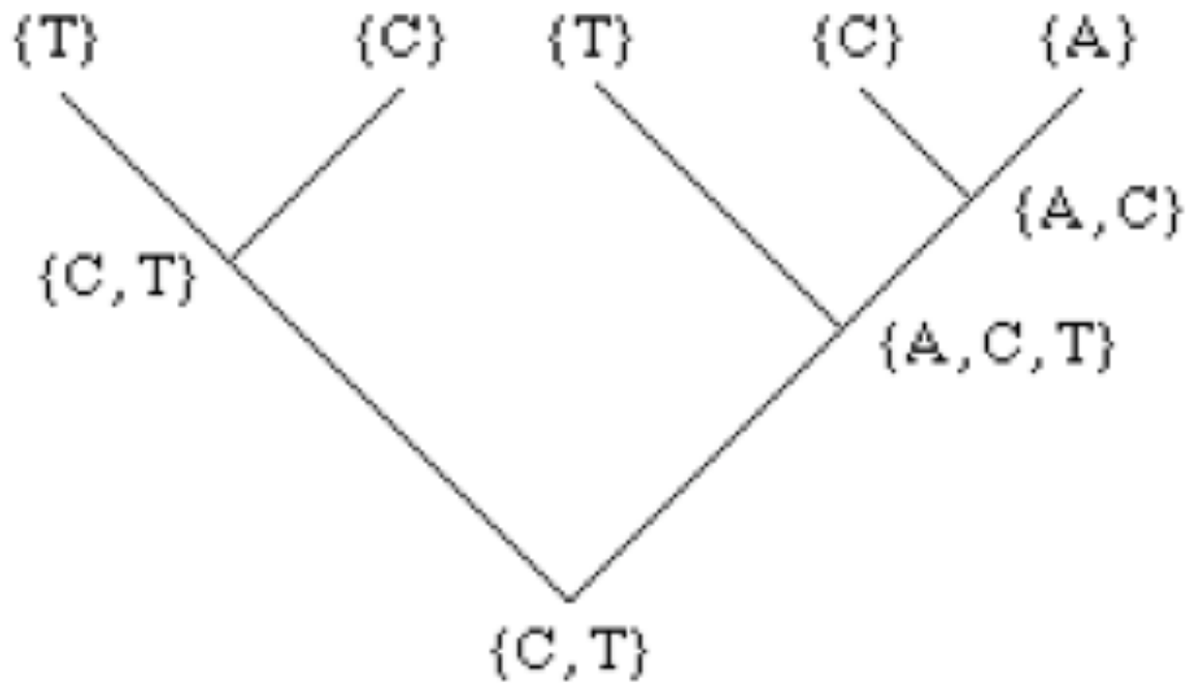
- **Input:** n sequences of length k
- **Output:** A tree $T = (V, E)$ and a sequence s_u of length k for each node u to minimize:

$$\sum_{(u,v)\in E} \text{Hamming}(s_u, s_v)$$

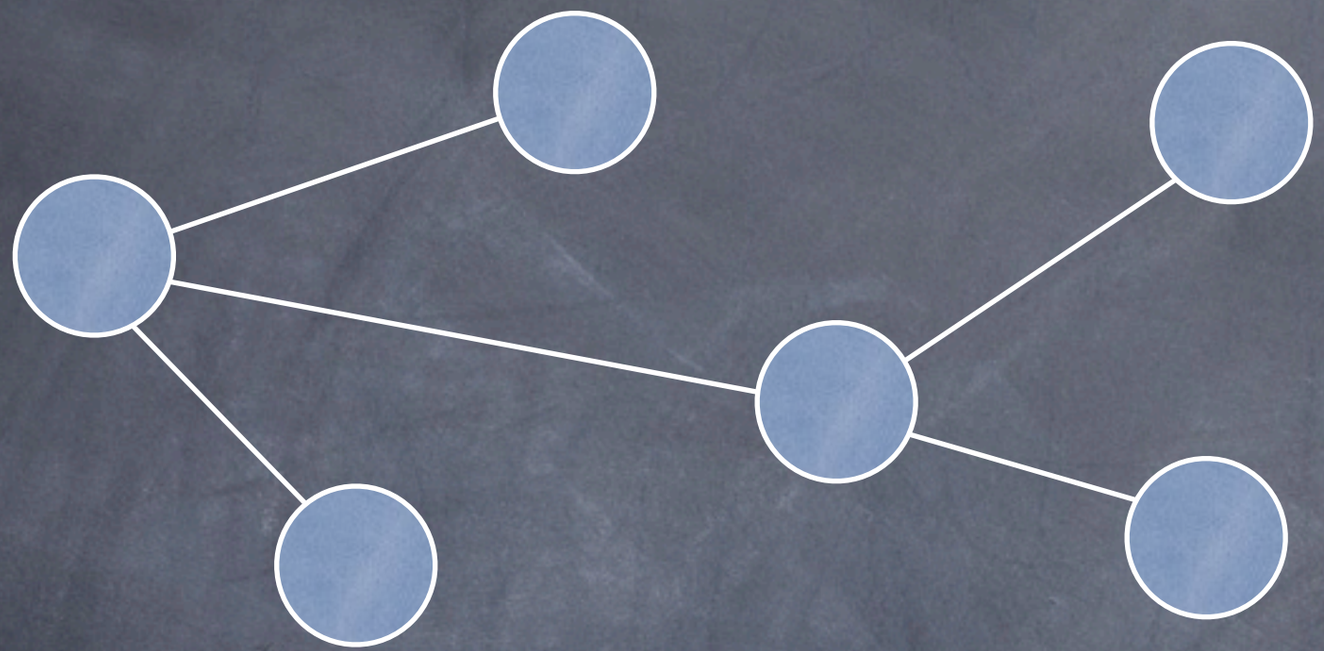
NP-hard (reduction from Hamming distance Steiner tree)
Can score a given tree in time $O(|\Sigma|nk)$.

Fitch's Algorithm

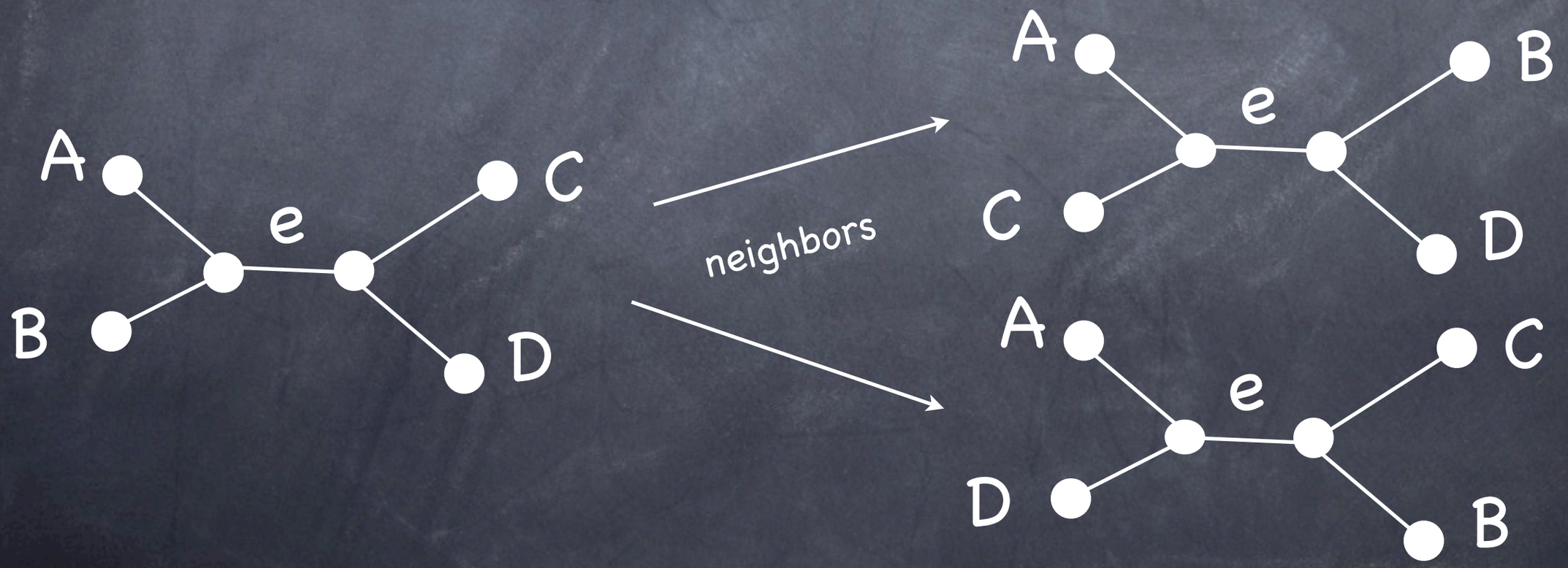
- Work from children to the root
  - $S_k = S_i \cup S_j$ if $S_i \cap S_j$ empty
  - $S_k = S_i \cap S_j$ otherwise
- Parsimony score = # of unions



Heuristic: Nearest Neighbor Interchange



Walk from tree T to its neighbors, choosing best neighbor at each step.



Maximum Likelihood

- **Input:** n sequences $S_1,\ldots,S_n$ of length k; choice of model
- **Output:** Tree T and parameters p_e for each edge to maximize:

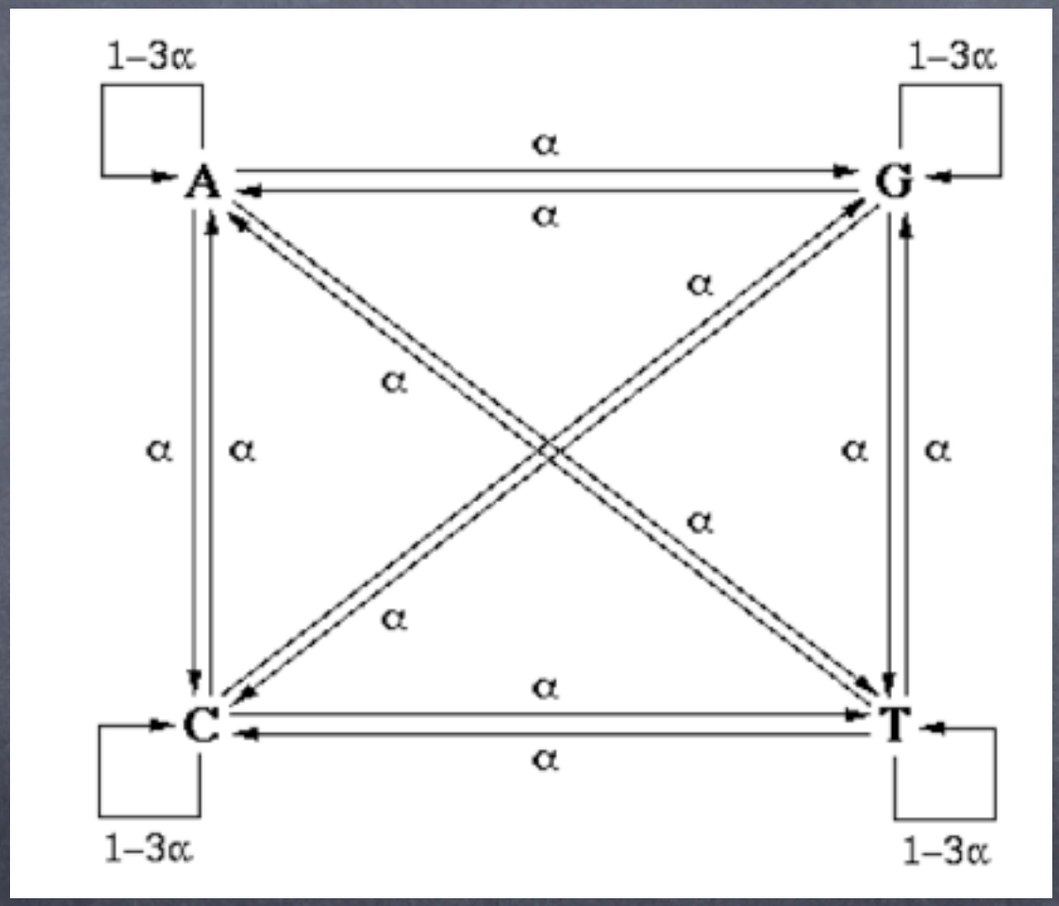
$\Pr[S_1,\ldots,S_n \mid T, p]$

NP-hard if model is Jukes-Cantor; probably NP-hard for other models.

Score of a fixed tree = $\sup_p\{\Pr[S1,\ldots,Sn \mid T, p]\}$; NP-hard to even compute this (even for trees with 4 leaves).

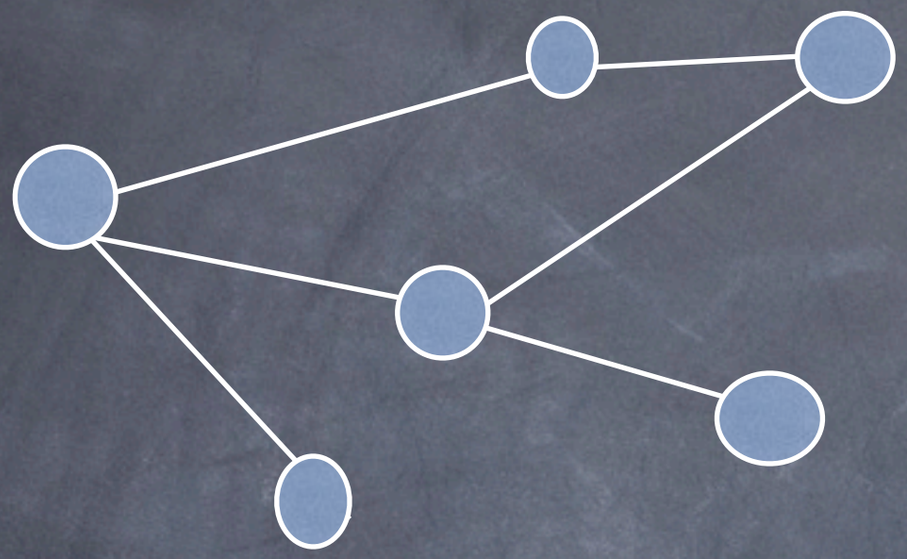
Jukes-Cantor Model of Sequence Evolution

- Simplest Model: if you mutate, you switch to each base with equal probability:



- GTR: General Time Reversible: Symmetric 4x4 matrix + edge lengths

Bayesian MCMC



Walk from tree T to its neighbors, choosing a particular neighbor at each step with probability related to its improved likelihood.

- # of times you visit a tree (after "burn in")= probability of that topology
- Outputs a distribution of trees, not a single tree.

Bootstrapping

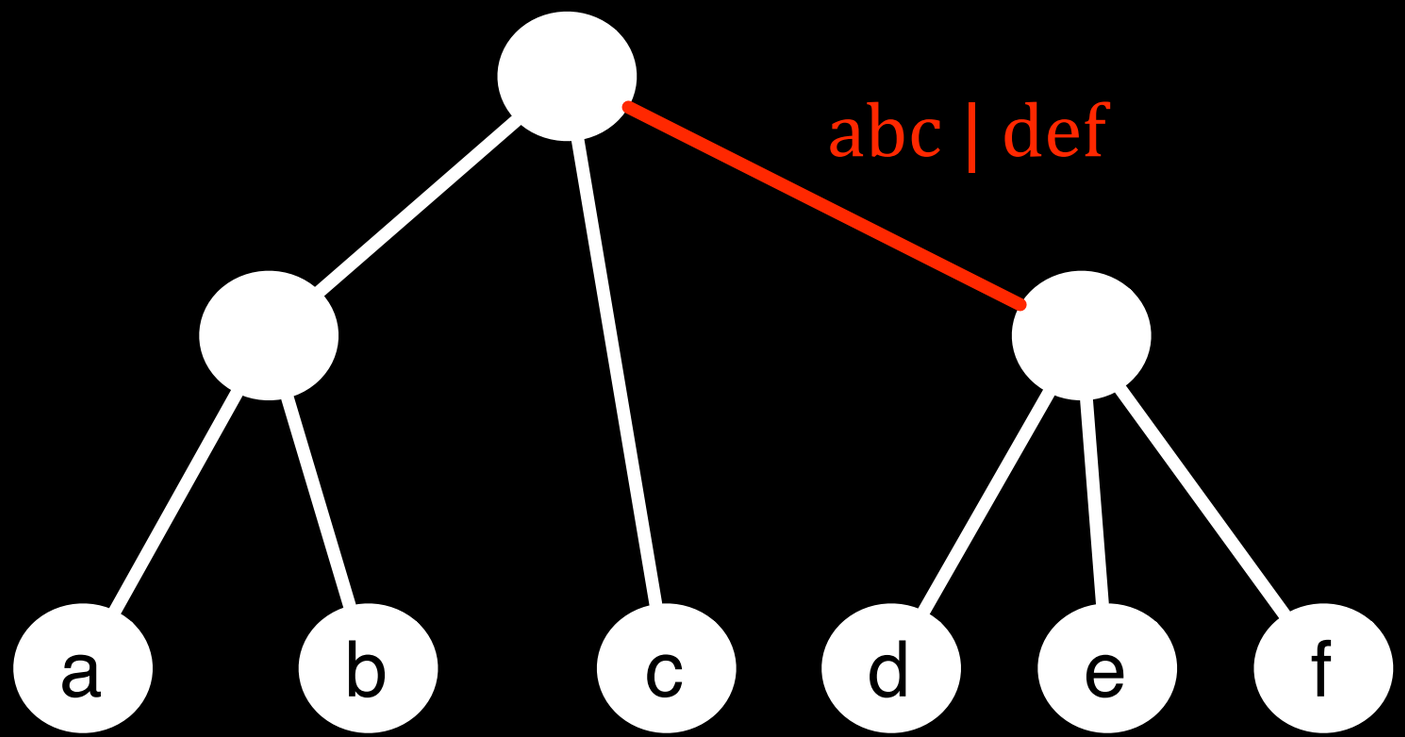
- How confident are we in a given edge?

- Bootstrapping:

1. Create (e.g.) 1,000 data sets of same size as input by sampling markers (MSA columns) with replacement.
2. Repeat phylogenetic inference on each set.
3. Support for edge is the % of trees containing this edge (bipartition).

- **Interpretation**: probability that edge would be inferred on a random data set drawn from the same distribution as the input set.

Splits



Every edge ⇒ a split, a bipartition of the taxa

- taxa within a clade leading from the edge
- taxa outside the clade leading from the edge

Example: this tree = {abc|def, ab|cdef + ‘trivial’ splits}

Consensus

- Multiple trees: from bootstrap, from Bayesian MCMC, trees with sufficient likelihood, same parsimony:

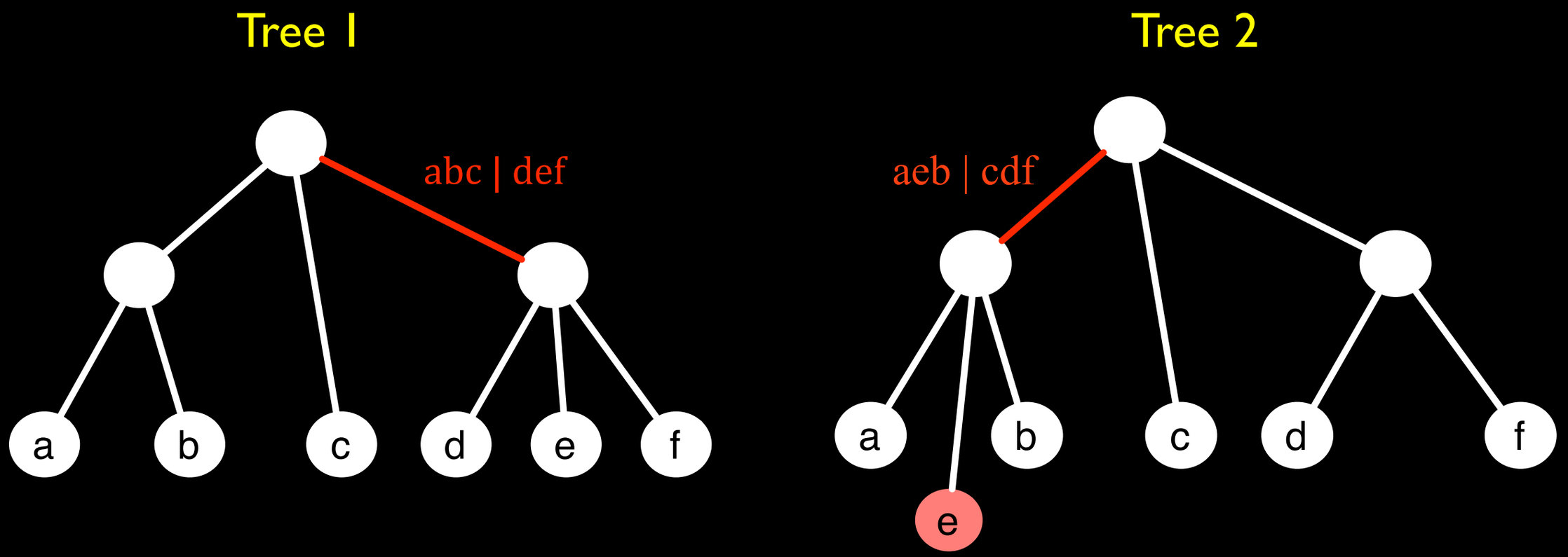
$$T = \{T_1, \ldots, T_n\}$$

- Splits of $T_i := C(T_i) = \{ b(e) : e \in T_i \}$

  $b(e)$ is the split (bipartition) for edge e.

- **Majority consensus:** tree given by splits which occur in > half inferred trees.

Incompatibility



Two splits are incompatible if they cannot be in the same tree.

Majority Consensus Always Exists

- **Proof**:

1. Let $\{s_i\}$ be the splits in > half the trees.

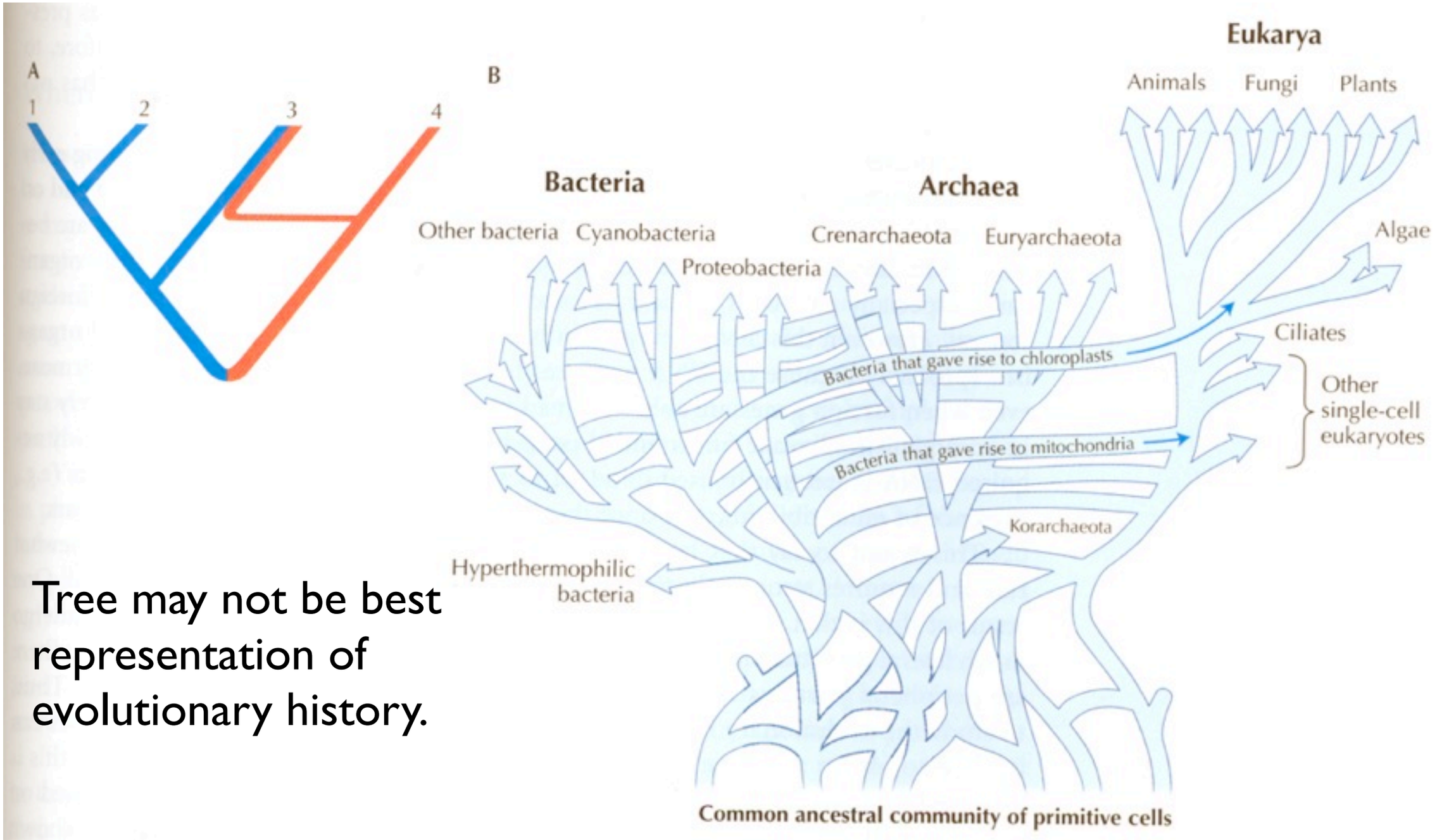
2. Pigeonhole: for each i, j there must be a tree containing both s_i and s_j.

3. If s_i and s_j are in same tree they are compatible.

4. Any set of compatible splits forms a tree.

$\Rightarrow$ The $\{s_i\}$ are pairwise compatible and form a tree.

Horizontal Gene Transfer



Tree may not be best representation of evolutionary history.

DNA uptake; retroviruses