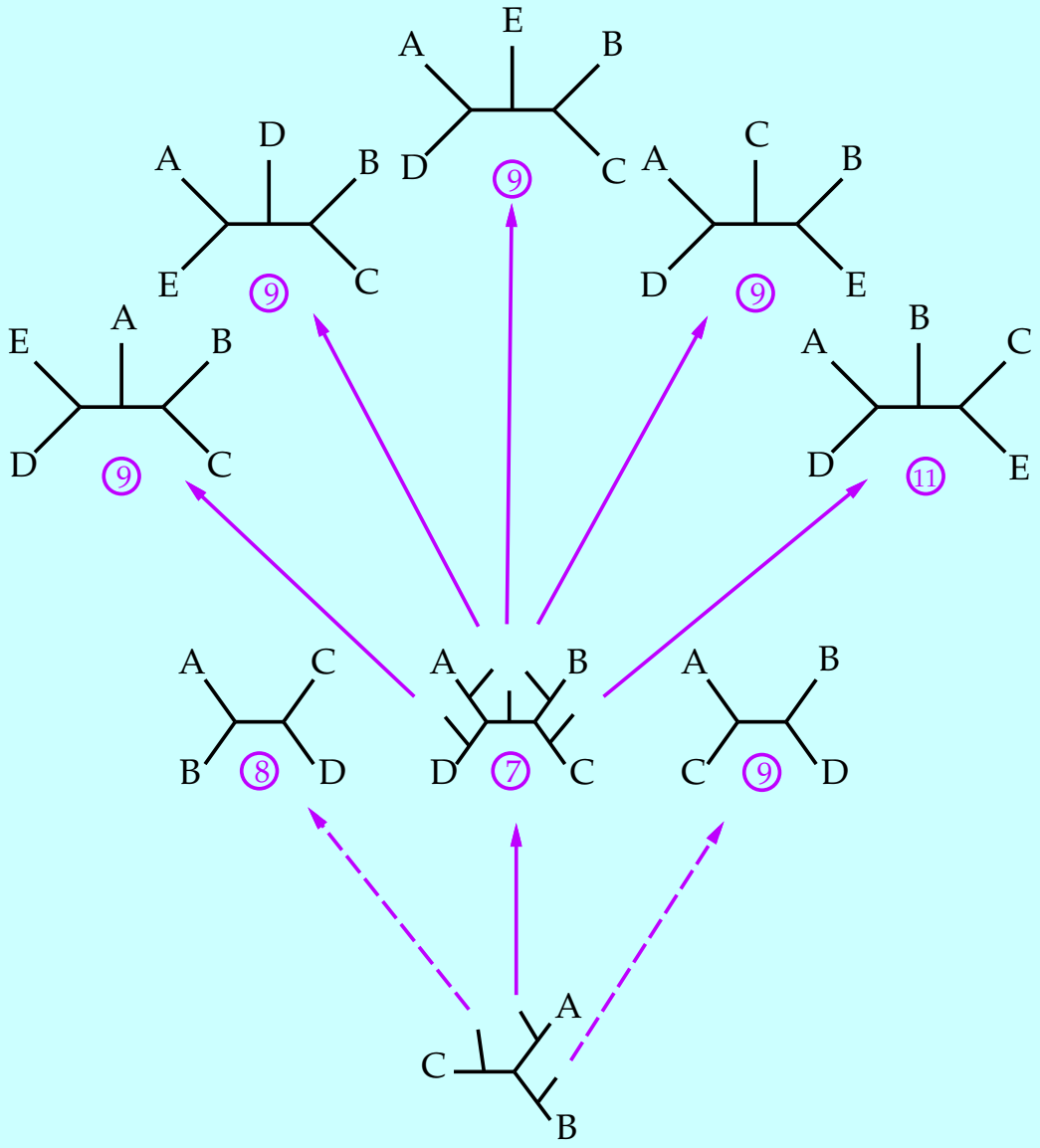
Lecture 3. Phylogeny methods: Branch and bound, distance methods

Joe Felsenstein

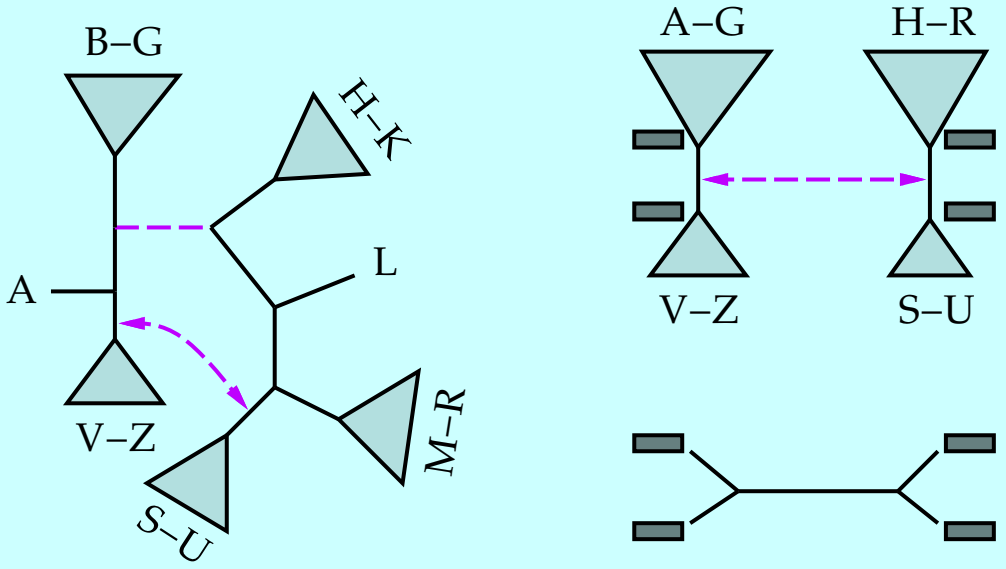
Department of Genome Sciences and Department of Biology

Greedy search by sequential addition



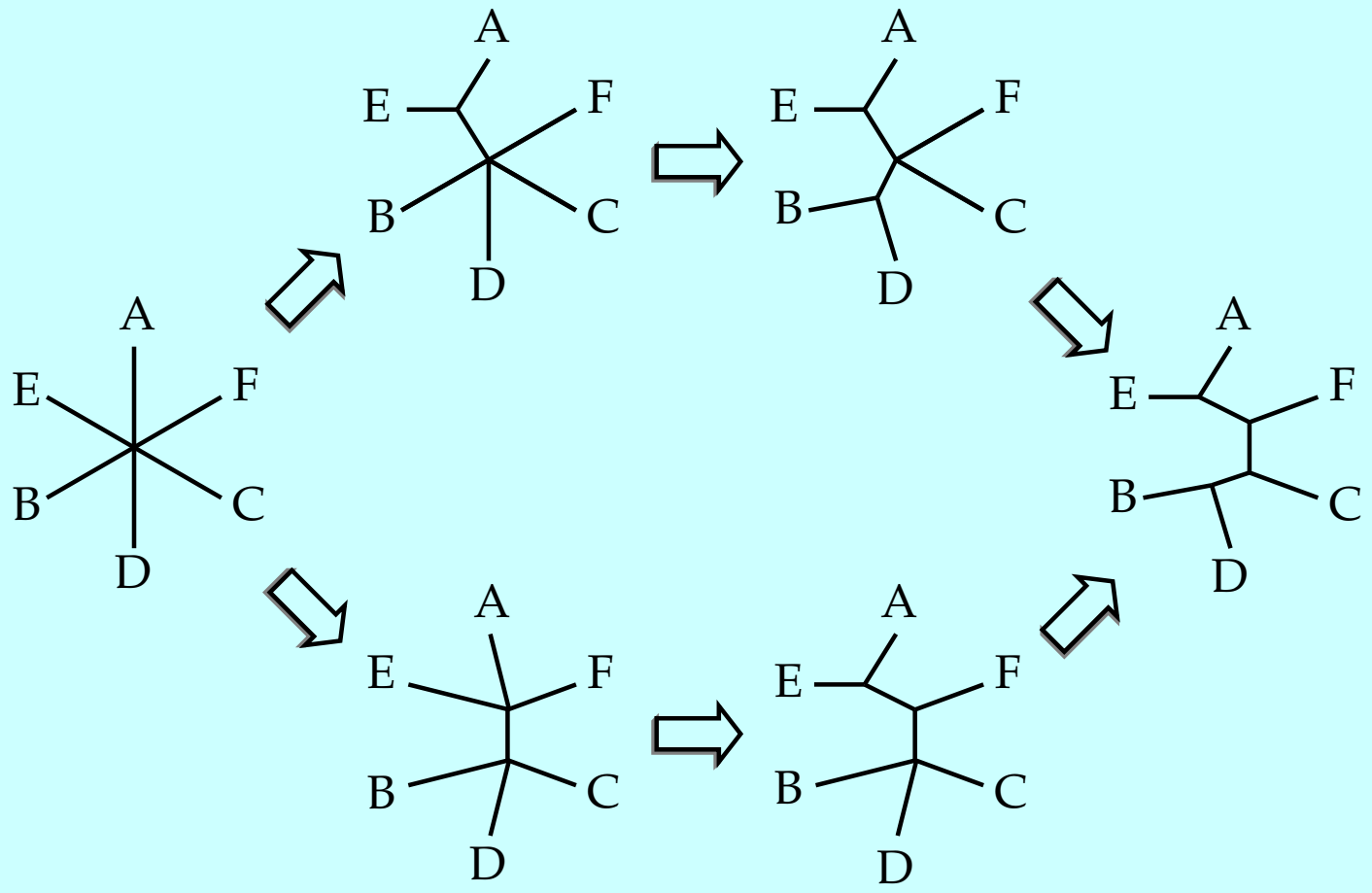
Greedy search by addition of species in a fixed order (A, B, C, D, E) in the best place each time.

Goloboff's time-saving trick



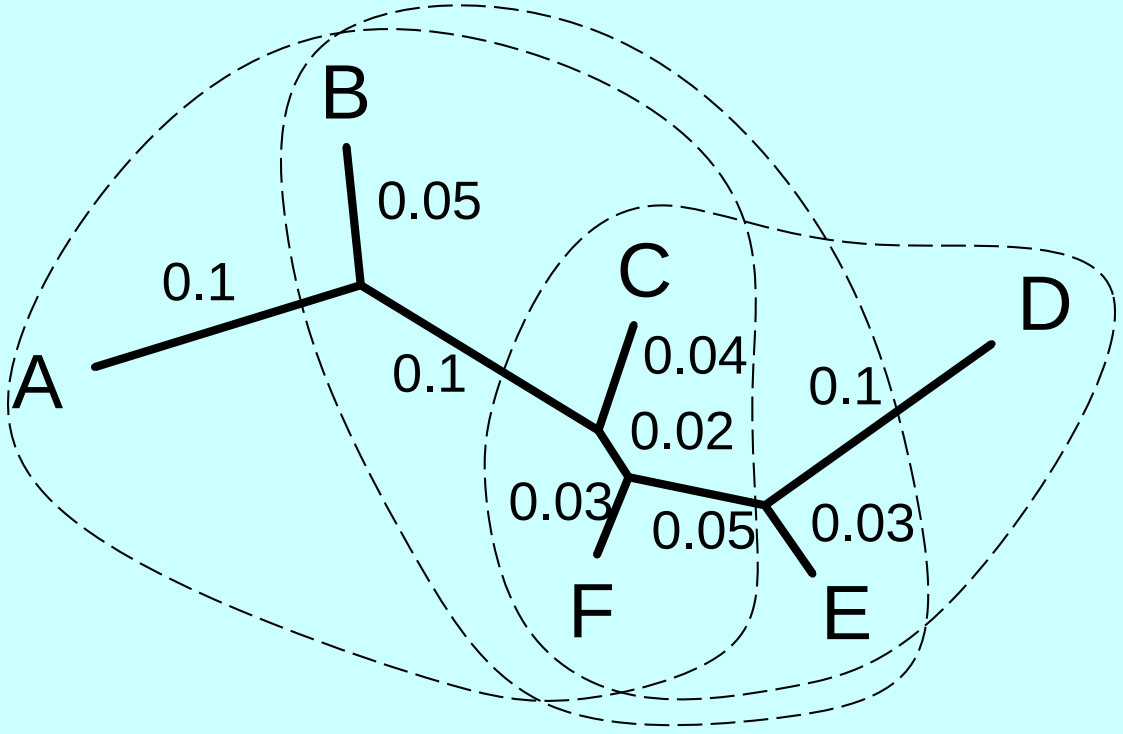
Goloboff's economy in computing scores of rearranged trees
Once the "views" have been computed, they can be taken to represent subtrees, without going inside those subtrees

Star decomposition



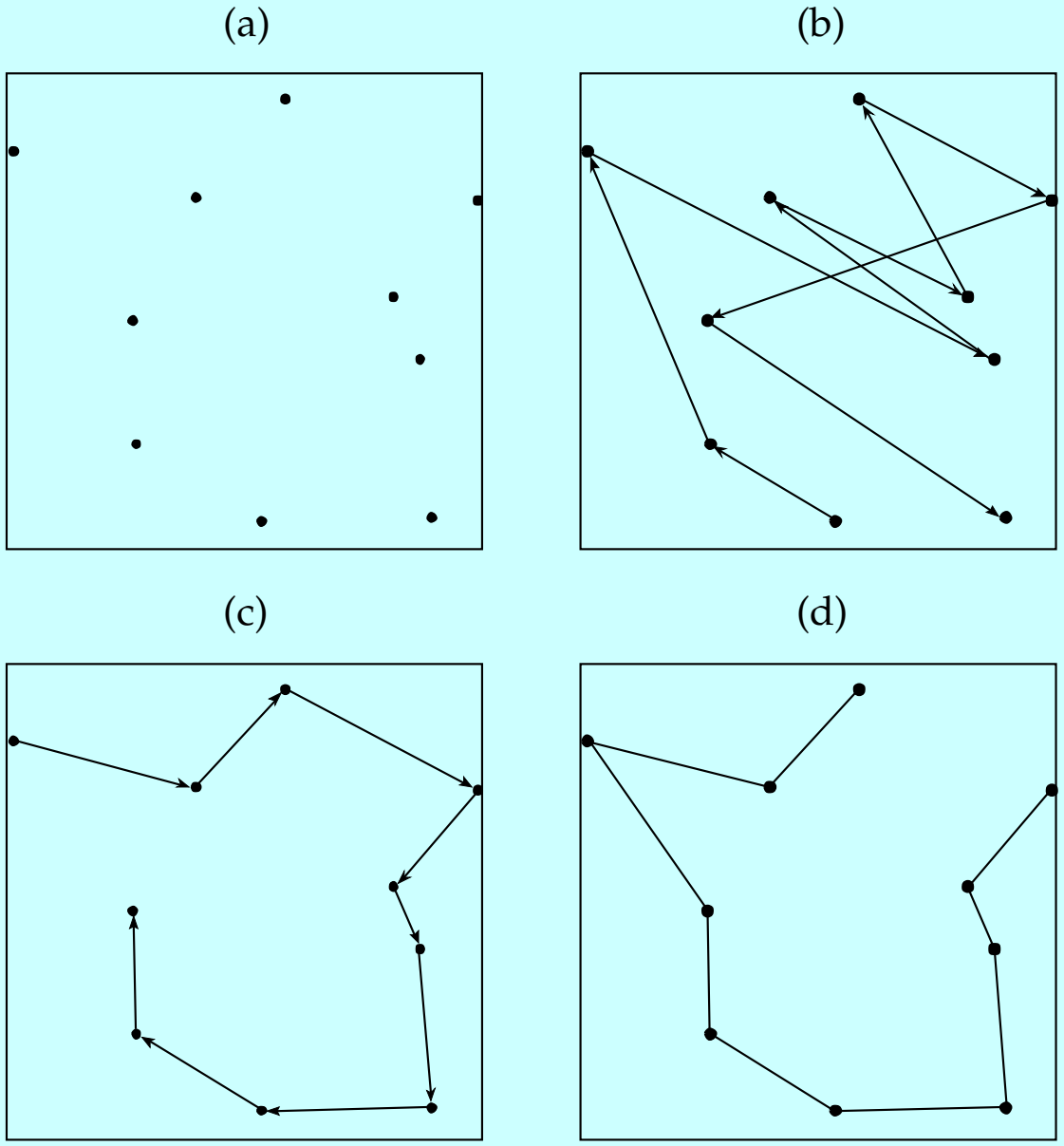
“Star decomposition" search for best tree can happen in multiple ways

Disk-covering

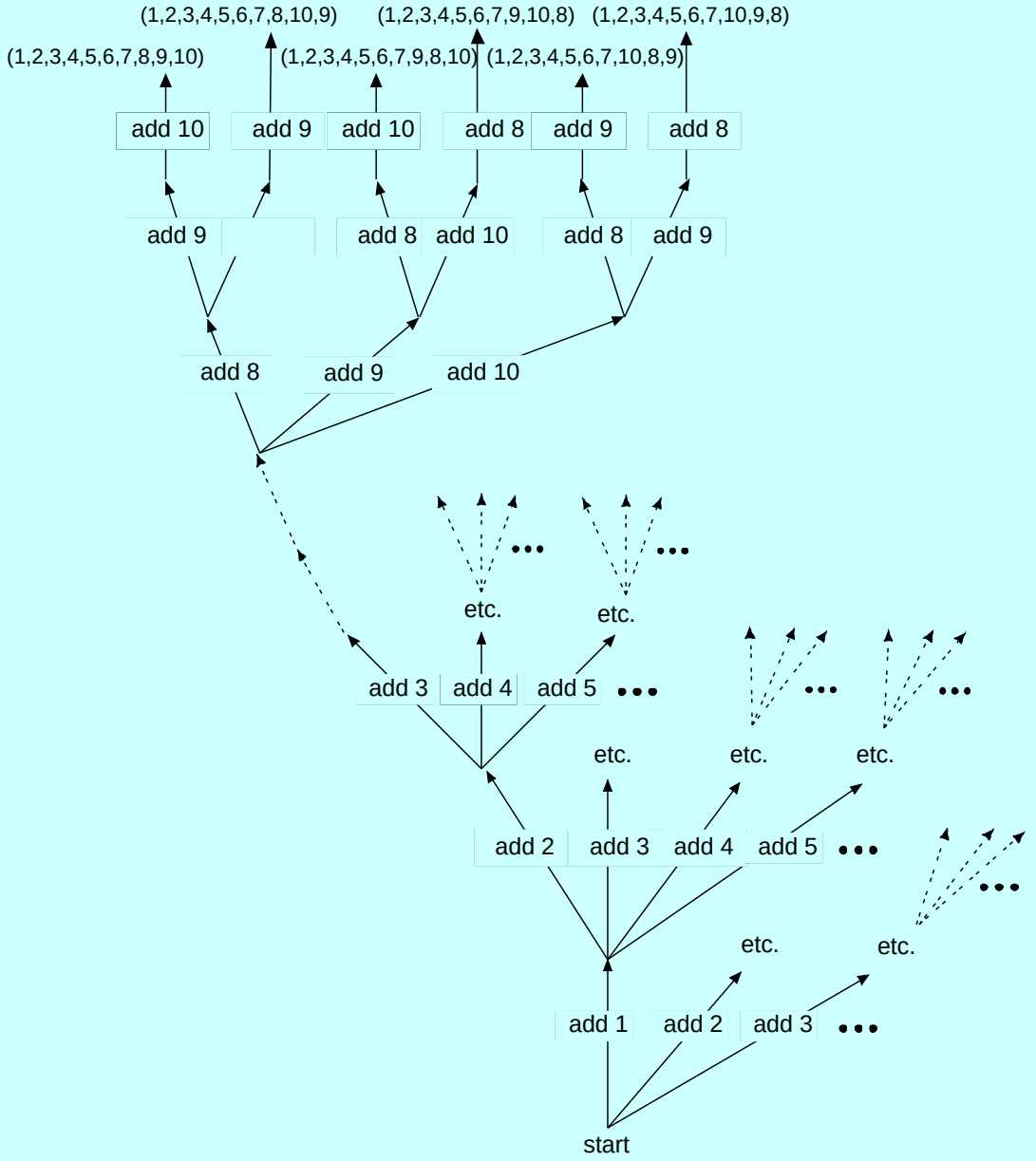


“Disk covering" – assembly of a tree from overlapping estimated subtrees

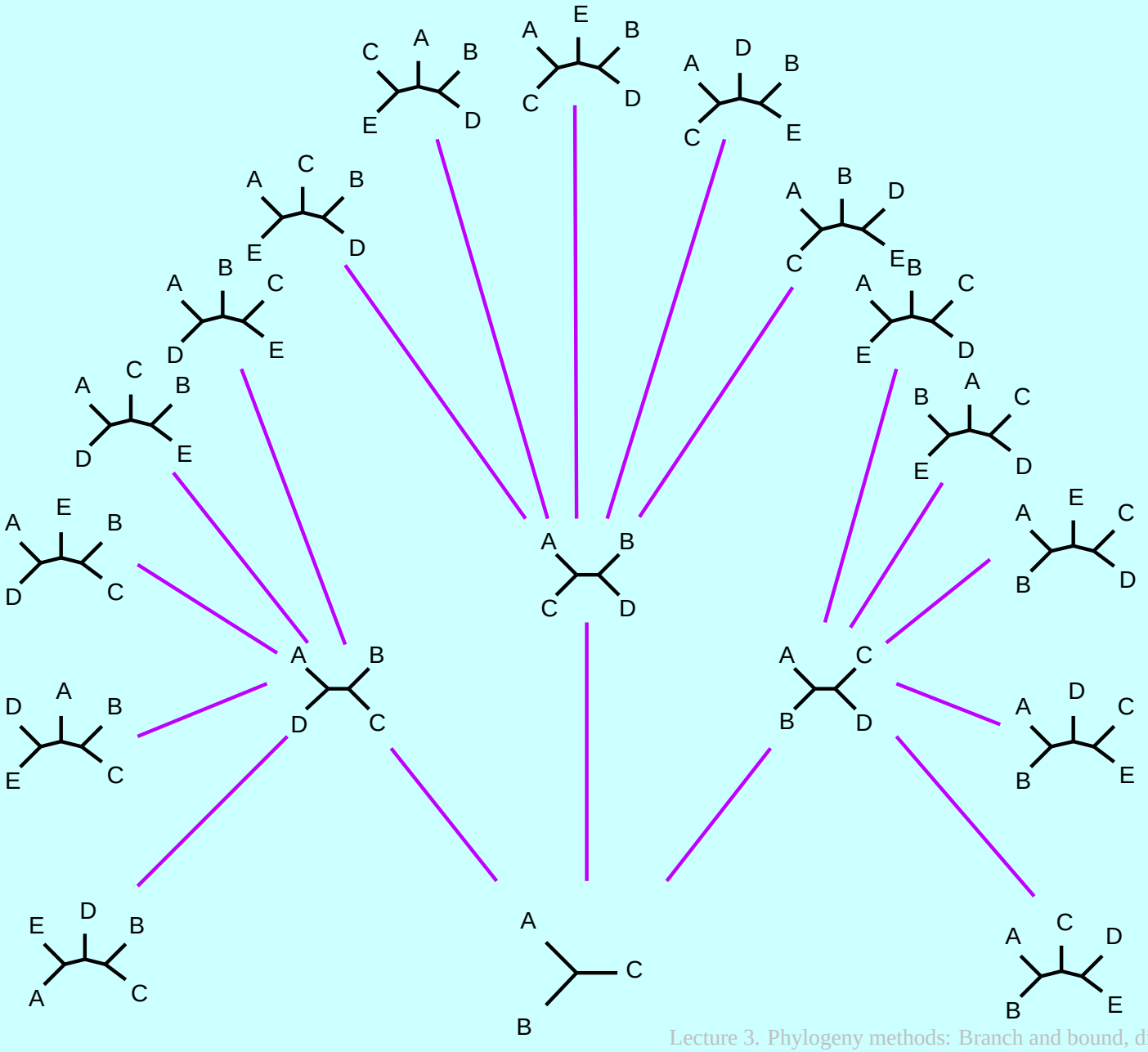
Shortest Hamiltonian path problem



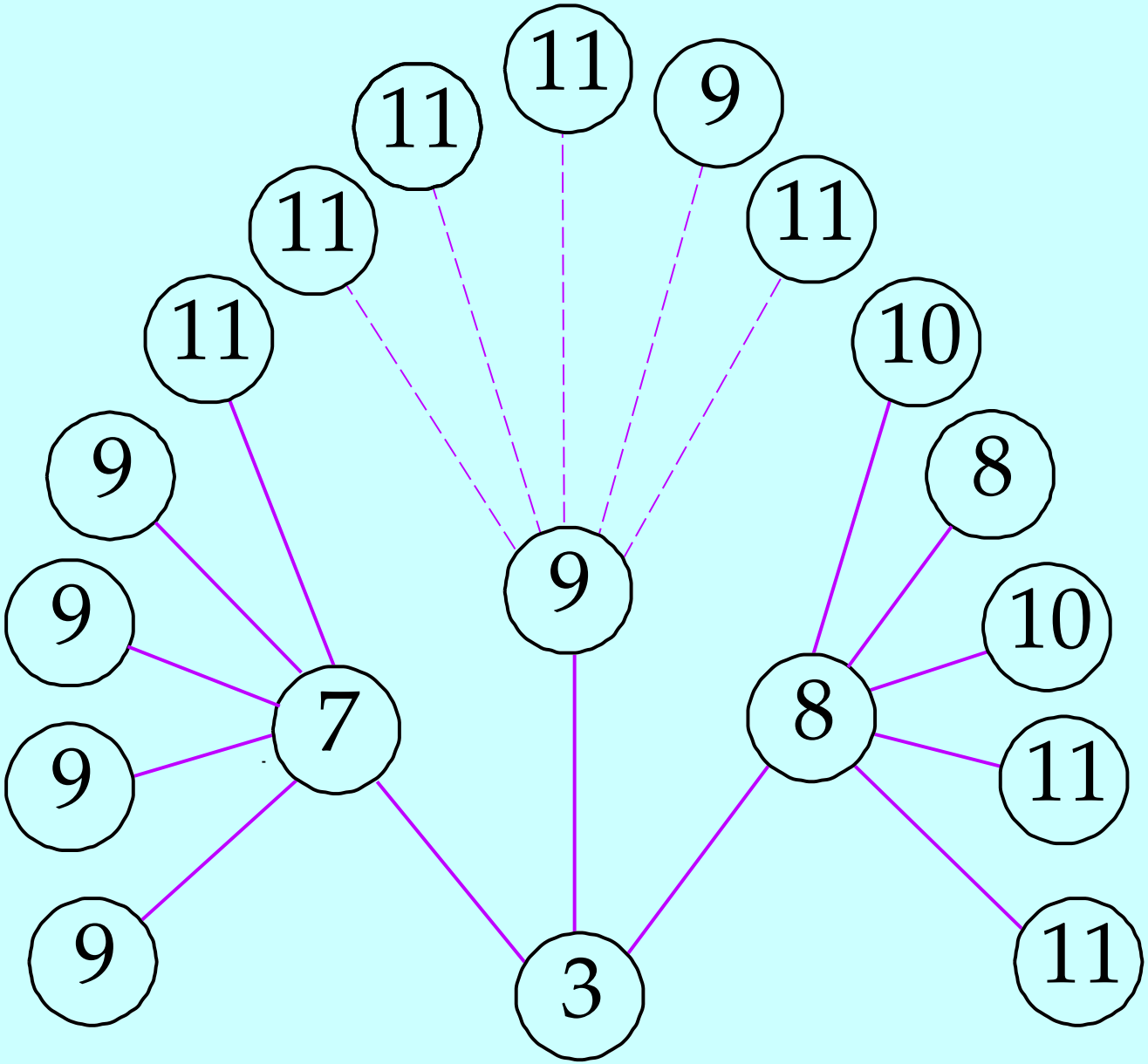
Search tree for this problem



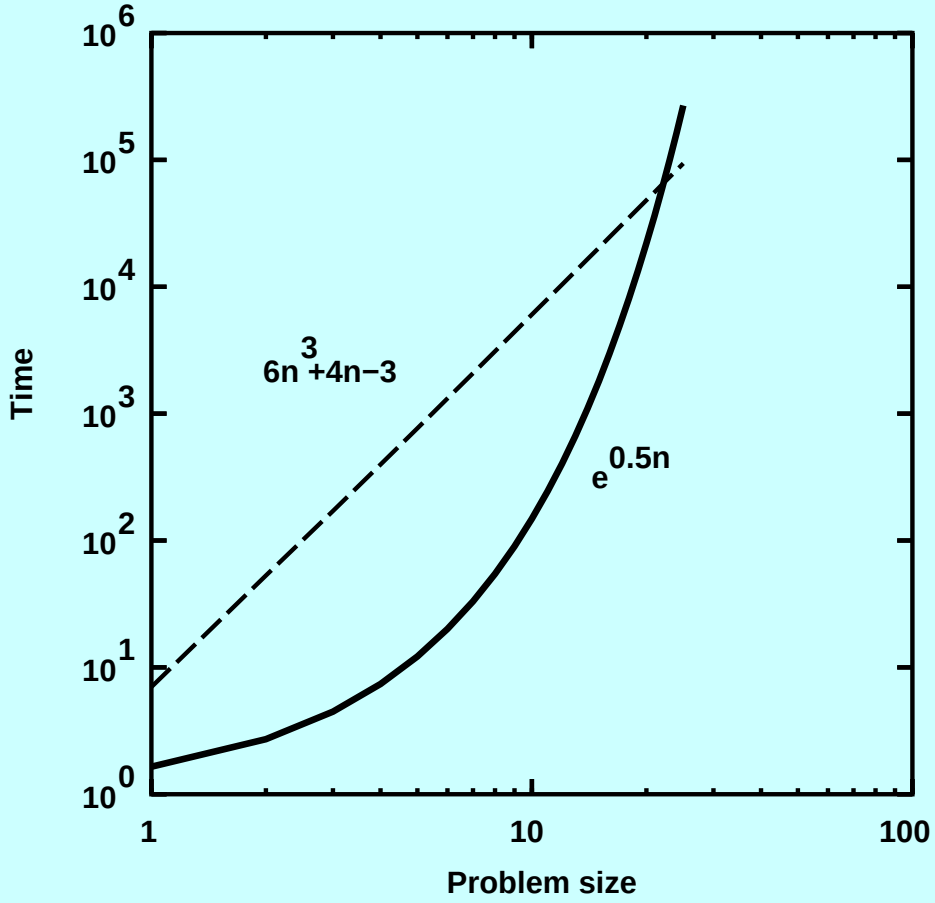
Search tree of trees



same, with parsimony scores in place of trees

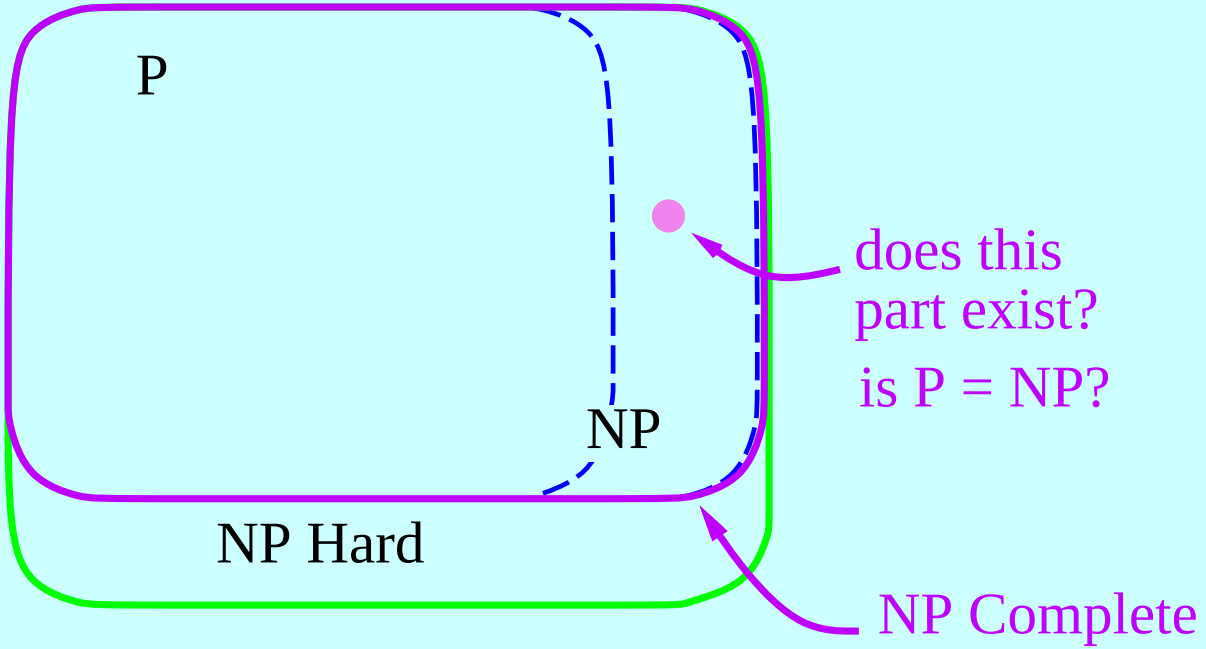


Polynomial time and exponential time



How does the time taken by an algorithm depend on the size of the problem? If it is a polynomial (even one with big coefficients), with a big enough case it is faster than one that depends on the size exponentially.

NP completeness and NP hardness



(This diagram is not quite correct – see the diagrams on the Wikipedia page for “NP-hard”).

P = problems that can be solved by a polynomial time algorithm

NP complete = problems for which a proposed solution can be checked in polynomial time but for which it can be proven that if one of them is in P, all are.

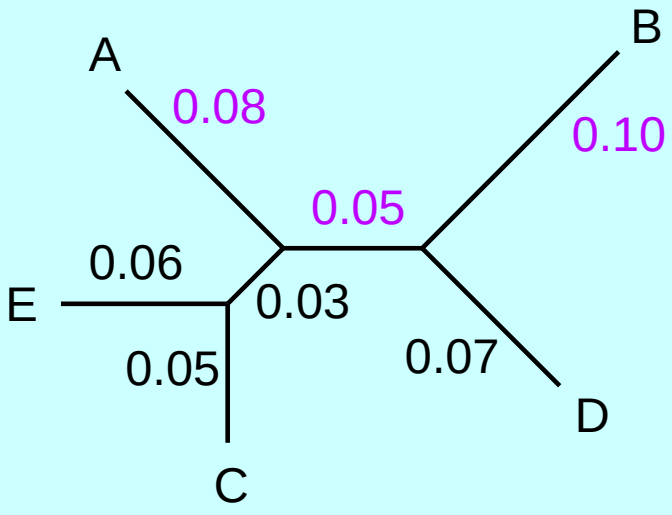
NP hard = problems for which a solution can be checked in polynomial time, but might be not solvable in polynomial time.

Distance methods

These have been attractive, particular to mathematical scientists who love geometry. This has its good and bad effects.

1. Take the sequences in all pairs.
2. For each pair compute a distance. (As we will see, this is best thought of as the length of the 2-species tree for those species).
3. Try to find that tree which best fits the table of distances.

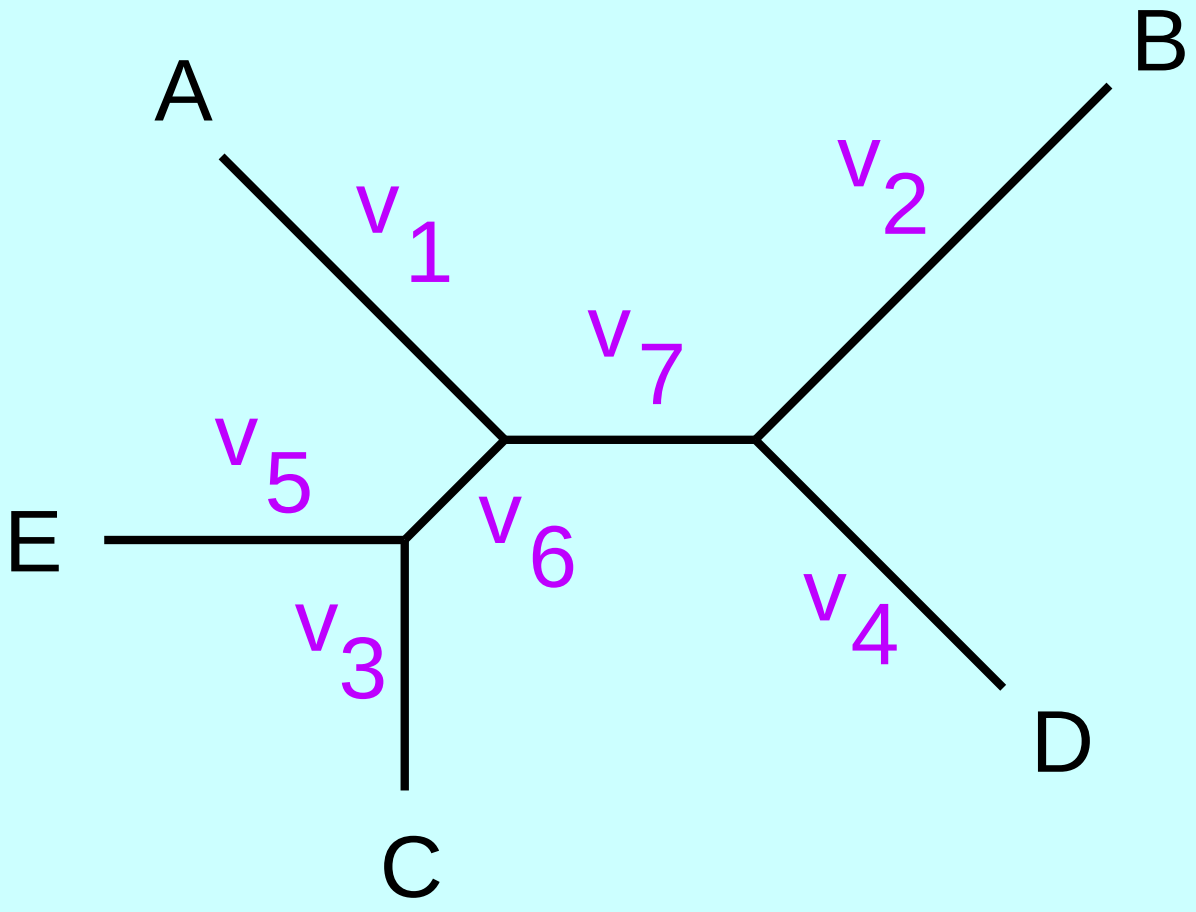
A phylogeny with branch lengths



	A	B	C	D	E
A	0	0.23	0.16	0.20	0.17
B	0.23	0	0.23	0.17	0.24
C	0.16	0.23	0	0.15	0.11
D	0.20	0.17	0.15	0	0.21
E	0.17	0.24	0.11	0.21	0

and the pairwise distances it predicts

A phylogeny with branch lengths



Least squares trees

Least squares methods minimize

$$Q = \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} (D_{ij} - d_{ij})^2$$

over all trees, using the distances d_{ij} that they predict. Cavalli-Sforza and Edwards suggested $w_{ij} = 1$, Fitch and Margoliash suggested $w_{ij} = 1/D_{ij}^2$.

Statistical assumptions of least squares trees

Implicit assumption is that distances are (independently?) Normally distributed with expectation d_{ij} and variance proportional to $1/w_{ij}^2$:

$$D_{ij} \sim \mathcal{N}(d_{ij}, K/w_{ij})$$

Thus the different weightings correspond to different assumptions about the error in the distances. Also, there is assumed to be no covariance of distances.

In fact, the distances will covary, since a change in an interior branch of the tree increases (or decreases) all distances whose paths go through that branch.

Matrix approach to fitting branch lengths

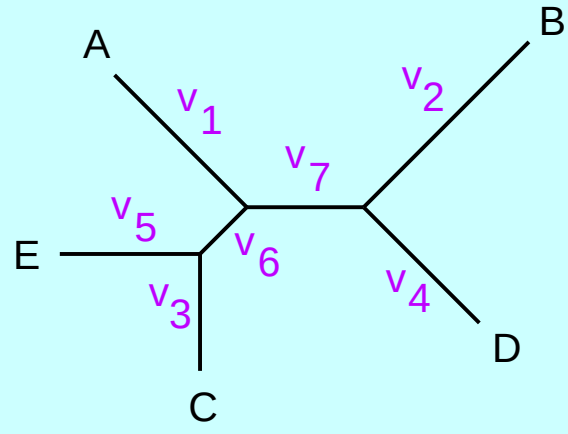
If we stack the distances up into a column vector $\mathbf{D}$, we can solve the least squares equation (obtained by taking derivatives of the quadratic form Q):

$$\mathbf{D}^{\mathsf{T}} = (\mathsf{D}_{12}, \mathsf{D}_{13}, \mathsf{D}_{14}, \mathsf{D}_{15}, \mathsf{D}_{23}, \mathsf{D}_{24}, \mathsf{D}_{25}, \mathsf{D}_{34}, \mathsf{D}_{35}, \mathsf{D}_{45})$$

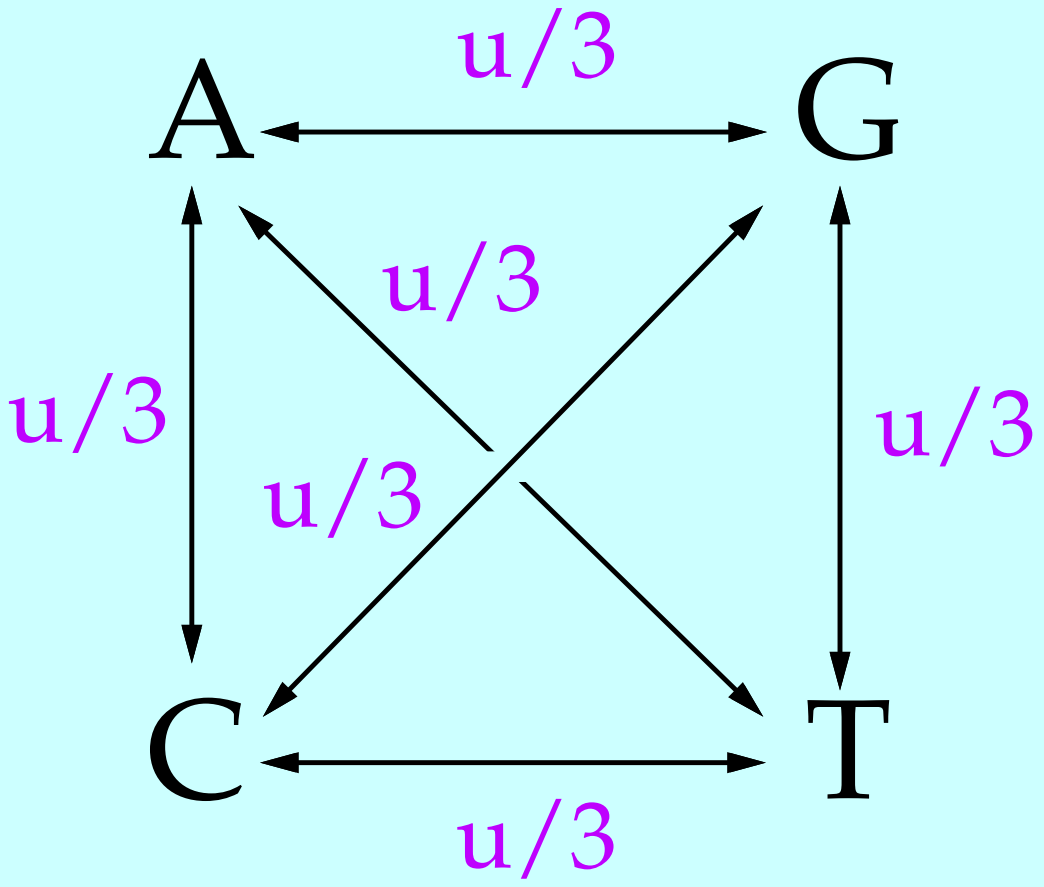
$$\mathbf{X}^{\mathsf{T}}\mathbf{D} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)\mathbf{v}.$$

where the "design matrix" $\mathbf{X}$ for the given tree topology has 1's whenever a given branch lies on the path between those two species. Here is the design matrix for the tree we just saw.

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccccccc} \multicolumn{7}{c}{\text{Branches}} \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \left[\begin{array}{ccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{array}\right] \end{array} \begin{array}{c} \text{which} \\ \mathsf{D} \\ 1,2 \\ 1,3 \\ 1,4 \\ 1,5 \\ 2,3 \\ 2,4 \\ 2,5 \\ 3,4 \\ 3,5 \\ 4,5 \end{array}$$



The Jukes-Cantor model for DNA



Derivation of the probability of change

1. Imagine events occuring at rate $\frac{4}{3}u$ per unit time which replace a base by one of the 4 bases chosen at random.

2. Persuade yourself that this is no different in outcome from events u per unit time that replace it by one of the other 3 chosen at random.

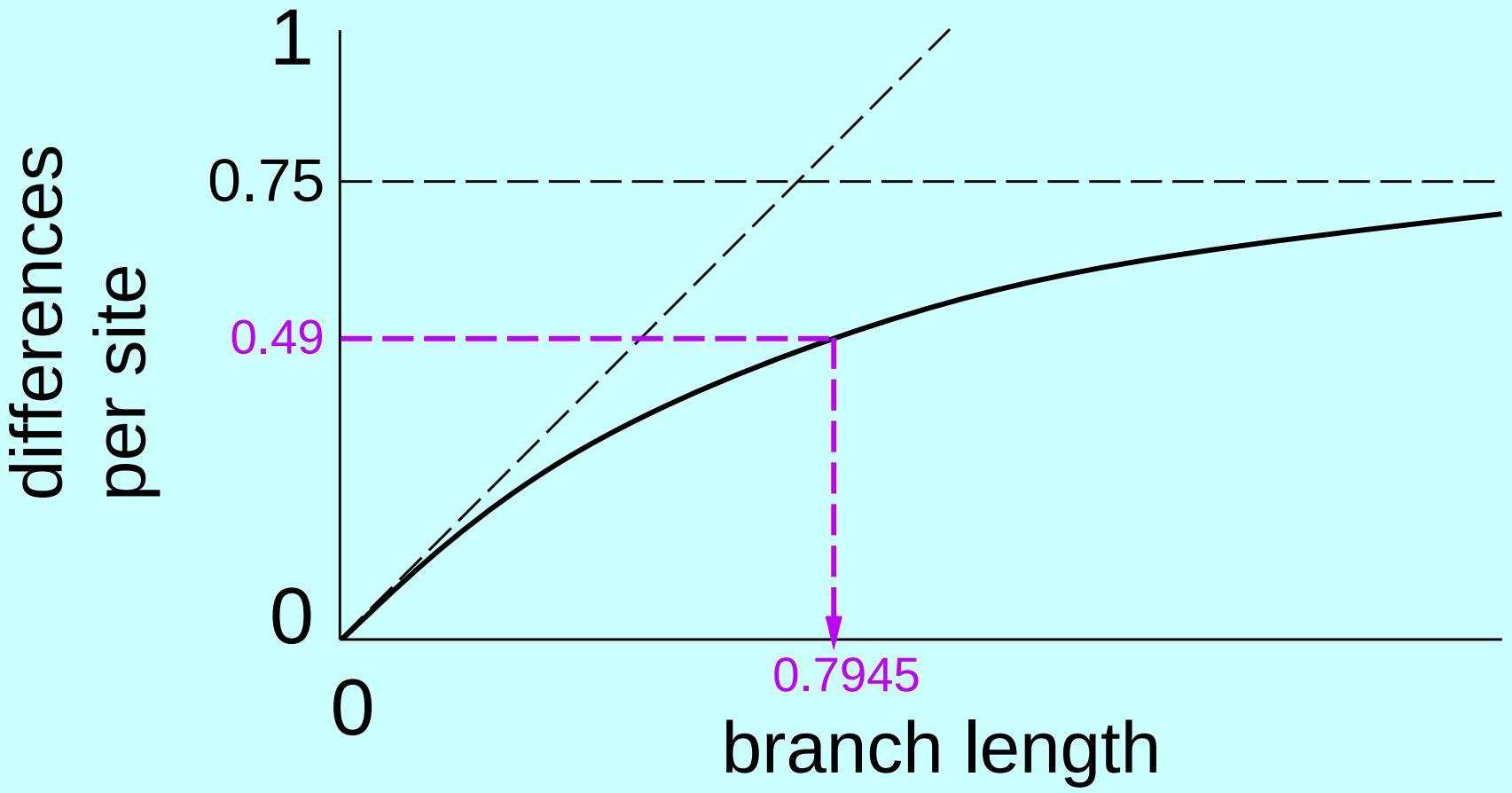
3. The probability a branch has none of these (first kind of) events if it is of length t is $\exp(-\frac{4}{3}ut)$. (Think the zero term of a Poisson distribution).

4. If it does have one or more of these events, you end up with one of the 4 bases chosen at random.

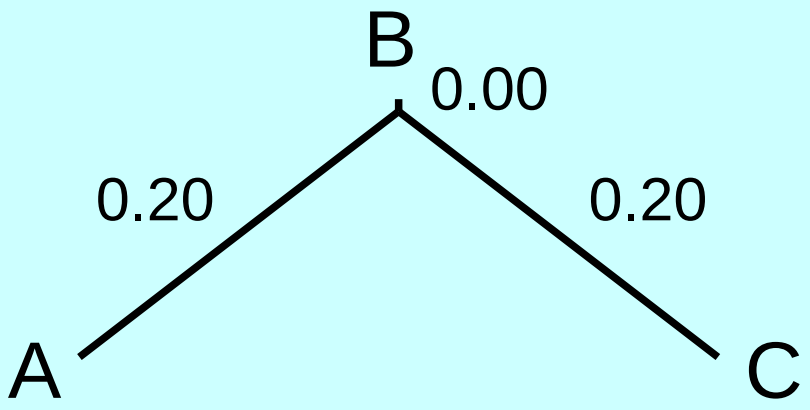
5. Therefore the probability of a net change is:

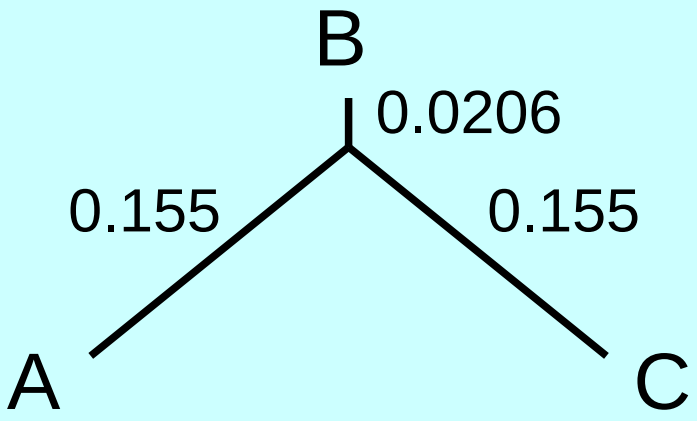
$$\frac{3}{4}\left(1-e^{(-\frac{4}{3}ut)}\right)$$

The distance for the Jukes-Cantor model



If you don’t correct for “multiple hits”





Left: the true tree. Right: a tree fitting the uncorrected distances

References, page 1

Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology* **40:** 315-328. **[Discusses heuristic search strategy involving ties, multiple starts]**

Farris, J. S. 1970. Methods for computing Wagner trees. *Systematic Zoology* **19:** 83-92. **[Early parsimony algorithms paper is one of first to mention sequential addition strategy]**

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4:** 406-425. **[First mention of star-decomposition search for best trees, sort of]**

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* **13:** 964-969. **[Assembles trees out of quartets]**

Huson, D., S. Nettles, L. Parida, T. Warnow, and S. Yooseph. 1998. The disk-covering method for tree reconstruction. pp. 62-75 in *Proceedings of "Algorithms and Experiments" (ALEX98), Trento, Italy, Feb. 9-11, 1998*, ed. R. Battiti and A. A. Bertossi. **["Disk-covering method" for long stringy trees]**

References, page 2

Foulds, L. R. and R. L. Graham. 1982. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics* **3:** 43-49. **[Parsimony is NP-hard]**

Graham, R. L. and L. R. Foulds. 1982. Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computat ional time. *Mathematical Biosciences* **60:** 133-142. **[... and more]**

Hendy, M. D. and D. Penny. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* **60:** 133-142 **[Introduced branch-and-bound for phylogenies]**

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts. **[For this lecture the material is chapters 4, and 5]**

Semple, C. and M. Steel. 2003. *Phylogenetics.* Oxford University Press, Oxford. **[Also covers search strategies]**

References, page 3

Felsenstein, J. 1984. Distance methods for inferring phylogenies: a justification. *Evolution* **38:** 16-24. [Argument for statistical interpretation of distance methods]

Farris, J. S. 1985. Distance data revisited. *Cladistics* **1:** 67 -85. [Reply to my 1984 paper]

Felsenstein, J. 1986. Distance methods: reply to Farris. *Cladistics* **2:** 130-143. [reply to Farris 1985]

Farris, J. S. 1986. Distances and statistics. *Cladistics* **2:** 1 44-157. [debate was cut off after this]

References, page 4

Bryant, D., and P. Waddell. 1998. Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Molecular Biology and Evolution* **15:** 1346-1359. [quicker least squares distance trees]

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts. [See chapter 11]

Semple, C. and M. Steel. 2003. *Phylogenetics*. Oxford University Press, Oxford. [See pp. 145-160]

Yang, Z. 2007. *Computational Molecular Evolution*. Oxford University Press, Oxford. [See pages 89-93]