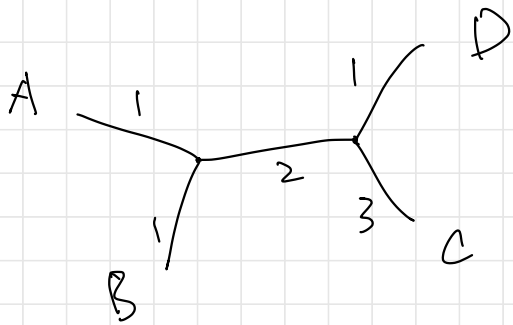
Review:

1. character, states
2. parsimony method
3. Ancestor construction
4. perfect phylogeny
5. maximum likelihood.

Distance Based Methods

- Input of distance based methods is an *n*x*n* distance matrix d(i,j).
- We want to compute a tree with n leaves, with edge weights. T(i,j) is the distance of two leaves on the tree.
- We want to minimize
- This is also NP-hard.
- So we use heuristics.

$$\sum_{i,j} |d(i,j) - T(i,j)|^2$$

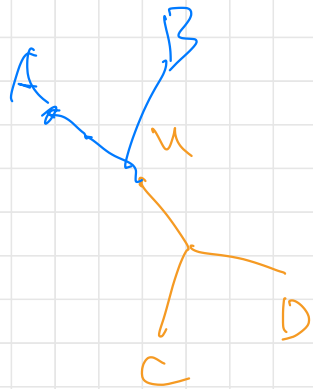


	A	B	C	D
A	0	2	6	4
B	2	0	6	4
C	6	6	0	4
D	4	4	4	0

A, B

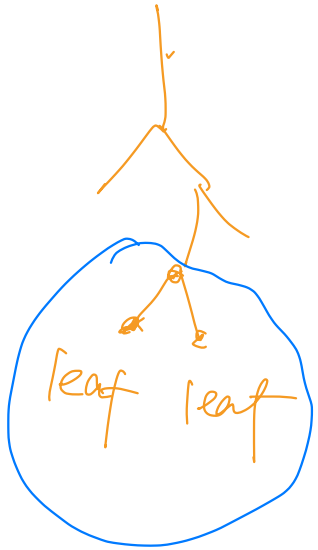
$$d(u,c) = \frac{d(A,C)+d(B,C)}{2}$$

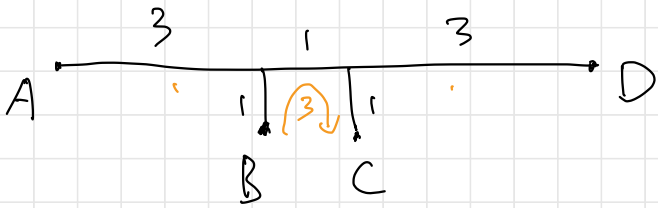
	U	C	D
U	0	6	4
C		0	4
D			0

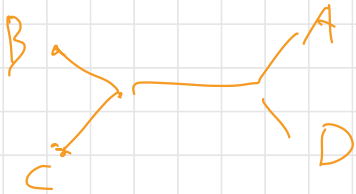


UPGMA

- Unweighted Pair Group Method with Arithmetic mean.
- A heuristic method with no performance guarantee.
- At each time, it finds *i,j* with the minimum distance. siblings.
- Merge the two taxa *i,j* into a new one *u*. Update the distance matrix.
- For any k, let: d(u,k) = (d(i,k)+d(j,k))/2.
- Recursion.




3
1
3
A
D
1
3
1
B
C


B
A
C
D

Neighbor Joining

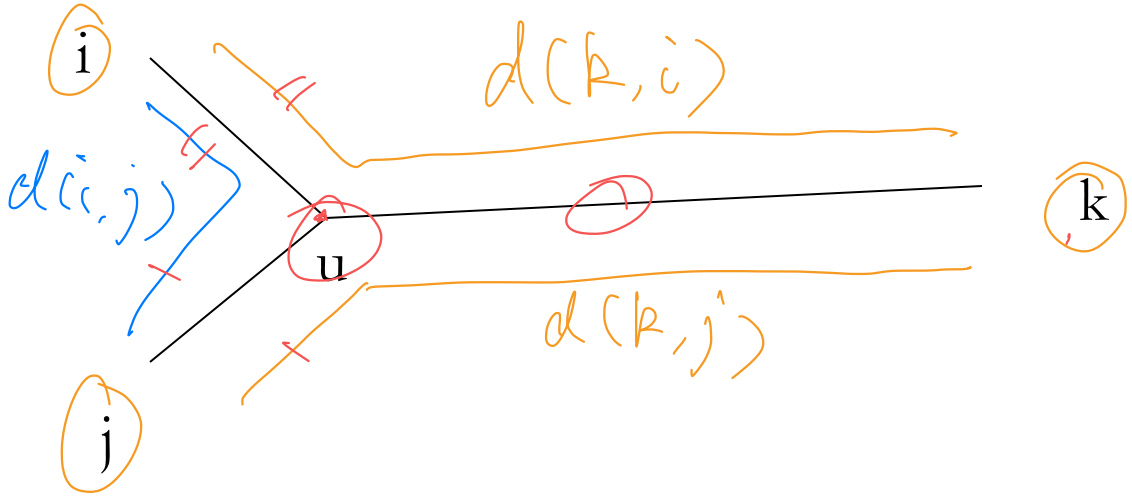
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987 Jul;4(4):406-25.
- Neighbor Joining uses a similar idea as UPGMA. But it uses a more sophisticated formula to determine the two neighbors to be joined.

$$Q(i,j) = (r-2)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k)$$

- Find the minimum Q(i,j), merge i,j to a new node u.
- Update d(k,u) = (d(k,i)+d(k,j)-d(i,j))/2.
- Recursion on the remaining r-1 nodes.

$$-Q(i,j) = \left(\sum_{k=1}^{r} \left(d(i,k) + d(j,k) - d(i,j) \right) \right) + 2d(i,j)$$

Neighbor Joining Idea



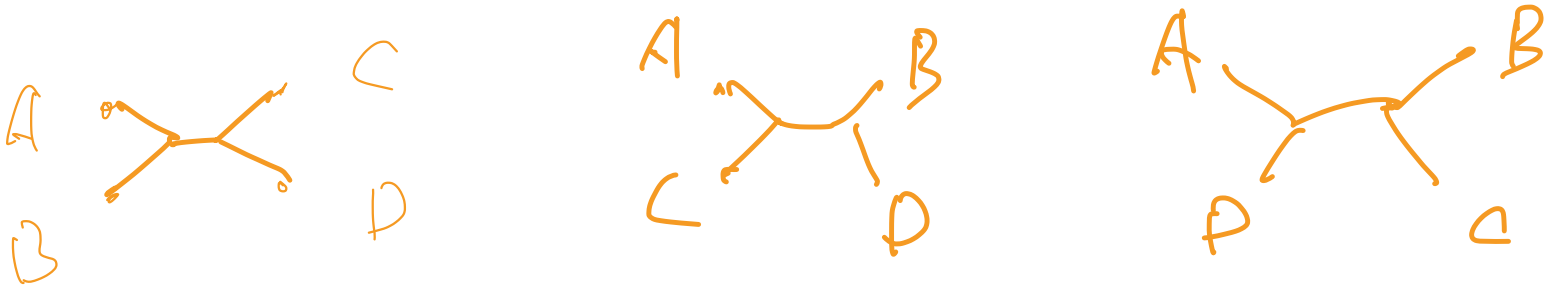
$$d(k,u) = (d(k,i)+d(k,j)-d(i,j))/2$$

Methods for constructing phylogeny

- Exhaustive Search
  - Parsimony Method: Find a tree topology so that the total number of mutations on the edges is the smallest.
  - Perfect Phylogeny.
  - Maximum Likelihood.
- Distance Based Method
  - UPGMA
  - Neighbour Joining
- Quartet Method
- Whole Genome Phylogeny.

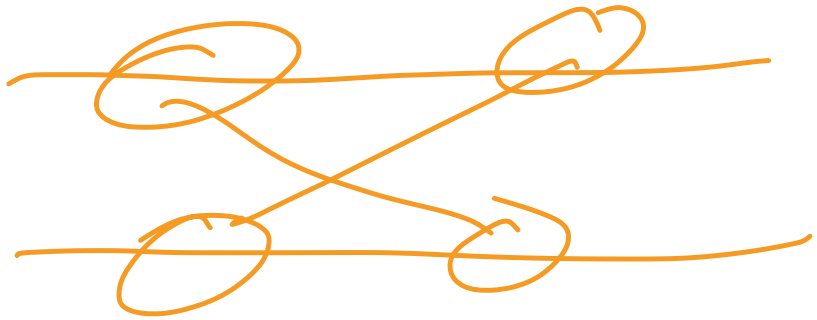
Quartet Methods

- For each group of four species, construct a tree of 4 (quartet), using your most favorite method, say maximum likelihood.
- Then find a tree that is most consistent with all the quartets.
- The problem is NP-hard (to find the tree with least error).
- There is a PTAS to do this (T. Jiang, P. Kearney, and M. Li. Orchestrating quartets: approximation and data (FOCS'98))



Challenges in Phylogeny of Chain Letter

- Parsimony or maximum likelihood: How do we know what is a "character" – a feature to look at?
  - In case of a chain letters?
  - In case of whole genome phylogeny?
- Distance based: What distance to use? edit distance doesn't work
- A "universal" solution -- information distance.



Information Distance

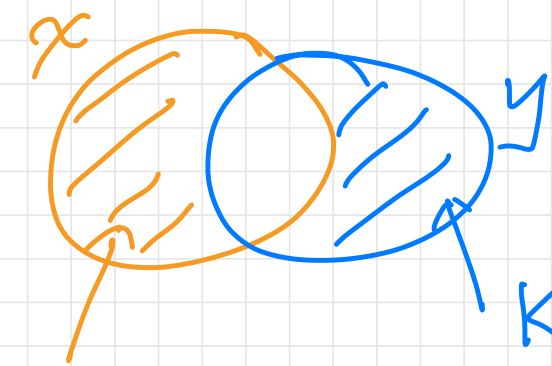
- Let x and y be two strings. P be the shortest program that takes x as input and computes y.
- The length of this shortest program defines K(y|x). K is called the Kolmogorov Complexity.
- The information distance between x and y is defined as
- d(x,y) = max { K(x|y), K(y|x)} / max{K(x),K(y)}.
- Kolmogorov complexity is incomputable, but we can use compression in practice.

$\underbrace{AAA \cdots A}_{10000}$

$K(x)$: shortest program length to output x.

$\pi = 3.14159265357\cdots$

$K(x|y)$: shortest program length to accept y as input and output x.



$$d(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

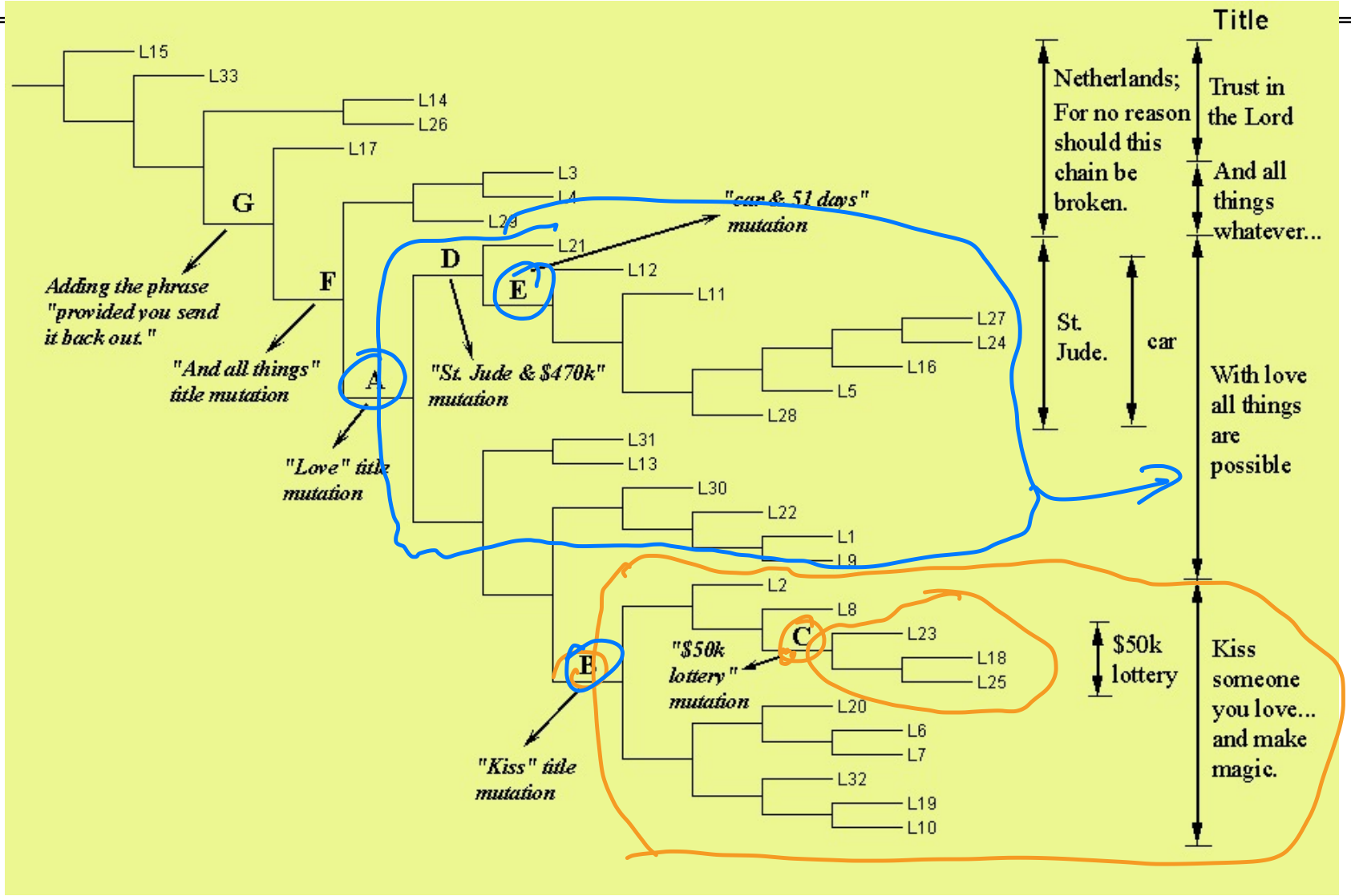
Reconstructing History of Chain Letters

$K(x|y) \approx K(xy) - K(y)$

$K(y|x) \approx K(xy) - K(x)$.

- For each pair of chain letters (x, y) we computed $d(x,y)$, hence a distance matrix.
- A DNA compression program is used to compute the information distance.
- Using Neighbor Joining to construct their evolutionary history based on the d(x,y) distance matrix.
- The resulting tree is an almost perfect phylogeny: distinct features are all grouped together.

Phylogeny of 33 Chain Letters



Summary

- Exhaustive Search
  - Parsimony Method: Find a tree topology so that the total number of mutations on the edges is the smallest.
  - Perfect Phylogeny.
  - Maximum Likelihood.
- Distance Based Method
  - UPGMA
  - Neighbour Joining
- Quartet Method
- Whole Genome Phylogeny.
- Chain Letter and 2019 nCoV