Phylogeny

Phylogeny

- A phylogeny (also known as a phylogenetic tree): a diagrammatic hypothesis about the history of the evolutionary relationships of a group of organisms.

Phylogenetic Tree of Life

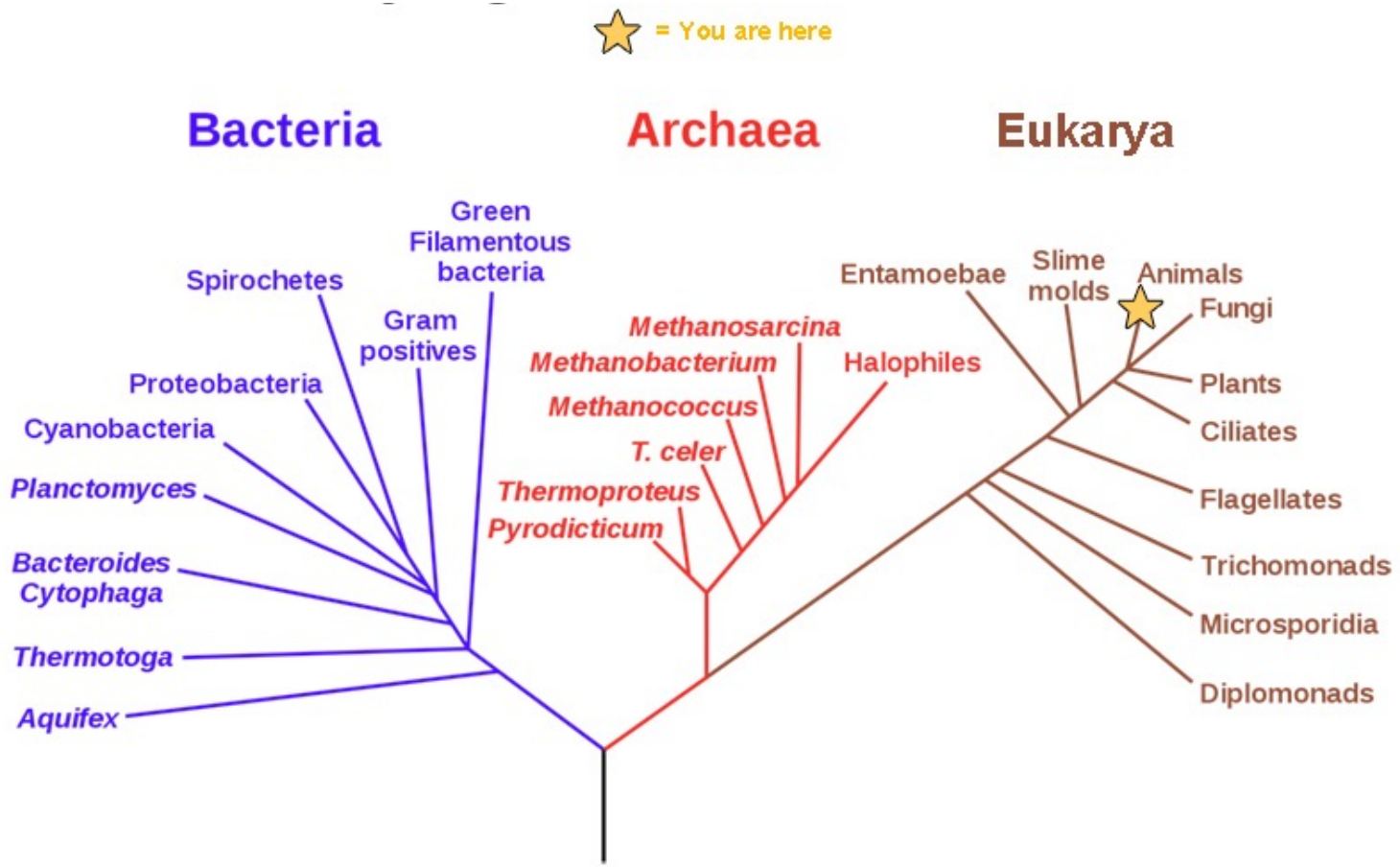


Image credit: https://courses.lumenlearning.com/wmopen-nmbiology1/chapter/phylogenetic-trees/

Phylogenetic Tree of Animals

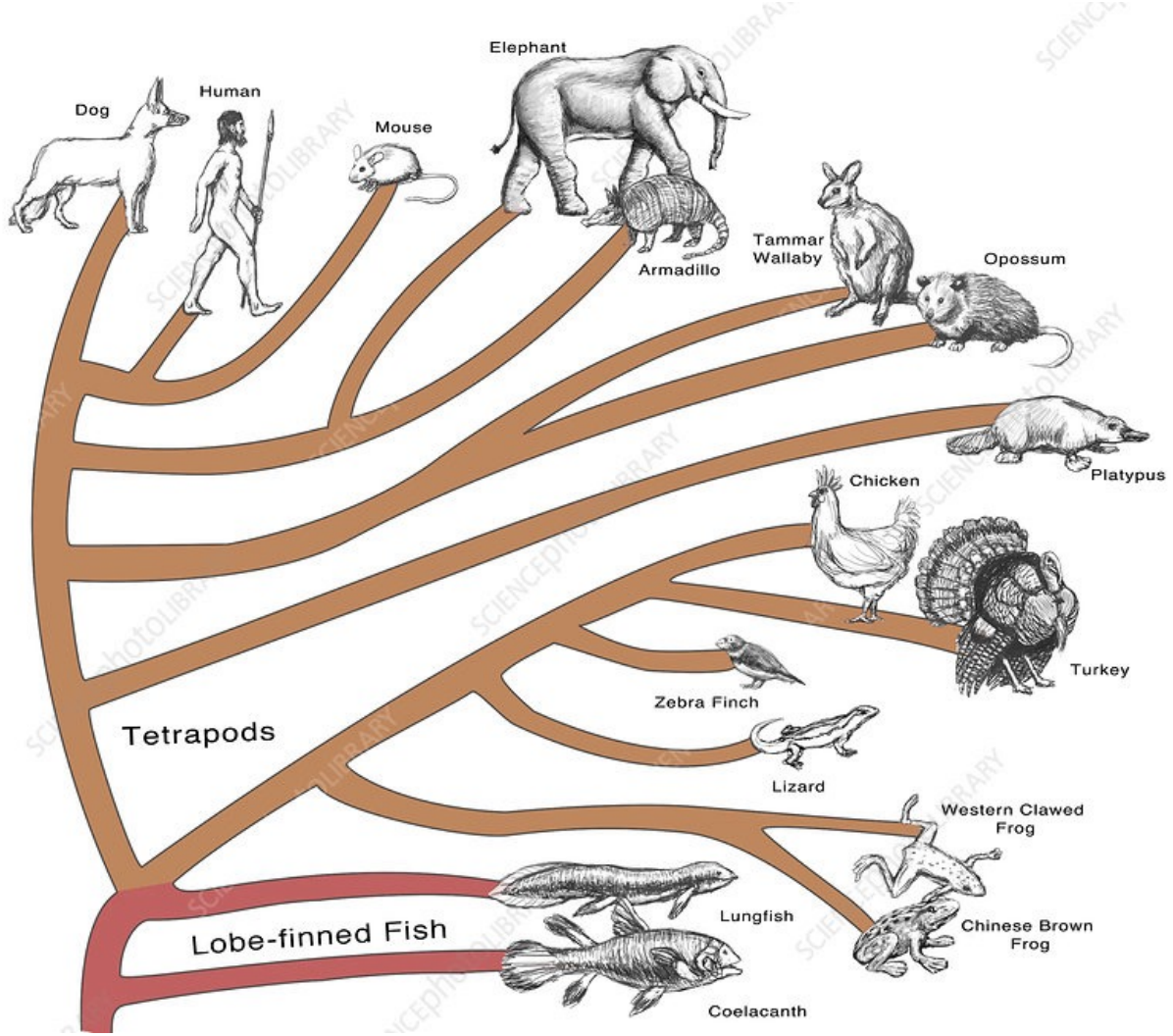
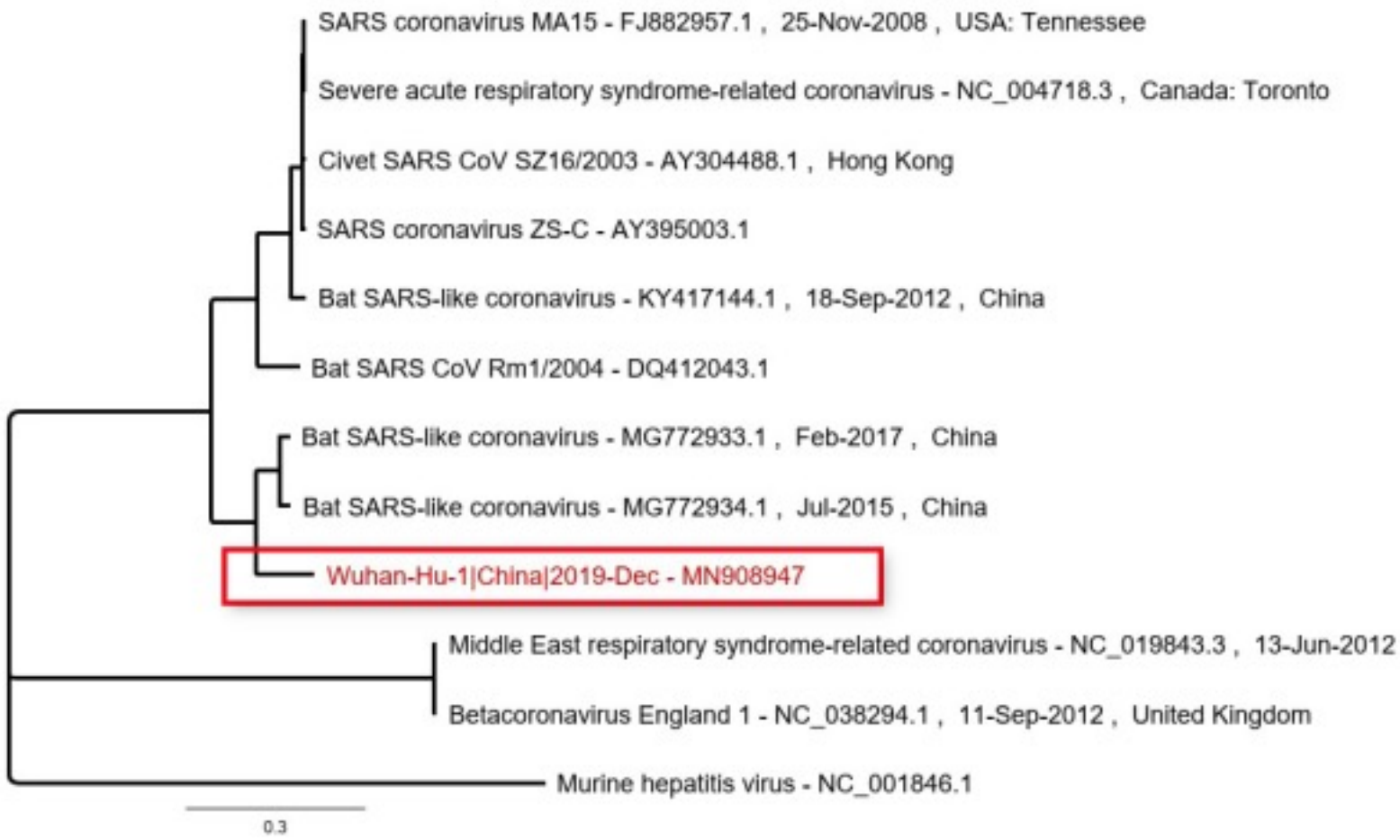


Image credit: https://www.sciencephoto.com/media/780514/view/evolutionary-tree-animals

Phylogeny Tree of Coronavirus



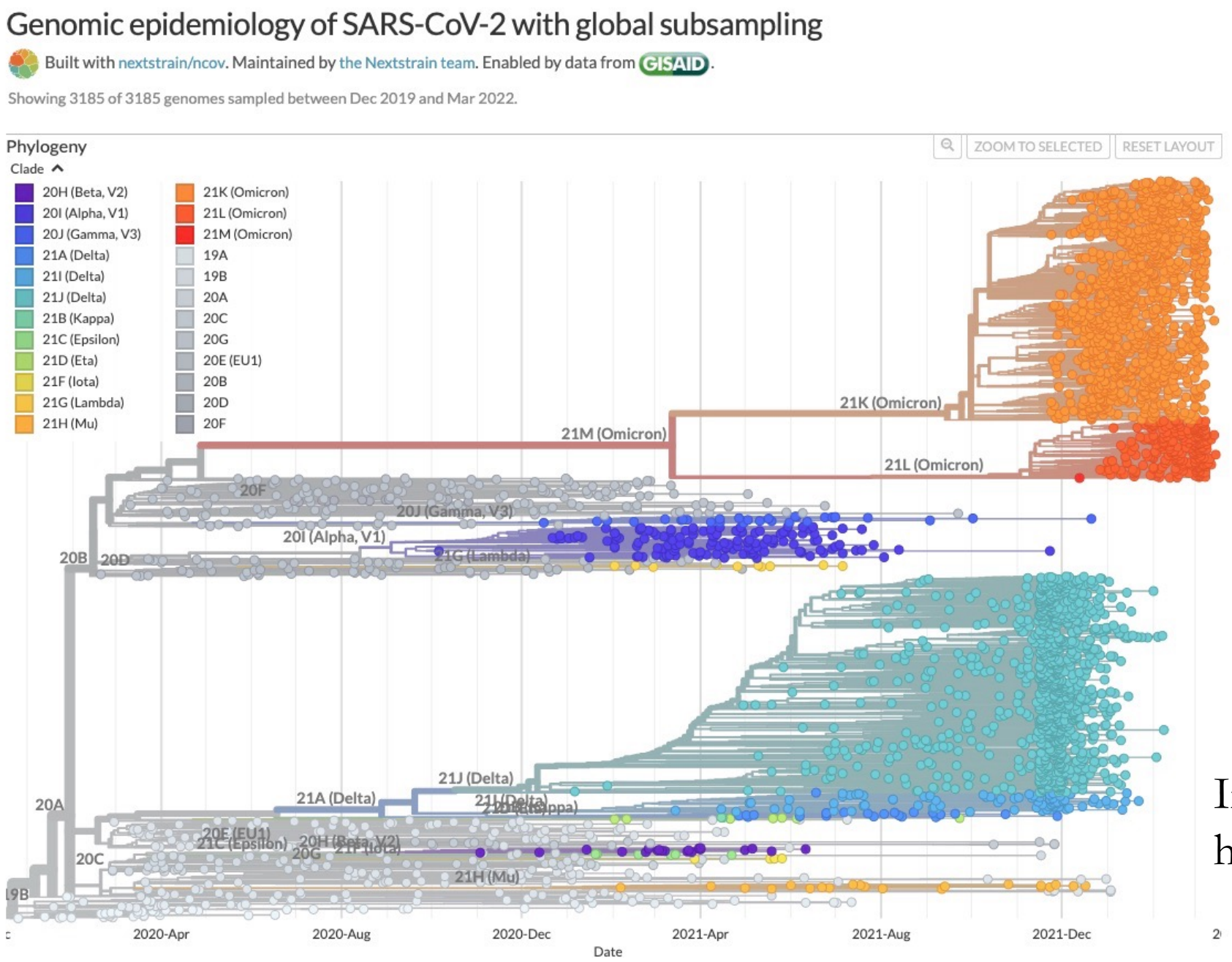
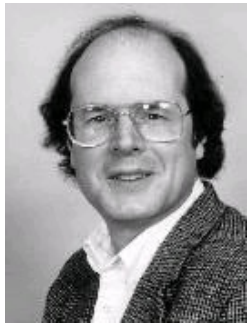


Image credit:
https://nextstrain.org/

Other Examples for Phylogenetic Analysis

- Examples outside of Biology:
  - Study how the news articles copy each other.
  - Study how chain letters evolve (our example to use).
- Terminologies:
  - Organisms: the studied subjects.
  - Taxon (taxa): A **taxon** is a group of one (or more) populations of organism(s).
  - Species: (usually) the largest group of organisms in which any two individuals of appropriate sexes can produce fertile offspring.

In 1997, on a Hong Kong mountain, the story begins …



Charles H. Bennett

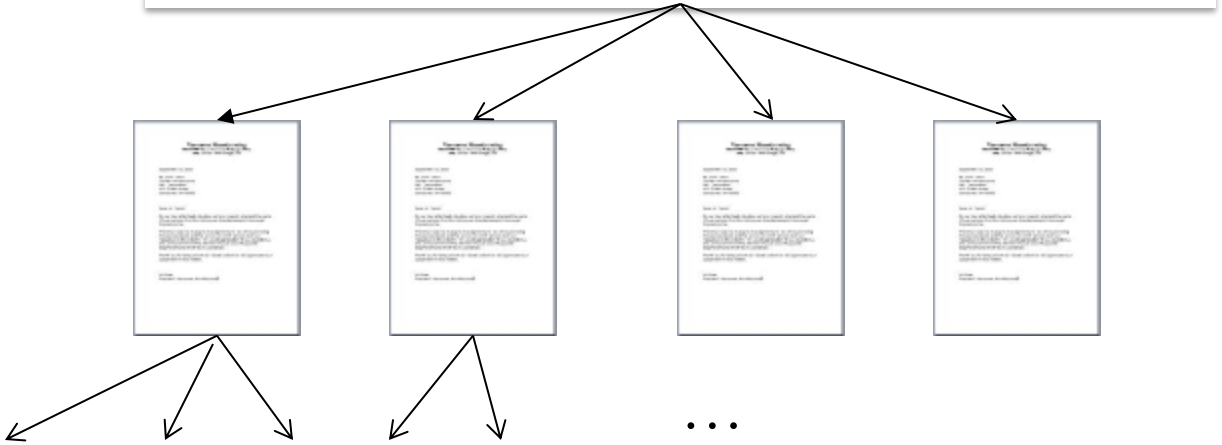
Ming Li

Lion Rock

Chain letters – old style

- Charles Bennett collected 33 copies of chain letters that were apparently from the same origin during 1980—1997.
- How does it work?

Make 20 copies and send to your friends.
Yes ➔ some good things will happen
No ➔ some bad things will happen.



Trust in the Lord with all your heart and he will acknowledge and He will light the way. This Prayer has been sent to you for good luck. The original copy is from the Netherlands. It has been around the world nine times. The luck has been brought to you. You are to receive good luck within four days of receiving this letter. This is nojoke. You will receive it in the mail. Send copies of this letter to people you think need good luck. Do not send money. Do not keep this letter. It must leave your hands within ninety six hours after you receive it. An RAF officer received $70,000. Don Elliott received $50,000 and lost it because he broke the chain. While in the Phillipines, General Welch lost his life six days after he received this letter. He failed to circulate the Prayer. However, before his death, he received $775,000. Please send twenty copies and see what happens to you on the fourth day. This chain comes from Venzuela and was written by Sol Anthony De Cadif, a missionary from South America. Since this chain must make a tour of the world. you must make twenty copies identical to this one and send it to your friends, parents, and acquantances. After a few days you will get a surprise. This is true, even if you are not superstitious. Take note of the following. Constantine Diaz received the chain in 1953. He asked his secretary to make twenty copies and send them. A few days later he won a lottery for two million dollars in his country. Carlo Craduit, and office employee, received the chain. He forgot it and in a few days lost his job. He found the chain and sent it to twenty people. Five days later he got an even better job. Dolin Moirchild received the chain and not believing in it, threw it away. Nine days later he died. For no reason what so ever should this chain be broken

An Example Chain Letter

A sample letter:

Trust in the Lord with all your heart

Trust in the Lord with all your heart and he will acknowledge and He will light the way. This Prayer has been sent to you for good luck. The original copy is from the Netherlands. It has been around the world nine times. The luck has been brought to you. You are to receive good luck within four days of receiving this letter. This is nojoke. You will receive it in the mail. Send copies of this letter to people you think need good luck. Do not send money. Do not keep this letter. It must leave your hands within ninety six hours after you receive it. An RAF officer received $70,000. Don Elliott received $50,000 and lost it because he broke the chain. While in the Phillipines, General Welch lost his life six days after he received this letter. He failed to circulate the Prayer. However, before his death, he received $775,000. Please send twenty copies and see what happens to you on the fourth day. This chain comes from Venzuela and was written by Sol Anthony De Cadif, a missionary from South America. Since this chain must make a tour of the world. you must make twenty copies identical to this one and send it to your friends, parents, and acquantances. After a few days you will get a surprise. This is true, even if you are not superstitious. Take note of the following. Constantine Diaz received the chain in 1953. He asked his secretary to make twenty copies and send them. A few days later he won a lottery for two million dollars in his country. Carlo Craduit, and office employee, received the chain. He forgot it and in a few days lost his job. He found the chain and sent it to twenty people. Five days later he got an even better job. Dolin Moirchild received the chain and not believing in it, threw it away. Nine days later he died. For no reason what so ever should this chain be broken

send money. Do not keep this letter. It must leave your hands within ninety six hours after you receive it. An RAF officer re

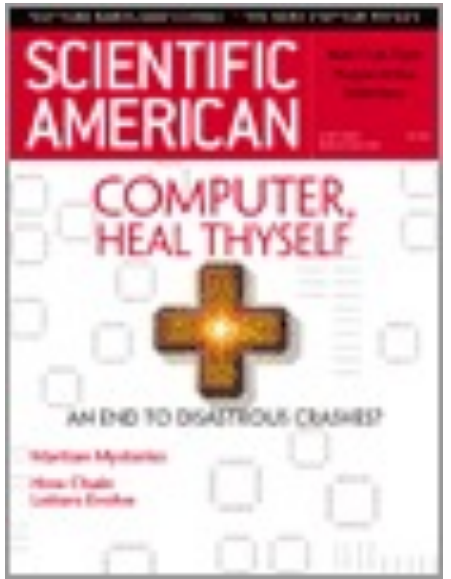
A few days later he won a lottery

the chain. While in the Phillipines, General Welch lost his life six days after he received this letter. He failed to circulate the Prayer. However, before his death, he received $775,000. Please send twenty

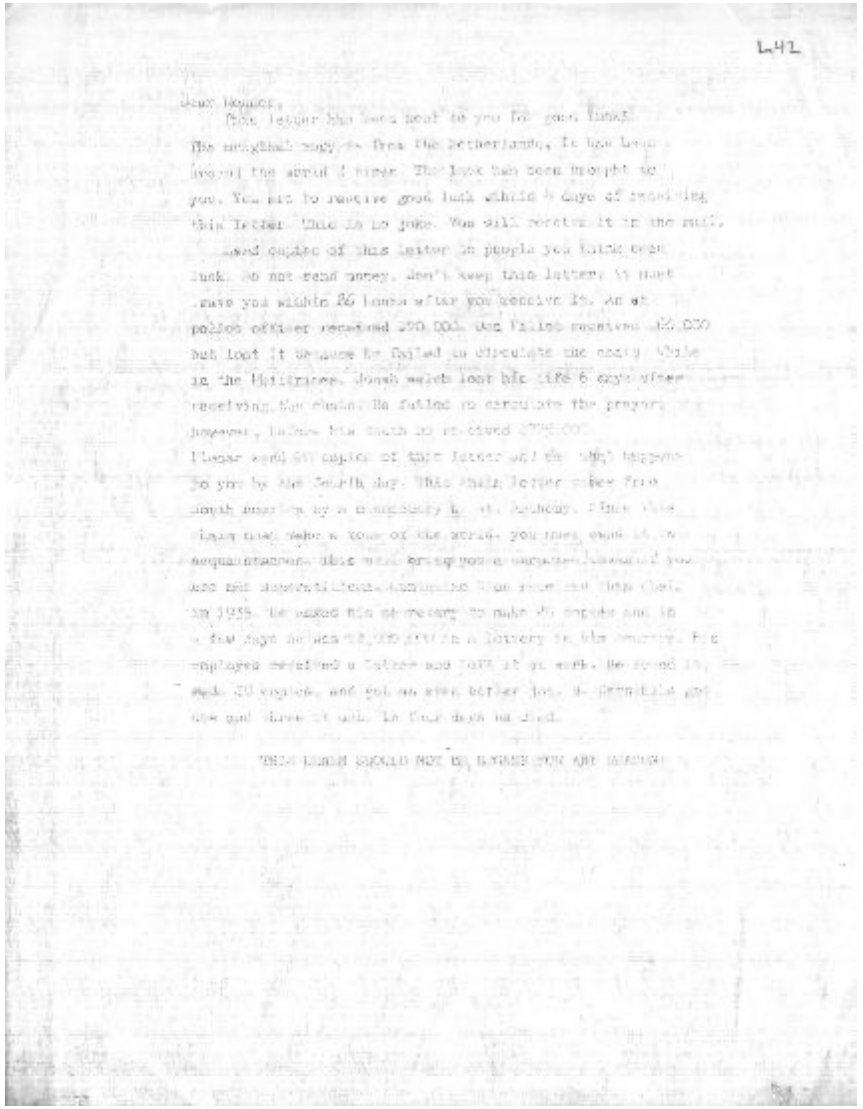
Chain letters – old style

- These letters are different but appear to have the same origin.
- We were interested in reconstructing the evolutionary history of these chain letters.
- Because these chain letters are readable, they provide a perfect tool for classroom teaching of phylogeny methods and test for such methods.
- *Scientific American*: Jun. 2003

C. Bennett, M. Li, B. Ma: Chain Letters & Evolutionary Histories



*An unclear letter reveals evolutionary path: ((copy)*mutate)**



Why bother with chain letters?

http://www.silcom.com/~barnowl/chain-letter/evolution.html

- Like a virus, it has reached billions of people, literally.
- Like a gene, they are about 2000 characters;
- It even resembles some subtle phenomenon in biological evolution!

cause he broke the chain. While in the Philippines, Gene Walsh lost his wife six days after receiving the letter. He failed to circulate the letter. However, before her death he received $7,755,000. Please

WITH LOVE ALL THINGS ARE POSSIBLE

This paper has been sent to you for good luck. The original copy is in New England. It has been around the world nine times. The luck has now been sent to you. You will receive good luck within four days of receiving this letter, providing, you in turn send it on. This is no joke. You will receive it in the mail. Send copies to people you think need good luck. Don't send money as fate has no price. Do not keep this letter. It must leave your hands within 96 hours. An RAF officer received $70,000. Joe Elliot received $40,000 and lost it because he broke the chain. While in the Philippines, Gene Walsh lost his wife six days after receiving the letter. He failed to circulate the letter. However, before her death he received $7,755,000. Please send 20 copies of the letter and see what happens in four days. The chain comes from Venezuela and was written by Saul Anthony Decroup, a missionary from South America. Since the copy must make a tour around the world, you must make 20 copies and send them to friends and associates. After a few days you will get a surprise. This is true even if you aren't superstitious. Do note the following: Constantion Dias received the chain in 1953. He asked his secretary to make 20 copies and send them out. A few days later he won the lottery of two million dollars. Carle Dadditt, an office employee, received the letter and forgot it had to leave his hands within 96 hours. He lost his job. Later, after finding the letter again, he mailed out the 20 copies. A few days later he got a better job. Dalan Fairchild received the letter and not believing, threw the letter away. Nine days later he died. Remember, send no money, and please don't ignore this.

IT WORKS

Coevolution

Life → wife

His → her

http://en.wikipedia.org/wiki/Coevolution

Methods for constructing phylogeny

- Charater Based Method
  - 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.

Character Based Method

- The first category of phylogeny methods do three things:
  - Define characters/features for each taxon.
  - Define a score function for each tree based on the characters.
  - Find the optimal tree

Basics

- A **character** is a "feature" in the species.
  - Vertebrate / invertebrate
  - Has hooves / does not.
  - A letter in multiple sequence alignment.
  - The title is "Trust in lord …" or "With love all things are possible".
- An **evolutionary tree** is a rooted and leaf-labeled binary tree.

More details

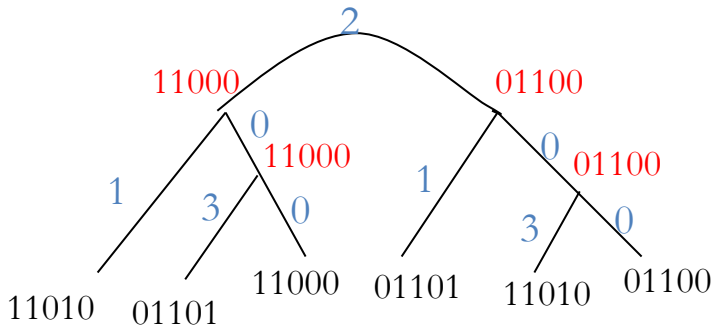
- Input: An *n* x *m* matrix of aligned characters. (*n* taxa, *m* characters)
  In the case of a multiple alignment, these are columns with no gaps.
- Output: A labeled tree with least number of mutations.

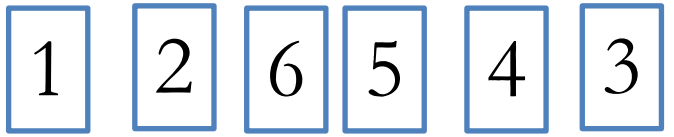
Characters (features)

taxa

	a	b	c	d	e
1	1	1	0	1	0
2	0	1	1	0	1
3	0	1	1	0	0
4	1	1	0	1	0
5	0	1	1	0	1
6	1	1	0	0	0

Value of character e of taxon 2.

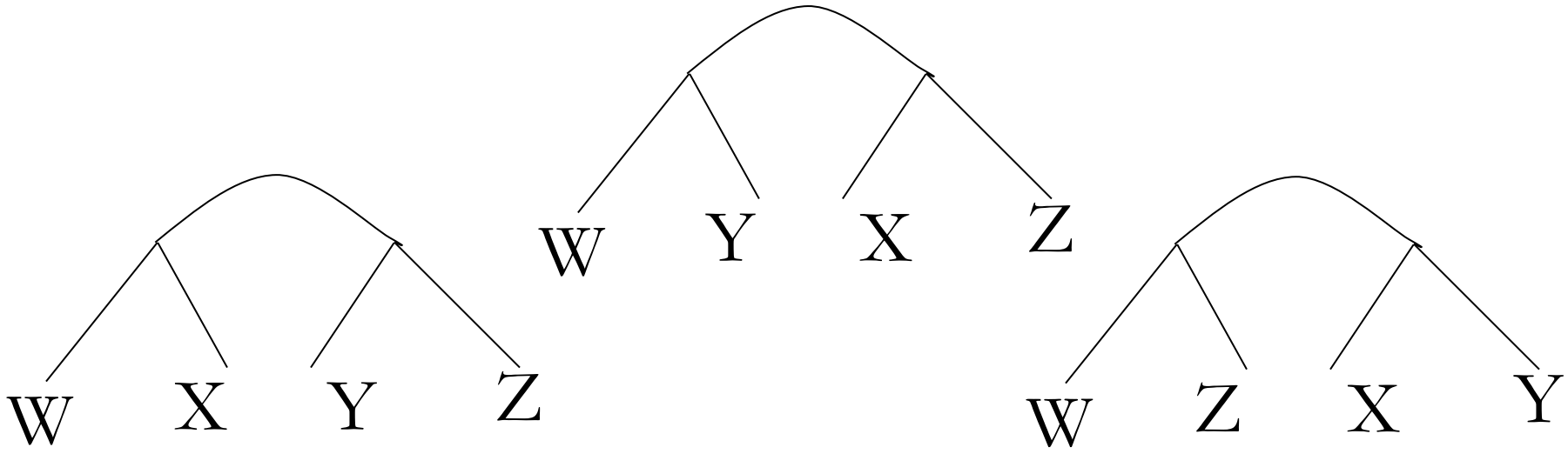




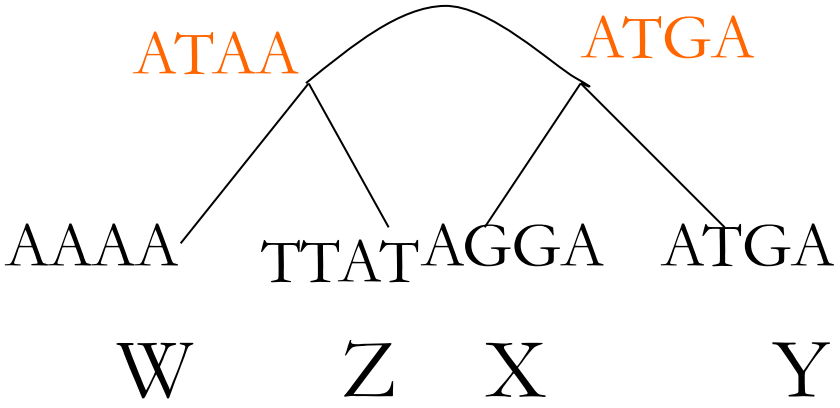
One solution (not optimal)

Example:

- Suppose we have four taxa:
- `W: AAAA`
- `X: AGGA`
- `Y: ATGA`
- `Z: TTAT`
- There're only 3 possible (unrooted) trees on 4 taxa. Which has the least mutations?



Parsimony example



- In this case, we need 5 mutations. The other 2 require 6.
- So the "cheapest" tree joins W and Z on one side and X and Y on the other.
- Where is the root of the tree?

Ancestor Reconstruction

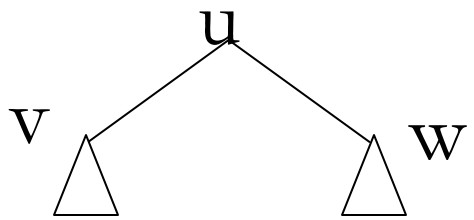
- For a given topology, how to construct the ancestors? (in order to calculate the score of the tree)
- First observation: We can solve each column separately. So we can just solve for 1-character strings.

- Algorithm by Sankoff: tree-based dynamic programming.

Tree DP, details

- For every node u of the tree and letter a of the alphabet Σ, let D[u, a] = min # of mutations in T_u if u's label is a.
- Let r be the root. We want $\min_x$ D[r,x].
- For a leaf node v, if the character at leaf v is a, then D[v,a] = 0, and D(v, b) = 1 for all other letters b.
- For an internal node u, with children v and w, suppose we know all of the values of D[v, *] and D[w,*].
- How to compute D[u,*]?

Tree DP details (end)



- If we put letter "a" at node u, the cost of the left branch of the tree is the minimum of
  - Case 1. D[v,a]
  - Case 2. $1 + \min_{b \neq a} D[v,b]$
- The same argument holds for the right branch. So
- $D[u,a] = \min (D[v,a], 1 + \min_{b \neq a} D[v,b]) +$
  $\min (D[w,a], 1 + \min_{b \neq a} D[w,b])$.
- Order of computation?
- Time complexity?

Total runtime

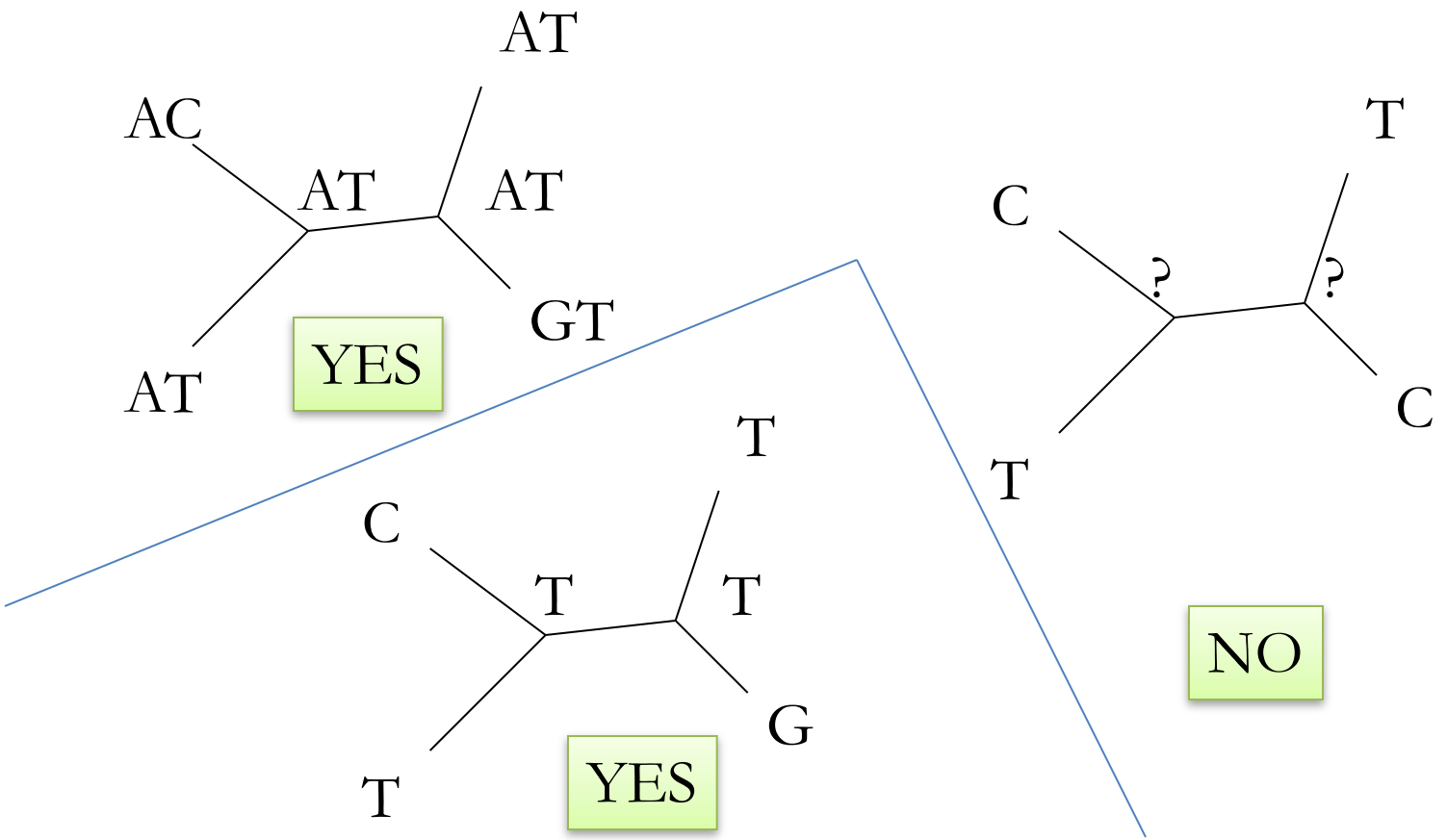
- We can ignore b≠a and minimize on all b without changing the value.
- Note: $\min_b D[v,b]$ only needs to be computed once, not once every letter a for $\min_{b \neq a} D[v,b]$.
- If the tree is binary, and the size of the alphabet is σ, this algorithm takes $O(n\sigma)$ time, since it's just $O(\sigma)$ time at each node

Parsimony

- Parsimony Method: Find a tree topology so that the total number of mutations on the edges is the smallest.
  - NP-hard.
  - Algorithm: For each possible tree topology, uses DP to compute cost. Output the best tree.
- Suppose there are $f(n)$ trees on n taxa.
- Total runtime: $O(nm\sigma f(n))$.
- Unfortunately, $f(n) = 1*3*\ldots*(2n-5)$. (Roughly $n!2^n$ or so).

Perfect Phylogeny

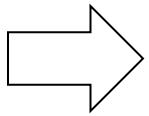
- A perfect phylogeny is such that for every character (every column), all species with the same state of that character is a connected component on the tree.

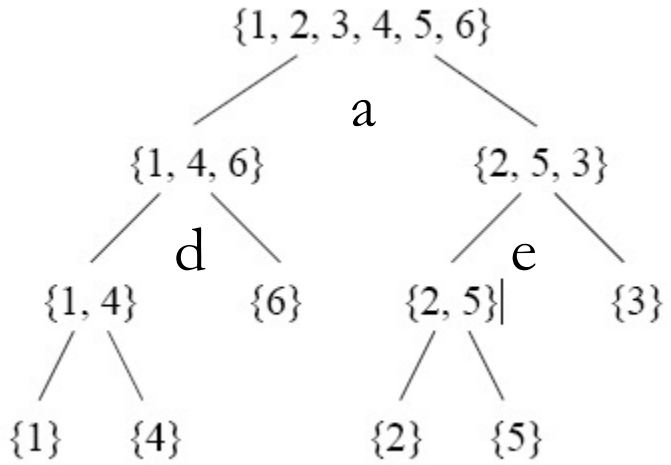


Algorithm for Binary Case

- Algorithm 1: Start with a set of all taxa. Find a character and split the set into two. Recursion until each set has only one taxon.

	a	b	c	d	e
1	1	1	0	1	0
2	0	1	1	0	1
3	0	1	1	0	0
4	1	1	0	1	0
5	0	1	1	0	1
6	1	1	0	0	0





Perfect Phylogeny

- For binary characters, Algorithm 1 is a polynomial time algorithm. If there is a perfect phylogeny, it outputs the perfect phylogeny.
  - Equivalently, if the output is not a perfect phylogeny, then there is no perfect phylogeny for the input.
- Theorem: If there is a perfect phylogeny for the input, and there are constant number of states for the characters, then a perfect phylogeny can be computed in polynomial time.
- r states, n taxa, m characters: $O(2^{2r} nm^2)$.

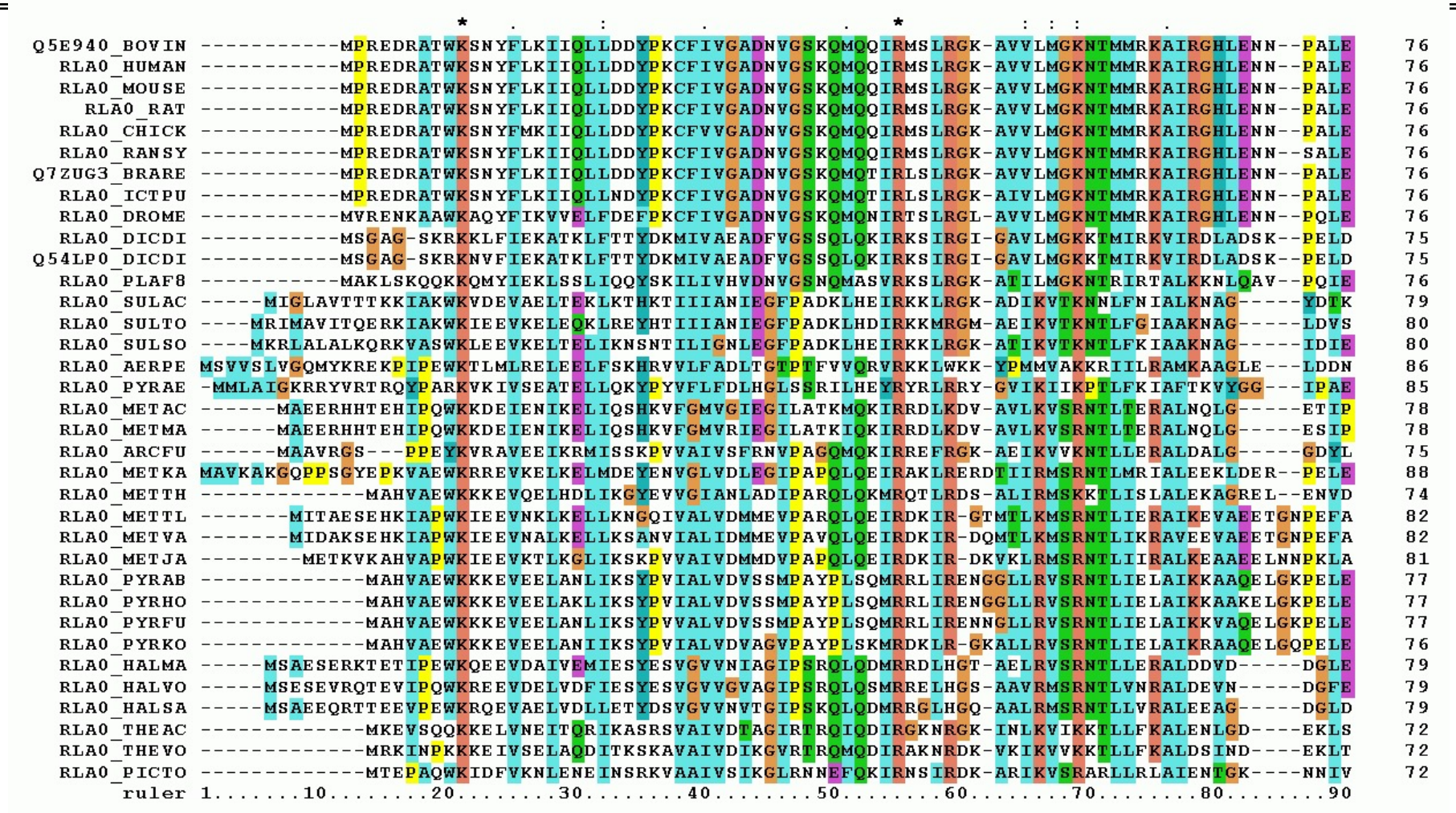
Important Fact

- If a column has k different states, then
  - any phylogeny requires at least k-1 mutations for the column.
  - a perfect phylogeny only has k-1 mutations for the column.
- Conclusion: a perfect phylogeny is the best you can get for parsimony.
- Not all input matrix can cause a perfect phylogeny.

Maximum Likelihood

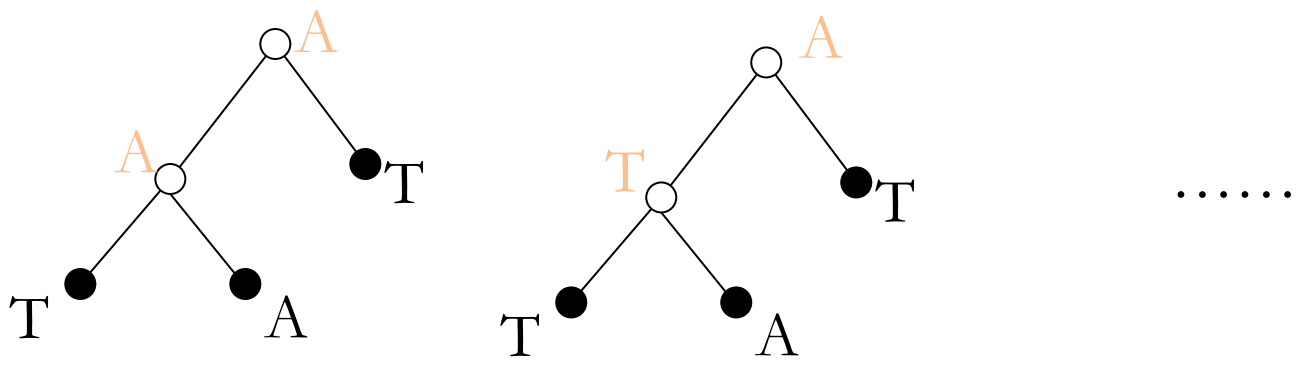
- The score function used in parsimony (and perfect phylogeny) is too simple, especially when sequencing data become available.
- For example: the multiple alignment of proteins can be used as the input
  - Thousands of columns (characters).
  - Mutations between different pairs of amino acids have different rates.
  - Different substitution matrices on different columns.
- The maximum likelihood method aims to provide a better scoring function.

A Multiple Alignment



Max likelihood method starts with a multiple alignment. Different columns may have different substitution freqeuncy matrix.

Maximum Likelihood



- For each possible tree topology T, for each possible internal node assignment, and calculate the probability based on the substitution matrix.
- For each tree T, add up all probabilities of all possible internal nodes. This is the likelihood of the input tree T. Figure shows a single column of the multiple alignment.
- Find the tree that maximizes the likelihood.

More Notes about Maximum Likelihood

- Often more accurate than other methods.
- Very time consuming. Usually heuristic algorithms and dynamic programming algorithms are used to assist the search and estimation of the likelihood.
- If desired, one can also allow the change of the edge length (the mutation rate at each edge).
- Software available: e.g. PhyML.

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.

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.

$$\sum | d(i,j) - T(i,j) |^2$$

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.
- 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.

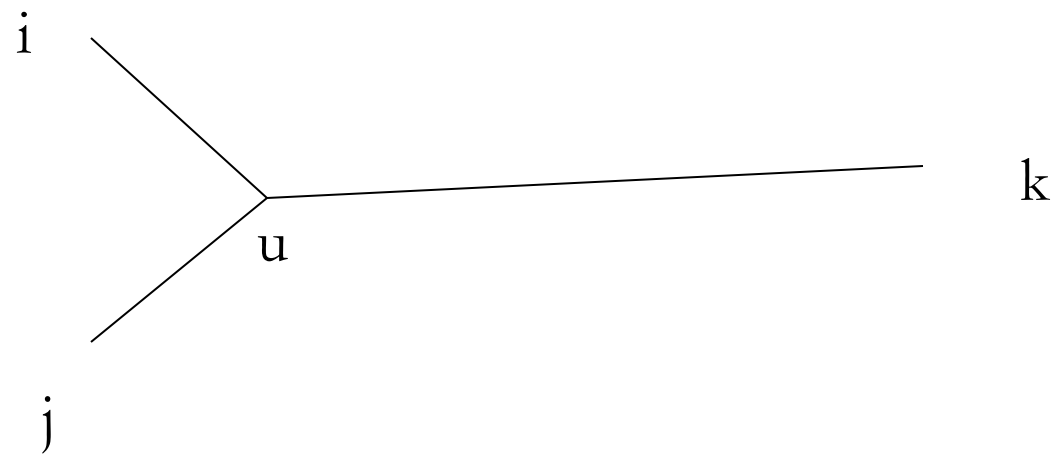
Neighbor Joining

- 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.

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?
- A "universal" solution -- information distance.

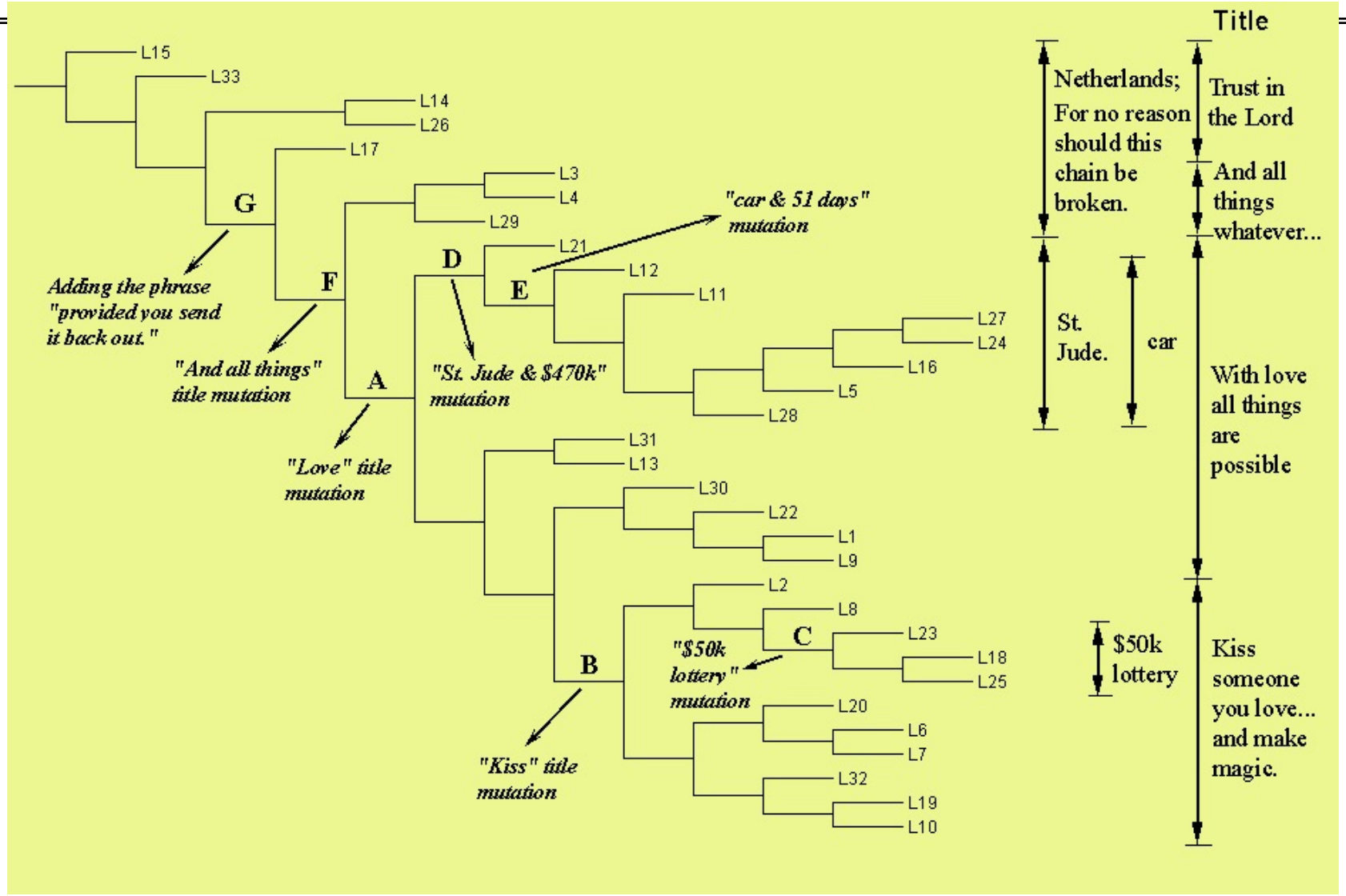
Information Distance

- 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.

Reconstructing History of Chain Letters

- 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