Phylogeny
Plan:

• Phylogeny is an important subject. We have 2.5 hours.

• So I will teach all the concepts via one example of a chain letter evolution. The concepts we will discuss include:
  • Evolutionary trees, characters.
  • Perfect phylogeny
  • Methods: NJ, parsimony, max likelihood, Quartet method
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.
A sample letter:

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 no joke. 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 Venezuela and was written by Sol Anthony De Cedif, 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 acquaintances. 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 Cruduit, 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 Koirichild 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.
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
A faded-out letter reveals evolutionary path:
((copy)*mutate)*
Why bother with chain letters?


• 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

http://en.wikipedia.org/wiki/Coevolution
Using a collection of over 600 dated paper chain letters, we have identified types and variations that appear and disappear over the years. Unexpectedly, it was discovered that, repeatedly, a single letter bearing some new innovation will propagate so abundantly and rapidly that within just a few years its descendants replace all similarly motivated letters in our collection.

“CHAIN LETTER EVOLUTION” by Daniel W. VanArsdale
Methods for constructing phylogeny

• 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.
• (Solving Chain Letter)
Methods for constructing phylogeny

• 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.
Basis

• A taxon is a group of one (or more) populations of organism(s), which a taxonomist adjudges to be a unit.
  • E.g. human, primate, mammal …
  • A taxon can be assigned a rank, usually (but not necessarily) when it is given a formal name.

The hierarchy of biological classification's eight major taxonomic ranks
Basics

• **A character** is a “feature” in the species.
  • Vertebrate / invertebrate
  • Has hooves / does not.
  • A DNA mutation.
  • 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

• The form we’ll have:

• Input: An $n \times m$ matrix of aligned characters. ($n$ taxa, $m$ characters)

In the case of a multiple alignment, these are columns with no gaps.
Example:

- Suppose we have four taxa:
  - W: AAAA
  - X: AGGA
  - Y: ATGA
  - Z: TTAT

- 3 trees on 4 taxa:

- Which is most parsimonious?
Parsimony example

• In this case, we need 5 mutations. The other 2 require 6.
• So the most parsimonious 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?

• 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 $\Sigma$, let $D[u, a] = \min \# \text{ of mutations in } T_u \text{ if } u\text{'s label is } a$.

• 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 $D[v, \ast]$ and $D[w, \ast]$.

• How to compute $D[u, \ast]$?
• If we put letter “a” at node u, the cost of the “left” half 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 half. So
• $D[u,a] = \min (D[v,a], 1 + \min_{b \neq a} D[v,b]) + \min (D[w,a], 1 + \min_{b \neq a} D[w,b])$.
• Time complexity?
Total runtime

• We can ignore $b \neq a$ 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 $\sigma$, this algorithm takes $O(n\sigma)$ time, since it’s just $O(\sigma)$ time at each node.

• What if mutations have a cost, so going from $a \rightarrow b$ costs $f(a,b)$? (Assume always non-negative…)

• That’s easy:
  • $D[u,a] = \min_b (f(b,a) + D[v,b]) + \min_b (f(b,a) + D[w,b])$.
  • Same argument as before, but now it takes $O(\sigma^2 n)$ time.
**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, $f(n)$ is huge: it’s $1*3*...*(2n-5)$. (Roughly $n!2^n$ or so) Parsimony is NP hard.
Methods for constructing phylogeny

• 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.
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

- Start with a set of all taxa. Find a character and split the set into two. Recursion until each set has only one taxonomy.
Perfect Phylogeny

• 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

- Perfect phylogeny is the best you can get for parsimony.
- If a column has k different states, then it only contributes cost k-1 in a perfect phylogeny tree.
Two practical techniques: Branch-and-bound and local search.

These are two common techniques for many NP-hard problems. So remember the idea.

Just like dynamic programming can be used in many problems too.
Branch-and-bound approach

- Idea by example: We exhaustively search all the trees by adding one taxon a time.
- Suppose we are working on 7 taxa (A…G), and we already know this tree exists:

```
2
1
2
1
2
1
2
4
```
(Total cost: 25)

- Then if a new tree structure on only A…E that requires at least 28 mutations for those 5 taxa, adding F and G anywhere is a waste of time, since things only get worse. This allows us to avoid looking at all trees.
Local improvement method

• Start with one reasonably decent tree.
• Create a graph on trees, where two trees are adjacent if a simple local change from one turns it into the other.
• (Examples: pull one taxon out of one tree and paste it into all possible other places. Or unroot one subtree and place it anywhere on what remains.)
• Keep doing this until no improvement can be achieved.
• This is not guaranteed to converge to a global optimum, just a local one. But it is used often.
Methods for constructing phylogeny

• 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.
Given a multiple sequence alignment, maximum likelihood method asks for a tree topology, with edge weights, such that the likelihood is maximized. NP-hard.

\[
\Pr(\text{Labels} \mid \text{Tree}) = 0.2 \times (1-0.1) \times 0.2 \times (1-0.1) = 0.0324
\]

\[
\Pr(\text{Labels} \mid \text{Tree}) = (1-0.2) \times 0.1 \times (1-0.2) \times (1-0.1) = 0.0576
\]
More Notes about Maximum Likelihood

• Multiplying likelihoods of all positions (bases or amino acids) is needed to estimate edge length.

• Instead of a uniform evolutionary rate for the whole length of the sequence, different sites may have different evolutionary rates.
  • More elaborated models are needed.

• Instead of using the most probable ancestor sequences, may use sum of all possible ancestor sequences.

• Need to enumerate all possible tree topologies to find the best one.
  • Very time consuming.
Distance Based Methods

- First compute the distances between pairs of taxa.
- So we have a $n \times n$ 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 $\sum |d(i,j) - T(i, j)|^2$
- This is also NP-hard.
Methods for constructing phylogeny

• 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.
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


- 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.
- More details can be found in Durbin’s book.
Neighbor Joining Idea
Methods for constructing phylogeny

• 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))
Methods for constructing phylogeny

- 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
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 $K(y \mid x)$. K is called the Kolmogorov Complexity.
• The information distance between x and y is defined as

$$d(x,y) = \max \{ K(x \mid y), K(y \mid x) \} / \max \{ K(x), K(y) \}.$$
• Kolmogorov complexity is incomputable, but we can use compression in practice.
It has been a disputed issue which of the two groups of placental mammals are closer: Primates, Ferungulates, Rodents.
Who is our closer relative?

Rodents  Ferungulates  Primates

Or

Rodents  Primates  Ferungulates
Eutherian Orders:

- It has been a disputed issue which of the two groups of placental mammals are closer: Primates, Ferungulates, Rodents.
- In mtDNA, 6 proteins say primates closer to ferungulates; 6 proteins say primates closer to rodents.
- Hasegawa’s group concatenated 12 mtDNA proteins from: rat, house mouse, grey seal, harbor seal, cat, white rhino, horse, finback whale, blue whale, cow, gibbon, gorilla, human, chimpanzee, pygmy chimpanzee, orangutan, sumatran orangutan, with opossum, wallaroo, platypus as out group, 1998, using max likelihood method in MOLPHY.
Li et al. used complete mtDNA genome of exactly the same species.
By using a DNA compression program to approximate the Kolmogorov complexity, information distance $d(x,y)$ can be computed.
They computed $d(x,y)$ for each pair of species, and used Neighbor Joining in Molphy package (and their own hypercleaning).
They constructed exactly the same tree. Confirming Primates and Ferungulates are closer than Rodents.
Evolutionary Tree of Mammals:
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 / Quartet phylogeny programs 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.
Confirmed by VanArsdale’s study, answers an open question
Summary

• A universal method
  • I did not know the “characters” – they are only inferred after the tree.
  • I did not have an evolutionary model.

• Phylogenetic methods:
  • Parsimony
  • Max likelihood
  • UPGMA
  • NJ
  • Quartet
  • (Whole genome phylogeny)