Phylogenetics

Bioinformatics Workshop

Jessica Kissinger
Trinidad
2010
Why do Phylogenetics?

• We make evolutionary assumptions in our everyday research life. For example, we need a drug that will kill the parasite and not us. Thus, we need a target that is present in the parasite and not us.

• We need a good model system, Which parasite (or host) is most closely related to *P. falciparum* or Humans?
Why Phylogenetics?

• This strain is resistant to drug and this one is sensitive, what has changed?

• Where did this parasite come from? Has it “co-evolved” with humans? Did it enter the human lineage from another source?

• Which other mosquitoes are likely to serve as a host for my parasite in nature?
Phylogenetics

• What is Phylogenetics?
  – Molecular Systematics
    • The use of molecular data to infer the relationships of the host species e.g. using rRNA to build trees to look at the relationship of the bacteria to the eukaryotes
  – Molecular Evolution
    • Use trees to infer how a molecule, protein, or gene has evolved (insertions, deletions, substitutions).
Gene Trees vs Species Trees
You Can Make Phylogenies of Many Things:

• Amino acid sequences
• Nucleotide sequences
• RFLP data
• Morphological data
• “Paper fastening devices”
Issues you had to deal with

1) Conflict - Size, color, material, shape
2) Direction of change, e.g. red to green?
3) Homology - these items have a similar function but do they have a similar origin?
4) Mixed materials - plastic coated metal
5) How do you assign weight, are some traits more important?
6) Lots of possibilities
   >8,2000,794,532,637,891,559,375 rooted trees!
Goals for this lecture

• Become familiar with concepts
• Become familiar with vocabulary
• Become familiar with the data analysis flow
• Reach the point where you can read the available literature on how to use these methods in greater detail
Assumptions made by Phylogenetic algorithms

- The sequences are correct
- The sequence are homologous
- Each position is homologous
- The sampling of taxa or genes is sufficient to resolve the problem of interest
- Sequence variation is representative of the broader group of interest
- Sequence variation contains sufficient phylogenetic signal (as opposed to noise) to resolve the problem of interest
- Each position in the sequence evolved independently
Availability of Sequenced Genomes

- Bacteria 74
  - Proteobacteria
  - Cyanobacteria
  - Flavobacteria
  - Thermotoga
  - Thermodesulfbacterium
  - Aquifex
  - Spirochetes
  - Gram+ bacteria
  - Green nonsulfur bacteria

- Archaea 16
  - Euryarcheota
    - Crenarcheota
      - Euryarcheota

- Eucarya 14
  - Animals
  - Fungi
  - Plants
  - Slime molds
  - Flagellates
  - Microsporidia
  - Giardia

Courtesy of Igor Zhulin
Apicomplexans

Giardia lamblia

Varimorpha necatrix

Trichomonas vaginallis

Trichomonas foetus

Physarum polycephalum

ENTAMOEBAE

D. histolytica

Entamoeba invadens

Naegleria gruberi

Bodonids

Kinetoplastids

Euglenoids

Physarum polycephalum

Trichomonas foetus

Trichomonas vaginallis

Cnidaria

ENTAMOEBAE

D. histolytica

Entamoeba invadens

Naegleria gruberi

Bodonids

Kinetoplastids

Euglenoids

Physarum polycephalum

Trichomonas foetus

Trichomonas vaginallis

Giardia lamblia

Varimorpha necatrix

adapted from Sogin et al (1991)
Circumsporozoite Phylogeny
(molecular systematics, host relationships)
How to do an analysis

- Define a question
- Select sequences appropriate to answer your question (not all sequences are equally good!)
- Make a multiple sequence alignment
- Edit your alignment to make it better
- Perform lots and lots of analyses
- Perform Bootstrap analyses to test confidence
Multiple Sequence Alignment
### Multiple Sequence Alignment

| 1  | pfacsnpkn | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 2  | pfacsnpnr | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 3  | psu09765ps | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 4  | pfacsapviv | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 5  | pfacsapvii | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 6  | pfacsapcy | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 7  | pfacsapcy | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 8  | pfacsapgly | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 9  | psu09766pm | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 10 | pfacsapfa | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 11 | pfacsapfa | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 12 | pfacsapkg | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 13 | pfacsapkg | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 14 | pfacsapkg | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 15 | pfacsapkg | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |
| 16 | pfacsapkg | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - | -QPAQGQDSGANAAG - |

| ruler | 0 | 120 | 130 | 140 | 150 | 160 | 170 | 180 | 190 |

---

*Note: The table above represents a multiple sequence alignment, where each row corresponds to a different sequence, and the columns represent aligned positions across the sequences.*
Study your Alignments!

**Region I**

- P. falciparum1: KPKHK--KLKQPCGDPNVPDNANPNVDPNANPNV
- P. falciparum2: KPKHK--KLKQPCGDPNVPDNANPNVDPNANPNV
- P. reichenowi: KPKHN--KLKQPCGDNVPDNANPNDPANPNDPANPND
- P. gallinaceum: YFRENVYNLNQPVGGNGGVPQAGGNGQVQAGGNGGV
- P. malariae1: KAVEN--KLKQPDGDDGAGNDANGNDAGNAAGNA
- P. malariae2: KAVEN--KLKQPDGDDGAGNDANGNDAGNAAGNA
- P. vivax1: NPREN--KLKQPCGDRADGPQAPAGDRADGQPA
- P. vivax2: NPREN--KLKQPCGDRADGPQAPAGDRADGQPA
- P. simium: NPREN--KLKQPCGDRADGPQAPAGDRADGQPA
- P. cynomolgi1: KPREN--KLKQPCGNNAAAGEAGNNAAAGEAGNNAAAGE
- P. cynomolgi2: KPREN--KLKQPCGNNAAAGEAGNNAAAGEAGNNAAAGE
- P. simiovale: KPHEN--KLKQPGANQEGGAAAPGANGGAAAP
- P. knowlesi1: KPNEN--KLKQPEGQPGAQGQGGAAGQGQGQGAAGQGQGGAAGQGQGQGAAGQGQGGAAG
- P. knowlesi2: KPNEN--KLKQPEGQAPAAGGAAPQAPAAGGAAPQAPAAGGAAPQAPAAGGAAP
- P. berghei1: IERNN--KLQKPPPCPPNPNDPPNPNDPPNPNDPPNPND
- P. berghei2: IERNN--KLQKPPPCPPNPNDPPNPNDPPNPNDPPNPND
- P. yoelii1: KEAQN--KLQKPVQADNVPQVGAPQPGAPQPGAP
- P. yoelii2: KEAQN--KLQKPVQADNVPQVGAPQPGAPQPGAP


A Word About Methods

- There are two overall categories of methods
  - **Transformed distance methods** (data are transformed into a distance matrix). The matrix is used to build a single tree. UPGMA and Neighbor-Joining are examples of this method. They are computationally simple and very fast.
  - **Optimality methods** (tree generation is separate from tree evaluation). Parsimony and Maximum-likelihood methods divorce the issue of tree generation from evaluating how good a tree is. For parsimony, there may be more than 1 “most parsimonious” or “shortest” tree found.
Distance methods

- **UPGMA**
  - Assume all lineages evolve at the same rate
  - Produces a root
  - Produces only one tree
  - Computationally very fast
  - Trees are additive

- **Neighbor-joining**
  - Permits variation in rates of evolution
  - Does not produce a root
  - Produces only one tree
  - Computationally very fast
  - Trees are additive
$1 \text{ ATTGCTCAGA}$

$2 \text{ AATGCTCTGA}$

$3 \text{ ATAGGACTGTA}$

1 vs 2 = 80% similar = 0.2 distance

1 vs 3 = 60% similar = 0.4 distance

2 vs 3 = 60% similar = 0.4 distance

Create a distance matrix

Can use scoring schemes to transform data into distances (e.g. do transitions occur more often than transversions)
The implementation of the UPGMA algorithm to produce the tree below. A new matrix is calculated at each iteration.

**Figure 29** Cluster analysis (UPGMA) of 5S rRNA evo-
An unrooted Neighbor-joining tree of the same dataset
Models of evolution: choosing parameters

Factors that Affect Phylogenetic Inference

1. Relative base frequencies (A,G,T,C)
2. Transition/transversion ratio
3. Number of substitutions per site
4. Number of nucleotides (or amino acids) in sequence
5. Different rates in different parts of the molecule
6. Synonymous/non-synonymous substitution ratio
7. Substitutions that are uninformative or obfuscatory
   1. Parallel substitutions
   2. Convergent substitutions
   3. Back substitutions
   4. Coincidental substitutions

In general, the more factors that are accounted for by the model (i.e., more parameters), the larger the error of estimation. It is often best to use fewer parameters by choosing the simpler model.
Some distance models: p-distance

• $p = \frac{n_d}{n}$, where $n$ is the number of sites (nucleotides or amino acids), and $n_d$ is the number of differences between the two sequences examined.
• Very robust when divergence times are recent and the affect of complicating phenomena is minor.
Some distance models: Jukes-Cantor

- Used to estimate the number of substitutions per site
- The expected number of substitutions per site is:
  \[ d = 3\alpha t = -(3/4)\ln[1-(4/3)p] \]
  where \( p \) is the proportion of difference between 2 sequences
- Variance can be calculated
- No assumptions are made about nucleotide frequencies, or differential substitution rates
Some distance models: Kimura two-parameter

- Used to estimate the number of substitutions per site
- $d = 2rt$, where $r$ is the substitution rate (per site, per year) and $t$ is the generation time; $r = \alpha + 2\beta$, so:
  - $d = 2\alpha t + 4\beta t$
- Accounts for different transition and transversion rates
- No assumptions are made about nucleotide frequencies, variance is greater than Jukes-Cantor

$\alpha = \text{transition rate}$
$\beta = \text{transversion rate}$
These are treated the same for long divergence times.
Other models

- Hasegawa, Kishino, Yano (HKY): corrects for unequal nucleotide frequencies and transition/transversion bias into account
- Unrestricted model: allows different rates between all pairs of nucleotides
- General Time Reversible model: allows different rates between all pairs of nucleotides and corrects for unequal nucleotide frequencies
- Many other models have been invented to correct for specific problems
- The more parameters are introduced, the larger the variance becomes
Optimality Methods

- All possible trees (or a heuristic sampling of trees) are generated and evaluated according to Parsimony or Maximum likelihood.
- Note: Tree generation is divorced from tree evaluation. More than one tree topology may be optimal according to your criteria.
## General differences between optimality criteria

<table>
<thead>
<tr>
<th>Minimum evolution</th>
<th>Maximum Parsimony</th>
<th>Maximum Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model based</td>
<td>“Model free”</td>
<td>Model based</td>
</tr>
<tr>
<td>Can account for many types of sequence substitutions</td>
<td>Assumes that all substitutions are equal</td>
<td>Can account for many types of sequence substitutions</td>
</tr>
<tr>
<td>Works well with strong or weak sequence similarity</td>
<td>Works only when sequence similarity is high</td>
<td>Works well with strong or weak sequence similarity</td>
</tr>
<tr>
<td>Computationally fast</td>
<td>Computationally fast</td>
<td>Computationally slow</td>
</tr>
<tr>
<td>Well understood statistical properties (easy to test)</td>
<td>Poorly understood statistical properties (hard to test)</td>
<td>Well understood statistical properties (easy to test)</td>
</tr>
<tr>
<td>Can accurately estimate branch lengths (important for molecular clocks)</td>
<td>Cannot estimate branch lengths accurately</td>
<td>Can estimate branch lengths with some degree of accuracy</td>
</tr>
</tbody>
</table>
A definite Beginning and Polarity, a root
In the world of trees, there are more rooted topologies for a given Number of taxa than unrooted.
Possible trees as function of number of Taxa

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Rooted Trees</th>
<th>Unrooted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>34,459,425</td>
<td>2,027,025</td>
</tr>
<tr>
<td>100</td>
<td>2 x 10^{182}</td>
<td></td>
</tr>
</tbody>
</table>

More trees than the number of atoms in the universe!
Tree search considerations

• Exhaustive searches are searches of all possible trees for the number of Taxa in your data set (15 Taxa or less)

• If you have more than 15 Taxa, then heuristic methods must be employed in which you search a sample of all possible trees. There are many algorithms for the generation of different populations of trees.
## Tree search considerations

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Stepwise addition</td>
<td>Algorithmic</td>
</tr>
<tr>
<td>• Star decomposition</td>
<td>Algorithmic</td>
</tr>
<tr>
<td>• <strong>Exhaustive</strong></td>
<td><strong>Exact</strong></td>
</tr>
<tr>
<td>• Branch &amp; bound</td>
<td>Exact</td>
</tr>
<tr>
<td>• <strong>Branch swapping</strong></td>
<td><strong>Heuristic</strong></td>
</tr>
<tr>
<td>• Genetic algorithm</td>
<td>Heuristic</td>
</tr>
<tr>
<td>• Markov Chain Monte Carlo</td>
<td>Heuristic</td>
</tr>
</tbody>
</table>
Parsimony basics & scores

• Based on shared derived characters (synapomorphies)
• Identical characters which evolve more than once are “homoplasies”
• Unique characters are “autapomorphies”
• The score of the tree is the total of all the changes needed to map the data. The scale bar is #of changes.
• Smaller, i.e. more parsimonious scores are better
• More than one tree topology may have the same score
An informative position is one that can favor one tree over another when some type of criteria are applied.

Position #5 is informative, it permits us to choose a shorter tree from among the options. It prefers the tree of length 1 over those of length 2.
Not all alignment positions can help pick a better tree

None of these characters can distinguish between the three possible unrooted topologies. They are uninformative.
Maximum Likelihood

- Is an optimality method, it is an algorithm which evaluates trees according to some criterion
- The algorithm searches for trees which maximize the probability of observing the data
- Trees are scored with Log likelihoods
- This is the most computationally intensive method available
- More tractable versions include (puzzle)
- Alternate approaches include Bayesian inference (Mr. Bayes)
Not all methods can be used with all types of data

- Parsimony can be used with all types of data, nucleotide, protein, binary, morphological, mixed data sets. States can be ordered.
- Distance can be used with nucleotide and protein data but you need a model to generate distances.
- Maximum likelihood, normally only nucleotide data, but PAML can do protein maximum likelihood (still a tricky and debatable approach).
- Bayesian - All types of sequence data
There are Many Types of Trees

• Cladogram vs. Phylogram
  – Cladograms have uniform branch lengths and only represent relationships
  – Phylograms have lengths proportional to change or distance

• Rooted vs. Unrooted
  – A defined origin as opposed to a network or relationships (most trees are unrooted because they are easier to calculate)

• Artistic license (slanted, rectangular, circle, “network”)
A Word about trees

A
B
C
D
E
F
G

A
B
C
D
E
F
G

A
B
C
D
E
F
G

A
B
C
D
E
F
G
A word about trees
(there are many types)

Slanted Cladogram

Unrooted Phylogram

Rectangular Phylogram
The Bootstrap

• The bootstrap is a method for assigning a measure of confidence to a particular node in tree.
• It is NOT a measure of the overall “goodness” of the tree.
• Rules of thumb: 70-100% = Good, 0-30% = bad, 30-70% = “gray zone” difficult to interpret.
The bootstrap process

<table>
<thead>
<tr>
<th>Original Data</th>
<th>Each column</th>
<th>Represented once</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>1ST sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 C C G A A A T G A</td>
</tr>
<tr>
<td>2 C C C G A C T G G</td>
</tr>
<tr>
<td>3 C C A G A T T A A</td>
</tr>
<tr>
<td>4 C C A G A G T A G</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2ND etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 9 6 2 1 3 4 8 7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3RD etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 3 3 1 6 5 7 4 9</td>
</tr>
</tbody>
</table>

Then build consensus of all trees produced by sample datasets. This provides support for nodes

100 or 1,000
A caution about alignments
characters in columns are homologous

<table>
<thead>
<tr>
<th>Task</th>
<th>Characters</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>beak</td>
<td>earSize</td>
<td>earShape</td>
<td>earPos</td>
<td>stance</td>
<td>bodyColor</td>
<td>Fingers</td>
<td>feet</td>
<td>Webbed</td>
<td>foreLimb</td>
<td>whitGlov</td>
<td>teeth</td>
<td>tail</td>
</tr>
<tr>
<td>1</td>
<td>stickman</td>
<td>0</td>
<td>1/3/4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Daffy</td>
<td>1</td>
<td>1/3/4/5/6/7</td>
<td>?</td>
<td>?</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Bugs</td>
<td>0</td>
<td>2/3/4/5/7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Wile E</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>RoadRunner</td>
<td>1</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Tweety</td>
<td>1</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Mickey</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Donald</td>
<td>1</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Pluto</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>Goofy</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
15 equally Parsimonious trees Of Disney characters. All trees have the same, smallest score.
Comparison of real trees

Assessment of support

Strict

Majority rule

Bootstrap

stickman

Daffy

RoadRunner

Tweety

Donald

Bugs

Goofy

Mickey

Wile E

Pluto

Daffy

RoadRunner

Tweety

Donald

Bugs

Goofy

Mickey

Wile E

Pluto

stickman

Daffy

RoadRunner

Tweety

Donald

Bugs

Goofy

Mickey

Wile E

Pluto

84

100

60

100

100

100

60

100

100

100

100
If 79% of the time this relationship holds, 29% it is something else.
Some points to consider for the paper fasteners:

We decided, in our evolutionary model that material was so important that we needed to give it extra weight, so we did (weight = 2).

Based on external information, such as the archeological record, we have learned that metal predates plastic, so, we ordered our characters: metal must precede plastic.

We decided to use as an “outgroup”, an unbent piece of metal, (taxon 21) to polarize the direction of evolution within our tree, i.e. we have evolved from a straight piece of metal into a “paper fastening device”. We will not allow reversion to this “unbent” state.

We will enforce the assumptions/decisions made above by using a constraint tree. By using this constraint tree, we reduce the number of possible rooted trees from $2.216431 \times 10^{20}$ to 273,922,023,375 and we reduce the number of unrooted trees from $6.332660 \times 10^{18}$ to 54,784,404,674 - a considerable savings!

We removed taxa 4 and 11 from the data set because they are non-homologous, i.e. the have a similar function but they do not share a common evolutionary descent or path. What we have here is a case of convergent evolution, i.e. independent origins of a paper fastening solution!
## Paper fastener Dataset

The dataset contains 21 taxa (paper fasteners). The table below contains 14 characters for each fastener. Each character (A – N) has multiple states, ranging from 2 states (0-1), to seven states (1-7).

<table>
<thead>
<tr>
<th>CLIP#</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3.</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5.</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6.</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7.</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8.</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9.</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10.</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12.</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13.</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14.</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15.</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16.</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17.</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18.</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19.</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20.</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21.</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Neighbor-joining analysis and bootstrap of clip dataset
Some of the >37,500 Trees generated by a Parsimony analysis of the clip dataset
Consensus of 5,000 parsimony Trees

Bootstrap of clips
Software and Books

• “How to make a phylogenetic Tree” by Barry Hall, comes with PAUP* CD, ~$30, Sinauer Press
• Phylip - Joe Felsenstein, Free via internet
• PAML - Free via internet
• Mr. Bayes - Free via internet
• ClustalW or ClustalX - Free via internet

* Best on a MAC, but also command line
Giving Credit

• Several slides in this presentation were provided by Mike Thomas, via a presentation he posted on the internet in 2002.