Lecture 22: 
Perfect Phylogeny

Not in textbook
• Thus far
  – distance-based evolutionary trees
    • Additive guarantees that the tree would reproduce all pairwise distances, but not all distance matrices are additive
    • Sequences $\rightarrow$ Distances $\not\rightarrow$ Sequences
  – character-based evolutionary trees
    • Trees directly from sequences
    • The most general version is hard (Large parsimony)

• Infinite Sites Model
• Perfect Phylogeny
• Local vs Global Phylogenetic Trees
• M has \( n \) rows (samples)
• M has \( m \) columns (characters)
• \( M_{ij} \) denotes the state object \( i \) has for character \( j \)
• Sequence Diversity Patterns (SDPs) often reoccur
Assumes mutations are rare events
Assumes DNA sequences are large
Multiple mutations at the same site are extremely rare
Infinite Sites Model assumes that multiple mutations never occur at the same sequence position
Thus, all states are “Binary” or “Biallelic”
A Different Kind of Tree

• Unrooted “Perfect Phylogeny” Tree
• Nodes correspond to sample sequences (haplotypes), both current and ancestral
• Edges correspond to actual mutations (SNPs)
• Removal of an edge creates a bipartition (each part is distinguished by a character at some position)
• SDPs can occur multiple times, and their frequency can be used as a edge weight
• Tree leaves correspond to mutations (allele variants) that are unique to a sequence, i.e. a SDP with only one minority allele instance, private
Unrooted Trees

- Unrooted phylogenetic trees are less specific than evolutionary trees
- The edges are undirected, thus the direction from ancestor to descendent are unknown
- All but one leaf, however, and possibly all leaves (if the root is an interior node) must be descendents
- Slightly fewer labeled unrooted trees than labeled rooted tree

\[
\begin{align*}
\text{uT}(n) &= \frac{(2n - 4)!}{2^{n-2}(n - 2)!} \quad \text{vs} \quad \text{T}(n) = \frac{(2n - 3)!}{2^{n-2}(n - 2)!}
\end{align*}
\]

- Moreover, any node can be an “observed” sample in a phylogenetic tree whereas only leaf nodes are observed an evolutionary tree
Three different evolutionary (rooted) trees that are consistent with a common phylogenetic (unrooted) tree.
Building a Phylogenetic Tree

- Assume we only have direct access to current haplotypes
- Construct a pair-wise distance matrix between haplotypes using Hamming distances
- Add smallest edge between all nodes which do not introduce a loop
- If the smallest distance is greater than 1 add d-1 “hidden” nodes between the pair so that adjacent nodes have a hamming distance of 1
- Augment the distance matrix with the new nodes and claim the introduced edges
- Repeat finding the smallest distance, and augmenting until the graph is connected

<table>
<thead>
<tr>
<th></th>
<th>S_1</th>
<th>S_2</th>
<th>S_3</th>
<th>S_4</th>
<th>S_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H_2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H_3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H_4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Four-Gamete Test

• Our tree construction method will not work for any arbitrary set of character sequences; it only works for those that satisfy the assumptions of the infinite sites model
• Under the assumption of the infinite sites model all SNP pairs exhibit the property no more that 3 out of the possible 4 allele combinations occur
• Direct consequence of only one mutation per site
• Showing that all SNP pair combinations satisfy the four gamete test is a necessary and sufficient condition for there to exist a perfect phylogeny tree

\[
\begin{array}{ccccc}
S_1 & S_2 & S_3 & S_4 & S_5 \\
H_1 & 1 & 1 & 0 & 0 & 0 \\
H_2 & 1 & 1 & 0 & 1 & 0 \\
H_3 & 0 & 0 & 0 & 0 & 1 \\
H_4 & 0 & 0 & 1 & 0 & 0 \\
\end{array}
\]
Questions

• Does there exist SDPs that are compatible with all others?

  Private SNPs are compatible with any other SNP

• Given N distinct haplotype sequences resulting from an infinite sites model what is minimum number of SDPs?

  N-1 edges are the fewest necessary to connect N haplotypes into a “linear” tree. How many singleton SNPs occur in such a tree? 2

• Given N distinct haplotype sequences resulting from an infinite sites model what is maximum number of SDPs?

  2N-3 edges, the number of edges in an unrooted tree with N leaves
Exercise

- Consider the following SNP panel

<table>
<thead>
<tr>
<th></th>
<th>S₁</th>
<th>S₂</th>
<th>S₃</th>
<th>S₄</th>
<th>S₅</th>
<th>S₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₁</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H₂</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H₃</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H₄</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H₅</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Satisfies the four gamete test?
- Construct the tree
- Is the SDP 11001ᵀ possible?
Complications

- There are two issues that limit the use of Perfect Phylogeny, both are violations of our infinite-sites model assumptions
  - In addition to mutations, haplotype diversity is generated by recombination, exchange of subsequences between haplotypes
  - Mutations reoccur at the same position (Homoplasy)
- Thus, global (over the entire genome) perfect phylogenies are rare, but local perfect phylogenies are common
- How do we locate recombinations and recurrent mutations?
Non-sequence Complications

• Evolutionary Convergence:
  – Wings on birds and bats
  – Fins on Seals and Fish

• Evolutionary Reversals:
  – Fish → Lizard → Snake
  – Fish → Mammal → Manatee
  – (gain and later loss of legs)

• Such paths also violate the infinite sites model
SNP Compatibility

- How do we find local genomic regions where our assumptions are valid?
- Apply 4-gamete test
- Issues
  - Can we efficiently find all compatibility intervals
  - How many intervals? (fewest necessary to cover the entire genome)
  - Unique?
  - Common properties
Algorithms

- Left-to-right scan
- Is this solution unique?
Algorithms

- Left-to-right scan
- Is this solution unique? No.
- Right-to-Left scan

- Given that the solution is not unique, which do we choose?
- The most parsimonious
• Questions
  • Of all scans, which has the fewest intervals?
  • Is there a solution with fewer intervals?
• What is a better solution?
  – Clearly the intervals could be larger
  – What is the maximal size of the intervals?
Theorem

- Left-to-right and right-to-left scans have the same number of intervals, \( k \)
- \( k \) is the minimum number of intervals possible
• The interval overlaps tell us something important
  • Pair the L-R and R-L scan intervals from left to right. The overlap of these pairs are the interval cores.
    - The $i^{th}$ core essentially is the SNPs that the $i^{th}$ interval of the L-R and R-L scan agree should be included in the $i^{th}$ interval of any minimal set of intervals

• A refinement of Parsimonious:
  - Use this to find the minimal set of maximally-sized intervals
Uber Scan

• But first, let's back up momentarily
  – The left-to-right scan found a minimal set of non-overlapping intervals
  – Can we find the set of all intervals of maximal size?
  – These were clearly not found in our left-to-right or right-to-left scans
Uber Scan

- Simple modification to the left-to-right scan algorithm
  - Instead of restarting when an incompatibility is found, only remove a portion of it
  - Specifically remove everything before (in the scanning direction) and including the closest newly introduced incompatibility
  - Open a new interval starting at the first SNP in the queue
  - Continue as before
Uber Scan

• Properties
  – Will contain more than the minimal number of intervals, $k$
  – Each interval is maximal in size (bounded on each side by an incompatibility)
  – Maintains a linear runtime
Max-\(k\) cover

- Minimal set of \(k\) maximally-sized intervals
  - Must be a subset of the Uber scan, since Uber includes all intervals of maximal size
  - Search all subsets of size \(k\)?

\[
\binom{|Uber|}{k}
\]

- No. Combinatorial Explosion
- Instead restructure the problem as a graph problem
Max-\( k \) cover

- Minimal set of \( k \) maximally-sized intervals
  - We know any minimal set must include the cores
  - Find all intervals from the Uber scan that overlap each core
  - Construct a \( k \)-partite graph
    - Vertices are intervals
    - Edges are weighted with the amount of overlap
  - Solve for maximal path (dynamic program)
Max-$k$ cover

- Properties
  - May not be unique
  - Theoretical runtime $O(ku)$, where $u$ is the number of intervals in Uber scan
  - In practice, we never see more than 3 intervals in any part, thus $O(k)$
Uses

• Phylogeny trees
  – Represent the data with the fewest possible trees
  – Maximal intervals provide maximal support for each tree

• Recombination
  – $k$ gives us a lower bound on the minimum number of recombinations needed to make the dataset
  – Although, not very tight
  – But it scales to large datasets
Critical SNPs

• How stable are these intervals?
  • If we remove any given SNP, will the minimal number of intervals needed, $k$, be reduced?

• Algorithm
  – Only consider the flagging SNPs of the Uber intervals
    • These intervals are bounded by incompatibilities, if they are not removed, the interval cannot change size
Some Context

346866 of 689472 Perlegen SNPs on Chr 1, 60 Billion pairwise relationships, >7.5 GBytes
Chromosome 14
15059098-15230790

Trees based on Perfect Phylogenies
Local to Global Trees

• Given a forest of local phylogeny trees, how do we construct a global tree?

• Generally, by combining tree metrics (Sum of distances from $i$ to $j$) across all trees and then applying either neighbor joining or UPMGA.

• Evolution is more complicated than a simple tree:
  – Common introgressions near species splits
  – Gene flows when branches interact