Other approaches to inferring phylogenetic trees
Basic construction approaches

Distance
Tree accounts for evolutionary distances estimated from data

Parsimony
Tree that requires minimum about of change to explain the data

Maximum likelihood
Tree that maximizes the likelihood of the data
Example

AAG
AAA
GGA
AGA

Notice: Sites are treated independently.

We can estimate the cost for each separately and sum over them.
Overview

Input:
- character-based data such as an alignment

Output:
- tree that requires minimal number of changes

Goal:
- Find right tree topology (how things branch) instead of the actual lengths of edges
- Hard part is finding optimal topology
Fitch’s algorithm

Published in 1971

Relies on the following assumptions:
- Any state can convert to any other state
- Positions in an input are independent
- Cost is uniform, i.e., $A \rightarrow T = G \rightarrow T$
Fitch's algorithm

We will do two traversals of the tree

Bottom - up (leaves to root)
- Determine set of possible states for each node

Top - down (root to leaves)
- Pick the ancestral state for each node from the set of possibilities
Step 1

Perform a post-order traversal of the tree

Compute:

# union operations = # of changes
Example

- {AG}
  - {G}
    - {CG}
    - {TG}
      - C
      - G
      - T
      - G
      - A
Step 2

Now do a preorder traversal of tree

Select state $r_j$ for an internal node $j$ with a parent node $i$ as follows:
Example

\[
\begin{align*}
\lambda \quad & \bullet \{CG\} \quad \{TG\} \\
& \bullet \{G\} \\
& \bullet \{AG\} \\
& \bullet C \quad G \quad T \quad \bullet G \\
& \bullet A
\end{align*}
\]
Weighted Parsimony

Sankoff & Cedergren (1983)

Rather than assume all state changes are equally likely, use different costs for different changes.

We’ll need to propagate costs up during first step of approach, but will not cover this in class.
Methods for exploring tree space

Consider any single internal edge in a tree.

There are 3 ways the four subtrees can be grouped:

AB - CD; AC - BD; AD - BC

Using this rule to look at trees is called nearest neighbor interchange.
Exact method

There is a branch and bound approach that can be used to calculate the best tree more efficiently.

In short, if a tree you build is worse than a previously discovered tree, you stop.

Keep track of all partial trees such that you can reuse information as much as possible.
Probabilistic methods

Input:
- Given an alignment and a mutation model (e.g., Jukes-Cantor)

Problem:
- Compute the likelihood of a tree as a product of the individual likelihoods from the alignment
  Assumes columns are independent
Conclusions

Many algorithms exist for these problems.

Parsimony generally does pretty well for most applications.

Likelihood, however, is becoming popular due to increased computational power.
Some future directions for sequencing

Organism sequencing
- Sequence a large fraction of all organisms
- Deduce ancestors
  - Reconstruct ancestral genomes
  - Synthesize ancestral genomes
  - Clone—Jurassic park!
- Study evolution of function
  - Find functional elements within a genome
  - How those evolved in different organisms
  - Find how modules/machines composed of many genes evolved
Sequencing technology

Sanger sequencing

Cost per finished bp:

- 1975
- 1980
- 1990
- 2000
- 2008

- $10.00
- $1.00
- $0.10
- $0.01

Read length:
- 15 – 200 bp
- 500 – 1,000 bp

Throughput:
- “grad-student years”
- $2 \cdot 10^6$ bp/day

Source: bioinformatics.org
$985 deCODEme (November 2007)

$399 Personal Genome Service (November 2007)

$2,500 Health Compass service (April 2008)

Genetic Information Nondiscrimination Act (May 2008)

$~1,000 Whole-genome sequencing (2013)
Sequencing technology
Next-generation sequencing

Read length: 450 bp “short reads”
Throughput: 600-800 Mb/day
Cost: ~20,000 bp/$
De novo: yes

Genome Sequencer / FLX
454 Sequencing

Roche (454) GSFLX Workflow:
Library construction

Emulsion PCR

PTP loading

Pyrosequencing reaction

DNA capture bead containing millions of copies of a single clonally amplified fragment

Sulfuryase

Luciferase

Signal image

Polymerase

APS

Annealed primer

PPi

ATP

Luciferin

Light + Oxy Luciferin

TRENDS in Genetics
Nanopore Sequencing

The Envisioned Device: A SOLID STATE NANOPORNE WITH EMBEDDED NANOTUBE SENSOR

http://www.mcb.harvard.edu/branton/index.htm
Demand for more sequencing

Sequencing technology improvement

Increase in sequencing data output

Demand for more sequencing

New sequencing applications
## Applications

<table>
<thead>
<tr>
<th>Whole-genome sequencing</th>
<th>Genotyping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparative genomics</td>
<td>Population genetics</td>
</tr>
<tr>
<td>Genome resequencing</td>
<td>Migration studies</td>
</tr>
<tr>
<td>Structural variation</td>
<td>Ancestry inference</td>
</tr>
<tr>
<td>analysis</td>
<td>Relationship inference</td>
</tr>
<tr>
<td>Polymorphism discovery</td>
<td>Genetic screening</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>Drug targeting</td>
</tr>
<tr>
<td>Environmental sequencing</td>
<td>Forensics</td>
</tr>
<tr>
<td>Gene expression profiling</td>
<td></td>
</tr>
</tbody>
</table>
Goal

• Sequencing a human genome
<table>
<thead>
<tr>
<th>Technology</th>
<th>Read length (bp)</th>
<th>Pairing</th>
<th>bp / $</th>
<th>de novo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanger</td>
<td>1,000</td>
<td>longish</td>
<td>1,000</td>
<td>yes</td>
</tr>
<tr>
<td>454</td>
<td>250</td>
<td>shortish</td>
<td>10,000</td>
<td>yes</td>
</tr>
<tr>
<td>Solexa/ABI</td>
<td>30</td>
<td>shortish</td>
<td>100,000</td>
<td>maybe</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Application</th>
<th>Sanger</th>
<th>454</th>
<th>Solexa/ABI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacterial sequencing</td>
<td>yes</td>
<td>yes</td>
<td>maybe</td>
</tr>
<tr>
<td>Mammalian sequencing</td>
<td>yes</td>
<td>sort of</td>
<td>probably no</td>
</tr>
<tr>
<td>Mammalian resequencing</td>
<td>Lots of $$</td>
<td>Lots of $$</td>
<td>yes</td>
</tr>
</tbody>
</table>
Overlap-Layout-Consensus

Assemblers: ARACHNE, PHRAP, CAP, TIGR, CELERA

Overlap: find potentially overlapping reads

Layout: merge reads into contigs and contigs into supercontigs

Consensus: derive the DNA sequence and correct read errors
Triazzle: A Fun Example

The puzzle looks simple

BUT there are repeats!!!

The repeats make it very difficult.

Try it – $10.99 at www.triazzle.com

iPhone version too!
Phenotype reconstruction based on mammoth nuDNA (possibly contaminated)
PHASE TWO: INTERPRETATION

I think I found a corner piece.

SHERMAN
The Steele Ledger
Whole Gene Shotgun Sequencing for Metagenomics

Multiple genomes

Random genomes fragmentation

Genomes assembly using overlaps
“Working with thousands of jigsaw puzzles"
What is Metagenomics?

- **Metagenomics** (Environmental Genomics or Community Genomics) is the study of genomes recovered from environmental samples

- **Pro**: No need to culture (grow in lab) them

- **Con**: Heavily uses bioinformatics tools to facilitate insight
Why is Metagenomics Important?

• Some reasons include:

  – Organisms can be studied directly in their environments

  – There are significant advantages for viral metagenomics, because of difficulties cultivating the appropriate host

  – Genomic information has advanced research in a diverse fields such as forensic science
Many projects, many fragments

• Examples:
  – Prokaryote:
    • Sargasso Sea (Venter et al 2004): 1.6 billion base pairs generated estimated to come from 1800 genomic species
  – Viral:
    • Marine water (Breitbart et al 2002) Mission Bay and Scripps Pier. 873 sequences for the Mission Bay and 1061 for Scripps Pier with respectively more than 65% and 73% of unknown
Many projects, many fragments

- Three years after the Marine Water project, most of sequences are still unique. Despite the fact that GenBank has more than doubled in size.

- All of the Metagenome projects have generated enormous amounts of data that still cannot be assembled or annotated.