Phylogenetic Tree Distance Metrics

Christopher Whidden
Dalhousie University
whidden@cs.dal.ca
www.cs.dal.ca/~whidden
Phylogenetic Tree

- Phylogenies represent evolutionary history
- Parents pass genes to their offspring
  - Vertical inheritance
  - Binary tree
- Labelled Leaves
  - Known organisms
- Internal nodes
  - Common ancestors
- Size
  - Number of leaves $n$
Building a Phylogenetic Tree

- **Input Data**
  - Multiple Sequence Alignment

- **Several Scoring Methods**
  - Maximum Parsimony
  - Maximum Likelihood

1: AAAC
2: AAAG
3: CCCCC
4: CCCCG
Exploring Tree Space

1. Start with a base tree
2. Score several modified trees
3. Continue with the best modified tree
Controversy!
- OK: Vertebrates
- Worse: Bacteria, Archaea, and Viruses
What is the Correct Phylogeny?

- Multiple tree-building methods
- Optimization criteria for creating phylogenies are NP-hard
- Many tree statistics
- Reticulation events
  - Hybridization
  - Lateral gene transfer
  - Recombination
Lateral Gene Transfer

- LGT/HGT
- Genes can be shared laterally among organisms
- Several Methods
  - Absorb stray DNA
  - Phages (Viruses)
  - Transfer Plasmids

(Hallet and Lagergren, 2001)
Lateral Gene Transfer

- Messes up the Tree of Life
  - Is it really a tree?
- Non-inherited spread of traits
  - Antibiotic resistance!
  - Hyperthermophiles?

(Hallet and Lagergren, 2001)
Hybridization

- Some closely related species can interbreed
- Most animal hybrids are sterile
- Common among plants

Zena the Zonkey, Oaklawn Farm Zoo Aylesford, Nova Scotia
Distance Metrics

- Distance $k$
- Determine how similar two trees are
- Identify the differences
- Prefer biological meaning
- Model reticulation events
Robinson Foulds Distance

- (Robinson and Foulds, 1981)
- Normalized count of the unique tree bipartitions
- \( A = \) bipartitions of \( T_1 \)
- \( B = \) bipartitions of \( T_2 \)
- \( RF(T_1, T_2) = \frac{(A \setminus B) + (B \setminus A)}{2} \)
- Can be calculated in linear time

- May be much larger than the number of LGT events
  - Does not model LGT well
Robinson Foulds distance = 3
Rooted SPR (rSPR) operation

1. Remove an edge $e$ and its subtree
2. Attach $e$ to a new vertex on another edge
3. Contract degree two vertices
Rooted SPR (rSPR) distance

\[ d_{rSPR}(T_1, T_2) \]

- Minimum Number of rSPR moves to turn \( T_1 \) into \( T_2 \)
- NP-hard
- Models lateral gene transfer
SPR Models LGT

LGT

SPR
SPR Models LGT

- One SPR is equivalent to one LGT
- SPR distance = minimum number of LGT events
  - Assuming no noise, statistical differences, etc
  - Minimum Set of SPRs provide potential LGT events
- Concerned about long SPR operations
Hybrid Number

- Minimum number of extra edges in a hybrid phylogeny
- NP-hard
- Models hybridization events
NP-hard problems

- Exponential running time assuming that $P \neq NP$
- Many NP-hard problems arise in bioinformatics
- Wide variety of methods for solving NP-hard problems in practice
  - Brute force
  - Simplify the problem
  - Heuristics
  - Convert to another problem
  - Approximation Algorithms
  - FPT Algorithms
Proposing SPR operations

- Try all possible combinations of \( x \) SPR operations
- Increment \( x \) if no solution is found
- Simple and obvious
- Slow!
  - \( O(n^n) \)
  - Use heuristics
EEEP

- (Beiko et al., 2006)
- Efficient Evaluation of Edit Paths
  - Uses evolutionarily reasonable constraints to eliminate unproductive search paths
  - Separates trees into regions of discordance
  - Bipartition-based
    - Can collapse bipartitions with poor support
    - Rooted reference tree and possibly Unrooted test tree
- Breadth-first search
  - High memory usage
Heuristics

- **Lattrans**
  - (Hallet and Lagergren, 2001)
  - A,B

- **HorizStory**
  - (MacLeod et al., 2005)
  - A,B
  - Phantom Sister C

- **EEEP**
  - A,B,D,E
  - C without time constraints
Encode as a different problem

- Use the principle of NP-completeness
  1. Encode the problem as a well-known problem
  2. Use an efficient solver
- As solvers improve, your solution improves
- Overhead from the conversion

- Sprdist
  - Convert rSPR distance to Integer Linear Programming
- TreeSAT
  - Convert rSPR distance to Boolean Satisfiability
Approximation Algorithms

- Provide an approximate solution quickly
- Key idea
  1. identify a small set of independent options such that one is correct
  2. apply all of the options
- K-approximation
  - Never better than the true solution
  - At most a factor of K worse than the true solution
- E.g. 3-approximation with distance $= 10$
  - Approximation is between 10 and 30
FPT Algorithms

- Move the exponential complexity to a parameter $k$ other than $n$
- E.g. $O(3^k n)$
  - $k$ is the rSPR distance
- Efficient if $k$ is small
- Two methods
  1. Reduction rules reduce the input to $O(k)$
  2. Bounded search tree
     a. identify a small set of independent options such that one is correct
     b. apply each of the options recursively
Maximum Agreement Forest

- Agreement Forest
  - Forest created by cutting edges in both trees
- Maximum Agreement Forest (MAF)
  - Maximum components
  - Minimum #edges cut
Maximum Acyclicity Agreement Forest

$T_1$ and $T_2$ are two trees. MAF (Maximum Agreement Forest) and MAAF (Maximum Acyclicity Agreement Forest) are shown below.
Agreement Forest Equivalences

- Minimum #edges cut = distance metric
- (Bordewich and Semple, 2005)
  - Rooted MAF: $\text{rSPR distance}$
- (Allen and Steel, 2001)
  - Unrooted MAF: TBR distance
- (Baroni et al., 2005)
  - Rooted MAAF: Hybridization distance
# New Results

<table>
<thead>
<tr>
<th></th>
<th>Approximation</th>
<th>Fixed-parameter (FPT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Previous</td>
<td>New</td>
</tr>
<tr>
<td>TBR</td>
<td>8-approximation</td>
<td>Linear-time 3-approximation</td>
</tr>
<tr>
<td>rSPR</td>
<td>$O(n^2)$ 3-approximation</td>
<td>Linear-time 3-approximation</td>
</tr>
</tbody>
</table>

|                   | Previous                              | New                                          |
| TBR               | $O(4^k k^5 + p(n))$                    | $O(4^k k + n^3)$                             |
| rSPR              | $O(3^k k^4 + n^3)$                     | $O(2.42^k k + n^3)$                         |
| Hybridization     | $O(28^k k^5 + n^3)$                   | $O(3.2^k k + n^3)$                          |

$(28 \times 10)^{10} \approx 2.96 \times 10^{24} = 2,961,967,666,954,240,000,000,000,000$

$(3.2^{10} \times 10) \approx 1.12 \times 10^{6} \approx 1,125,900$
Strategy
rSPR 3-approx Algorithm

- $R_d =$ roots of $F$
- $R_t =$ roots of subtrees that agree in $\hat{T}_1$ and $\hat{F}_2$
- Terminate if $|R_t| < 3$
  - $F_2$ is an AF of $T_1$ and $T_2$
1. Move agreeing roots of components from $R_t$ to $R_d$
2. Propagate labels of $R_t$ upward
3. Cut small edge sets
Case 1

- $c$ from $R_t$ is a root in $\hat{F}_2$
  - Cut in $T_1$
  - Move to $F$ by putting $c$ in $R_d$
Case 2

- Sibling pair \((a, c)\) in \(\hat{T}_1\) and \(\hat{F}_2\)
  - Add parent \(r_{ac}\) with label "\((a, c)\)" to \(R_t\)
  - Remove \(a\) and \(c\) from \(R_t\)
Case 3

- Sibling pair \((a,c)\) in \(\hat{T}_1\) but not \(\hat{F}_2\)
  - Cut \(e_a\), \(e_b\), and \(e_c\)
Edge Shifting

- Bordewich et al
- \( F - (E \cup \{e\}) \) creates a component with no labels
- Can remove any edge from \( E \) bordering that set
- \( F - (E \setminus \{f\} \cup \{e\}) \) and \( F - E \) yield the same forest
Case 3 Proof Sketch

- Choose edge set $E$
- $F_2 - E$ yields an MAF
  1. $e_a$, $e_b$, or $e_c$ is in $E$
     - Done
  2. No leaf $b' \sim b$ in $F_2 - E$
     - Shift edge from $E$ to $e_b$
  3. Same for $a$ and $c$
Case 3 Proof Sketch

4. \(a' \sim b'\) in \(F_2 - E\)
   - \(a \nsim c\)
   - \(c' \nsim x\), for all \(x \in X \setminus X_c\)
   - Shift edge from \(E\) to \(e_c\)
Linear Time

- $O(n)$ cases
  - Each case removes an edge or vertex
- $O(1)$ time per case
  - Nodes with pointers
  - Preprocess to match labels in $T_1$ to labels in $T_2$
  - List of sibling pairs in $\hat{T}_1$
  - List of labeled roots in $\hat{F}_2$
- $O(n)$ running time
rSPR FPT Algorithm

- Is the rSPR distance $\leq k$?
- Exact answer
- Case 3 branches into three computational paths
  - Cut $e_a$, cut $e_b$, cut $e_c$
  - reduced $k$
- $O(3^k n)$ running time
- Only store current path
  - Low memory usage
- Iterative deepening $k=0,1,...$
Case 3

- Sibling pair \((a,c)\) in \(\hat{T}_1\) but not \(\hat{F}_2\)
- Three branches cutting \(e_a\), \(e_b\), and \(e_c\)
Branch-and-bound

- Before branching calculate the 3-approximation
- If approx > $3k$
  - Return “no”
- More work per step but far fewer steps
- Worst case running time unchanged
MAAF Refinement

- rSPR algorithm works for hybridization distance
  1. Apply rSPR algorithms to get agreement forest $F$
  2. Refine $F$ to an MAAF
- $O(3.2^k \cdot n)$ FPT
Improved Algorithm

- Identify better branching rules
  - Split Case 3 into three subcases
  - Reduce the branching factor to 2.42
  - Much better performance in practice
Case 3.1: Separate Components (SC)

- $a$ and $c$ in separate components
- Two branches cutting $e_a$ and $e_c$
Case 3.2: One Pendant Node (COB)

- One pendant node $b$
  - Cut $e_b$
Case 3.3: Multiple Pendant Nodes (CAB)

- Multiple pendant nodes $b_1, b_2, ..., b_q$
- Branch cutting $\{e_a\}, \{e_{b_1}, e_{b_2}, ..., e_{b_q}\}$ and $\{e_c\}$
Improved Algorithm

- **SC**
  - Cut 2
  - $T(k) = 2T(k-1)$

- **COB**
  - Cut 1
  - $T(k) = T(k-1)$

- **CAB**
  - Cut 3
  - $T(k) = 2T(k-1) + T(k-2)$

- $O(2 \cdot 42^k n)$
MAAF

- One Pendant Edge Improvement may create a cycle!
  - Cut $e_b$, $e_c$ instead
  - Analysis still holds
- $O(3 \cdot 2^k n)$
Experimental Evaluation
rSPR Datasets

**Random Dataset**

<table>
<thead>
<tr>
<th>Leaves</th>
<th>rSPR distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1-20, 25, 30, ... 50</td>
</tr>
<tr>
<td>1000</td>
<td>1-20, 25, 30, ... 1000</td>
</tr>
</tbody>
</table>

- 10 pairs of trees for each

**Protein Tree Dataset**

- Reference supertree
  - 144 taxa
  - Constructed using MRP
  - Rooted to separate Bacteria and Archaea
- 22432 unrooted test trees
  - Derived from a microbial data set with MrBayes
- Used 5689 test trees with 10 or more taxa
- Rooted using approximation
rSPR results

- 4GB memory limit
- 5 hour running time
- TreeSAT
  - Solves rSPR distance using a SAT solver
- Sprdist
  - Solves rSPR distance using an ILP solver
  - Previously shown to be faster than EEEP, Lattrans, etc
Fig. 5. Number of protein trees solved by rSPR distance.
Fig. 6. Mean running time of the FPT, sprdist, and treeSAT protein tree runs.
**Fig. 7.** Mean running time of the FPT algorithm on the data set of randomly permuted trees. The trees with 100 and 1000 leaves are shown separately.
**Fig. 8.** Mean running time for combinations of the new cases on the protein tree runs.
Conclusions

- Determining the correct phylogeny is difficult
- Distance metrics
  - Compare the similarity of trees
  - Model reticulation events
- Many methods are used to solve NP-hard problems in practice
Conclusions

- Can efficiently handle moderate rSPR distances
- Protein Trees
  - $N=144$, $K=20$
    - less than a second on average
    - Limit of the other algorithms in 5 hours
  - $N=144$, $K=46$
- Random
  - $N=100$, $K=42$
  - $N=1000$, $K=99$
- A hybridization implementation should be similar
Future Work

- **Open Problems**
  - Multifurcating trees
  - Multiple tree agreement forests

- **Implementations**
  - rSPR approx and FPT
    - [http://kiwi.cs.dal.ca/Software/RSPR](http://kiwi.cs.dal.ca/Software/RSPR)
  - TBR approx and FPT
  - Hybridization approx and FPT

- **Partitioning**
  - Identifying LGT
  - Rooting trees