New Techniques for Building Large-Scale Evolutionary Trees

Tiffani L. Williams

Department of Computer Science
Texas A&M University
tlw@cs.tamu.edu
Dr. David Acer, a Florida dentist

- In 1990, a Florida woman contracted AIDS without exposure to the established risks of HIV.
- The CDC launched an extensive investigation to establish the cause of her HIV contraction.
- They suspected her dentist, David Acer, who was suffering from AIDS, was responsible for her AIDS contraction.
- Was the dentist responsible for infecting the Florida woman? If so, how could it be proven?
Figure: Phylogeny of HIV transmission in the Florida dentist case.
Dr. Richard Schmidt, a Louisiana physician

- In 1998, a Louisiana woman claimed that her ex-lover and physician, Richard Schmidt, deliberately injected her with HIV-tainted blood.
- Evolutionary trees proved that the physician did indeed infect his ex-lover with HIV-tainted blood.
- This case marks the first time that evolutionary trees were admitted as evidence in a US criminal court.
- Schmidt is serving a 50-year sentence for attempted second-degree murder.
Evolutionary Trees and Relationships

- Here, relationship refers strictly to connections based on genealogy.
- In other systems, relationship may be equated with similarity without any evolutionary implications.
- Things which ”look the same are the same” can lead to incorrect classifications.
- Degree of similarity is never equated with degree of relatedness in a phylogenetic system.
- Evolutionary trees are hypotheses and not facts. Their ”accuracy” depends upon both the quality and quantity of data which support them as well as the methods used to reconstruct them.
Figure: Phylogeny: Illustration.
New Techniques for Building Large-Scale Evolutionary Trees

Introduction

Tree-Building Timeline

- **1980s**: Most published trees consisted of less than 50 taxa.
- **1990s**: Evolutionary trees with 500 taxa could be constructed in a few hours!
  - Previously required years on a cluster!
  - The parsimony ratchet lead to faster and more accurate algorithms.
- **Today**: Phylogenetic trees with 16,000 taxa can be constructed in less than a day!
- **Future**: To reconstruct the Tree of Life!
Figure: The Tree of Life.
Phylogenetic Methods

- **Distance-based methods**
  - **Software:** PAUP*, BIONJ

- **Parsimony-based methods**
  - **Software:** PAUP*, Phylip, Mega, TNT

- **Likelihood methods**
  - **Software:** PAUP*, Phylip, fastDNAML, GAML, MrBayes

- **Meta-methods**
  - **Software:** TREE-PUZZLE, PAUP*, Rec-I-DCM3
Our focus will be on Maximum Parsimony

- **Small Parsimony Problem (Tree Evaluation):** Compute MP score for a given tree $T$.
  - $O(nk)$, where $n$ is the number of sequences and $k$ is the sequence length.

- **Large Parsimony Problem (Tree Search):** Find the trees with minimal cost by searching the tree space.
  - Very difficult! It’s an NP-hard problem.
  - For $n$ taxa, there are $(2n - 5)(2n - 3) \cdots (5)(3)$ trees.
  - Over 13 billion possible trees for 13 taxa!
New Techniques for Building Large-Scale Evolutionary Trees

Maximum Parsimony

Figure: Evaluation of a fixed tree.

A C A C G C T T C A G
A T T C G A T T T C G
C G A G G A A T G A G
C G G C G A A T T C G

Fixed tree

Random Assignment

Optimal Assignment
New Techniques for Building Large-Scale Evolutionary Trees

Maximum Parsimony

Figure: Scoring a fixed tree.
Figure: The space of all 15 possible unrooted trees with for 5 organisms.
Figure: The parsimony score of the trees in previous figure. Dataset is A=100110, B=001000, C=110000, D=110111, E=001110.
Figure: Overview of DCMs.
DCM Decompositions

- Construct a triangulated graph and then apply either a max-clique or a separator-component decomposition to the graph to obtain the subproblems.
  - Threshold graph decompositions
  - Guide tree decompositions
- Combination of base method, choice of triangulated graph, and decomposition technique on the triangulated graph impact the behavior the resultant "DCM-boosted" method.
Figure: DCM decomposition.
Figure: The leaves that from the short subtree around edge $e$ are $A, D, E, G, I$. 
DCM3 and its Variants

1. Static decomposition + TNT-Ratchet
   - **DCM3**: Uses a "guide tree" for the dataset decomposition.
   - **Rec-DCM3**: Recursively applies a DCM3 decomposition until each subproblem is at most $m$.

2. New decomposition every iteration.
   - **I-DCM3**: Iteratively computes a DCM3 tree.
   - **Rec-I-DCM3**: Iteratively constructs a Rec-DCM3 tree.
Rec-I-DCM Performance

Figure: Performance of methods on a dataset of 4114 Actinobacteria sequences (1263 sites).
Using a Population of Trees

- Rec-I-DCM3 is guided by a single tree.
- Good performance from a search algorithm is due to a balance of *exploration* and *exploitation*.
- Unclear if above objectives can be met by a local search heuristic that manipulates a single individual.
Cooperative Rec-I-DCM3

- Maintains a population of Rec-DCM3 trees to search the tree landscape.
- Cooperative in the sense that the individuals in the population share pieces (i.e., subtrees) of themselves to create new individuals.
- The presence of diverse solutions in the population represent a collective effort in finding accurate trees quickly.
Overview of Cooperative Rec-I-DCM3

1. Create \( p \) initial trees (e.g., random trees, GreedyMP trees), which represents the initial population of solutions.

2. Run Rec-I-DCM3 on each of the \( p \) trees.

3. Create a new population of \( p \) trees.
   - Selection
   - Recombination

4. Repeat steps 2 and 3 for the required number of iterations.
Creating a new population for iteration $i + 1$

- **Best tree found during search**: A
- **Top-scoring trees found in iteration $i$**: B
- **Random selection of remaining trees from iteration $i$**: C

Tree population: $T_0, T_1, T_2, T_3, \ldots, T_{p-3}, T_{p-2}, T_{p-1}$
Experimental Methodology

■ **Platforms:**
  ■ 4-node Apple Workgroup Cluster for Bioinformatics; each node consists of two, 2 GHz PowerPC G5 processors. 16 GB total RAM across the cluster.
  ■ 4-node Linux Beowulf cluster; each node contains two, 2 GHz Intel Xeon processors. 8 GB total RAM across the cluster.

■ **Datasets:**
  ■ 921 aligned Avian Cytochrome b DNA sequences. Best known score of 40494 established by Cooperative Rec-I-DCM3.
  ■ 2000 aligned Eukaryotic sRNA sequences from the Gutell Lab at UT-Austin. Best known score of 74536 established by Cooperative-Rec-I-DCM3.

■ **Implementation:** TCP-Linda from Scientific Computing Associates (www.lindaspaces.com)
Phylogenetics Meets Linda

Tuple Space

- `eval("recidcm3", recidcm3(id))`
- `rd("tree", ?tid, ?str, 40)`
- `out("tree", 3, "((D,A,B,C))", 40)`
- `recidcm3(id=1)`
- `startup(id=0)`
- `recidcm3(pid=2)`
- `recidcm3(pid=3)`

Trees:
- `("tree",0,"((A,C),(B,D)), 39")`
- `("bestSore", 35)`
- `("tree",10,"((A,B),(C,D)), 43")`
Figure: Average performance of Rec-I-DCM3 and Cooperative Rec-I-DCM3 on a dataset of 921 Avian Cytochrome b DNA sequences.
Figure: Extending the runs of Rec-I-DCM3 for 1000 iterations on a dataset of 921 Avian Cytochrome b DNA sequences.
Figure: Average performance of Rec-I-DCM3 and Cooperative Rec-I-DCM3 on a dataset of 2000 Eukaryotic sRNA sequences.
Figure: Extending the runs of Rec-I-DCM3 for 1000 iterations on a dataset of 2000 Eukaryotic sRNA sequences.
## Running Times

<table>
<thead>
<tr>
<th>popsize</th>
<th>Dataset 1 iteration (secs)</th>
<th>Dataset 1 total (hrs)</th>
<th>Dataset 2 iteration (secs)</th>
<th>Dataset 2 total (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54.00</td>
<td>15.00†</td>
<td>157.64</td>
<td>43.74†</td>
</tr>
<tr>
<td>2</td>
<td>55.08</td>
<td>1.53</td>
<td>157.68</td>
<td>4.38</td>
</tr>
<tr>
<td>4</td>
<td>58.68</td>
<td>1.63</td>
<td>184.32</td>
<td>5.12</td>
</tr>
<tr>
<td>6</td>
<td>72.00</td>
<td>2.00</td>
<td>219.96</td>
<td>6.11</td>
</tr>
<tr>
<td>8</td>
<td>114.12</td>
<td>3.17</td>
<td>335.16</td>
<td>9.31</td>
</tr>
</tbody>
</table>

**Table:** Average running times for Rec-I-DCM3 ($popsize=1$) and Cooperative Rec-I-DCM3. † is the time to run Rec-I-DCM3 for 1000 iterations.
Benefits of Cooperation to Phylogenetics

- Cooperation is a promising approach to finding good trees in the exponentially-sized tree space.

- Why limit the tree-building process to one technique?
  - Distance approaches
  - Parsimony
  - Maximum likelihood
  - Bayesian analysis

- Why not run multiple tree-building techniques simultaneously?
  - There is no best search algorithm. (No Free Lunch Theorem)
  - Leverage the strengths of the different techniques as a unified approach to tree-building.
Figure: Cooperating phylogenetic methods.
Summary

■ Large evolutionary trees are extremely difficult to reconstruct.

■ Need new approaches for establishing support for our inferences on extremely large trees.
  ■ Disk-Covering Methods
  ■ Cooperative tree searches

■ Plenty of room for future work
  ■ Heterogeneous mix of methods
  ■ Population creation strategies
  ■ Stopping criteria
  ■ ... and much more
Acknowledgements

- Collaborators
  - Jessica Jones, Harvard University
  - Bernard Moret, U of New Mexico
  - Usman Roshan, New Jersey Institute of Technology
  - Marc Smith, Colby College
  - Tandy Warnow, U of Texas at Austin

- Radcliffe Institute of Advanced Study
- CiPRES project: www.phylo.org
- Webpage: faculty.cs.tamu.edu/tlw
THANKS FOR LISTENING!
New Techniques for Building Large-Scale Evolutionary Trees

Concluding Remarks