Computational Challenges from the Tree of Life

Tiffani L. Williams

Department of Computer Science
Texas A&M University
tlw@cs.tamu.edu
Figure: The Tree of Life.
Figure: Sequence evolution.
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.
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 14,000 taxa can be constructed in less than a day!
- **Future:** To reconstruct the Tree of Life!
  - Consists of approximately 10 to 100 million taxa.
  - Current methods can handle a few thousand taxa.
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!
Figure: Small Parsimony Problem: Evaluation of a fixed tree.
Figure: Large Parsimony Problem: Tree search under MP.
Maintaining a Population of Solutions

- The best known algorithm for MP is Rec-I-DCM3.
- Rec-I-DCM3 is guided by a single tree (or solution).
- 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.
Cooperative Rec-I-DCM3 Algorithm

1. Obtain $p$ initial trees (e.g., random trees, GreedyMP trees), which represent 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
- Top-scoring trees found in iteration $i$
- Random selection of remaining trees from iteration $i$

$T_0 \rightarrow A \rightarrow B \rightarrow C \rightarrow T_1 \rightarrow T_2 \rightarrow T_3 \rightarrow \ldots \rightarrow T_{p-3} \rightarrow T_{p-2} \rightarrow T_{p-1}$
**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

```
out("tree", 3, "((D,A),B,C)")
in("tree", ?tid, ?str, ?score)
eval("recidcm3", recidcm3(id))
("tree",10,"((A,B),(C,D))", 43)
("tree",0,"((A,C),(B,D))", 39)
("bestSore", 35)
("tree",10,"((A,B),(C,D))", 43)
recidcm3(id=1)
startup(id=0)
recidcm3(pid=3)
recidcm3(pid=2)
```

Tuple Space

```
("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>total (hrs)</th>
<th>Dataset 2 iteration (secs)</th>
<th>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. Rec-I-DCM3 and Cooperative Rec-I-DCM3 ran for 1000 and 100 iterations, respectively.
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.
Additional Challenges

- Statistical validation of results
- Consensus methods
- Constrained reconstruction
- Simulation studies
Summary

- Reconstructing the TOL is the computational grand-challenge of evolutionary biology.
- Research challenges include designing new heuristics, foundational work in statistics, and basic algorithm and data structure design.
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!