Cooperating Evolutionary Trees

Tiffani L. Williams

Department of Computer Science
Texas A&M University
tlw@cs.tamu.edu
Figure: Phylogeny: Illustration.
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: The Tree of Life.
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!
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 $(2n - 5)(2n - 3) \cdot \cdot (5)(3)$ trees.
  - Over 13 billion possible trees for 13 taxa!
Cooperating Evolutionary Trees

Maximum Parsimony

Figure: Evaluation of a fixed tree.
Cooperating Evolutionary Trees

Maximum Parsimony

Figure: Exhaustive search under MP.
Figure: The space of all 15 possible unrooted trees with for 5 organisms.
Disk-Covering Methods (DCMs)

1. Decompose the original problem into overlapping subproblems.
2. Solve subproblems with a phylogenetic method of choice.
3. Merge the partial solutions into a complete solution.
4. Refine the merged tree into a binary tree.
Figure: Overview of DCMs.
Rec-I-DCM Performance

Figure: Performance of methods on a dataset of 4114 Actinobacteria sequences (1263 sites).
Motivating Question

- Search algorithms attempt to balance two objectives.
  - **Exploration**: Generation of new individuals in untested regions of the search space
  - **Exploitation**: Concentration of the search in the vicinity of known good solutions

- Can both of these objectives be met by a local search heuristic that manipulates a single solution to guide its search?
Cooperative Rec-I-DCM3 Algorithm

1. Create a population of $\mu$ starting trees, which represents the initial population of solutions.
2. Run Rec-I-DCM3 on each of the $\mu$ trees.
3. Create a new population of $\mu$ trees by performing selection and recombination on the trees returned from the Rec-I-DCM3 search.
4. Repeat steps 2 and 3 for the required number of iterations.
**Figure**: Selection operator. Here, $\mu = j + k + l$. 

<table>
<thead>
<tr>
<th>Best trees found during search</th>
<th>Top ranking trees from iteration $i$</th>
<th>Low ranking trees from iteration $i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_0$</td>
<td>$\ldots$</td>
<td>$A_j$</td>
</tr>
</tbody>
</table>
Figure: Recombination operator.
Figure: Evolutionary trees cooperating in tuple space.
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.

■ Experiments:
  ■ All experiments consisted of 5 runs of Rec-I-DCM3 and Cooperative Rec-I-DCM3.
  ■ Rec-I-DCM3 ran for 1,000 iterations.
  ■ Cooperative Rec-I-DCM3 ran for 100 iterations.

Figure: Average performance of Rec-I-DCM3 and Cooperative Rec-I-DCM3 on a dataset of 2,000 aligned Eukaryotic sRNA sequences from the Gutell Lab at UT-Austin.
Figure: Wall-clock performance of Rec-I-DCM3 and Cooperative Rec-I-DCM3 on a dataset of 2,000 aligned Eukaryotic sRNA sequences from the Gutell Lab at UT-Austin.
Figure: Wall-clock performance of Rec-I-DCM3 and Cooperative Rec-I-DCM3 on a dataset of 2,000 aligned Eukaryotic sRNA sequences from the Gutell Lab at UT-Austin.
Can we do better?

- **Diverse strategy:**
  - Incorporates a heterogeneous mix of solutions (in terms of tree scores) into the population.
  - Results show that such strategy outperforms Rec-I-DCM3.

- **Elitist strategy:**
  - Each solution in the population matches that of the best found so far in the search.
  - Motivated, in part, by its use in traditional phylogenetic searches.
  - This strategy only uses selection. There is no recombination.
Figure: Diverse population strategy. Average performance of Rec-I-DCM3 and Cooperative Rec-I-DCM3 on a dataset of 2,000 aligned Eukaryotic sRNA sequences from the Gutell Lab at UT-Austin.
Figure: Elitist population strategy. Average performance of Rec-I-DCM3 and Cooperative Rec-I-DCM3 on a dataset of 2,000 aligned Eukaryotic sRNA sequences from the Gutell Lab at UT-Austin.
**Figure: Diverse population strategy.** Average performance of Rec-I-DCM3 and Cooperative Rec-I-DCM3 on a dataset of 1,127 aligned large subunit ribosomal RNA sequences (1,078 sites) obtained from the Ribosomal rRNA database.
Figure: Elitist population strategy. Average performance of Rec-I-DCM3 and Cooperative Rec-I-DCM3 on a dataset of 1,127 aligned large subunit ribosomal RNA sequences (1,078 sites) obtained from the Ribosomal rRNA database.
Summary

- Large evolutionary trees are extremely difficult to reconstruct.
- Need new approaches for establishing support for our inferences on extremely large trees.
- Experimental results demonstrate that both population size ($\mu$) and selection have a huge impact of the performance of Cooperative Rec-I-DCM3.
- For diverse populations, a population of cooperating trees leads to increased performance as well as establishing best scores on several datasets.
- Performance gains are negligible when an elitist population strategy is used.
Figure: Cooperating phylogenetic methods.
Spring 2006: Computational Phylogenetics

Sample topics include:

- Phylogenetic methods (e.g., maximum parsimony, maximum likelihood, Bayesian)
- Consensus and supertree methods
- Models of DNA evolution
- High-performance phylogenetics
- Phylogenetic networks
- Multiple sequence alignment
- Other uses of phylogeny reconstruction (e.g., linguistics, chain letters)
Acknowledgements

- Collaborators
  - 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, Harvard University
- CiPRES project: www.phylo.org
- Webpage: faculty.cs.tamu.edu/tlw
THANKS FOR LISTENING!