CURRENT METHODS FOR AUTOMATED FILTERING OF MULTIPLE SEQUENCE ALIGNMENTS FREQUENTLY WORSEN SINGLE-GENE PHYLOGENETIC INFERENCE

TAN ET AL., SYST. BIOL. 2015

CS 581 Paper Presentation
Jinfeng Xiao
3/29/2018
Intro: MSA Filtering

- **Filtering**: Removing unreliable *columns* from a multiple sequence alignment (MSA) *before* downstream tasks, e.g. tree reconstruction

- This paper provides the 1st systematic and comprehensive comparison of various filtering methods on real data
## Filtering Methods in This Study

<table>
<thead>
<tr>
<th>Filtering methods</th>
<th>Type of “undesirable” sites filtered out by the method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gblocks</td>
<td>Gap-rich and variable sites</td>
</tr>
<tr>
<td>TrimAl</td>
<td>Gap-rich and variable sites</td>
</tr>
<tr>
<td>Noisy</td>
<td>Homoplasic sites</td>
</tr>
<tr>
<td>Aliscore</td>
<td>Random-like sites</td>
</tr>
<tr>
<td>BMGE</td>
<td>High entropy sites</td>
</tr>
<tr>
<td>Zorro</td>
<td>Sites with low posterior</td>
</tr>
<tr>
<td>Guidance</td>
<td>Sites sensitive to the alignment guide tree</td>
</tr>
</tbody>
</table>
Test #1: Species Tree Discordance Test

- Orthologous sequences from species with an undisputed branching order should conform to the species tree.
- A particular filter is beneficial if it leads to trees more similar to the species tree.
- 10,999 sets of six orthologs were sampled from the Sep 2008 Orthologous MAtrix (OMA) database release on three taxonomic ranges: fungi, eukaryotes, and bacteria.
- For each set, up to 24 additional homologs were sampled from the SwissProt database with a threshold E-value of $10^{-10}$.
- Run once on amino acid and once on nucleotide data.
Enriched Species Tree Discordance Test

<table>
<thead>
<tr>
<th>Step</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Sequence sampling</td>
</tr>
<tr>
<td>2.</td>
<td>Alignment</td>
</tr>
<tr>
<td>3.</td>
<td>Filtering</td>
</tr>
<tr>
<td>4.</td>
<td>Tree building</td>
</tr>
<tr>
<td>5.</td>
<td>Pruning</td>
</tr>
<tr>
<td>6.</td>
<td>Tree evaluation</td>
</tr>
</tbody>
</table>

1. **Sequence sampling**
   - Orthologs
   - Additional homologs

2. **Alignment**
   - Prank
   - Enriched alignment

3. **Filtering**
   - TrimAl
   - GBlocks
   - Noisy

4. **Tree building**
   - PhyML

5. **Pruning**

6. **Tree evaluation**
   - 50%
   - 100%
   - 66%
Test #2: Minimum Duplication Test

- More accurate gene trees tend to require fewer duplications to explain the history of gene families
- Up to 60 homologs from eukaryotic and fungal genomes
- Run once on amino acid and once on nucleotide data
- Observations are very similar to the Species Tree Discordance Test
Minimum Duplication Test on Nucleotides
Test #3: Ensembl Compara Test

- Infers gene trees reconciled with a reference species tree (i.e. gene trees whose inner nodes are labeled as speciation or duplication events)
- Some resulting reconciled trees contain very poorly supported duplication nodes: Differential gene loss
- Uses the average number of losses per branch over all trees as indicator for the contribution of filtering
- Based on Ensembl data version 66, which contains 19,491 homologous clusters (with MSAs/trees) containing 969,577 protein-coding genes from 57 species
Results of Ensembl Compara Test
Test #4: Simulation Test

- Performed on simulated data
- Tree accuracy
- Alignment accuracy: Calculate precision and recall by comparing the homologs pairs in the filtered and reference alignments
- Two sets of 500 30-sequence MSAs were simulated using Artificial Life Framework (ALF)
  - Sequence length was drawn from a Gamma distribution
  - Sequences were evolved along 30-taxa birth–death trees
  - Characters were substituted according to WAG substitution matrices
  - Insertions and deletions were applied
Results of Simulation Test
Similar Results Were Observed Under Other Settings

- **Aligners**
  - *Mafft 6.843*
  - *Prank 100802*
  - *ClustalW 2.0.10*
  - *T-Coffee v.5.72 & v.10.00*
  - *M-Coffee*

- **Tree reconstruction methods**
  - *Maximum likelihood: PhyML 3.0*
  - *Least-squares distance: Darwin*

- >1M CPU hours in total!
Take-Home Messages

- Alignment filtering does not in general improve phylogenetic tree inference
- Default settings are usually suboptimal
- More aggressive filtering generally results in poorer trees
- Modest amounts of filtering have little impact on the reconstructed trees but decrease computational time