GSR
The Gene Sequence evolution model
with iid Rate variation over tree

Örjan Åkerborg, KTH
Lars Arvestad, KTH
Jens Lagergren, KTH
Bengt Sennblad
What?

- Gene evolution through duplication and loss
- Sequence evolution
Why?

• Base reconciliation analysis directly on data
  – Avoid information loss
  – Addresses uncertainty better

• Gene tree reconstruction should mirror generation
When?

- Arvestad et al. 2003,
  - MrBayes + GEM
  - Flawed model
60s ribosomal data

Table 2. Probabilities and bootstrap support values for RLA1 and RLA2 using three different methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>RLA1</th>
<th>RLA2</th>
</tr>
</thead>
<tbody>
<tr>
<td>OrthoStrapper</td>
<td>0.35</td>
<td>0.46</td>
</tr>
<tr>
<td>RIO</td>
<td>0.12</td>
<td>0.21</td>
</tr>
<tr>
<td>Our method</td>
<td>0.88</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Fig. 6. Species tree for the 60s ribosomal testcase. Numbers indicate million years ago and are taken from Wang et al. (1999).
When?

• Arvestad et al. 2003,
  – MrBayes + GEM
  – Flawed model

• Arvestad et al. 2004
  – Intergrated GEM + Substitution model
  – Mathematically correct model
  – Molecular clock
  – Sampling algorithm - slow
MHC revisited

Table 2: Orthology and paralogy probabilities for the MHC-data using different methods of orthology analysis. (Abbreviations used: GEM = Gene Evolution Model, GSEM = Gene Sequence Evolution Model)

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Parsimony</th>
<th>GEM</th>
<th>GSEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthology</td>
<td>Yes</td>
<td>0.89</td>
<td>0.75</td>
</tr>
<tr>
<td>Paralogy</td>
<td>No</td>
<td>0.11</td>
<td>0.25</td>
</tr>
</tbody>
</table>
When?

• Arvestad et al. 2003,
  – MrBayes + GEM
  – Flawed model

• Arvestad et al. 2004
  – Intergrated GEM + Substitution model
  – Mathematically correct model
  – Molecular clock
  – Sampling algorithm - slow

• Åkerborg et al. Submitted (GSR)
  – Integrated GEM + SRT model
How?

• **GSR**
  – **GEM**
    • Reconciled trees – duplication and loss
  – **SRT**
    • Relaxed clock model (iid)
    • Substitution model

– **Fast algorithm**
  • Discretized time space
Self-consistency

$Pr[D, T]$

Sequence data, $F$

Trees, $T$
Self-consistency – the X%-test
Application to Yeast data

• Compare to prev. Results
  – YGOB
  – Orthogrups (SYNERGY)
    • Both synteny-based
Synteny -- gene order
Application to Yeast data

• Compare to prev. Results
  – YGOB
  – Orthogroups (SYNERGY)
    • Both synteny-based
    – Whole genome duplication
      • Challenge!

• Genome-wide analysis!
  – 4809 gene families (orthogroups)
Comparison YGOB results

<table>
<thead>
<tr>
<th>Vertex</th>
<th>Vertex recovery comparison between modes*</th>
<th>Mode 4 vertex recovery loss with different number of post WGD genes removed†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mode 2</td>
<td>mode 4</td>
</tr>
<tr>
<td>Pre-WGD/post-WGD</td>
<td>159</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td>86.4%</td>
<td>89.7%</td>
</tr>
<tr>
<td>Both</td>
<td>105</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>57.1%</td>
<td>62.5%</td>
</tr>
</tbody>
</table>

*Absolute number and percent of total number of 184 gene families given.
†Absolute number and percent of the number of genes recovered with all genes included.
Molecular clock?

Mean and CV of substitution rates

- YGOB mean
- YGOB CV
- Orthogroup mean
- Orthogroup CV
Comparison SYNERGY results

Table 2. The SYNERGY comparison.

<table>
<thead>
<tr>
<th>Orthogroup size</th>
<th>Analyzed orthogroups</th>
<th>SYNERGY trees ranked</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th+</th>
<th>SYNERGY trees not ranked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>1,477</td>
<td></td>
<td>894</td>
<td>97</td>
<td>34</td>
<td>45</td>
<td>407</td>
</tr>
<tr>
<td></td>
<td>60.5%</td>
<td></td>
<td>6.6%</td>
<td>2.3%</td>
<td>3.0%</td>
<td>27.6%</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1,580</td>
<td></td>
<td>912</td>
<td>67</td>
<td>23</td>
<td>19</td>
<td>559</td>
</tr>
<tr>
<td></td>
<td>57.7%</td>
<td></td>
<td>4.2%</td>
<td>1.5%</td>
<td>1.2%</td>
<td>35.4%</td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>774</td>
<td></td>
<td>131</td>
<td>24</td>
<td>9</td>
<td>23</td>
<td>587</td>
</tr>
<tr>
<td></td>
<td>16.9%</td>
<td></td>
<td>3.1%</td>
<td>1.2%</td>
<td>3.0%</td>
<td>75.8%</td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>978</td>
<td></td>
<td>614</td>
<td>41</td>
<td>8</td>
<td>6</td>
<td>309</td>
</tr>
<tr>
<td></td>
<td>62.8%</td>
<td></td>
<td>4.2%</td>
<td>0.8%</td>
<td>0.6%</td>
<td>31.6%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4,809</td>
<td></td>
<td>2,551</td>
<td>229</td>
<td>74</td>
<td>93</td>
<td>1,862</td>
</tr>
<tr>
<td></td>
<td>53.0%</td>
<td></td>
<td>4.8%</td>
<td>1.5%</td>
<td>1.9%</td>
<td>38.7%</td>
<td></td>
</tr>
</tbody>
</table>

The table shows the number of orthogroups for which the gene tree suggested by SYNERGY was ranked first, second, third, fourth or lower, and not ranked in the PrIME-GSR posterior distribution.
Sequence vs. synteny data

• No sequence diff
  – 36% of data sets are >85% similar

• Strong divergence
  – 25% of data set are <40% similar
  – Long-branch attraction

• Conflicting sequence-synteny signal
Sequence vs. synteny data

- Spom_SPAC1002.13
- Anid_AN6697.2
- Fgra_FG01351.1
- Mgri_MG00505.4
- Ncra_NCU02668.2
- Ylip_YAL10E33539g
- Kwal_Kwal4800
- Klac_KLLA0F004433g
- Agos_AER372C
- Cgla_CAGL0J09922g
- Cgla_CAGL0G08668g
- Scas_Scas694.17
- Scas_Scas662.3
- 1 2
  - Smik_YNL066W-18618
  - Spar_YNL066W-19097
- 3 4
  - Smik_YIL123W-10746
  - Spar_YIL123W-10680
- Dhan_DEHA0F22242g
- Calb_orf19.3642
Summary

• primeGSR
  – Integrated model
    • Reconciliation – gene duplication loss
    • Relaxed clock
    • Sequence evolution
  – Efficient algorithms
  – Improved gene tree reconstruction

• Future prospects
  – Divergence time estimates (MAP)
  – Species tree reconstruction
  – Include synteny