## GeneMarkS-2 accuracy assessment

<table>
<thead>
<tr>
<th>Gene Finder</th>
<th>Number of missed genes (with COG annotation) in 145 prok genomes</th>
<th>Number of FP in simulated intergenic sequence (1 MB for each genome)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glimmer</td>
<td>2,053</td>
<td>74,061</td>
</tr>
<tr>
<td>Prodigal</td>
<td>1,350</td>
<td>23,919</td>
</tr>
<tr>
<td>GeneMarkS-2</td>
<td>837</td>
<td>10,931</td>
</tr>
</tbody>
</table>

Zoom in on COG genes:

All missed COG genes

- GMS-2: 248
- Prodigal: 589
- GMS-2 ∩ Prodigal: 761

Missed COG genes longer than 600 bp

- GMS-2: 42
- Prodigal: 67
- GMS-2 ∩ Prodigal: 92
ORFs supported by at least two peptides are considered as genes. 87,359 ORFs with peptide support in the set

Predictions incompatible with peptide supported ORFs (full out of frame overlap) are considered as false predictions

Test on 58 genomes with proteomics data.
Accuracy of gene start prediction. Test on genes with starts confirmed by N-terminal sequencing of protein products

<table>
<thead>
<tr>
<th></th>
<th>Number of validated genes</th>
<th>Prodigal</th>
<th>GeneMarkS-2</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. pernix</td>
<td>130</td>
<td>127</td>
<td>124</td>
<td>-3</td>
</tr>
<tr>
<td>E. coli</td>
<td>769</td>
<td>751</td>
<td>744</td>
<td>-7</td>
</tr>
<tr>
<td>H. salinarum</td>
<td>530</td>
<td>514</td>
<td>520</td>
<td>6</td>
</tr>
<tr>
<td>M. tuberculosis</td>
<td>701</td>
<td>620</td>
<td>628</td>
<td>8</td>
</tr>
<tr>
<td>N. pharaonis</td>
<td>315</td>
<td>309</td>
<td>313</td>
<td>4</td>
</tr>
<tr>
<td>Synechocystis</td>
<td>96</td>
<td>92</td>
<td>89</td>
<td>-3</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>2,541</strong></td>
<td><strong>2,413</strong></td>
<td><strong>2,418</strong></td>
<td><strong>5</strong></td>
</tr>
</tbody>
</table>
GeneMarkS-2 is available to run & test at
http://topaz.gatech.edu/GeneMark/genemarks2.cgi