Accounting for Contamination in Sequence Data

Chloe Lee
How do you know what organism your DNA sequence is from?
Terminology

- GC content
- GC3
- ORF
- contig
- chimera
How do you know what organism your DNA sequence is from?

- BLAST, but if your sequence is contaminated, what do you get out?
- GC content varies among different organisms
- Translational selection for highly expressed genes influences GC content
- Mutational selection
Reasoning for project:

- *Childonella uncinata* as a model organism for genomic architecture and genome evolution
- MAC single gene chromosomes
- Paralogs
- Endosymbionts?
Procedure

- Take assembly from SPAdes
  - Identify contaminants
  - Assess validity of assembly
- Analyze GC content of sequence
- Identify potential protein coding regions
- Determine the effective number of codons for those ORFs
GC content

- Genes are characterized as having higher GC content than genome as a whole
- Longer coding sequences associated with higher GC content
- GC content variable among organisms

Implementation:
- Sliding window script in python
ZOOMING IN ON AREAS OF LOW OR HIGH GC

Area of low GC content

Sliding Window Start site (bp)

GC content (%)
Glimmer
Gene Locating Interpolated Markov Model

- Linear combination of 8 markov chains
- 3 different Markov Models
- Sequence score is based on probabilities of bases in sequence
- Designed for working with gene rich sequences
<table>
<thead>
<tr>
<th>ORF ID</th>
<th>Start point</th>
<th>End point</th>
<th>Frame</th>
<th>Raw Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>orf00001</td>
<td>3295745</td>
<td>3297883</td>
<td>+1</td>
<td>14.89</td>
</tr>
<tr>
<td>orf00002</td>
<td>153</td>
<td>881</td>
<td>+3</td>
<td>12.32</td>
</tr>
<tr>
<td>orf00004</td>
<td>1893</td>
<td>997</td>
<td>-1</td>
<td>17.83</td>
</tr>
<tr>
<td>orf00005</td>
<td>2349</td>
<td>1900</td>
<td>-1</td>
<td>18.73</td>
</tr>
<tr>
<td>orf00006</td>
<td>3615</td>
<td>2359</td>
<td>-1</td>
<td>13.95</td>
</tr>
<tr>
<td>orf00008</td>
<td>3811</td>
<td>3602</td>
<td>-2</td>
<td>16.59</td>
</tr>
<tr>
<td>orf00010</td>
<td>3846</td>
<td>5303</td>
<td>+3</td>
<td>15.28</td>
</tr>
<tr>
<td>orf00011</td>
<td>5371</td>
<td>5955</td>
<td>+1</td>
<td>16.80</td>
</tr>
<tr>
<td>orf00013</td>
<td>5989</td>
<td>7014</td>
<td>+1</td>
<td>19.15</td>
</tr>
<tr>
<td>orf00014</td>
<td>7017</td>
<td>7799</td>
<td>+3</td>
<td>17.83</td>
</tr>
<tr>
<td>orf00016</td>
<td>7822</td>
<td>8442</td>
<td>+1</td>
<td>17.23</td>
</tr>
<tr>
<td>orf00018</td>
<td>8445</td>
<td>9629</td>
<td>+3</td>
<td>15.11</td>
</tr>
<tr>
<td>orf00021</td>
<td>9645</td>
<td>10415</td>
<td>+3</td>
<td>17.59</td>
</tr>
<tr>
<td>orf00022</td>
<td>10419</td>
<td>11012</td>
<td>+3</td>
<td>16.08</td>
</tr>
</tbody>
</table>
**Nc**

**Effective number of codons**

- Evaluated using the program chips on the ORFs found by glimmer
- Low values in highly biased genes
- High values in lowly biased genes
Effective number of codons vs. GC content of glimmer predicted orfs
Examples
Blast of ORFs with high raw score
Goals

- Make scripts user-friendly
- Learn how to use matplotlib
- Calculate GC3 instead of GC
- Plot results of different contigs together
- Train Glimmer using both *Chilo* and contaminant clade predictions
- Investigate Glimmer-MG
- BLAST results – convert to protein sequences
- Isolate *Chilodonella uncinata* DNA from the assembly.
Sources


NCBI BLAST

Glimmer 3.02

Emboss Chips

CodonW Sourceforge

Supek, Frank et al. **Translational Selection is Ubiquitous in Prokaryotes.** PLOS Genetics. 2010 June 25.

Xyrus Maurer-Alcala, Laura Katz
Cusp

# CdsCount: 50065

# Coding GC 41.09%
# 1st letter GC 41.07%
# 2nd letter GC 41.31%
# 3rd letter GC 40.90%

<table>
<thead>
<tr>
<th>Codon</th>
<th>AA</th>
<th>Fraction</th>
<th>Frequency</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCA</td>
<td>A</td>
<td>0.344</td>
<td>16.901</td>
<td>17035</td>
</tr>
<tr>
<td>GCC</td>
<td>A</td>
<td>0.236</td>
<td>11.604</td>
<td>11696</td>
</tr>
<tr>
<td>GCG</td>
<td>A</td>
<td>0.140</td>
<td>6.874</td>
<td>6929</td>
</tr>
<tr>
<td>GCT</td>
<td>A</td>
<td>0.280</td>
<td>13.730</td>
<td>13839</td>
</tr>
<tr>
<td>TGC</td>
<td>C</td>
<td>0.506</td>
<td>16.912</td>
<td>17046</td>
</tr>
<tr>
<td>TGT</td>
<td>C</td>
<td>0.494</td>
<td>16.510</td>
<td>16641</td>
</tr>
<tr>
<td>GAC</td>
<td>D</td>
<td>0.251</td>
<td>6.241</td>
<td>6291</td>
</tr>
<tr>
<td>GAT</td>
<td>D</td>
<td>0.749</td>
<td>18.647</td>
<td>18795</td>
</tr>
<tr>
<td>GAA</td>
<td>E</td>
<td>0.734</td>
<td>21.273</td>
<td>21442</td>
</tr>
<tr>
<td>GAG</td>
<td>E</td>
<td>0.266</td>
<td>7.718</td>
<td>7779</td>
</tr>
</tbody>
</table>