IDENTIFYING A TYPE OF GENETIC CODE IN AN ANONYMOUS, PROKARYOTIC DNA SEQUENCE

AARON PFENNING1, ALEXANDER LOMSADZE2, MARK BORODOVSKY1, 2, 3
1School of Biological Sciences, 2Wallace H. Department of Biomedical Engineering, 3School of Computational Science and Engineering

INTRODUCTION

- It was commonly believed that genetic code was universal when it was discovered.
- Due to technological advances variations of the canonical code have been discovered.
  → Calling for an ab-initio approach
- Recently, phages with two different genetic codes have been described.
  → Computational tool must be able to predict potential switching points

Why is gene prediction important?
- If an outbreak of a new virus has happened: → accurate gene prediction is required to help identify potential drug targets in downstream analysis
- Code switch point predicted with a mean error of 0.53 genes ± 6.47 genes
  → Utilizes prediction of canonical and non-canonical models to refine switching point predictions in difficult cases

CONCLUSION

- First tool of its kind
- Accurate on complete genomes and contigs greater than 10Kbp
  → makes use of other codon frequencies to determine to which amino acid a stop codon is reassigned
- If reassigned: frequency is significantly increased

RESULTS

Table 1: Results on dataset 1, 2 & 3. No misclassifications were made in the simple mode. When employing the complex mode there is one genome predicted to have a partial reassignment of ~8% in dataset 1. In dataset 2 & 3 some partial reassigments of less than 10% are predicted and hence should be considered as artifacts.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
<th>Dataset 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>No reassignment</td>
<td>0</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td>Complete reassignment</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Accuracy</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Complex mode</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No reassignment</td>
<td>0</td>
<td>88</td>
<td>99</td>
</tr>
<tr>
<td>Partial reassignment</td>
<td>1</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Complete reassignment</td>
<td>99</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.99</td>
<td>0.93</td>
<td>0.99</td>
</tr>
</tbody>
</table>

1. Simple mode
- Tested on dataset 4:
  - 5 genomes annotated as genetic code 4 but predicted with genetic code 11 at NCBI
  - All Acholeplasma sp.
- Av. Gene length ~1000nt → NCBI agreed and changed code assignment

2. Complex mode
- Merged contigs from dataset 1 & 3 to simulate switching points
- Mean error of 0.53 genes ± 6.47 genes
  → in reality it might be less when strand information can be utilized for refinement of the prediction → simulated genomes do not show change in encoding as observed in the phages of dataset 5

REFERENCES

1. S. Osawa and T. H. Jukes, Codon reassignment (codon capture) in evolution
2. A. Rabin and S. Rachmany, The Universal Genetic Code and Non-Canonical Variants
3. R. D. Papp, Codon usage bias in the wild
4. N. N. Varsha et al., Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut
5. E. Guerin et al., Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus of less than 10% are predicted and hence should be considered as artifacts
6. K. Zitronella et al., An Unsupervised, New-Canonical, Reverse-Complement, and Very High GC% Code
7. J. Zhou, Z. Lomadze, and M. Borodovsky, 60-nt wide gene identification in metagenomic assemblies
8. A. Sentmanal, E. Gutierrez, S. Tang, and M. Borodovsky, Modulating leaderless transcription and opitcal gene models in high-accuracy gene prediction in prokaryotes
9. A. Ya et al., Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms

CONTACT

E-mail: apfennig3@gatech.edu
Phone: +1 (404) 649 2817

MATERIALS AND METHODS

Workflow of Genetic Code Identifier

Simple mode

<table>
<thead>
<tr>
<th>Code length</th>
<th>Simple mode</th>
<th>Mean error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000nt</td>
<td>1.0</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Complex mode

<table>
<thead>
<tr>
<th>Code length</th>
<th>Complex mode</th>
<th>Mean error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000nt</td>
<td>0.93</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Figure 1: Modified Stop codons frequencies. TGA is reassigned in genetic code 4, its frequency is significantly increased.

Figure 2: Av. gene lengths drops from 1000nt to 400nt if incorrect model is employed (dataset 1 & 2).

Figure 3: The analysis of a genome of a phage in dataset 5. The phage is predicted to have two different genetic codes. The predictions are concordant with the literature.

Figure 4: Evaluation of the complex mode and the switch point prediction on simulated contigs.