Abstract:
The rise of antibiotic-resistant bacteria has increased interest in bacteriophages (viruses that kill bacteria) in recent years. Due to the decreasing cost of genome sequencing, the number of sequenced phage genomes is growing at a geometric rate. Sequencing is followed by annotation, in which genes, start codons, and putative protein functions are identified. Most phage genomes are auto-annotated with programs designed for prokaryotes. Accuracy metrics for these programs with regard to phage genomes are not available. The genome of Escherichia coli phage Lambda was used to benchmark the accuracy of several genome annotation methods and programs. Discovered in 1951, Lambda is the most well-studied phage, with nearly all gene functions and start sites demonstrated experimentally. Eight programs were used to annotate the Lambda genome: Glimmer, BASys, RAST, GeneMark, GeneMarkS, GeneMarkS2, and GeneMark with Heuristic models. Calls were compared to the reference genome from the literature.

Goal: To determine the accuracy of the eight selected programs in regard to bacteriophage genome annotation.

Hypothesis: Manual curation and compilation of auto-annotation results obtained from several programs will yield more accurate gene feature and start codon prediction than auto-annotation alone.

Methods:

1. Lambda raw nucleotide sequence obtained from Genbank.
2. Auto-annotation with each program under investigation.
3. Results tabulated and compared to protein interaction map to find TN, TP, FN, and FP.

Step 2 Random Sequence Analysis
- Random nucleotide sequence (40,000 bases) generated in R.
- Auto-annotation with each program.
- Results tabulated and compared to obtain a clearer picture of the TN and FP.

Step 3 Manual Annotation
- Auto-annotation with Glimmer and 5 GeneMark programs.
- Gene calling (features called by ≥6 programs).

Step 4 PseudoGene Analysis
- PseudoGenes called by ≥1 non-Glimmer program were scrutinized (all called starts, 8 pseudogenes and 9 starts total).
- Evaluated on ntBLAST, coding potential (CPM), calls, length, and overlap.
- Start Codon Calling (all potential starts for each gene)
- Evaluated on pBLAST, program calls, length, codons, overlap, CPM, and likelihood of translational start (RBS).

Results:

Table 1 Calling programs, coordinates, length, CPM for all PseudoORFs generated by non-Glimmer programs in the randomly generated sequence.

Table 2 Positively and negatively. 73 genes, 545 true negatives in the reference.

Calculations:

**Sensitivity**

**True Positive Rate (TPR) Describes the proportion of genes called correctly**

\[
TPR = \frac{TP}{TP + FN}
\]

**Precision**

**Positive Predictive Value (PPV) Describes the probability that a call is a gene**

\[
PPV = \frac{TP}{TP + FP}
\]

**Figure 1 Feature designations in regard to phage genomics. A pseudoORF is a continuous reading frame ≥75bp long.**

Conclusions:

- Manual annotation is slightly more accurate, particularly in start calling.
- No gene called by all programs was a false positive.
- Some genes were not detected by any annotation method.
- Glimmer is disproportionately prone to nonsense FN.
- Pseudogenes called by any program other than Glimmer represent borderline features, which, during manual annotation, would be:
  - Deleted in the absence of any additional evidence.
  - Kept if they satisfied one of the following conditions:
    1. Filled a gap completely without generating overlap or direction change.
    2. Created 4bp overlap on one or both ends, suggesting an operon.
- Pseudogenes called in the random sequence have:
  1. No plateaus of coding potential.
  2. No significant ntBLAST result.
  3. No significant pBLAST results.
  4. Coding potential does not align well with start and STOP coordinates.

Future Research:

- Uncalled genes will be scrutinized and compared to pseudogenes called in the random sequence.

Acknowledgements:

This research is supported by the McNair Scholars Summer Research Institute.

Thanks to the Office of Undergraduate Research Summer Undergraduate Research Funding Scholarship (OUR SURF), and to the Mechanisms of Evolution Research Experience for Undergraduates (MOE REU).