Partial gene predictions on unassembled reads: evaluating the Good, the Bad and the slightly ORF [Posters / Short talks]

Genome annotation is a difficult computational challenge that is often reliant on the observation of previously discovered genes, both putative and predicted. Statistical analysis conducted on these genes and their host genomes is used to build representative models for describing their characteristics. However, much of this work was carried out around the turn of the century when the collection of representative genomes and experimentally validated genes were limited. Previous work on evaluating annotation techniques and attempts to improve contemporary annotations and the methods behind them have highlighted a number of shortfalls and areas for improvement. Specifically, these include the reliance on prior knowledge and assumptions of genes such as start codon selection, minimum length, GC content and gene overlap. Additionally, this work has mostly been conducted on assembled genomes where the tools can observe the entirety of the target gene, and its genomic neighbourhood context.

The numerous challenges associated with genome assembly, whether for cultured isolates or environmental DNA, introduce a host of additional complexities, particularly when dealing with metagenomic samples. As the assembly process itself introduces errors and uncertainties, impacting the accuracy of gene prediction, it can result in both erroneous gene annotation and increase the difficulty of detecting such errors. Additional challenges lie not only in genome annotation but also in overcoming the incomplete utilisation of sequenced reads. As sequencing depth and costs have caught up and even surpassed computational resources, it is now common for large metagenomic assembly projects to be unable to incorporate large proportions, often up to half, of their read collection. Therefore, while tools to study unassembled reads have been used to study function and taxonomy, they most often rely on alignments of the entire read or k-mers to a precomputed database. This does not allow for the investigation of genes without database similarities or for the future reconstruction of the full gene product.

Predicting gene content directly from unassembled reads can help overcome several variables such as assembly error and reduce computational complexity. Therefore, here we describe a set of metrics for the systematic evaluation of the performance of prokaryotic gene annotation tools applied to unassembled DNA sequencing reads. Our previous work on the ORForise platform provided 72 helpful and explanatory metrics for the evaluation of CoDing Sequence (CDS) prediction tools. However, while the ORForise platform was able to apply all metrics to each CDS prediction we evaluated, we found that read analysis creates additional obstacles. When evaluating the read-based predictions against the known genes present in a ground truth genome the obstacles include handling the directionality and location of the read mapping in relation to the genome and the directionality of the prediction made on the read. This is further complicated by the potentially truncated nature of the prediction on a short read. Therefore, this requires a different approach. To evaluate correctness, we must examine multiple aspects: direction, frame, start position, stop position, and the scope of the partial prediction.

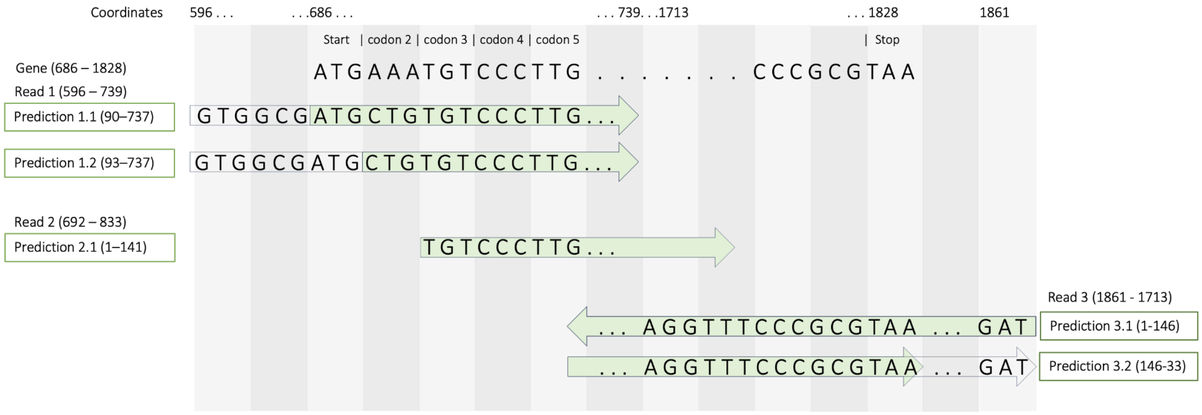
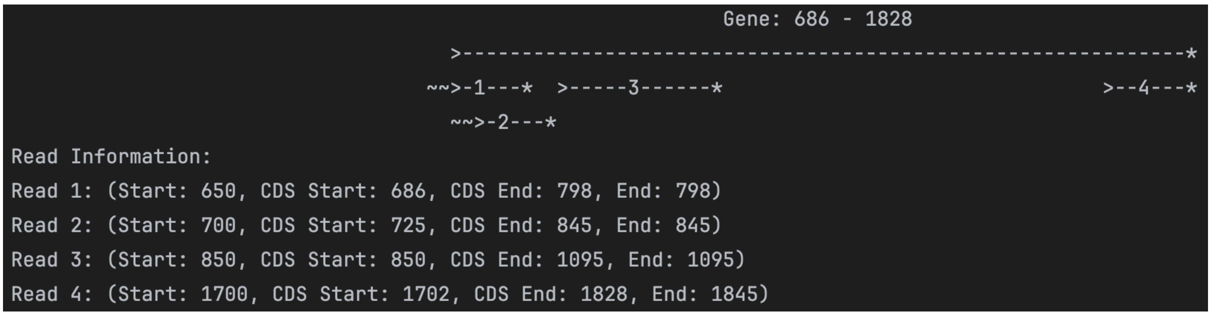


Figure 1: An illustration of some of the complexity when evaluating read-level gene predictions. Many reads map to a gene, and these may span the gene, overhang the ends of the gene or be contained within the gene. The reads may also map in reverse. Each read contains a partial prediction for the coding region, which may or may not align with the boundaries of the gene.

We provide a comprehensive evaluation framework for the prediction of CDS regions (both fragmented and whole) direct from DNA reads. We demonstrate that the insight into annotation correctness provided by this framework suggests that established tools should be re-evaluated. We found that, as with our previous analysis of gene prediction from assembled-reads, tool performance appears to be species and gene-specific.



```
                                                            Gene: 686 - 1828
                                     >-------------------------------------------------------------*
                                   ~~>-1---*  >-----3------*                              >--4---*
                                     ~~>-2---*
Read Information:
Read 1: (Start: 650, CDS Start: 686, CDS End: 798, End: 798)
Read 2: (Start: 700, CDS Start: 725, CDS End: 845, End: 845)
Read 3: (Start: 850, CDS Start: 850, CDS End: 1095, End: 1095)
Read 4: (Start: 1700, CDS Start: 1702, CDS End: 1828, End: 1845)
```

Figure 2: A visualisation interface provided as part of the evaluation framework that can be used to show how genes are covered by partial predictions in order to fully understand the results of the evaluation. Four reads are shown here, each making a partial CDS gene prediction.

Furthermore, by studying the characteristics of the miscalled or omitted gene fragments, we are able to learn how they perform under different circumstances and for different types of genes and identify areas of improvement. To harness this information, we developed additional read annotation approaches which intentionally used naive assumptions and found that their performance was sometimes comparable to or better than state-of-the-art methods. These results can be used to further our understanding of prediction success or failure and provide insight to improve prediction tools.