1.



2.a.i

2.a.ii

2.b.i

Basic Assumptions:

1. An ORF starts with start codon(TAC) and ends with stop codons (ATT, ACT, ATC)
2. The minimum length of genes in prokaryotes is 200bp, 300bp in eukaryotes (excluding the stop codons)
3. In the same reading frame, if there are overlapping genes sharing the same stop codons. The longest ORF is the one we are looking for.
4. In different reading frames, the maximum overlapping length is 60bp.
5. In different reading frames, if maximum overlapping length is greater than 60bp, we keep the longer ORFs.

2.b.ii



2.b.iii

The prediction evaluation contains two parts. Firstly, we compared our prediction result with the prediction result of GLIMMER. Then we compared our predicted genes with the proteome of the five species.

1) Comparison with GLIMMER

True positive: Nucleotides predicted both in GLIMMER and our own predictor.

False positive: Nucleotides that appears only in our predictor instead of GLIMMER

True negative: Nucleotides that presents in GLIMMER instead of our own predictor.

False negative: Nucleotides that don’t show up in both predictions.

Results:



2) Comparison with Uniprot proteome

To compare the prediction result with the Uniprot proteome, we firstly searched for the proteome of five species in Uniprot. Then, we translated our predicted genes in to proteins to compare.

To identify the proteins that were correctly predicted, we performed two blastp test. Firstly ,we used the Uniprot proteome as database and the predicted proteins as query to blast. Hits with e-value less than 0.001 were selected as the predicted proteins that appears in the Uniprot proteome. Then, we used the Uniprot proteins as queries and the predicted proteome as the dataset and ran blast again. Proteins that shows up in both blast tests are considered as true positive predictions.

True positive: Proteins that appears in both the real proteome and the predicted proteome.

False positive: Proteins that appears only in the predicted proteome instead of the Uniprot proteome.

False negative: Proteins that appears only in the Uniprot proteome instead of the predicted proteome.

The sensitivity and specificity is calculated using formulas above.

Results:



2.b.iv

1) Adding promoter information

In prokaryotes, there are Pribnow boxes near ORFs. The Pribnow box locates at 10bp upstream the transcrptional start site and the transcrition start site is 20 to 40 nucleotides upsteam the start codon. We can check if there is pribnow box near the start codon to help with ORF prediction or ORF validation. (ref)

In eukaryotes, TATA boxes information can be used in the same way as the Pribnow box.

2) The minimum gene length

The average gene length is 991bp and there are around 10% of the genes that shorter than 300bp in E.coli. (ref) So we selected 200bp as our minimum gene length threshold. While the minimum gene length varies among different species and each species should be treated differently.

3) The maximum overlapping length

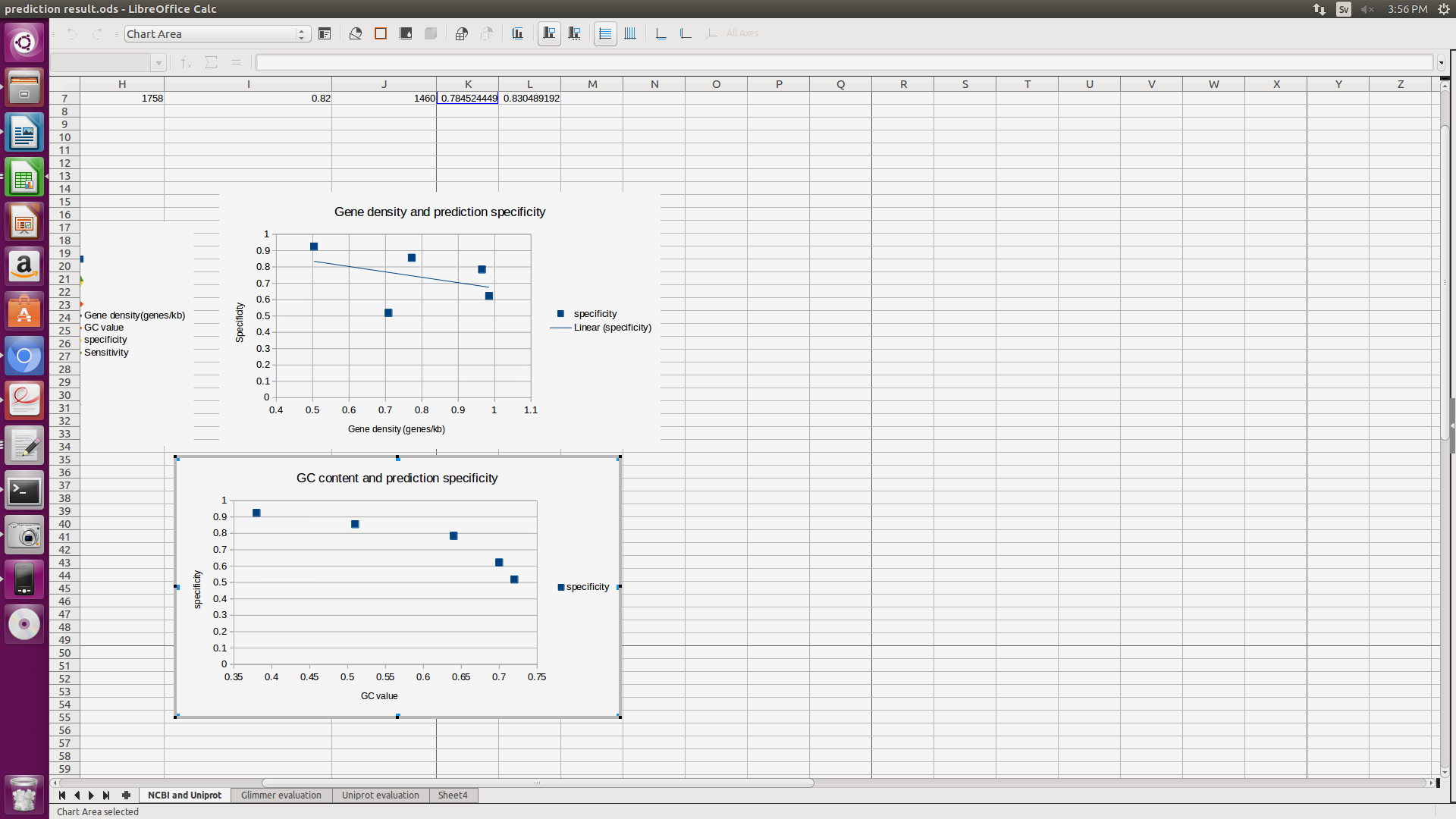
Overlaps in different reading frames that are longer than 60bp are forbidden in our predictor. In this case, we only select the longest ORF in the overlapping genes. However, the maximum overlapping length and the selection of overlapping distance depends on the species and should be treated differently in different species.

4) Gene length distribution

Our predictor selects the longest gene among all overlapping genes(genes that shares the same ORFs). However, the longest gene is not always the most likely ORF. Instead, we can find information of the length distribution of ORFs and calculates the probability of each gene lengthin the genome and select the most likely ORF.

5) Different start codons

In prokaryotes, the start codon could be ATG, GTG and TTG. Our preditor only takes ATG as the start codon. However,some genes starts with Val instead of Met in the real genome and in the prediction of GLIMMER. We found that the prediction specificity decreases with the increase of GC content in our prediction.



The higher the GC content is, the more likely the ORFs starts with GTG instead of ATG. We should also include other possible start codons in the predictor.