Prokaryotic gene prediction on assembly graphs

Egor Barsukov
Advisor: Sergey Nurk

1 Introduction

Gene prediction is an essential part of almost any assembly-based metagenomic analysis. Unfortunately, many genes in metagenomic assemblies end up fragmented across multiple contigs due to differences between individual strains/species sharing particular gene. Since all gene prediction tools available today only work with contiguous nucleotide sequences, such variable genes completely evade straightforward analysis. In this project we tried to perform gene prediction directly on assembly graph of metagenomic assembly and realize additional filtering using information from common gene prediction tools to handle false positive predictions.

2 Report

All the experiments were carried out on SYNTH dataset. It is synthetic metagenomic mixture of 64 prokaryotes and archeas with different coverage. It contains 109 million 100 bp paired-end reads with mean insert size of 206 bp. The assembly was carried out with the size of k-mers 55. Assembly graph comprise 59 659 nodes and 63 395 edges. All genes are annotated, consequently we can use this information to assess the quality of our work.

2.1 Number of scattered genes

To find out the scale of the gap in gene finding only on contigs we use QUAST to find scattered genes. The Figure 1 shows that scattered genes usually passes 3 or less contigs. Also we modified online-vis in SPAdes to see that of scattered genes have complete path on assembly graph and, accordingly, are of interest to us.

1. **182 thousands** genes total in our dataset
2. **12 thousands** (6% of total) genes are scattered by contigs
3. **5.5 thousands** (3% of total) scattered genes have complete path on assembly graph

Also we compute distribution of length of genes in edges (Figure 1). In order to count the number of edges that genes have passed we also used modified online-vis. In Figure 1 we can see that most genes passed less than 14 edges.

2.2 ORF's finding

Since prokaryotic genes contains in open reading frames (ORF), the first our step is implement ORF's search on assembly graph.

Assuming a uniform random distribution, a stop codon is expected to be observed every 21 codons (since there are 3 stop codons). Average proteins are much longer, being coded by about 1000 base pairs. Therefore we consider only long ORFs.

To prevent exponential growth of ORFs number and execution time we add limit on search depth. In order to determine this limit we examine distribution of genes length (in passed edges). In out dataset was found that almost all genes placed less that on 14 edges.

Table 1: ORFs search benchmarking

Minimum length	1000	300
Maximum depth	8	15
ORFs found	22 553	158 916
Partial genes found	3292/5500	5343/5500

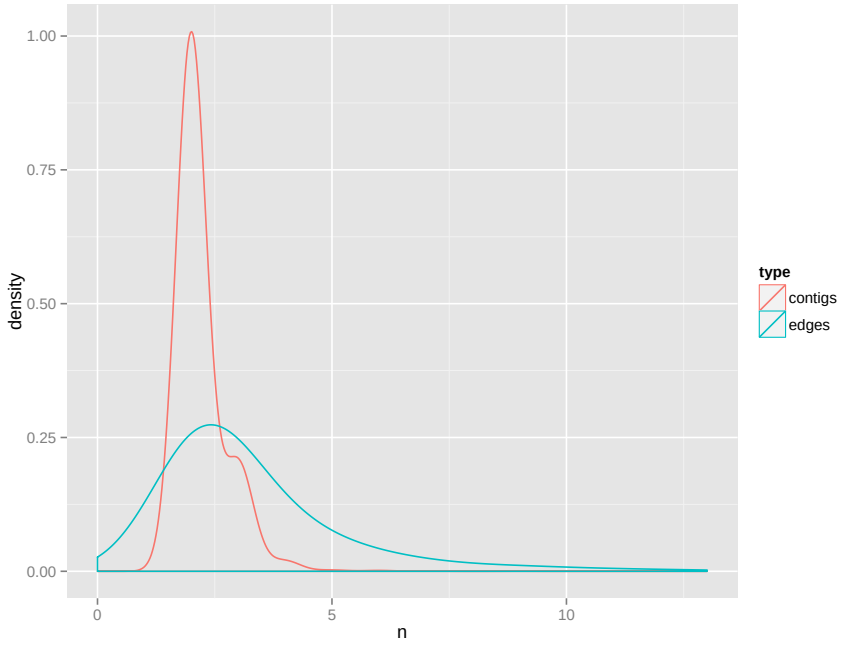


Figure 1: Number of passed edges or contigs for scattered genes

2.3 Clustering

ORF's search gives a lot of possible genes that doesn't exist in genome. In order to deal with this problem we perform hierarchical clustering the results of the previous step. The distance between two sequences is defined as the ratio of the length of identical sections to the sum of the lengths of both sequences.

As we can see on Figure 2 we can choose the largest possible threshold of clustering, since with its increase the cluster that contains partial genes almost do not merge, unlike clusters that do not contain partial genes.

2.4 Additional filtering

As a temporary solution to further reduce the number of clusterswe only consider ORFs containing "partial" genes reported by Prodigal in contigs. Thus, we significantly reduce the number of false positives (but is still large) which can be seen bellow.

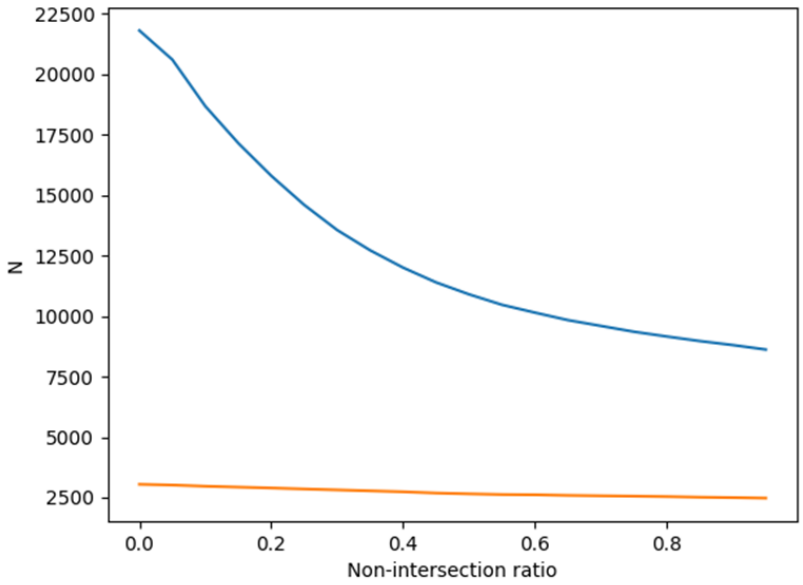
This is benchmarking of additional filtering for 158 916 ORFs:

1. Total found 39 302 possible genes in 6150 clusters
2. 3925 clusters contains real gene
3. 5163 real genes was found total (of 5.5k possible)

Figure 2: Clustering benchmarking for 22 553 ORFs

Clustering distance	No clustering	0.4	1-
Number of clusters	22 553	12 008	8326
Clusters with partial (on contigs) genes	3292	2738	2459
Clusters with partial (on edges) genes	5636	4686	3989
Partials genes (contigs) found	3292/5500		

(a) Number of clusters for different threshold



(b) Blue line — number of cluster. Orange line — number of clusters containing genes.