Introduction to Prokaryotic gene prediction (CDS and rRNA)

Frank Aylward

Abstract

Citation: Frank Aylward Introduction to Prokaryotic gene prediction (CDS and rRNA). **protocols.io** dx.doi.org/10.17504/protocols.io.pigdkbw
Published: 17 Apr 2018

Protocol

Download a Prokaryotic genome to start analyzing

Step 1.

We'll be working with Prochlorococcus marinus MED4 today.

wget -O med4.fna.gz ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/007/925/GCF_000007925.1_ASM792v1/GCF_000007925.1_ASM792v1_genomic.fna.gz

gunzip med4.fna.gz

Predict protein-coding genes using Prodigal

Step 2.

Prodigal will predict genes from chromosomes (or contigs), translate those genes into amino acids, and produce annotation summary files such as gff, depending on what options you use.
prodigal -i med4_genome.fna -a med4.proteins.faa -d med4.genes.fna -f gff -o med4.prodigal.gff

You can also make genbank format files this way

Step 3.

or use GenBank output file for a summary
prodigal -i med4_genome.fna -a med4.proteins.faa -d med4.genes.fna -f gbk -o med4.prodigal.gbk

What about rRNA genes?

Step 4.

Prodigal is only useful for predicting protein coding genes. What other kind of genes are there in genomes?

Barrnap is useful for predicting rRNA genes
barrnap med4_genome.fna > med4.rRNA.gff

What about rRNA genes?

Step 5.

unfortunately barrnap only provides the summary files (in this case gff). So we need to do a bit more legwork to get the actual sequences
bedtools getfasta -fi med4_genome.fna -bed med4.rRNA.gff -fo med4.rRNA.fasta

16S genes are extremely useful for classification. If you ever have a genome and you don't know what it is, a good first step is to identify any 16S ribosomal genes in the chromosome and use them for classification.