Assignment 10 Review

Bio5488
3/29/2019
Don’t hesitate and ask questions early!
Big Idea

• Looking for resistance genes in a human gut microbiota sample.
• You have done functional metagenomics to enrich for resistance genes in your sample.
• You have assembled these short reads into longer contigs.
• Data looks like: Metagenomic shotgun sequencing contigs that putatively contain resistance genes
Contig assembly

Bacterial genomes present in a sample

Genomes cut into small fragments

Sequentialing of many random fragments from pool of fragments

DNA sequences

Computer-assembled consensus sequence

Alignment of DNA sequences with a computer program to create a larger consensus sequence
Task 1

• Annotate (aka “identify”) the Open Reading Frames (ORFs) in these contigs
“Longest ORFs”? 

- ORFs are mutually exclusive -> 2 ORFs in a certain frame cannot share the same start or stop codon

- One possible strategy: start searching for a new start codon (ATG) only after the position you encountered the last stop codon (start codons between a start and a stop are ignored).
Task 2

- Run MetaGeneMark on command line

*Ab initio* gene identification in metagenomic sequences

Wenhan Zhu¹, Alexandre Lomsadze² and Mark Borodovsky²,³,⁴,*

¹School of Biology, ²Wallace H. Coulter Department of Biomedical Engineering, ³School of Computational Science and Engineering and ⁴Center for Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, GA 30332, USA
Is my call_orfs.py working?

Online tools that do similar work?

Notes for this tool:
1: an ORF is allowed to run to the end of the contig (no STOP codon required)
2: it only reports the longest ORF

More tools?
Task 3

• Compare approaches for ORF finding
• MetaGeneMark vs. your homemade script
Take home messages:

• Small changes in prediction model input
  -> large changes in output

• Gene prediction
  -> identify new potential genes from unknown

• Many methods exist for annotation
  -> performance comparison + check for assumptions

• Non-ATG start codons exist in prokaryotic genomes! They still code for a modified Met (fMet)