SciLifeLab

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

Lucile Soler PhD
Nima Rafati PhD

Bacterial

Genome Annotation

https://nbisweden.github.io/workshop-genome_annotation_elixir/labs/prokaryote_annotation

elixir

excelerate

Genome assembly and annotation course
2021

Bacterial genome characteristics

- A bacterial genome is a single "circular” DNA molecule with several million base pairs in size
- Bacteria can contains plasmids (small and circular DNA molecules, that contain (usually) non-essential genes)
- Genomes contain a few thousand genes.
- ”Gene density” is much higher than in humans, one million base pairs of bacterial DNA contains about 500 to 1000 genes.
  - bacterial genes have no introns (intron-less),
  - the average number of codons in bacterial genes is less than in human genes,
  - neighboring genes are very close together throughout the genome

Bacterial feature types

- protein coding genes and associated features
  - promoter (-10, -35)
  - ribosome binding site (RBS)
  - coding sequence (CDS)
    - signal peptide, protein domains, structure
  - terminator

- non coding genes
  - transfer RNA (tRNA)
  - ribosomal RNA (rRNA)
  - non-coding RNA (ncRNA)

- other
  - repeat patterns, origin of replication, ...

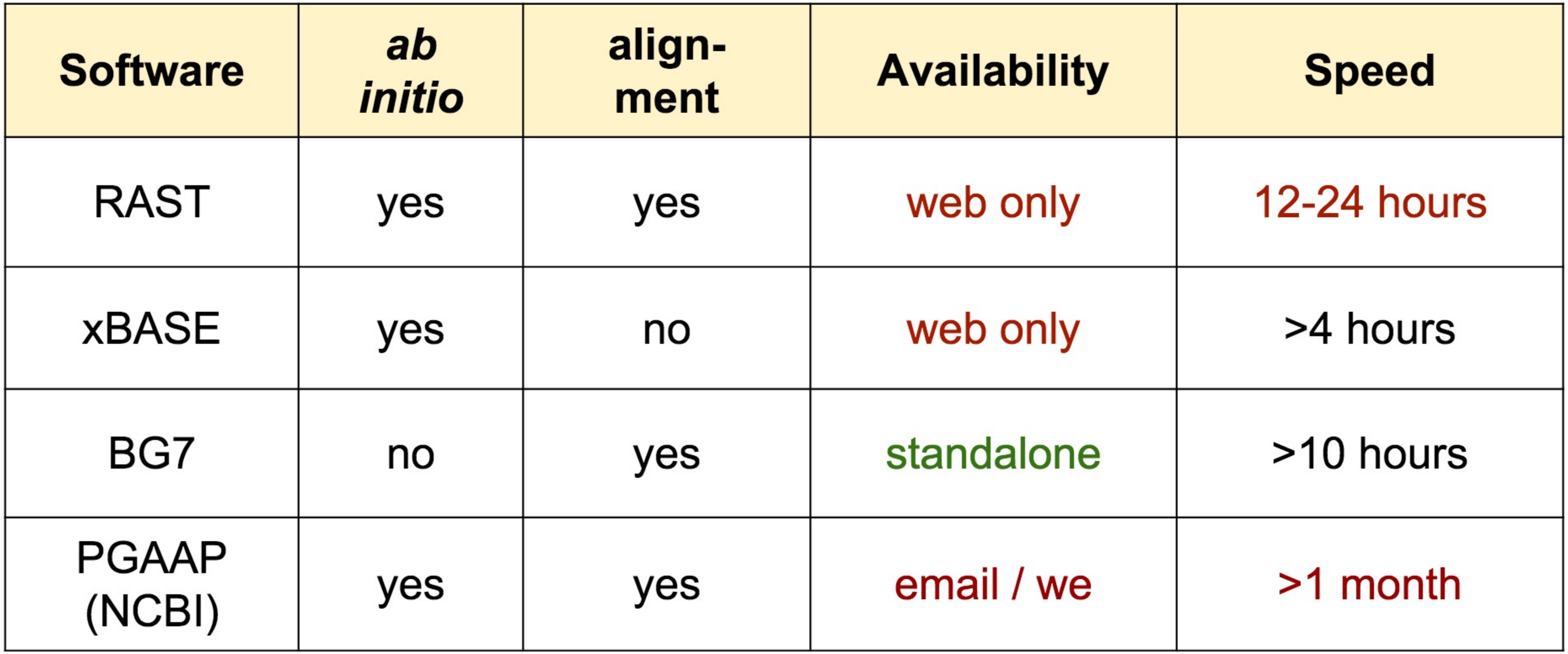
Automatic annotation

Two strategies for identifying coding genes:

- **sequence alignment**
  - find known protein sequences in the contigs
    - transfer the annotation across
  - will miss proteins not in your database
  - may miss partial proteins

- ***ab initio* gene finding**
  - find candidate open reading frames
    - build model of ribosome binding sites
    - predict coding regions
  - may choose the incorrect start codon
  - may miss atypical genes, overpredict small genes

Some good existing tools



Software	*ab initio*	align-ment	Availability	Speed
RAST	yes	yes	web only	12-24 hours
xBASE	yes	no	web only	>4 hours
BG7	no	yes	standalone	>10 hours
PGAAP (NCBI)	yes	yes	email / we	>1 month

Seemann T et al. *Bacterial genome annotation,* **presentation 2016**

Prokka

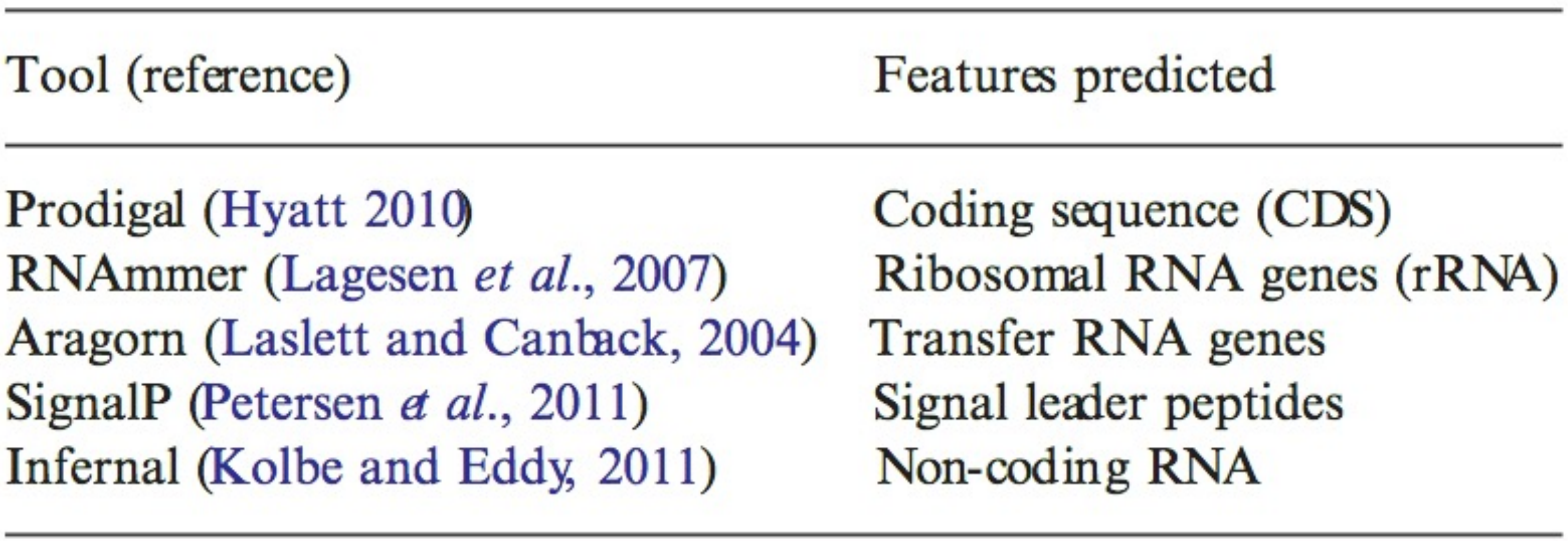
- Fast
  - exploits multi-core computers (aim < 15min)
- Convenient
  - Does structural and functional annotation in one go
  - Help submitting to NCBI and ENA
- Standards compliant
  - GFF3/GBK for viewing, TBL/FSA for Genbank.
- Provenance
  - Keep record of where/how/why it was annotated
- Also annotates archaea, mitochondria, and viruses

https://github.com/tseemann/prokka

Prokka

- Complicated to install
  - many dependencies (available on conda)

Feature prediction tools used by Prokka :



Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen *et al*., 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen *et al*., 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA

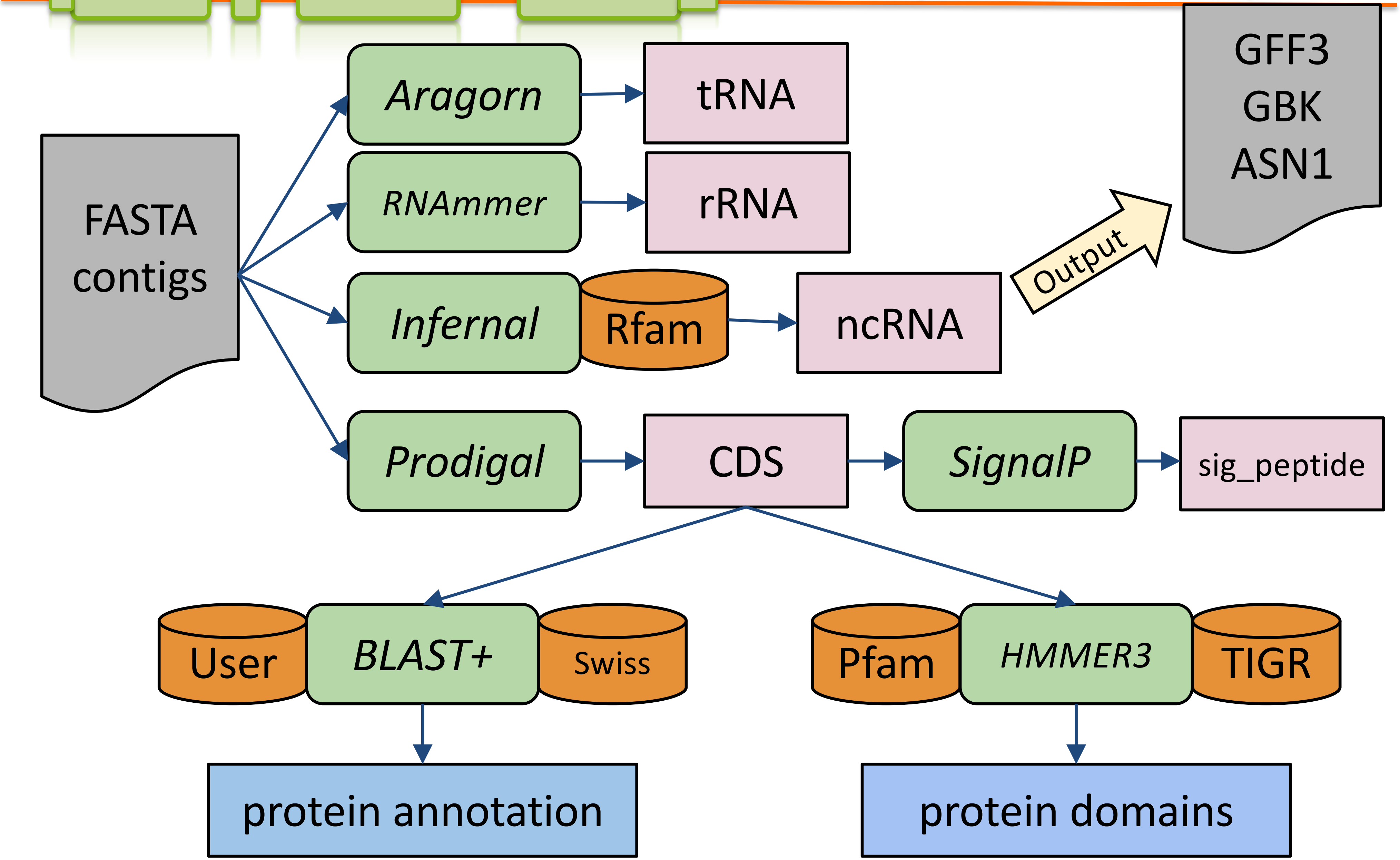
Seemann T. *Prokka: rapid prokaryotic genome annotation.* **Bioinformatics**. 2014 Jul 15;30(14):2068-9. PMID:24642063

Prokka : method

- Prodigal identifies the coordinates of candidates genes

- Compares with a database of known sequences
  - Small trustworthy database: the user provides a set of annotation proteins (optional)
  - Genus-specific proteome (optional)
  - Medium-size domain specific database: Uniprot-Swissprot
  - Curated model of protein families: all proteins from finished bacterial genomes in Refseq
  - HMMs profile: Pfam, TIGRFAMS (with HMMER)
  - If nothing is found, is labeled as ‘hypothetical protein’

Prokka pipeline (simplified)

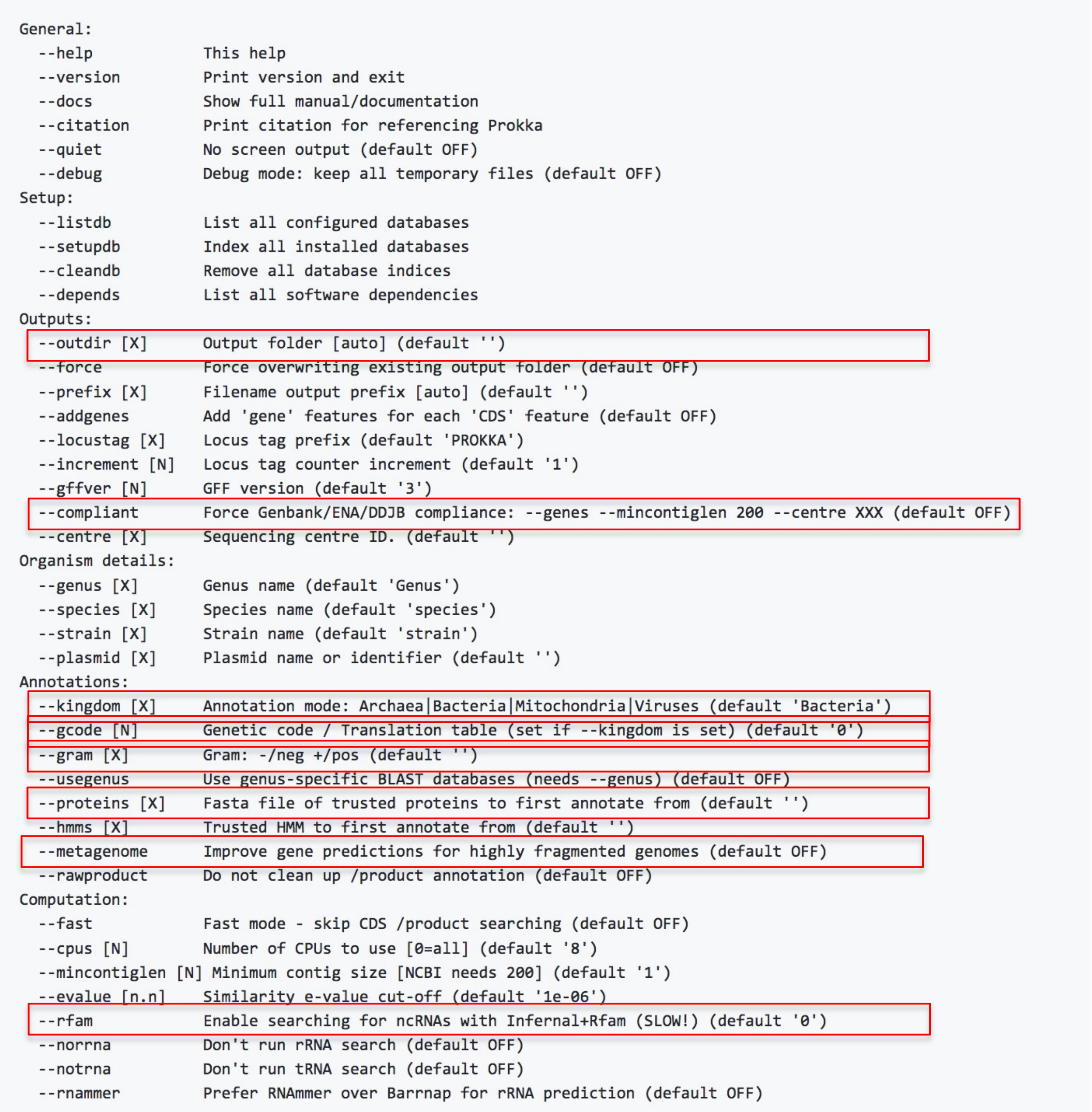


Seemann T et al. *Bacterial genome annotation,* **presentation 2016**

Prokka options

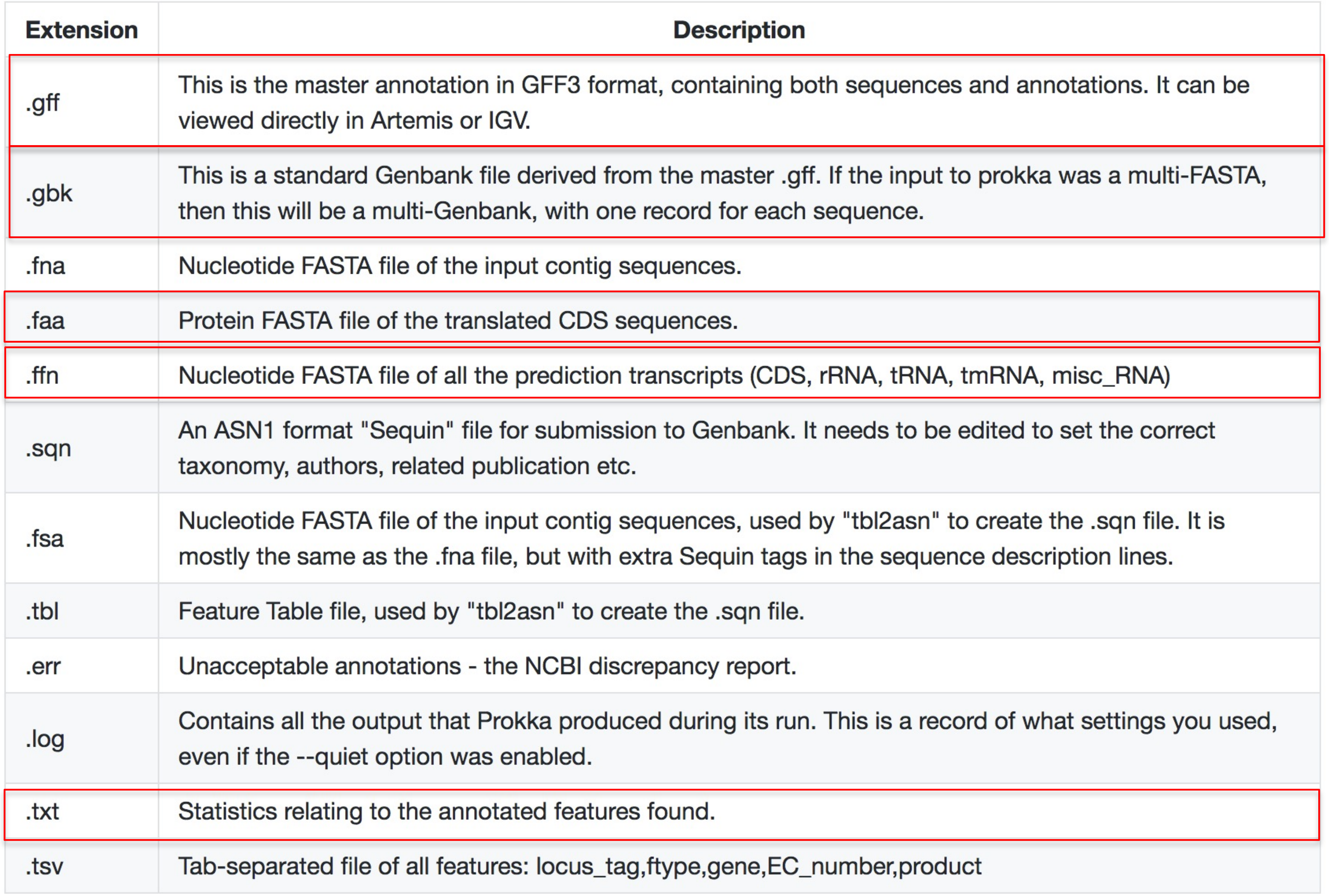
- Only one parameter mandatory :
  Input fasta format
  - prokka [options] <contigs.fasta>
- More than 30 different options available
  - prokka --help

Command line options



```
General:
  --help            This help
  --version         Print version and exit
  --docs            Show full manual/documentation
  --citation        Print citation for referencing Prokka
  --quiet           No screen output (default OFF)
  --debug           Debug mode: keep all temporary files (default OFF)
Setup:
  --listdb          List all configured databases
  --setupdb         Index all installed databases
  --cleandb         Remove all database indices
  --depends         List all software dependencies
Outputs:
  --outdir [X]      Output folder [auto] (default '')
  --force           Force overwriting existing output folder (default OFF)
  --prefix [X]      Filename output prefix [auto] (default '')
  --addgenes        Add 'gene' features for each 'CDS' feature (default OFF)
  --locustag [X]    Locus tag prefix (default 'PROKKA')
  --increment [N]   Locus tag counter increment (default '1')
  --gffver [N]      GFF version (default '3')
  --compliant       Force Genbank/ENA/DDJB compliance: --genes --mincontiglen 200 --centre XXX (default OFF)
  --centre [X]      Sequencing centre ID. (default '')
Organism details:
  --genus [X]       Genus name (default 'Genus')
  --species [X]     Species name (default 'species')
  --strain [X]      Strain name (default 'strain')
  --plasmid [X]     Plasmid name or identifier (default '')
Annotations:
  --kingdom [X]     Annotation mode: Archaea|Bacteria|Mitochondria|Viruses (default 'Bacteria')
  --gcode [N]       Genetic code / Translation table (set if --kingdom is set) (default '0')
  --gram [X]        Gram: -/neg +/pos (default '')
  --usegenus        Use genus-specific BLAST databases (needs --genus) (default OFF)
  --proteins [X]    Fasta file of trusted proteins to first annotate from (default '')
  --hmms [X]        Trusted HMM to first annotate from (default '')
  --metagenome      Improve gene predictions for highly fragmented genomes (default OFF)
  --rawproduct      Do not clean up /product annotation (default OFF)
Computation:
  --fast            Fast mode - skip CDS /product searching (default OFF)
  --cpus [N]        Number of CPUs to use [0=all] (default '8')
  --mincontiglen [N] Minimum contig size [NCBI needs 200] (default '1')
  --evalue [n.n]    Similarity e-value cut-off (default '1e-06')
  --rfam            Enable searching for ncRNAs with Infernal+Rfam (SLOW!) (default '0')
  --norrna          Don't run rRNA search (default OFF)
  --notrna          Don't run tRNA search (default OFF)
  --rnammer         Prefer RNAmmer over Barrnap for rRNA prediction (default OFF)
```

Prokka output

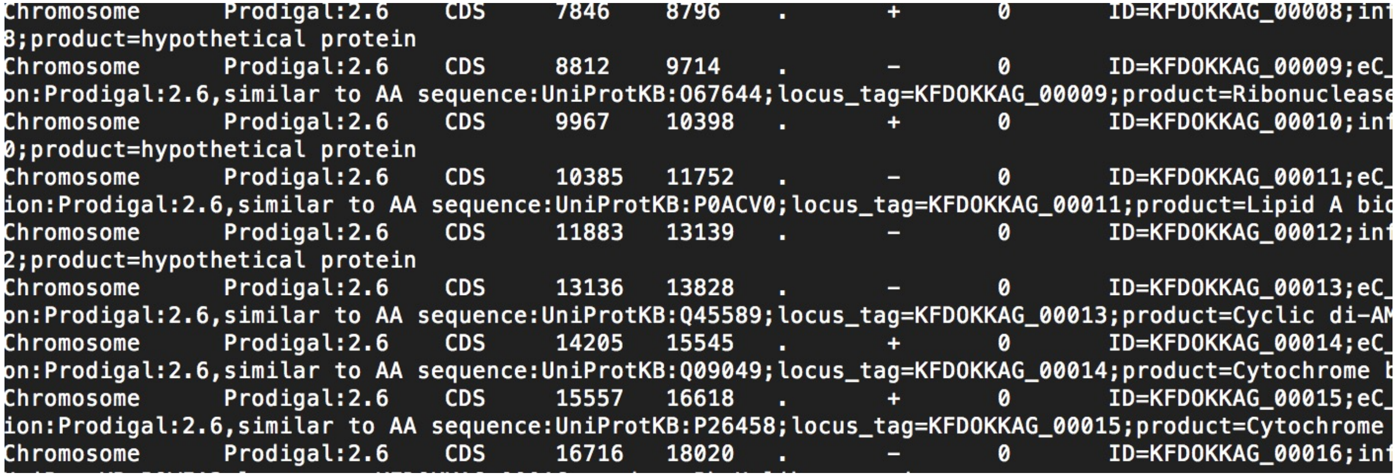


Extension	Description
.gff	This is the master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV.
.gbk	This is a standard Genbank file derived from the master .gff. If the input to prokka was a multi-FASTA, then this will be a multi-Genbank, with one record for each sequence.
.fna	Nucleotide FASTA file of the input contig sequences.
.faa	Protein FASTA file of the translated CDS sequences.
.ffn	Nucleotide FASTA file of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)
.sqn	An ASN1 format "Sequin" file for submission to Genbank. It needs to be edited to set the correct taxonomy, authors, related publication etc.
.fsa	Nucleotide FASTA file of the input contig sequences, used by "tbl2asn" to create the .sqn file. It is mostly the same as the .fna file, but with extra Sequin tags in the sequence description lines.
.tbl	Feature Table file, used by "tbl2asn" to create the .sqn file.
.err	Unacceptable annotations - the NCBI discrepancy report.
.log	Contains all the output that Prokka produced during its run. This is a record of what settings you used, even if the --quiet option was enabled.
.txt	Statistics relating to the annotated features found.
.tsv	Tab-separated file of all features: locus_tag,ftype,gene,EC_number,product

https://github.com/tseemann/prokka#output-files

Prokka output

GFF format



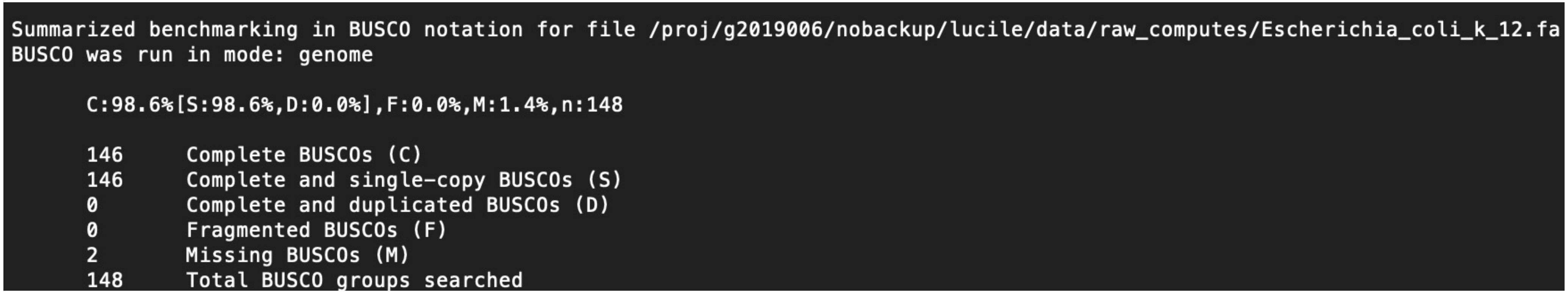
```
Chromosome      Prodigal:2.6    CDS     7846    8796    .       +       0       ID=KFDOKKAG_00008;inf
8;product=hypothetical protein
Chromosome      Prodigal:2.6    CDS     8812    9714    .       -       0       ID=KFDOKKAG_00009;eC_
on:Prodigal:2.6,similar to AA sequence:UniProtKB:O67644;locus_tag=KFDOKKAG_00009;product=Ribonuclease
Chromosome      Prodigal:2.6    CDS     9967    10398   .       +       0       ID=KFDOKKAG_00010;inf
0;product=hypothetical protein
Chromosome      Prodigal:2.6    CDS     10385   11752   .       -       0       ID=KFDOKKAG_00011;eC_
ion:Prodigal:2.6,similar to AA sequence:UniProtKB:P0ACV0;locus_tag=KFDOKKAG_00011;product=Lipid A bic
Chromosome      Prodigal:2.6    CDS     11883   13139   .       -       0       ID=KFDOKKAG_00012;inf
2;product=hypothetical protein
Chromosome      Prodigal:2.6    CDS     13136   13828   .       -       0       ID=KFDOKKAG_00013;eC_
on:Prodigal:2.6,similar to AA sequence:UniProtKB:Q45589;locus_tag=KFDOKKAG_00013;product=Cyclic di-AM
Chromosome      Prodigal:2.6    CDS     14205   15545   .       +       0       ID=KFDOKKAG_00014;eC_
on:Prodigal:2.6,similar to AA sequence:UniProtKB:Q09049;locus_tag=KFDOKKAG_00014;product=Cytochrome b
Chromosome      Prodigal:2.6    CDS     15557   16618   .       +       0       ID=KFDOKKAG_00015;eC_
ion:Prodigal:2.6,similar to AA sequence:UniProtKB:P26458;locus_tag=KFDOKKAG_00015;product=Cytochrome
Chromosome      Prodigal:2.6    CDS     16716   18020   .       -       0       ID=KFDOKKAG_00016;inf
```

Seqid	source	type	start	end	score	strand	phase	attributes
Chr1	Prodigal	exon	234	1543	.	+	.	gene_id “gene1”; transcript_id “transcript1”; “prediction:.., protein motif…”
Chr1	Snap	CDS	577	1543	.	+	0	gene_id “gene1”; transcript_id “transcript1”;

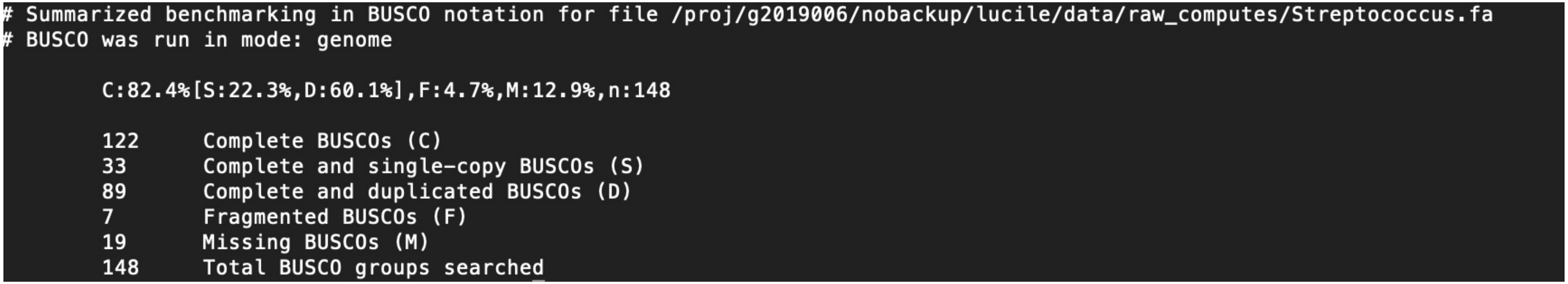
https://github.com/tseemann/prokka#output-files

Bacterial Genome Annotation

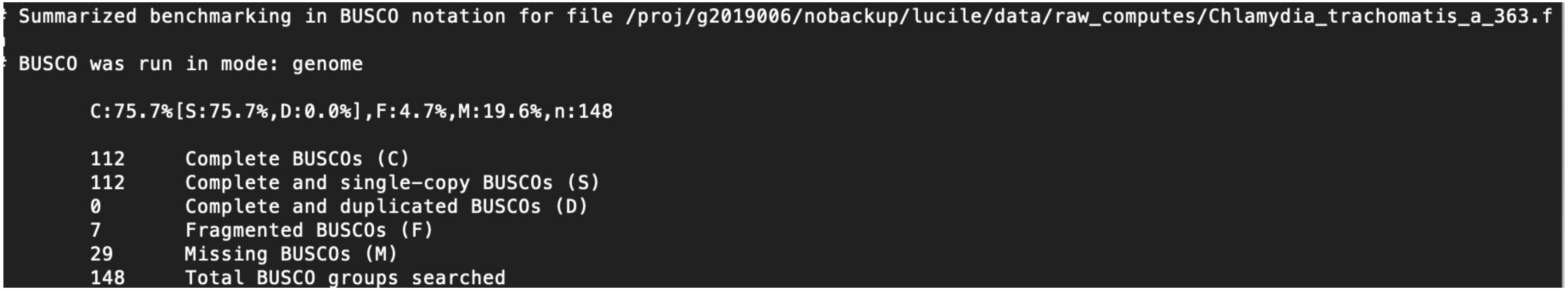
Exercises



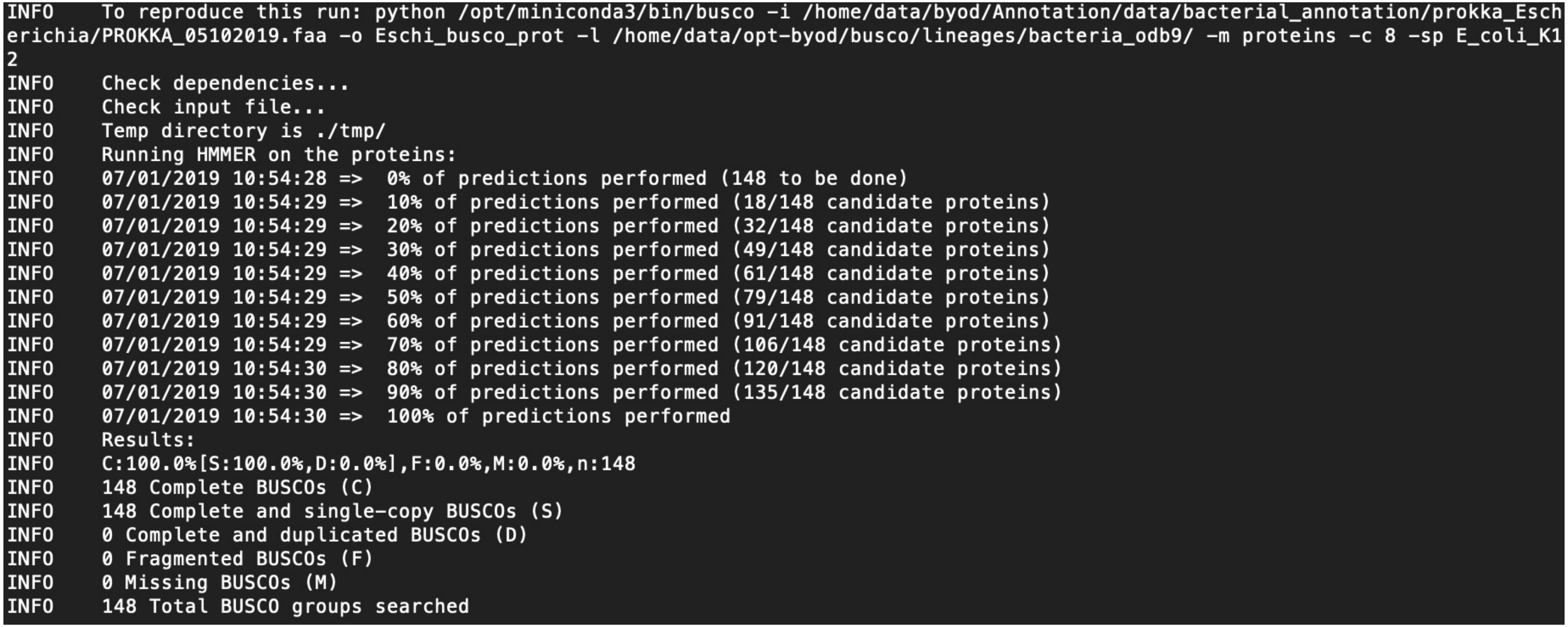
```
Summarized benchmarking in BUSCO notation for file /proj/g2019006/nobackup/lucile/data/raw_computes/Escherichia_coli_k_12.fa
BUSCO was run in mode: genome

      C:98.6%[S:98.6%,D:0.0%],F:0.0%,M:1.4%,n:148

      146     Complete BUSCOs (C)
      146     Complete and single-copy BUSCOs (S)
      0       Complete and duplicated BUSCOs (D)
      0       Fragmented BUSCOs (F)
      2       Missing BUSCOs (M)
      148     Total BUSCO groups searched
```



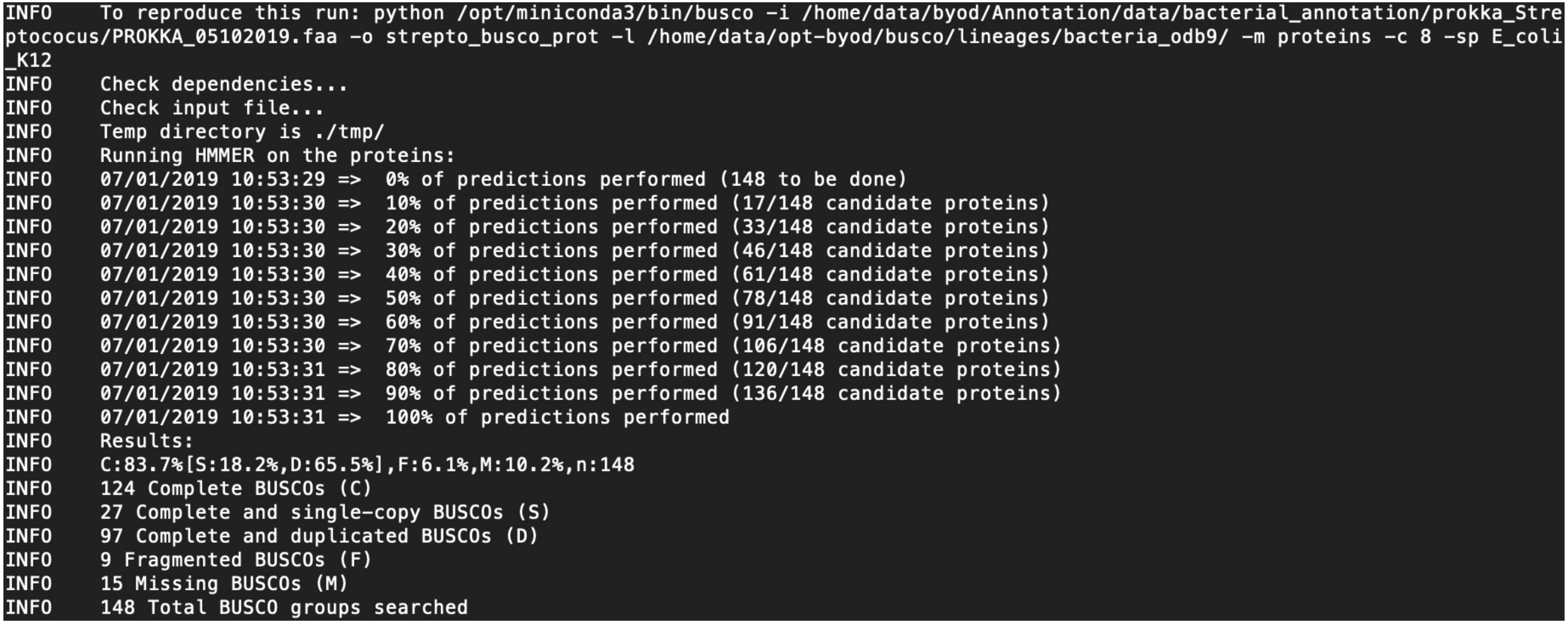
```
# Summarized benchmarking in BUSCO notation for file /proj/g2019006/nobackup/lucile/data/raw_computes/Streptococcus.fa
# BUSCO was run in mode: genome

        C:82.4%[S:22.3%,D:60.1%],F:4.7%,M:12.9%,n:148

        122     Complete BUSCOs (C)
        33      Complete and single-copy BUSCOs (S)
        89      Complete and duplicated BUSCOs (D)
        7       Fragmented BUSCOs (F)
        19      Missing BUSCOs (M)
        148     Total BUSCO groups searched
```



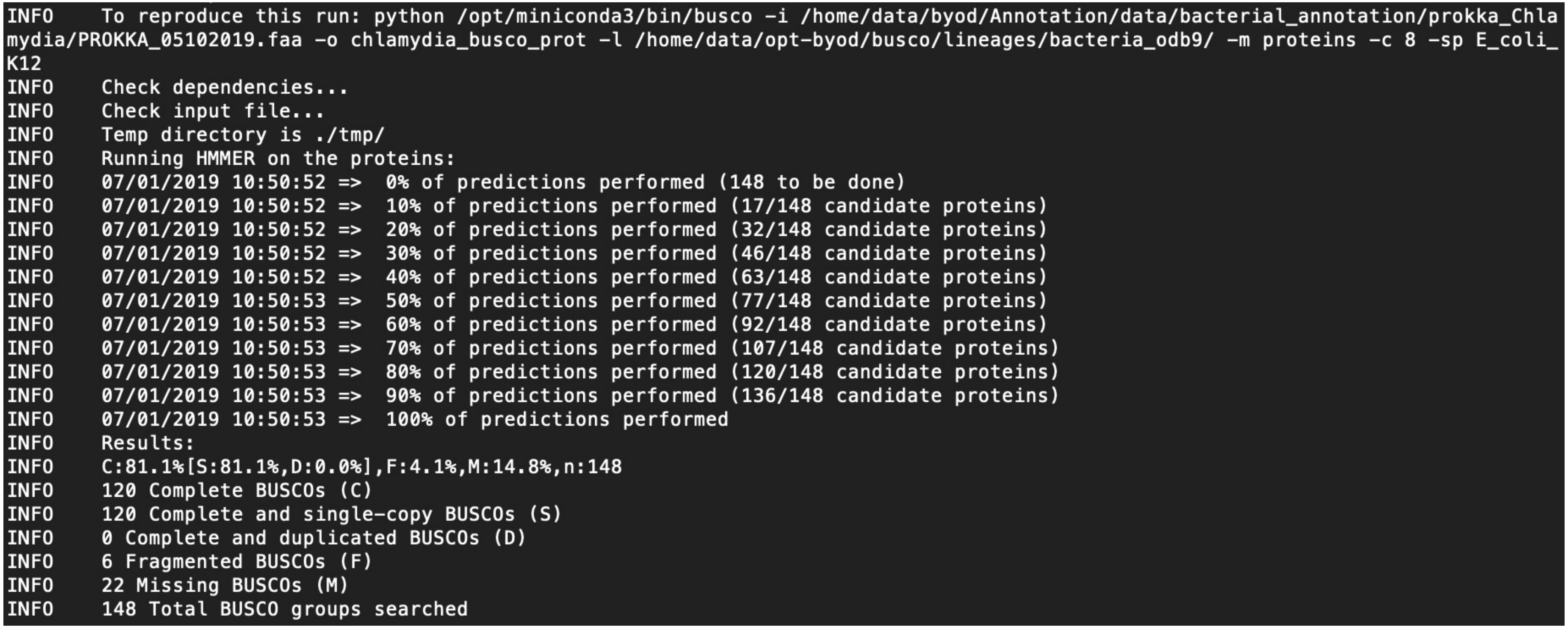
```
# Summarized benchmarking in BUSCO notation for file /proj/g2019006/nobackup/lucile/data/raw_computes/Chlamydia_trachomatis_a_363.f

# BUSCO was run in mode: genome

       C:75.7%[S:75.7%,D:0.0%],F:4.7%,M:19.6%,n:148

       112     Complete BUSCOs (C)
       112     Complete and single-copy BUSCOs (S)
       0       Complete and duplicated BUSCOs (D)
       7       Fragmented BUSCOs (F)
       29      Missing BUSCOs (M)
       148     Total BUSCO groups searched
```



```
INFO    To reproduce this run: python /opt/miniconda3/bin/busco -i /home/data/byod/Annotation/data/bacterial_annotation/prokka_Esch
erichia/PROKKA_05102019.faa -o Eschi_busco_prot -l /home/data/opt-byod/busco/lineages/bacteria_odb9/ -m proteins -c 8 -sp E_coli_K1
2
INFO    Check dependencies...
INFO    Check input file...
INFO    Temp directory is ./tmp/
INFO    Running HMMER on the proteins:
INFO    07/01/2019 10:54:28 =>  0% of predictions performed (148 to be done)
INFO    07/01/2019 10:54:29 =>  10% of predictions performed (18/148 candidate proteins)
INFO    07/01/2019 10:54:29 =>  20% of predictions performed (32/148 candidate proteins)
INFO    07/01/2019 10:54:29 =>  30% of predictions performed (49/148 candidate proteins)
INFO    07/01/2019 10:54:29 =>  40% of predictions performed (61/148 candidate proteins)
INFO    07/01/2019 10:54:29 =>  50% of predictions performed (79/148 candidate proteins)
INFO    07/01/2019 10:54:29 =>  60% of predictions performed (91/148 candidate proteins)
INFO    07/01/2019 10:54:29 =>  70% of predictions performed (106/148 candidate proteins)
INFO    07/01/2019 10:54:30 =>  80% of predictions performed (120/148 candidate proteins)
INFO    07/01/2019 10:54:30 =>  90% of predictions performed (135/148 candidate proteins)
INFO    07/01/2019 10:54:30 =>  100% of predictions performed
INFO    Results:
INFO    C:100.0%[S:100.0%,D:0.0%],F:0.0%,M:0.0%,n:148
INFO    148 Complete BUSCOs (C)
INFO    148 Complete and single-copy BUSCOs (S)
INFO    0 Complete and duplicated BUSCOs (D)
INFO    0 Fragmented BUSCOs (F)
INFO    0 Missing BUSCOs (M)
INFO    148 Total BUSCO groups searched
```



```
INFO    To reproduce this run: python /opt/miniconda3/bin/busco -i /home/data/byod/Annotation/data/bacterial_annotation/prokka_Stre
ptococus/PROKKA_05102019.faa -o strepto_busco_prot -l /home/data/opt-byod/busco/lineages/bacteria_odb9/ -m proteins -c 8 -sp E_coli
_K12
INFO    Check dependencies...
INFO    Check input file...
INFO    Temp directory is ./tmp/
INFO    Running HMMER on the proteins:
INFO    07/01/2019 10:53:29 =>  0% of predictions performed (148 to be done)
INFO    07/01/2019 10:53:30 =>  10% of predictions performed (17/148 candidate proteins)
INFO    07/01/2019 10:53:30 =>  20% of predictions performed (33/148 candidate proteins)
INFO    07/01/2019 10:53:30 =>  30% of predictions performed (46/148 candidate proteins)
INFO    07/01/2019 10:53:30 =>  40% of predictions performed (61/148 candidate proteins)
INFO    07/01/2019 10:53:30 =>  50% of predictions performed (78/148 candidate proteins)
INFO    07/01/2019 10:53:30 =>  60% of predictions performed (91/148 candidate proteins)
INFO    07/01/2019 10:53:30 =>  70% of predictions performed (106/148 candidate proteins)
INFO    07/01/2019 10:53:31 =>  80% of predictions performed (120/148 candidate proteins)
INFO    07/01/2019 10:53:31 =>  90% of predictions performed (136/148 candidate proteins)
INFO    07/01/2019 10:53:31 =>  100% of predictions performed
INFO    Results:
INFO    C:83.7%[S:18.2%,D:65.5%],F:6.1%,M:10.2%,n:148
INFO    124 Complete BUSCOs (C)
INFO    27 Complete and single-copy BUSCOs (S)
INFO    97 Complete and duplicated BUSCOs (D)
INFO    9 Fragmented BUSCOs (F)
INFO    15 Missing BUSCOs (M)
INFO    148 Total BUSCO groups searched
```



```
INFO    To reproduce this run: python /opt/miniconda3/bin/busco -i /home/data/byod/Annotation/data/bacterial_annotation/prokka_Chla
mydia/PROKKA_05102019.faa -o chlamydia_busco_prot -l /home/data/opt-byod/busco/lineages/bacteria_odb9/ -m proteins -c 8 -sp E_coli_
K12
INFO    Check dependencies...
INFO    Check input file...
INFO    Temp directory is ./tmp/
INFO    Running HMMER on the proteins:
INFO    07/01/2019 10:50:52 =>  0% of predictions performed (148 to be done)
INFO    07/01/2019 10:50:52 =>  10% of predictions performed (17/148 candidate proteins)
INFO    07/01/2019 10:50:52 =>  20% of predictions performed (32/148 candidate proteins)
INFO    07/01/2019 10:50:52 =>  30% of predictions performed (46/148 candidate proteins)
INFO    07/01/2019 10:50:52 =>  40% of predictions performed (63/148 candidate proteins)
INFO    07/01/2019 10:50:53 =>  50% of predictions performed (77/148 candidate proteins)
INFO    07/01/2019 10:50:53 =>  60% of predictions performed (92/148 candidate proteins)
INFO    07/01/2019 10:50:53 =>  70% of predictions performed (107/148 candidate proteins)
INFO    07/01/2019 10:50:53 =>  80% of predictions performed (120/148 candidate proteins)
INFO    07/01/2019 10:50:53 =>  90% of predictions performed (136/148 candidate proteins)
INFO    07/01/2019 10:50:53 =>  100% of predictions performed
INFO    Results:
INFO    C:81.1%[S:81.1%,D:0.0%],F:4.1%,M:14.8%,n:148
INFO    120 Complete BUSCOs (C)
INFO    120 Complete and single-copy BUSCOs (S)
INFO    0 Complete and duplicated BUSCOs (D)
INFO    6 Fragmented BUSCOs (F)
INFO    22 Missing BUSCOs (M)
INFO    148 Total BUSCO groups searched
```