NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Karolinska Institutet

KTH VETENSKAP OCH KONST

Stockholms universitet

UPPSALA UNIVERSITET

SciLifeLab

Bacterial Genome Annotation

Lucile Soler

Annotation course 9th-11th may 2017

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,
  - 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
  - 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, operons, 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
BG7	no	yes	standalone	>10 hours
PGAAP (NCBI)	yes	yes	email / we	>1 month

Seemann T. *Prokka: rapid prokaryotic genome annotation,* **presentation 2013**

Prokka

- Fast
  - exploits multi-core computers (aim < 15min)
- Convenient
  - Does structural and functional annotation in one go
- Standards compliant
  - GFF3/GBK for viewing, TBL/FSA for Genbank.
- Also annotates Archaea, fungi, mitochondria, and viruses

Prokka

- Complicated to install
  - many dependencies

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

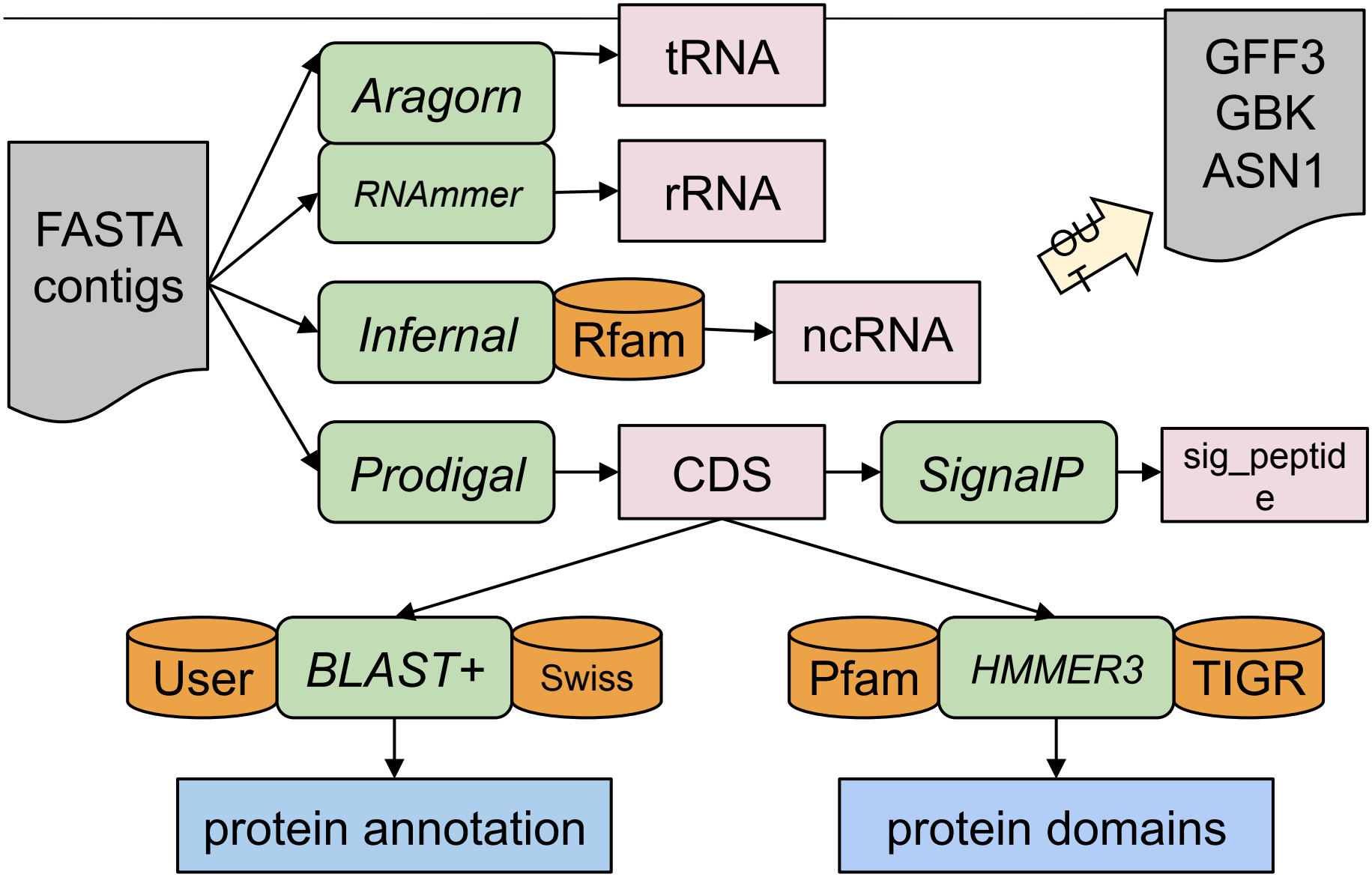
Seemann T. *Prokka: rapid prokaryotic genome annotation.* **Bioinformatics**. 2014 Jul 15;30(14):2068-9. PMID:24642063

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)
  - Medium-size domain specific database: Uniprot
  - Curated model of protein families: all proteins from finished bacterial genomes in Refseq
  - HMMs profile: Pfam, TIGRFAMS (with HMMER)
  - If nothing is found, label as ´hypothetical protein’

Prokka pipeline (simplified)



Seemann T. *Prokka: rapid prokaryotic genome annotation,* **presentation 2013**

Prokka options

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

Prokka output

Suffix	Description of file contents
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FASTA file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	Sequin editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

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

Practical 1

- Annotate 3 bacteria
- Use BUSCO to check genes completeness
- Use Prokka to annotate the assemblies