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 |

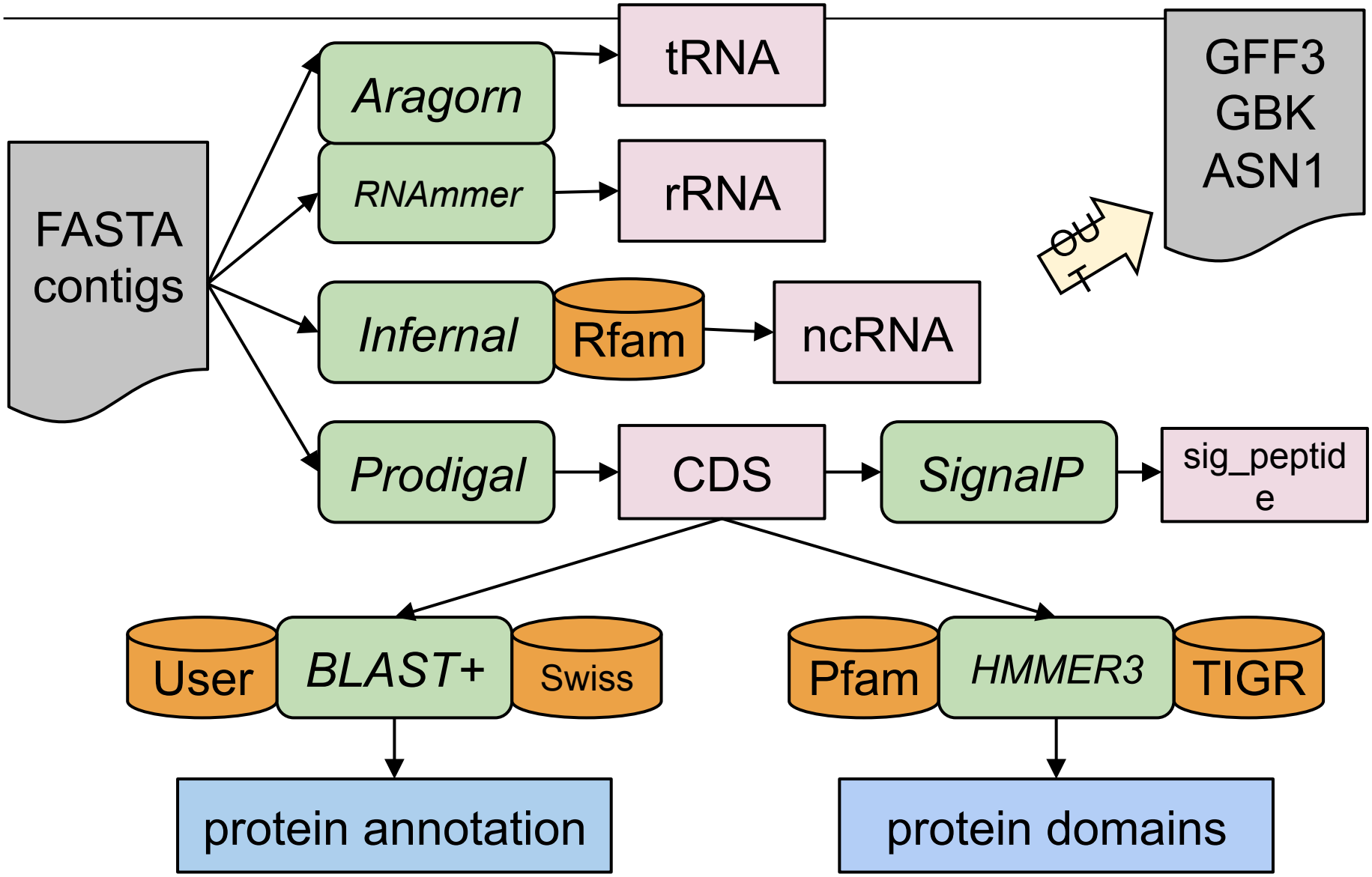
Seemann T. *Prokka: rapid prokaryotic genome annotation.* **Bioinformatics**. 2014 Jul 15;30(14):2068-9. PMID:24642063

# 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 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

# Practical 1

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