Genome annotation

Endrews Delbaje
29.03.2022

Genome annotation

The standardized identification and registry of functional elements in a genome sequence.

It requires:

- Identification of all potentially coding regions (CDS);
- Start and stop coordinates of the genes/structure in the genome;
- Associated function (or if the function is unknown).

Identification of coding regions

Finding ORFs (Open Reading Frames)

Example:

```
   ATGAGGTGACACCGCAAGCCTTATATTAGCTAA

 3 ATG AGG TGA CAC CGC AAG CCT TAT ATT AGC TAA
 2 A TGA GGT GAC ACC GCA AGC CTT ATA TTA GCT AA
 1 AT GAG GTG ACA CCG CAA GCC TTA TAT TAG CTA A

-1 TA CTC CAC TGT GGC GTT CGG AAT ATA ATC GAT T
-2 T ACT CCA CTG TGG CGT TCG GAA TAT AAT CGA TT
-3 TAC TCC ACT GTG GCG TTC GGA ATA TAA TCG ATT
```

Genome annotation formats

GFF (GFF3) (general feature format)

One line per feature and 9 columns

Example:

seqname	source	feature	start	end	score	strand	phase	attribute
scaffold1	prokka	CDS	12000	12980	.	+	.	Amoa
scaffold1	prokka	tRNA	13000	13082	.	-	.	tRNA-Leu

…

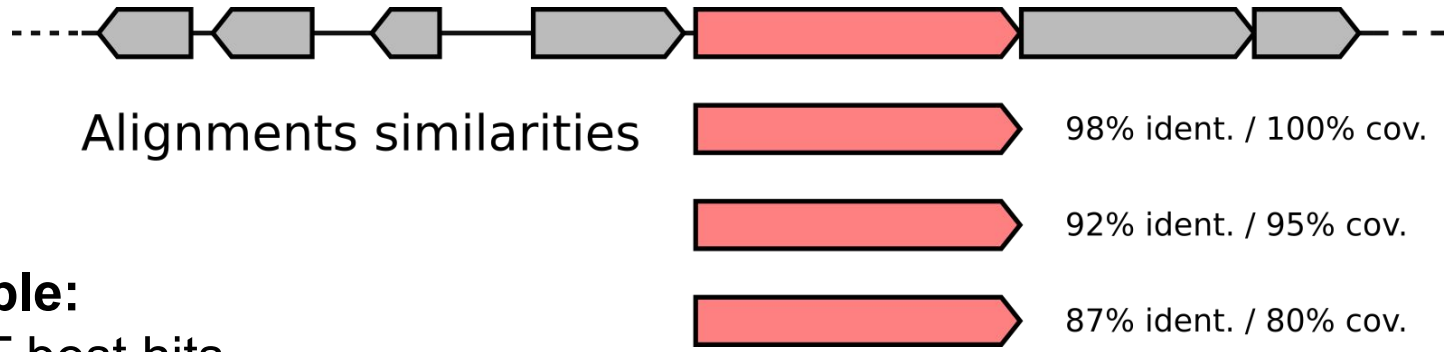
Other formats:

GBK

Tables

Function assignment - Database search

Functional assignment by homology using the best database hits:



Example:
BLAST best hits

Public available databases

NCBI (GenBank): Varied sequence community oriented database;
ENA: Varied sequence community oriented database;
KEGG: Gene database curated and organized for pathways;
Pfam: Protein database organized by protein families;
UniProt: Partially curated protein database;
…

Nowadays the process is automatized - Genome annotation by programs/platforms:

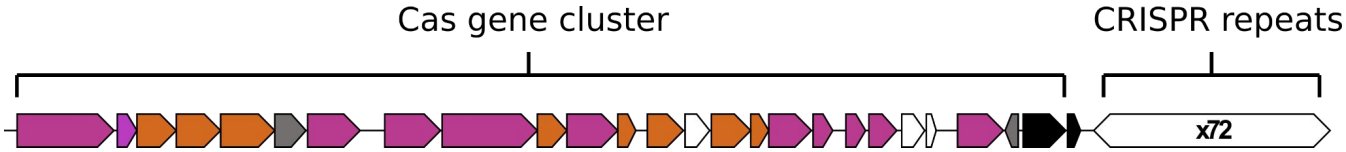
PROKKA
Genome Annotation Pipeline (PGAP)
EggNOG
…

Specialized annotation

Normally for complex regions or meta-features

Biosynthetic gene clusters (e.g. AntiSMASH program):

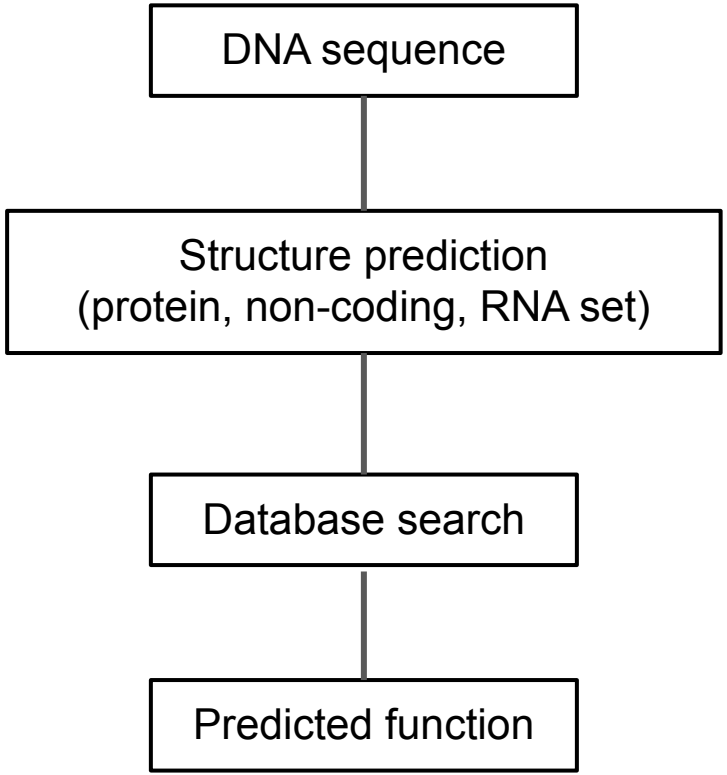
CRISPR/Cas (e.g. CRISPRone program):



Viral sequences, transposons, microRNA, etc.

Prokka: rapid prokaryotic genome annotation

Workflow:



Prodigal: ORF finding and translation;
Aragorn: tRNA;
Barrnap: rRNA.

Search with **BLAST+** and **HMMR3** in the databases:
ISfinder: transposases;
NCBI Bactarial antimicrobial;
UniProtKB: curated protein database.

PROKKA: Results summary

```
contigs: 1
bases: 4495168
CDS: 3873
gene: 3927
rRNA: 8
repeat_region: 9
tRNA: 45
tmRNA: 1
```

PROKKA: Annotation table (.tbl)

```
5776    5234    CDS
            EC_number   7.1.1.6
            db_xref COG:COG0723
            gene    petC_1
            inference   ab initio prediction:Prodigal:002006
            inference   similar to AA sequence:UniProtKB:P0C8N8
            locus_tag   GNOHDOCP_00006
            product Cytochrome b6-f complex iron-sulfur subunit
```