Genômica Computacional

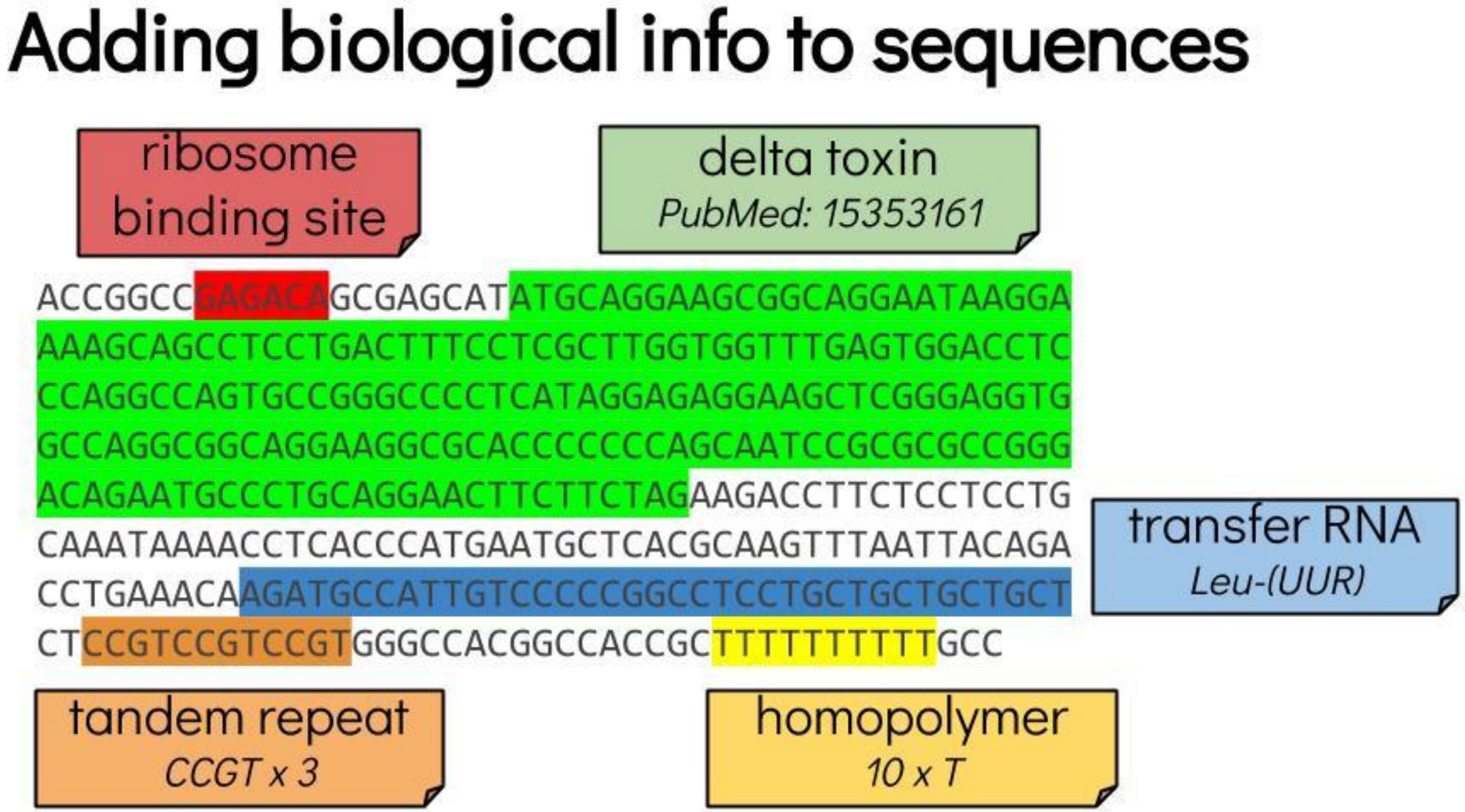
Anotação de genômas

Professor: Ricardo A. Vialle

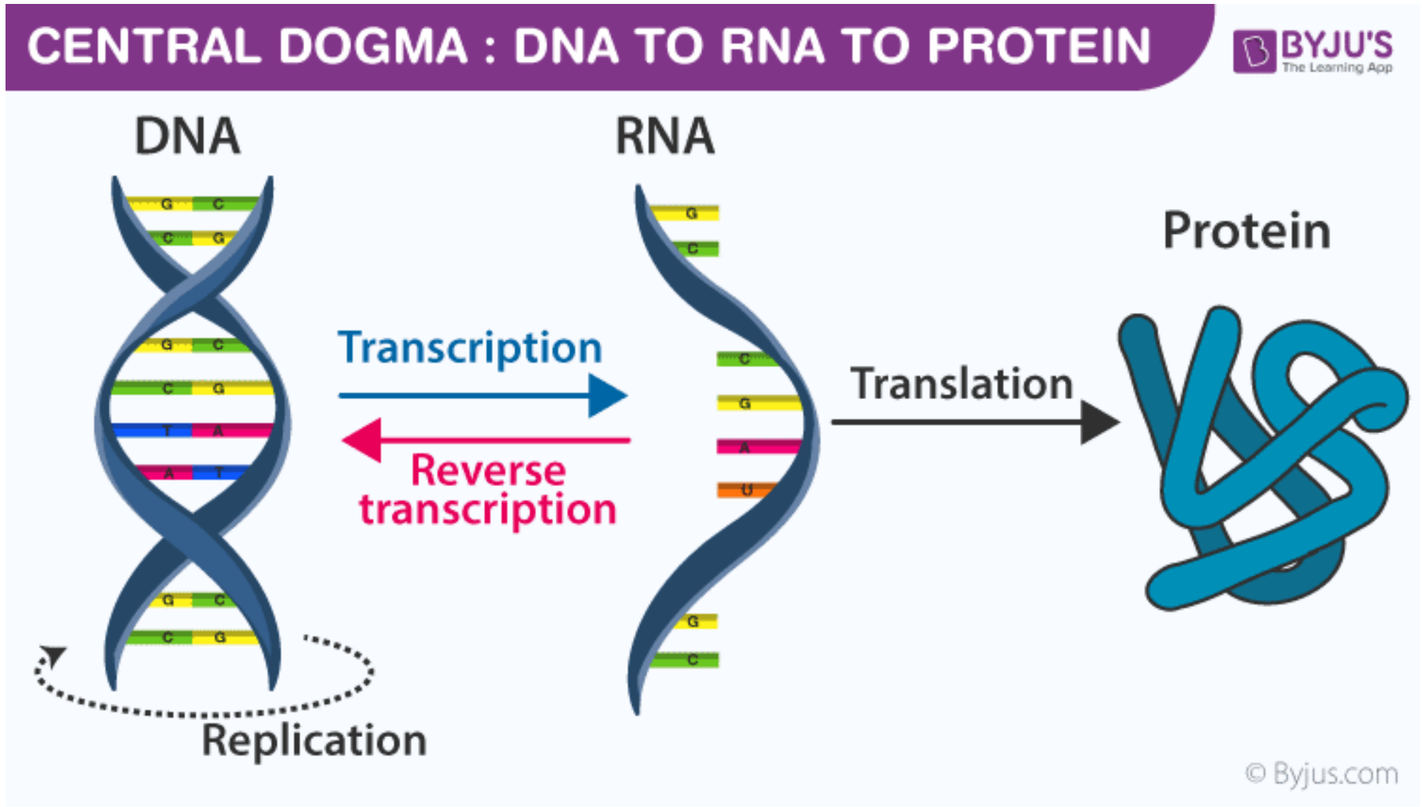
CS31 - Genômica Computacional

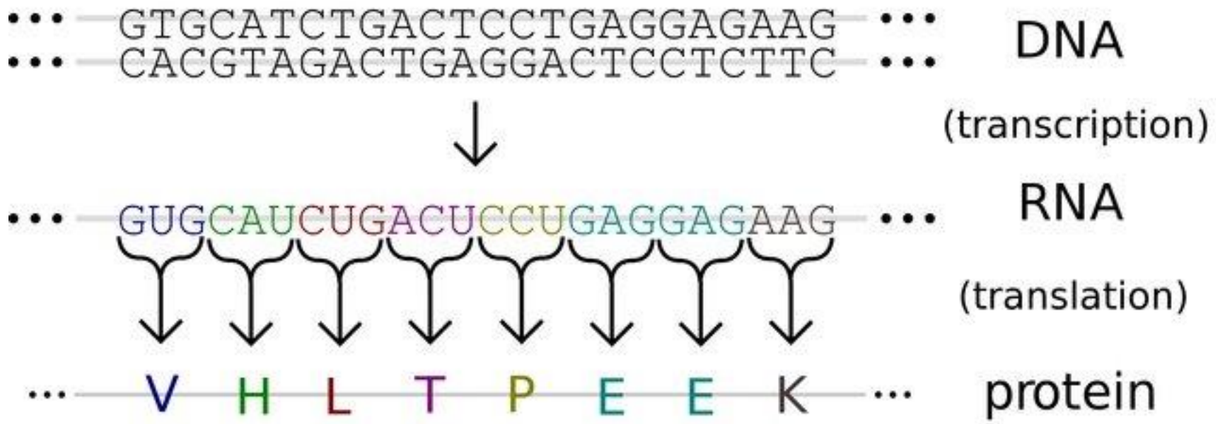
16 de Julho de 2025

Genome assembled, what next?



https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/slides-plain.html

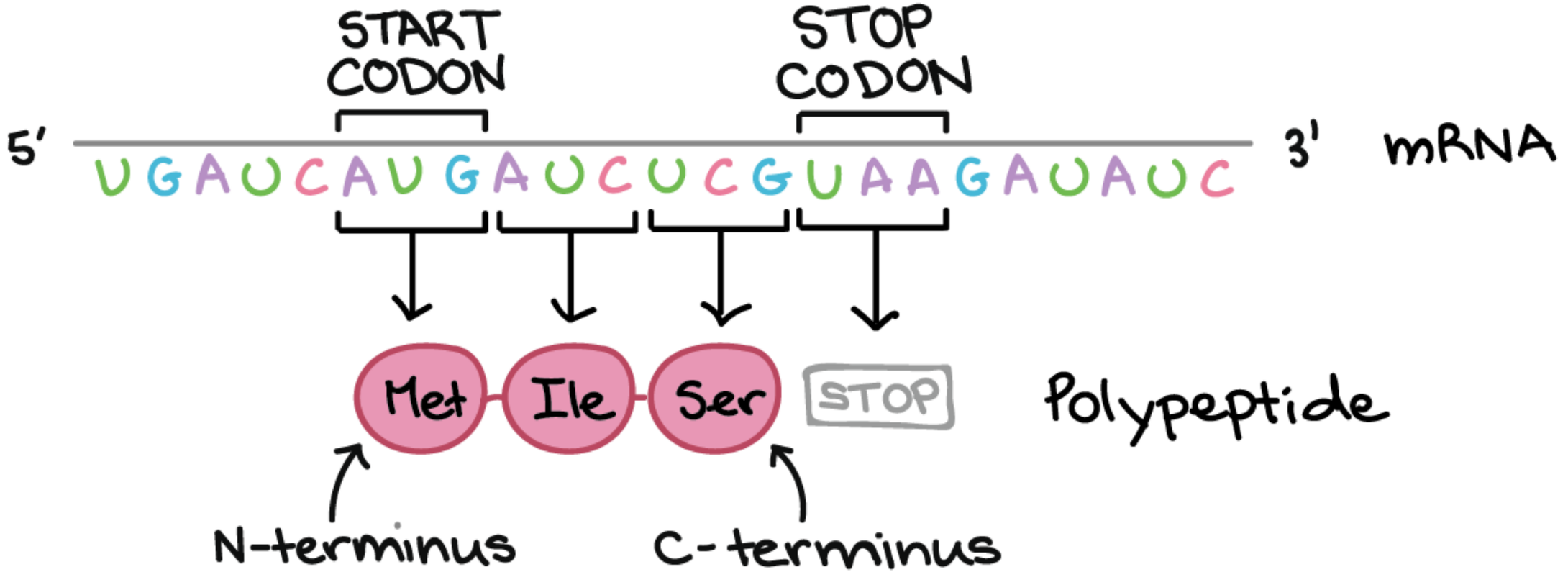




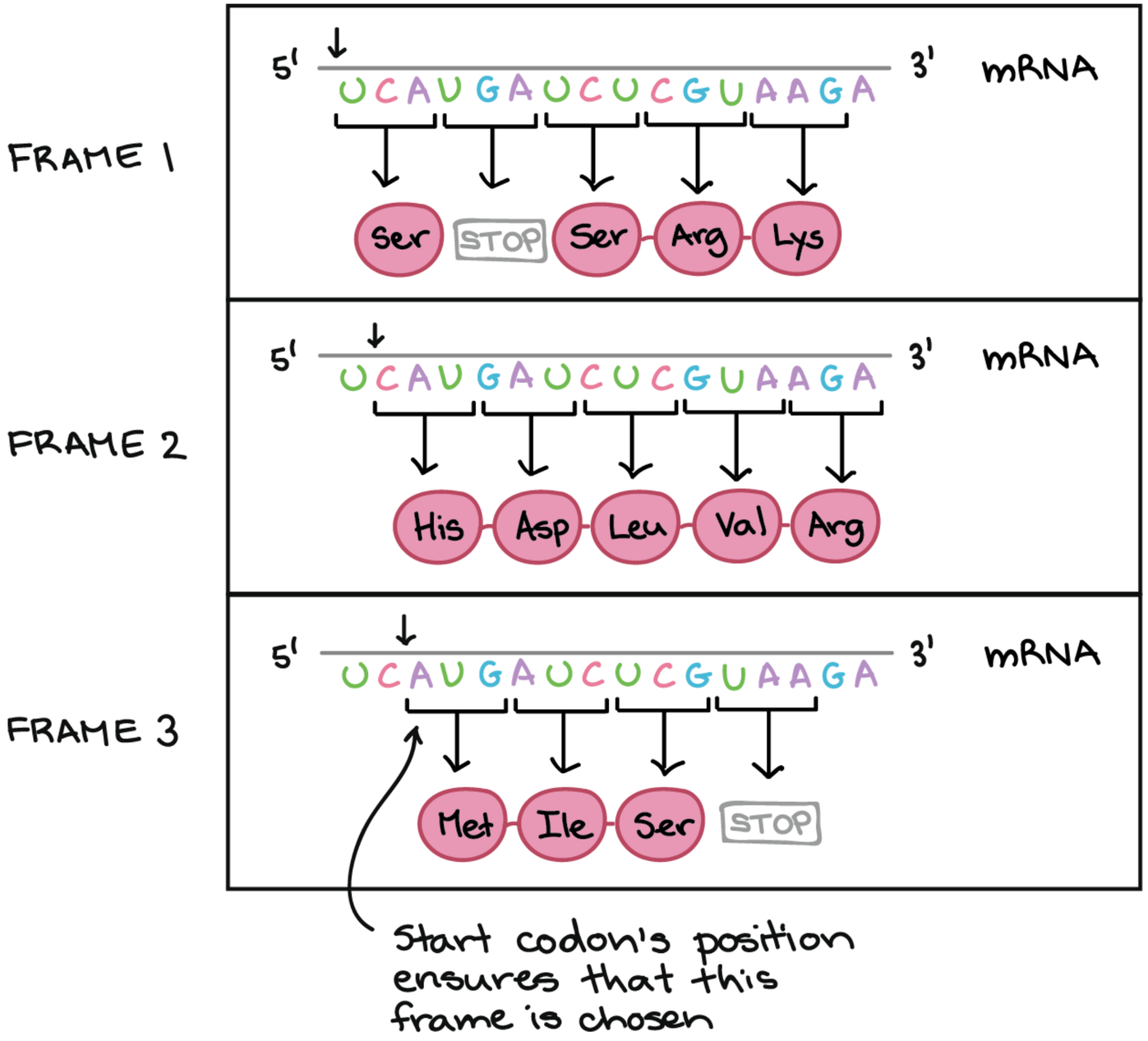
Computational Statistics for the Identification of Transcriptional Gene Regulatory Networks DOI: 10.13140/RG.2.1.4791.8567

Primera base	Segunda base: U	Segunda base: C	Segunda base: A	Segunda base: G	Tercera base
U	UUU, UUC: Fenilalanina (Fen) UUA, UUG: Leucina (Leu)	UCU, UCC, UCA, UCG: Serina (Ser)	UAU, UAC: Tirosina (Tir) UAA, UAG: Codón de terminación	UGU, UGC: Cisteína (Cis) UGA: Codón de terminación UGG: Triptófano (Tri)	U C A G
C	CUU, CUC, CUA, CUG: Leucina (Leu)	CCU, CCC, CCA, CCG: Prolina (Pro)	CAU, CAC: Histidina (His) CAA, CAG: Glutamina (Gln)	CGU, CGC, CGA, CGG: Arginina (Arg)	U C A G
A	AUU, AUC, AUA: Isoleucina (Ileu) AUG: Metionina (Met) Inicio	ACU, ACC, ACA, ACG: Treonina (Tre)	AAU, AAC: Asparagina (Asn) AAA, AAG: Lisina (Lis)	AGU, AGC: Serina (Ser) AGA, AGG: Arginina (Arg)	U C A G
G	GUU, GUC, GUA, GUG: Valina (Val)	GCU, GCC, GCA, GCG: Alanina (Ala)	GAU, GAC: Ácido aspártico (Asp) GAA, GAG: Ácido glutámico (Glu)	GGU, GGC, GGA, GGG: Glicina (Gli)	U C A G

https://e1.portalacademico.cch.unam.mx/alumno/biologia1/unidad2/sintesisdeproteinas/codigo-genetico

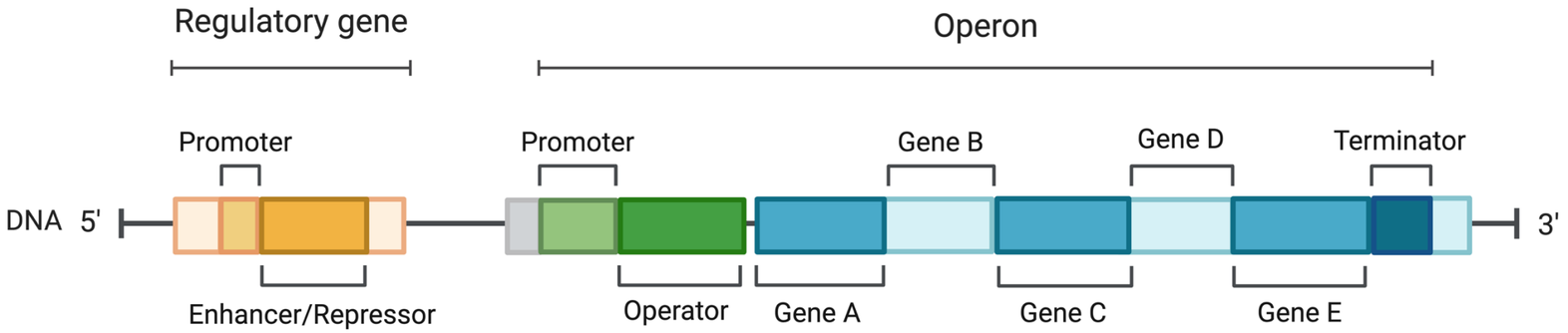


https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/translation/a/the-genetic-code-discovery-and-properties

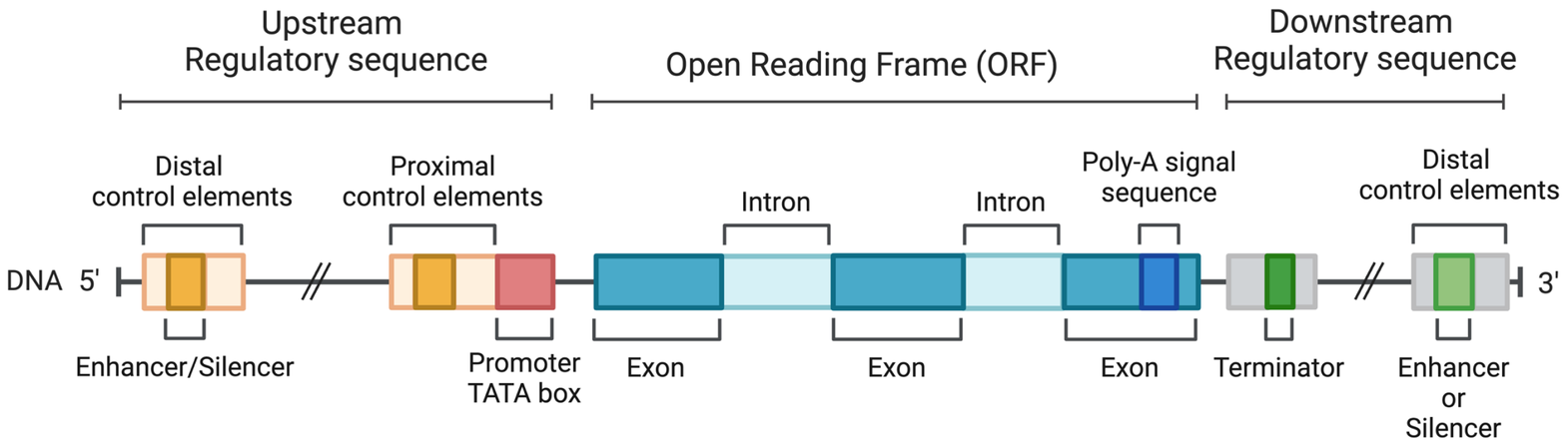


https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/translation/a/the-genetic-code-discovery-and-properties

Prokaryotic Gene Structure



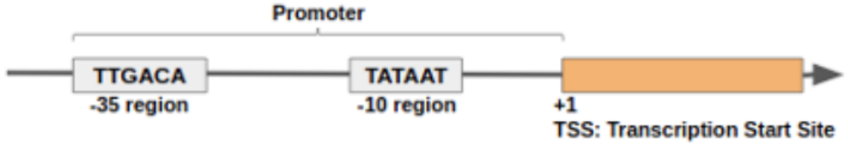
Eukaryotic Gene Structure



Prokaryotic Genes

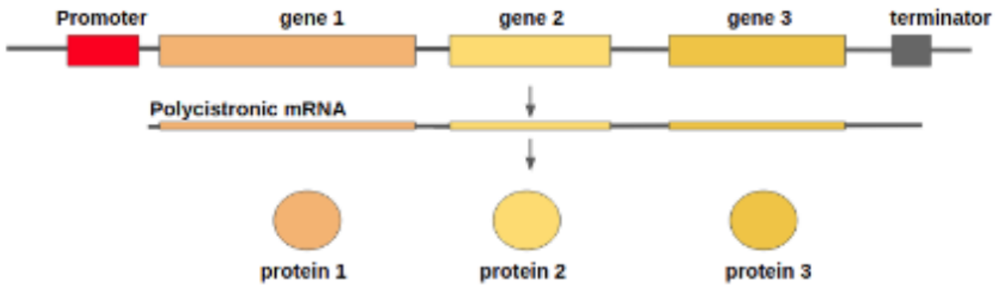
Promoter:

- -35 Region
- TATA Box
- Initiation site (TSS)

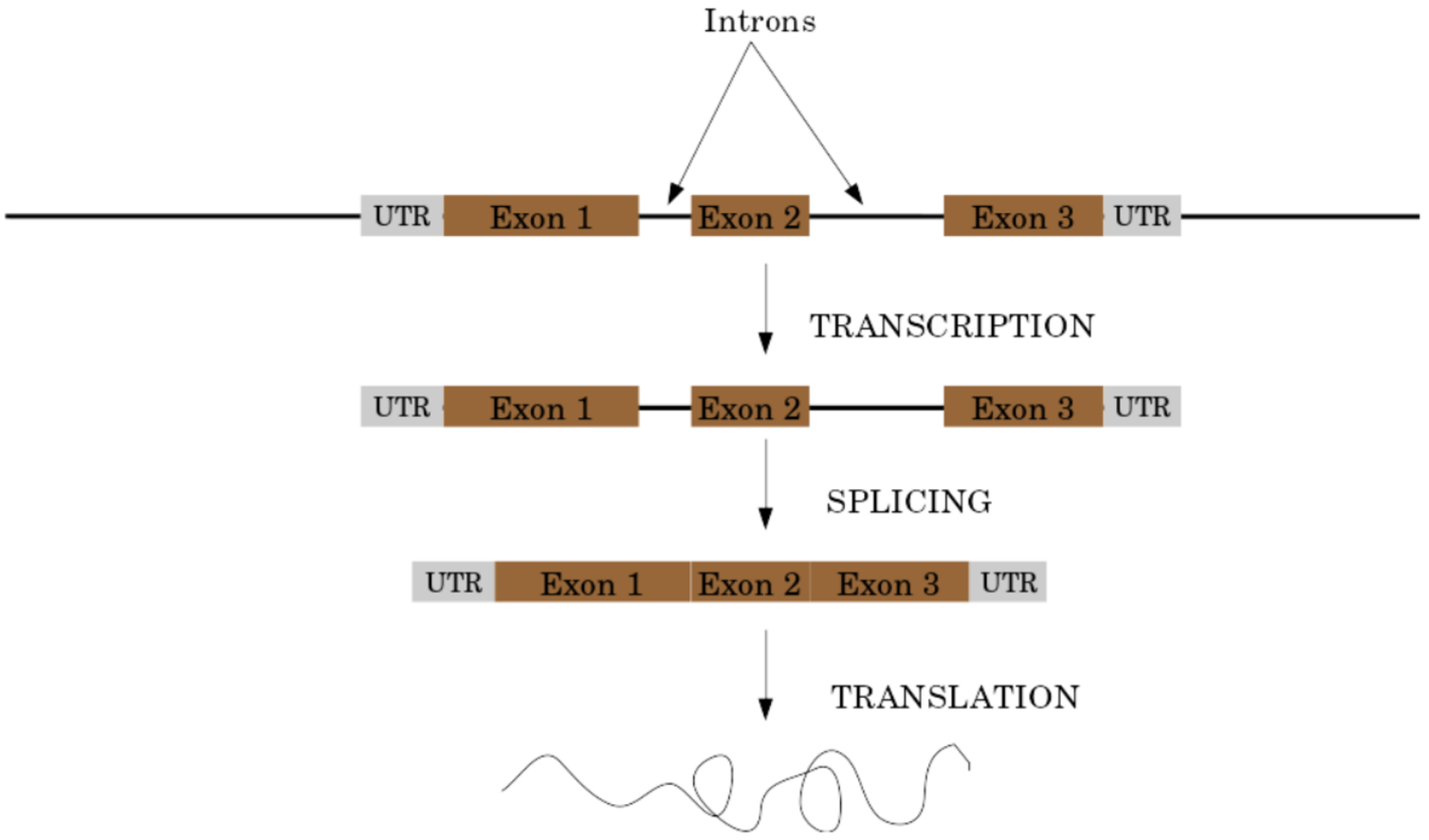


Operons:

- Promoter
- Some genes
- A terminator



Eukaryotic Genes

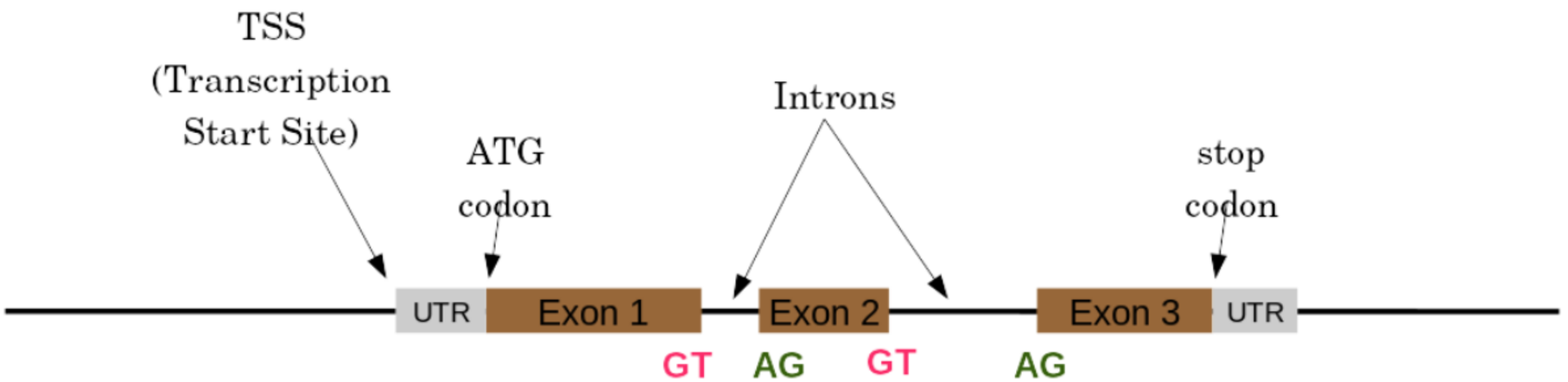


https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/slides-plain.html

Automatic Structural Annotation

Very difficult problem

- Short, variable, unspecific motifs
- Need data to support predictions



https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/slides-plain.html

Strategies for identifying coding genes - Structural annotation

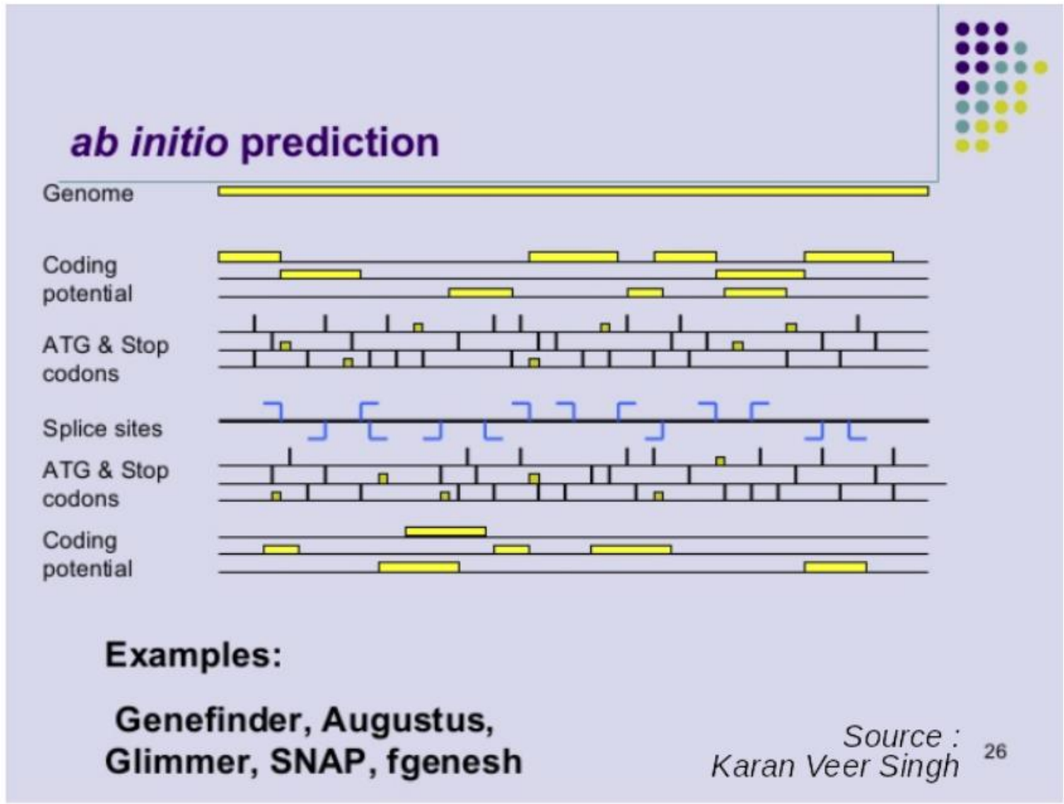
Ab initio Gene Calling

Predictions using:

- Genome sequence
- Statistical model (specific to organism)

Models:

- Training on the best evidence-based gene calls
- "Best" = strong evidence, highly conserved
- Training can be iterative:
  - train, predict, select best genes, retrain, etc



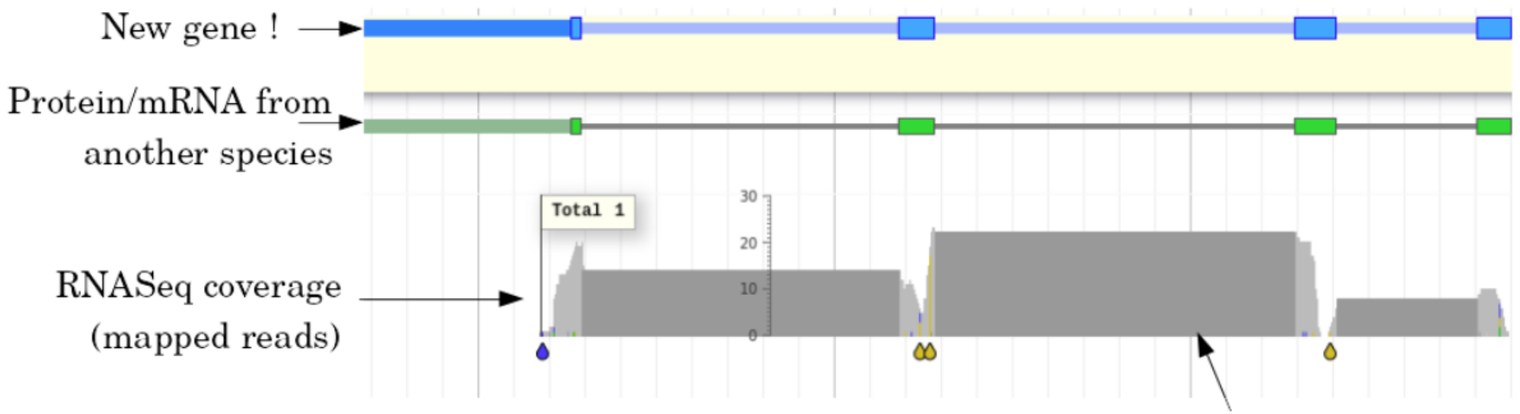
https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/slides-plain.html

Strategies for identifying coding genes - Structural annotation

Evidence

Multiple pieces of evidence

- Alignment of RNASeq reads
- Alignment of EST or transcripts (same species or closely related species)
- Alignment of proteins (closely related species)



But data unavailable for novel or very distant genes, or unexpressed genes

Strategies for identifying coding genes - Structural annotation

Alignment

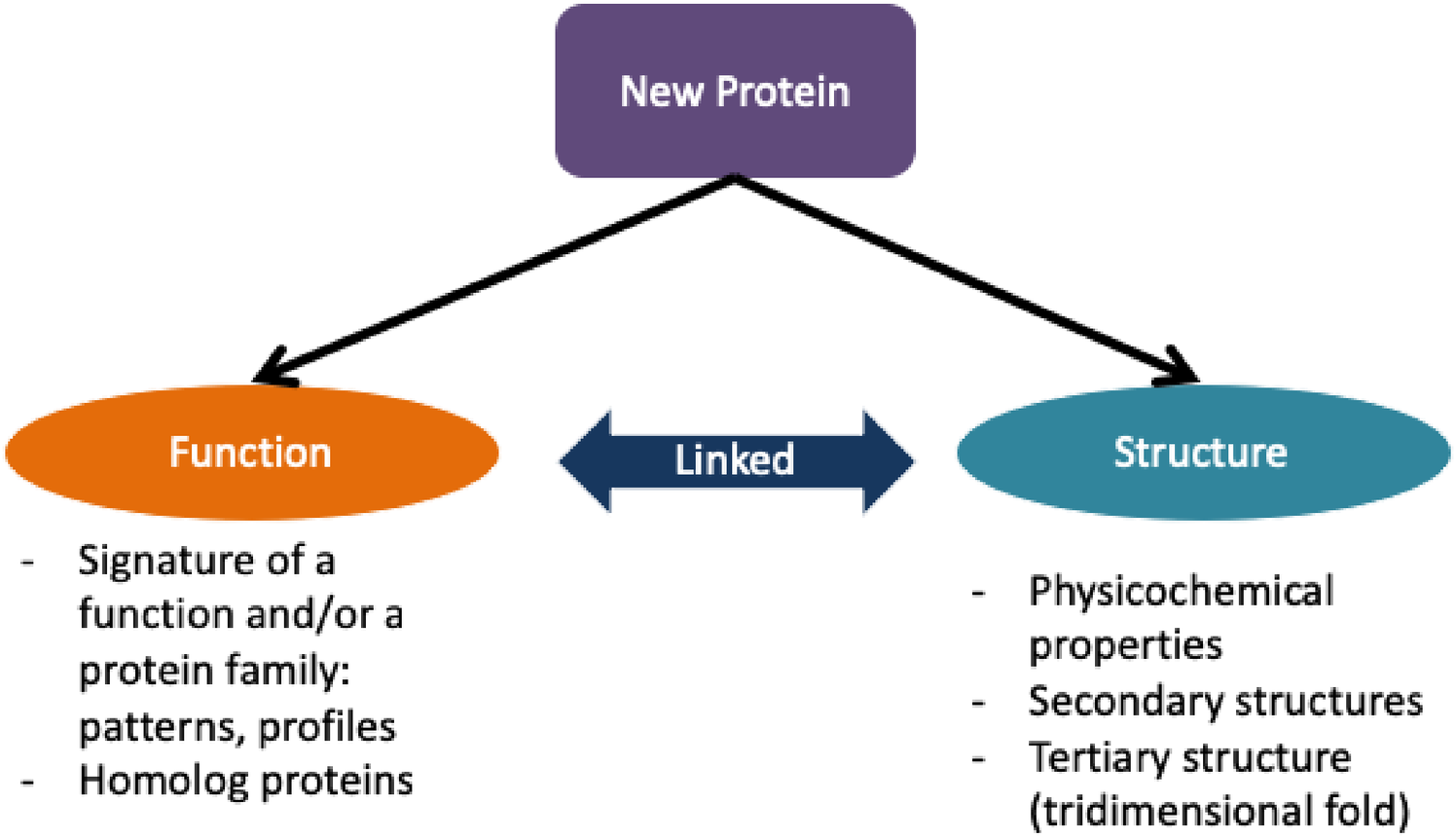
- Homology based
- Find similar sequences in DBs
- Might miss proteins that are not in the DB

Ab initio

- Find candidate ORFs
  - Model of RBS
  - Prediction of CDS
- May choose the incorrect start codon
- May miss atypical genes

https://www.hadriengourle.com/tutorials/annotation/

Functional annotation of the identified proteins



similar sequences => similar structures => similar functions

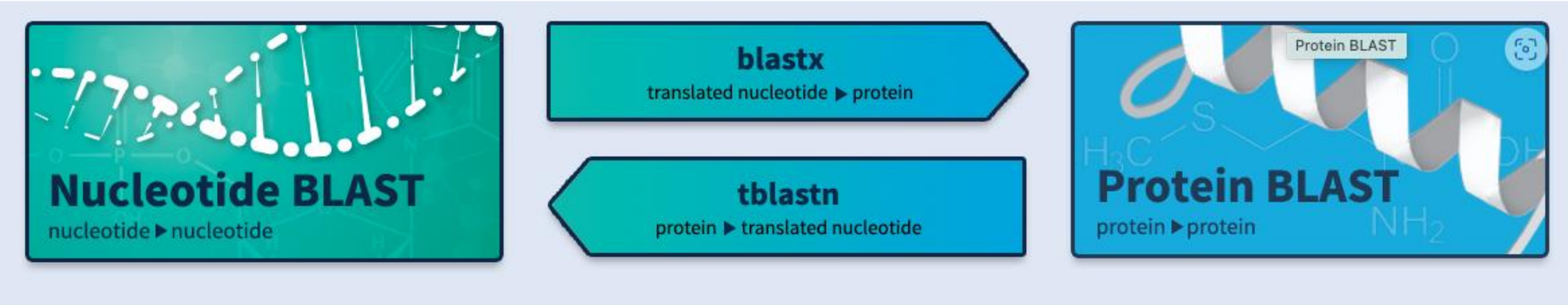
https://www.hadriengourle.com/tutorials/annotation/

Functional annotation

Methods

- Similarity search / homology
- Pattern search
- Orthologies
- Comparison against databases:
  - GenBank, NR: sequence databanks
  - InterPro: pattern databank (active sites, protein families, peptide signal ...)
  - EggNOG: databank of orthology relationships + functional annotation

NCBI Blast

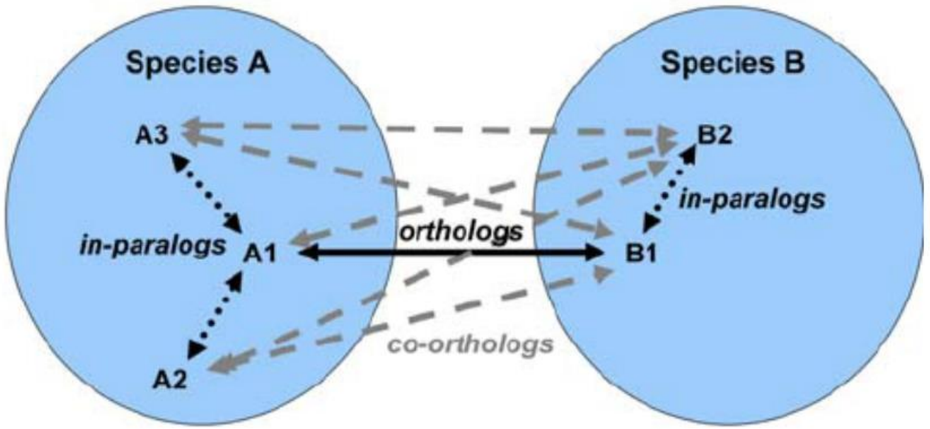


- Blast against **NR**
  - For each protein (or CDS) of the annotation
  - Find the best xx hits

- Huge database, good chances to have a match
- Risk:
  - Spread of "putative xx protein"
  - Spread of low-evidence annotations

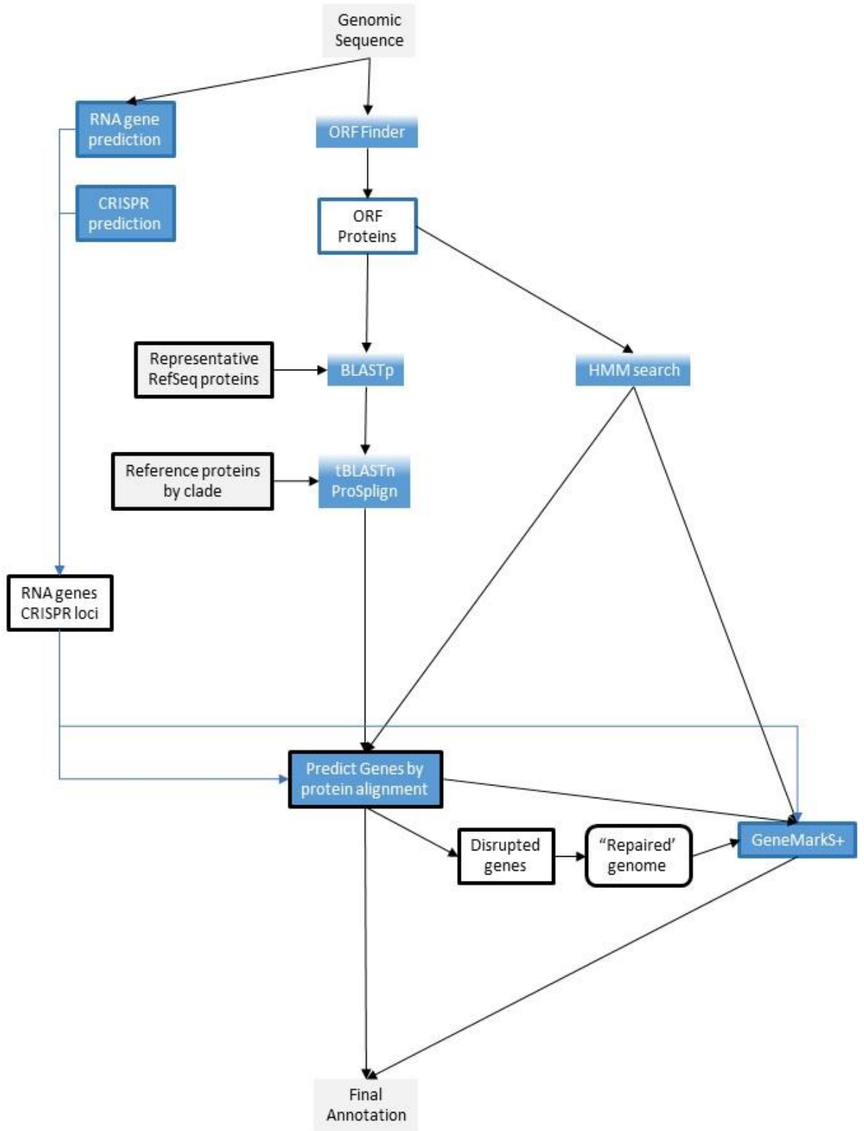
Functional annotation

- For each annotated gene
  - Search of orthologous genes in related species
  - Search for paralogues
- Bioinformatics method:
  - Blast all against all transcripts
  - Filtering the best hits
  - Clustering
  - OrthoFinder, OrthoMCL, ...



https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/slides-plain.html

Functional annotation



NCBI Prokaryotic Genome Annotation Pipeline

- Structural annotation by comparing ORFs to:
  - libraries of protein models (HMMs),
  - representative RefSeq proteins
  - proteins from well characterized reference genomes.
- GeneMark S+: *ab initio* coding region predictions for regions lacking HMM or protein evidence

https://www.hadriengourle.com/tutorials/annotation/

Prokka

- Fast
- Structural and Functional annotation
- Standards compliant
  - GFF3/GBK, TBL/FSA for GenBank
- Archaea, mitochondria, viruses

https://www.hadriengourle.com/tutorials/annotation/

Prokka

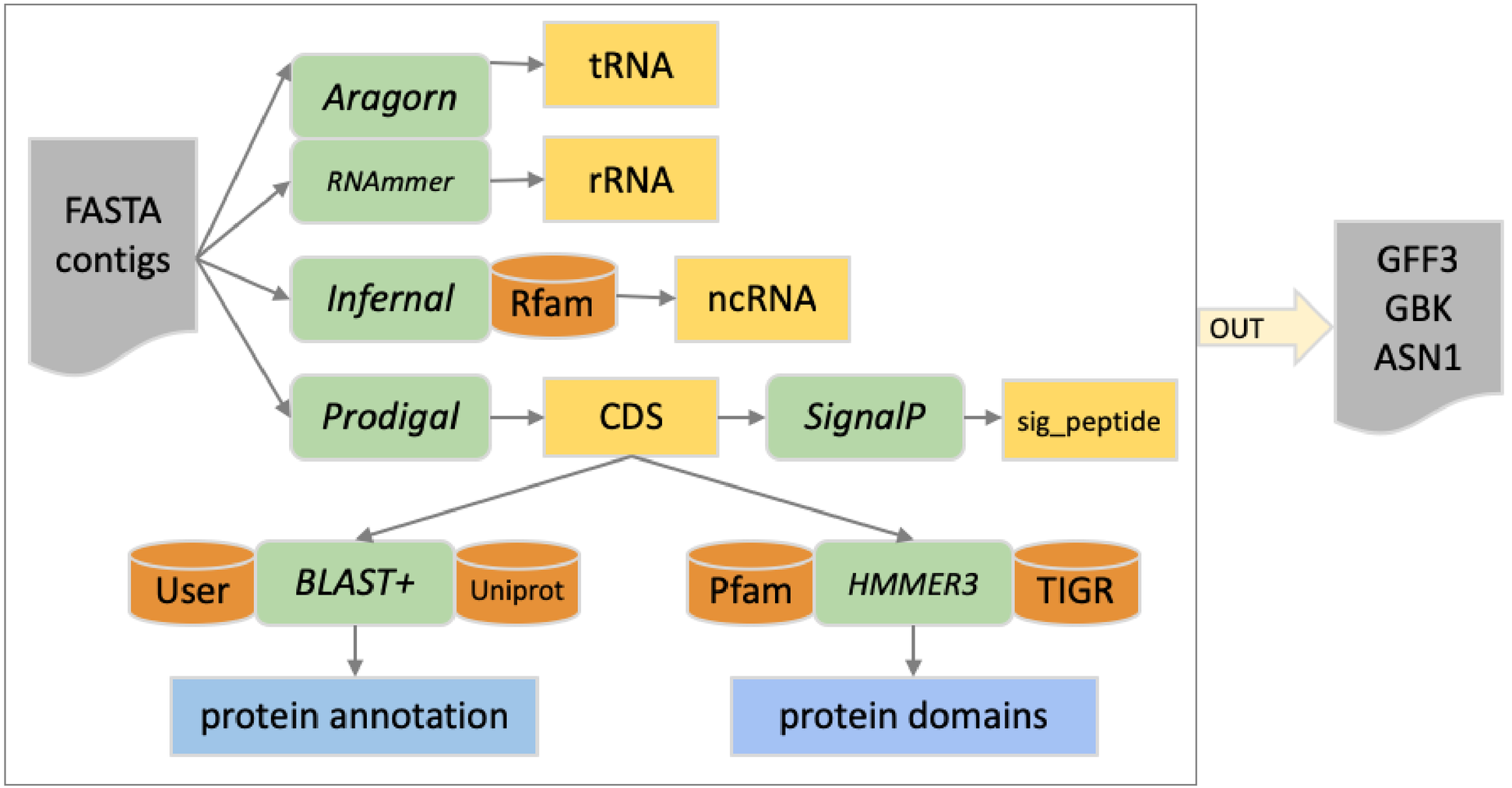
Seemann T. *Prokka: rapid prokaryotic genome annotation.* **Bioinformatics**. 2014 Jul 15;30(14):2068-9. PMID:24642063

Feature prediction tools used:

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

Prokka

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

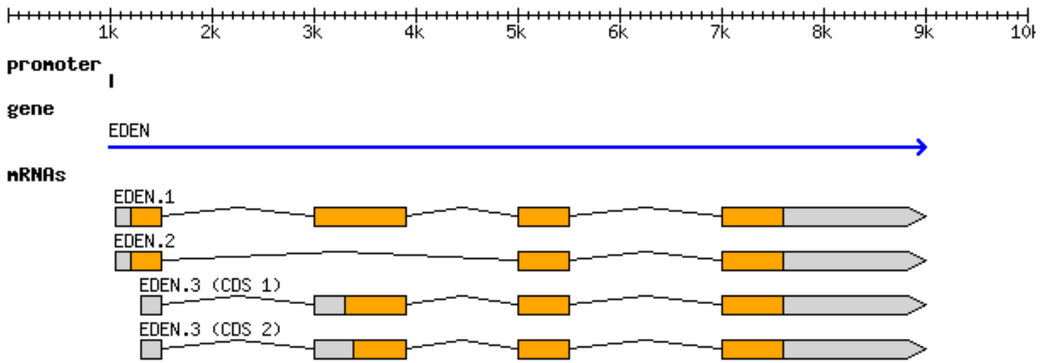


https://www.hadriengourle.com/tutorials/annotation/

GFF3

GFF3 - The GFF (General Feature Format) format consists of one line per feature, each containing **9** columns of data, plus optional track definition lines.

1.seqid - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. Important note: the seq ID must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
2.source - name of the program that generated this feature, or the data source (database or project name)
3.type - type of feature. Must be a term or accession from the SOFA sequence ontology
4.start - Start position of the feature, with sequence numbering starting at 1.
5.end - End position of the feature, with sequence numbering starting at 1.
6.score - A floating point value.
7.strand - defined as + (forward) or - (reverse).
8.phase - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
attributes - A semicolon-separated list of tag-value pairs, providing additional information about each feature. Some of these tags are predefined, e.g. ID, Name, Alias, Parent - see the GFF documentation for more details




```
0  ##gff-version 3.2.1
1  ##sequence-region ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16 ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19 ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21 ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22 ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23 ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24 ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md