

Methods in genome annotation



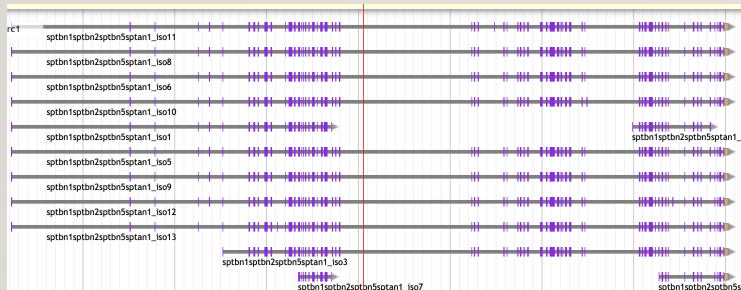
22/09/2025



Structural annotation:

Find out where the regions of interest
(usually genes) are in the sequence data
and what they look like.

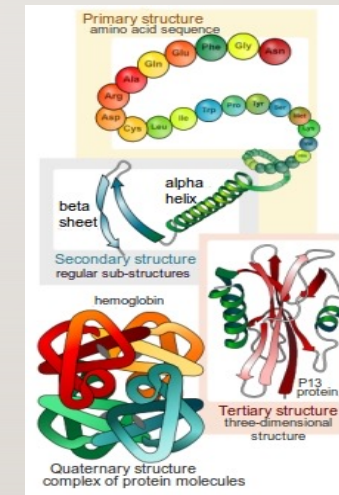
=> **Gene prediction / Gene Finding**



VS

functional annotation:

Find out what the regions do.
What do they code for?





Structural annotation + functional annotation + ...



Experimental

(ESTs, cDNAs, RNA-seq)

Isolate and clone cognate transcripts (as cDNA) sequence them and compare cDNA with genomic DNA

=> It's the **ONLY** secure method but:

- Cloning is time consuming.
- Lowly expressed genes are difficult to detect.
- ...

Predictive

computational methods

- **Intrinsic (Machine Learning)**

Use only the DNA sequence itself to predict genes based on statistical models of gene structure

- **Extrinsic**

Use **external biological evidence**, like RNA-Seq, ESTs, protein alignments, or known homologs, to guide annotation

- **Hybrid**

Modern methods often integrate intrinsic and extrinsic signals, blurring the line...

- Similarity-based methods :

These use similarity to annotated sequences like proteins and transcripts

- *Ab initio* prediction :

Machine learning based methods (probabilistic: Likelihood, Bayesian | deterministic: Neural networks)

- Hybrid approaches :

Ab initio tools with the ability to integrate external evidence/hints

- Comparative (homology) based gene finders :

These align genomic sequences from different species and use the alignments to guide the gene predictions

- Chooser / Combiner :

Selects among existing predictions according to evidence or confidence.

Constructs consensus gene models by integrating evidence from different sources.

- Pipelines :

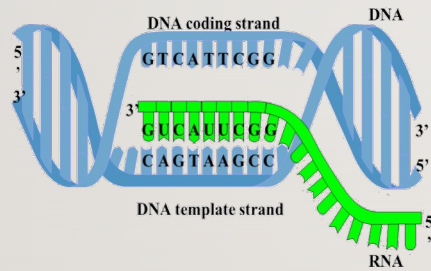
These combine multiple approaches, can annotate other types of features, work on functional annotation, etc.

Extrinsic

Intrinsic

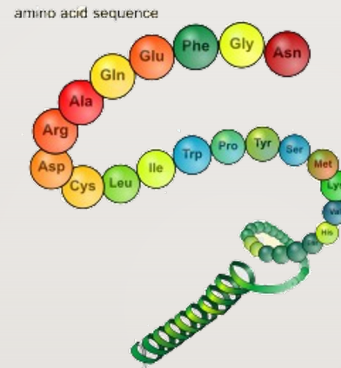
Transcripts

Assembled from RNA-seq or
downloaded ESTs



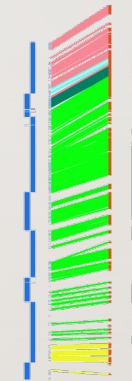
Proteins

Known amino acid sequences
from other organisms



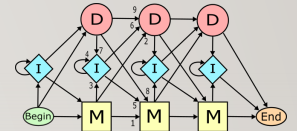
Genomes

Genome from close related species

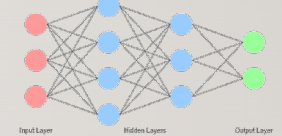


ML models

probabilistic (HMM, GHMM)



deterministic (CNN, RNN i.e deep learning)



Annotation approach

Similarity	X	X		
Comparative			X	
Pure ab initio				X
Hybrid	X	X	X	X
Chooser/Combiner	X	X		X
Pipeline	X	X		X

Extrinsic approaches

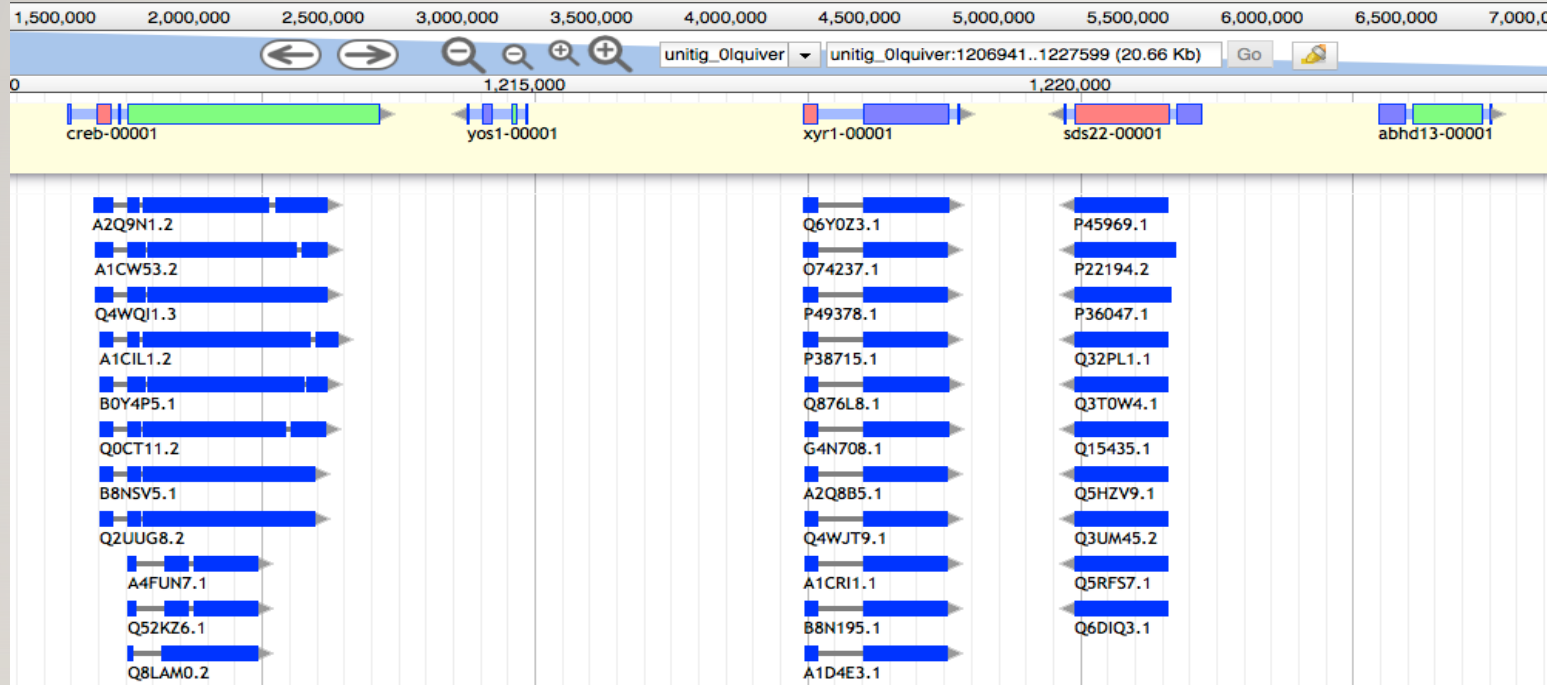
Similarity-based methods

These use similarity to annotated sequences like

- Proteins
- Transcripts
- ESTs

❑ Evidence sequences (protein, est, etc.) are aligned to the genome

Conserved in sequence => conserved annotation with little noise



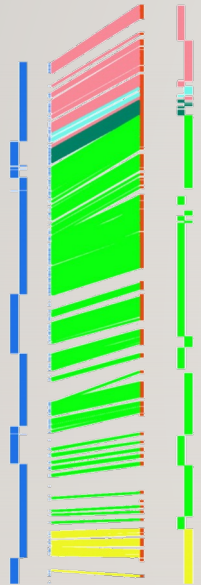
Limits :

- Related to pre-existing data
- Data from model organisms often used => bias?
- Data can be incomplete
- Data can be wrong

Extrinsic approaches

Comparative methods

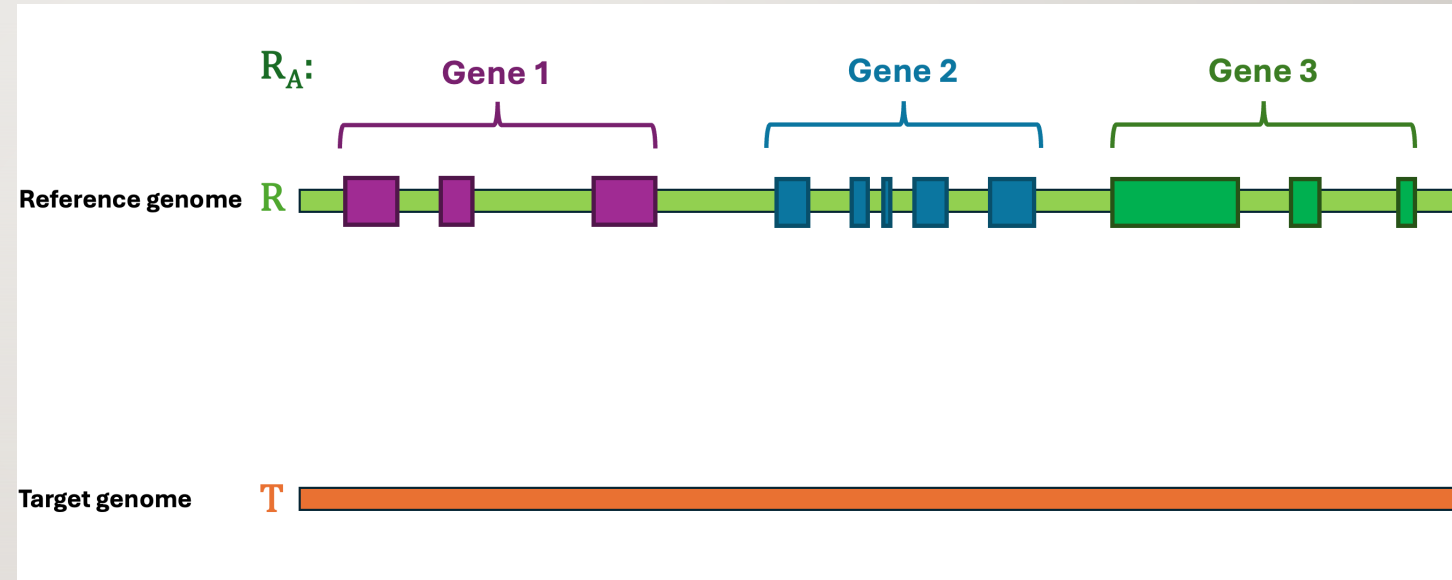
Tools that use **whole-genome alignments** between closely related species.
Transfer gene models based on conserved synteny, gene structure, and alignment blocks.



LASTZ
Mummer
Satsuma
SynMap
Pecan
HAL

chained alignments

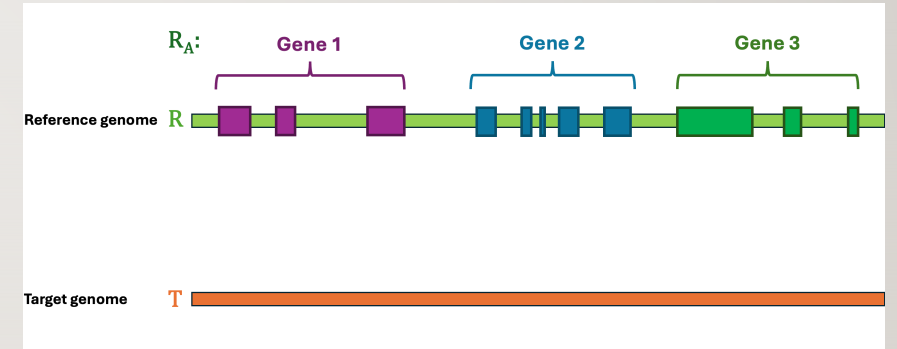
graph alignments



Popular tools

- **Kraken (not the kmer tool one)**
- **TransMap (UCSC)**

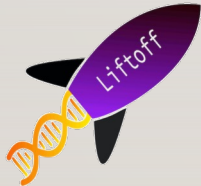
Tools that use **local alignments** between species.
Transfer gene models based on gene alignment.



Popular tools

LiftOff

transfer annotation between closely related species



minimap2

LiftOn

transfer annotation between close and distant related species



Miniprot

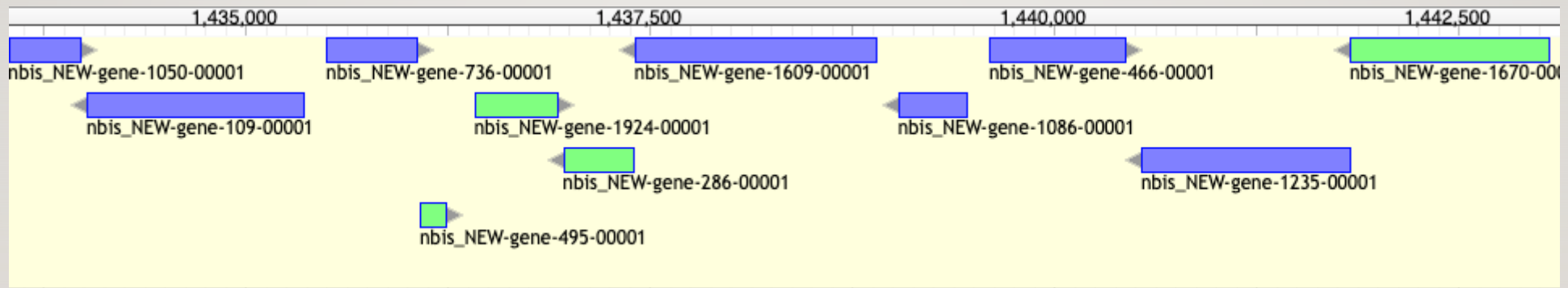
Intrinsic approaches

ab initio

Using Machine Learning approaches (probabilistic and/or deterministic models) to predict features

- **Bacterial gene prediction**

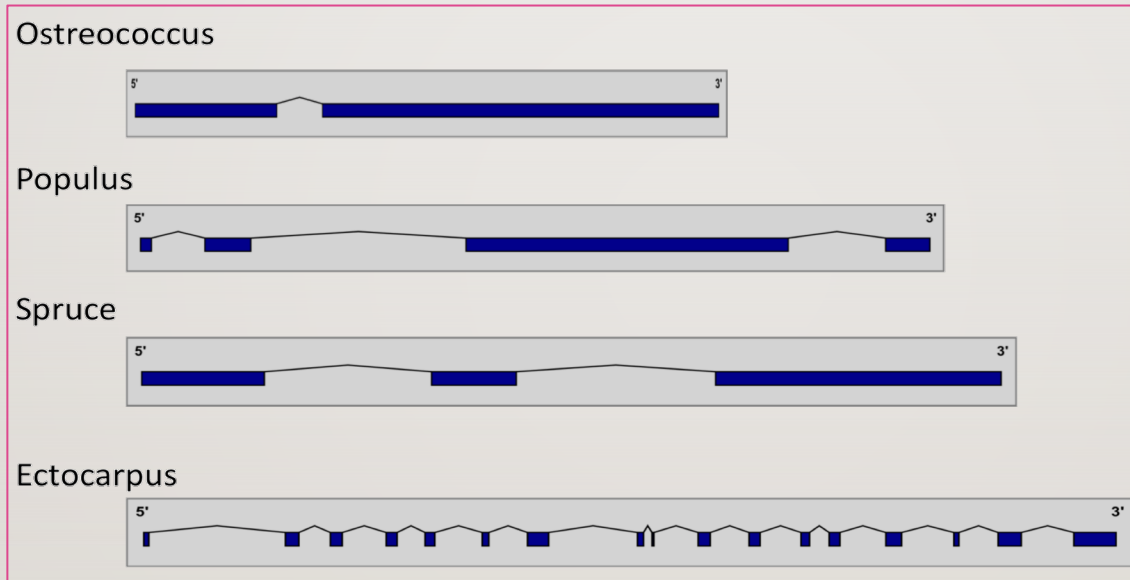
- Short intergenic regions
- Uninterrupted ORFs (No intron)
- Very conserved signals



⇒ easy problem: Accuracy > 90%

- **Eukaryotic gene prediction**

- Presence of intron => structures differ
- Isoforms
- New features (e.g. lncRNAs)



=> Hard problem: Accuracy < 70%

=> Requires training

- **gene content**

Statistical properties of protein-coding sequence learn from existing data.

- codon usage
- hexamer usage
- GC content
- compositional bias between codon positions
- nucleotide periodicity
- exon/intron size
- ...

- **signal detection**

These are like **landmarks** in the genome, they tell the model *where* to look.

- Promoter
- ORF
- Start codon
- Splice site (Donor and acceptor)
- Stop codon
- Poly(A) tail
- CpG islands
- ...

=> *Ab initio* tools will combine this information through different Probabilistic models (e.g. HMM, GHMM, WAM) to calculate the most probable combination that represents a valid gene

Popular tools (Prokaryote)

Complete bacterial or archaeal genomes

- **Prodigal** a widely used baseline for protein-coding gene prediction
- **Pyrodigal** a reimplementation of Prodigal in Python
- **GeneMarkS-2**
- **GLIMMER 3** (requires a training step for a new organism)

Metagenomes assemblies

- **FragGeneScan** (particularly relevant when sequences contain sequencing errors or frameshifts)
- **MetaGeneMark**
- **MetaGene**
- **MetaGeneAnnotator**

Neural networks (CNN, bLSTM) predicts **for each base**, the probability of being:

- in an exon
- in an intron
- at a splice site
- in UTRs, etc.

+

HMM postprocessor

Captures complex, long-range sequence dependencies.

=> struggles to precisely locate exon boundaries

=> base-level predictions can be

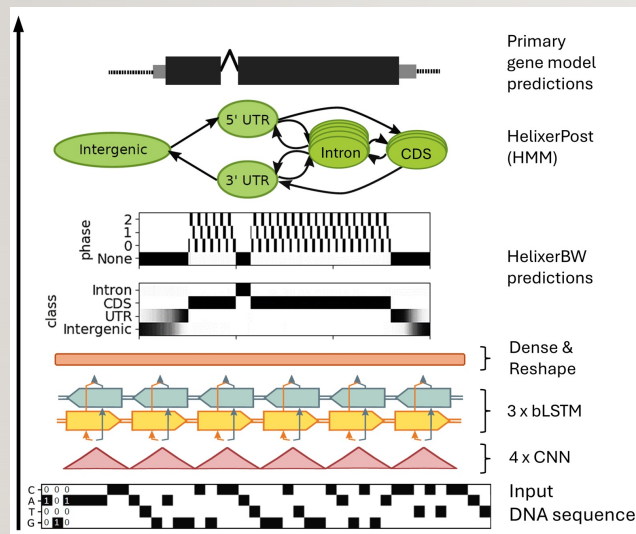
- **Noisy** (e.g., false positives),
- **Biologically inconsistent** (e.g., a stop codon in the middle of a predicted CDS),
- **Missing structure** (e.g., incorrect start/stop, fragmented exons).

Ensures coherent gene structures by filtering noisy outputs.

Popular tools (Eukaryote)

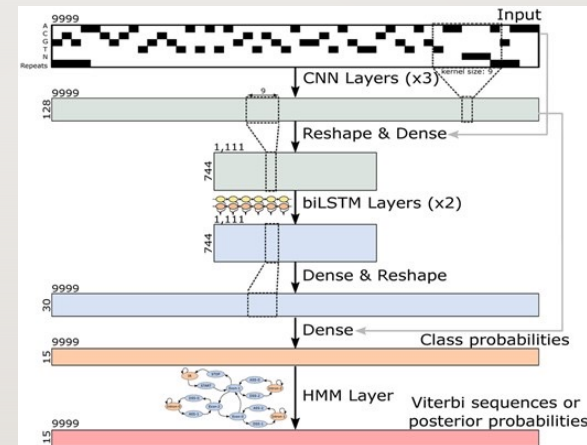
Helixer

Helixer—de novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model

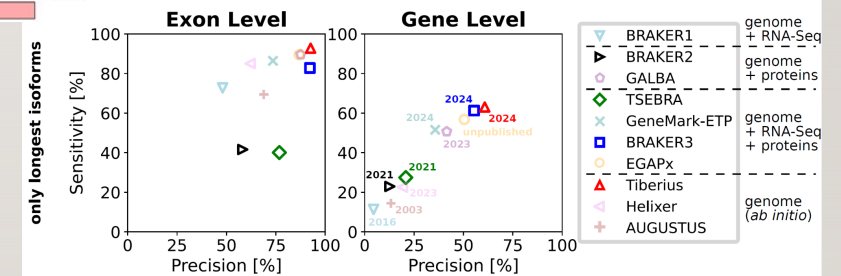


Tiberius

End-to-End Deep Learning with an HMM for Gene Prediction



Gene prediction accuracy for *Homo sapiens*



<https://doi.org/10.1101/2023.02.06.527280>

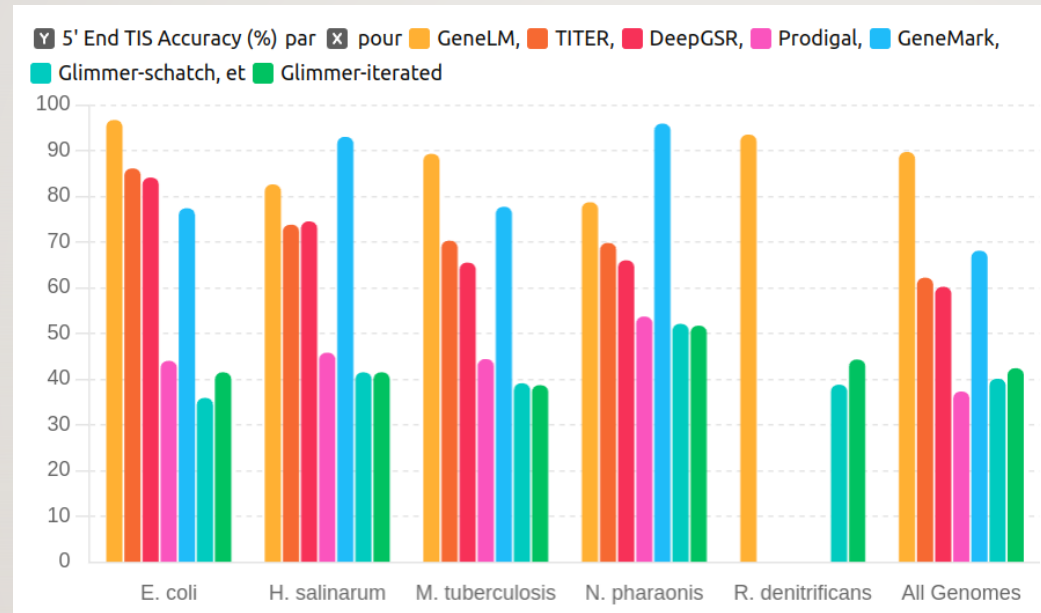
<https://github.com/weberlab-hhu/Helixer>

<https://academic.oup.com/bioinformatics/article/40/12/btae685/7903281>

<https://github.com/Gaius-Augustus/Tiberius>

Popular tools (prokaryote)

GeneLM



=> needs training

Strengths :

- Fast and easy
- Annotate unknown genes
- Sensitivity ok
- Need no external evidence

Limits :

- Split single gene into multiple predictions
- Fused with neighboring genes
- May be less accurate than homology-based methods (Exon boundaries)
- Some tools need training

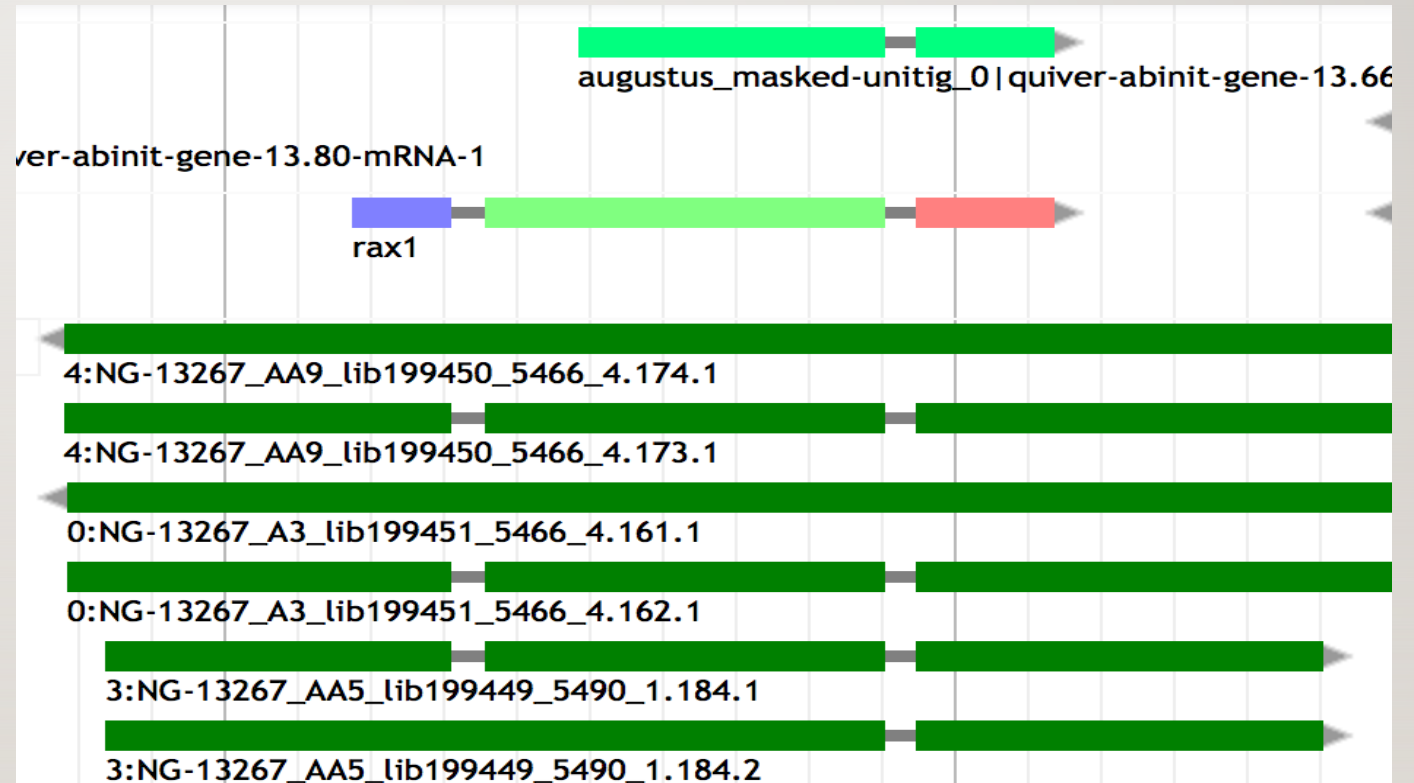
Hybrid approaches

Evidence-drivable gene predictors incorporate hints/extrinsic information in the form of alignments (transcript, protein, whole genome) to increase the accuracy of the gene prediction.

- *Can predict locus without extrinsic information*
- *Improve prediction when extrinsic information available*

Evidence-drivable gene predictors / Ab initio evidence-based approaches incorporate hints (e.g. protein) alignments to increase the accuracy of the gene prediction.

- **Can improve :**
 - Split single gene into multiple predictions
 - Fused with neighboring genes
 - Exon boundaries
- **Tools :**
 - PGAP (NCBI) - constrains or influences gene prediction
 - Bakta - gene boundaries refinement



Chooser / Combiner

- *Selection of gene models*
- *Combining heterogeneous data into gene models*

Use battery of gene finders and evidence (transcript, protein) alignments and:

Chooser: Selects among existing predictions according to evidence or confidence.

Combiner: Constructs consensus gene models by integrating evidence from different sources.

Tool	Consensus based chooser	Evidence based chooser	weight of different sources	Comment
GeMoMa		X		You can weight on presence in multiple genome
JIGSAW	X			
EVM Evidencemodeler	X	X	X	User can set the expected evidence error rate manually or/and learn from a training set
Evigan	X		X	Unsupervised learning method
=> Ipred		X		Does not require any a priori knowledge Can also combine only evidences to create a gene model

Pipelines

(The ultimate step)

- *Align evidence (proteins)*
- *Annotation of other features (Repeat, tRNA, etc)*
- *Add functions*
- *And more...*

Pipeline

Tool	Protein	<i>Ab initio</i>	Function	Other annotated features	Comment
PGAP	X	X	X	tRNA, rRNA, ncRNA, pseudogenes, CRISPR, mobile elements	Integrates protein alignments, protein-family HMMs and GeneMarkS-2+ for structural and functional annotation
Prokka	X	X	X	tRNA, tmRNA, rRNA, ncRNA, CRISPR; signal peptides when supported by the relevant dependencies	Uses Pyrodigal for CDS prediction.
Bakta	X	X	X	tRNA, tmRNA, rRNA, ncRNA, regulatory regions, CRISPR, pseudogenes, replication origins, sORFs	Uses Prodigal for CDS prediction. Huge database used.
DFAST	X	X	X	RNA genes and other genomic features, depending on the modules used	
RAST / RASTtk	X	X	X	RNA features, metabolic functions and other features, depending on the workflow	
BV-BRC	X	X	X	tRNA, rRNA, repeat regions, prophages, antimicrobial resistance genes, virulence factors, protein families	CDS prediction uses Prodigal followed by Glimmer

The End

Thank you for your attention