Bioinformatics

CS300

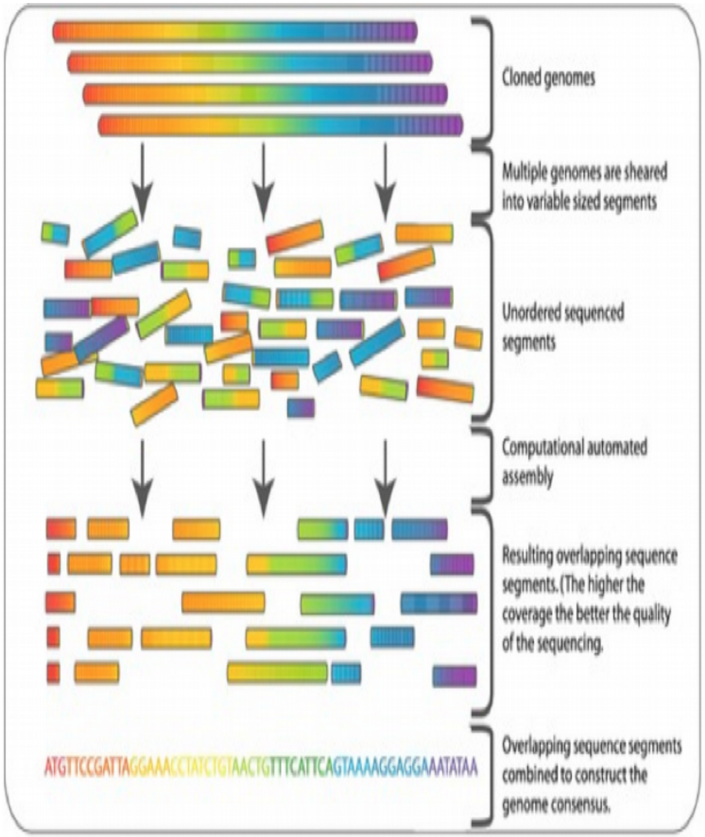
Genome annotation and sequence-based gene prediction

Fall 2017

Oliver Bonham-Carter

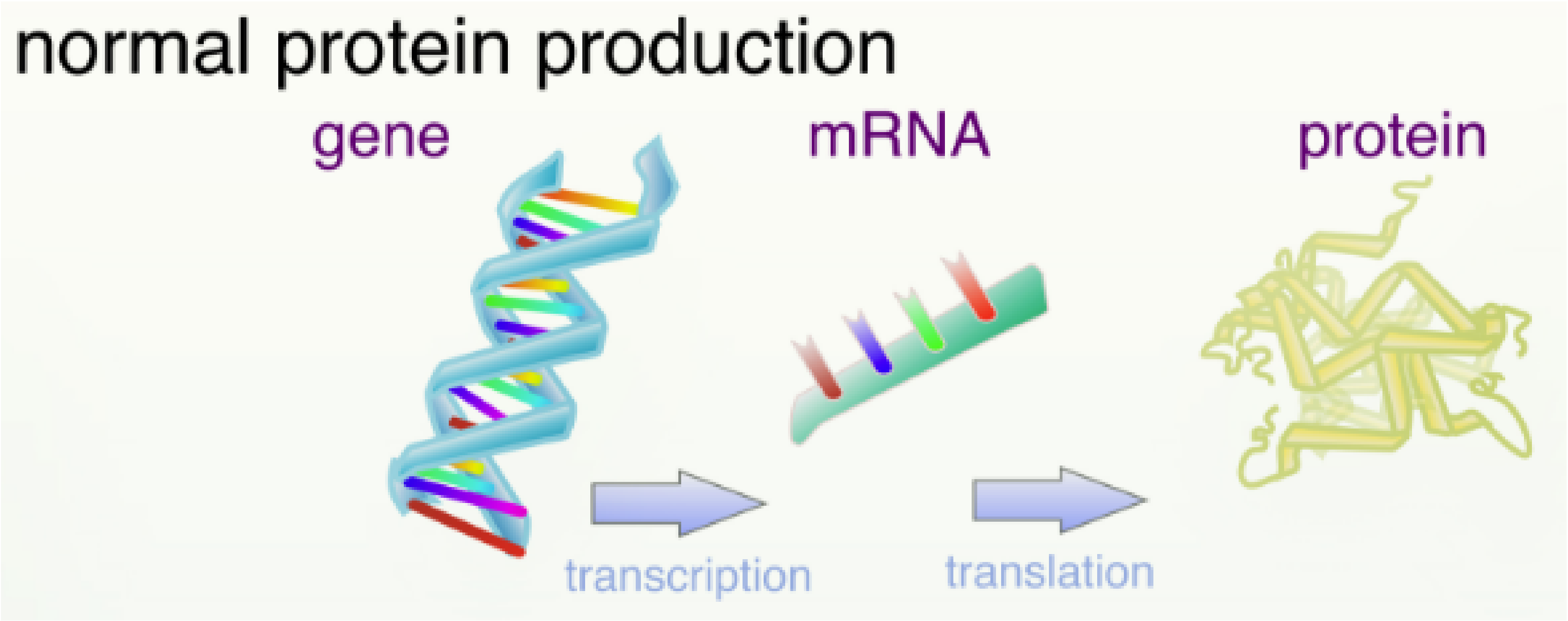
Genome Projects

- **Goals**:
  - Determine complete genome sequence of an organism
  - Annotate protein-coding genes and other important genome-encoded features
    - find
    - identify
    - characterize
    - describe
    - computational predictions later confirmed at the lab bench


Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequence segments combined to construct the genome consensus.

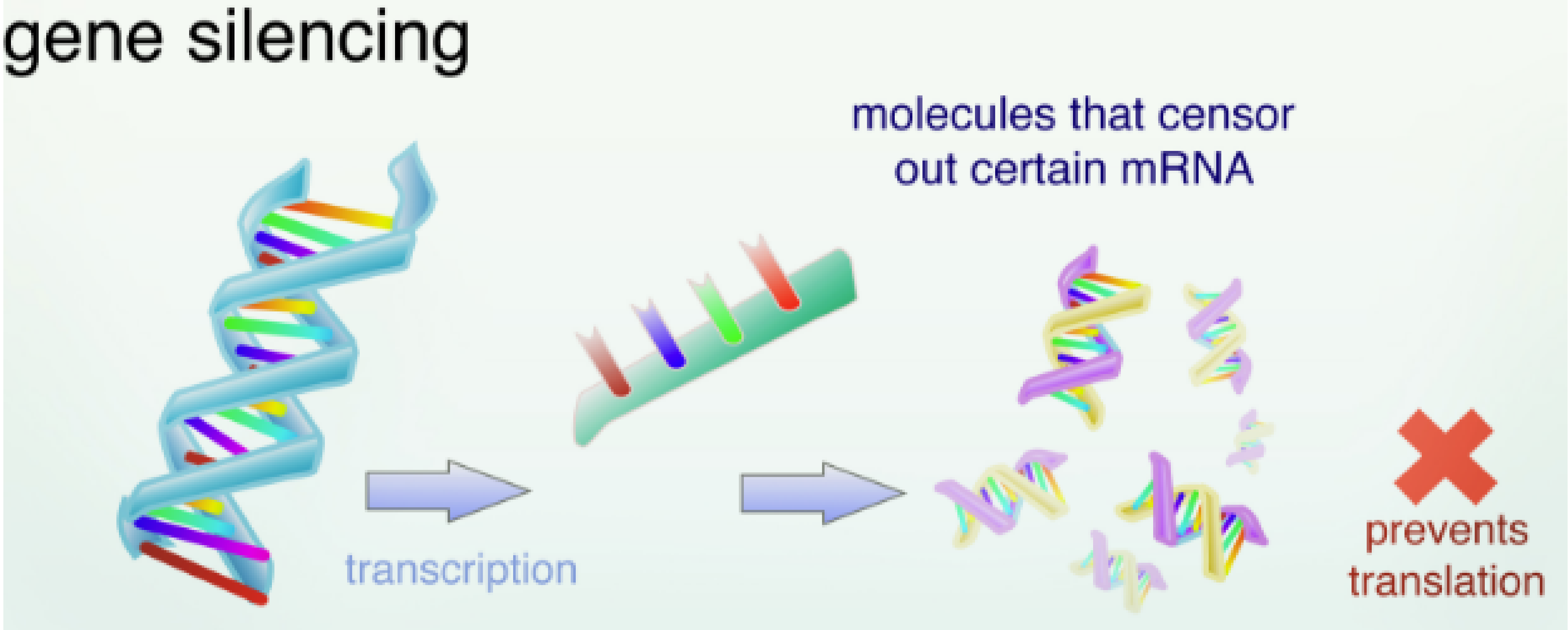
Gene Prediction

- Sequence-based – find features based on specific sequences
- What does a gene look like?
  - Qualities?
  - Behaviors?
  - Sequence trends?



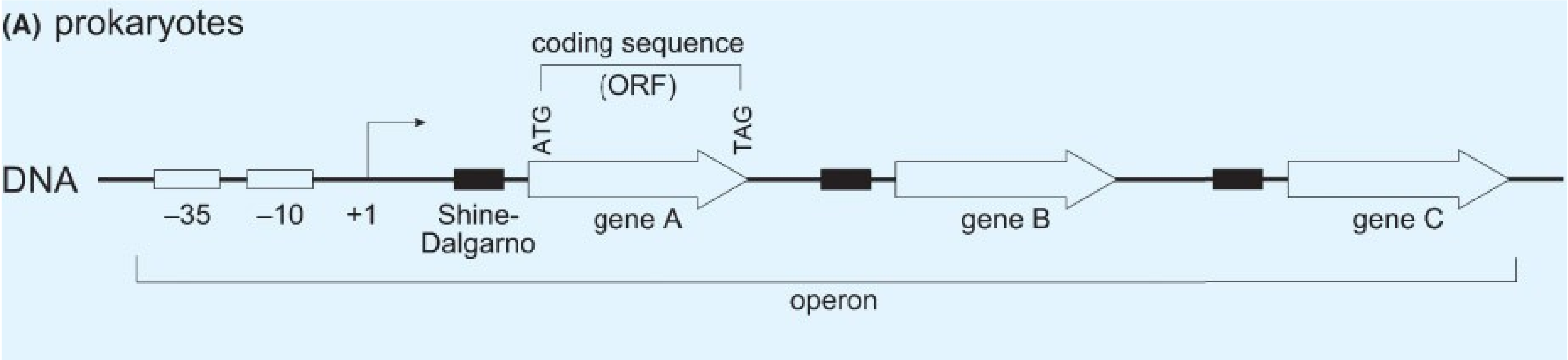
Gene Prediction

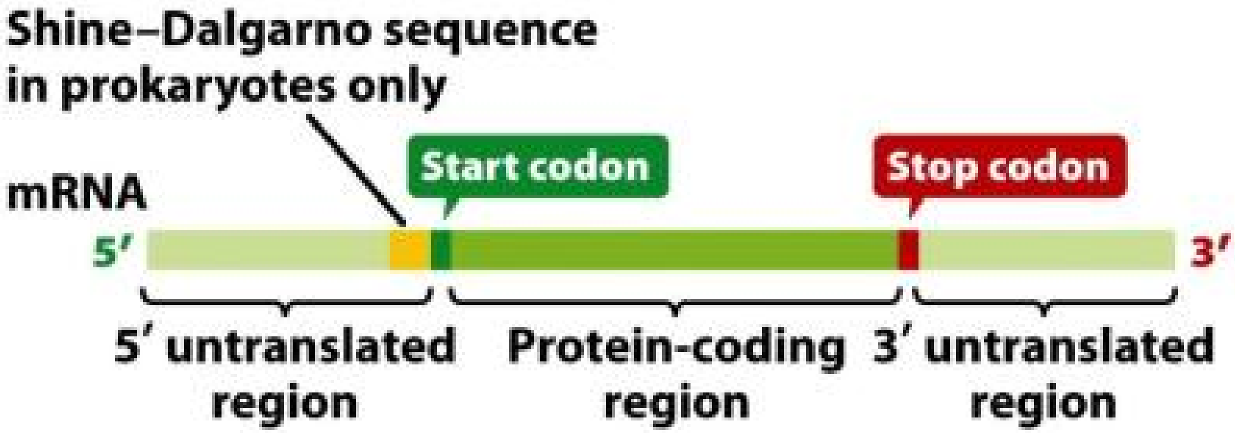
- Two obvious questions:
- **Why not just look to see what proteins are available?**
- **Could that tell us what gene must be there to make the protein?**



Gene Prediction

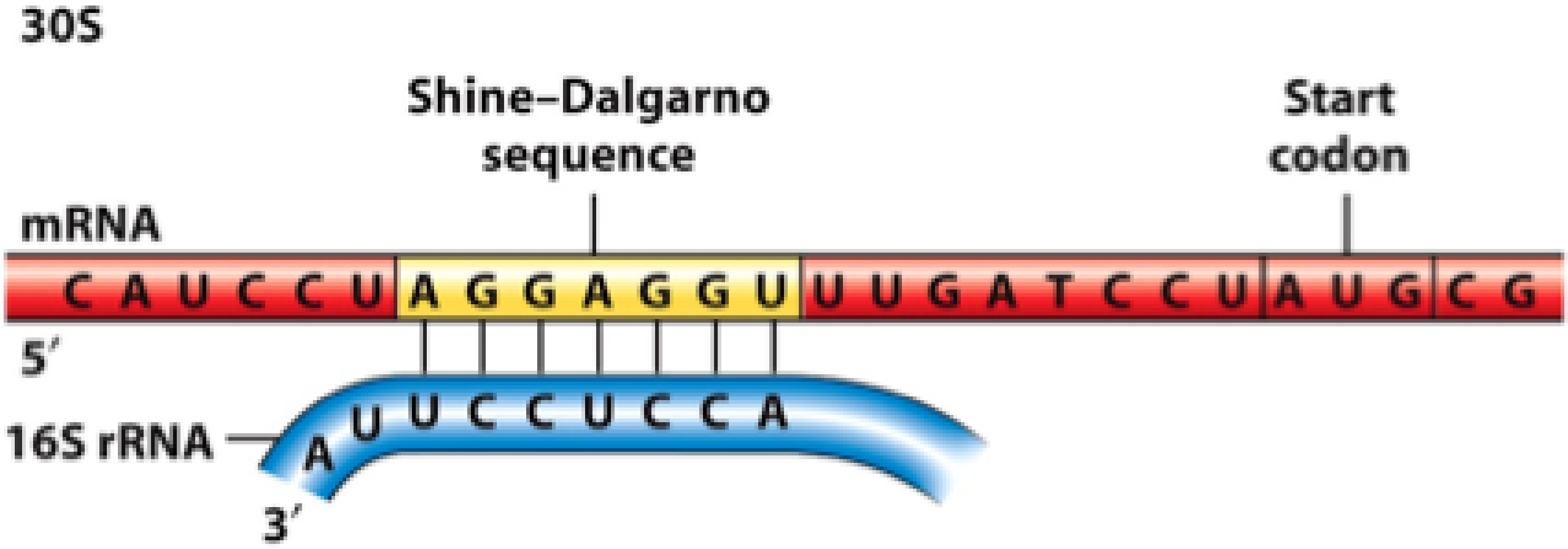
- We look for specific features or *land-marks* in a sequence that may *suggest* that there is a gene at play.
  - The Shine-Dalgarno: found of a upstream of a DNA start codon: ATG





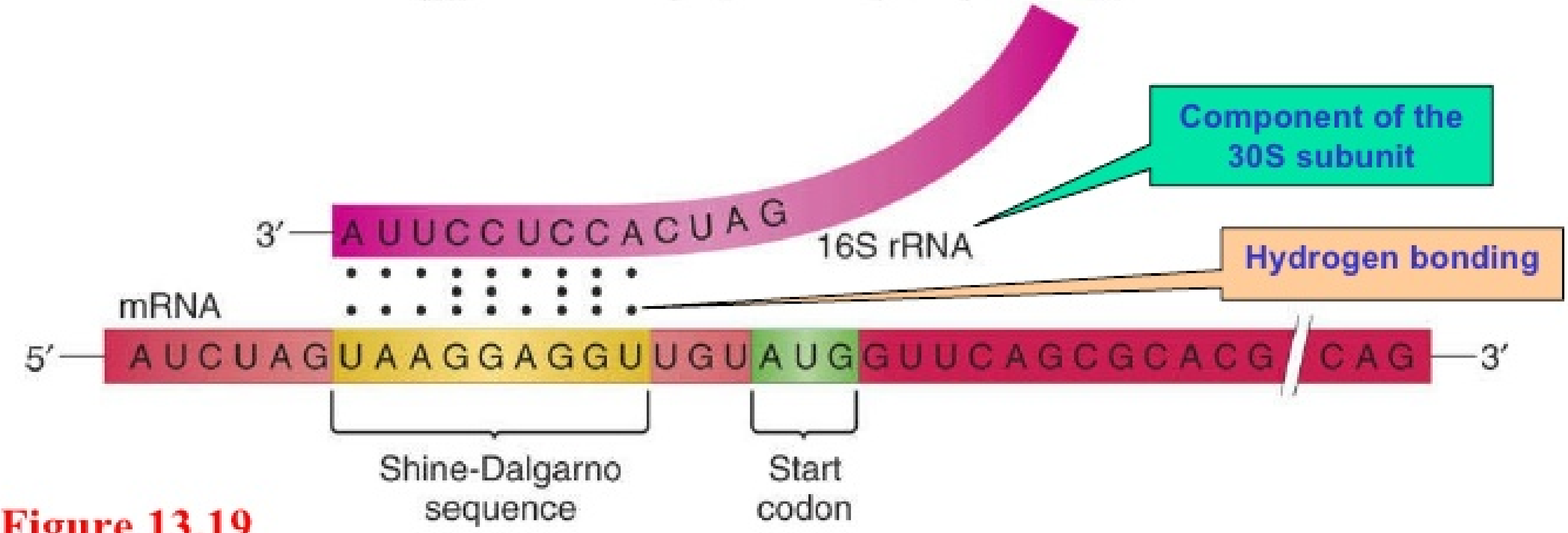
Shine-Dalgarno Sequence

- Shine and Dalgarno showed that the nucleotide tract at the 3' end of E. coli 16S ribosomal RNA (rRNA) is pyrimidine-rich and has the sequence: **Py-*ACCUCCU*UA-3'OH**.
- They proposed that these ribosomal nucleotides recognize the complementary purine-rich sequence ***AGGAGGU***, which is found upstream of the start codon AUG in a number mRNAs found in viruses that affect E. coli.



Shine-Dalgarno Sequence

- The binding of mRNA to the 30S subunit is facilitated by a ribosomal-binding site or Shine-Dalgarno sequence
  - This is complementary to a sequence in the 16S rRNA


3′
A U U C C U C C A C U A G
16S rRNA
mRNA
5′
A U C U A G U A A G G A G G U U G U A U G G U U C A G C G C A C G C A G
3′
Shine-Dalgarno sequence
Start codon

Component of the 30S subunit

Hydrogen bonding

Figure 13.19

- Figure 13.18 outlines the steps that occur during translational initiation in bacteria

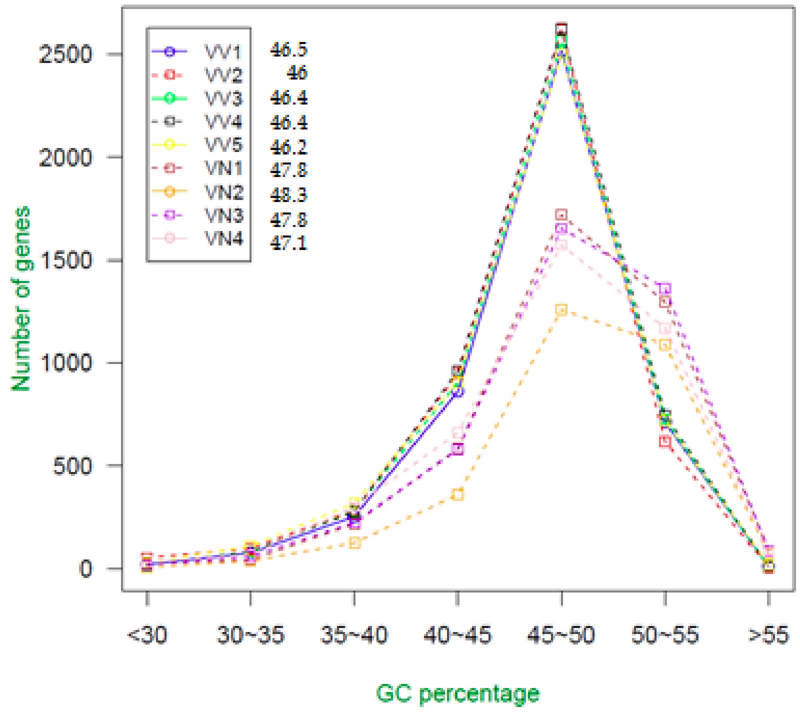
Prediction Algorithms

- Alignment-based – find genes/features based on conserved sequences is well-studied organisms (database searching)
  - Automatic assignment based on sequence similarity (best BLAST hit): gene name, protein name, function
  - Quality vs Quantity

Prediction Algorithms

- Content-based – consider overall properties of the sequence when making predictions
- nucleotide frequency
- Codon frequency/codon bias

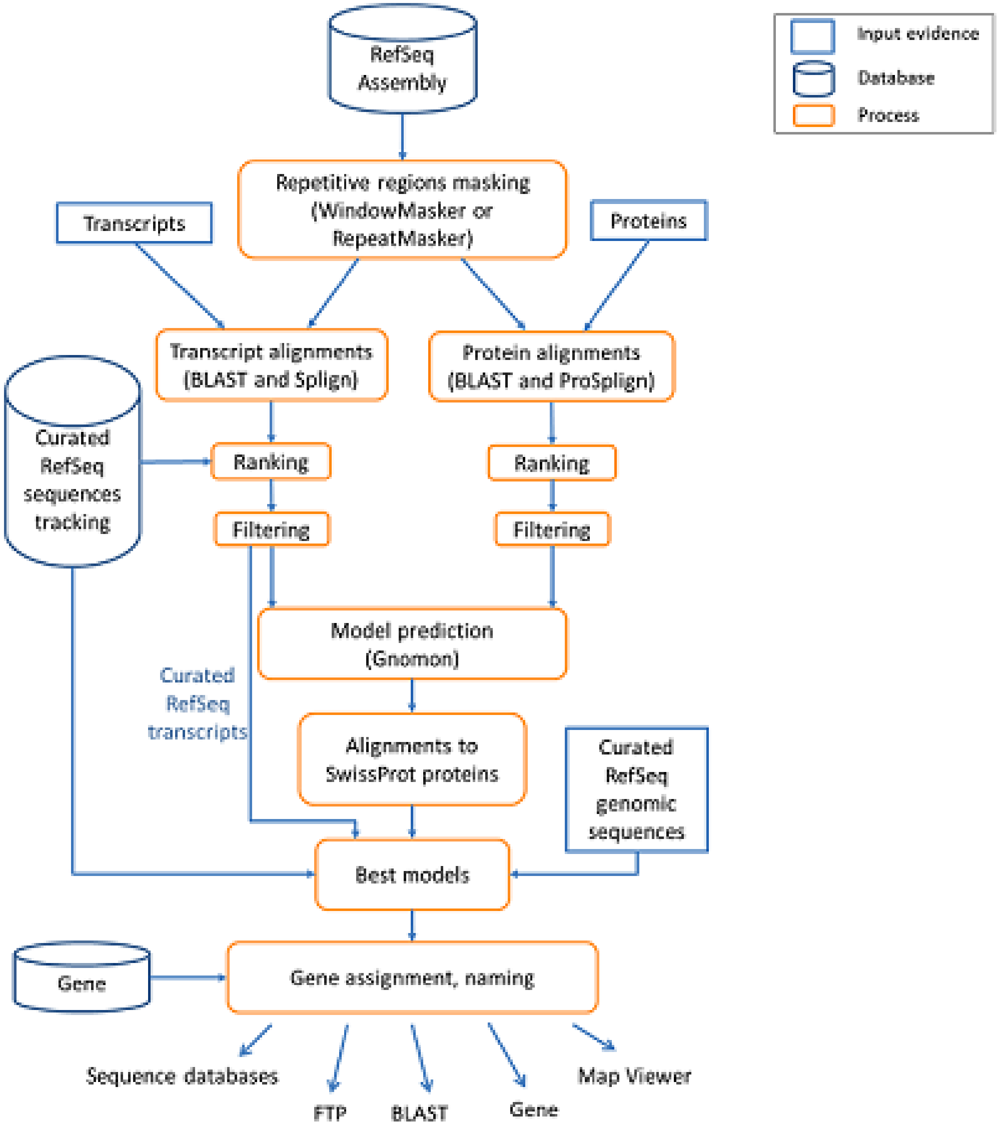
DISTRIBUTION OF GC CONTENT



- GC Content for all *V. vulnificus* and *V.naverensis* gene predictions
- Most of the genomes contained a high percentage of genes with GC contents between 45-50%.

Prediction Algorithms

- Probabilistic – combination of sequence-based and content-based plus probability
- *“annotation pipeline”*



RefSeq Assembly

Repetitive regions masking (WindowMasker or RepeatMasker)

Transcripts

Proteins

Transcript alignments (BLAST and Splign)

Protein alignments (BLAST and ProSplign)

Curated RefSeq sequences tracking

Ranking

Ranking

Filtering

Filtering

Model prediction (Gnomon)

Curated RefSeq transcripts

Alignments to SwissProt proteins

Curated RefSeq genomic sequences

Best models

Gene

Gene assignment, naming

Sequence databases

FTP

BLAST

Gene

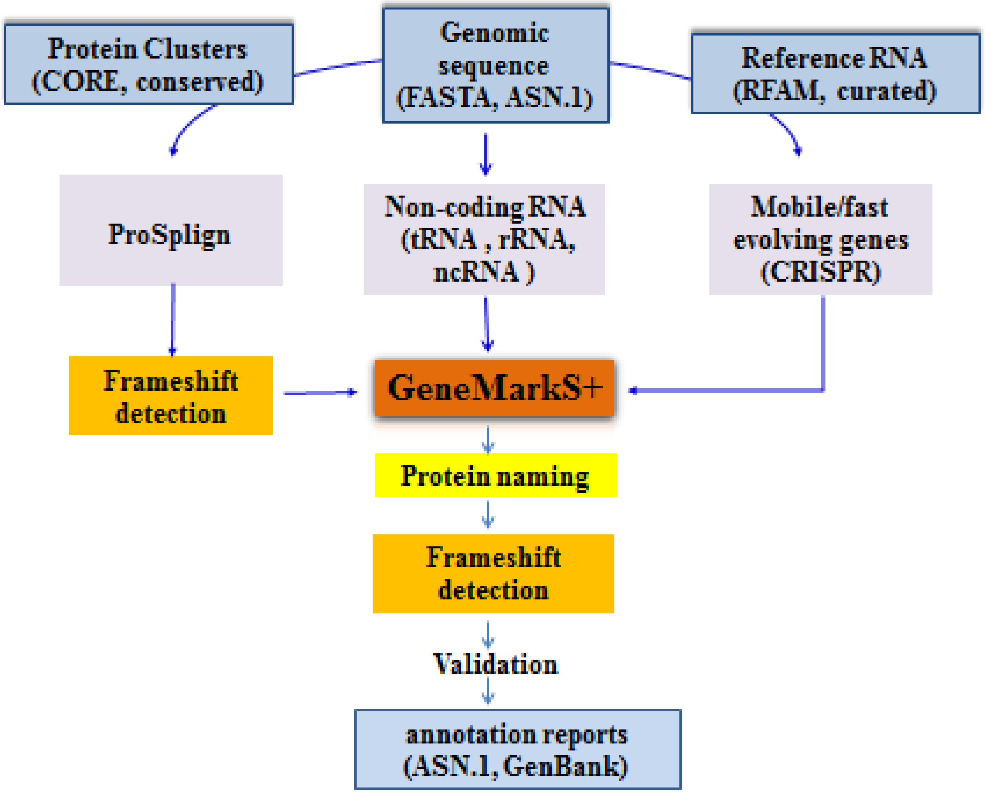
Map Viewer

Input evidence

Database

Process

NCBI Prokaryotic Annotation Pipeline

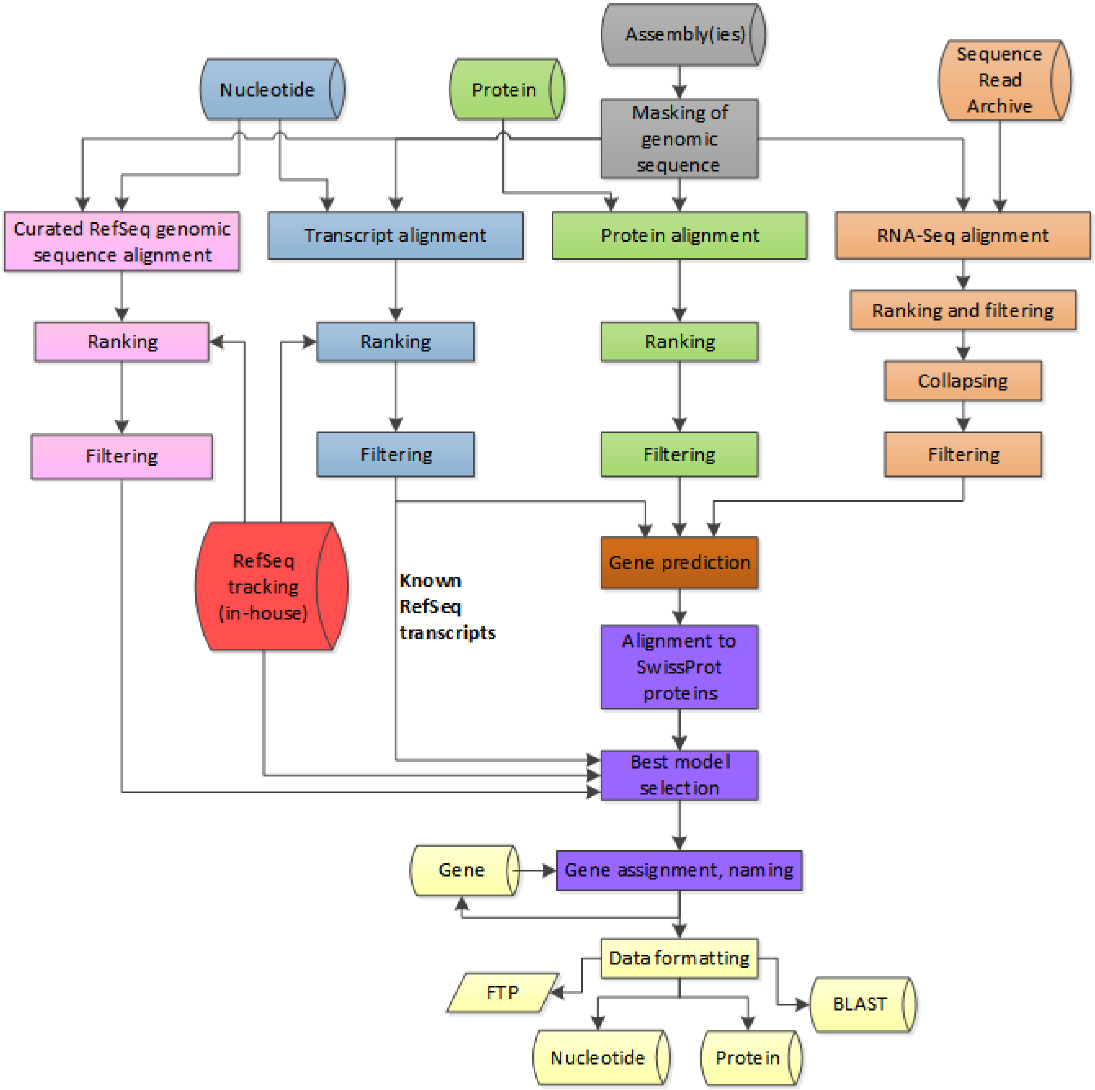

Protein Clusters (CORE, conserved)
Genomic sequence (FASTA, ASN.1)
Reference RNA (RFAM, curated)
ProSplign
Non-coding RNA (tRNA , rRNA, ncRNA)
Mobile/fast evolving genes (CRISPR)
Frameshift detection
GeneMarkS+
Protein naming
Frameshift detection
Validation
annotation reports (ASN.1, GenBank)

- Combines sequence-based algorithm with alignment-based approach
  - Protein-coding genes
  - Structural RNAs (5S, 16S, 23S)
  - Transfer RNAs
  - Small non-coding RNAs
- Rely only on properties of DNA and training set of genes

http://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/

NCBI Eukaryotic Annotation Pipeline

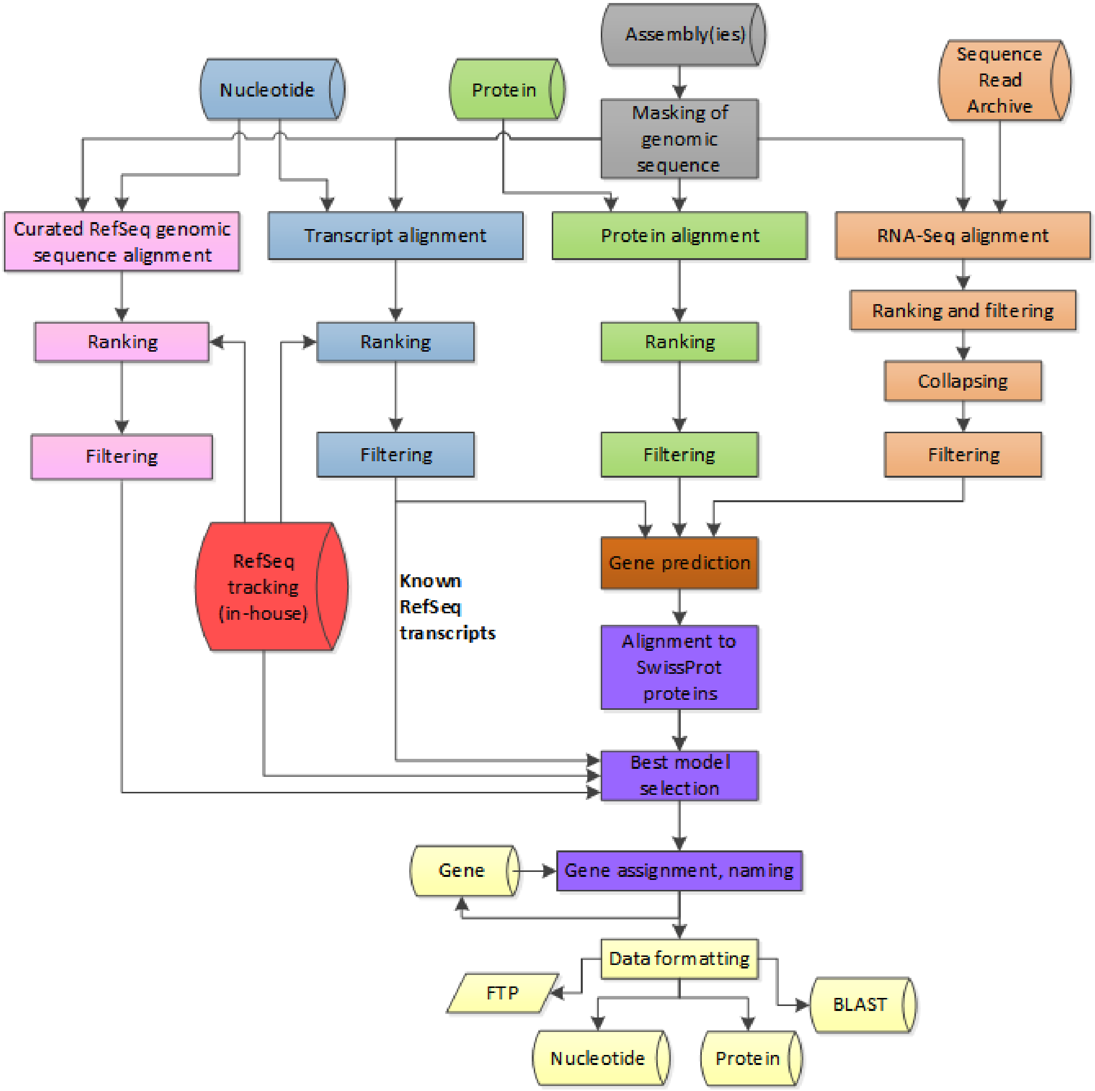
1. Masking
   - try to identify and ignore non-coding regions
2. Alignment-based predictions
   - Where have we seen this sequence before?
3. Sequence/content-based predictions from alignment-based
4. Best selected (probability), named, and released



https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/#assemblies

NCBI Eukaryotic Annotation Pipeline

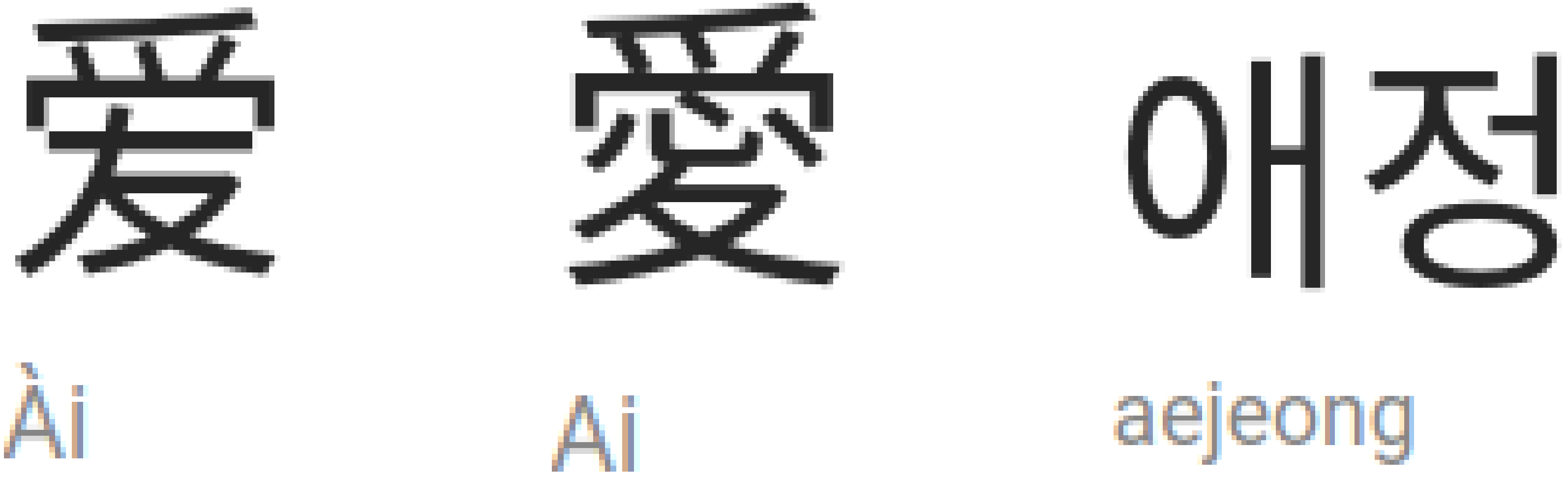
- The best models are selected among the RefSeq and the predicted models, named and accessioned (purple).
- At the end, the annotation products are formatted and deployed to public resources (yellow).



https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/#assemblies

Natural Differences

- We can use the general differences in genetic presentation between types of organisms to find meaningful regions (which could be genes)



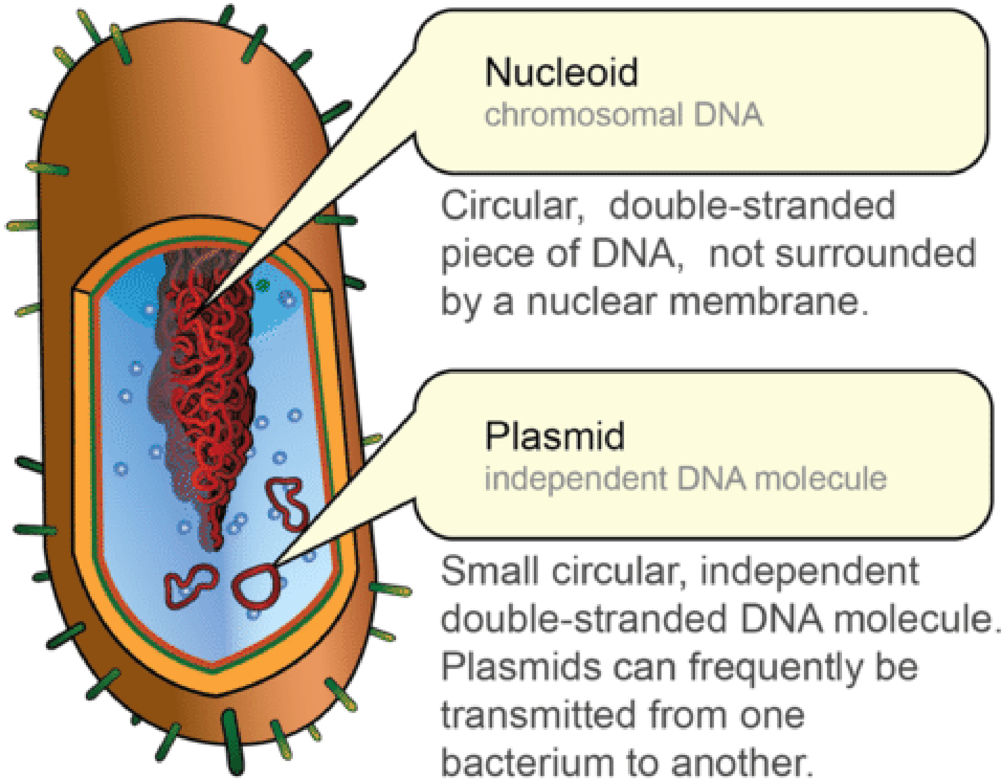
爱 Ài

愛 Ai

애정 aejeong

“Love” in Chinese, Japanese and Korean

Prokaryotic versus Eukaryotic Genomes



Nucleoid
chromosomal DNA

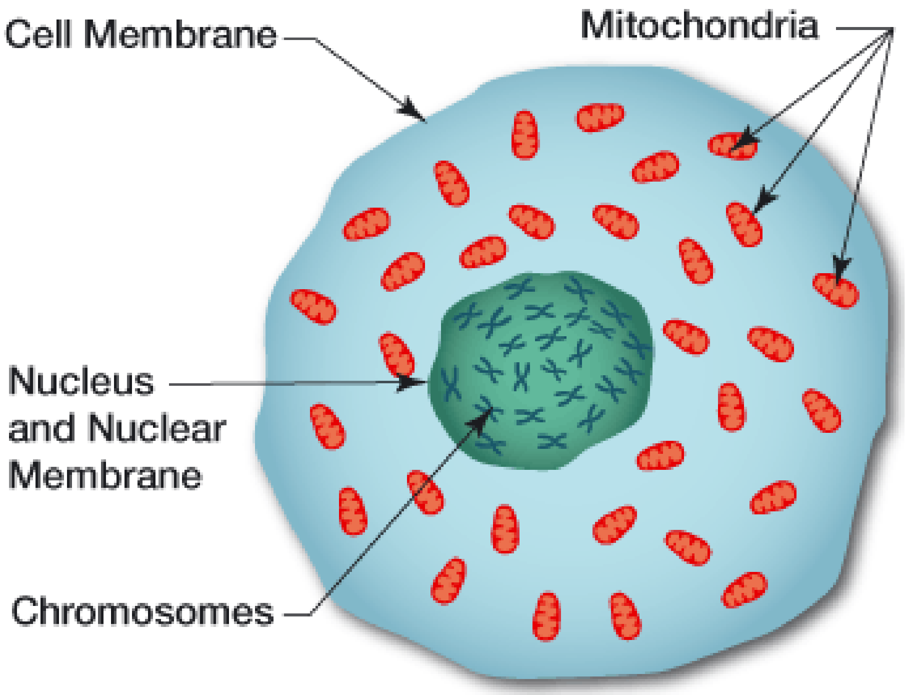
Circular, double-stranded piece of DNA, not surrounded by a nuclear membrane.

Plasmid
independent DNA molecule

Small circular, independent double-stranded DNA molecule. Plasmids can frequently be transmitted from one bacterium to another.

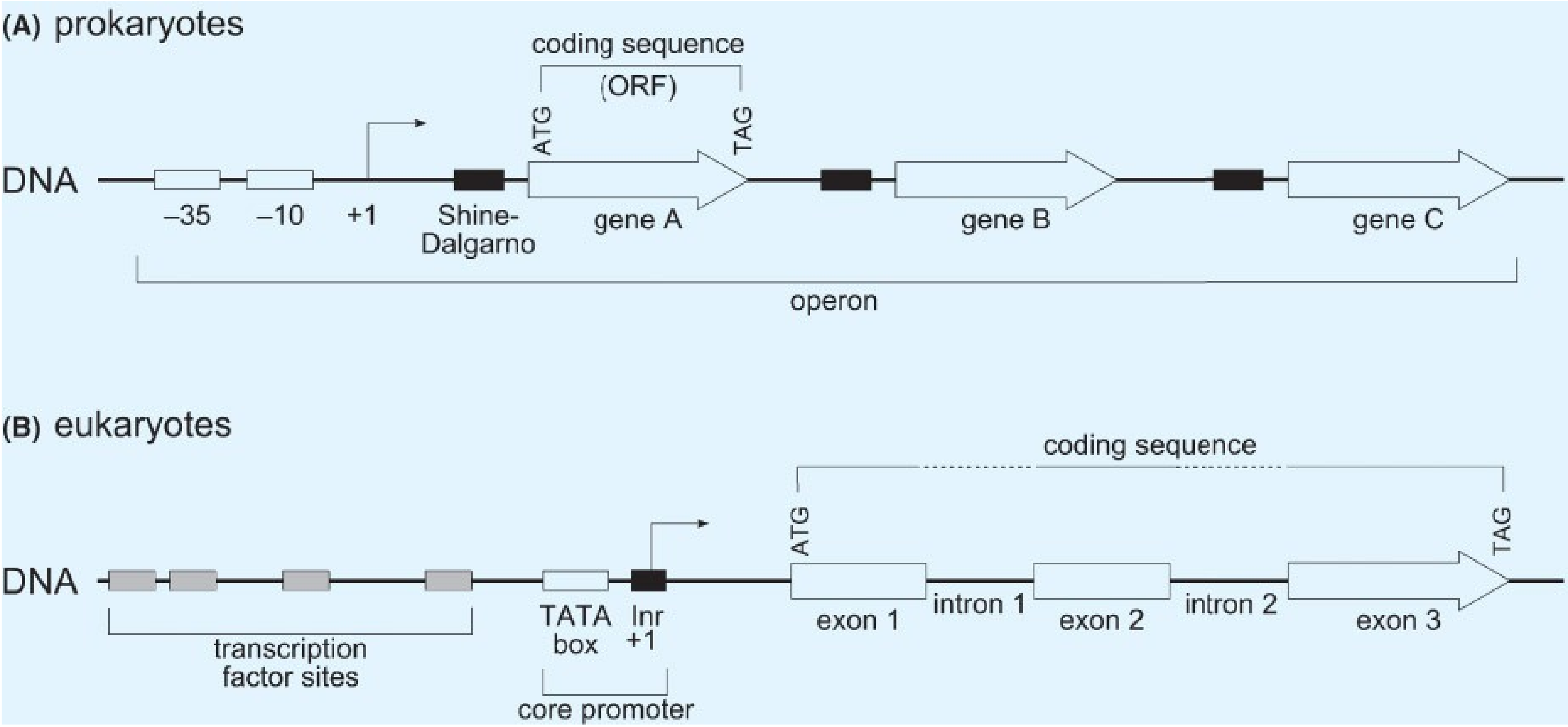
- Prokaryotes
  - A circular chromosome
    - "Genome"
  - Extra DNA in plasmids
    - smaller, self-replicating

Prokaryotic versus Eukaryotic Genomes



- Eukaryotes
  - Multiple linear Chromosomes
    - “Genome”
  - Extra DNA in Mitochondria/chloroplast

Need to know feature structure



Comparison of Prokaryotes vs Eukaryotes transcription unit structures

Prokaryotic versus Eukaryotic Genomes

Organism	Amount of DNA (bp)	# of genes	Genes per million bases
🟥 *Escherichia coli*	4,600,000	4,400	950
🟦 *Saccharomyces cerevisiae*	12,000,000	5,800	480
🟦 *Drosophila melanogaster*	180,000,000	13,700	76
🟦 *Mus musculus*	2,600,000,000	25,000	11
🟦 *Homo sapiens*	2,900,000,000	25,000	10

🟦 **Eukaryotic cells**

🟥 **Prokaryotic cells**

Consensus Sequences

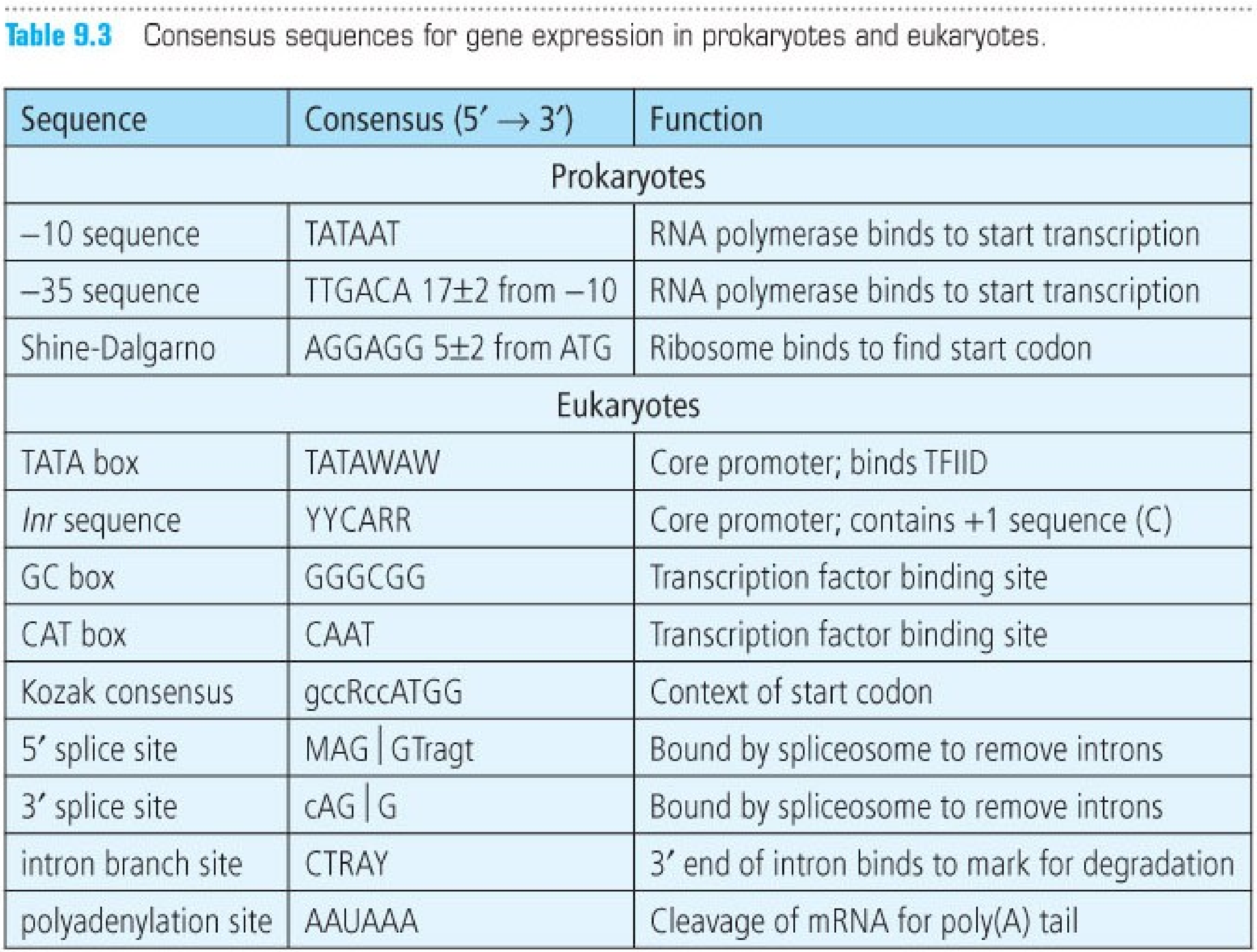
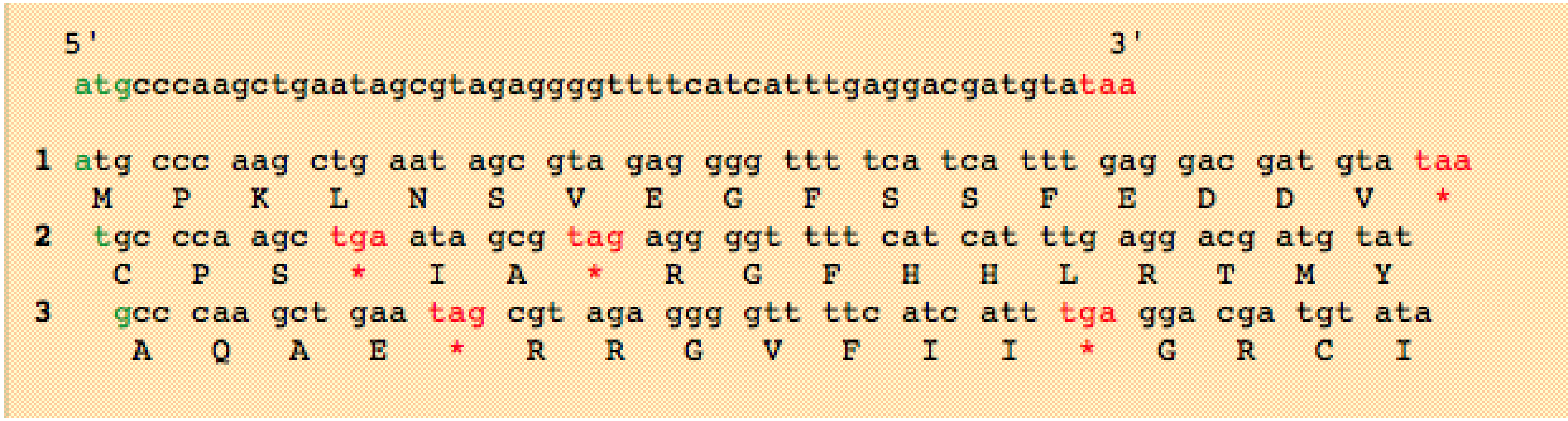


Table 9.3 Consensus sequences for gene expression in prokaryotes and eukaryotes.

Sequence	Consensus (5′ → 3′)	Function
	Prokaryotes	
−10 sequence	TATAAT	RNA polymerase binds to start transcription
−35 sequence	TTGACA 17±2 from −10	RNA polymerase binds to start transcription
Shine-Dalgarno	AGGAGG 5±2 from ATG	Ribosome binds to find start codon
	Eukaryotes	
TATA box	TATAWAW	Core promoter; binds TFIID
Inr sequence	YYCARR	Core promoter; contains +1 sequence (C)
GC box	GGGCGG	Transcription factor binding site
CAT box	CAAT	Transcription factor binding site
Kozak consensus	gccRccATGG	Context of start codon
5′ splice site	MAG | GTragt	Bound by spliceosome to remove introns
3′ splice site	cAG | G	Bound by spliceosome to remove introns
intron branch site	CTRAY	3′ end of intron binds to mark for degradation
polyadenylation site	AAUAAA	Cleavage of mRNA for poly(A) tail

Open Reading Frame (ORF)

- Online tools:
  - NCBI:
  - https://www.ncbi.nlm.nih.gov/orffinder/
- Sequence Manipulation Suite:
  - http://www.bioinformatics.org/sms2/orf_find.html

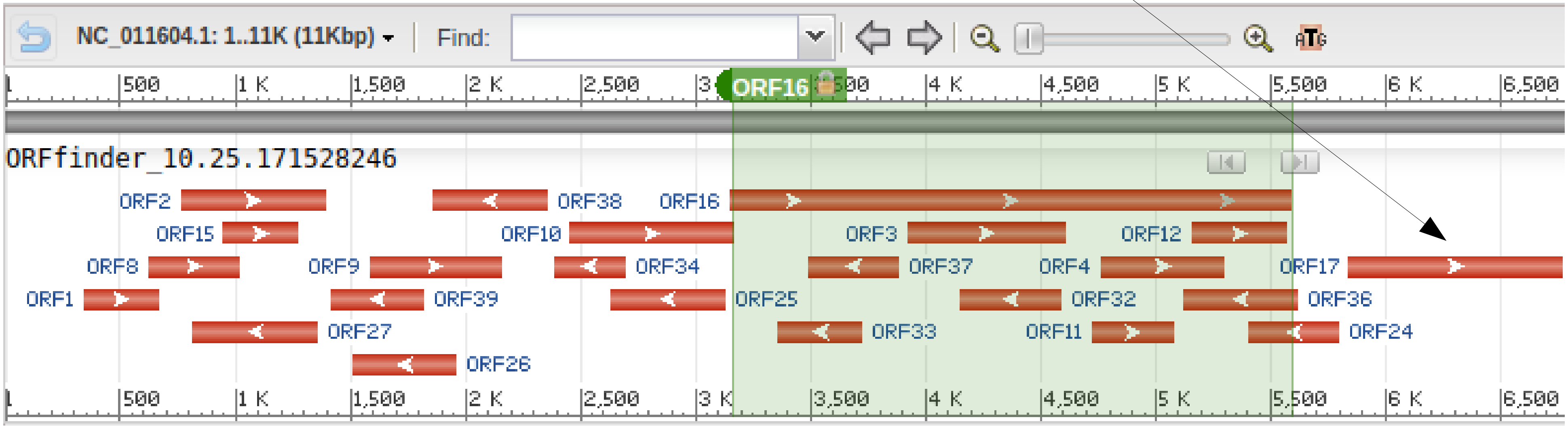


```
 5'                                                     3'
  atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa

1 atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
2  tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
3   gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

Class Activity: NCBI – ORFfinder

- Use NCBI ORF Finder to annotate a plasmid
  - https://www.ncbi.nlm.nih.gov/orffinder/
- Try: NC_011604
  - Salmonella enterica subsp. enterica serovar Westhampton plasmid pWES-1, complete sequence
  - What are the red rectangles with the arrows?



NC_011604.1: 1..11K (11Kbp)
Find:
1
500
1 K
1,500
2 K
2,500
3 K
3,500
4 K
4,500
5 K
5,500
6 K
6,500
ORFfinder_10.25.171528246
ORF2
ORF38
ORF16
ORF15
ORF10
ORF3
ORF12
ORF8
ORF9
ORF34
ORF37
ORF4
ORF17
ORF1
ORF39
ORF25
ORF32
ORF36
ORF27
ORF33
ORF11
ORF24
ORF26