Bioinformatics

CS300

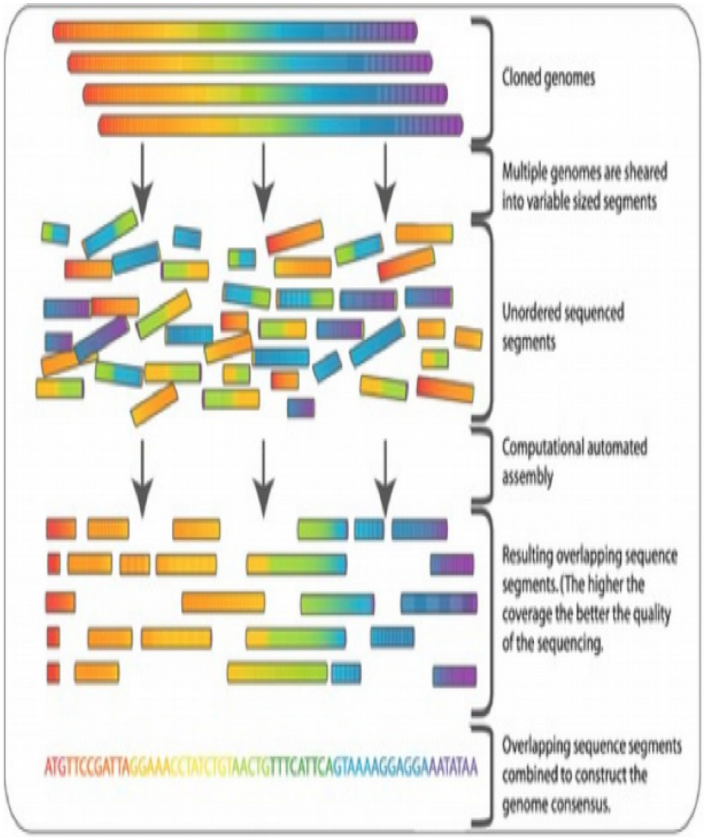
Genome annotation and sequence-based gene prediction

Fall 2019

Oliver BONHAM-CARTER

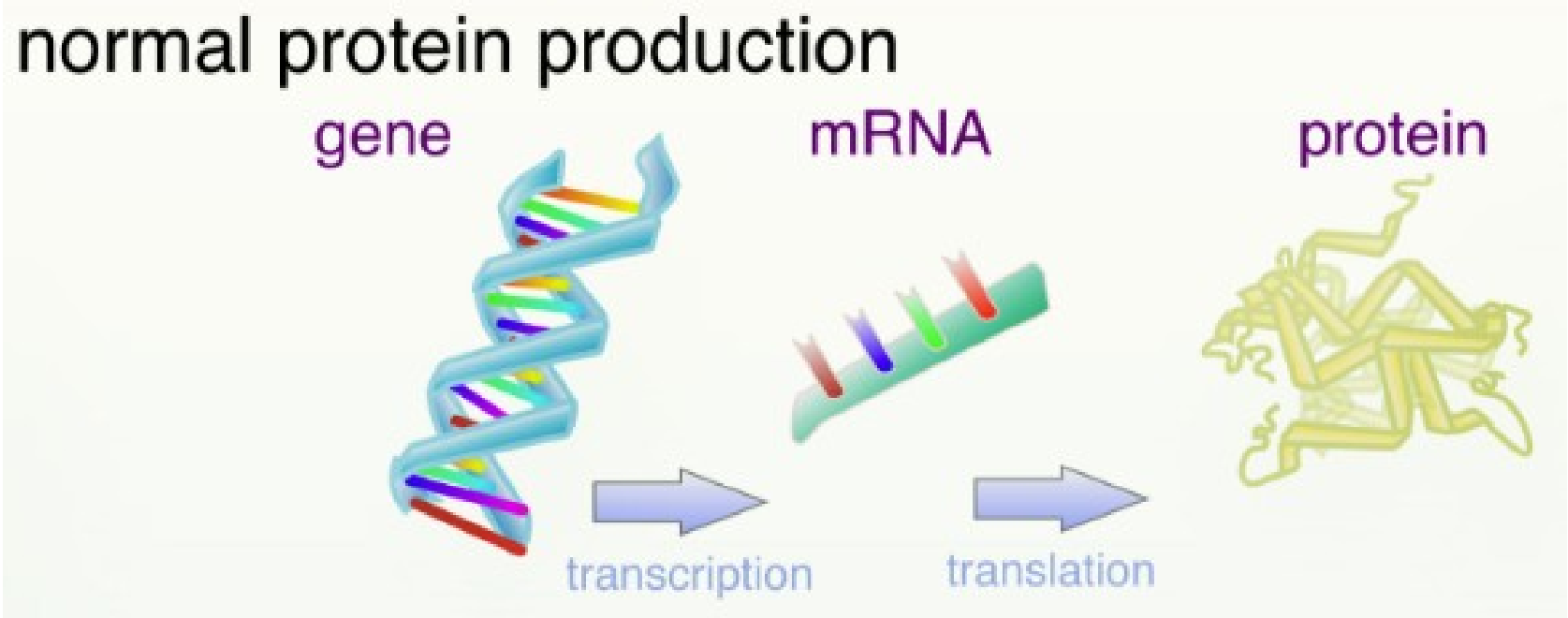
Genome Projects

- **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



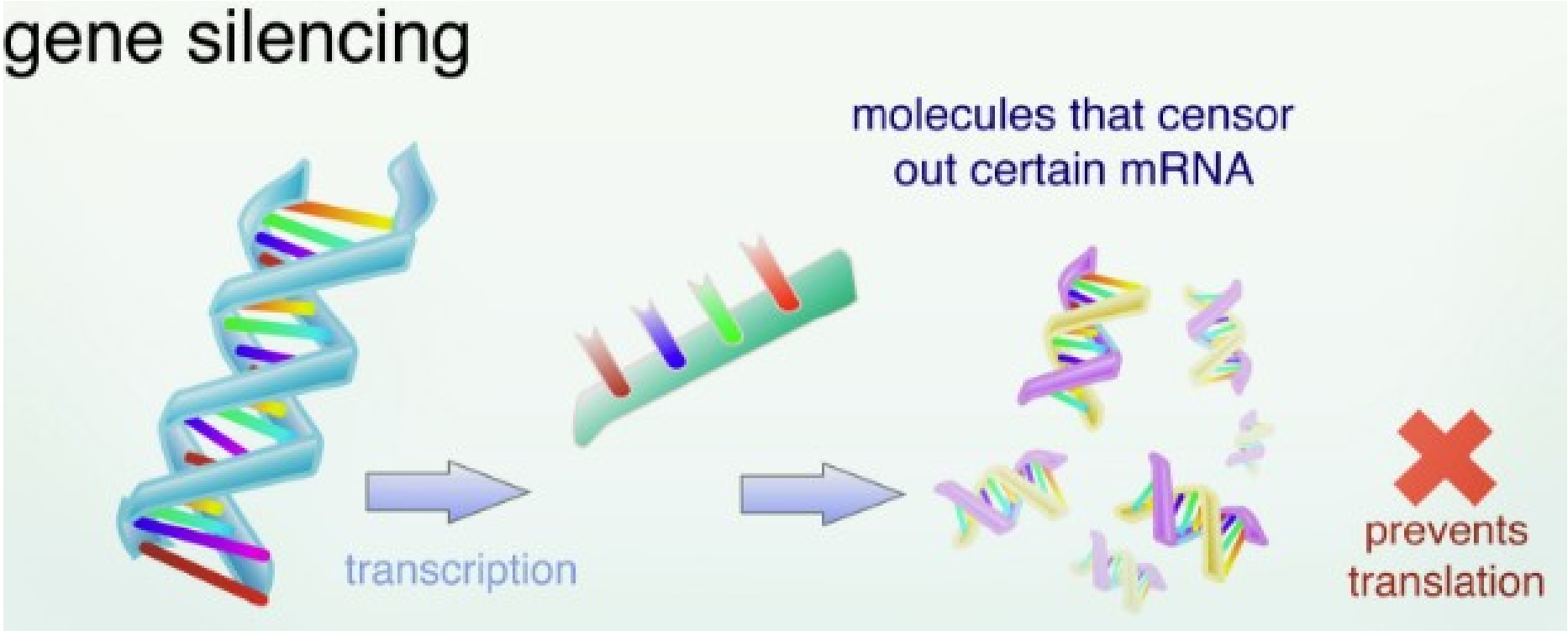
Gene Prediction

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

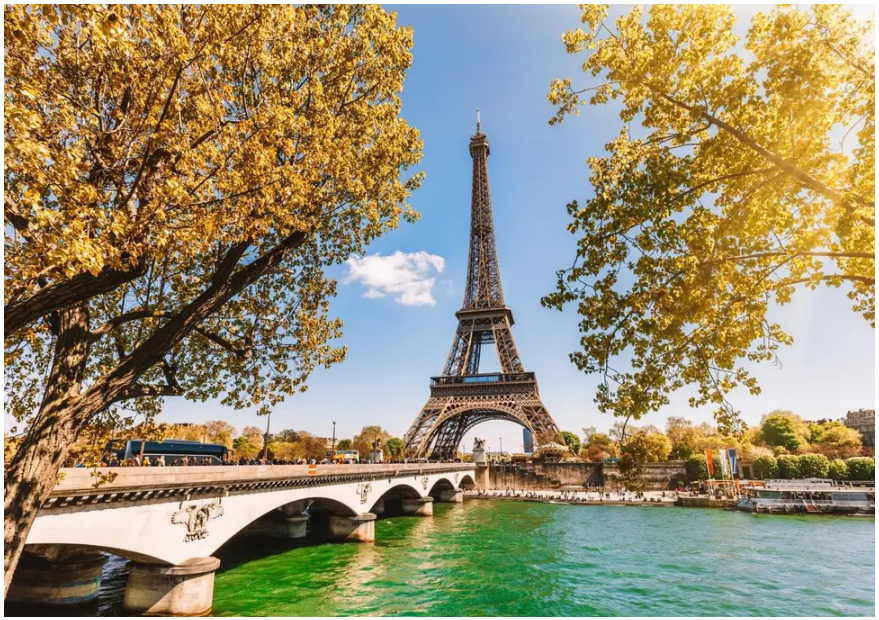


Gene Prediction

- 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?**

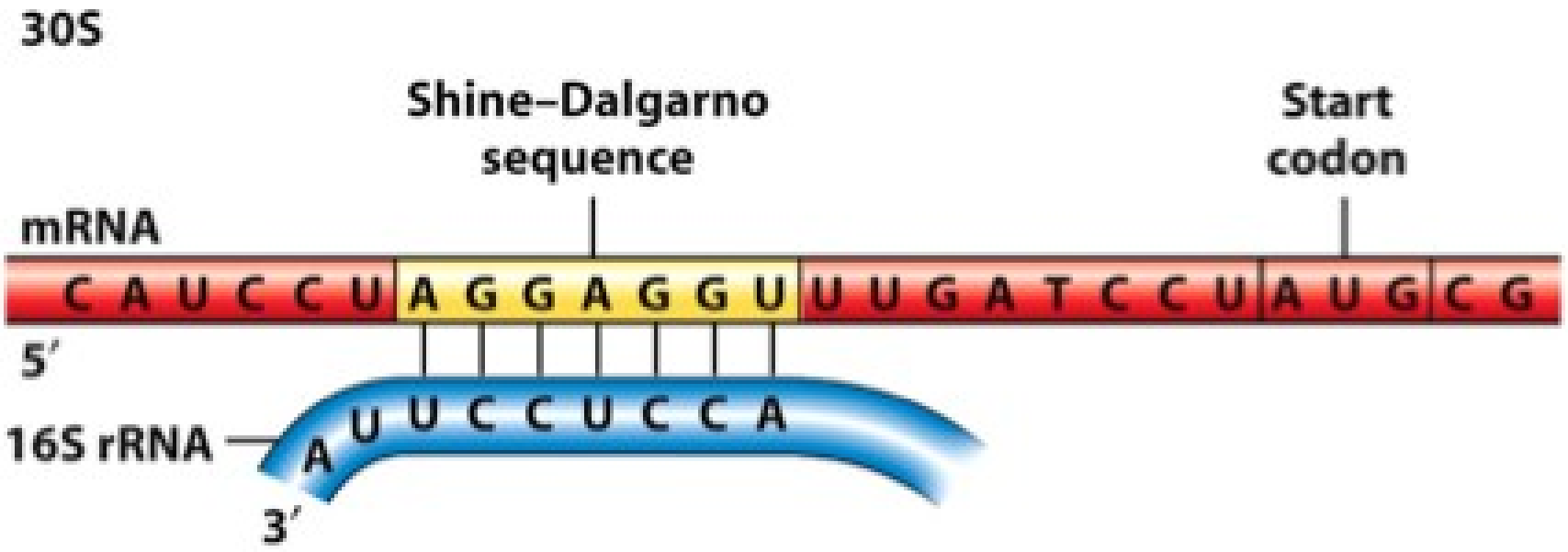


What are *Land Marks*?



Shine-Dalgarno Sequence

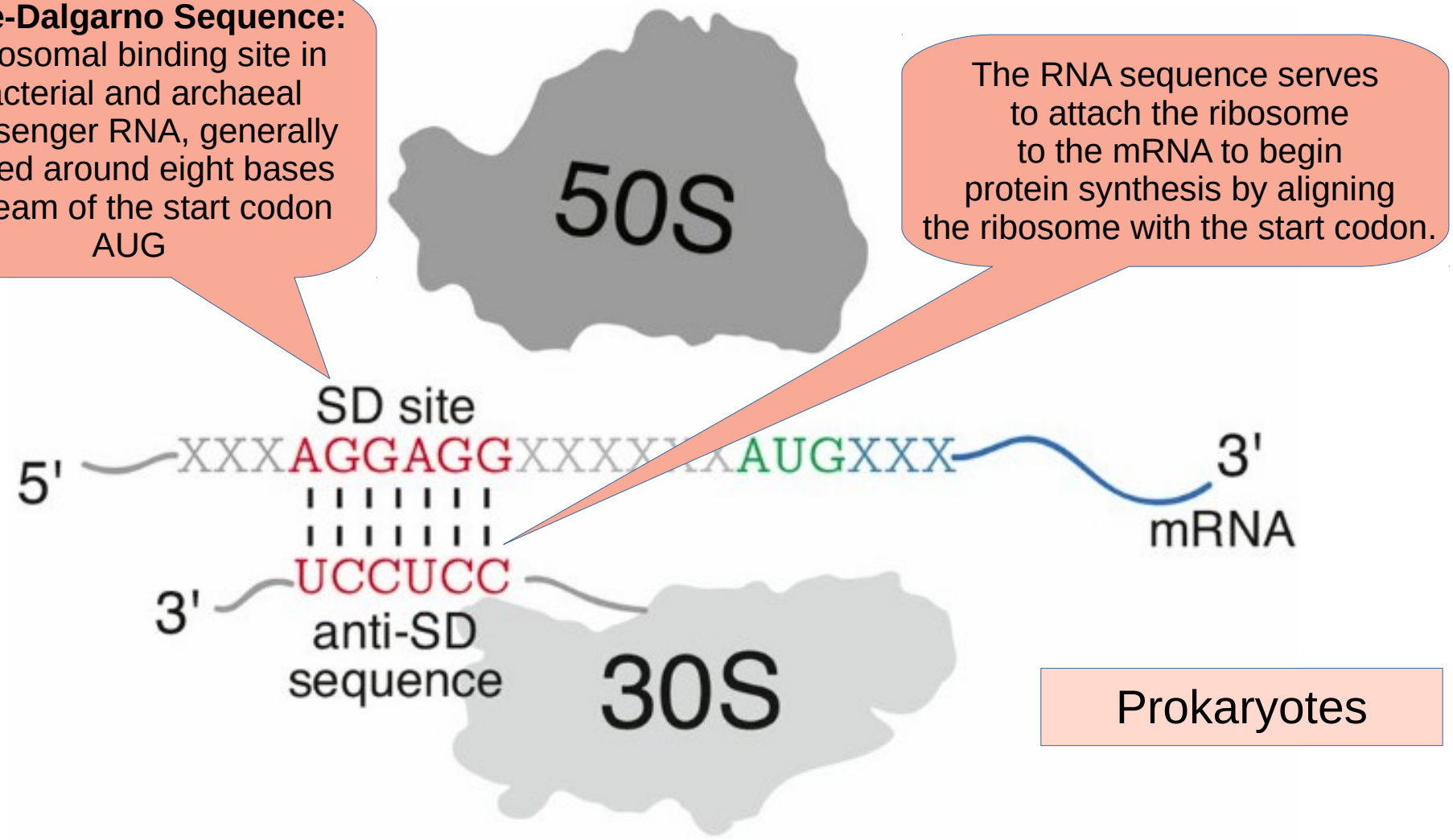
- Shine and Dalgarno showed that the nucleotide tract at the 3' end of E. coli 16S ribosomal RNA (rRNA) is **pyrimidine-rich** and has the sequence: **Py-*ACCUCCU*UA-3'OH**.
- They proposed that these ribosomal nucleotides recognize the complementary purine-rich sequence ***AGGAGGU***, which is found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



Genetic Land Marks?

Shine-Dalgarno Sequence: a ribosomal binding site in bacterial and archaeal messenger RNA, generally located around eight bases upstream of the start codon AUG

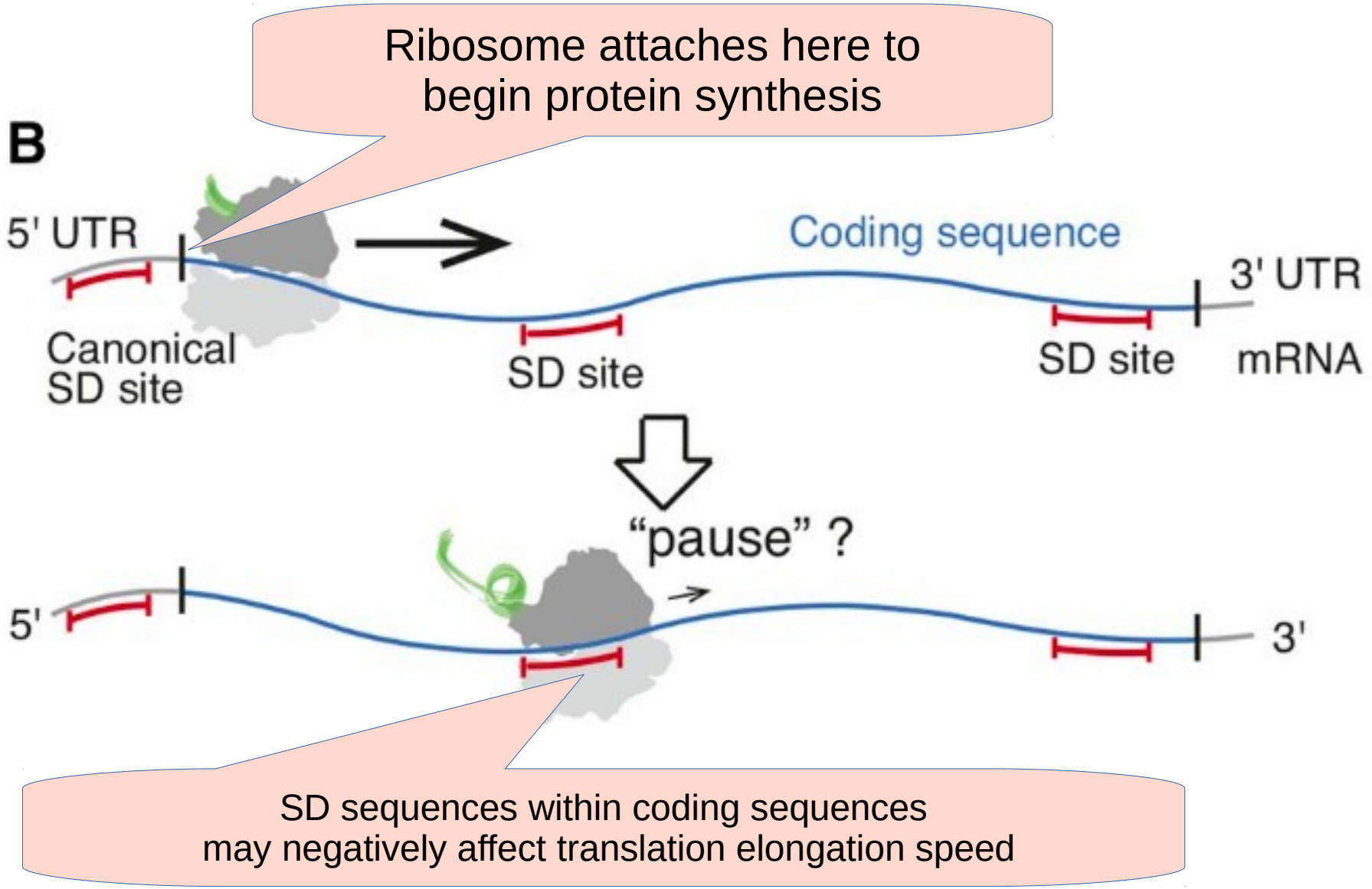
The RNA sequence serves to attach the ribosome to the mRNA to begin protein synthesis by aligning the ribosome with the start codon.



Prokaryotes

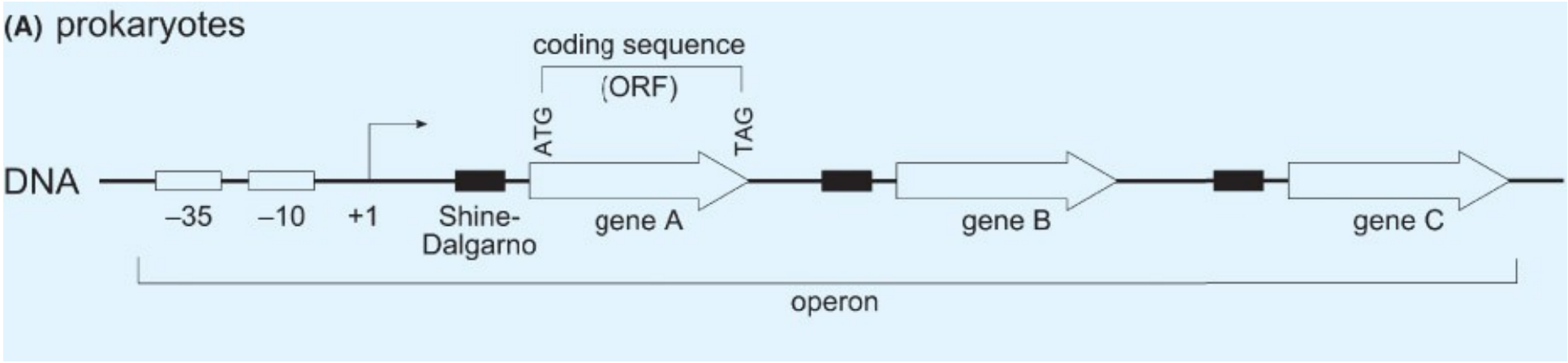
Depletion of Shine-Dalgarno Sequences Within Bacterial Coding Regions Is Expression Dependent, Chuyue Yang, Adam J. Hockenberry, Michael C. Jewett and Luís A. N. Amaral, https://www.g3journal.org/content/6/11/3467

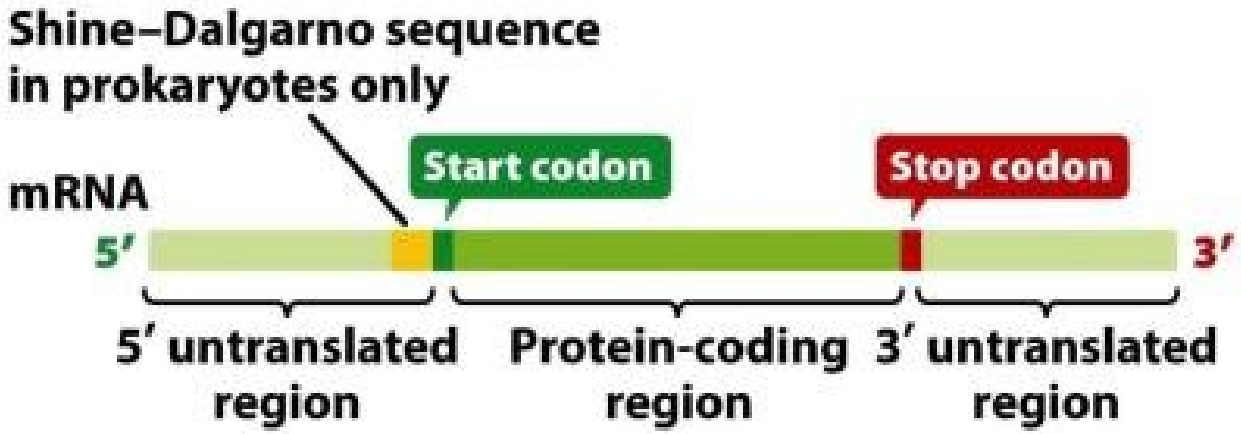
Genetic Land Marks?



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





Prediction Algorithms

- Can you find any sense in the below sequence?

Lo gicwi llg etyo ufro mAt oB.
Ima ginat ion wi llge ty
oue ve rywh ere.

- How did you find the meaning here?
- How would an algorithm do it?

Prediction Algorithms

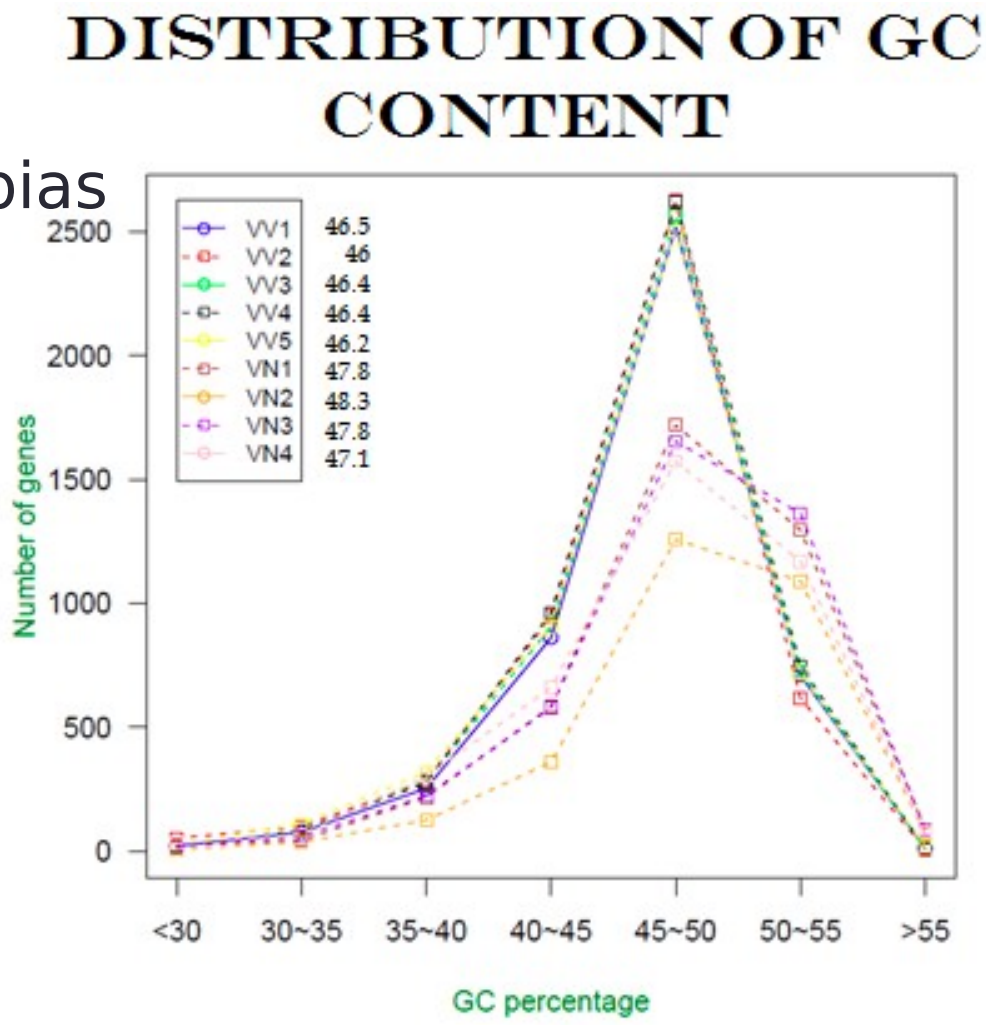
- Alignment-based – find genes/features based on conserved sequences is well-studied organisms (database searching)
  - Automatic assignment based on sequence similarity (best BLAST hit): gene name, protein name, function
  - Quality vs Quantity: How much time do you have to find this gene? Heuristic-based, or exhaustive search

Prediction Algorithms

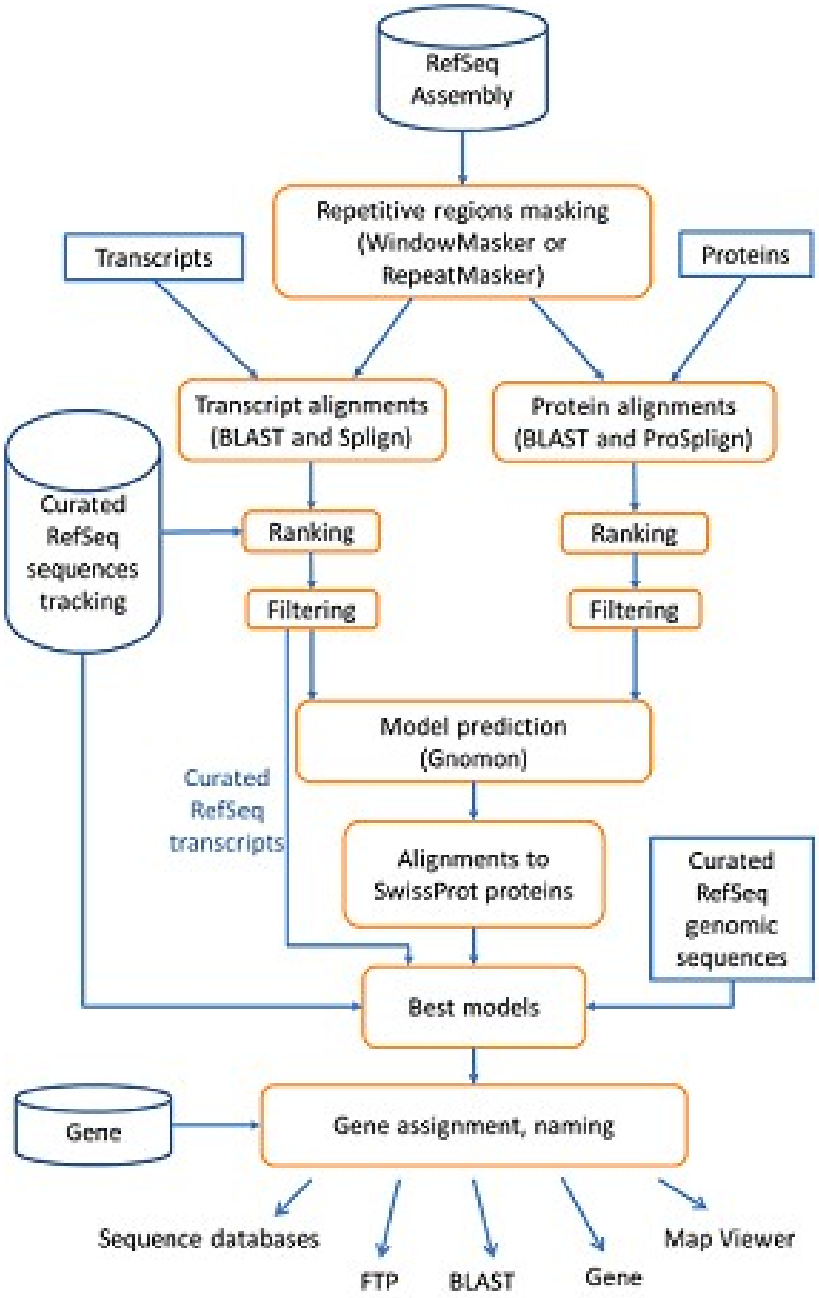
- Content-based – consider overall properties of the sequence when making predictions
- Nucleotide frequency
- Codon frequency/codon bias

- GC content for all *V. vulnificus* and *V.naverensis* gene predictions (Figure)

- Most of the genomes contained a high percentage of genes with GC contents between 45-50%.

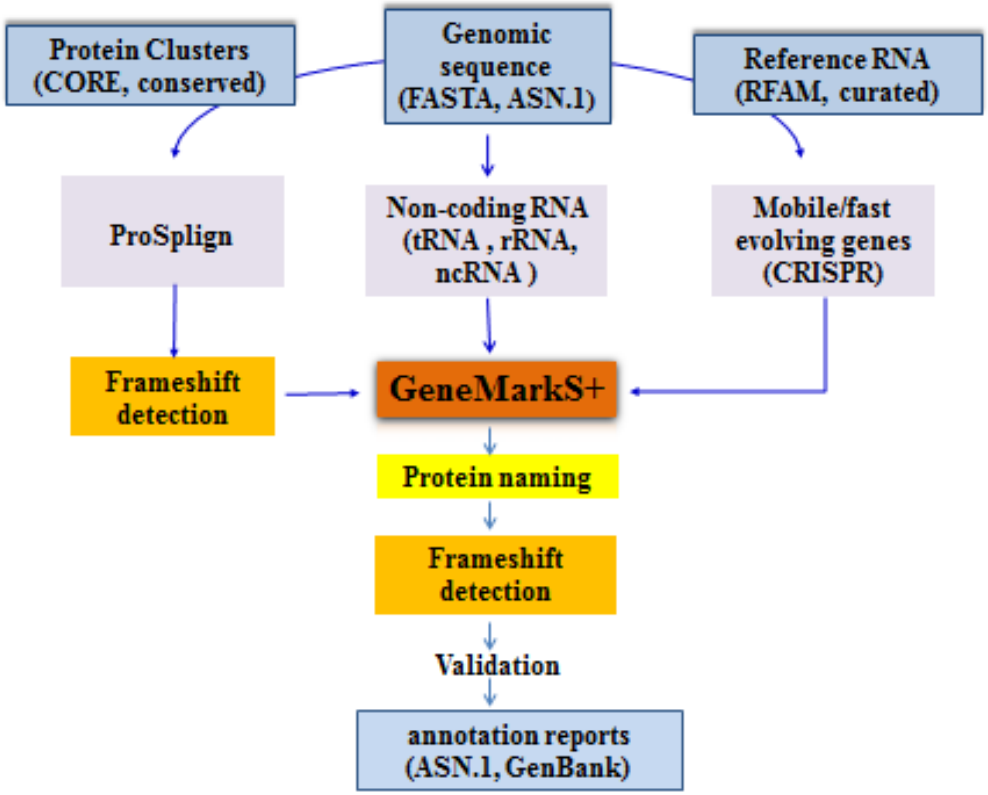


Prediction Algorithms



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

NCBI Prokaryotic Annotation Pipeline

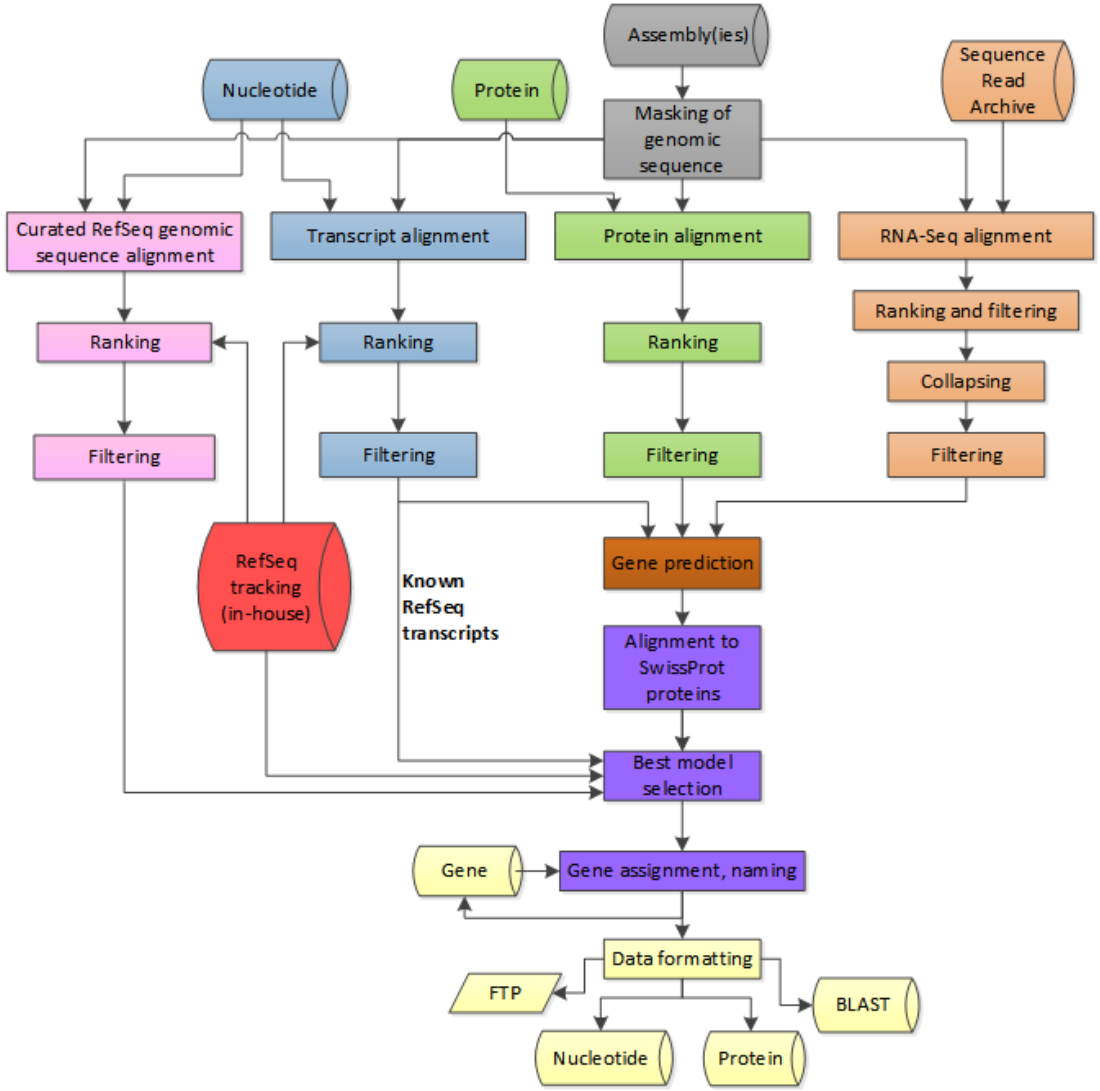


- Combines sequence-based algorithm with alignment-based approach
  - Protein-coding genes
  - Structural RNAs (5S, 16S, 23S)
  - Transfer RNAs
  - Small non-coding RNAs
- Rely only on properties of DNA and training set of genes

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

NCBI Eukaryotic Annotation Pipeline

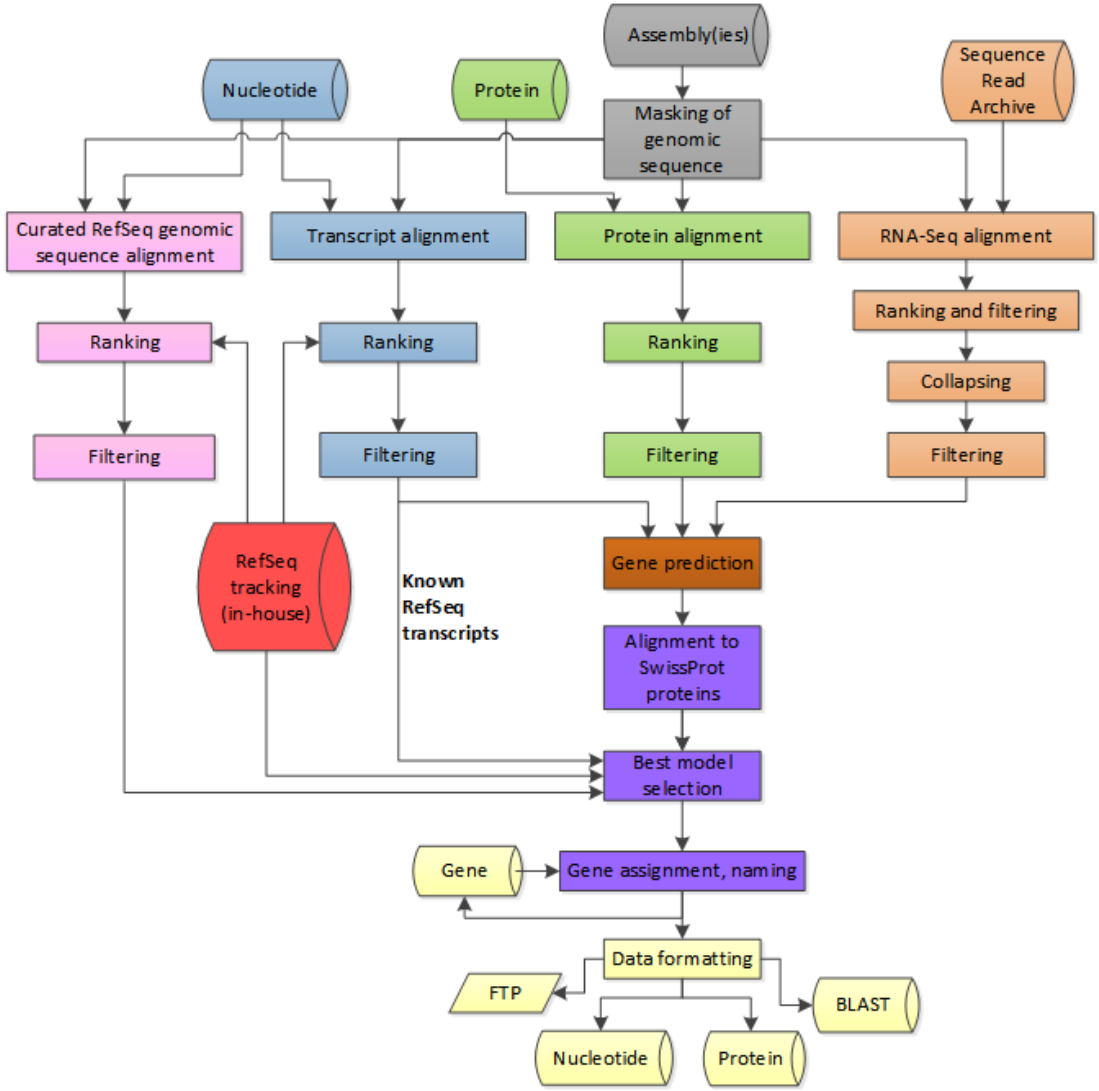
1. Masking
   - Try to identify and ignore non-coding regions
2. Alignment-based predictions
   - *Ask where we have seen this sequence before.*
3. Sequence/content-based predictions from alignment-based
4. Best selected (probability), named, and released



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

NCBI Eukaryotic Annotation Pipeline

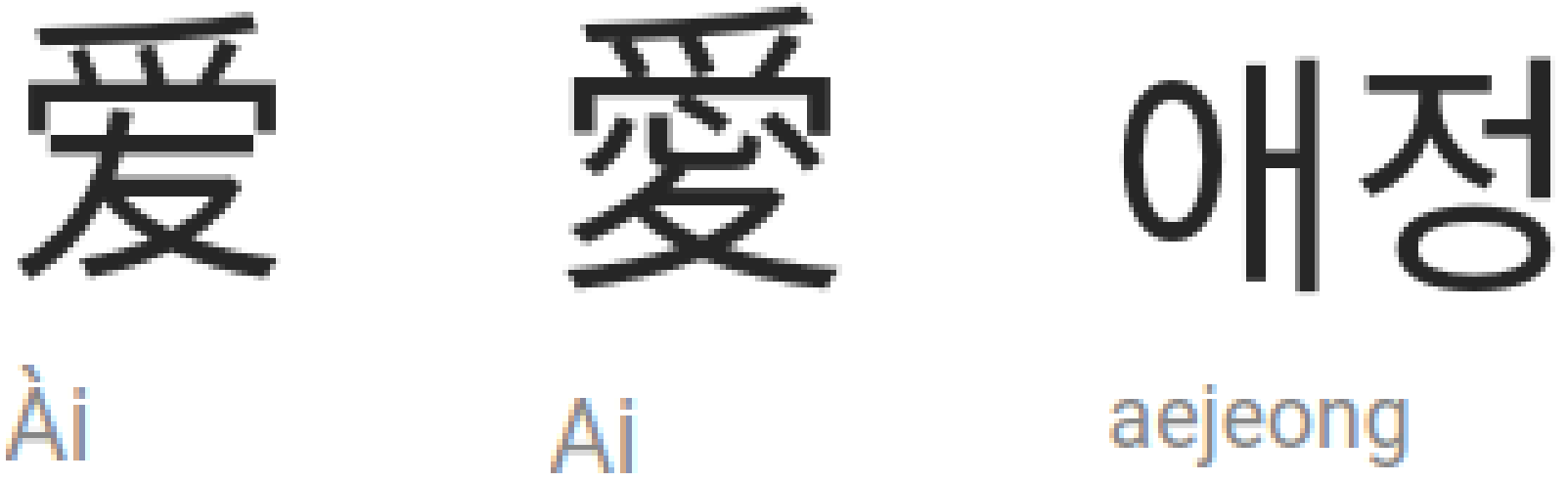
- The best models are selected among the RefSeq and the predicted models, named and accessioned (purple).
- At the end, the annotation products are formatted and deployed to public resources (yellow).



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

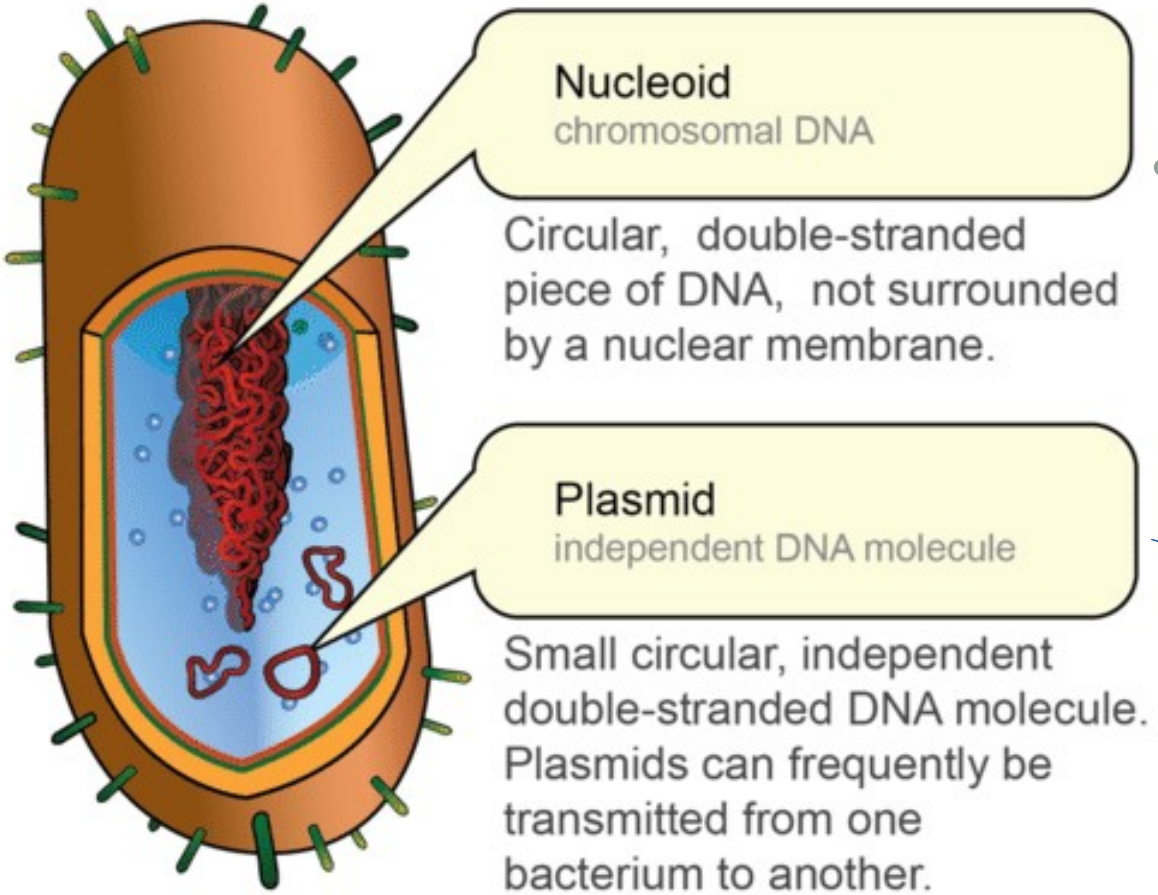
Natural Differences

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



“Love” in Chinese, Japanese and Korean

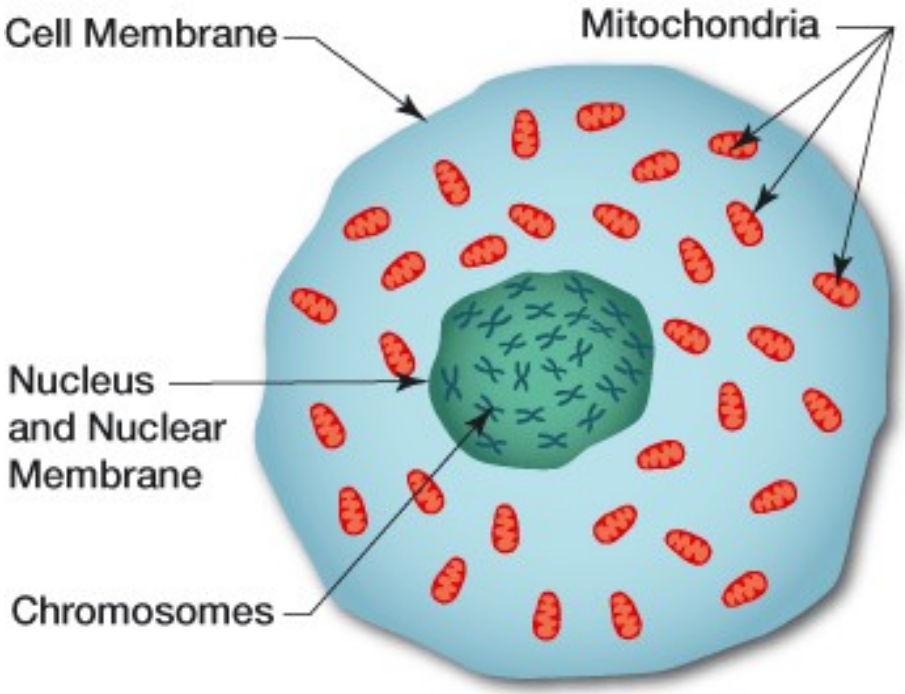
Prokaryotic versus Eukaryotic Genomes



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

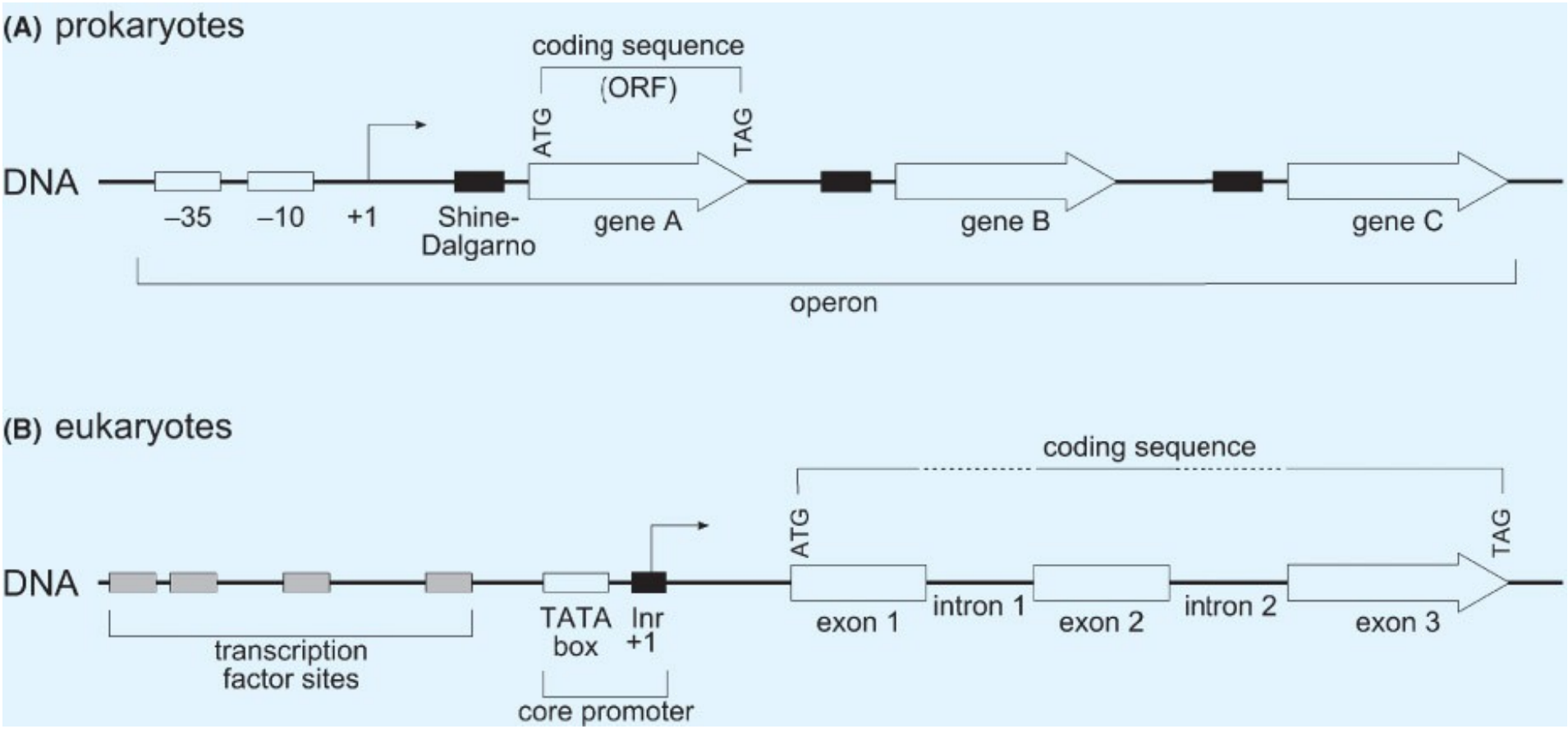
Different types of genomes require different approaches to find differences...

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)

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

https://www.ncbi.nlm.nih.gov/orffinder/

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?