

Bioinformatics

CS300

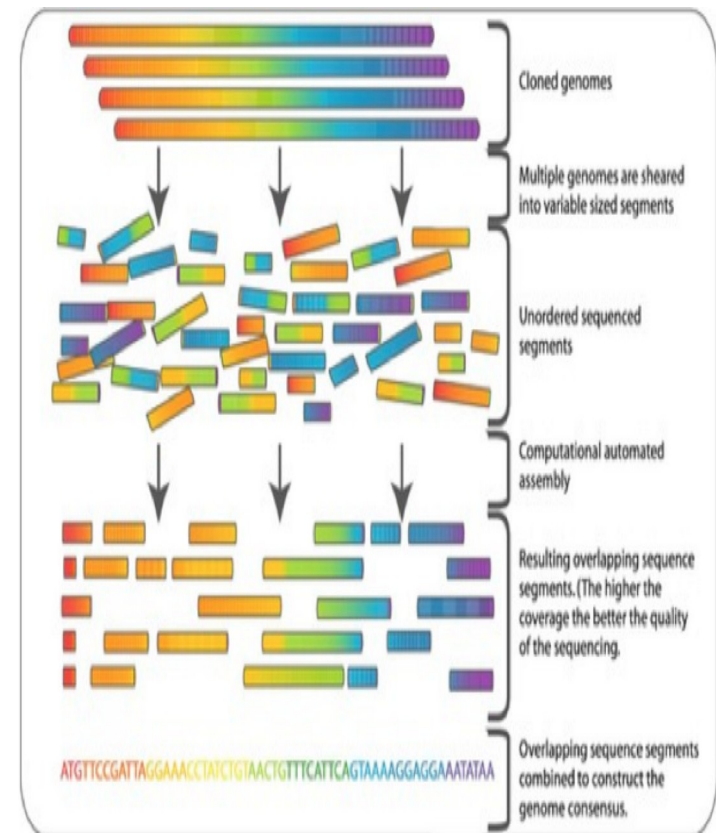
**Genome annotation
and sequence-based
gene prediction**

Spring 2021

Oliver BONHAM-CARTER

Genome Projects

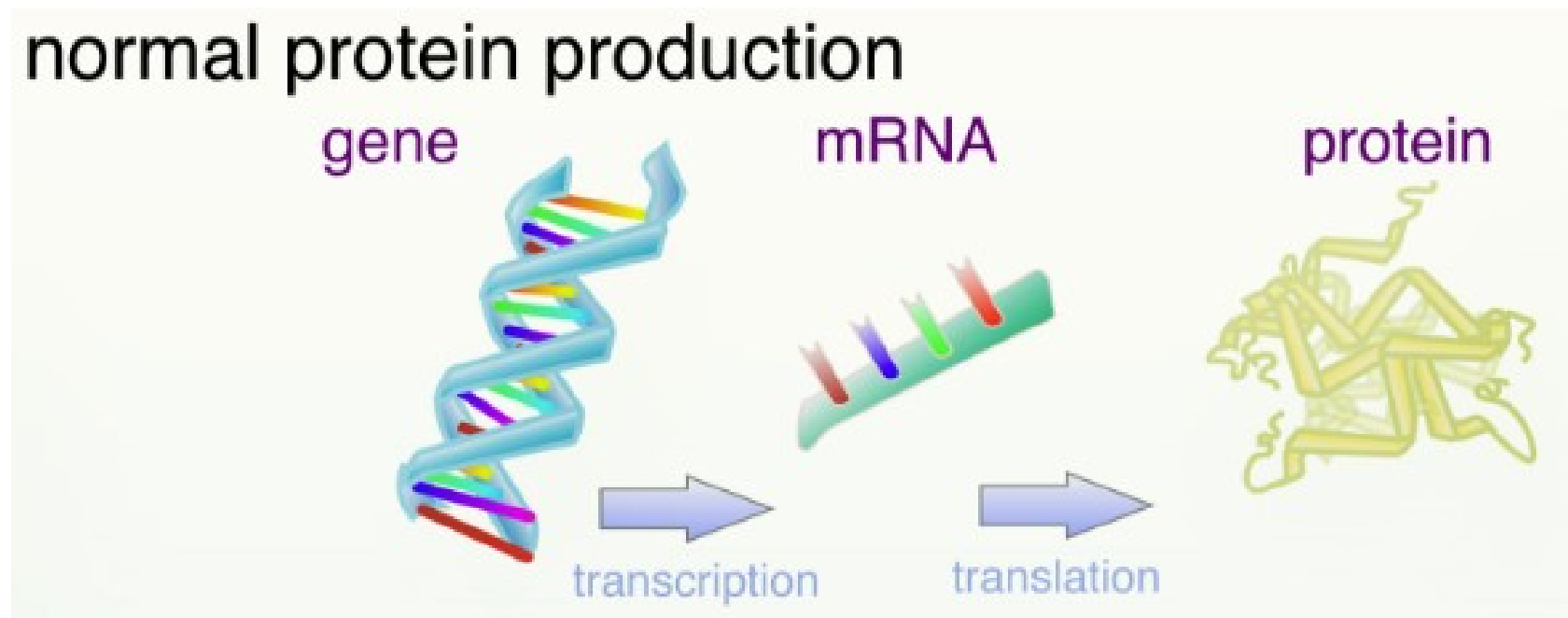
- **Goals:**
 - Determine complete genome sequence of an organism
 - **Annotate protein-coding genes and other important genome-encoded features**
 - find
 - identify
 - characterize
 - describe
 - computational predictions later confirmed at the lab bench



Gene Prediction

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

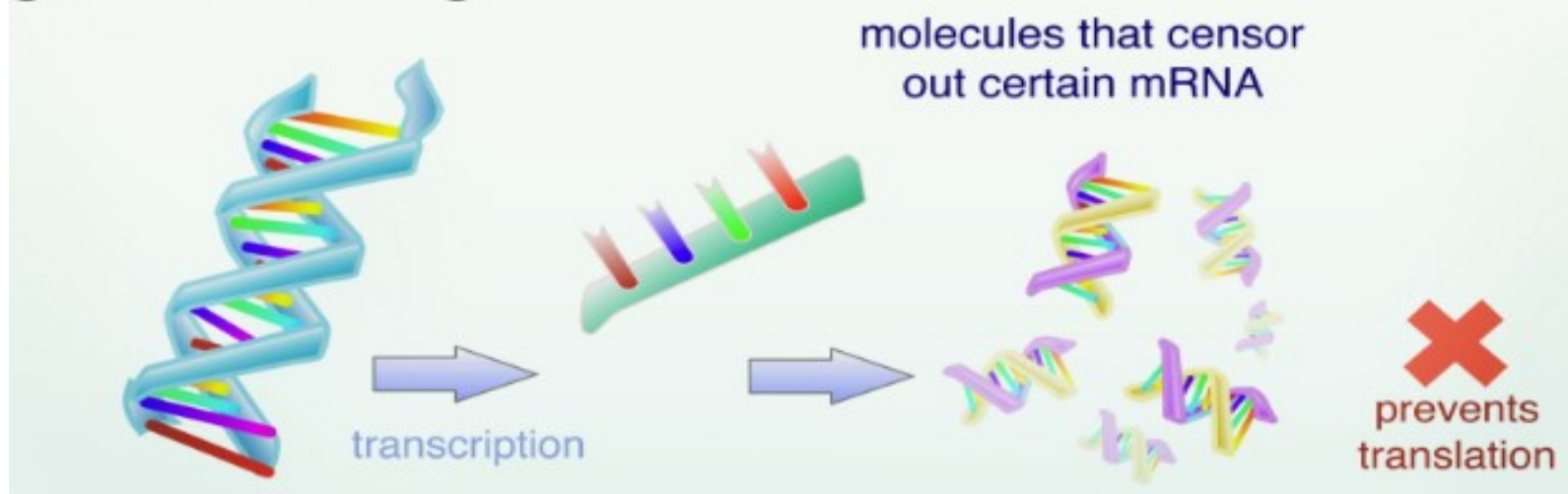
normal protein production



Gene Prediction

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

gene silencing



Not all present genes will make a protein ...



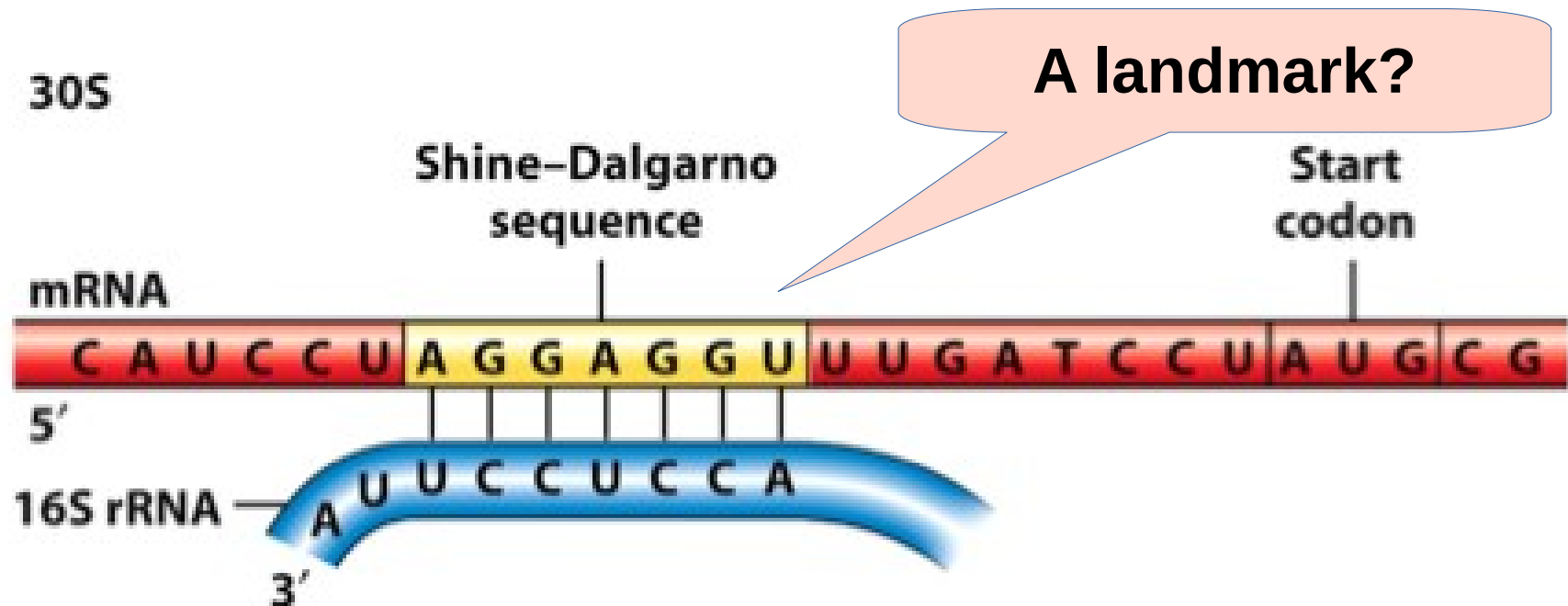
ALLEGHENY
COLLEGE

What are *Land Marks*?



Shine-Dalgarno Sequence

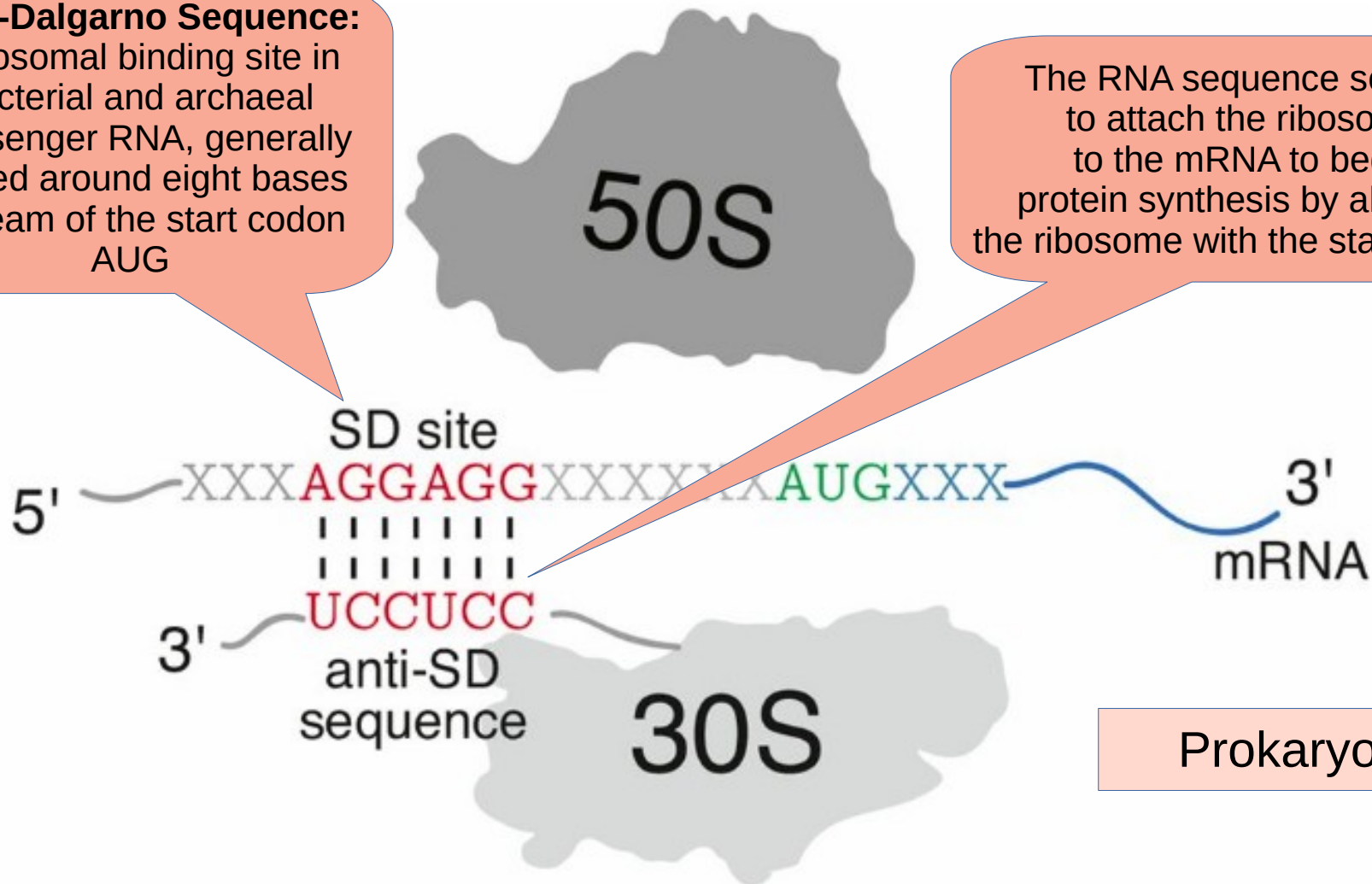
- Shine and Dalgarno showed that the nucleotide tract at the 3' end of E. coli 16S ribosomal RNA (rRNA) is **pyrimidine-rich** and has the sequence: **Py-ACCUCCUUA-3'OH**.
- They proposed that these ribosomal nucleotides recognize the complementary purine-rich sequence **AGGAGGU**, which is found upstream 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.

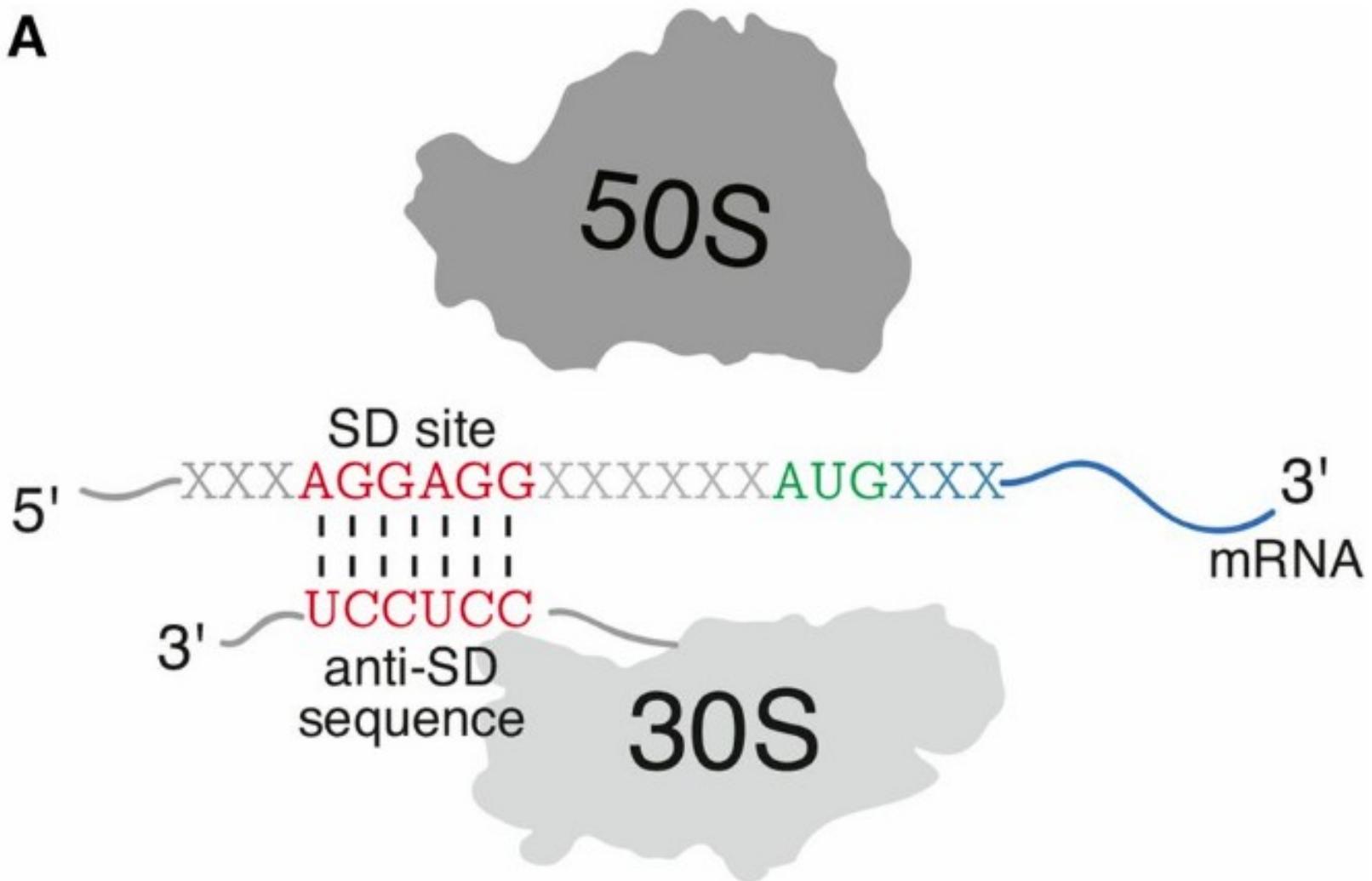


Prokaryotes

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?

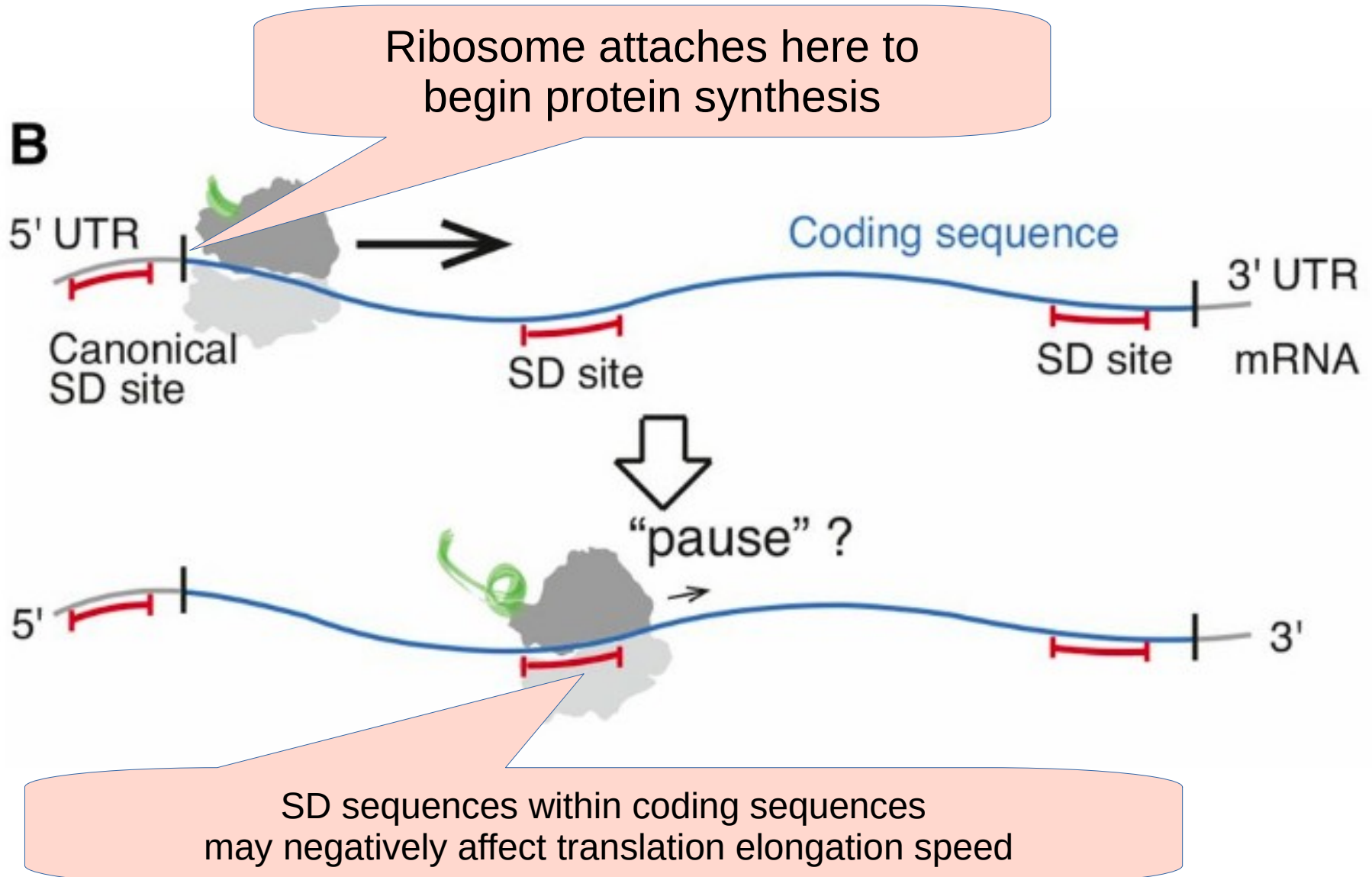
A



More details:

https://en.wikipedia.org/wiki/Shine%E2%80%93Dalgarno_sequence

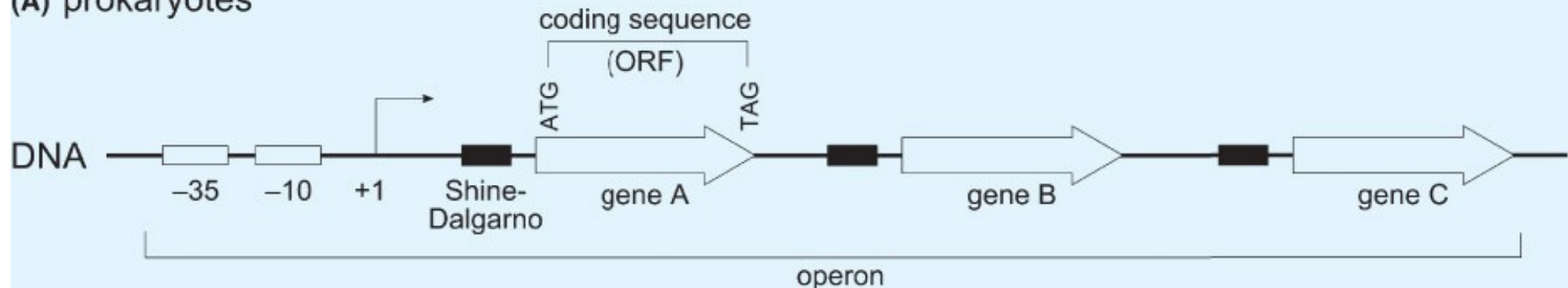
Genetic Land Marks?



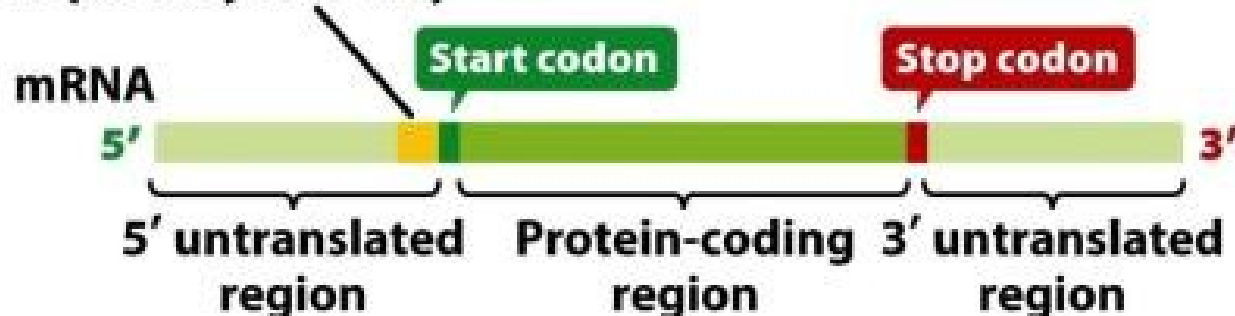
The Familiar to Find the Unfamiliar

- 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

(A) prokaryotes



**Shine-Dalgarno sequence
in prokaryotes only**





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 rywhe re.

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



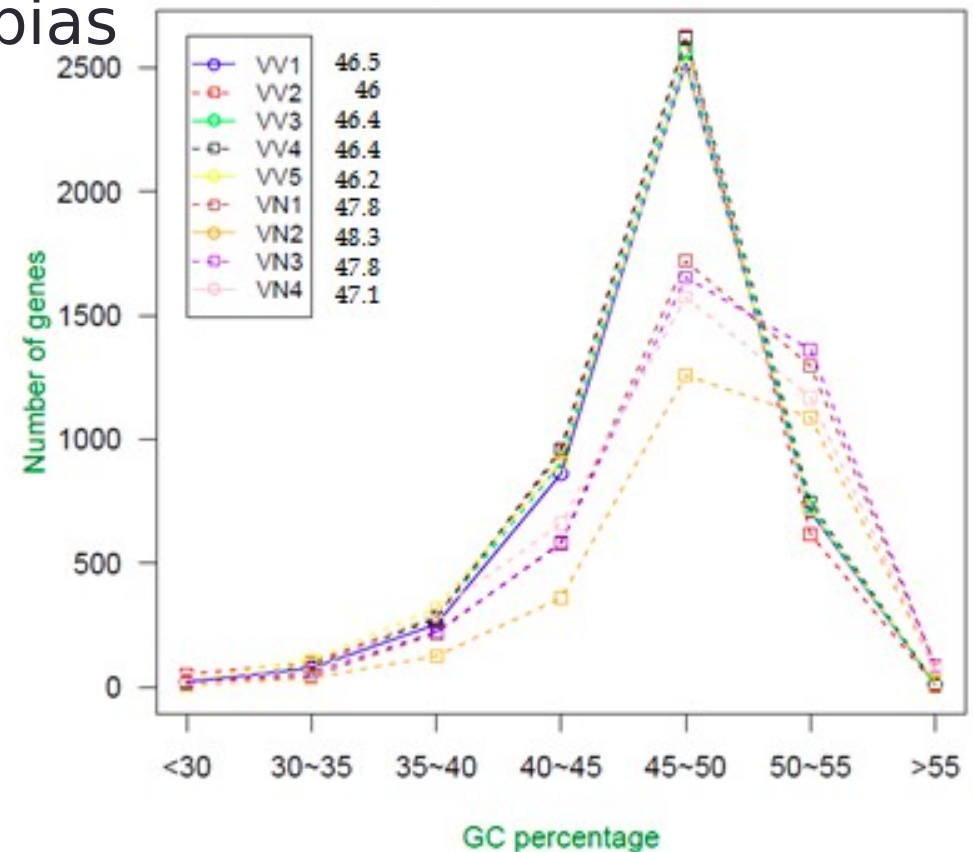
Prediction Algorithms

- Alignment-based – find genes/features based on conserved sequences in 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

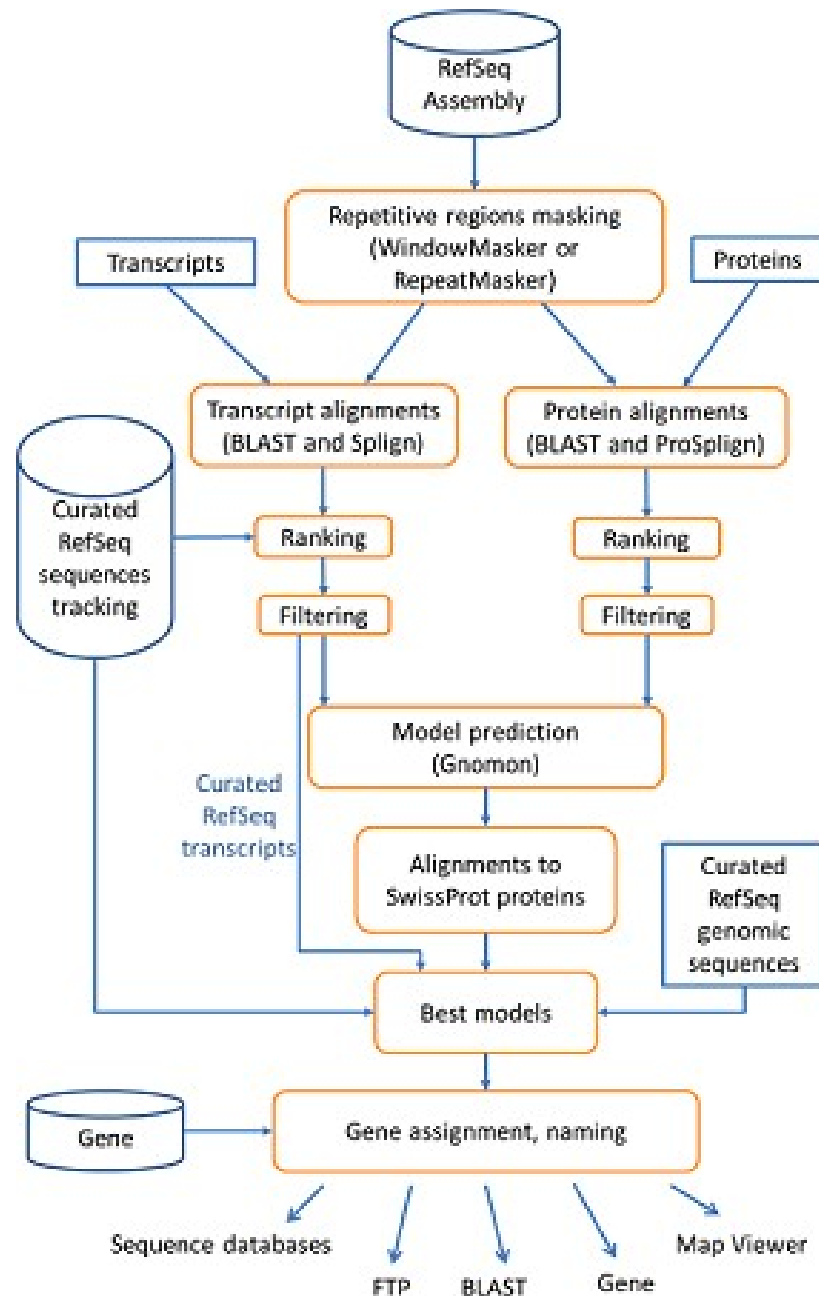
Prediction Algorithms

- 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%.

DISTRIBUTION OF GC CONTENT

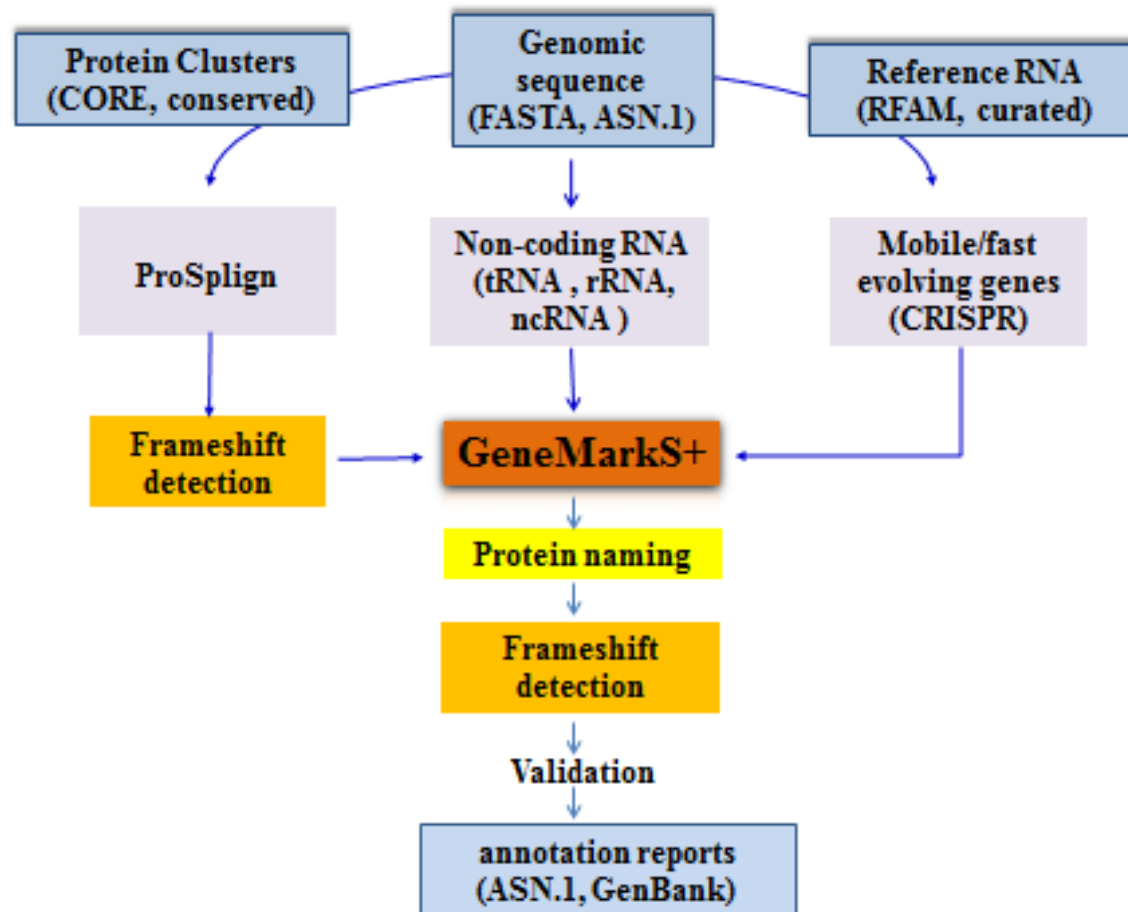


Prediction Algorithms



- **Many stages of analysis to determine genomic artifacts ...**
- Probabilistic – combination of sequence-based and content-based plus probability
- An analysis by the “*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

NCBI Prokaryotic Annotation Pipeline

1. Masking

- Try to identify and ignore non-coding regions

2. Alignment-based predictions

- Ask where we have seen this sequence before (BLAST)*

3. Sequence/content-based predictions from alignment-based

4. Best selected (probability), named, and released



NCBI Eukaryotic Annotation Pipeline

- The best predictions are selected to describe observed artifacts (purple).
- At the end, the annotation products are formatted and deployed to public resources (yellow).



Natural Differences

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

爱

Ài

愛

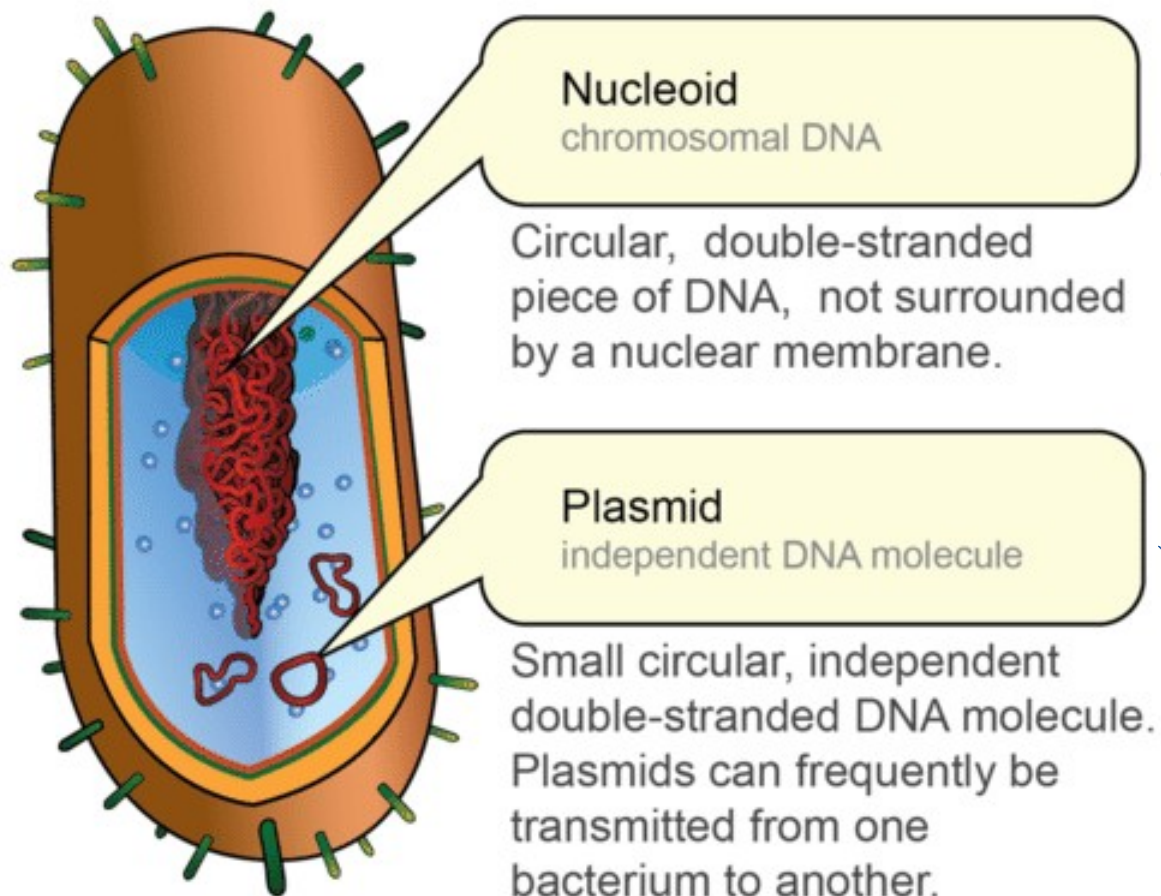
Ai

애정

aejeong

“Love” in Chinese, Japanese and Korean

Prokaryotic versus Eukaryotic Genomes

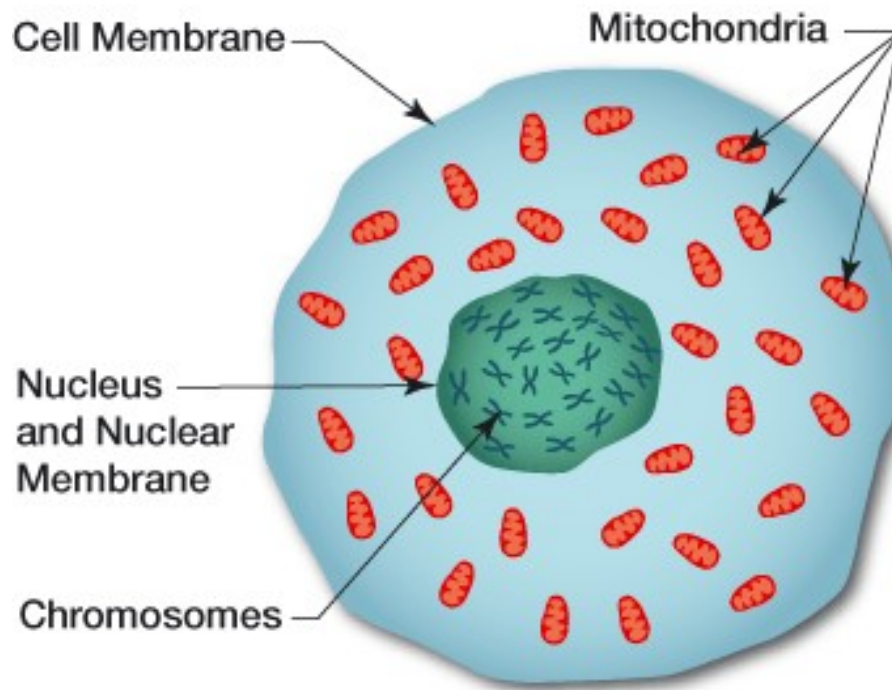


• Prokaryotes

- A circular chromosome
 - “Genome”
- Extra DNA in plasmids
 - smaller, self-replicating

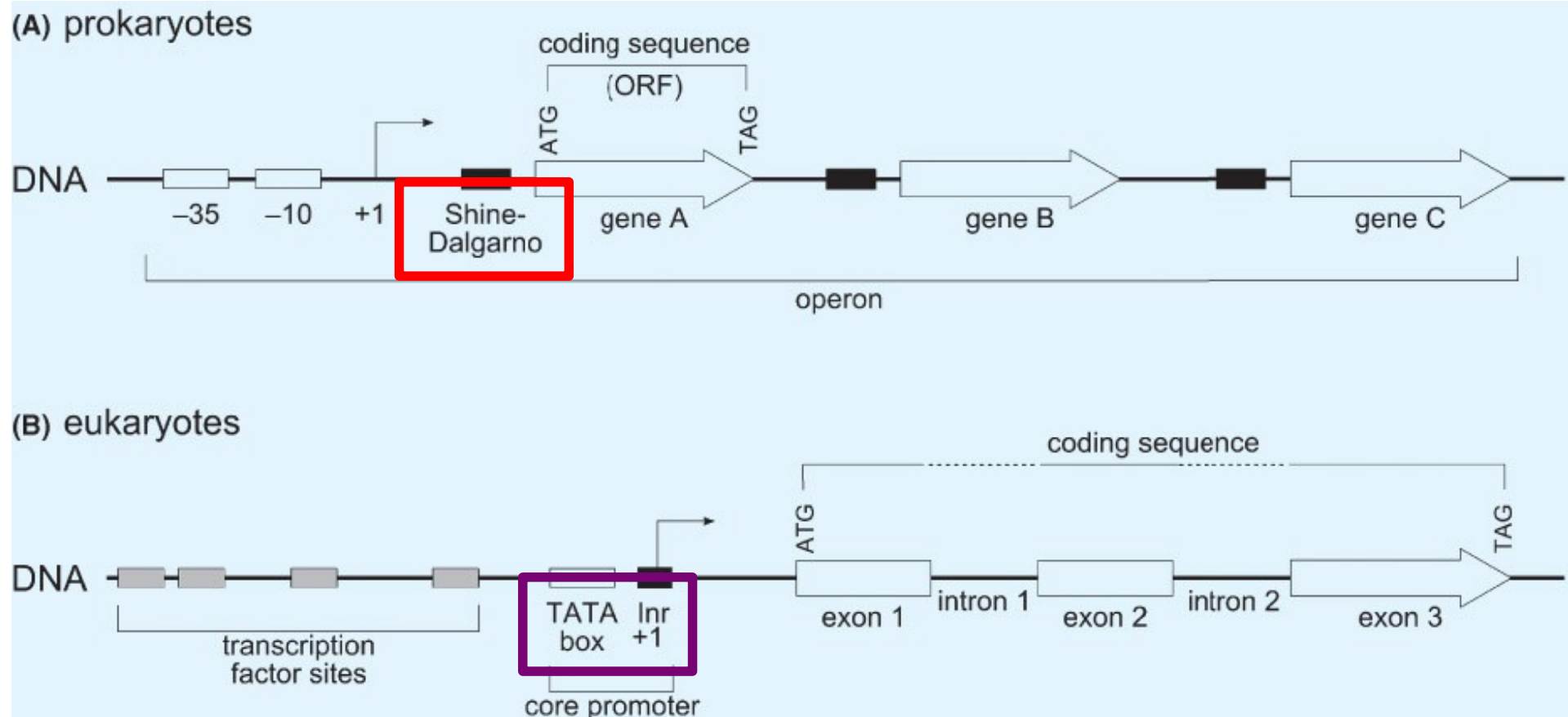
Different types of genomes require different approaches to find genetic differences...

What kind of DNA?!



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

Feature Structures



Comparison of Landmarks in
Prokaryotes vs Eukaryotes



Prokaryotic versus Eukaryotic Genomes

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

Landmarks!



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'
atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataaa

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
```



ALLEGHENY
COLLEGE

Bring the Tool!



Up Next!



Viewing Annotations in NCBI's Open Reading Frame

Cyprinus carpio SIRT1 mRNA, partial cds

GenBank: [KF881970.1](#)

[FASTA](#) [Graphics](#)

[Go to:](#) ☐

LOCUS	KF881970	375 bp	mRNA	linear	VRT 06-NOV-2014
DEFINITION	Cyprinus carpio SIRT1 mRNA, partial cds.				
ACCESSION	KF881970				
VERSION	KF881970.1				
KEYWORDS	.				
SOURCE	Cyprinus carpio (common carp)				
ORGANISM	Cyprinus carpio Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Cyprininae; Cyprinus.				

SIRT1 Accession



Go to NCBI's ORF Tool



Quick link:

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



Enter the Accession Number

Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

KF881970.1

From: To:

Choose Search Parameters

Minimal ORF length (nt): 75 ▼

Genetic code: 1. Standard ▼

ORF start codon to use:

☒ "ATG" only

☐ "ATG" and alternative initiation codons

☐ Any sense codon

Ignore nested ORFs: ☐

Start Search / Clear

Click submit

NCBI ORFfinder tool interface showing the analysis of sequence KF881970.1. The tool identifies three Open Reading Frames (ORFs): ORF1, ORF2, and ORF3. ORF3 is highlighted, and its details are displayed in a pop-up window.

ORF3 Details:

- CDS: ORF3
- Qualifiers: Partial stop
- Name: CDS
- Location: complement(2..316)
- Span on KF881970.1: 315 nt
- Protein length: 105 aa
- KF881970.1 position: 156
- CDS position: 161
- Protein position: 54
- Protein sequence: HLIFILLHGSVEKLW[K]VLAKEDDVWFHDGI

ORF Summary Table:

Label	Strand	Frame	Start	Stop	Length
ORF3	-	3	316	>2	
ORF1	+	2	122	>373	
ORF2	-	1	255	82	



Protein Locations and Options

Cyprinus carpio SIRT1 mRNA, partial cds

ORFs found: 3 Genetic code: 1 Start codon: 'ATG' only

(U) ORFfinder_4.19.224388

ORF3

CDS: ORF3
Qualifiers: Partial stop
Name: CDS
Location: complement(2..316)
[Length]
Span on KF881970.1: 315 nt
Protein length: 105 aa
[Positional Info]
KF881970.1 position: 156
CDS position: 161
Protein position: 54
Protein sequence: HLIFILLHGSVEKLW[K]VLAKEDDVWFHDGI

Download FASTA: ORF3

Links & Tools
BLAST nr: KF881970.1 (2..316)
FASTA record: KF881970.1 (2..316)
GenBank record: KF881970.1 (2..316)

Mark subset... Marked: 0 Download marked set as Protein FASTA ▼

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF3	-	3	316	>2	315 104
ORF1	+	2	122	>373	252 83
ORF2	-	1	255	82	174 57

Get more info about the embedded proteins