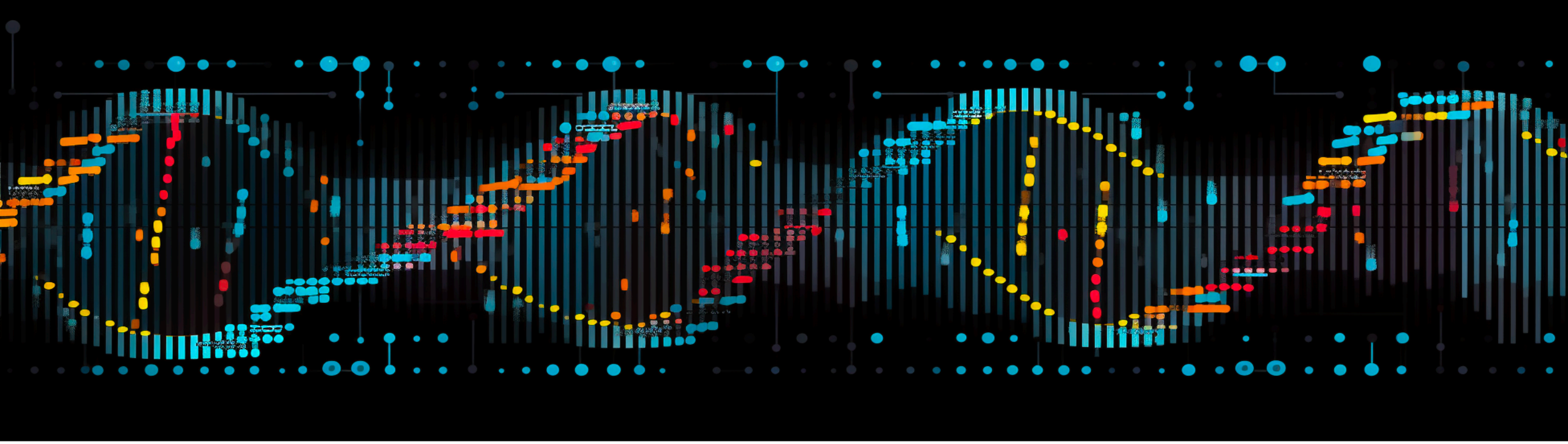
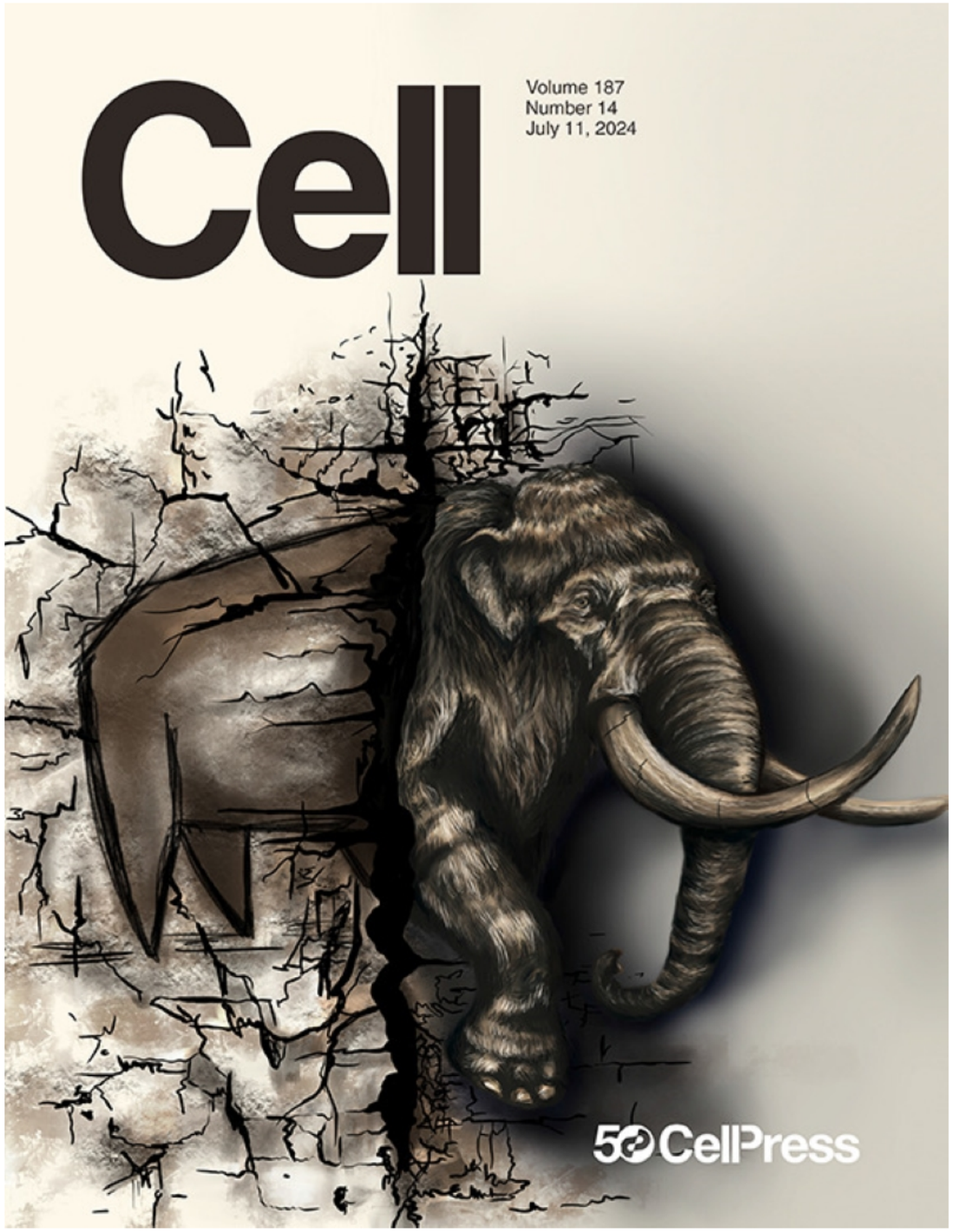


Introduction to bioinformatics

Chapter2 Part1- Gene and RNA

Jiaxing Chen


JURASSIC WORLD
EVOLUTION 3



Mammoth fossil → Ancient DNA → Mammoth genome

How to get the genome?

?

Outline-Chapter 2

- **Gene and RNA**
- Genome and Transcriptome, NGS, Assembly
- Alignment and BLAST
- Protein and CRISPR

Outline-Chapter2-Part1

- **Gene**
- Computational Gene prediction
- RNA
- Noncoding RNA world
- RNA secondary structure

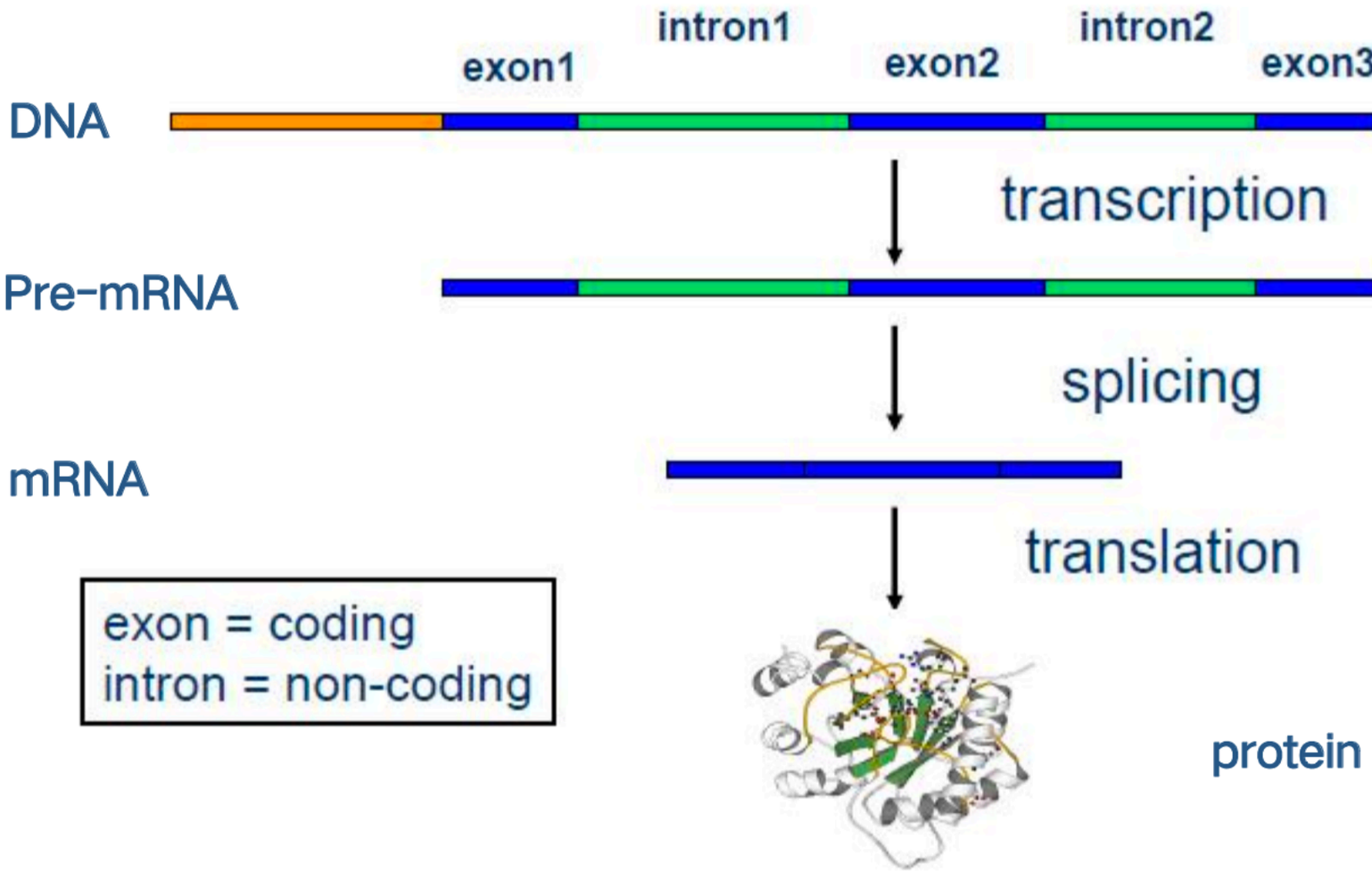
What is a gene?

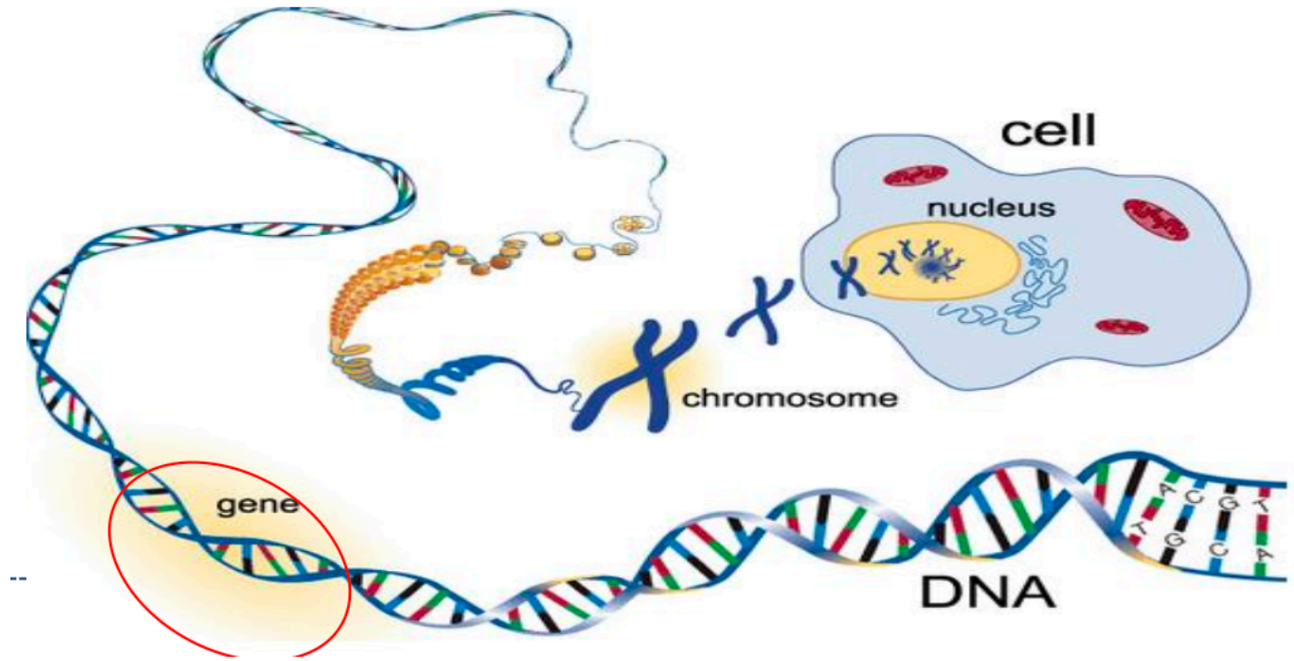
The word *gene* was first used by Wilhelm Johannsen in 1909, based on the concept developed by Gregor Mondel in 1866.

New definition: 

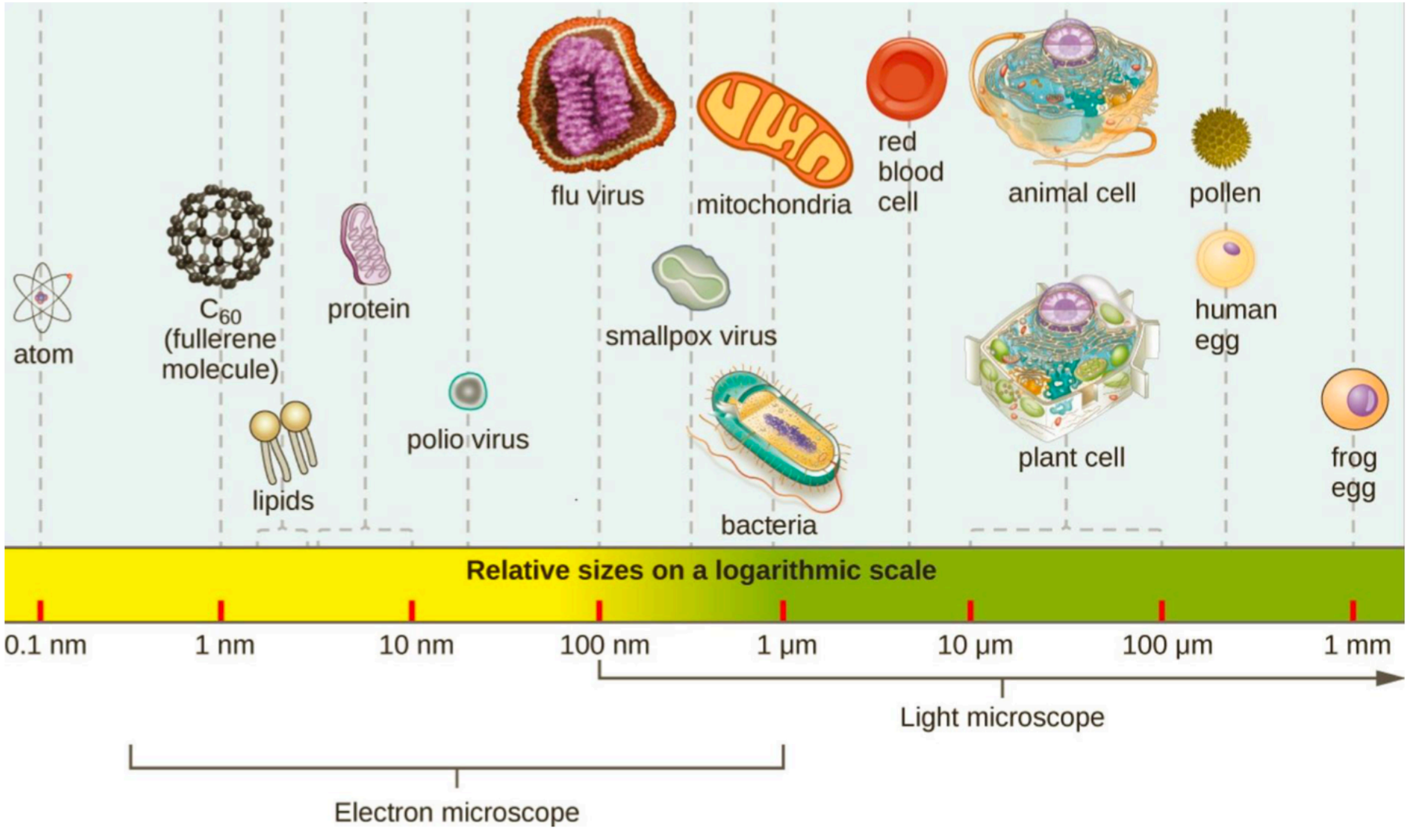
- A gene is a locus (or region) of DNA that encodes a functional protein or RNA product, and is the molecular unit of heredity.

The central dogma and splicing




cell
nucleus
chromosome
gene
DNA

How big are human cells?



How long is our genome?

Data

> Estimated number of eukaryotic (human) cells in the human body: 1.0×10^{13} （十兆）

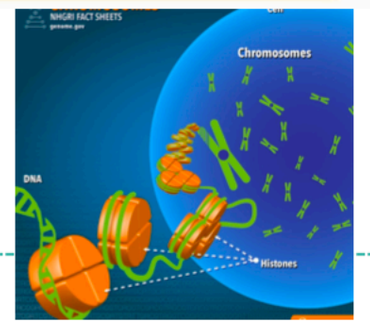
—Wolfram Alpha

> The haploid human genome (23 chromosomes) is estimated to be about 3.2 billion base pairs long.

—Human Genome – Wikipedia

The full DNA content of a cell is therefore 6.4 Gbp.

> Average base pair size: one bp corresponds to approximately 3.4 Å of length along the strand



Codon Table

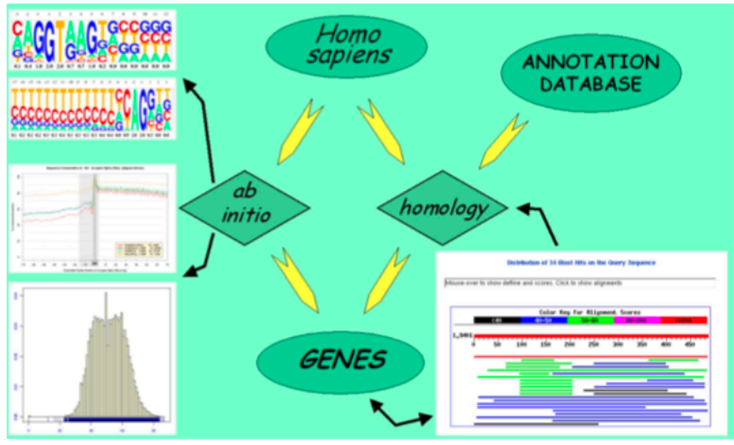
Seconed Position

First Position	U	code	Amino Acid	C	code	Amino Acid	A	code	Amino Acid	G	code	Amino Acid	Third Position
	U	UUU	phe		UCU	ser		UAU	tyr		UGU	cys	U
		UUC			UCC			UAC			UGC		C
		UUA	leu		UCA			UAA	STOP		UGA	STOP	A
		UUG			UCG			UAG	STOP		UGG	trp	G
	C	CUU	leu		CCU	pro		CAU	his		CGU	arg	U
		CUC			CCC			CAC			CGC		C
		CUA			CCA			CAA	gln		CGA		A
		CUG			CCG			CAG			CGG		G
	A	AUU	ile		ACU	thr		AAU	asn		AGU	ser	U
		AUC			ACC			AAC			AGC		C
		AUA			ACA			AAA	lys		AGA	arg	A
		AUG	met		ACG			AAG			AGG		G
	G	GUU	val		GCU	ala		GAU	asp		GGU	gly	U
		GUC			GCC			GAC			GGC		C
		GUA			GCA			GAA	glu		GGA		A
		GUG			GCG			GAG			GGG		G

Finding all genes in a genome could be hard

- Finding all the genes is hard

  - Mammalian genomes are large ~3 billion bp
  - Only < 2% coding proteins
  - Non-coding RNAs are more difficult to be predicted



COMPUTATIONAL GENE PREDICTION

- A **gene:** a nucleotide sequence that codes for a protein
- **Gene prediction**: given a genome, locate the beginning and ending position of every gene.

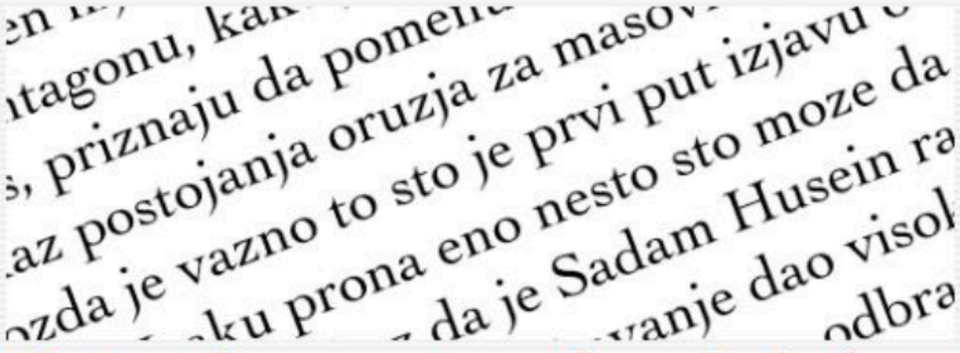
```
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcg
gctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgg
gatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttgga
atatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagc
tgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatcctgcggctatgctaatgaatggtcttgggatttaccttggaatgct
aagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcgg
ctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggct
atgcaagctgggatccgatgactatgcttaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgct
aagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaag
ctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatggtct
tgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttacctt
ggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgc
taagctcatgcgg
```

GENE FINDING APPROACHES

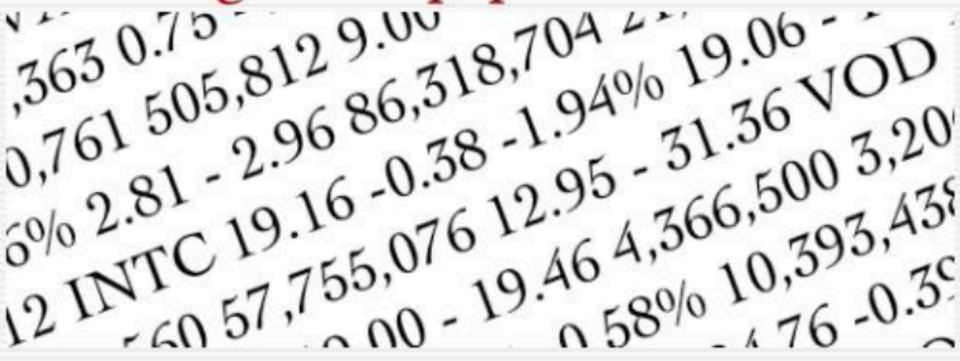
Computational Methods:

- Something that matches statistical patterns common to all genes (ab initio)
- Something that matches an already known gene (homology)
- Hybrid

STATISTICAL APPROACH: METAPHOR IN UNKNOWN LANGUAGE



- Noting the differing frequencies of symbols (e.g. *'%'*, '.', '-') and numerical symbols could you distinguish between a story and a stock report in a foreign newspaper?



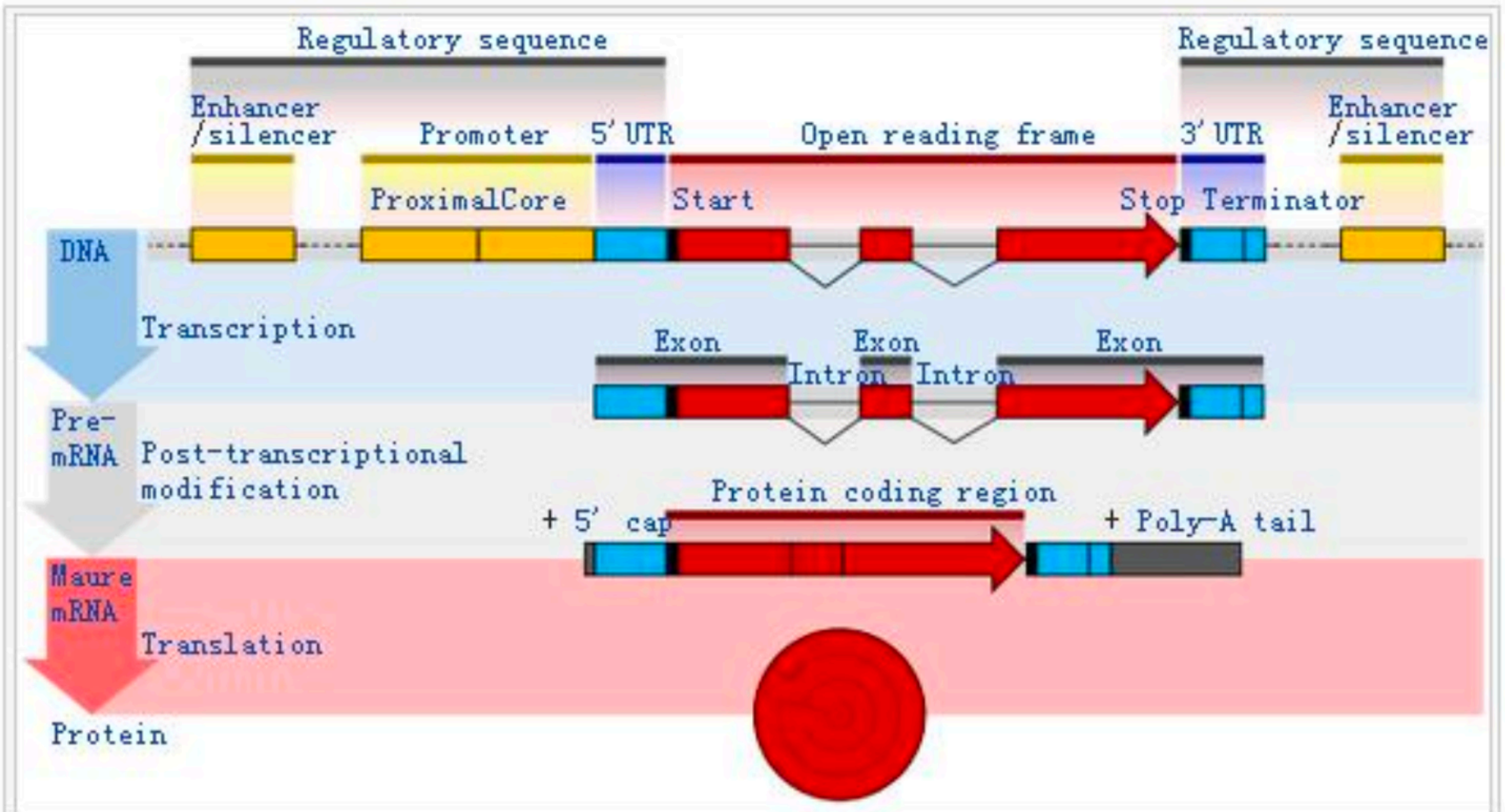
WHAT CAN WE MEASURE ABOUT GENES?

- ORF (Open Reading Frame): a sequence started by ATG and terminated by a stop codon (a.g TAA, TAG, TGA)

- Codon Usage: the preference for using specific synonymous codons most frequently measured by CAI (Codon Adaptation Index)

- Features and motifs
  - Promoters, splice sites, enhancers, untranslated regions (UTRs)

The structure of eukaryotic (真核生物的) genes

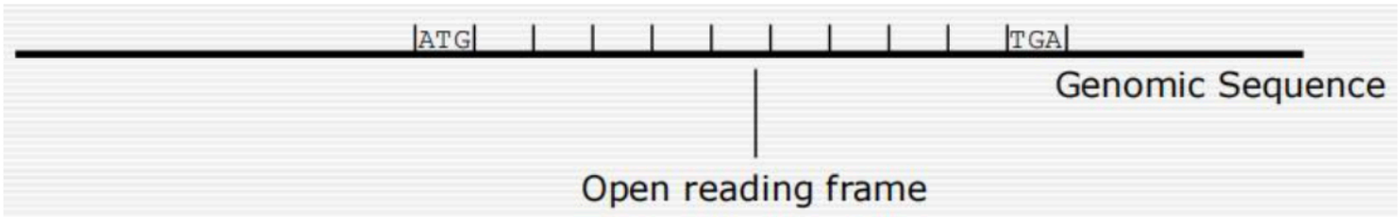


The structure of a eukaryotic protein-coding gene. Regulatory sequence controls when and where expression occurs for the protein coding region (red). Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified to add a 5′ cap and poly-A tail (grey) and remove introns. The mRNA 5′ and 3′ untranslated regions (blue) regulate translation into the final protein product.

OPEN READING FRAMES

Detect potential coding regions by looking at ORFs

- A genome of length n is comprised of (n/3) codons
- Stop codons break genome into segments
- The subsegments of these that start from the Start codon (ATG) are ORFs
- Some ORFs can overlap and code for different genes!



USING KNOWN GENES TO PREDICT NEW GENES

- Some genomes may be very well-studied, with experimentally verified genes.

- Closely-related organisms may have similar genes

- Unknown genes in one species may be compared to genes in a sufficiently closely-related species

- The idea is that gene structure is on average quite stable.

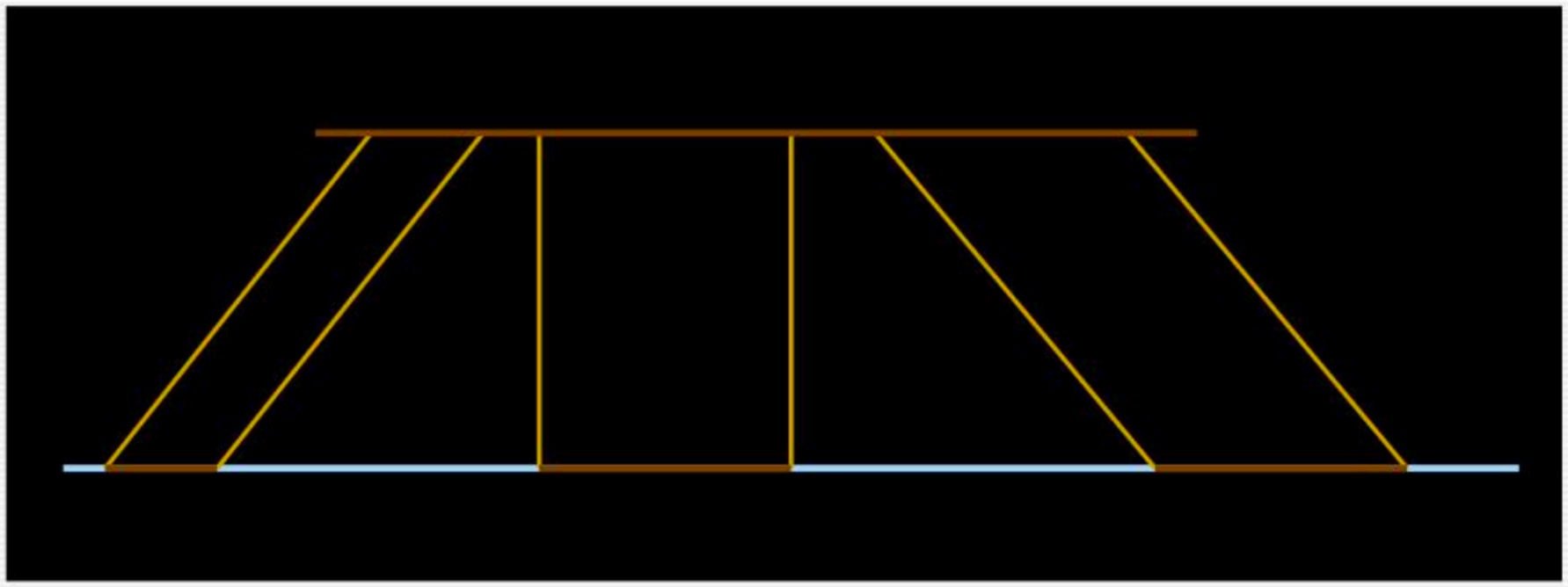
SIMILARITY-BASED APPROACH TO GENE PREDICTION

- Genes in different organisms are similar

- The similarity-based approach uses known genes in one genome to predict (unknown) genes in another genome

- **Problem:** Given a known gene and an un-annotated genome sequence, find a set of substrings of the genomic sequence whose concatenation best fits the known gene

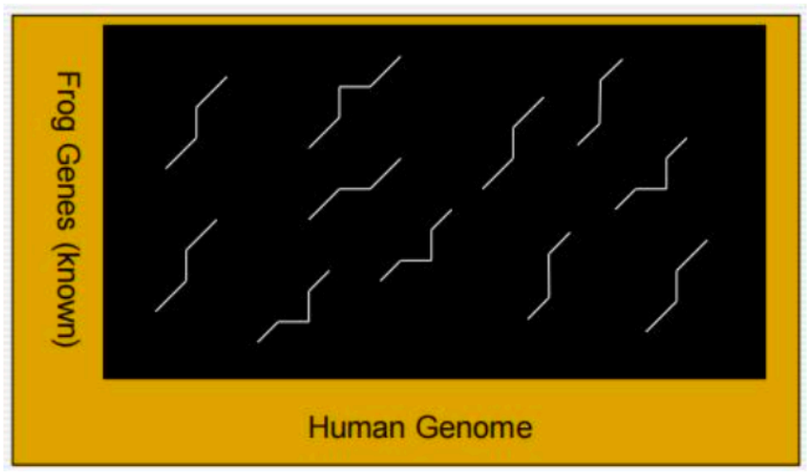
COMPARING GENES IN TWO GENOMES



SMALL ISLANDS OF SIMILARITY
CORRESPONDING TO SIMILARITIES
BETWEEN EXONS

USING SIMILARITIES TO FIND THE EXON STRUCTURE

- The known frog gene is aligned to different locations in the human genome
- Find the "best" path to reveal the exon structure of human gene
- Start with a local alignment to find putative exons



CHAINING LOCAL ALIGNMENTS

- Find substrings that match a given gene sequence (candidate exons); use a cutoff to define significance.

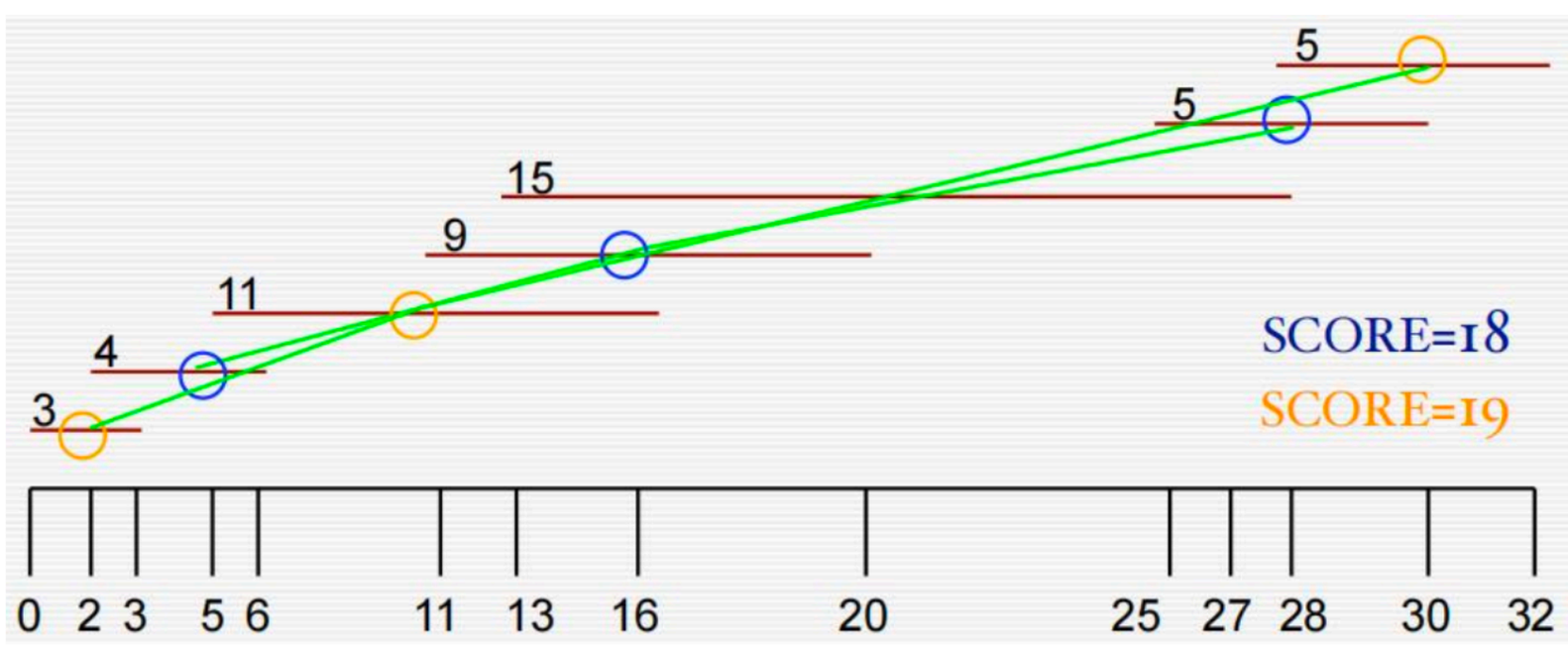
- Define a candidate exon as (l, r, w): left position, right position, weight defined as score of local alignment

- Look for a maximum chain of substrings, i.e. a set of non-overlapping non-adjacent intervals.

EXON CHAINING PROBLEM

- Locate the number and beginning and end of each interval (2n points)

- Find the “best”, i.e. maximum weight path



EXON CHAINING PROBLEM: FORMULATION

- Exon Chaining Problem: Given a set of putative exons, find a maximum set of non-overlapping putative exons

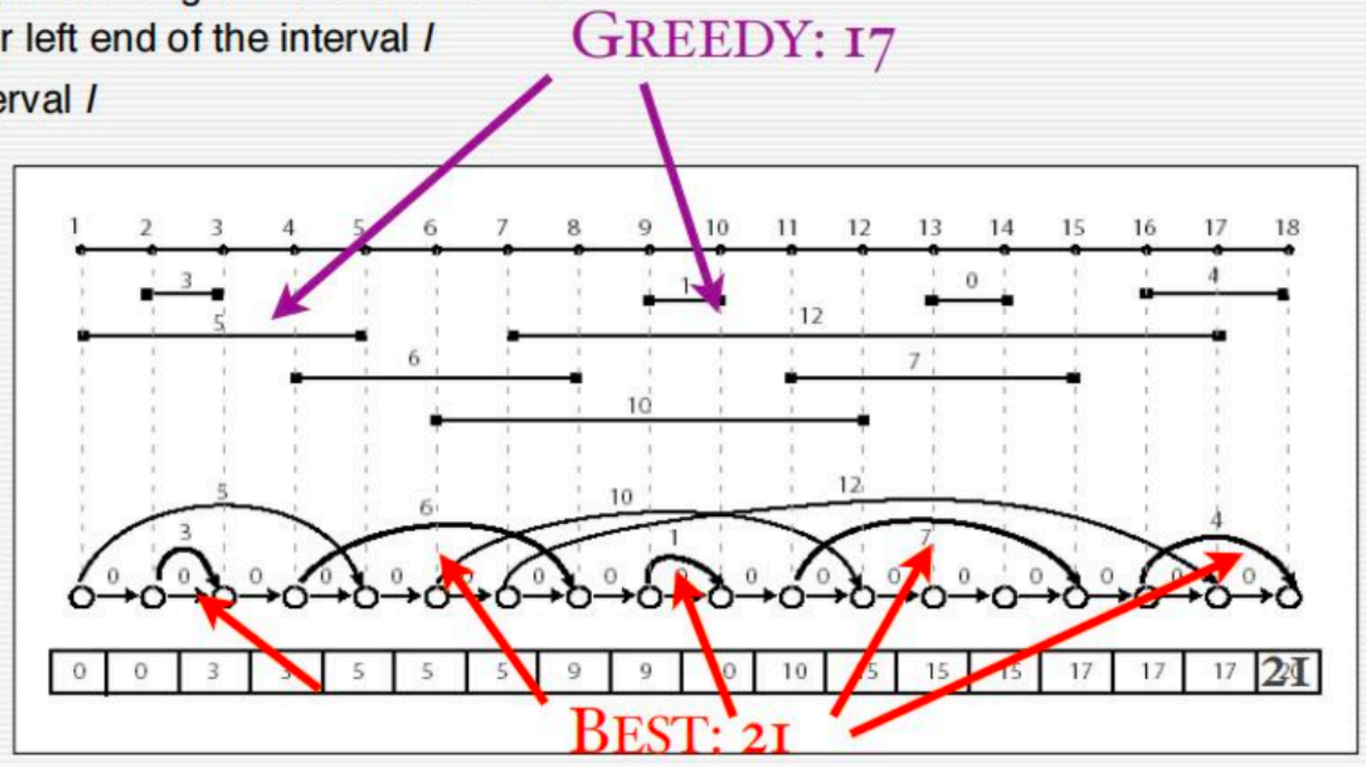
- Input: a set of weighted intervals (putative exons)

- Output: A maximum chain of intervals from this set

- Would a greedy algorithm solve this problem?

Solution:

ExonChaining (*G, n*) //Graph, number of intervals
for $i \leftarrow$ to $2n$
  $s_i \leftarrow 0$
for $i \leftarrow 1$ to $2n$
  if vertex v_i in G corresponds to right end of the interval I
    $j \leftarrow$ index of vertex for left end of the interval I
    $w \leftarrow$ weight of the interval I
    $s_i \leftarrow max\{s_j + w, s_{i-1}\}$
  else
    $s_i \leftarrow s_{i-1}$
return s_{2n}

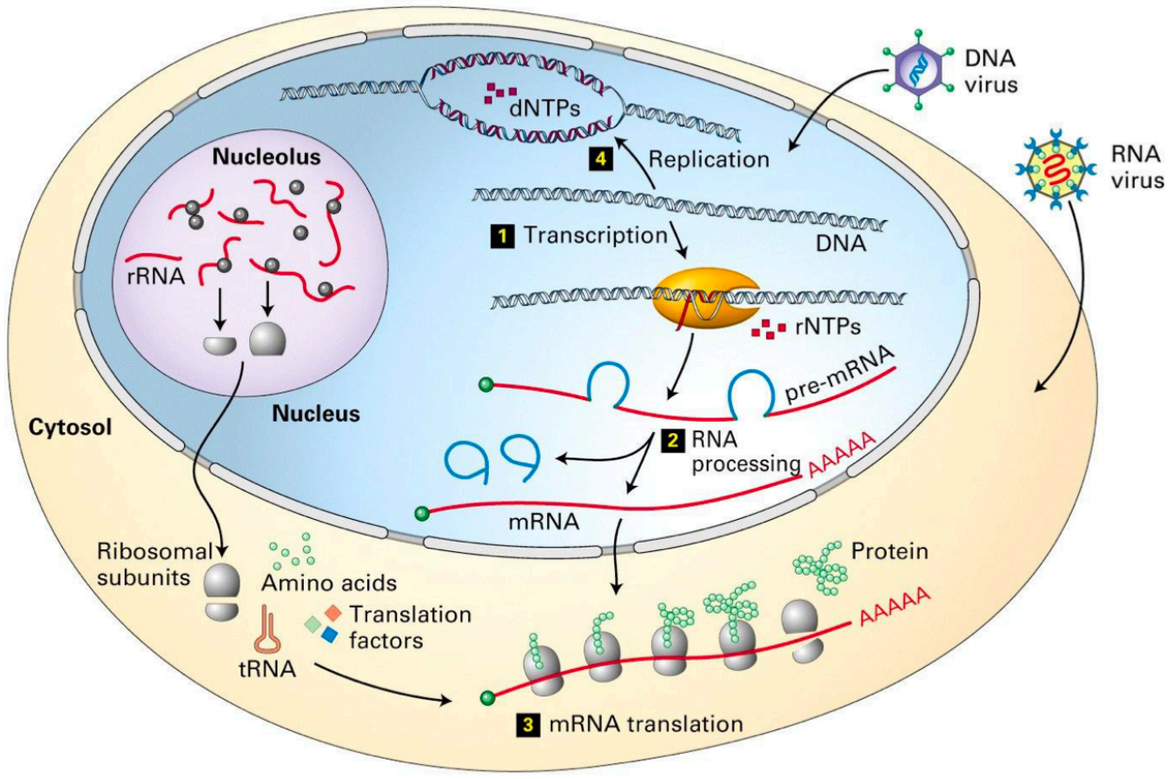


- Use a graph representation of the exon chaining problem
- Can be solved in O(n) time using dynamic programming

Outline

- Gene
- Gene prediction
- **RNA**
- Noncoding RNA world
- RNA secondary structure

The Central Dogma of Molecular Biology



DNA ⇌ RNA → protein

mRNA
3-5% of total RNAs

→ Remanents for degradation

↓ Non-coding RNAs

Non-coding RNA (ncRNA)

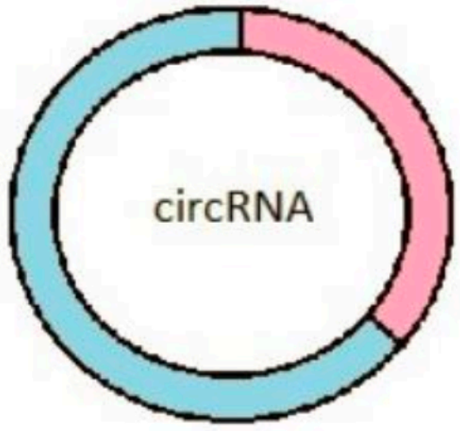
- What are non-coding RNAs?

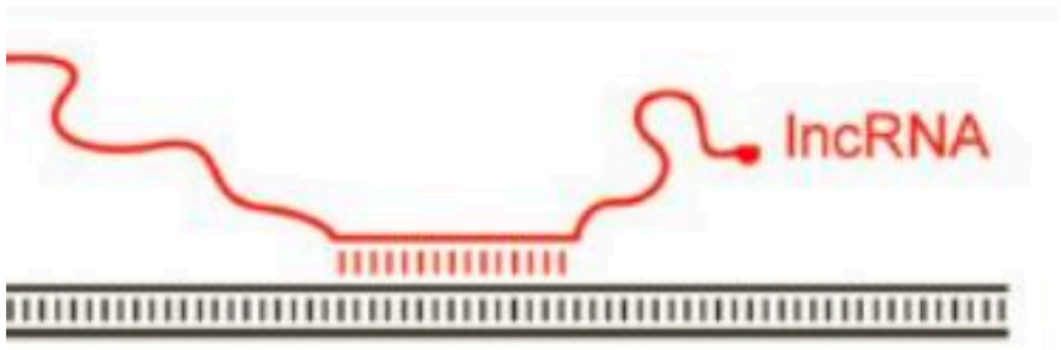
  - RNA molecules that function without being translated into a protein

Are non-coding RNAs always non-translatable?

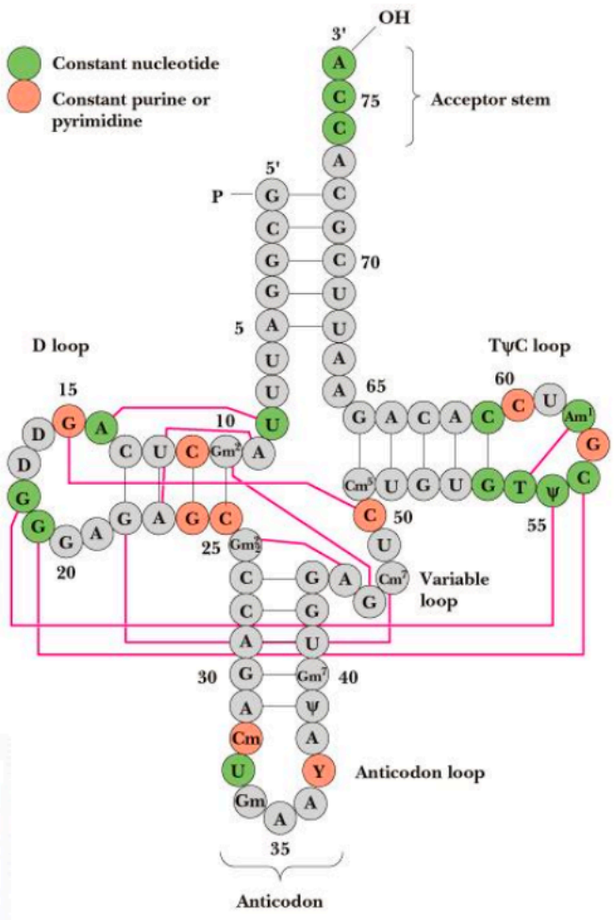
Commonly used rules for non-coding RNA prediction

- Without long open reading frame (ORF)
- Without codon-related nucleotide composition
- May have specific secondary structures

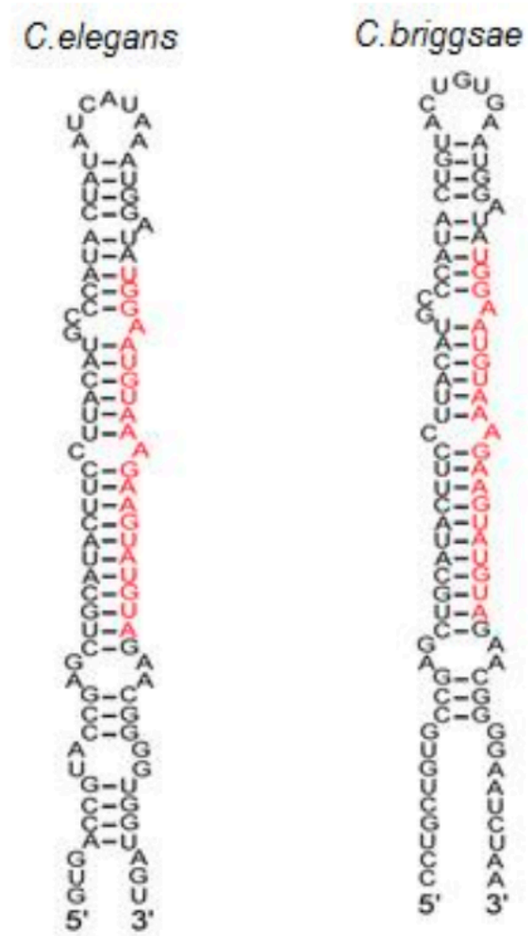




Structures are the key features for non-coding RNA identification



tRNA



miRNA

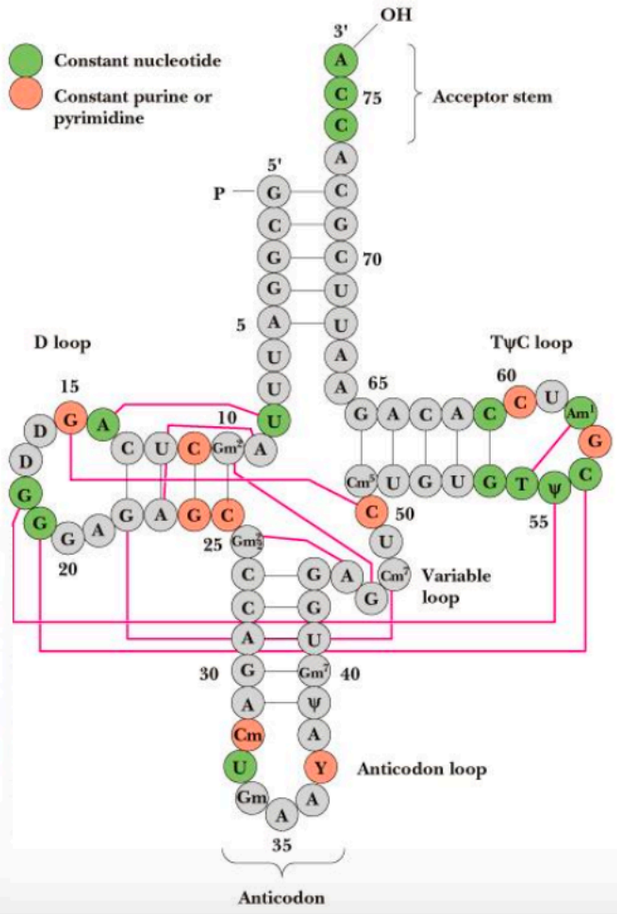
The Non-coding RNA World

- **transfer RNAs (tRNAs)**
- ribosomal RNAs (rRNAs)
- small nucleolar RNAs (snoRNAs)
- small nuclear RNAs (snRNAs)
- small interfering RNAs (siRNAs)
- microRNAs (miRNAs)
- PIWI-interacting RNA (piRNA)
- long non-coding RNA (lncRNA)
- antisense RNAs
- pseudogenes
- circRNAs

Transfer RNAs (tRNAs)

- The first identified RNA class
- Proposed as intermediaries between DNA and amino acids during protein synthesis before it was identified
- ~80 nt in length
- Clover leaf-like secondary structure

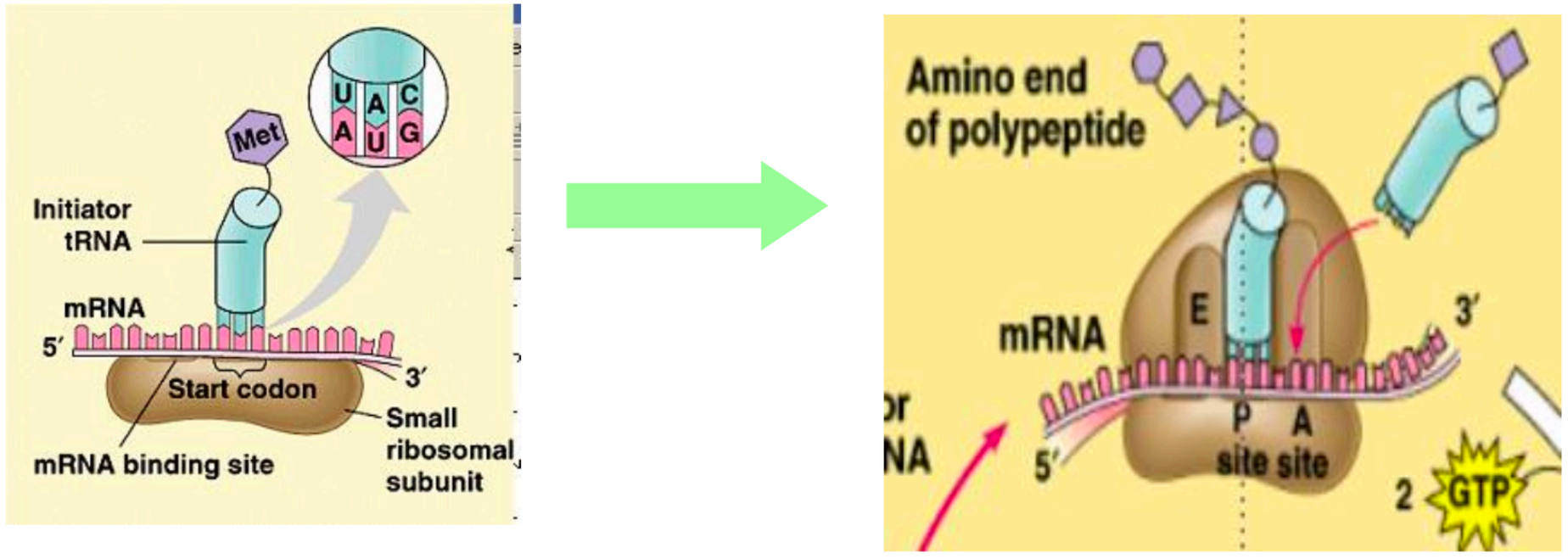
  - four short double-helical elements
  - three loops
    (D, anti-codon, and T loops)




a clover leaf

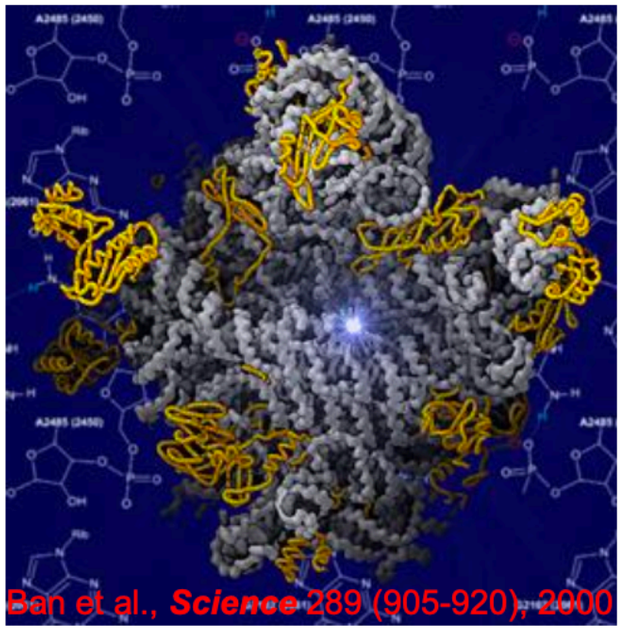
Functions of Transfer RNAs (tRNAs)

- tRNAs function as adaptors between mRNA and amino acids



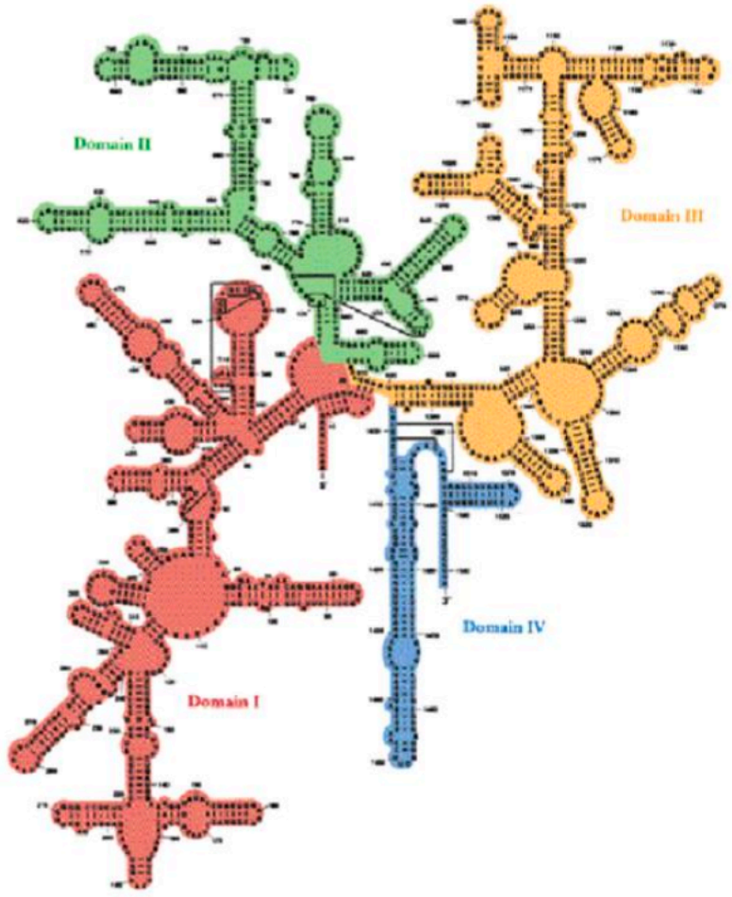
Ribosomal RNAs (rRNAs)

- Basic structure component of ribosomes
- Catalyze protein synthesis



Ban et al., *Science* 289 (905-920), 2000

Tertiary Structure Of 16S rRNA

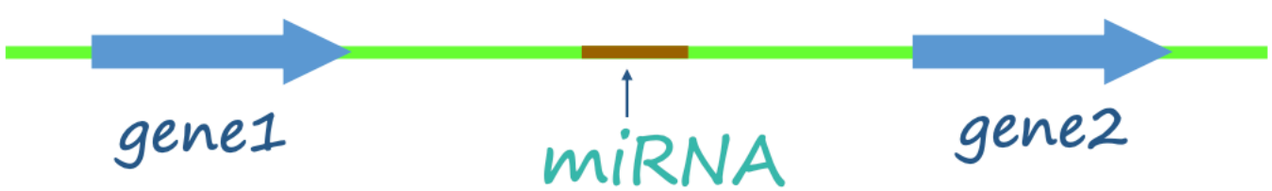


Secondary structure of 16S rRNA

microRNA (miRNA)

- ~22 nt long non-coding RNAs
- First identified in *C.elegans*

  lin-4 (1993) let-7 (2000)
- Present in most eukaryotes
- usually arose from non-protein coding genomic regions or introns



miRNAs Have Hairpin-like Precursors

miR-1

C.elegans

C.briggsae

Drosophila

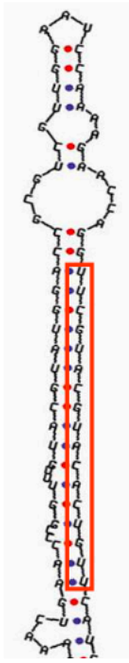
H. sapiens

Lau et al 2001

Prediction of microRNAs

Methods:

hairpin-like secondary structure
sequence conservation
nucleotide composition



Predict miRNAs

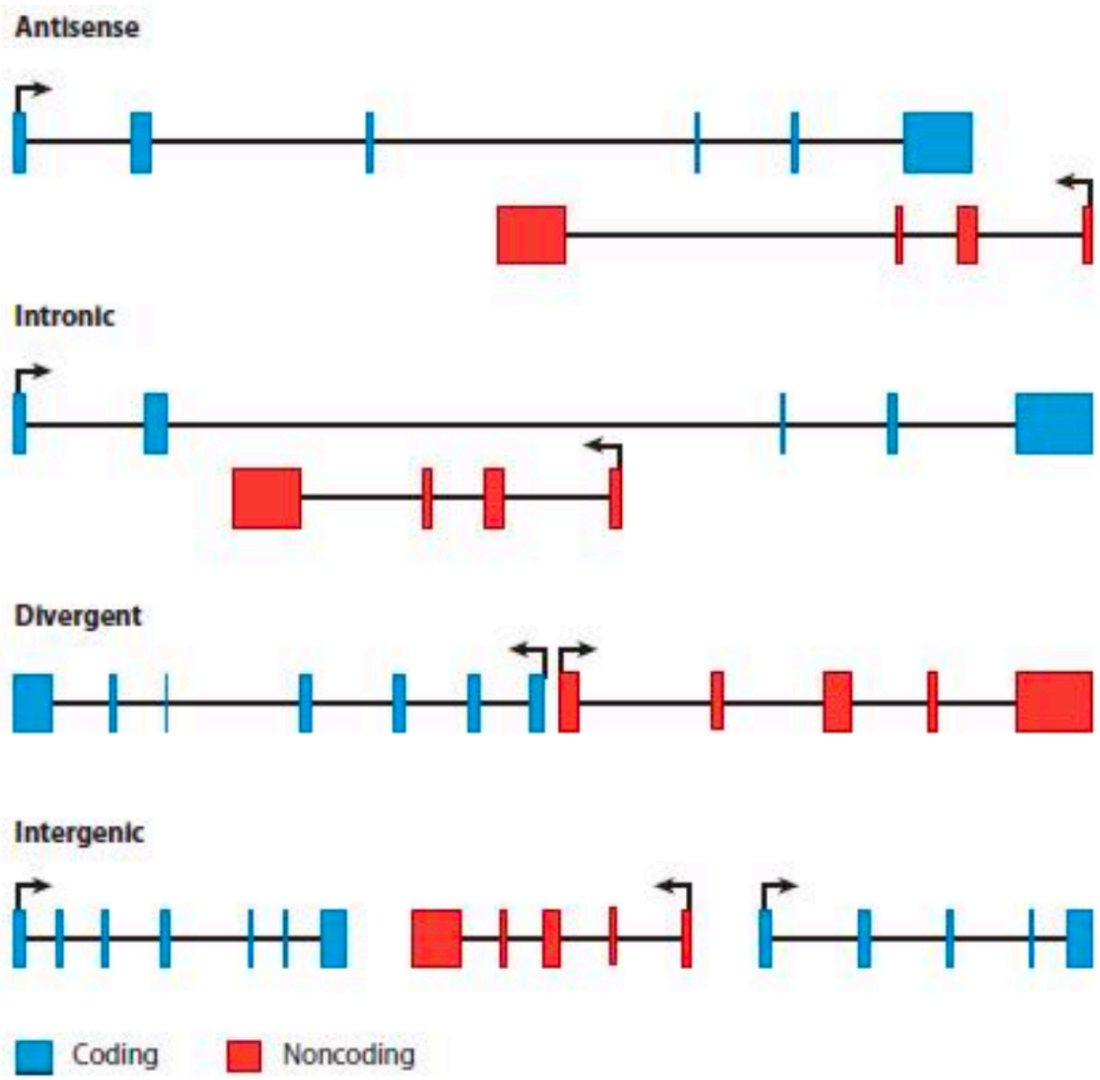
The Plant Cell, Vol. 20: 3186–3190, December 2008, www.plantcell.org © 2008 American Society of Plant Biologists

COMMENTARY

Criteria for Annotation of Plant MicroRNAs

Blake C. Meyers,[a,1] Michael J. Axtell,[b,1] Bonnie Bartel,[c] David P. Bartel,[d] David Baulcombe,[e] John L. Bowman,[f] Xiaofeng Cao,[g] James C. Carrington,[h] Xuemei Chen,[i] Pamela J. Green,[a] Sam Griffiths-Jones,[j] Steven E. Jacobsen,[k] Allison C. Mallory,[l] Robert A. Martienssen,[m] R. Scott Poethig,[n] Yijun Qi,[o] Herve Vaucheret,[l] Olivier Voinnet,[p] Yuichiro Watanabe,[q] Detlef Weigel,[r] and Jian-Kang Zhu[i]

Long non-coding RNA (lncRNA)



lncRNA

Annu. Rev. Biochem. (2012)
81:145－66

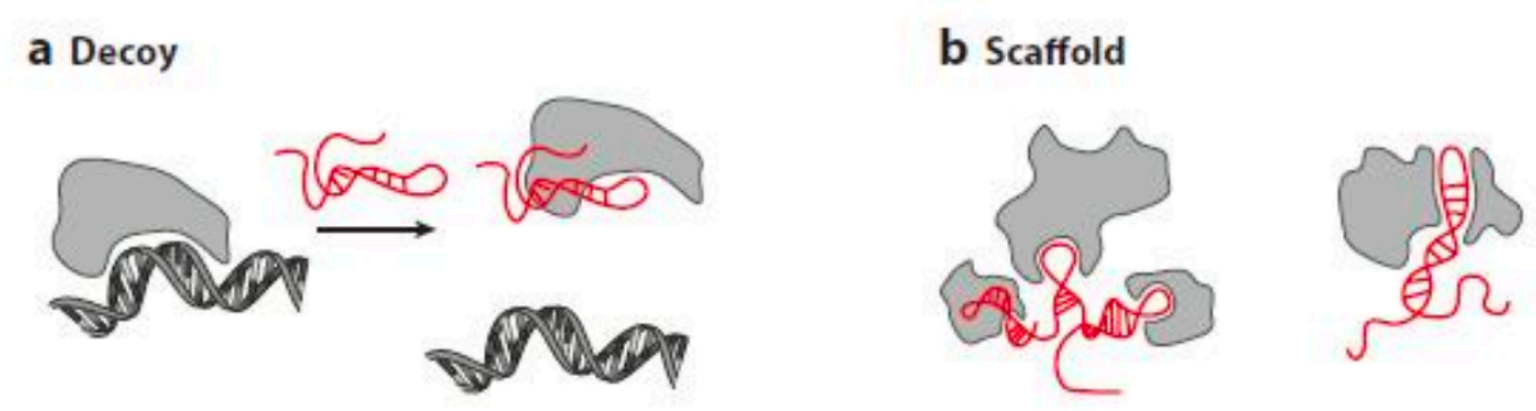
How to define a lncRNA?

- Size: At least > 200 nt, usually > 500 nt

- Without the capacity to encode normal proteins
  (Without long ORFs (usually > 33 aa))

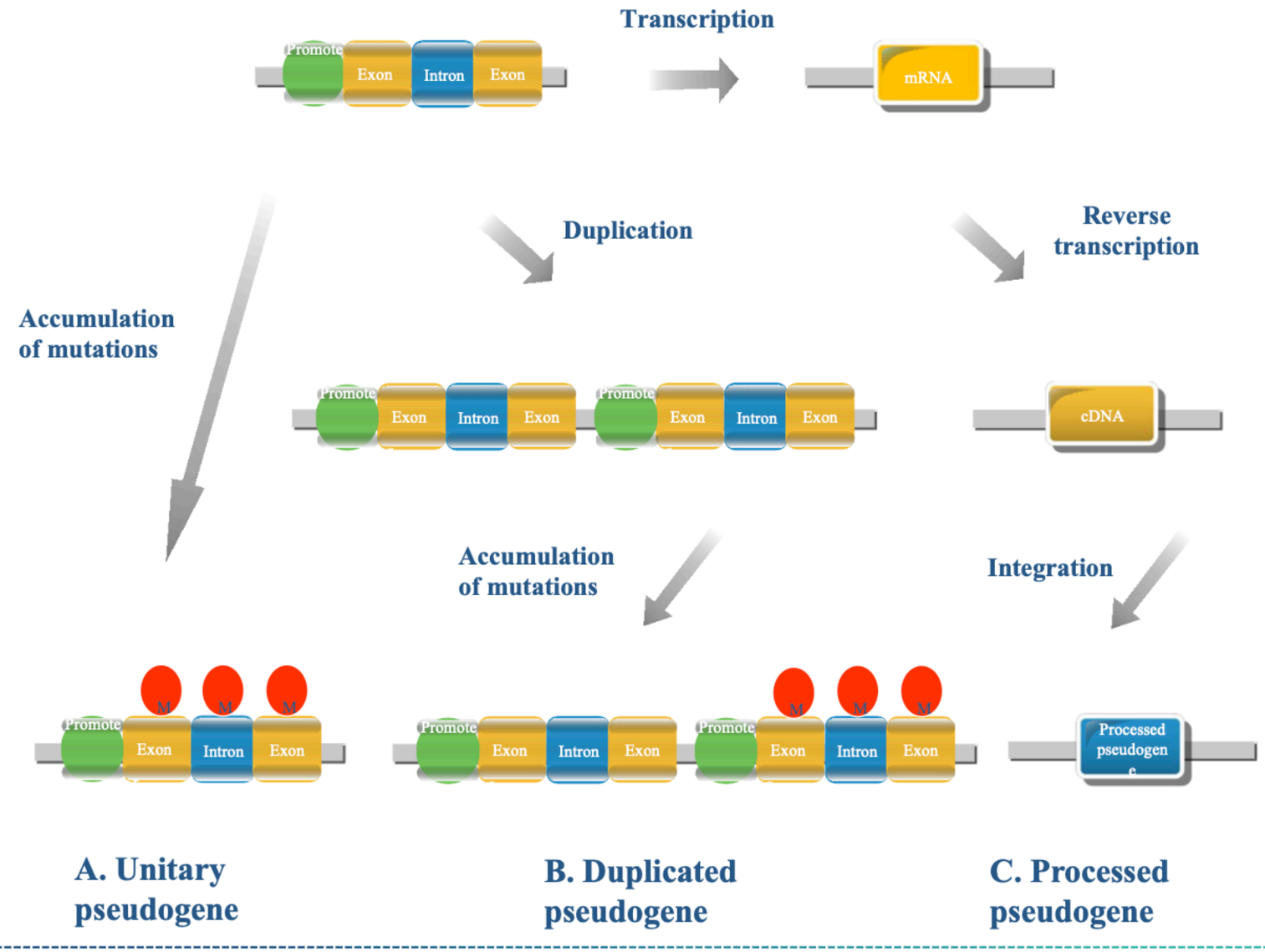
- Could encode small peptide

Known functions of lncRNA



Annu. Rev. Biochem. 2012. 81:145–66

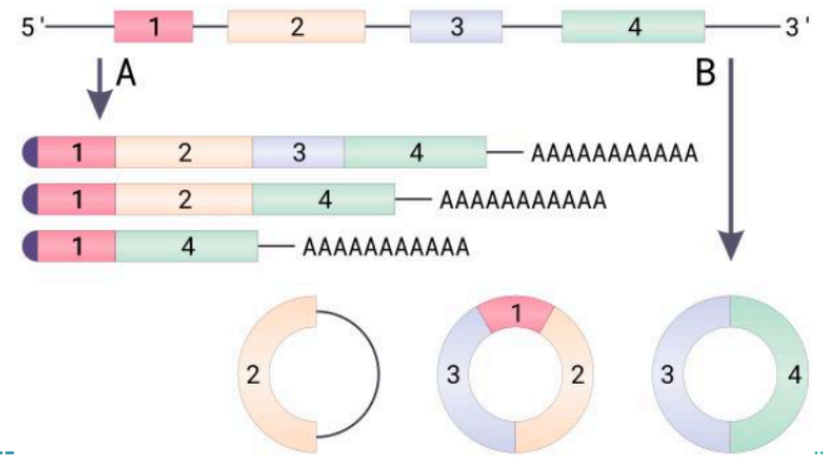
Three Types of Pseudogenes



Li, Yang and Wang, JGG (2013)

Circular RNA (CircRNA)

- A type of RNA which forms a covalently closed continuous loop.
- Many circular RNAs arise from otherwise protein-coding genes.
- Come could encode proteins.
- Do not have 5' or 3' ends.
- Resistant to exonuclease-mediated degradation.



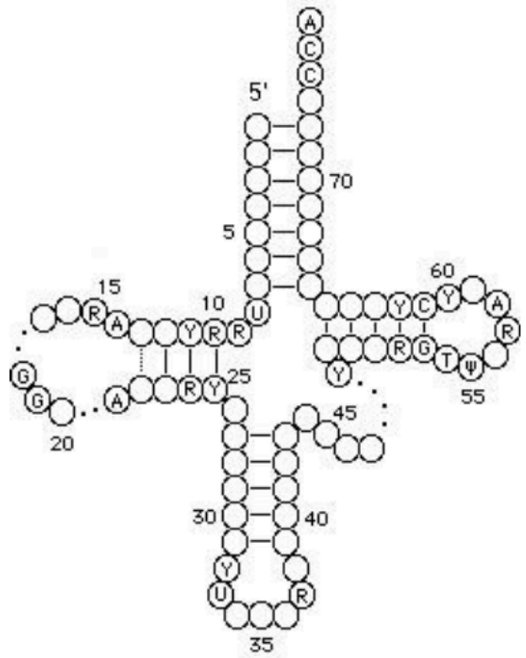
A little More about RNA secondary structures

Hierarchical organization of RNA molecules

Primary structure:

5’ ACCACCUGCUGA 3’ —**Covalent bonds (共价键)**

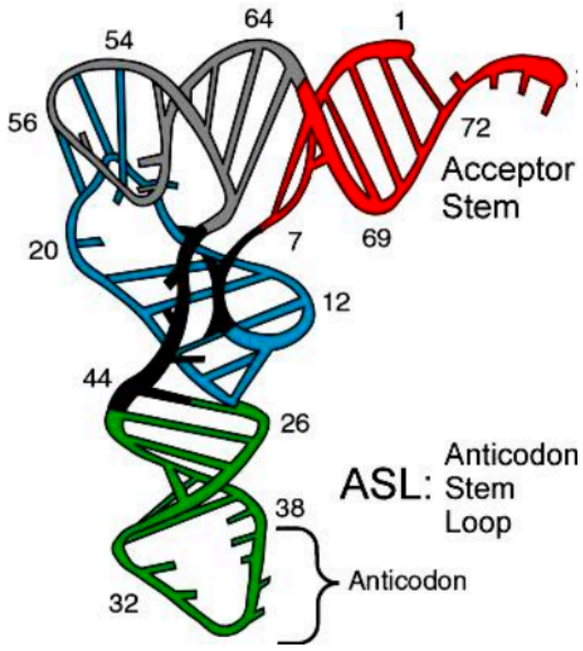
Secondary structure:

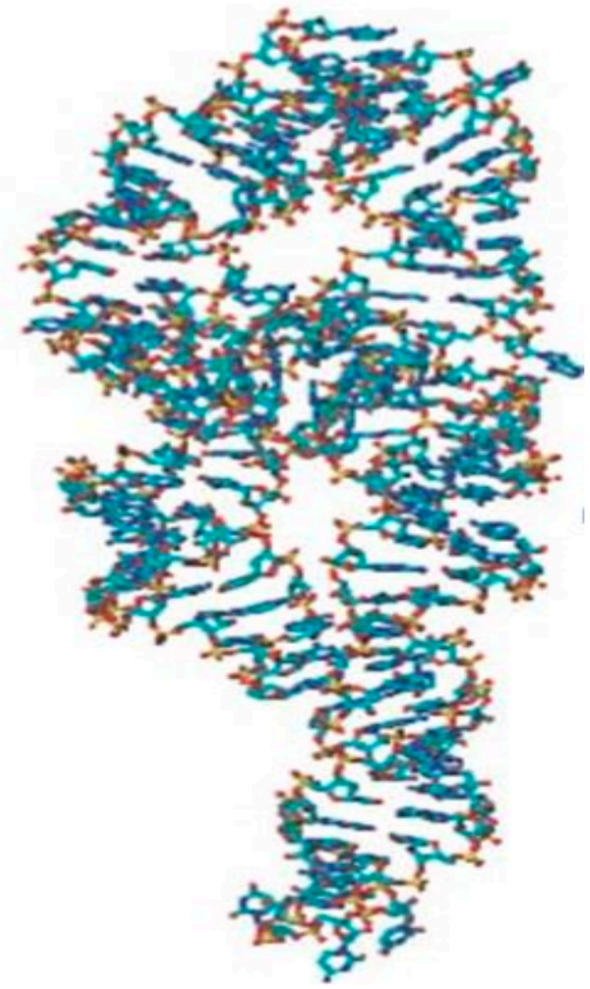


Hierarchical organization of RNA molecules

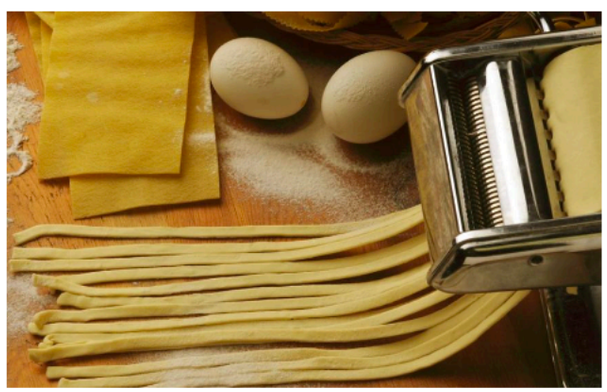
Tertiary structure* (三级结构)*:

List of interactions between secondary structures





Hierarchical organization of RNA molecules

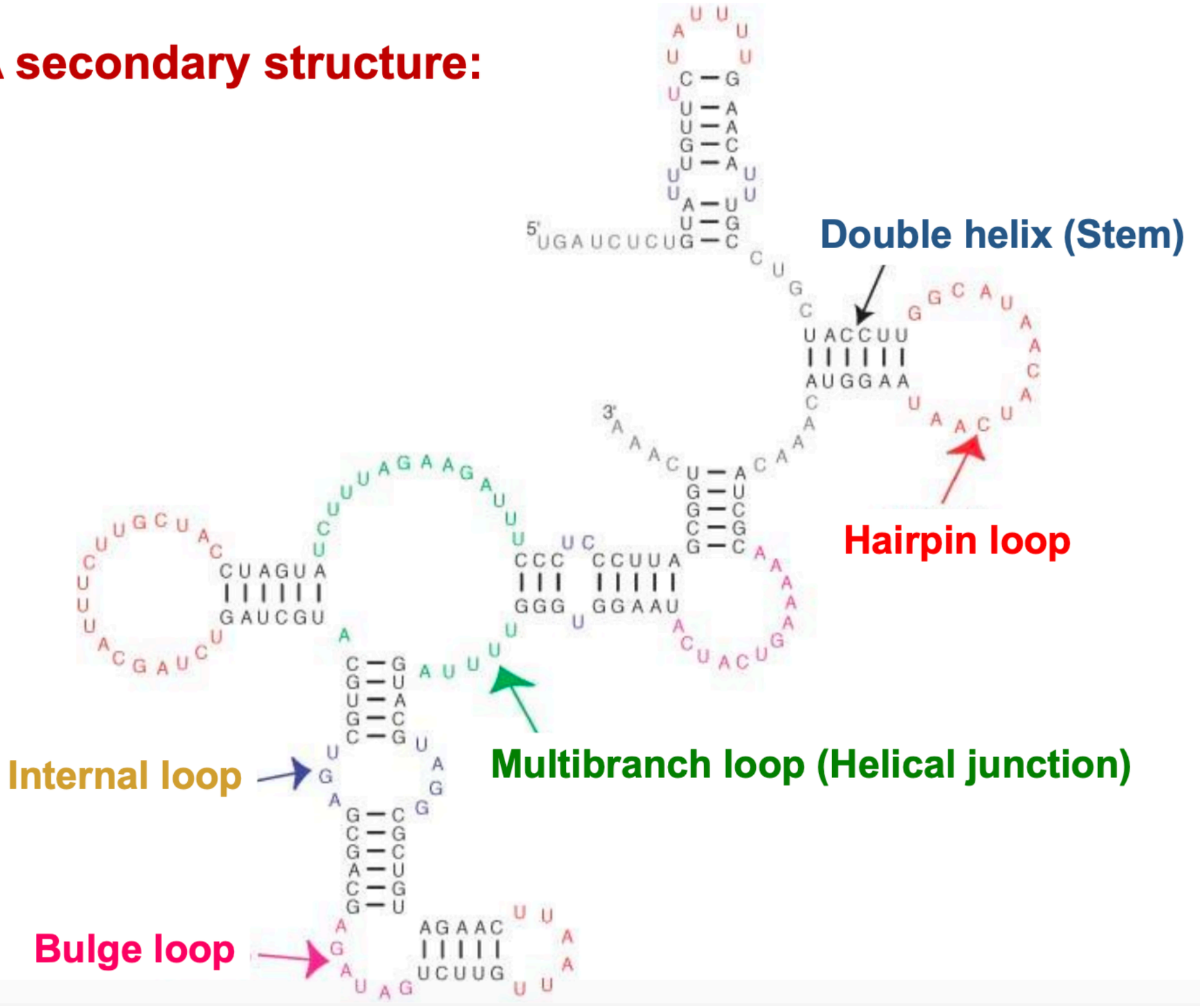


Primary structure
（一级结构）

Secondary structure
（二级结构）

Tertiary structure
（三级结构）

RNA secondary structure:



RNA Secondary Structure Prediction

Two primary methods for RNA secondary structure prediction:

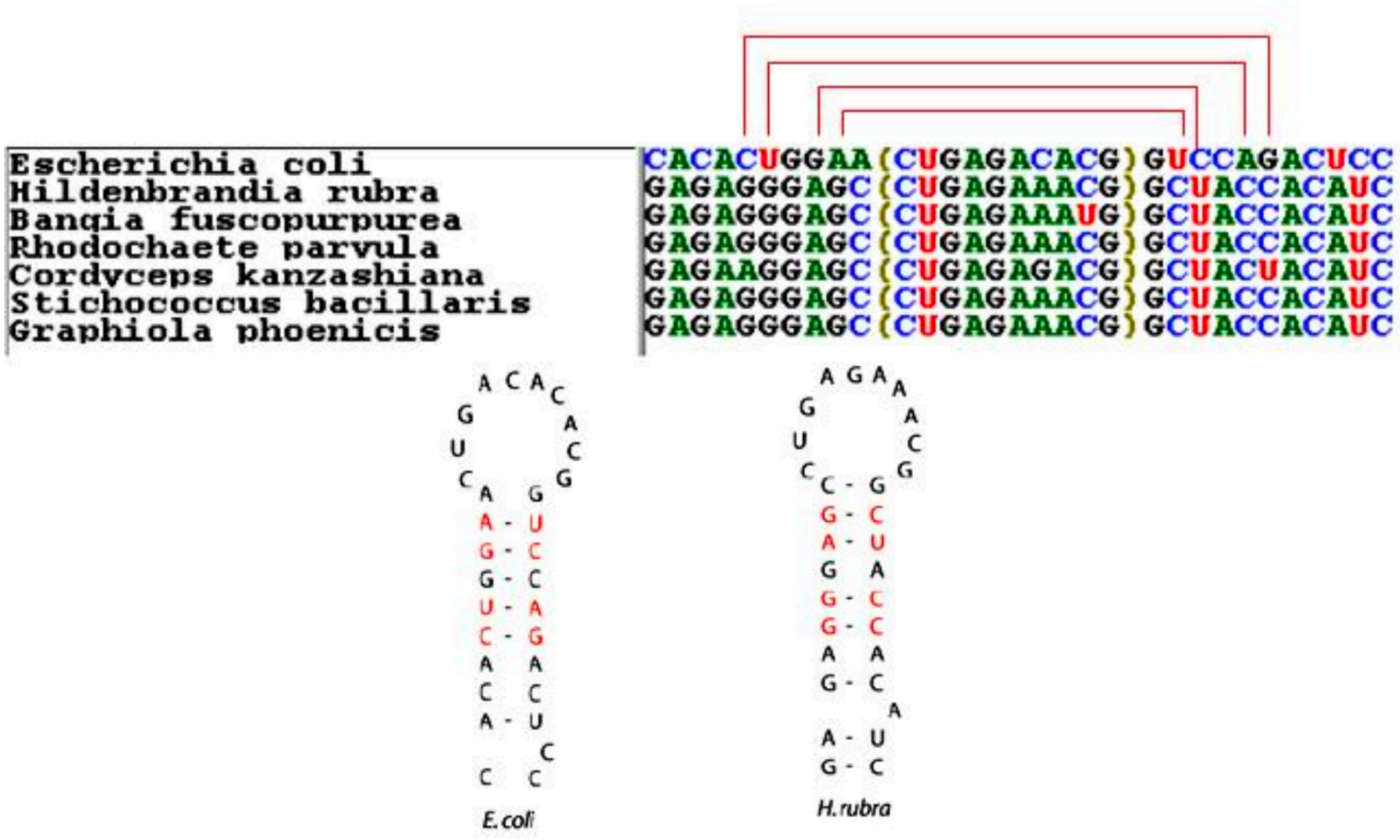
- **Co-variation analysis** (comparative sequence analysis)

  Takes into account of conserved patterns of basepairs during evolution (2 or more sequences)

- **Single sequence prediction**

  Predicts secondary structures with minimum free-energy

Co-variation Analysis



THANK YOU