Protein Synthesis and Gene Finding

Day 2

DNA Analysis

- 23andMe
- Personalized medicine
- Genetic screens for disease
- Etc.

Bio Crash Course: DNA

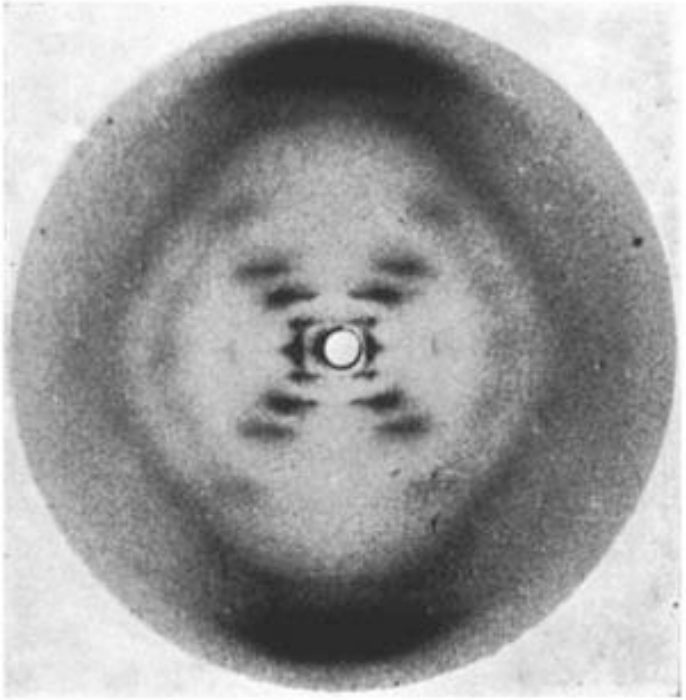
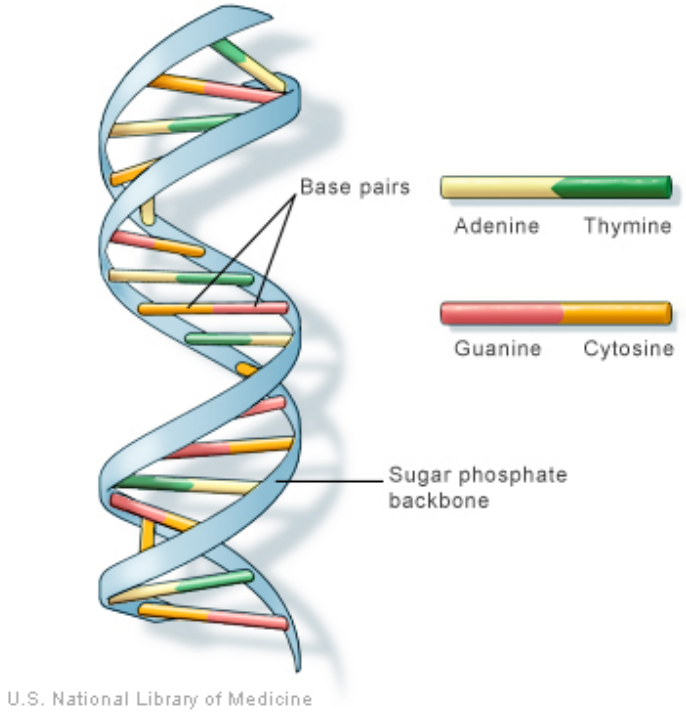


Photo 51: X-ray diffraction created by Raymond Gosling under the supervision of Rosalind Franklin May, 1952

Double Helix Structure: Watson and Crick

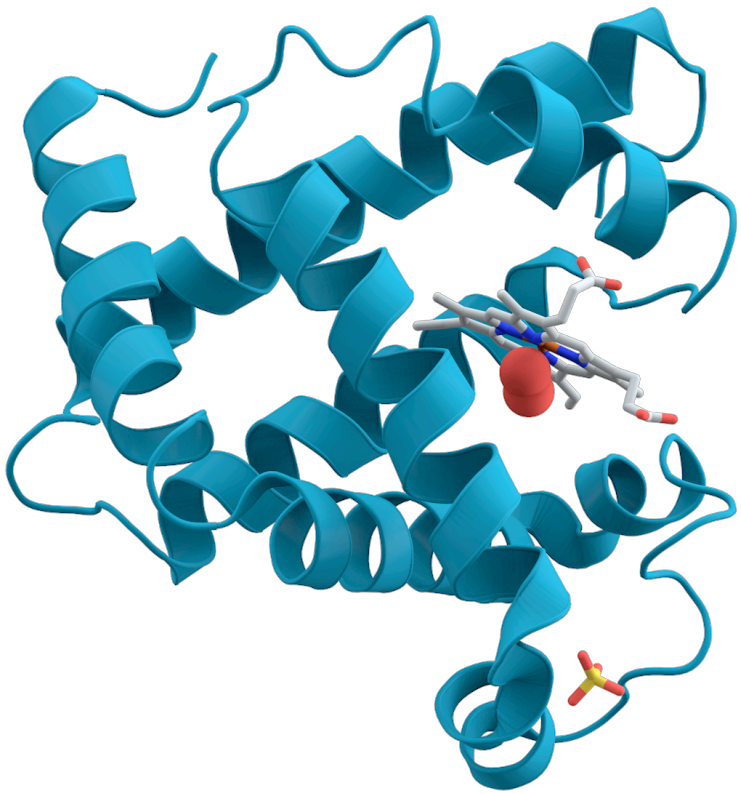
DNA and the Central Dogma



The blueprints for making all the proteins necessary for life are contained within DNA.

Or, more succinctly: DNA Makes RNA, RNA makes Protein

Proteins

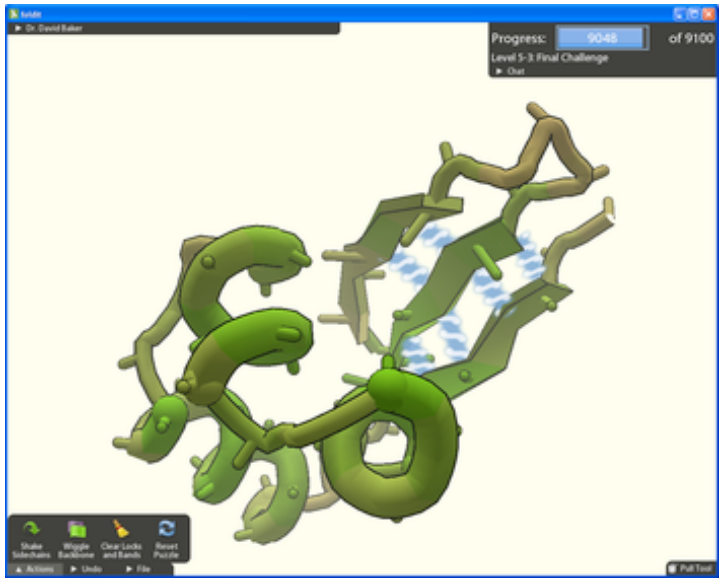


Proteins are chains of Amino Acids

Aside: Protein Secondary Structure Prediction

Goal: find the lowest energy configuration of an amino acid chain

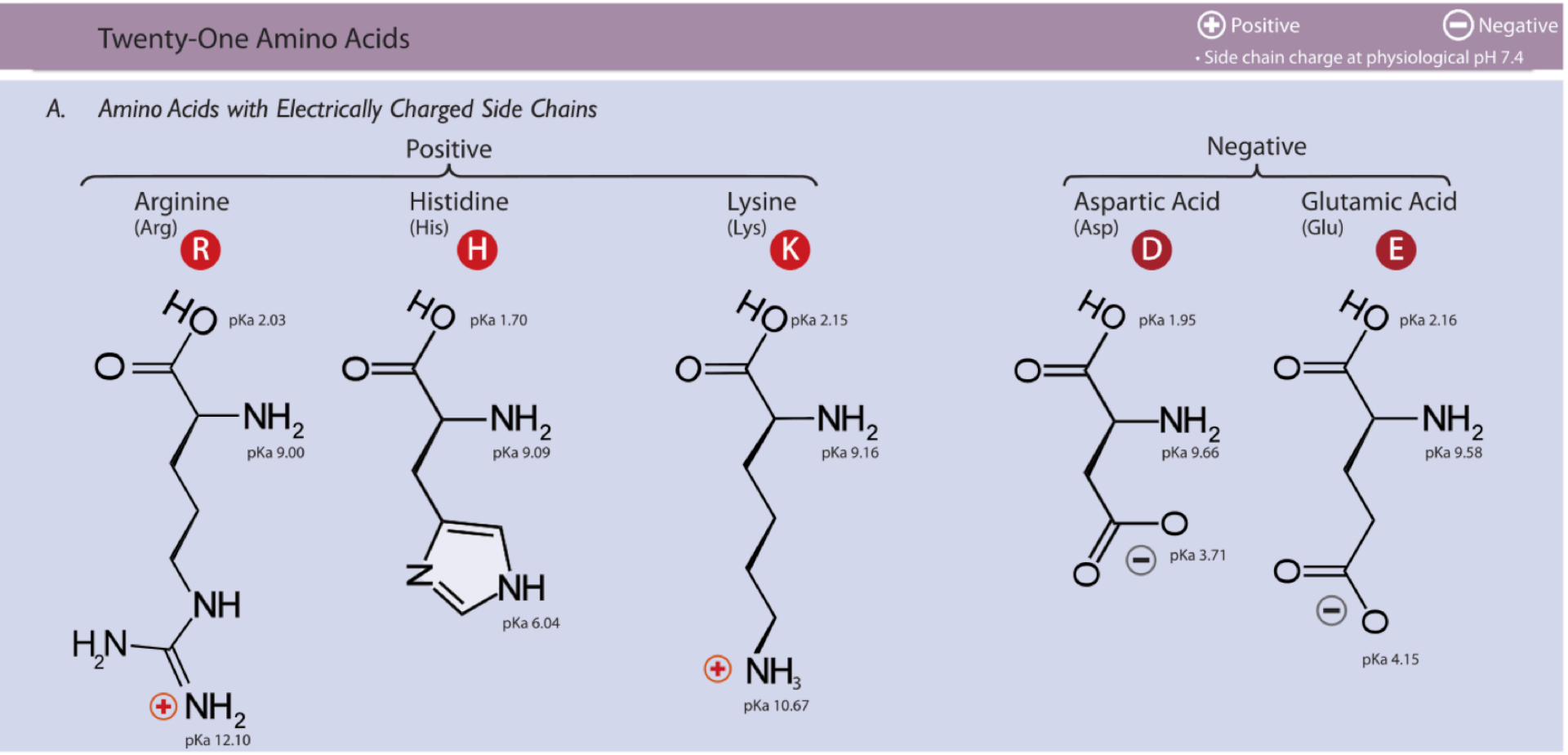
Unfortunately this problem is hard!
How hard is it?
This hard: http://www.cs.berkeley.edu/~christos/hp.ps



https://www.youtube.com/watch?v=lGYJyur4FUA

Fold-it: a gameified approach to computing this structure

Proteinogenic Amino Acids



http://en.wikipedia.org/wiki/Amino_acid

DNA and Protein Synthesis

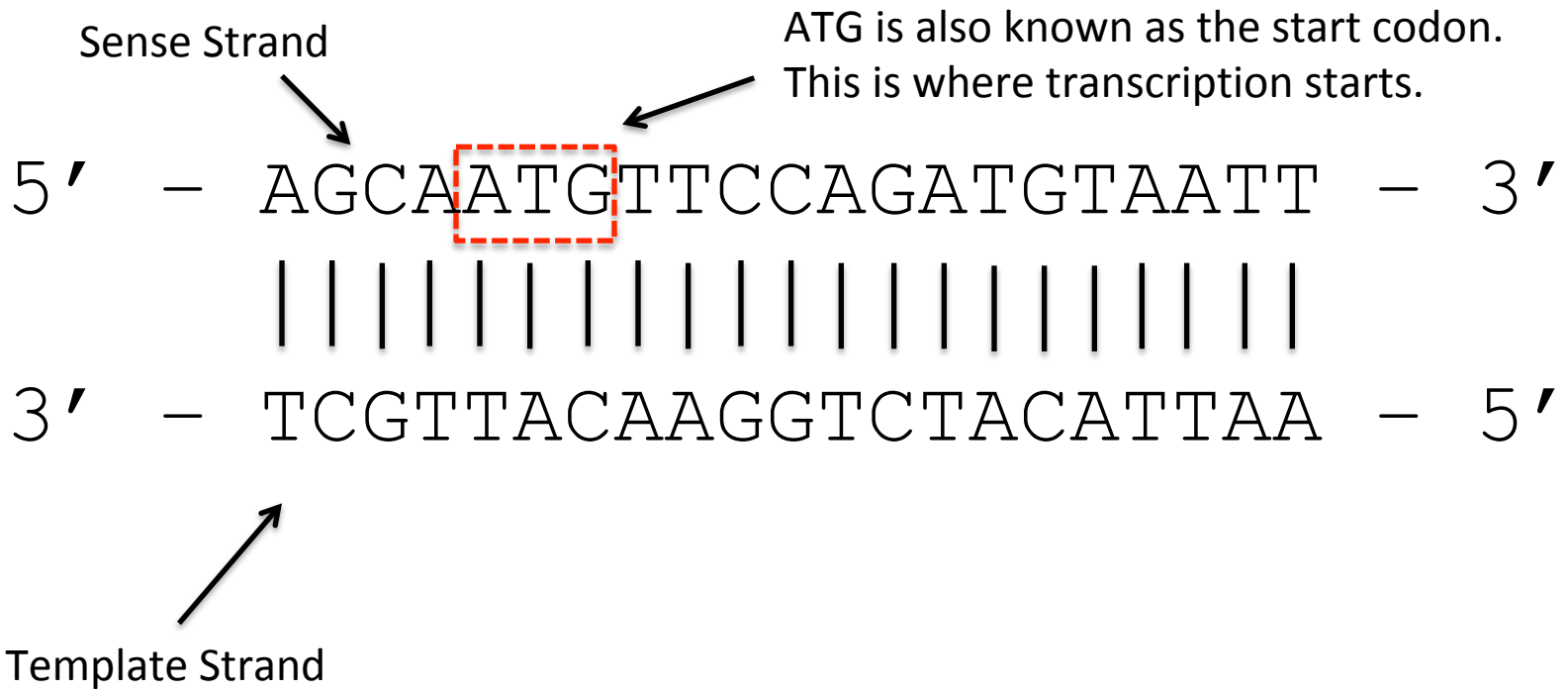
- **Stage 1:** an enzyme “decides” that it is time to synthesize a protein (epigenetics)
- **Stage 2:** DNA is unzipped and paired with a complementary strand of RNA (transcription)
- **Stage 3:** tRNA attaches amino acids to complementary segments of RNA (translation)

Protein Synthesis Detailed Example

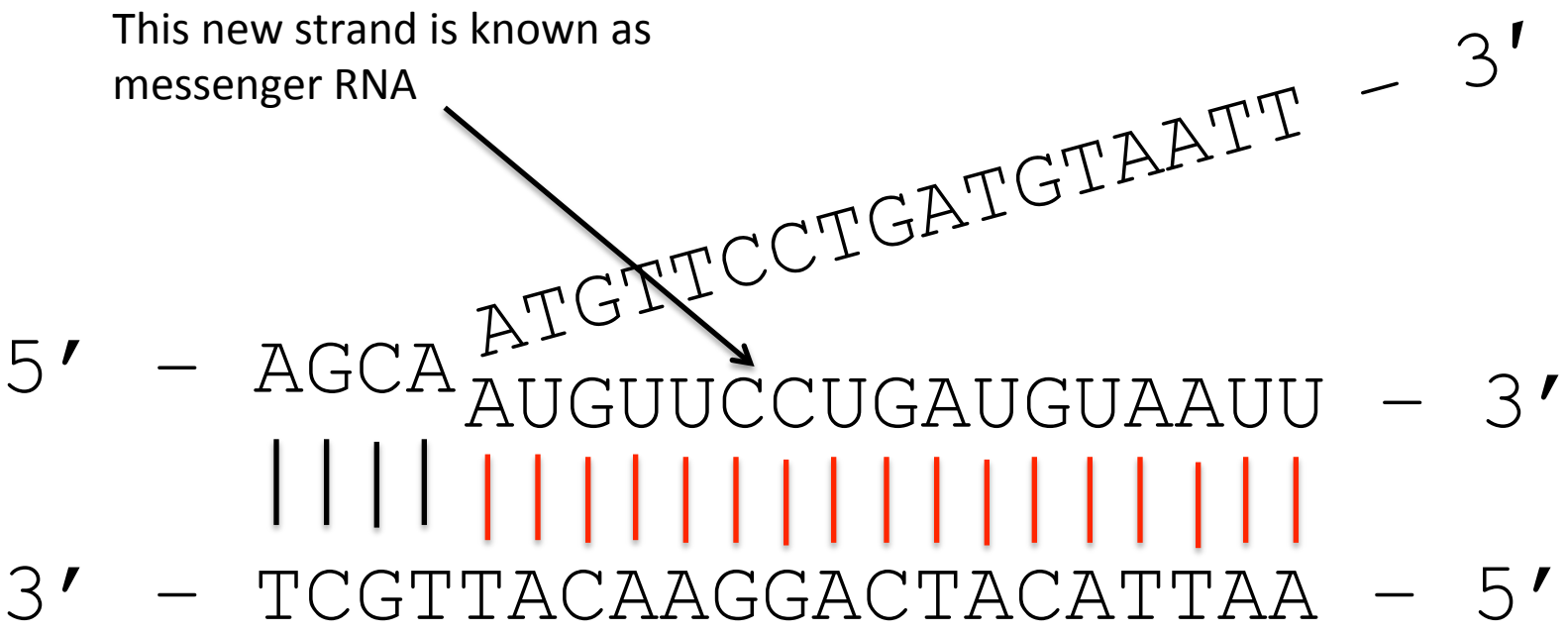
```
5′ – AGCAATGTTCCAGATGTAATT – 3′
     ||||||||||||||||||||
3′ – TCGTTACAAGGTCTACATTAA – 5′
```

Aside: The 5’ and 3’ are useful when talking about the directionality of certain processes related to protein synthesis. More later.

Protein Synthesis
Step 1: Transcription Initiation



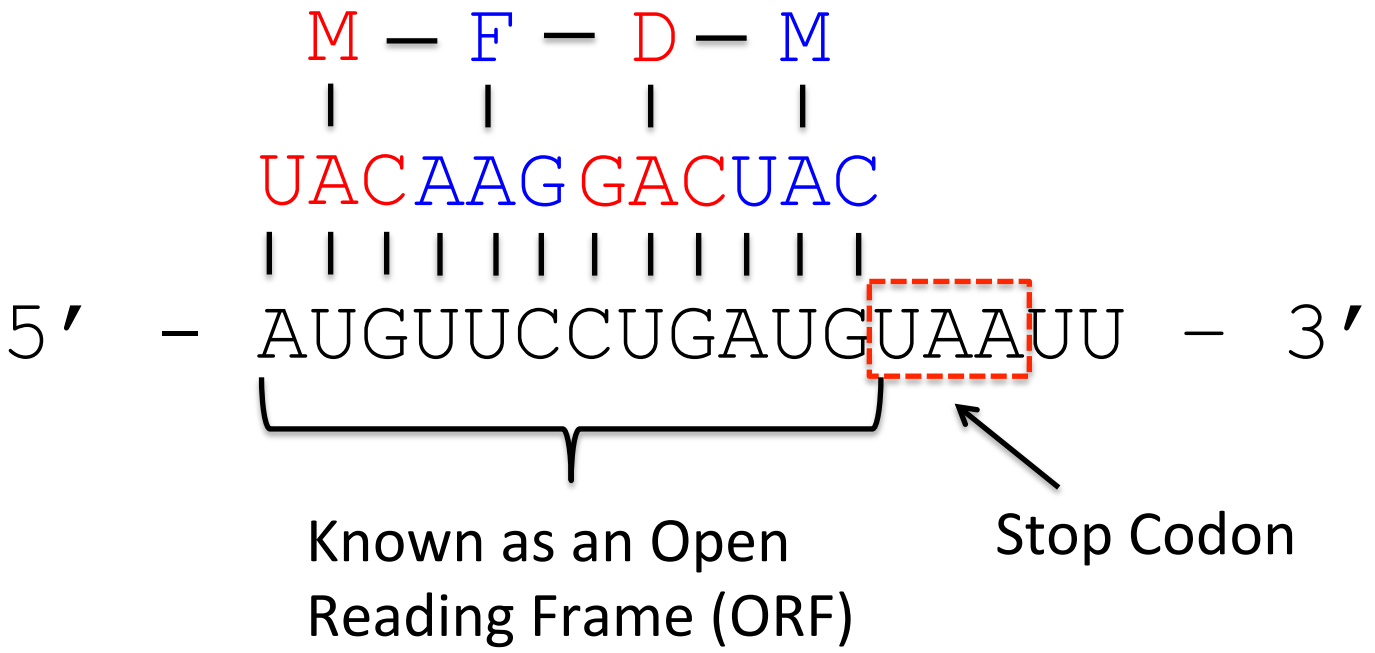
Protein Synthesis
Step 2: Messenger RNA Attaches



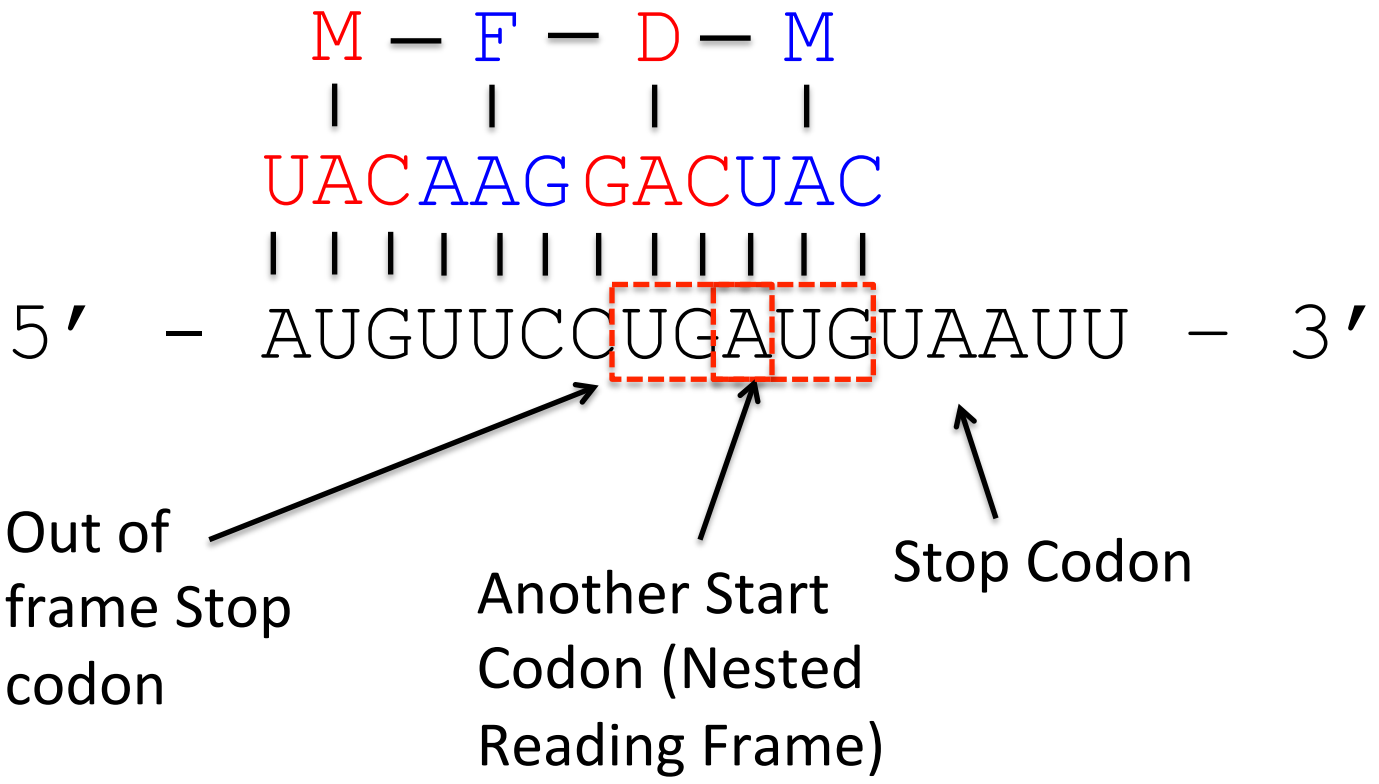
Protein Synthesis
Step 3: Messenger RNA Breaks Away

5′ – AUGUUCCUGAUGUAAUU – 3′

Protein Synthesis
Step 4: Amino Acid Chain Forms



Protein Synthesis
Some things to Notice



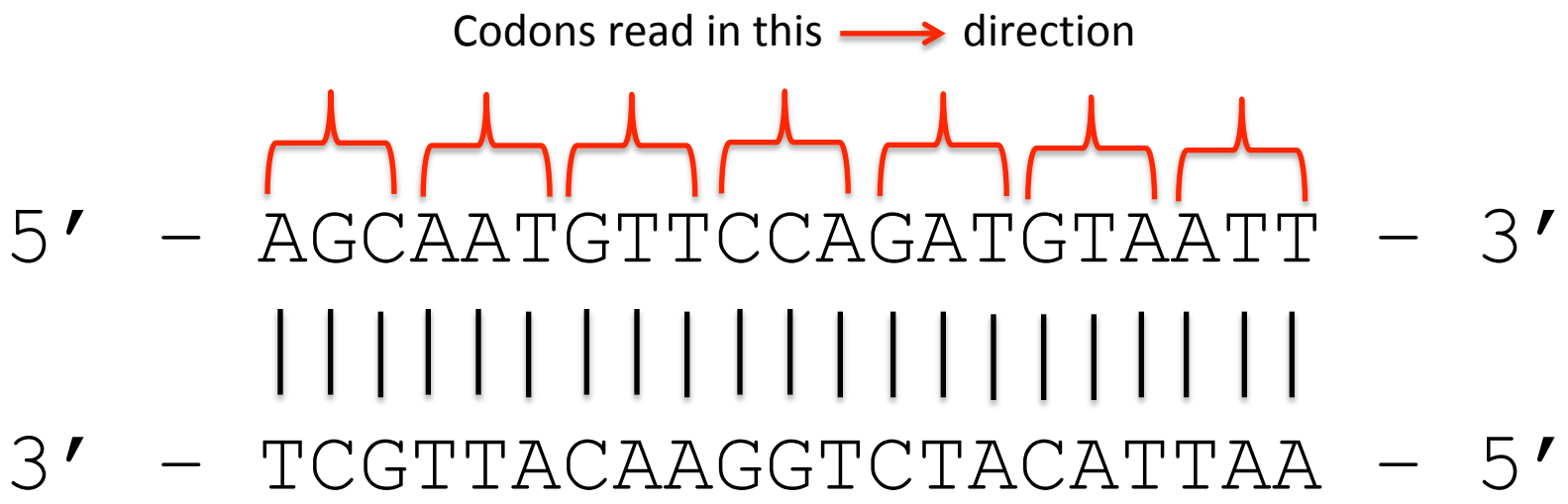
DNA Codon Table

Standard genetic code

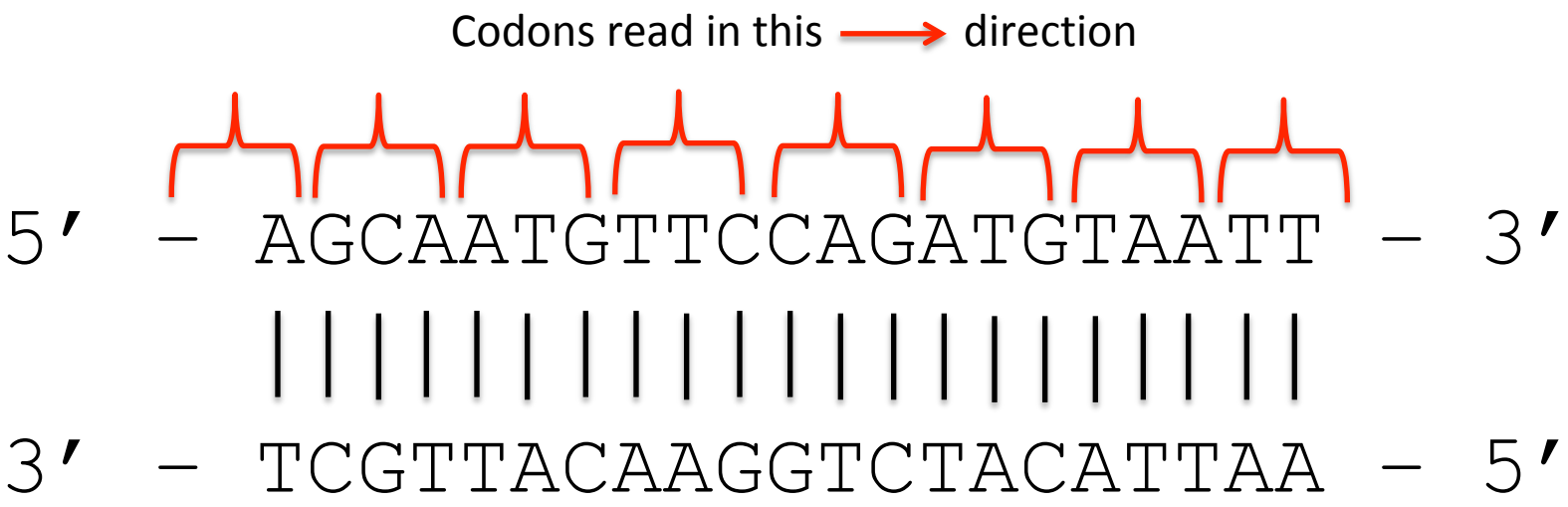
1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)	TGA	Stop (*Opal*)	A
	TTG		TCG		TAG	Stop (*Amber*)	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

This code is read from 5’ to 3’ on the DNA strand

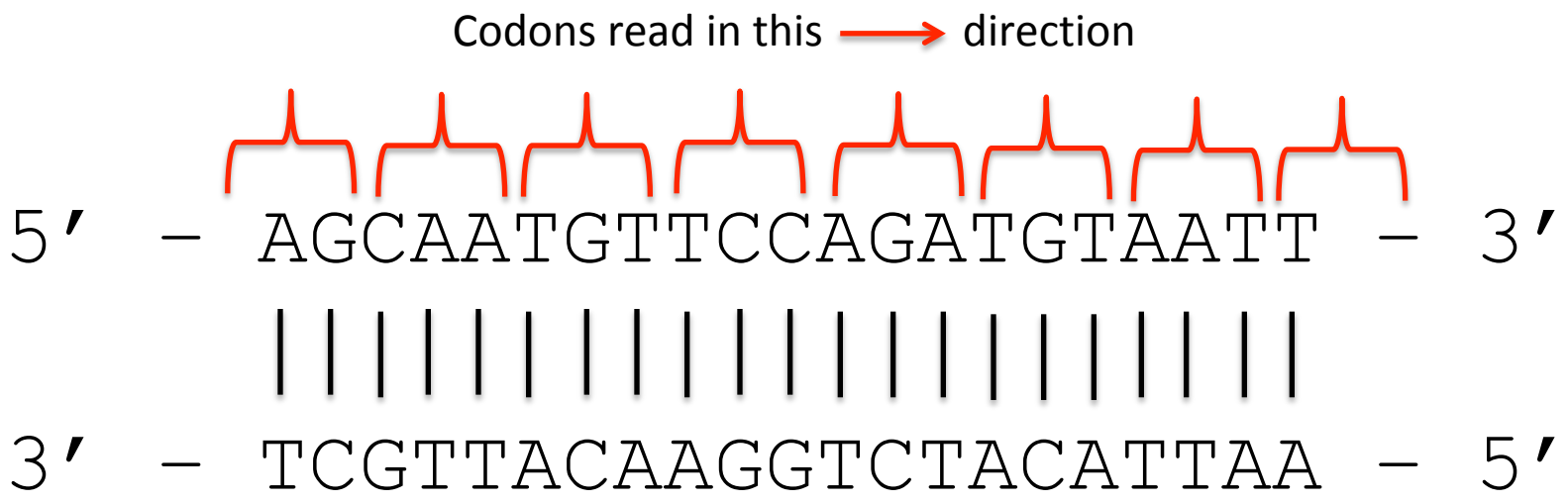
6 Reading Frames: Reading Frame 1



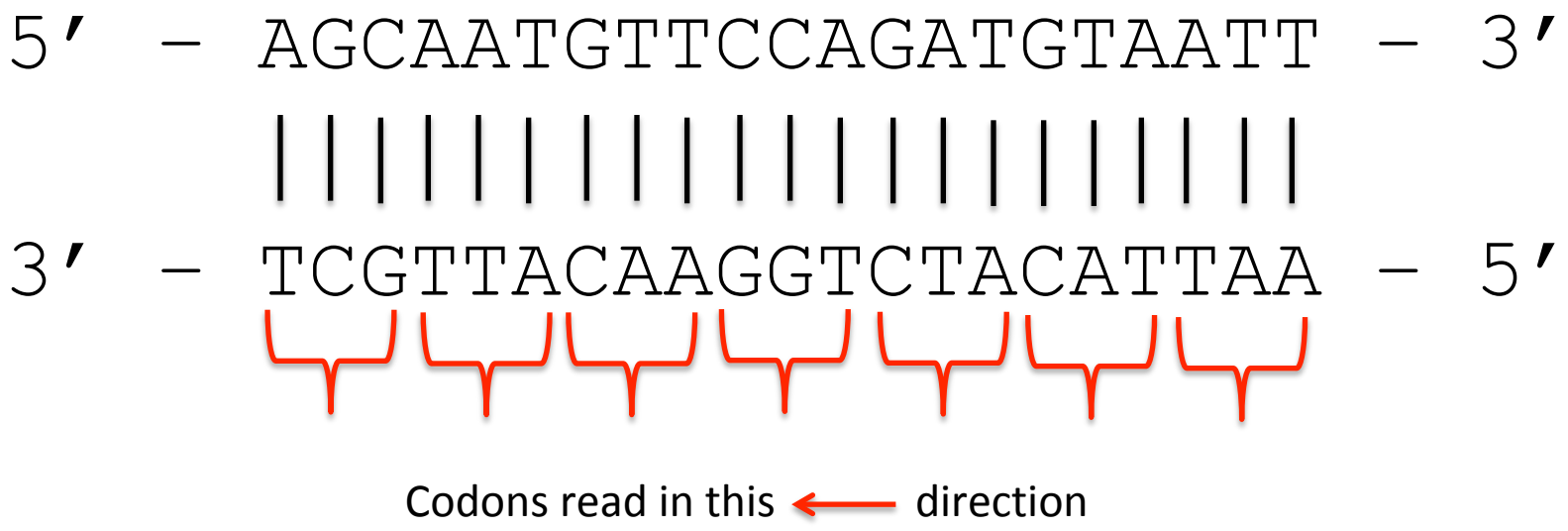
6 Reading Frames: Reading Frame 2



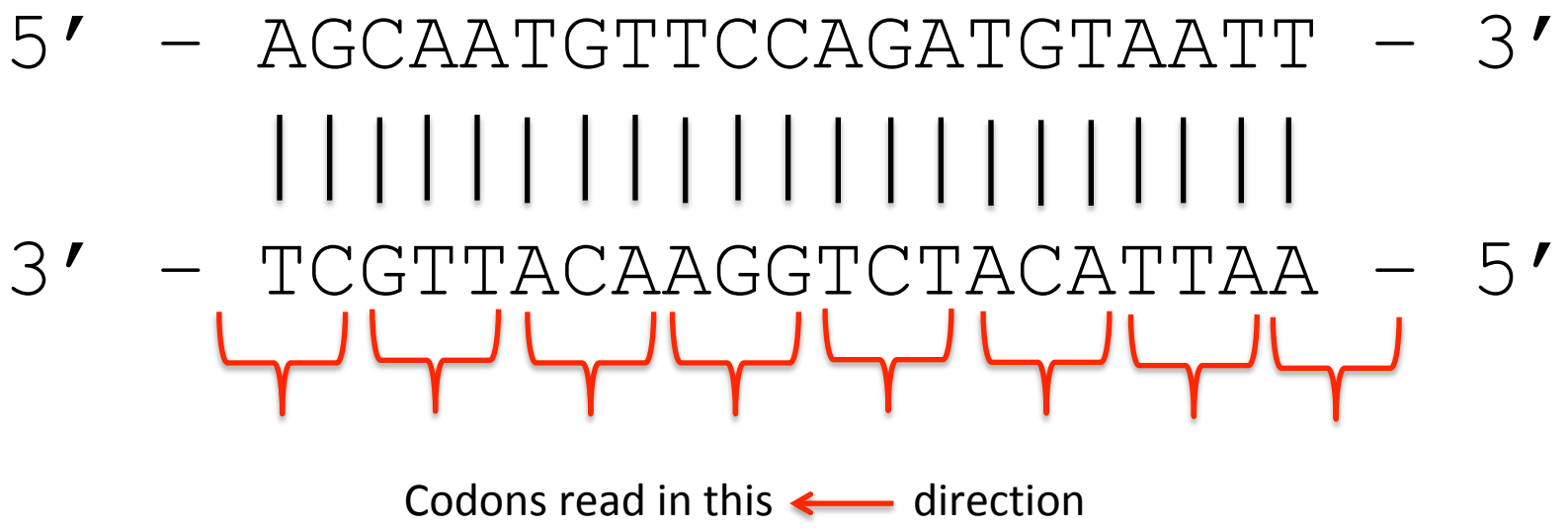
6 Reading Frames: Reading Frame 3



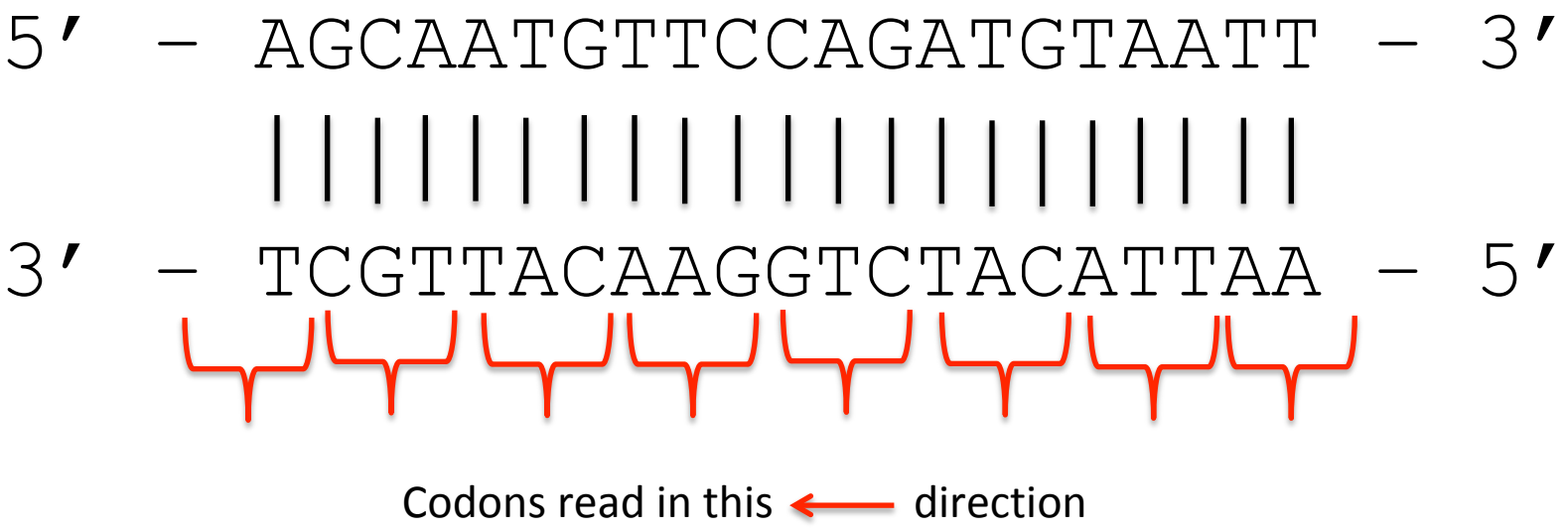
6 Reading Frames: Reading Frame 4



6 Reading Frames: Reading Frame 5



6 Reading Frames: Reading Frame 6



Mini-Project: Gene Finder!

- Create a python program that performs ab initio gene finding by
  - Determining which DNA segments likely code for proteins
  - Outputting the amino acid sequences coded by these regions of DNA
- Run these amino acid sequences through a search engine to determine their function!

Ab-Initio Gene Finding Strategy

- Tell-tale marker: suspiciously long ORFs
- Suspicious is defined as being improbable to find in non-coding DNA

(part of) Next time

Joanne Pratt will guest lecture on a genetic search engine called BLAST (and its uses) as well as the role of BLAST in studying pathogenesis.