Bioinformatics

CS300

Chap 2

Computational Manipulation of DNA

Fall 2019

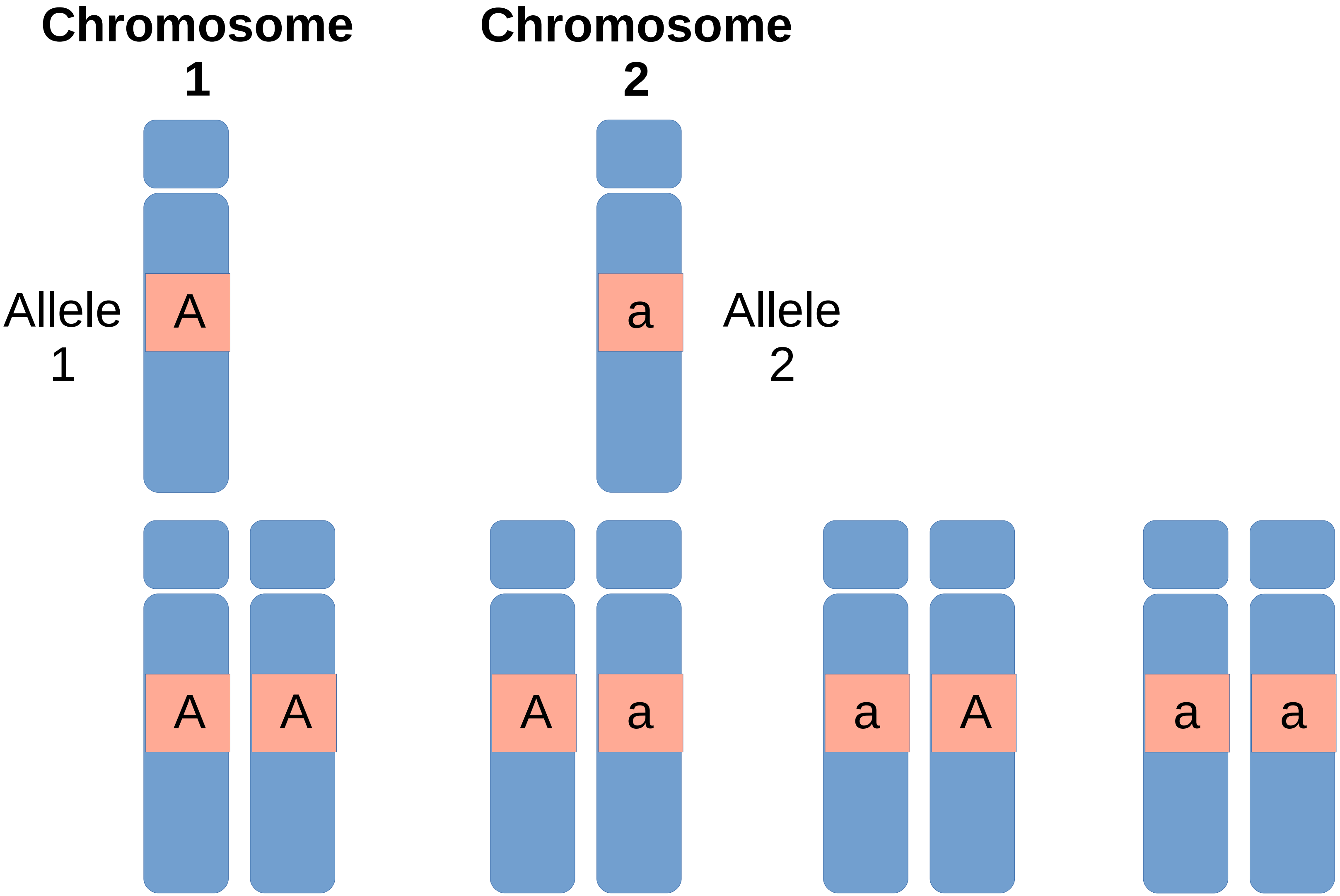
Oliver BONHAM-CARTER

Genes and Alleles

- **Gene**: A distinct sequence of nucleotides forming a piece of a chromosome. In biology, a gene is a sequence of nucleotides in DNA or RNA that codes for a molecule (a *protein*) that has a function. During gene expression, the DNA is first copied into RNA which is then transcribed into protein.

- **Allele**: One of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome.

Patterns of Inheritance by Alleles



Chromosome 1

Chromosome 2

Allele 1

A

a

Allele 2

A A

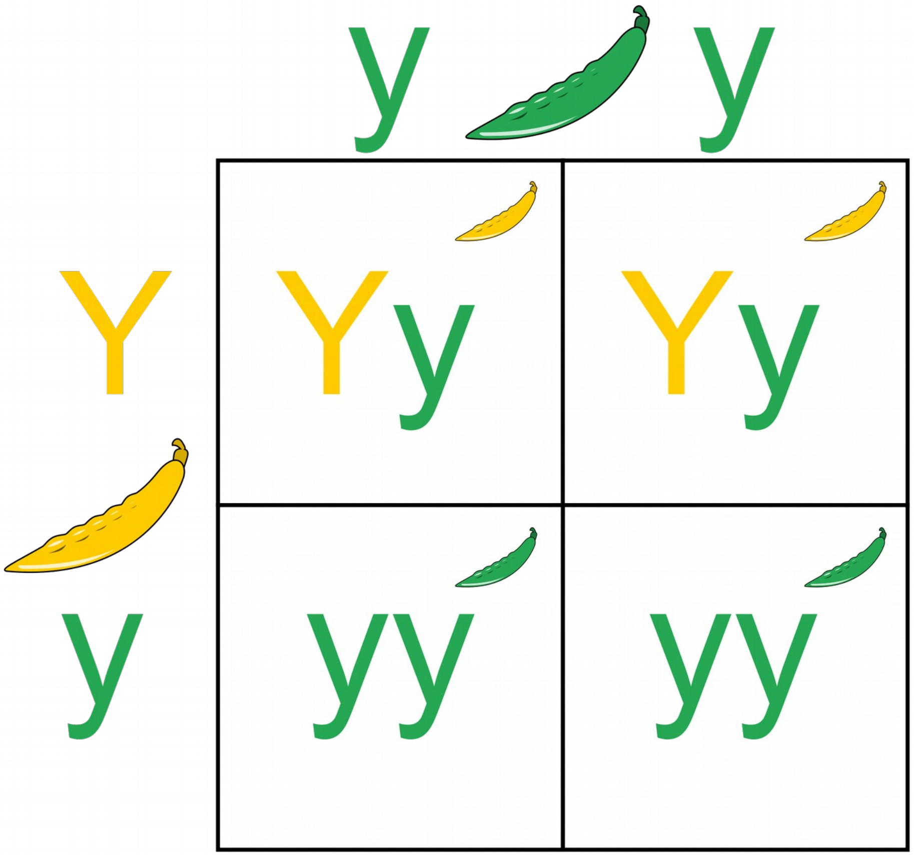
A a

a A

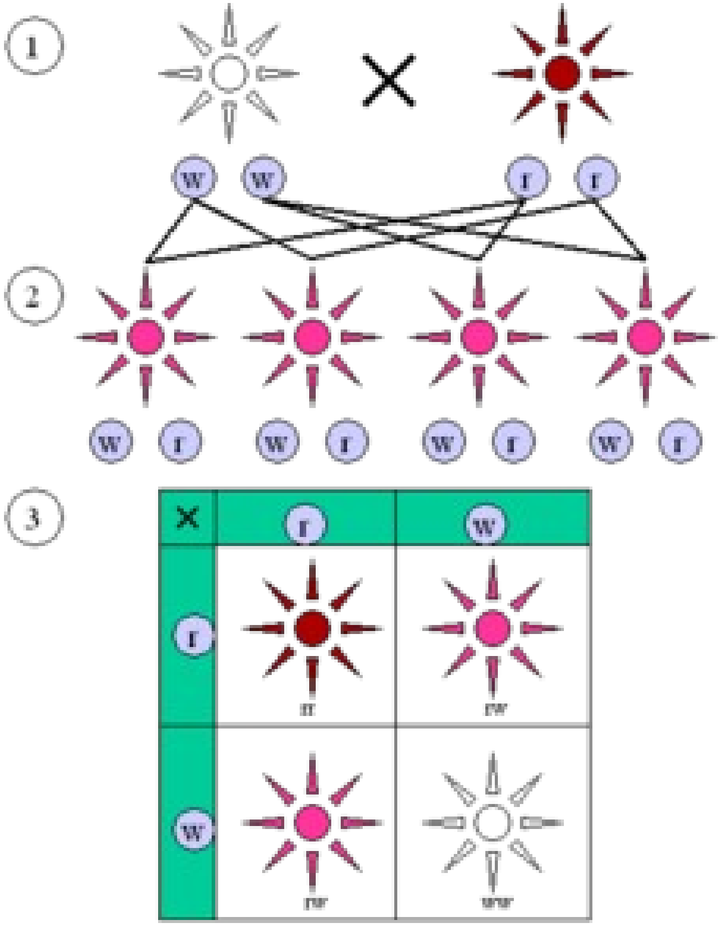
a a

Understanding Alleles

- What is the difference between a *gene* and an *allele*?
- Answer: In the context of cystic fibrosis and the *CFTR* gene
- *Mendelian Genetics studies the alleles*



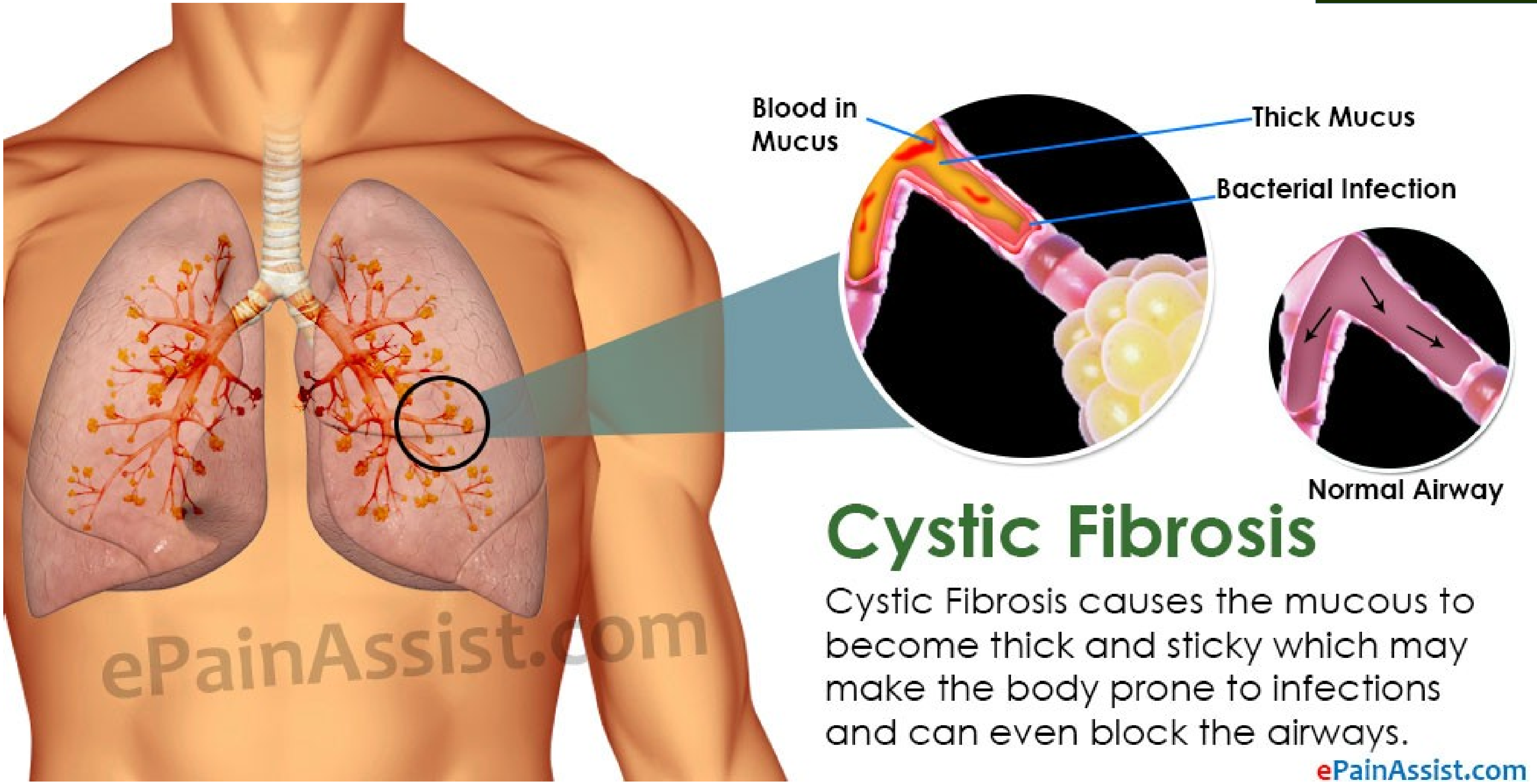
	y	y
Y	Yy	Yy
y	yy	yy



The Cystic Fibrosis Gene

- Cystic Fibrosis Transmembrane conductance: **CFTR**
- Gene product is a bad regulator which fails to move water after displacing chloride ions in epithelial (thin tissue) cells
- Water follows chloride ions by osmosis.
- **What if water regulation were not possible in the cells and organs?**

Cystic Fibrosis



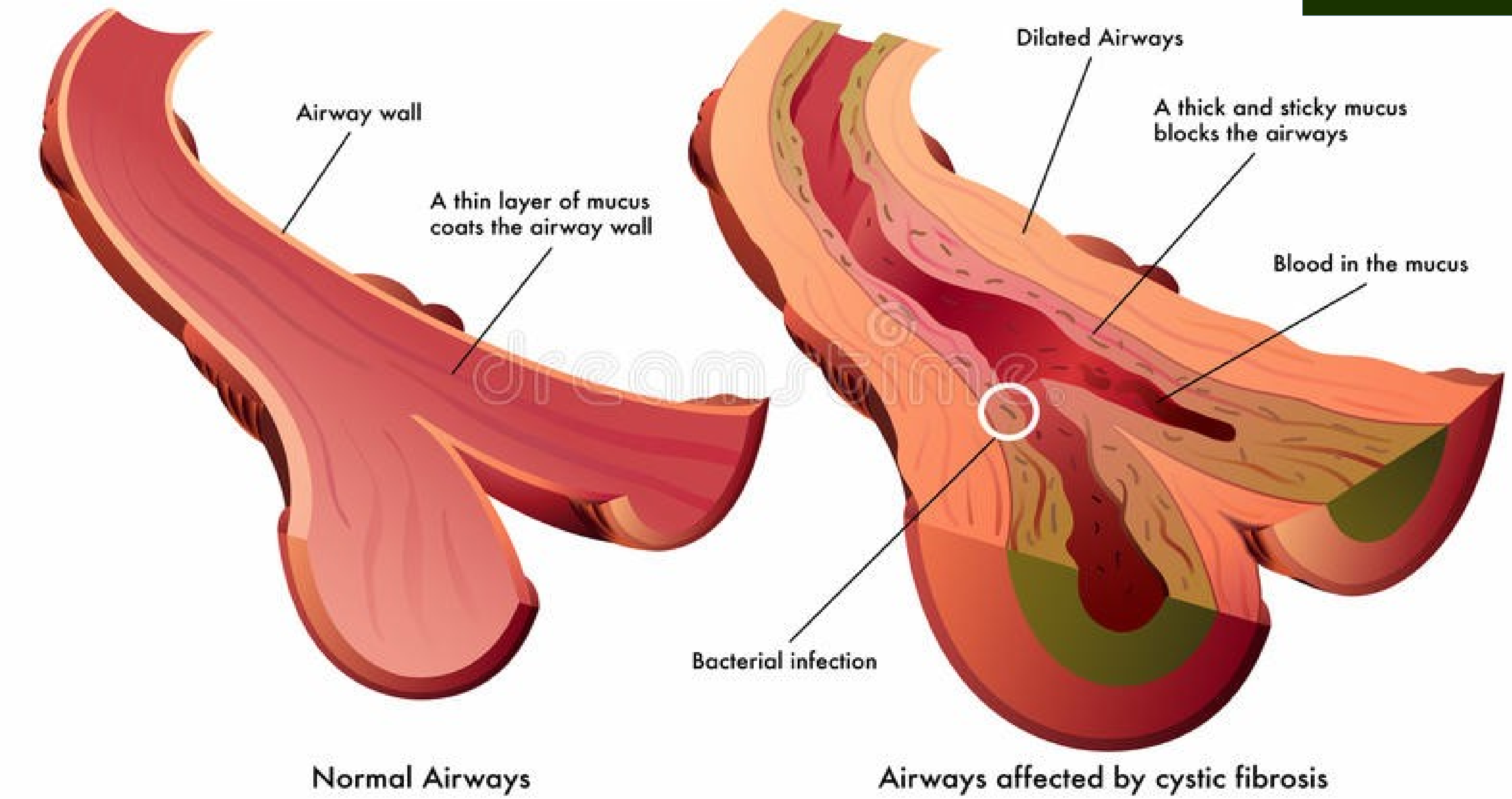
Cystic Fibrosis

Cystic Fibrosis causes the mucous to become thick and sticky which may make the body prone to infections and can even block the airways.

ePainAssist.com

- Inherited medical condition of the secretory glands (producers of mucous and sweat)

Cystic Fibrosis: Symptoms

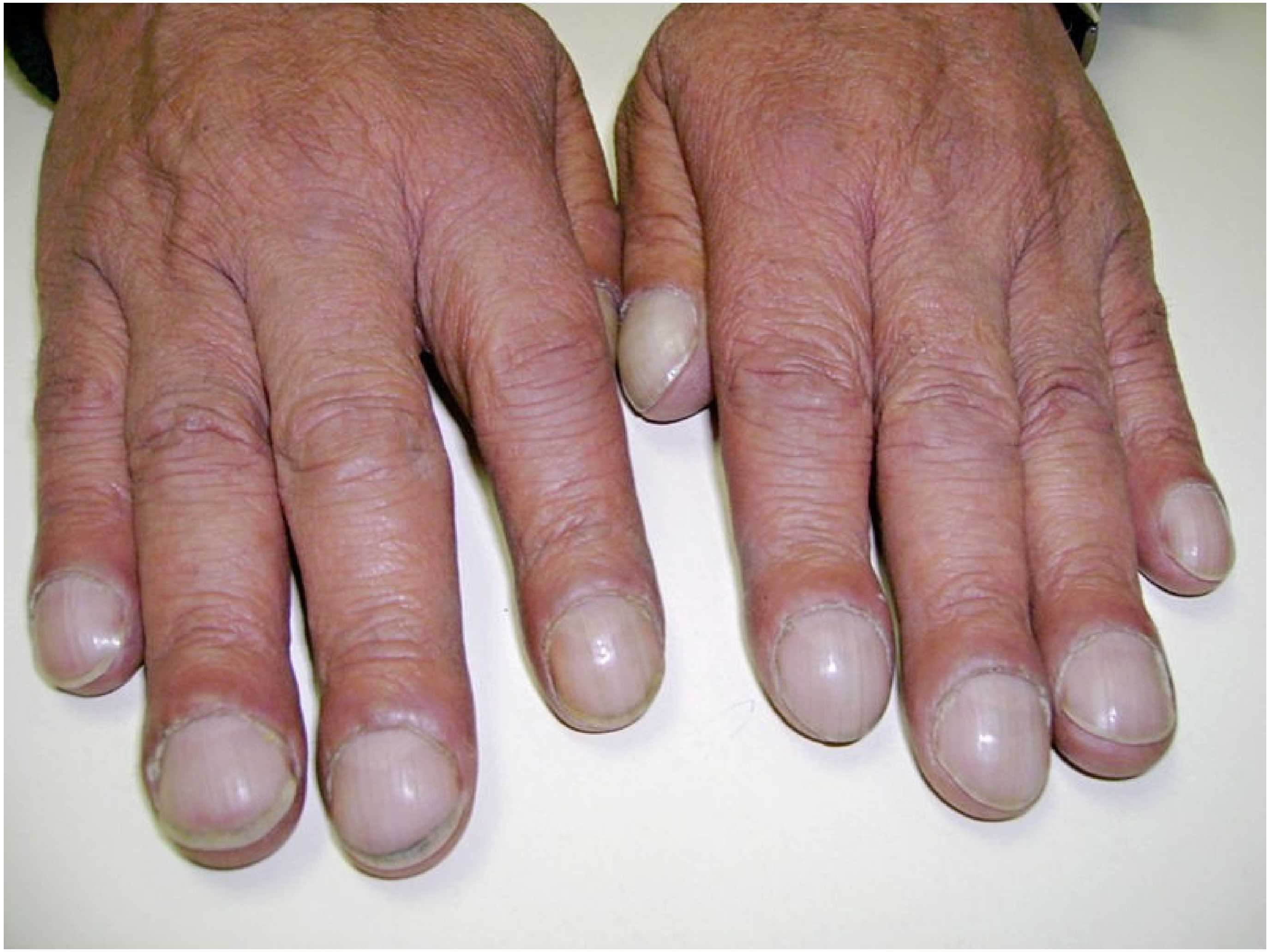

Airway wall
A thin layer of mucus coats the airway wall
Normal Airways
Dilated Airways
A thick and sticky mucus blocks the airways
Blood in the mucus
Bacterial infection
Airways affected by cystic fibrosis

- Restricted flow in airways from mucous build-ups.
- Suffocation

A Build-Up of Anything is Bad

- What if the the garbage collection crews in Paris went on strike (as they did in 2016)?

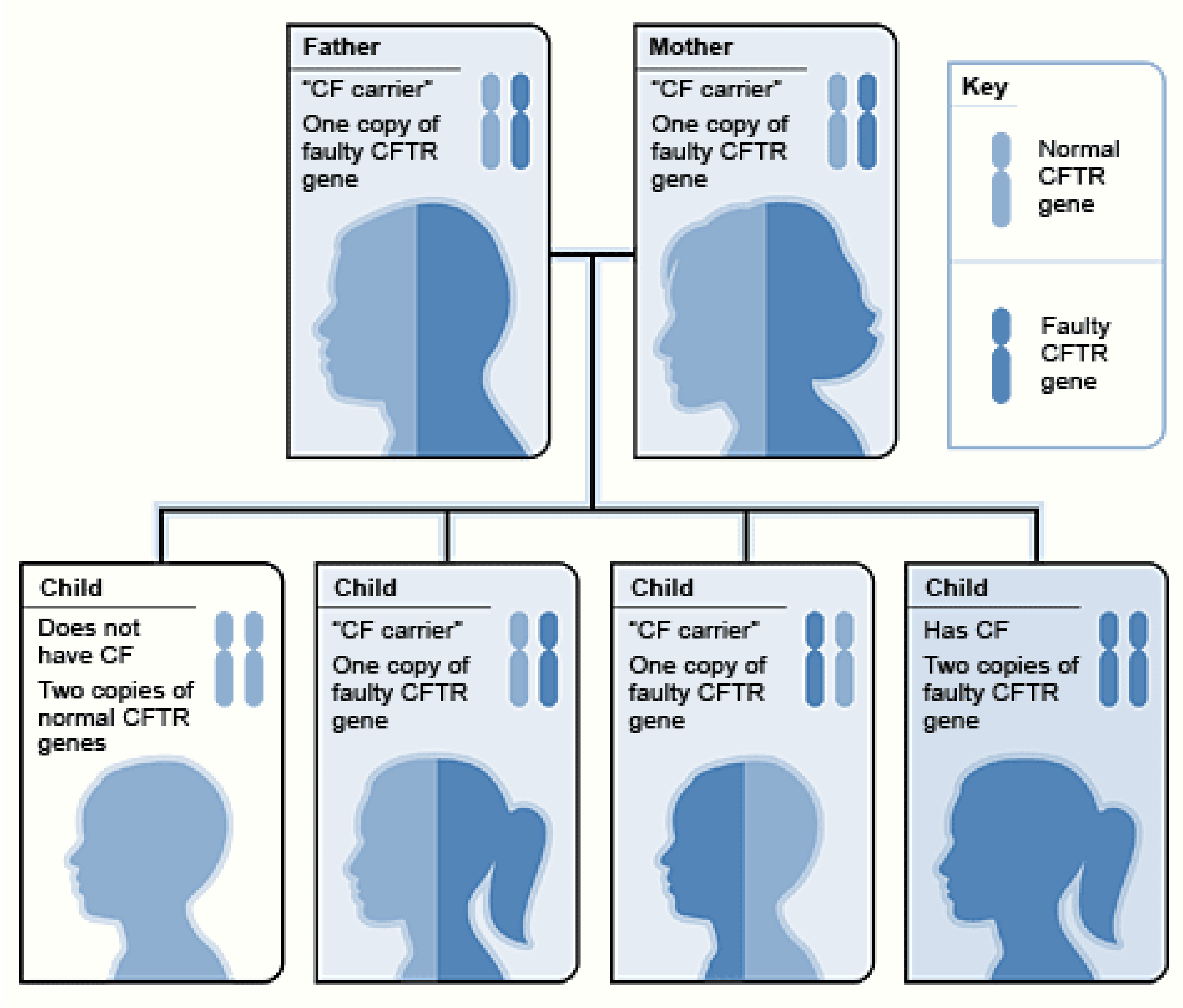
Cystic Fibrosis: Symptoms



- Clubbed fingers: occurs in heart and lung diseases that reduce the amount of oxygen in the blood

Cystic Fibrosis: Inheritance

- Autosomal recessive type condition: one faulty gene is inherited from both parents (together) in order for the offspring to get this condition
- Modeled via Mendelian Genetics
- Impossible to know that someone is sure to get a condition.



Father
"CF carrier"
One copy of faulty CFTR gene
Mother
"CF carrier"
One copy of faulty CFTR gene
Key
Normal CFTR gene
Faulty CFTR gene
Child
Does not have CF
Two copies of normal CFTR genes
Child
"CF carrier"
One copy of faulty CFTR gene
Child
"CF carrier"
One copy of faulty CFTR gene
Child
Has CF
Two copies of faulty CFTR gene

The Cystic Fibrosis Gene

- Cystic Fibrosis Transmembrane conductance: **CFTR**
- Gene product is a bad regulator which fails to move water after displacing chloride ions in epithelial (thin tissue) cells
- Water follows chloride ions by osmosis.
- What happens if water regulation is impossible in the cells and organs?

Three Bad Proteins From the Four

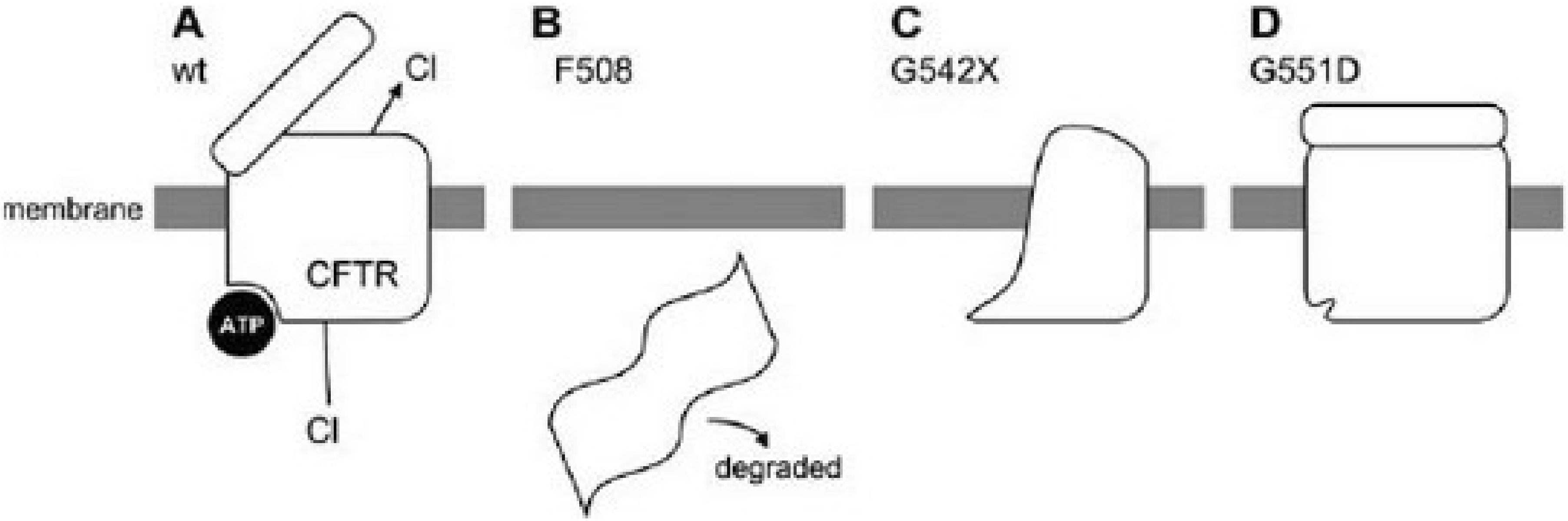
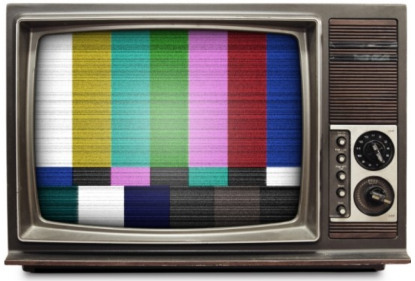


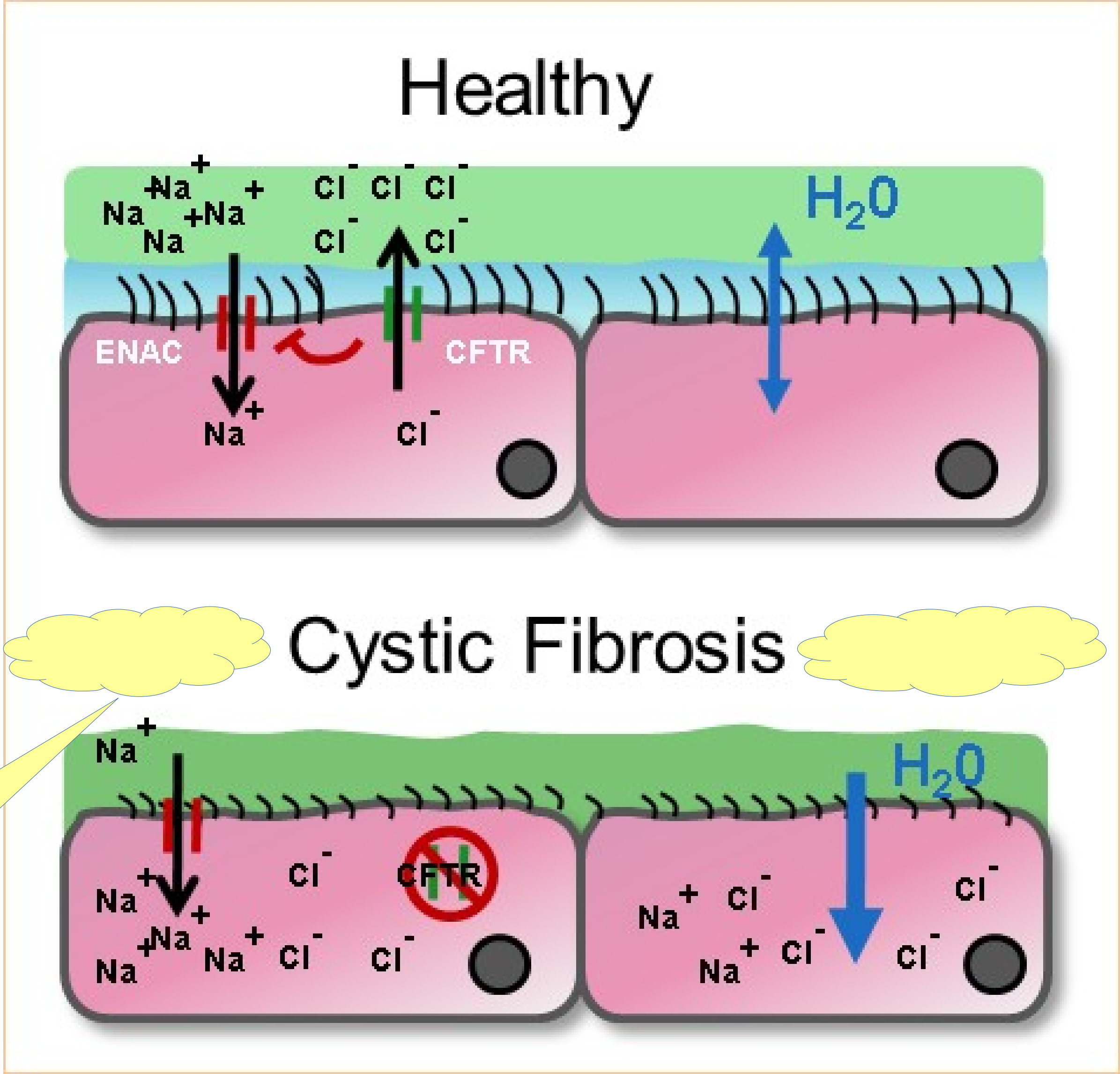
Figure 2.2 The wild-type allele (A) of the CFTR gene produces a chloride transport protein localized in the membrane; three different common CF alleles illustrated here result in variant proteins that are folded incorrectly (ΔF508; B), truncated (G542X; C), or unable to transport chloride (G551D; D).

- Short video of membrane transport proteins https://www.youtube.com/watch?v=EuLVCYrurok



The Cystic Fibrosis Gene

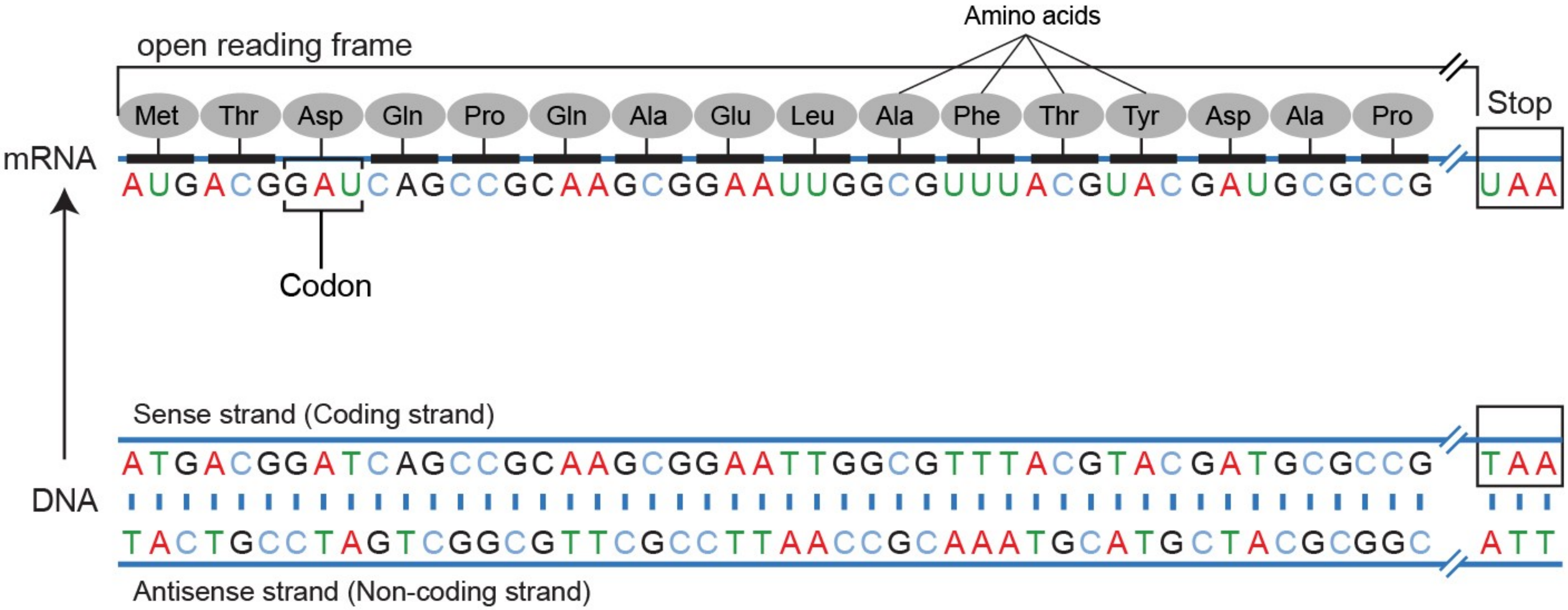
- Gene codes for four different proteins: only one working type to move chloride ions and enable water displacement.



Mucous build-up

Open Reading Frames

- An open reading frame (ORF) is the part of a reading frame that has the ability to be translated into protein.
- An ORF is a continuous stretch of codons that begins with a **start** codon (usually AUG) and ends at a **stop** codon (usually UAA, UAG or UGA).



Amino acids
open reading frame
Met
Thr
Asp
Gln
Pro
Gln
Ala
Glu
Leu
Ala
Phe
Thr
Tyr
Asp
Ala
Pro
Stop
mRNA
AUGACGGAUCAGCCGCAAGCGGAAUUGGCGUUUACGUACGAUGCGCCG
UAA
Codon
Sense strand (Coding strand)
ATGACGGATCAGCCGCAAGCGGAATTGGCGTTTACGTACGATGCGCCG
TAA
DNA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCAAATGCATGCTACGCGGC
ATT
Antisense strand (Non-coding strand)

Cite:
https://www.genome.gov/genetics-glossary/Open-Reading-Frame

Open Reading Frames: Simple Example

- Pam Can See The Man and Dog
- Frame shift by one letter!
- ~~P~~ amC anS eeT heM ana ndD ~~og~~
- Frame shift by two letters!
- ~~Pa~~ mCa nSe eTh eMa nan dDo ~~g~~
- Frame shift by three letters
- ~~Pam~~ Can See The Man and Dog

Reading by triplets

Notice how the code changes depending on where you start reading? (That is a *frameshift*.)

Open Reading Frames: DNA Example

Note: RF means *reading frame, where you start reading the words.*

Original: **CAATGGCGAATCGACGTGTATAAA**

RF1 - 5' - CAA TGG CGA ATC GAC GTG TAT AAA – 3'

RF2 - 5' - C AAT GGC GAA TCG ACG TGT ATA AA – 3'

RF 3 - 5' - CA **ATG** GCG AAT CGA CGT GTA **TAA** A – 3'

3' - CAA TGG CGA ATC GAC GTG TAT AAA - 5' – RF 4

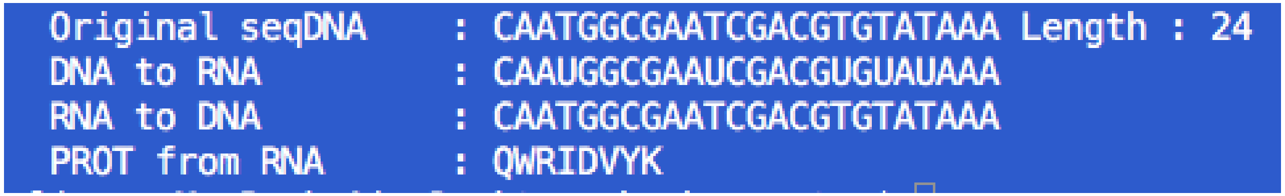
3' - C AAT GGC GAA TCG ACG TGT ATA AA - 5' – RF 5

3' - CA ATG GCG **AAT** CGA CGT **GTA** TAA A - 5' – RF 6

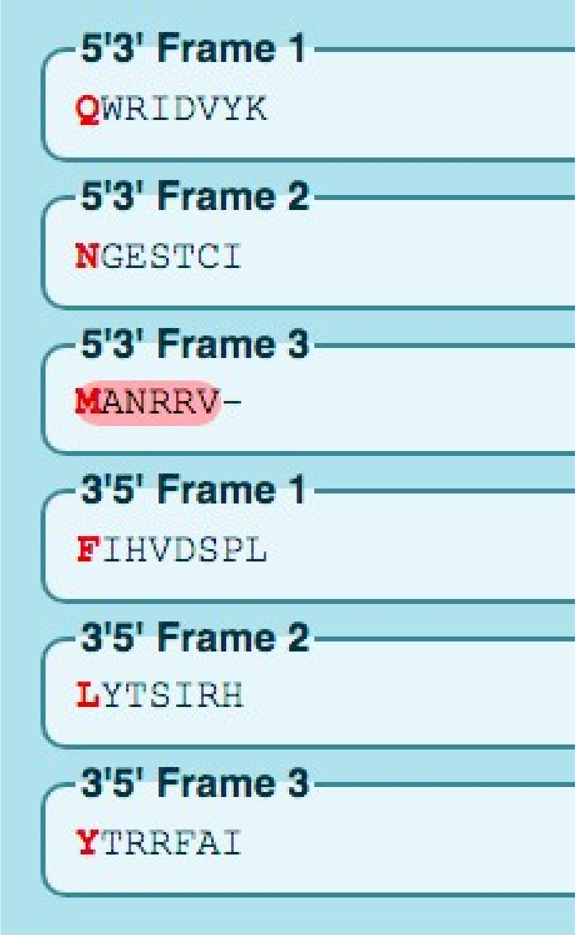
Open Reading Frames: Online

- Original: **CAATGGCGAATCGACGTGTATAAA**
- Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.
  - https://web.expasy.org/translate/

Biopython:: SmallTranslator_i.py



```
 Original seqDNA     : CAATGGCGAATCGACGTGTATAAA Length : 24
 DNA to RNA          : CAAUGGCGAAUCGACGUGUAUAAA
 RNA to DNA          : CAATGGCGAATCGACGTGTATAAA
 PROT from RNA       : QWRIDVYK
```



5'3' Frame 1

QWRIDVYK

5'3' Frame 2

NGESTCI

5'3' Frame 3

MANRRV-

3'5' Frame 1

FIHVDSPL

3'5' Frame 2

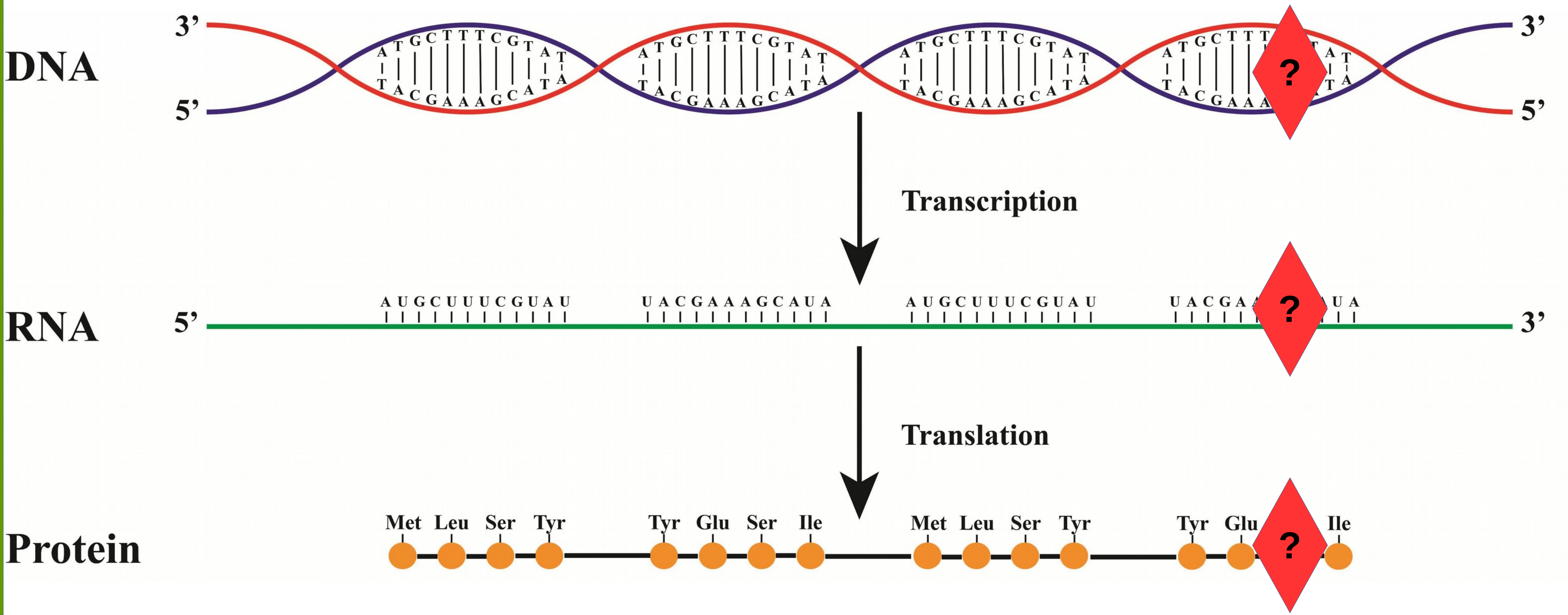
LYTSIRH

3'5' Frame 3

YTRRFAI

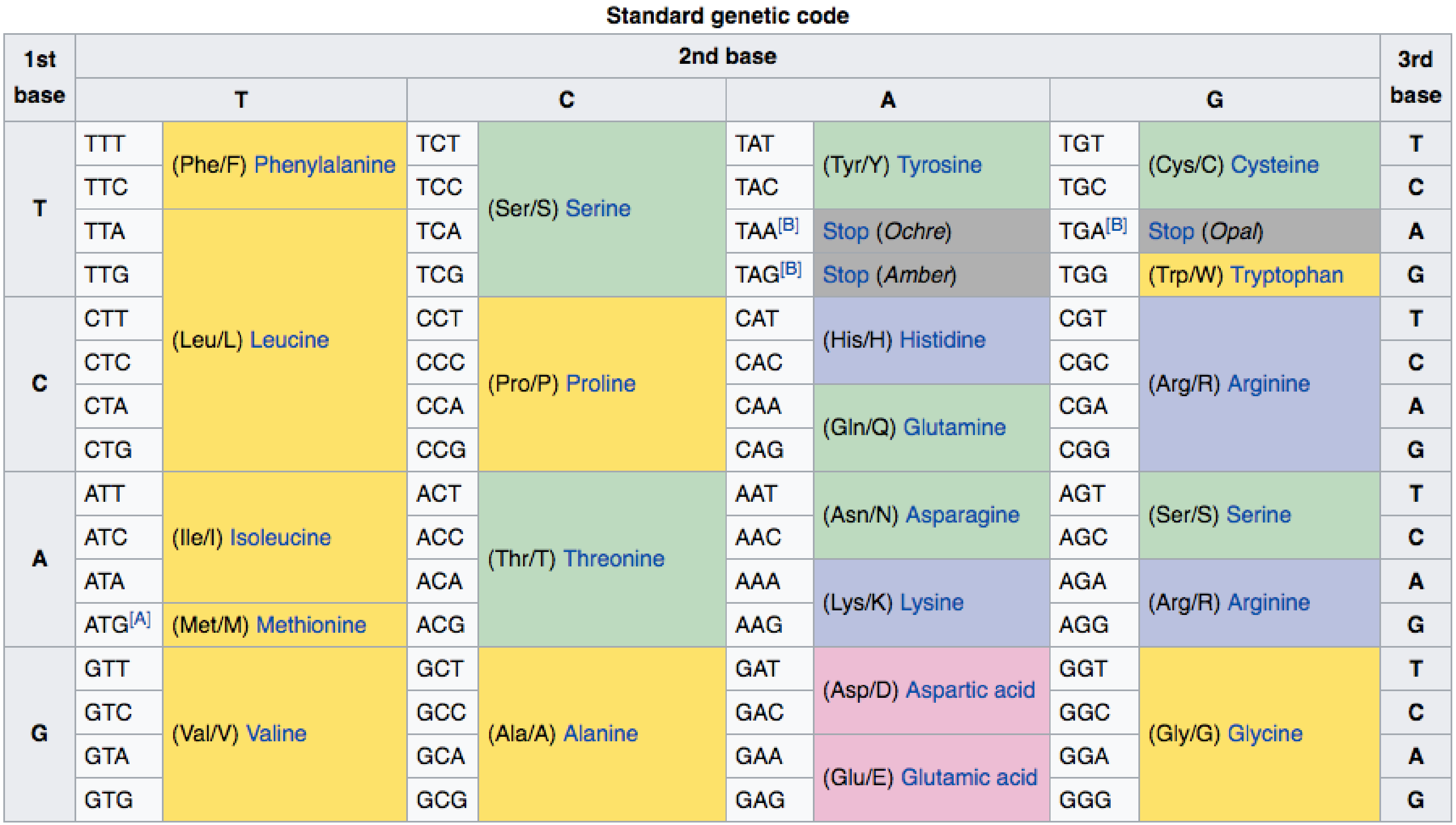
Sequence is Carrier?

- How do we determine if a sequence carries the Cystic Fibrosis allele?
- Get DNA sample and translate into protein. Then compare product protein sequence to that of a “working protein”
- Is there a difference between the protein sequences?



Remember the Codon Table?

- DNA triplets read in groups of three called codons, code amino acids
- T’s from DNA are read as U’s as RNA after transcription

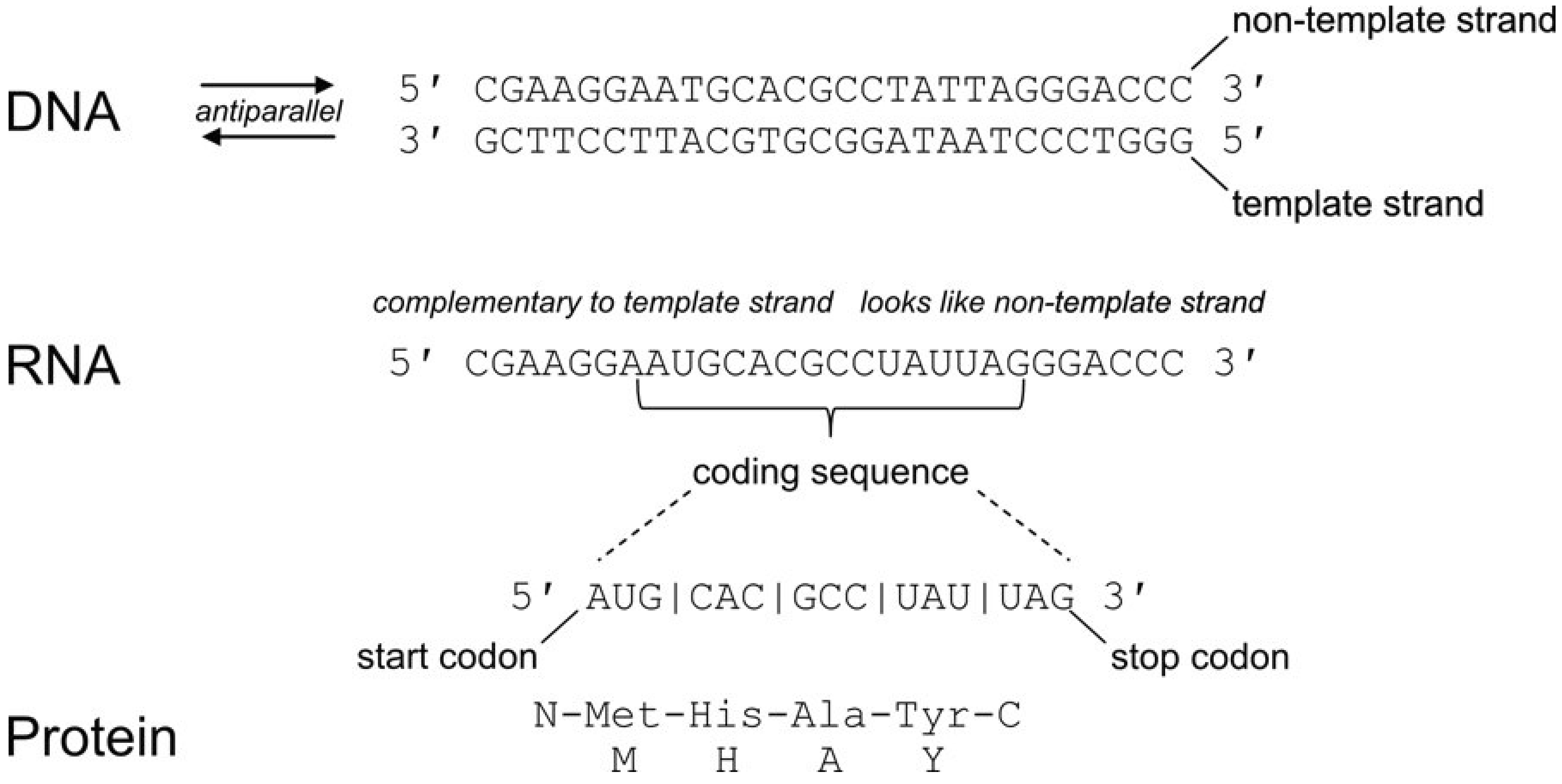


Standard genetic code

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[B]	Stop (*Ochre*)	TGA[B]	Stop (*Opal*)	A
	TTG		TCG		TAG[B]	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

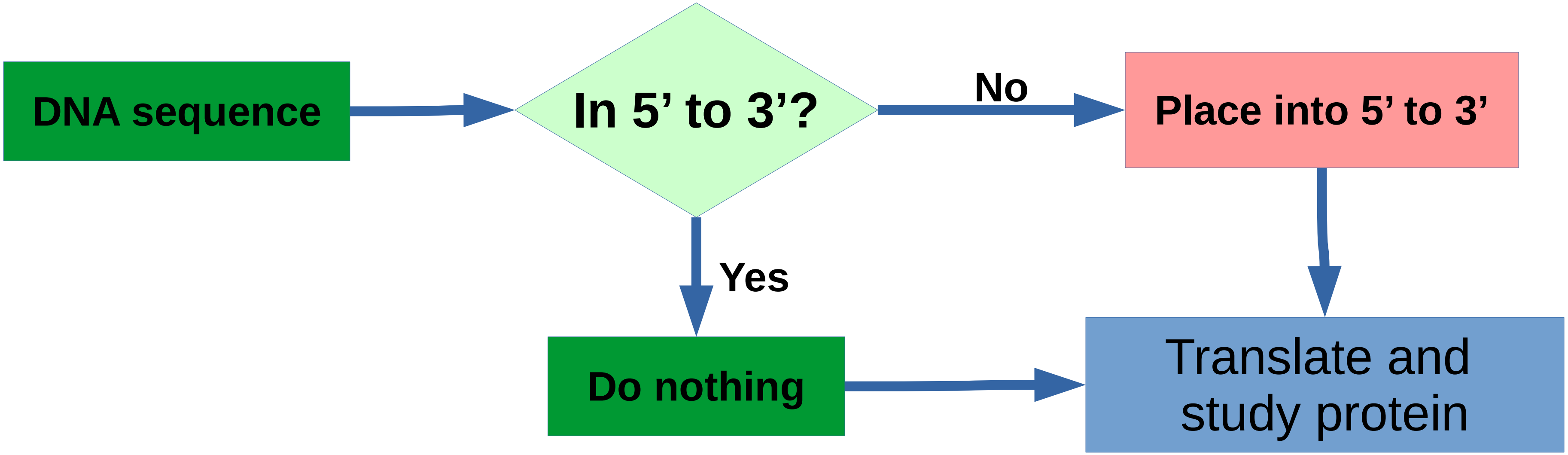
Summary:
The Steps to Study Protein

- Translating DNA to find defects in the protein


DNA
antiparallel
non-template strand
5′ CGAAGGAATGCACGCCTATTAGGGACCC 3′
3′ GCTTCCTTACGTGCGGATAATCCCTGGG 5′
template strand
complementary to template strand
looks like non-template strand
RNA
5′ CGAAGGAAUGCACGCCUAUUAGGGACCC 3′
coding sequence
5′ AUG|CAC|GCC|UAU|UAG 3′
start codon
stop codon
Protein
N-Met-His-Ala-Tyr-C
M H A Y

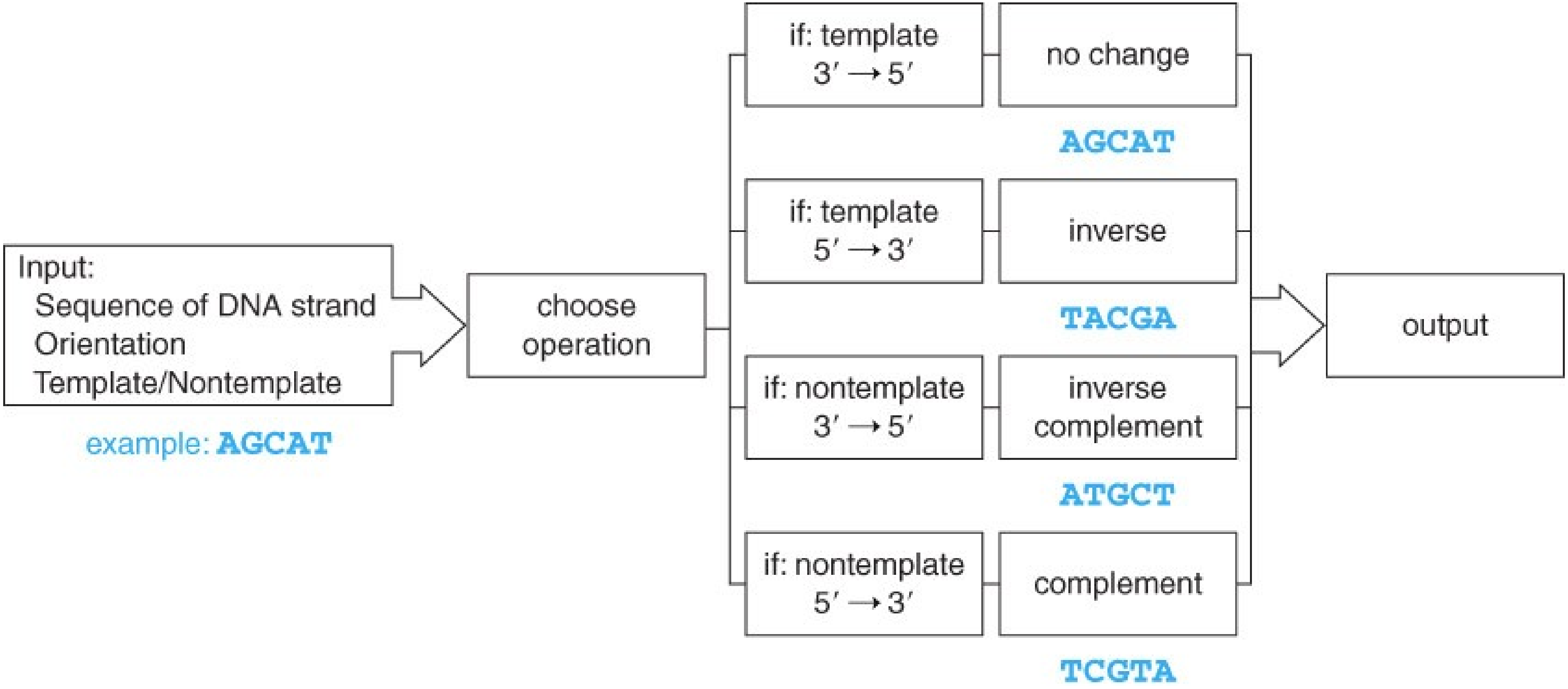
Remember: DNA Must Be In 3’ to 5’ Direction To Find The Sequence

- Unlabeled strands of DNA are assumed to be in the 5’ to 3’, (left to right) direction.
- A new sequence is given to us for analysis.
- What are the steps to place this sequence into a format for use with bioinformatics tools?



DNA Manipulation Algorithm

- A series of steps when handling DNA



Output DNA in 3’ to 5’

Translation Algorithm

- Input: mRNA strand in the **5’** → **3’** orientation
- Output: amino acid sequence
  - Traverse the string looking at one codon at a time
  - Add one amino acid corresponding to the protein sequence.

WAIT! Why is the 5’ to 3’
direction so important?!
Remember the carbon atoms on DNA?

Review Question 1

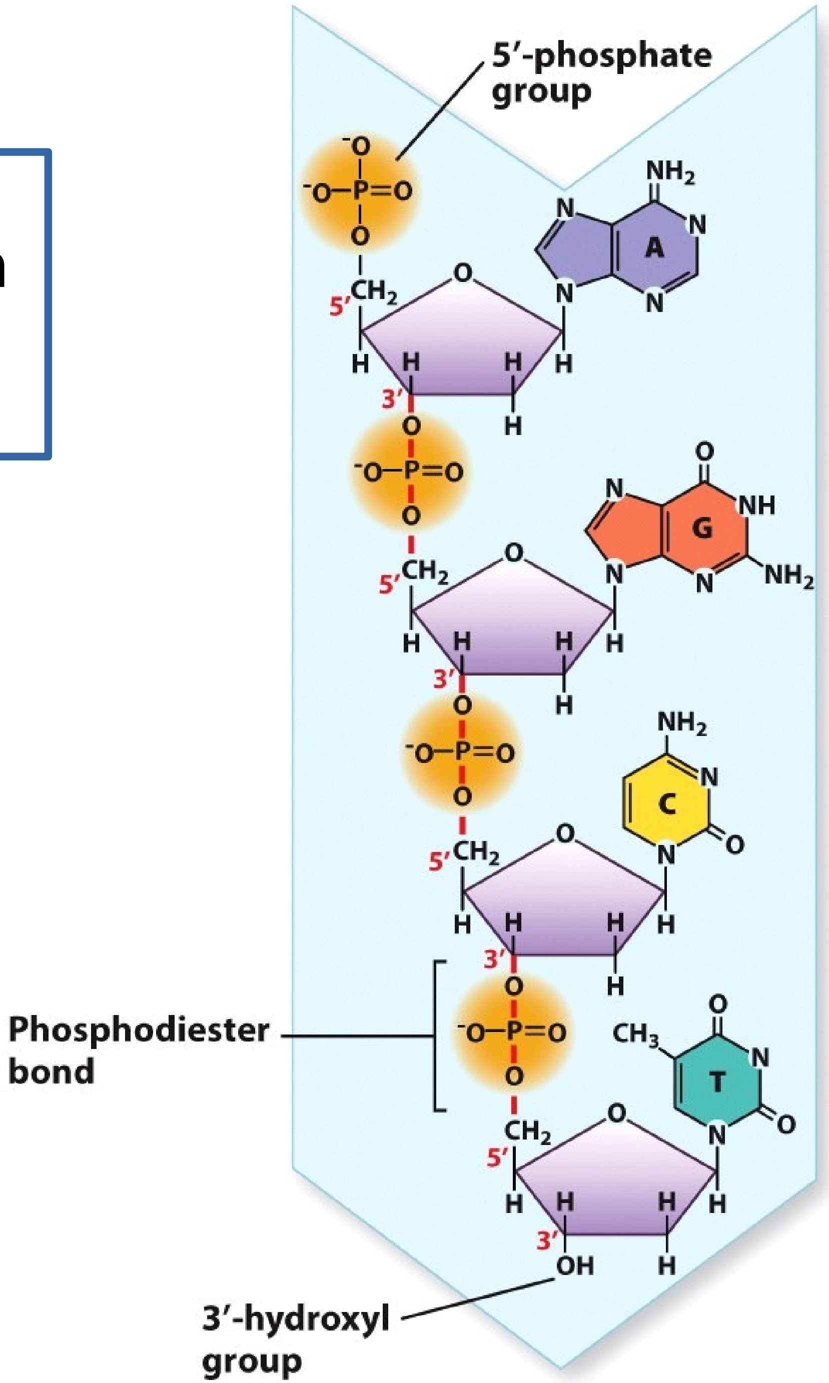
In the DNA sequence 5'–AGCT–3', the phosphodiester linkage between the adenine and the guanine connects:

A. The 2' end of the adenine to the 4' end of the guanine.

B. The 5' end of the adenine to the 3' end of the guanine.

C. The 5' end of the guanine to the 1' end of the adenine.

D. The 3' end of the adenine to the 5' end of the guanine.



Review Question 1

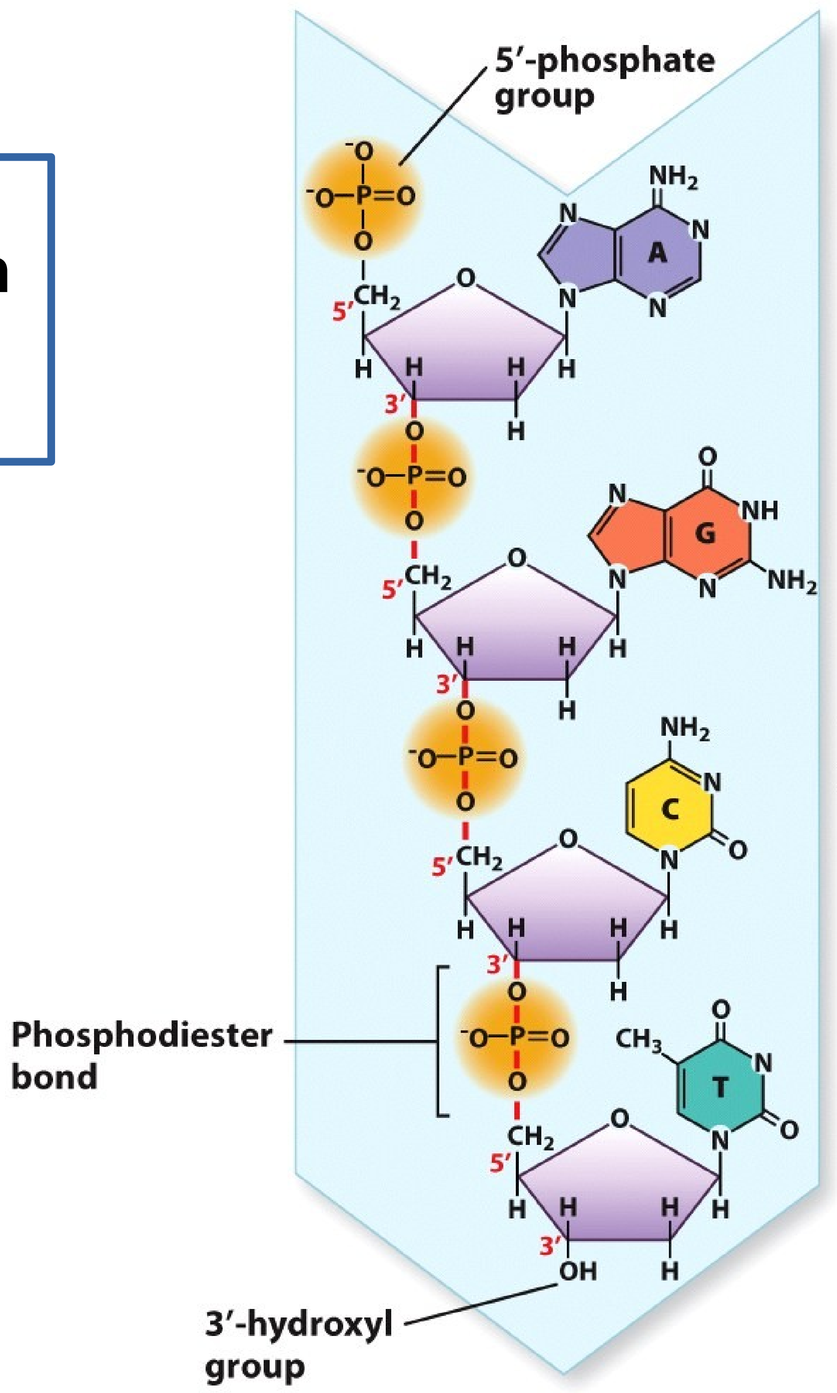
In the DNA sequence 5'–AGCT–3', the phosphodiester linkage between the adenine and the guanine connects:

A. The 2' end of the adenine to the 4' end of the guanine.

B. The 5' end of the adenine to the 3' end of the guanine.

C. The 5' end of the guanine to the 1' end of the adenine.

D. The 3' end of the adenine to the 5' end of the guanine.



5′-phosphate group
A
G
C
T
Phosphodiester bond
3′-hydroxyl group