Bioinformatics

CS300

Chap 2

Computational Manipulation of DNA

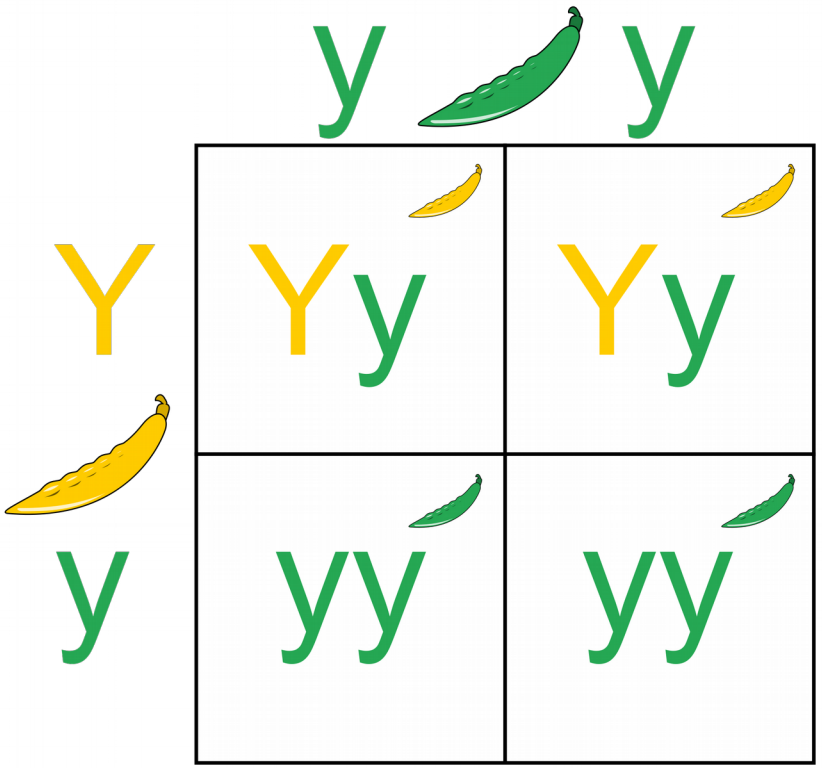
Fall 2017

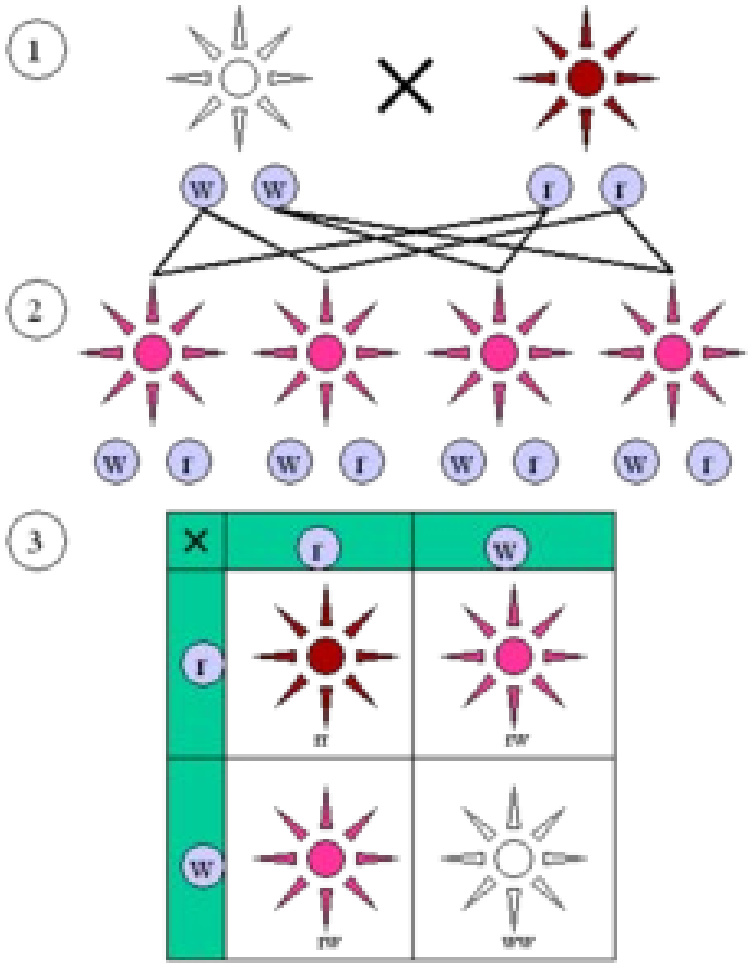
Oliver Bonham-Carter

Consider this...

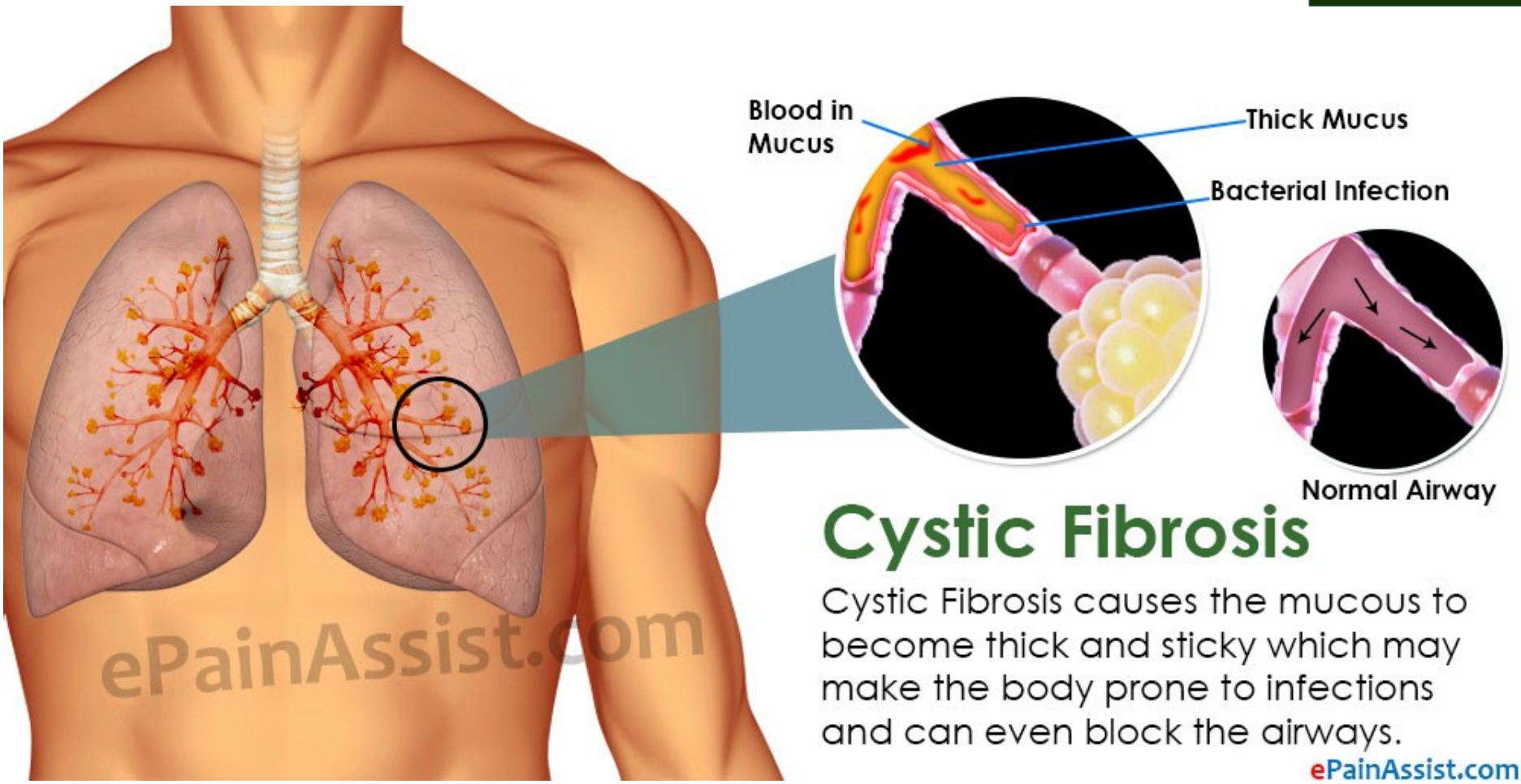
- What is the difference between a gene and an allele?
- Answer in the context of cystic fibrosis and the *CFTR* gene

  Hint: *Think Mendelian Genetics*





Cystic Fibrosis



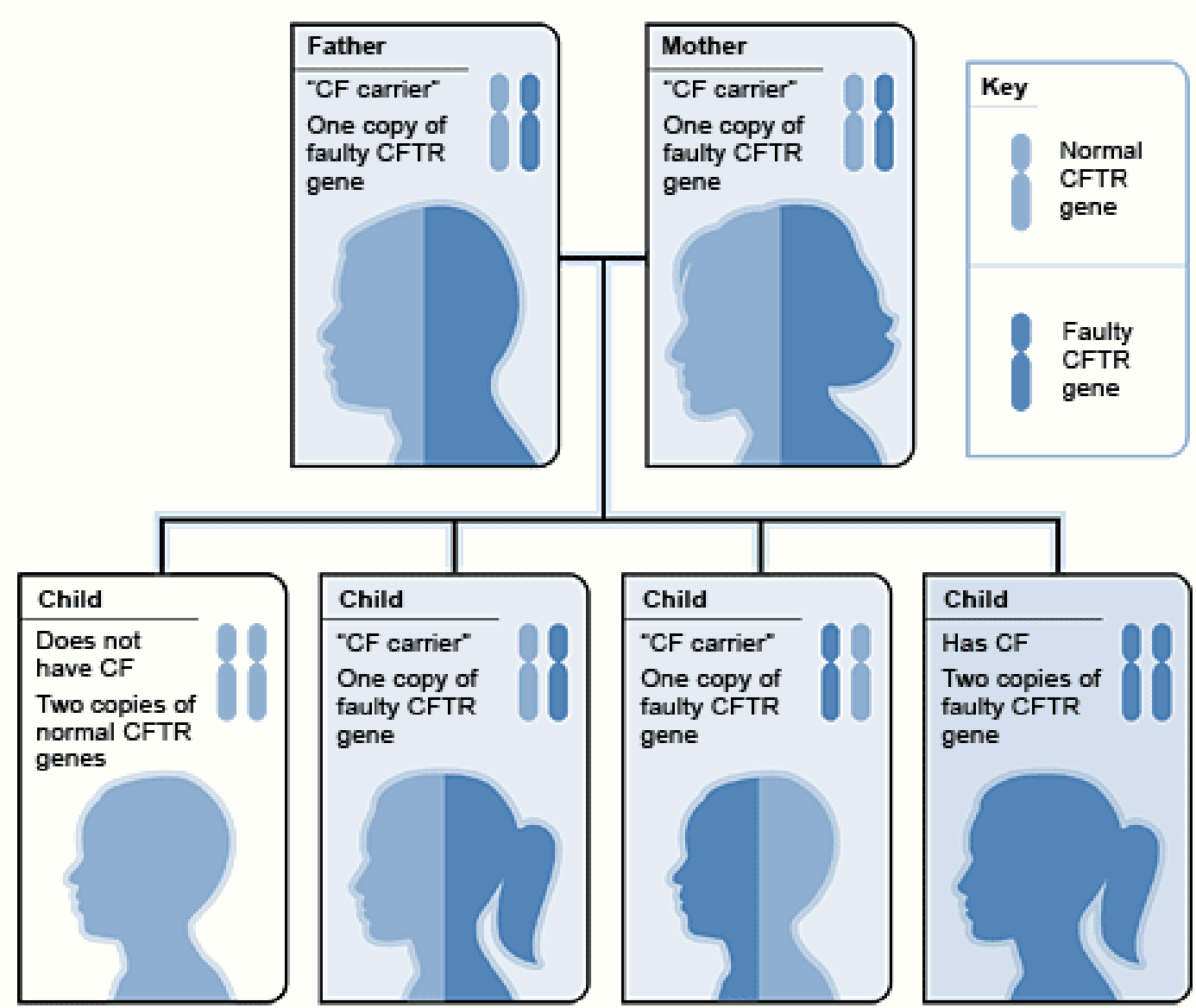
- Inherited medical condition of the secretory glands (producers of mucous and sweat)

Cystic Fibrosis

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

Cystic Fibrosis

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

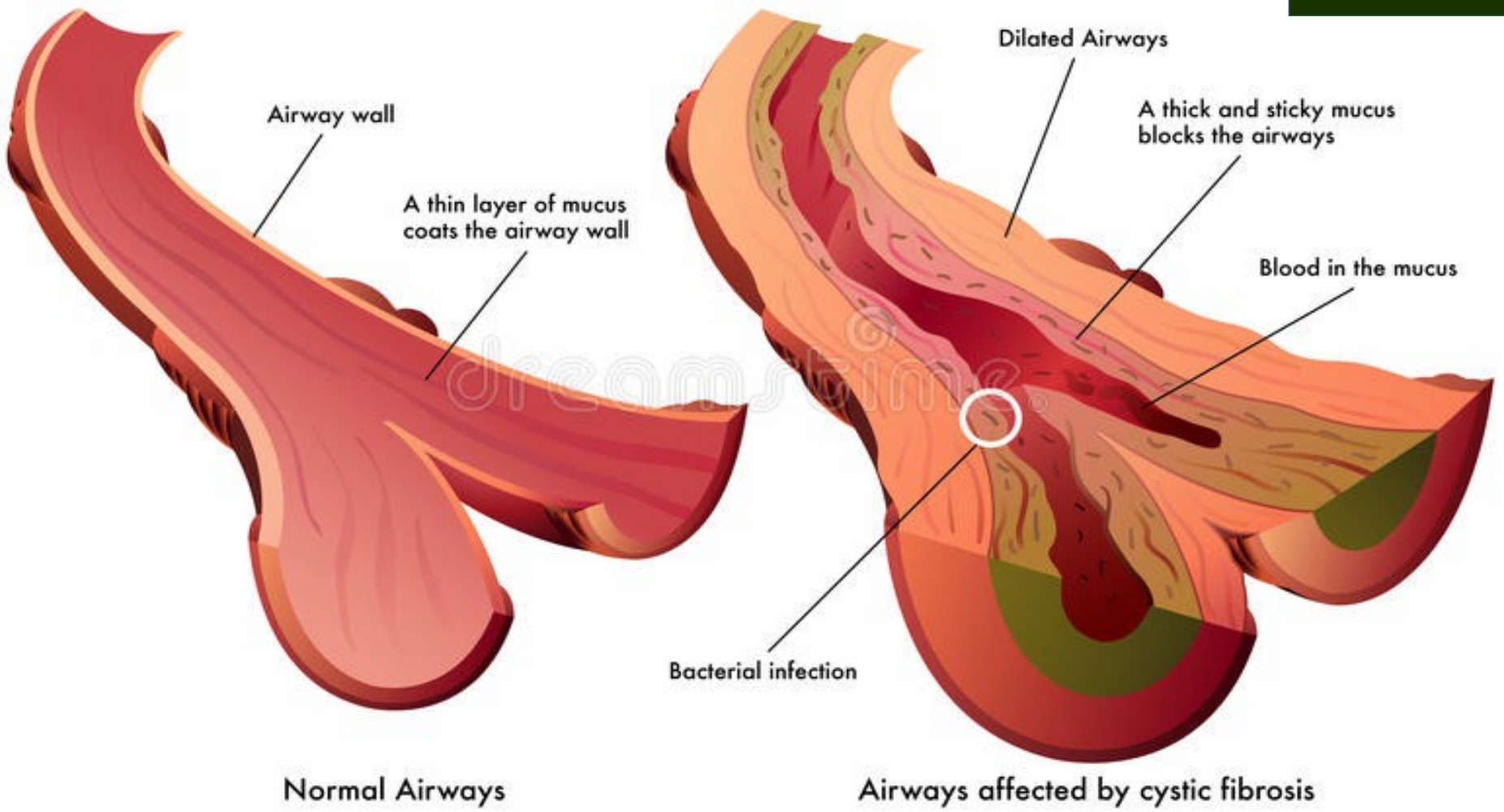


The Cystic Fibrosis Gene

- 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



- 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)?

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.
- https://www.youtube.com/watch?v=EuLVCYrurok

- **What if water regulation were not possible in the cells and organs?**

Three Bad Proteins From the Four

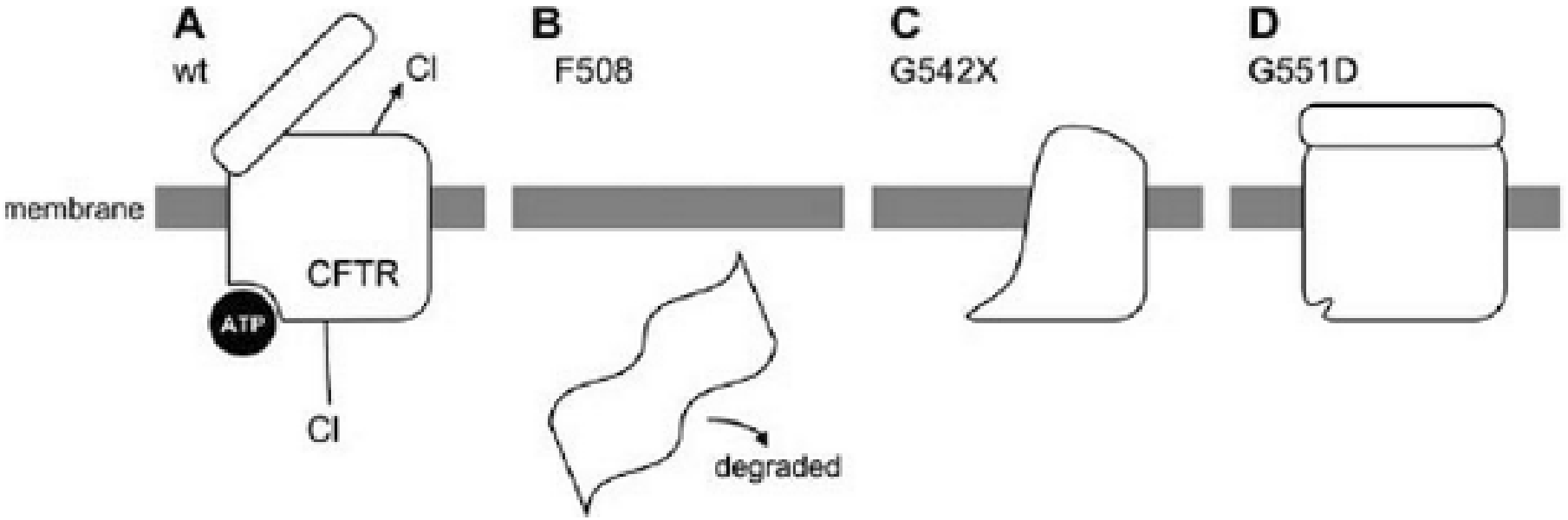
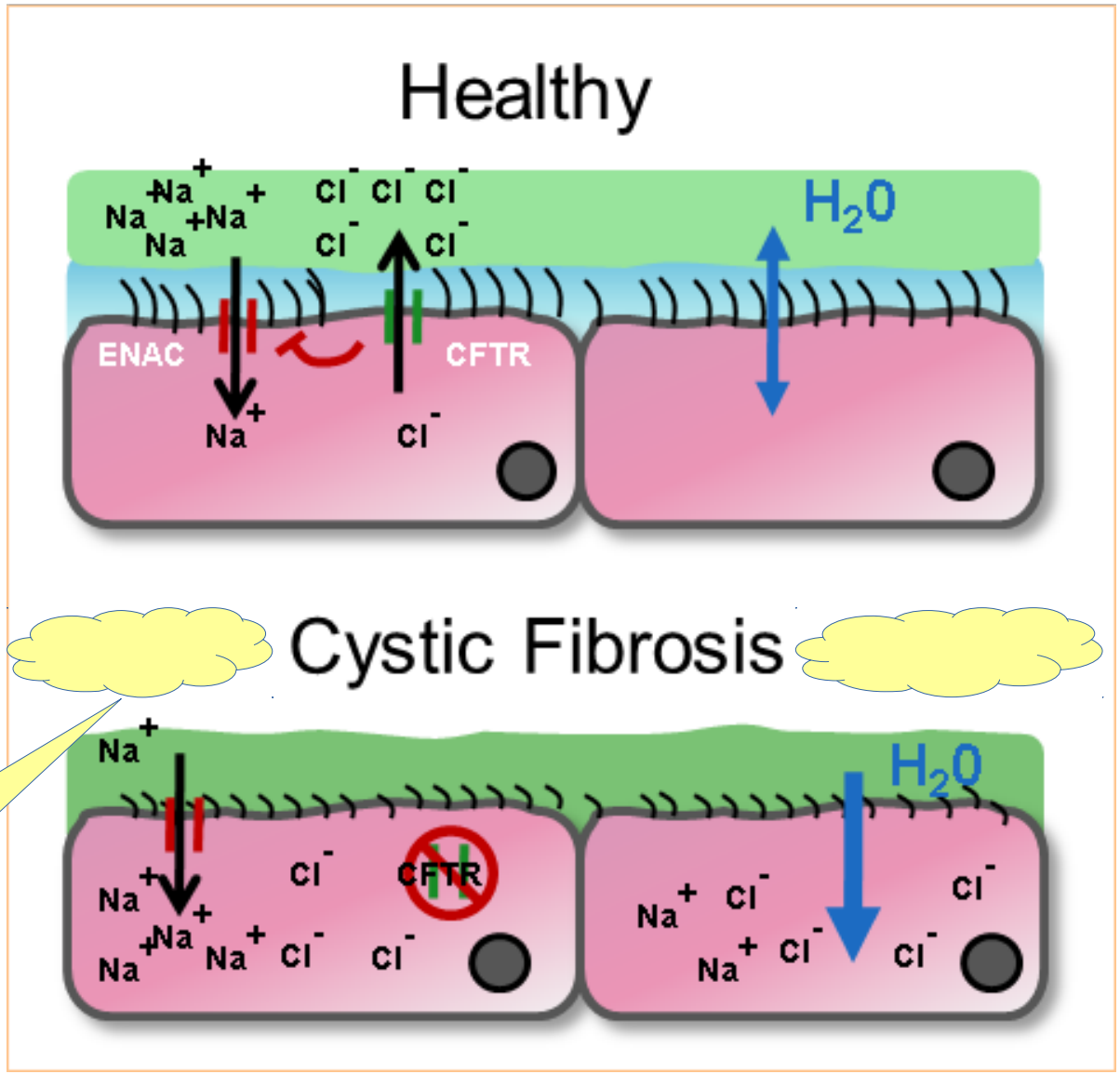


Figure 2.2 The wild-type allele (A) of the CFTR gene produces a chloride transport protein localized in the membrane; three different common CF alleles illustrated here result in variant proteins that are folded incorrectly (ΔF508; B), truncated (G542X; C), or unable to transport chloride (G551D; D).

The Cystic Fibrosis Gene

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



Open Reading Frames: Explaining Disorders?

- 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

Notice how the code changes depending on where you start reading?

Open Reading Frames

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’

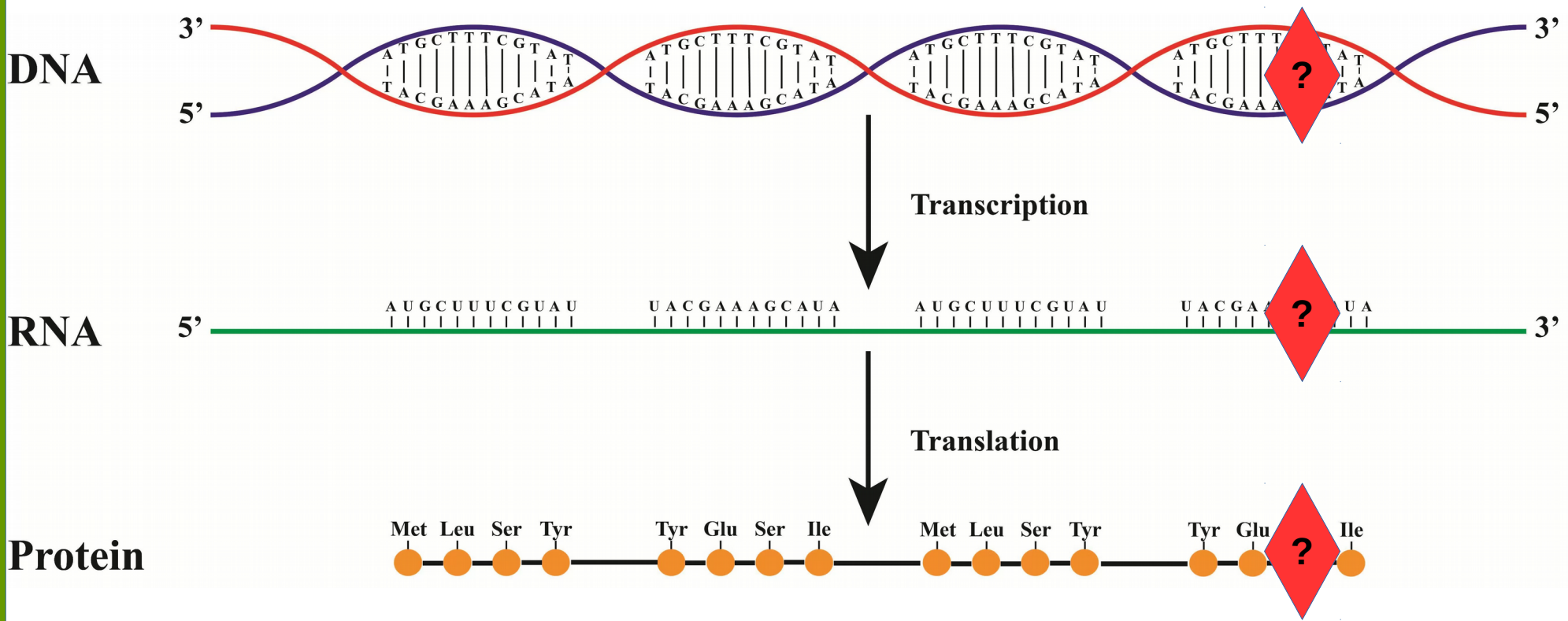
3’ - CAA TGG CGA ATC GAC GTG TAT AAA - 5’ – RF 4

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

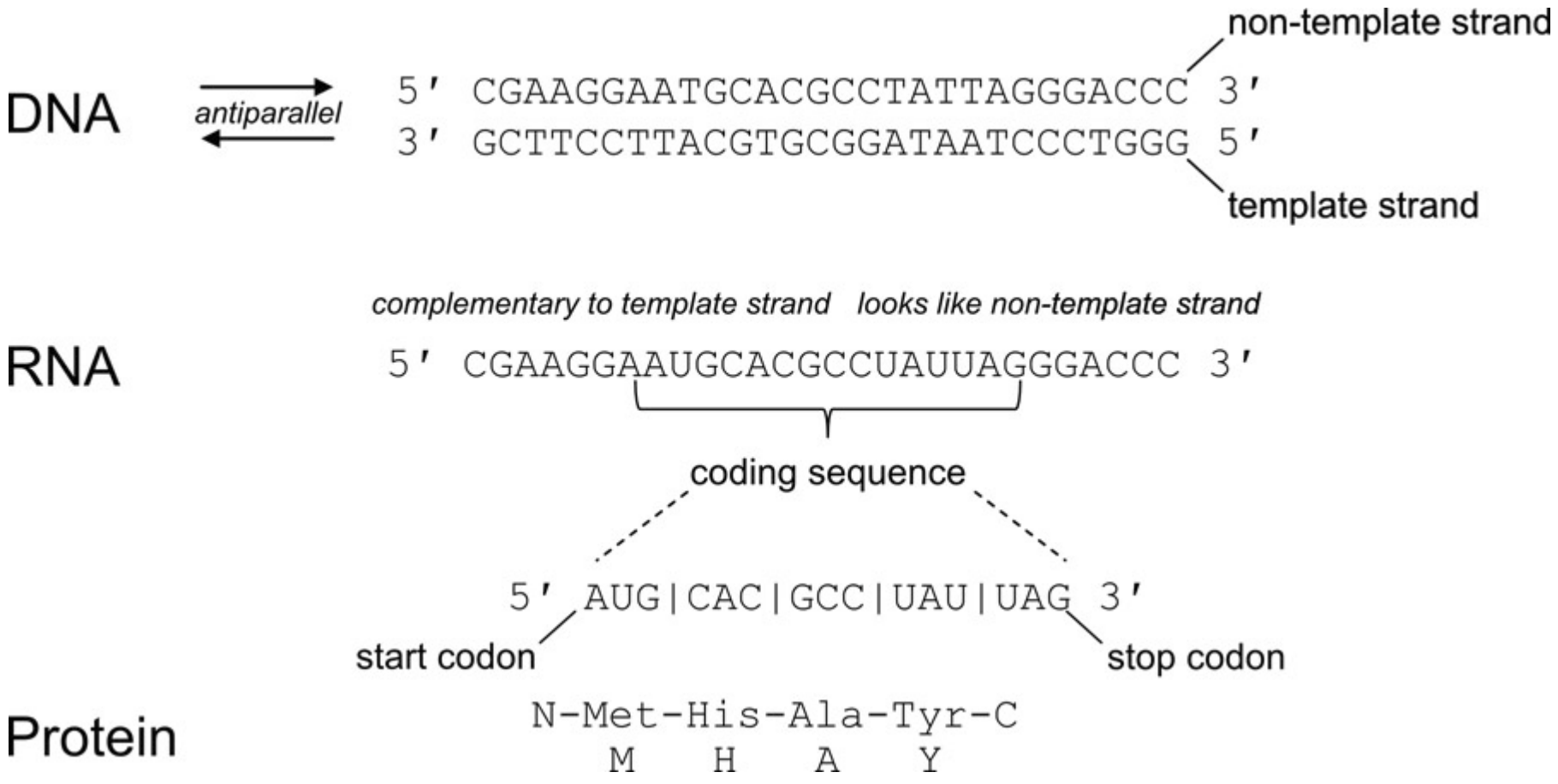
Sequence is Carrier?

- How do we determine if a sequence carries the CF allele?
- Get DNA sample
- Translate DNA to protein: Compare this seq to seq of a “working protein”
- Is difference found between both proteins?



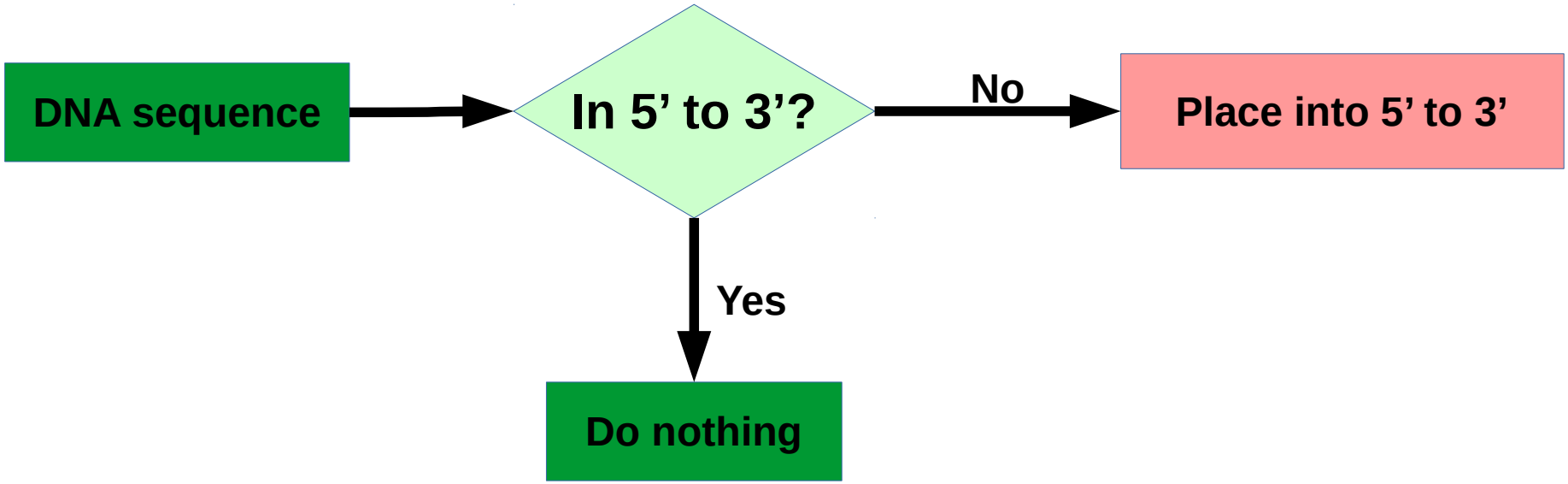
Analyze the Protein

- Translating DNA to find defects



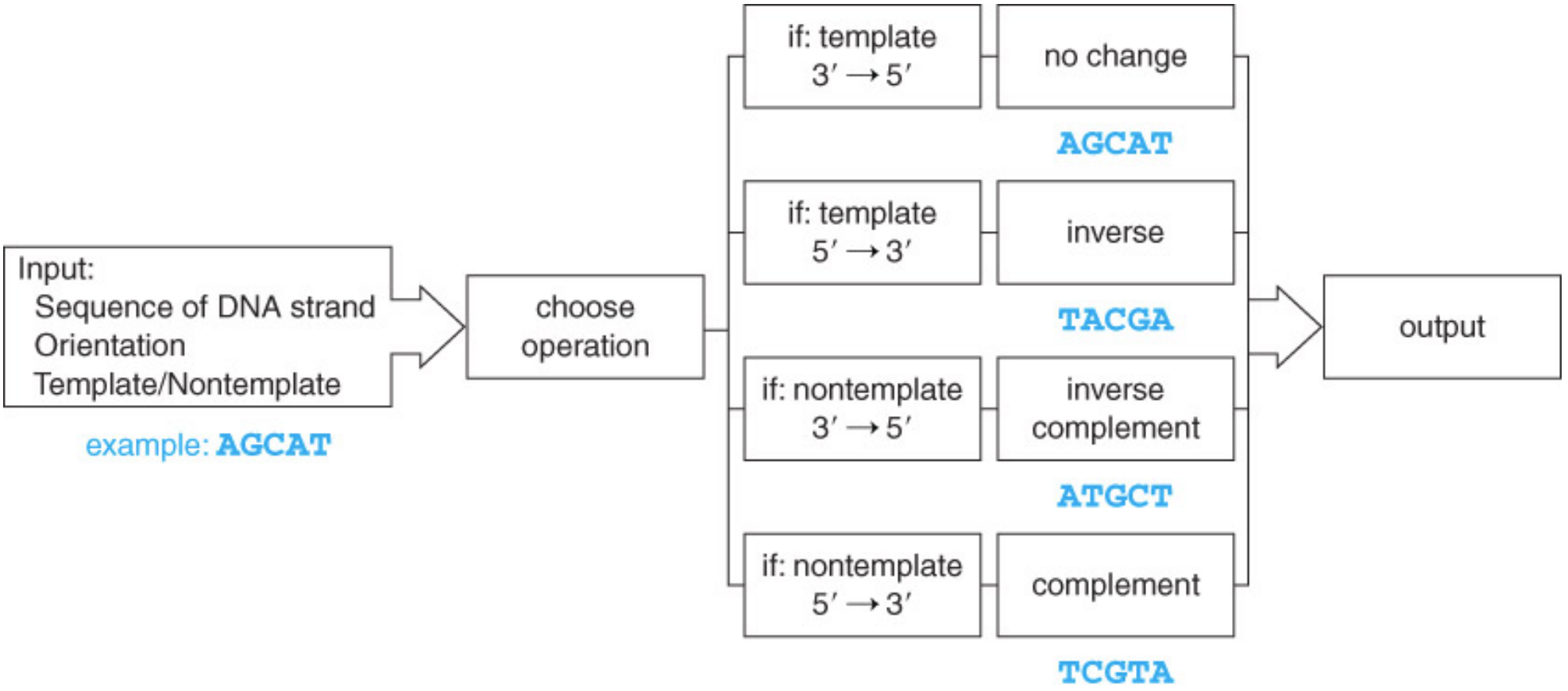
The Unnamed Sequence

- 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'

The DNA Manipulation Algorithm

1. Input a DNA sequence, including details of being a template or non-template strand as well as its orientation
2. Convert to all uppercase
3. Choose the appropriate operation:
    1. If it is the template strand and oriented 3’ -> 5’, simply output the same sequence
    2. If it is the template strand and oriented 5’ -> 3’, **inverse** the sequence (traverse the string from right to left and add each character to output the string)
    3. If it is the non-template strand and oriented 3’ -> 5’, generate the **inverse complement** sequence ((i.) traverse the string from right to left and (ii) for each character, add the complement to the output string)
    4. If it is the non-template strand and oriented 5’ -> 3’, generate the **complement** ((i.) traverse the string from left to right and (ii) for each character add the complement to the output string)
4. Output the completed sequence, including 5’ and 3’ end labels

Transcription Algorithm

- Input: **template** strand in the **3’ → 5’** orientation
- Output: mRNA strand in the 5’ → 3’ orientation
  - Traverse the string from left to right
    - add complementary base to the output string
    - (note T is now U)

Alternative Transcription Algorithm

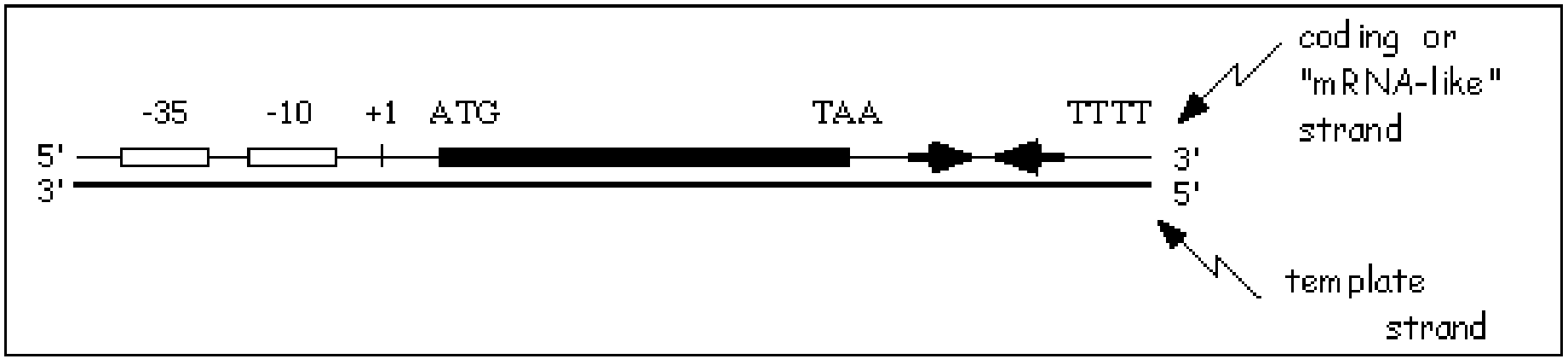
- Input: **non-template** strand in the **5’ → 3’** orientation
- Output: mRNA strand in the **5’ → 3’** orientation
  - Traverse the string from left to right
    - Replace all the T’s with U’s

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.

Template vs nonTemplate

- Input:
  - DNA sequence – AGCAT
  - Strand – template (used to make mRNA) or non-template (the compliment of this strand that looks like mRNA)
  - Orientation – 3’ -> 5’ or 5’ -> 3’
- Output:
  - Template strand in 3’ -> 5’ orientation ready for transcription



Codon Table to Translate the Protein Product

- DNA triplets read in groups of three called codons and represent an amino acid

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

Python Programming

ALLEGHENY COLLEGE

- Biopython
- Translation functions
  - DNA → RNA
  - RNA → DNA
  - RNA → Protein
- Gives a protein sequence to compare to the wild type protein sequence

Follow along in class and save your notes in a text file!!

python