Lecture 3
Protein Structure and Function, Carbohydrates, Nucleic Acids

- Protein Structure and Stability
- Ligand Binding
- Proteins as enzymes (PKU disease)
- Carbohydrates
- Nucleic Acid Technologies

Summary - Interactions that Stabilize Folded Proteins.

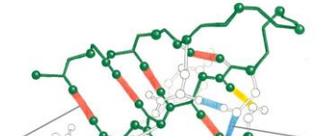
H-bonds
van der Waals
Hydrophobic effect

Chain disorder

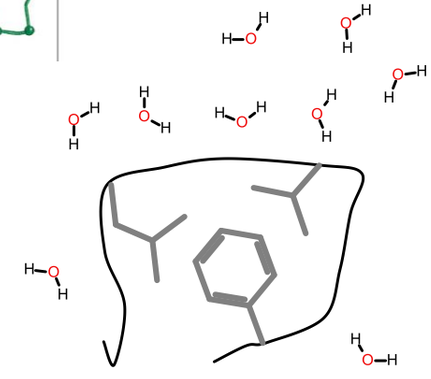
- **Hydrogen bonds** form between hydrogen atoms (NH) and the carbonyl group in the peptide backbone (mainchain), and between and donors and acceptors on sidechains. *Mainchain-mainchain H-bonds are responsible for secondary structures.*
- **Hydrophobic interactions** within a protein increase stability of the folded state by *increasing entropy due to the release of water that was ordered by the exposed non-polar groups in the unfolded protein.*
- **van der Waals interactions** are *optimized in the well packed core of the protein*.

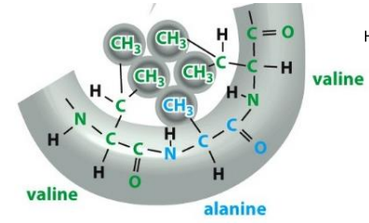
- **Covalent disulfide bonds** form between sulfur-containing cysteine residues stabilizing them (usually only exported, secreted proteins).

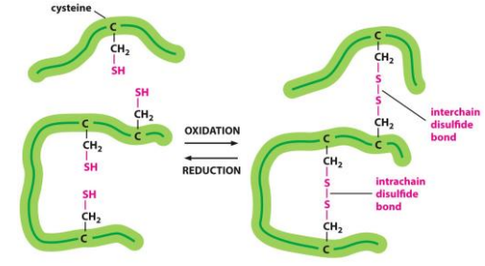
Native



Unfolded







A single change in the amino acid sequence can change the function of a protein, and often affecting how it folds – Producing Inactive Proteins.

frontiers in Molecular Biosciences

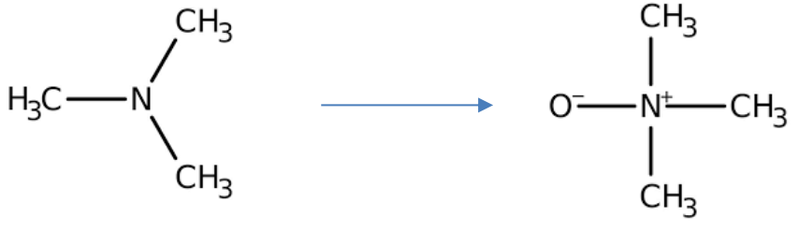
ORIGINAL RESEARCH
published: 07 January 2021
doi: 10.3389/fmolb.2020.626363

Solvent Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein Structures to Protein Sequences

Castrense Savojardo [1], Matteo Manfredi [1], Pier Luigi Martelli [1] and Rita Casadio [1,2]*

[1] Biocomputing Group, Department of Pharmacy and Biotechnologies, University of Bologna, Bologna, Italy, [2] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies of the National Research Council, Bari, Italy

Mutations in dimethylaniline monooxygenase 3 cause trimethylaminuria (high levels of trimethylamine)



dimethylaniline monooxygenase 3

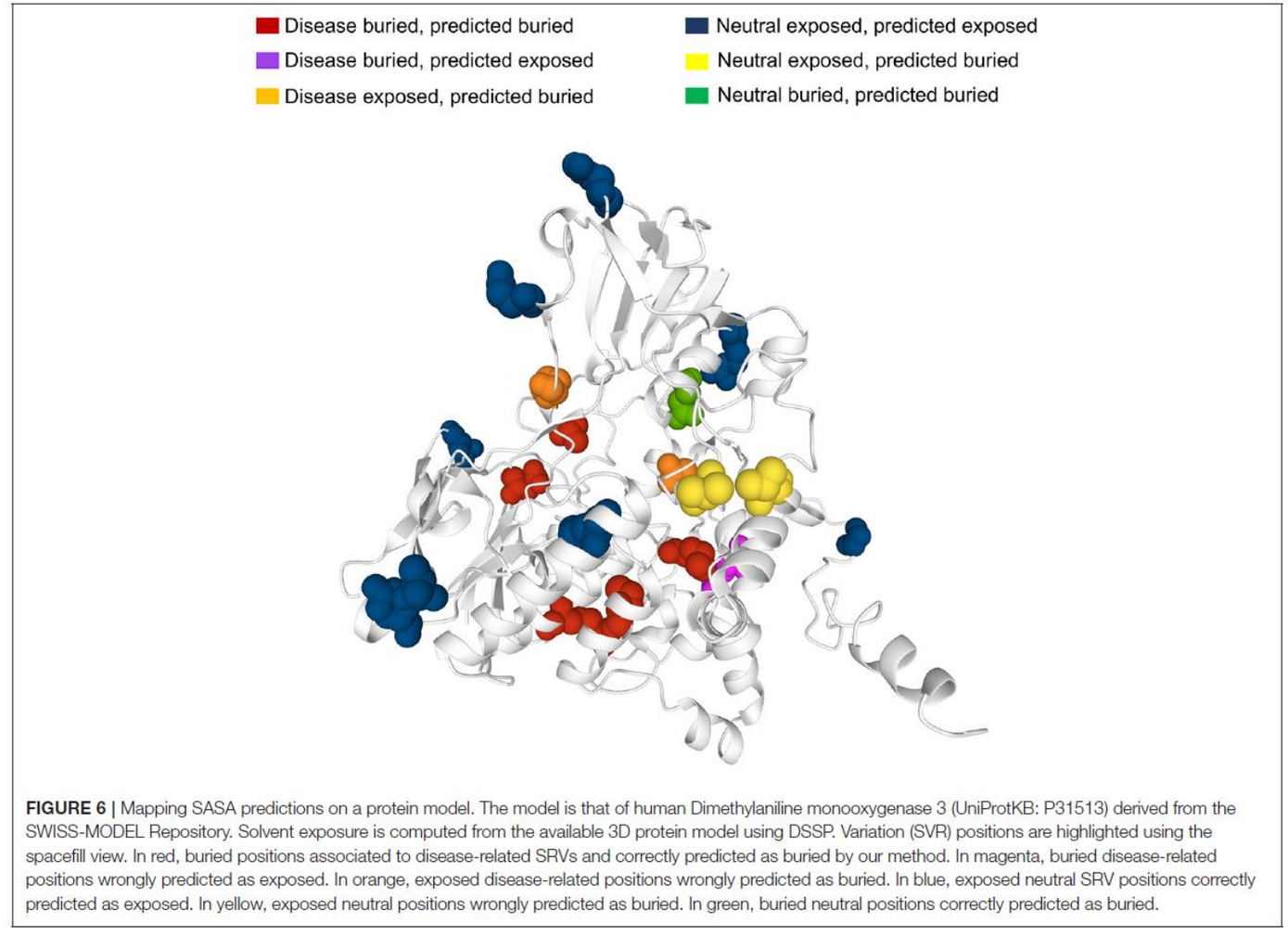
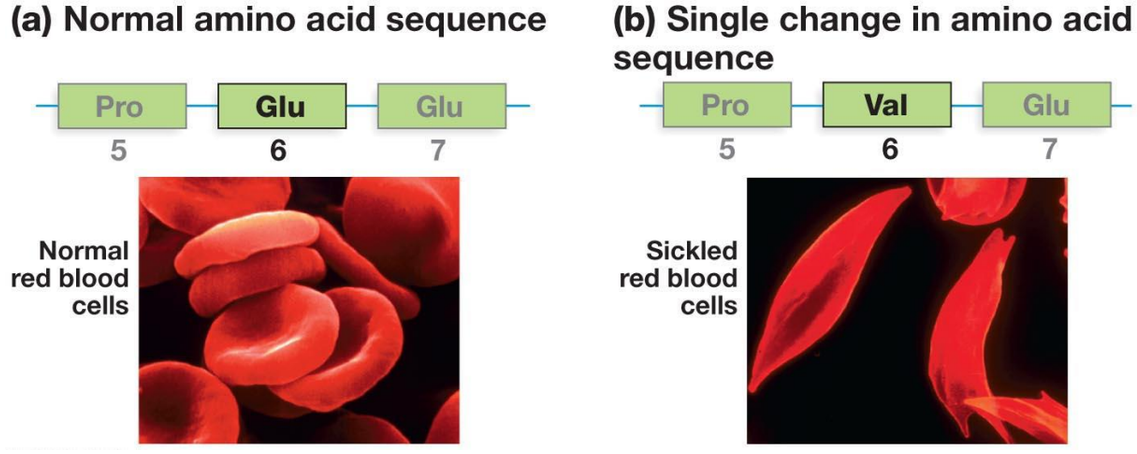


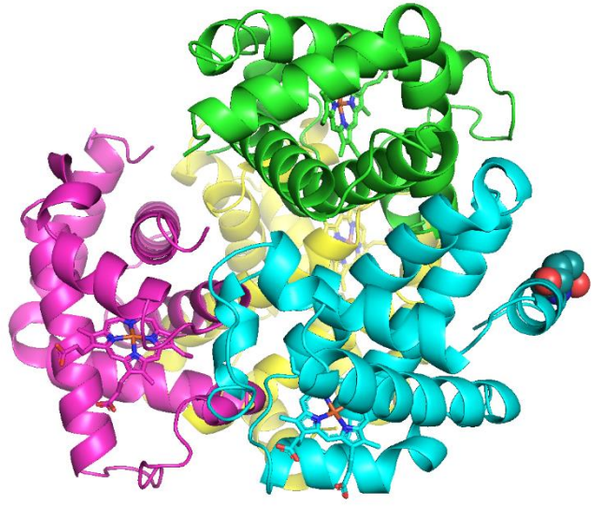
FIGURE 6 | Mapping SASA predictions on a protein model. The model is that of human Dimethylaniline monooxygenase 3 (UniProtKB: P31513) derived from the SWISS-MODEL Repository. Solvent exposure is computed from the available 3D protein model using DSSP. Variation (SVR) positions are highlighted using the spacefill view. In red, buried positions associated to disease-related SRVs and correctly predicted as buried by our method. In magenta, buried disease-related positions wrongly predicted as exposed. In orange, exposed disease-related positions wrongly predicted as buried. In blue, exposed neutral SRV positions correctly predicted as exposed. In yellow, exposed neutral positions wrongly predicted as buried. In green, buried neutral positions correctly predicted as buried.

Surface Mutations May Also Lead to Disease

Effect of mutations on protein folding – sickle cell anemia



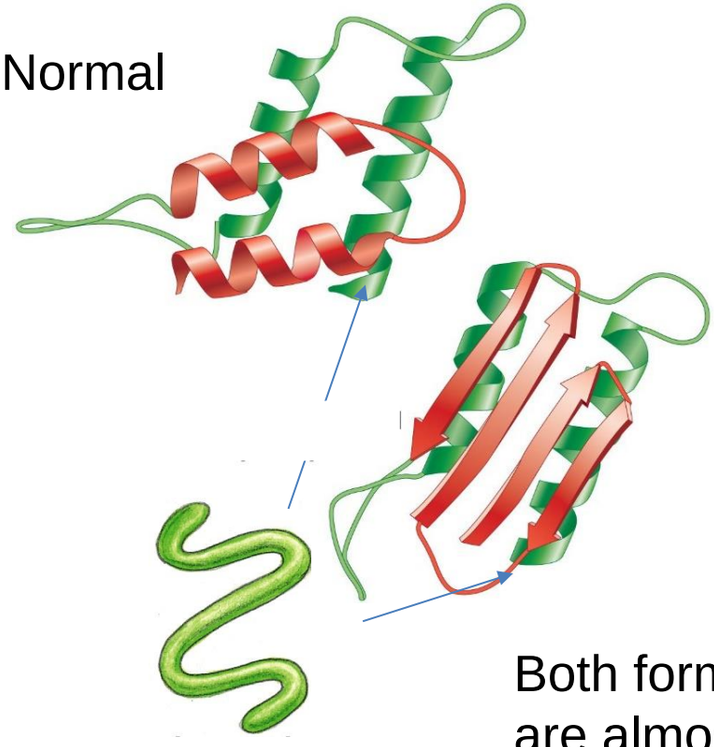
A single change in the amino acid sequence can change the function of a protein



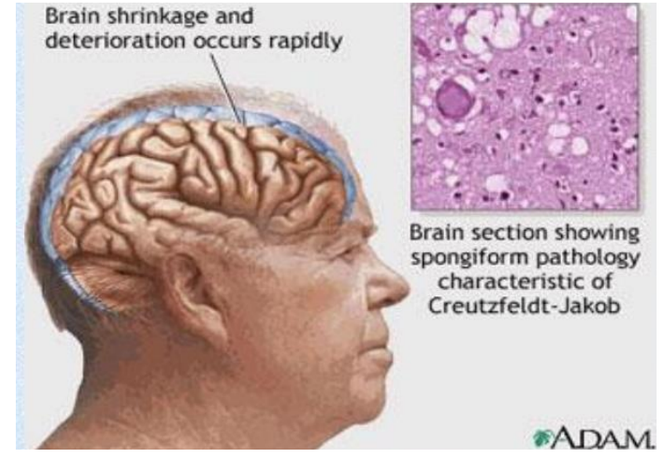
Glu6

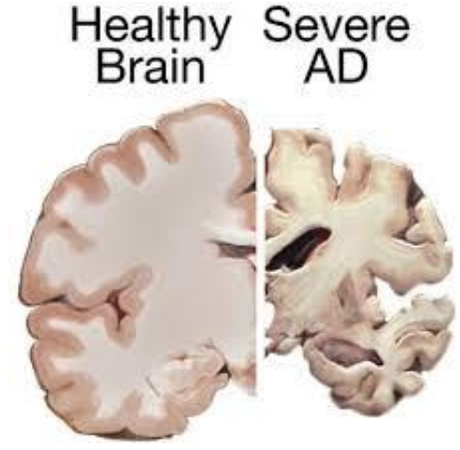
What Happens When Proteins Fold Into Different Structures?

Prions are improperly folded proteins that cause neurodegenerative diseases



Both forms are almost equally stable.





What is the effect on the brain?

Unfolded protein response (UPR): The presence of unfolded proteins can trigger the UPR, which can turn off protein synthesis in the cell, leading to cell death.

Why do the brain cells die?

Protein Structure - Summary and Expectations

Primary Structure:

- Can you describe the mechanism of peptide bond formation
- Can you draw structure of peptides.
- Can you identify amino terminus and give the sequence of amino acids, N -> C

Secondary structure:

- Identify helical and sheet secondary structures,
- know that they are stabilized by **mainchain** hydrogen bonds between N-H and O=C.
- Location of H-bonds and sidechains

Tertiary Structure:

- Can you describe and identify role of the following in stabilizing the folded state.
  - H-bonds,
  - van der Waals,
  - hydrophobic effect
- Can you predict, based on sidechain, which amino acids are found in the core of the protein and which are found on the surface.

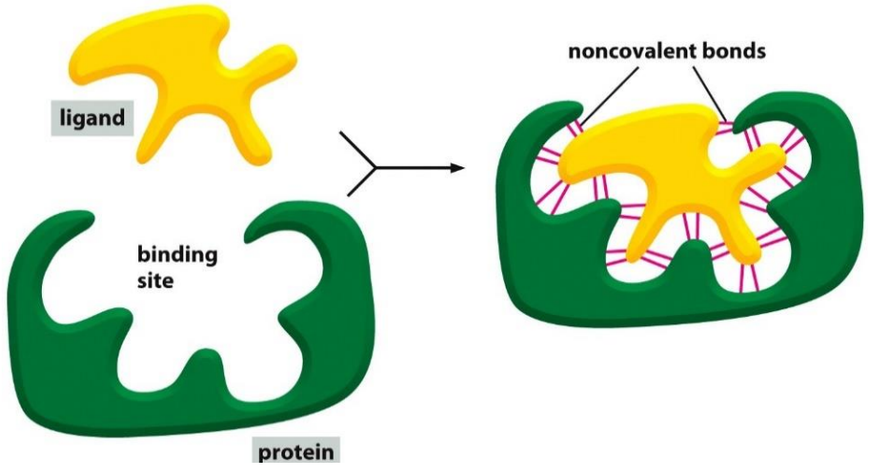
Quaternary Structure:

- Multiple chains, stabilized by non-covalent and covalent (disulfide bonds) interactions.

Diseases related to protein structure:

- Core mutations - affect folding
- Surface mutations - affect protein-protein (and Protein-DNA) interactions
- Stable isoforms – toxic to the cell

Ligand Binding: Most Proteins Bind to Other Molecules in Biological Interactions:



Ligand: Something that binds to a protein, usually small molecules (e.g. cyclicAMP, cAMP).

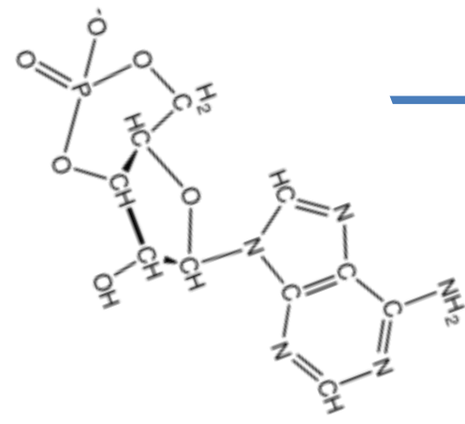
Binding site allow a protein to interact with specific **ligands**

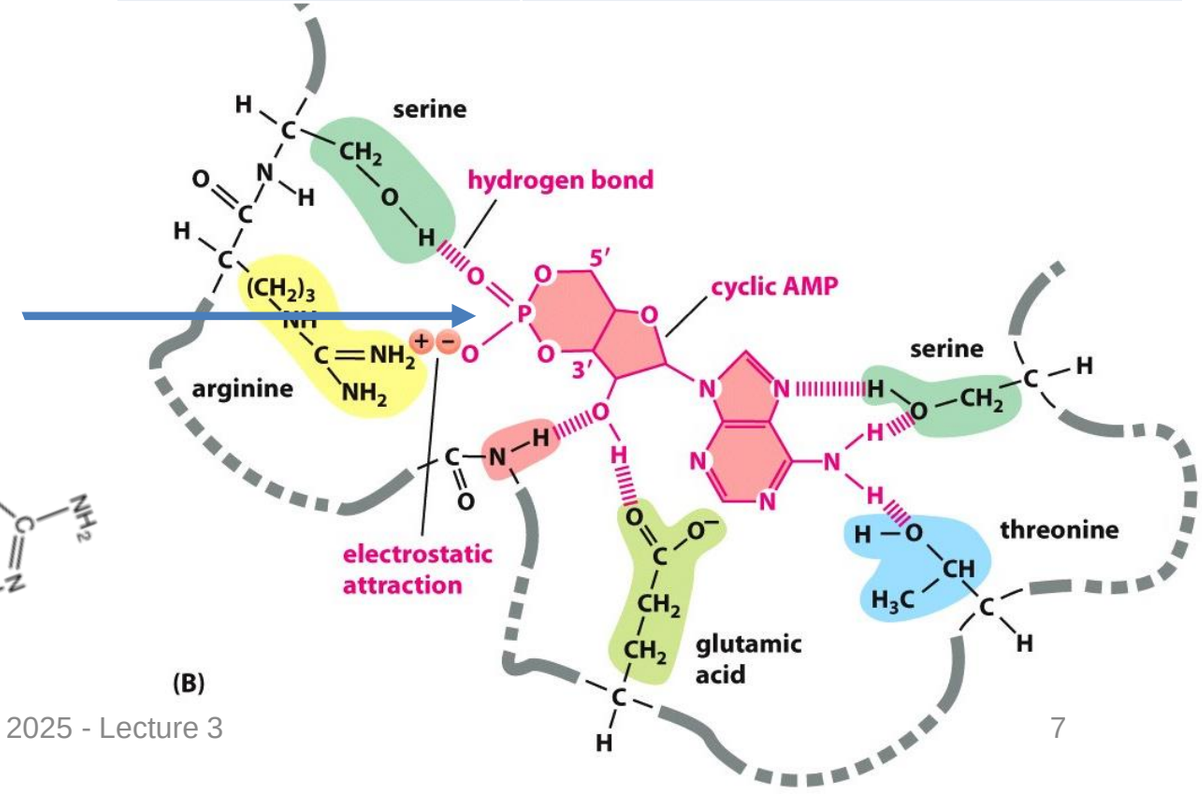
Binding site is generated by the **folded** form of the protein.

The bound ligand can be stabilized by any and all of the following interactions:

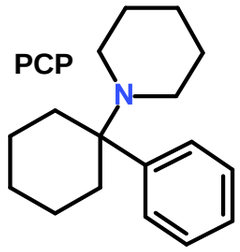
Interaction	Which stabilize cAMP Binding?
Electrostatic	
van der Waals	
H-Bonding	
Hydrophobic effect	

cAMP

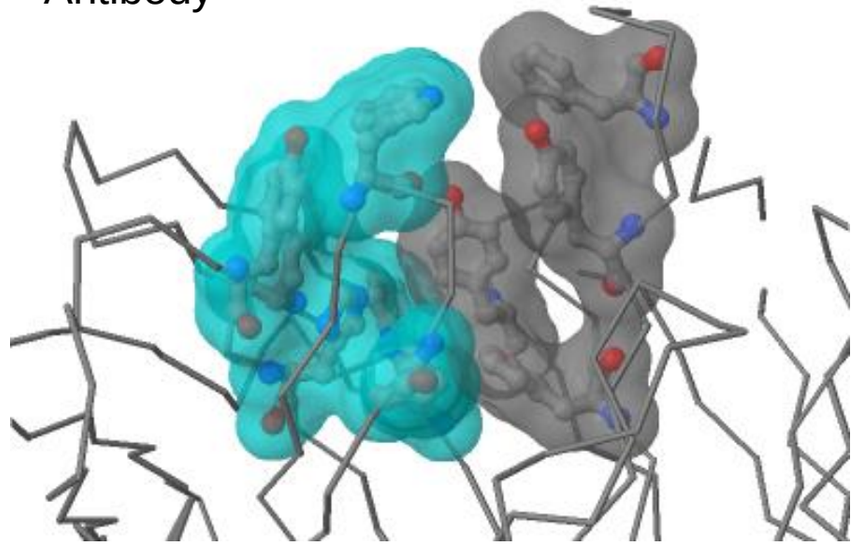




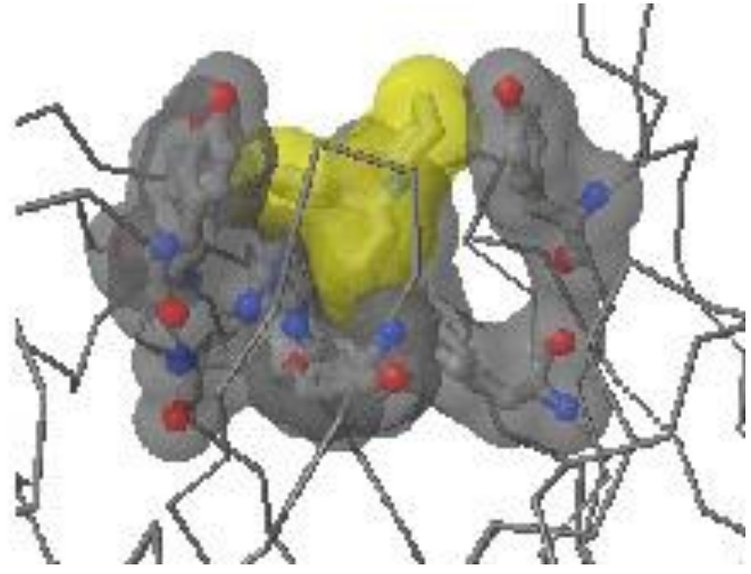
Hydrophobic Effect Drives Binding of Non-polar Ligands

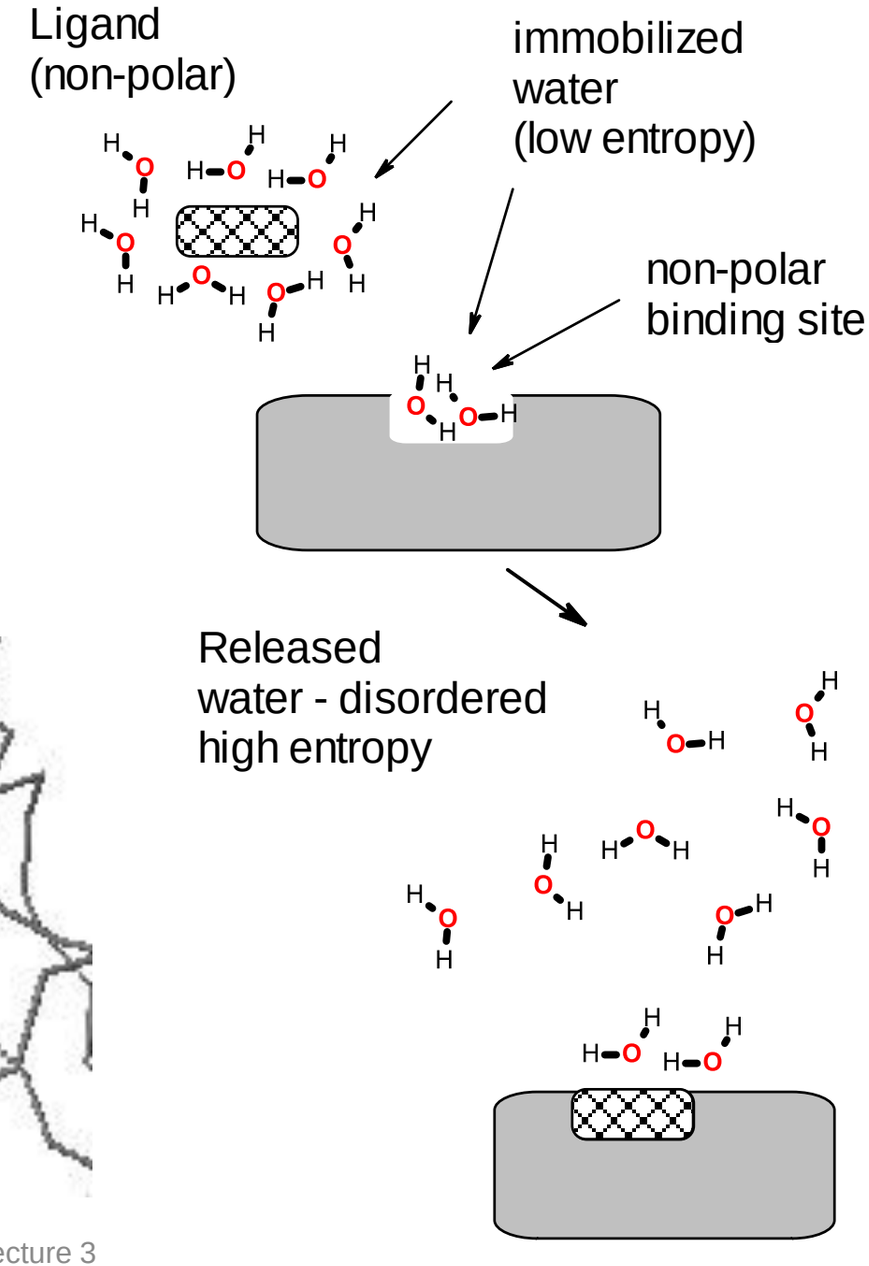


Antibody



Antibody-PCP complex

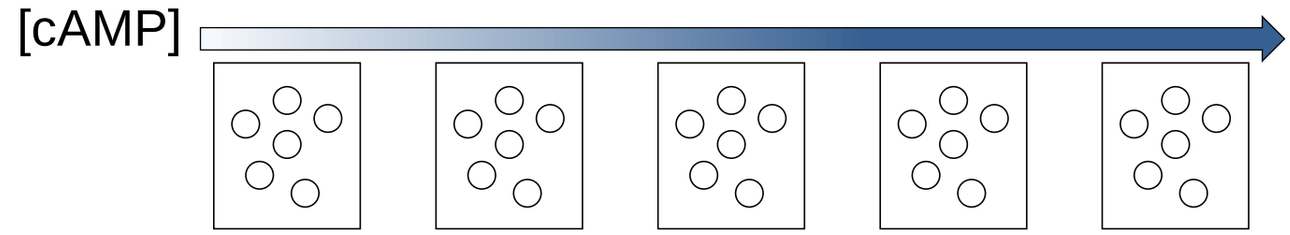


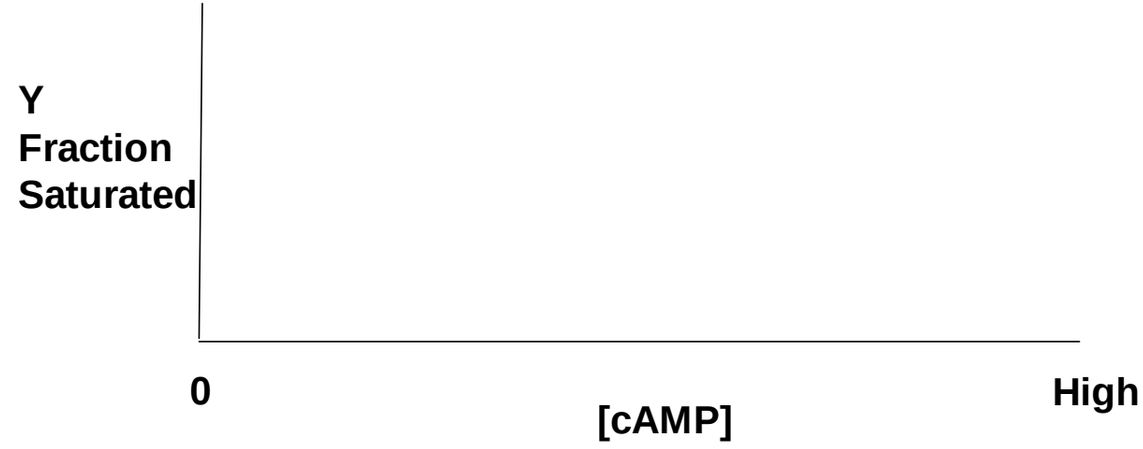


Ligand Binding & Saturation:

Define fraction saturated: $Y = \frac{[ML]}{[M] + [ML]}$

[M] = free macromolecule (e.g. antibody with no antigen).
[ML] = macromolecule with ligand bound (e.g. antibody with antigen bound).





The boxes with circles represent proteins with no cAMP bound, each box (left to right) is at a higher [cAMP]. Filled circles indicate bound ligand.

1. How will the number of filled circles depend on the cAMP concentration?

2. Plot the fraction saturated data point for each box.

Key Points:

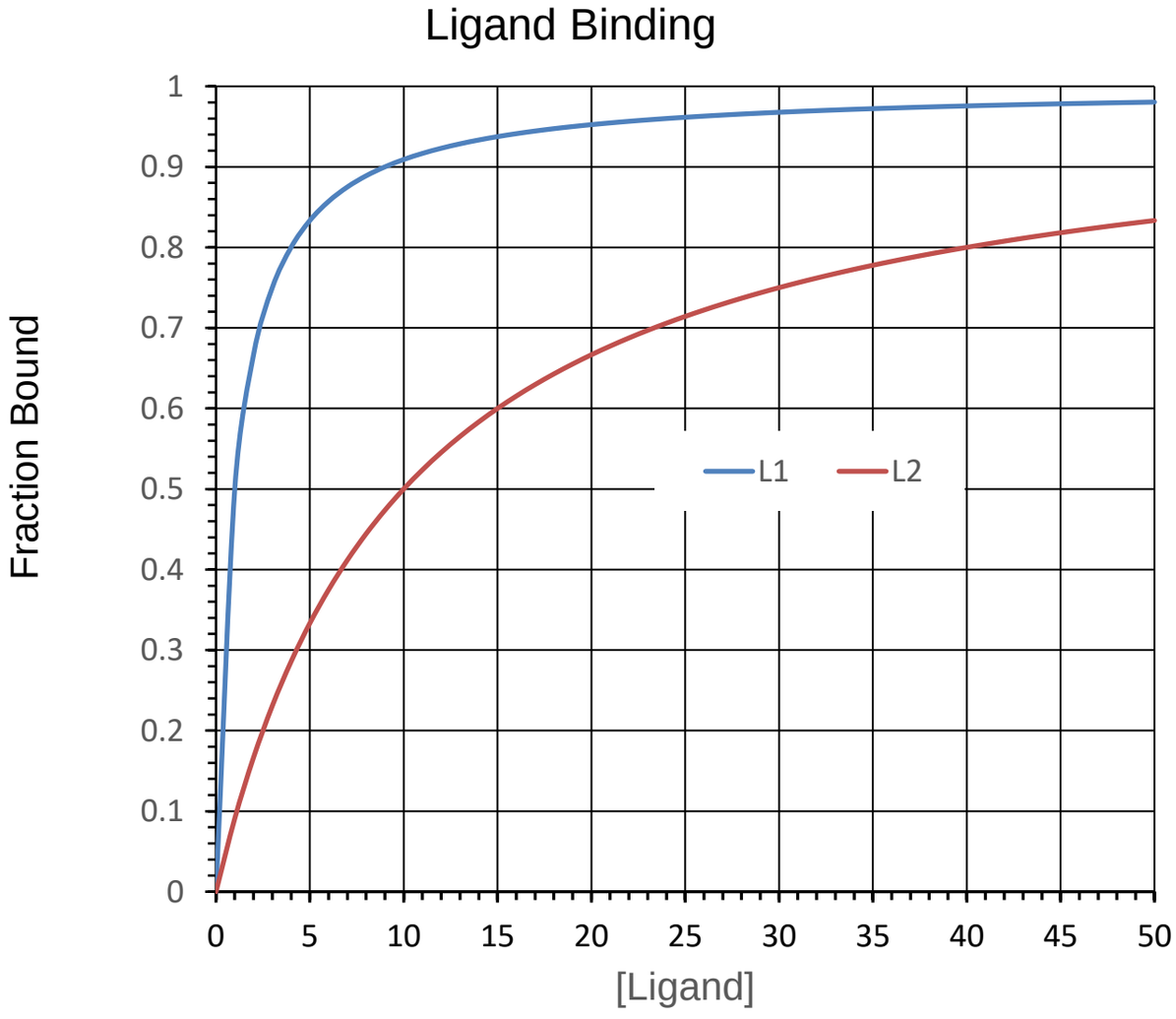
1. The binding sites saturate, when all are full no more ligand can bind.
2. There is a ligand concentration, [L], where ½ the sites are full. This [L] is K_D
3. K_D is the equilibrium constant for ligand dissociation:

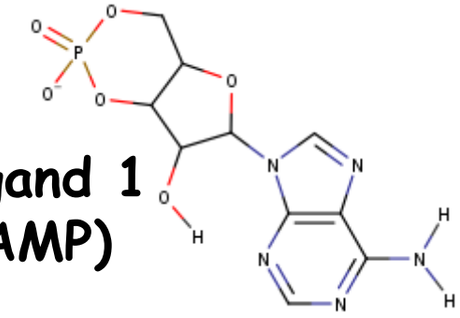
$$K_{Eq} = \frac{[products]}{[reactants]}$$

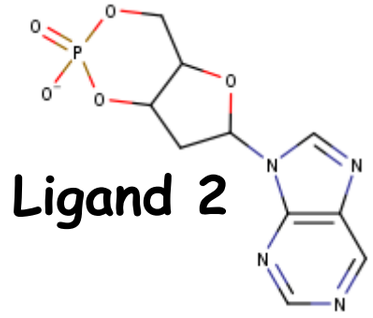
$$(ML) \rightleftharpoons (M) + (L)$$

$$K_D = \frac{[M][L]}{[ML]}$$

Using K_D to Compare Ligand Binding







The binding of two different molecules to the same protein was measured and the data is shown on the right. L1 is cAMP, L2 is similar to cAMP

Which ligand has a K_D of 1? L1 or L2?

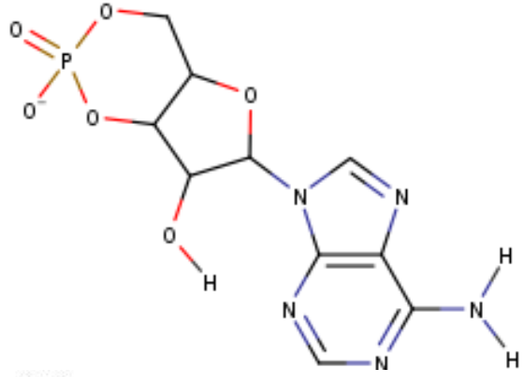
Which ligand has a K_D of 10? L1 or L2?

Which ligand binds more tightly to the protein (higher affinity)? L1 or L2?

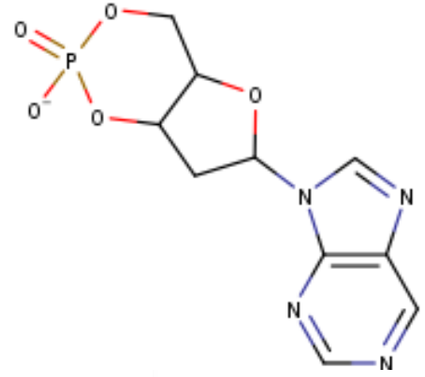
__Why__ does L1 bind more tightly (higher affinity)?

1. What are the chemical differences between L1 and L2 (Upper diagram)

Ligand 1 (cAMP)

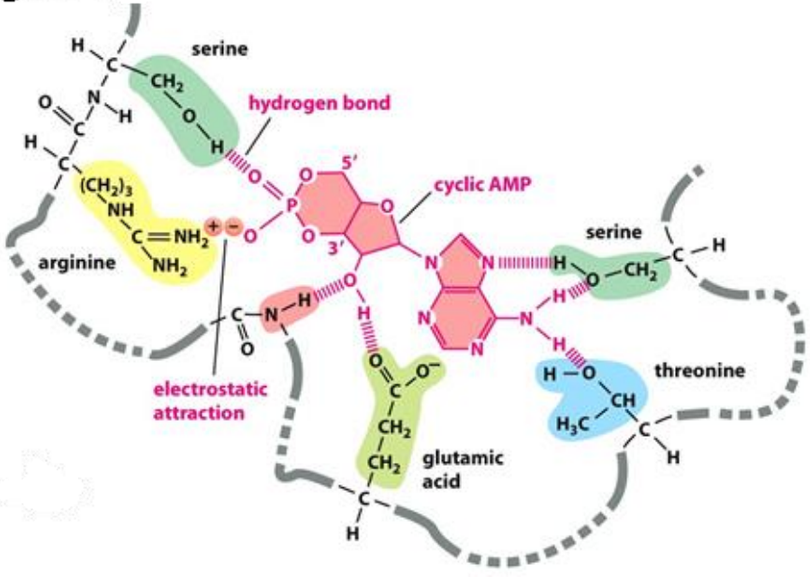


Ligand 2

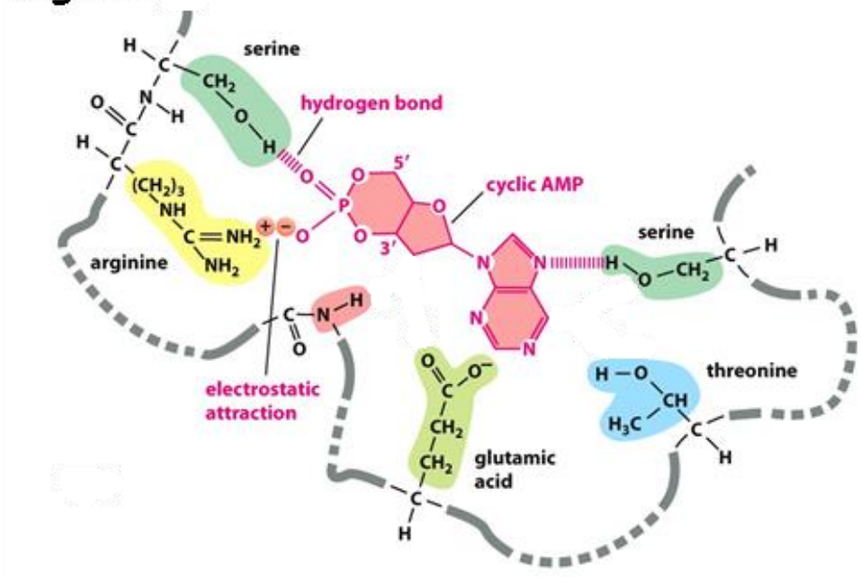


2. How do these differences affect the interactions with the protein (lower diagram)?

Ligand 1



Ligand 2



3. How do the differences affect K_D?

Key Points:

Binding:

Folded proteins have ***binding sites*** that recognize other molecules (***ligands***) using ***any and all*** of the following:

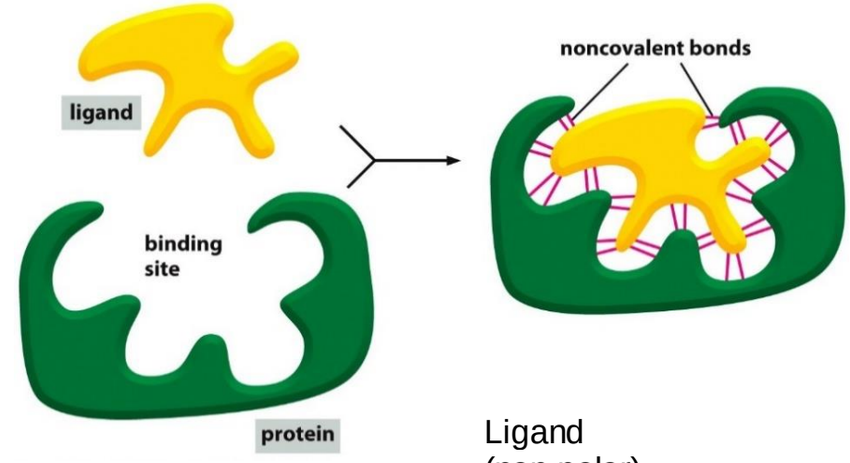
- H-bonds,
- van der Waals,
- Electrostatic,
- Non-polar interactions (hydrophobic)

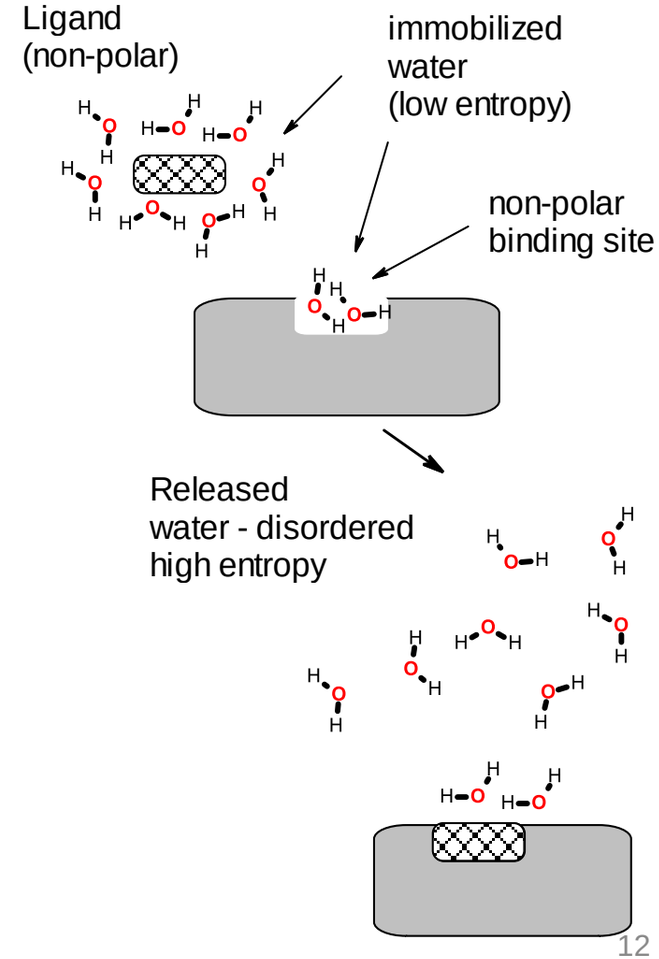
Binding is **reversible**

Binding is **saturable**

Binding ½ point (Y=0.5) occurs at K_D

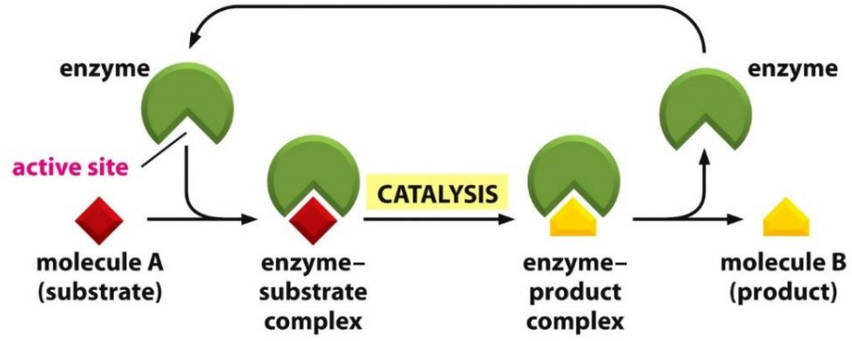
The higher the affinity (strength of interaction), the lower the K_D

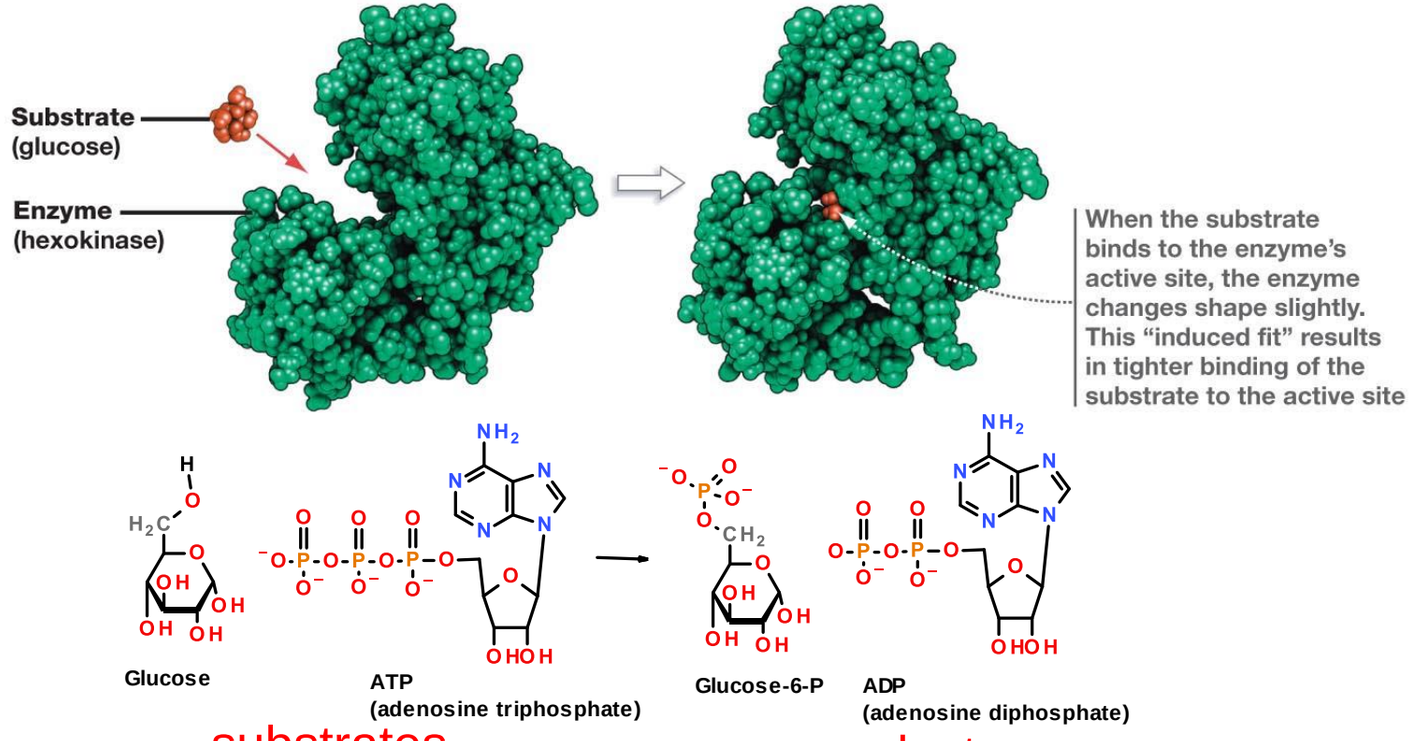




Enzymes

- **Enzymes** are protein or RNA catalysts. They increase the rate of the reaction.
- They bind “substrates” and convert them to “products”. Usually, the substrate undergoes a chemical reaction and is changed in its structure.
- Most biological chemical reactions occur at meaningful rates only in the presence of an enzyme.
- Substrates bind specifically to the enzyme’s **active site,** interacting with amino acid side chains (or RNA bases). Usually, an enzyme binds one substrate.
- The chemical change caused by the enzyme is catalyzed by additional functional groups in the active site.
- Many enzymes undergo a conformational change when the substrates are bound to the active site; this change is called an **induced fit**.





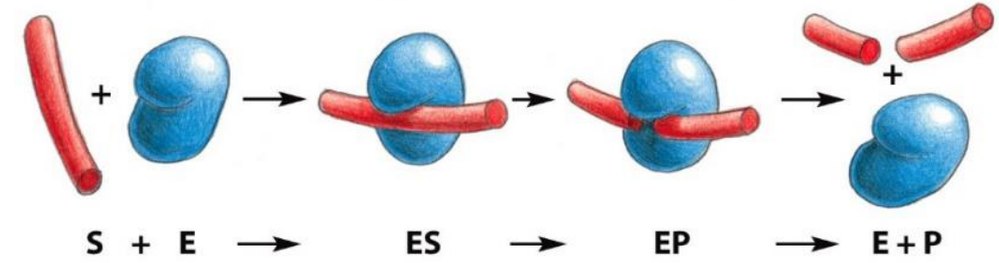
Enzyme – Chemical Diversity

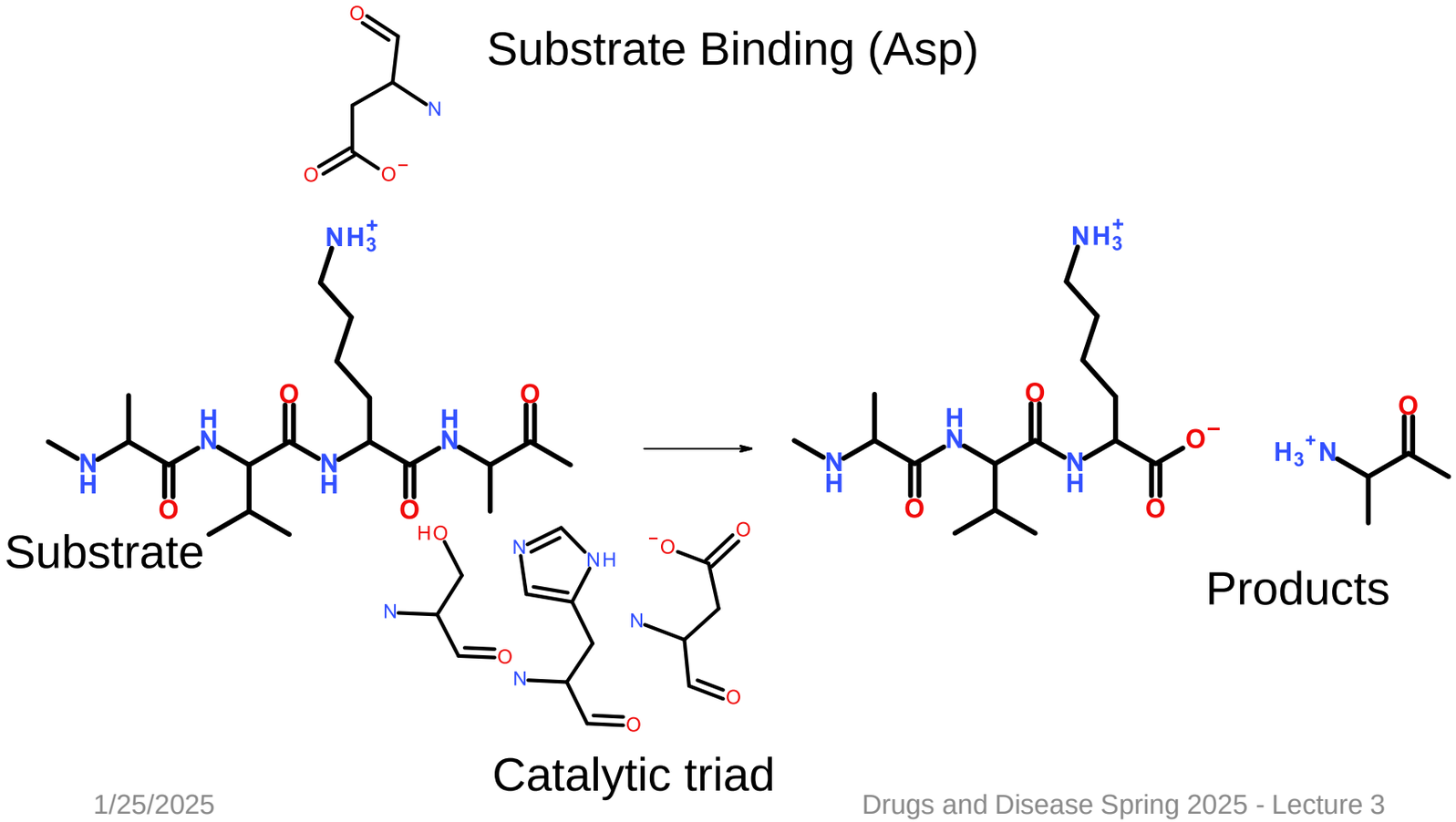
TABLE 4–1 SOME COMMON FUNCTIONAL CLASSES OF ENZYMES	
ENZYME CLASS	**BIOCHEMICAL FUNCTION**
Hydrolase	General term for enzymes that catalyze a hydrolytic cleavage reaction.
Nuclease	Breaks down nucleic acids by hydrolyzing bonds between nucleotides.
Protease	Breaks down proteins by hydrolyzing peptide bonds between amino acids.
Synthase	General name used for enzymes that synthesize molecules in anabolic reactions by condensing two molecules together.
Isomerase	Catalyzes the rearrangement of bonds within a single molecule.
Polymerase	Catalyzes polymerization reactions such as the synthesis of DNA and RNA.
Kinase	Catalyzes the addition of phosphate groups to molecules. Protein kinases are an important group of kinases that attach phosphate groups to proteins.
Phosphatase	Catalyzes the hydrolytic removal of a phosphate group from a molecule.
Oxido-reductase	General name for enzymes that catalyze reactions in which one molecule is oxidized while the other is reduced. Enzymes of this type are often called oxidases, reductases, or dehydrogenases.
ATPase	Hydrolyzes ATP. Many proteins with a wide range of roles have an energy-harnessing ATPase activity as part of their function, including motor proteins such as myosin and membrane transport proteins such as the sodium–potassium pump.

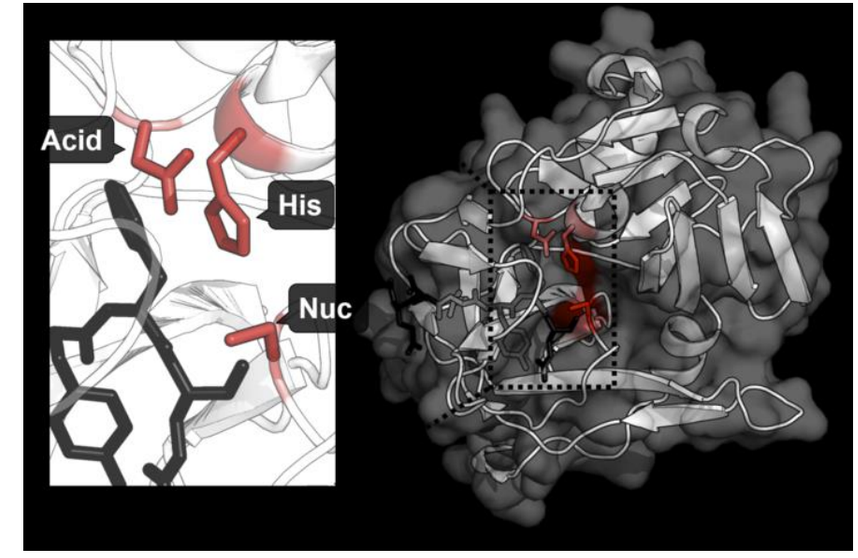
- Most enzyme names end in “-ase”
- Usually named by their substrates and the reactions they catalyse, i.e. glucose kinase

Example of Active Site Functional Groups:

- Catalytic triad (Asp, His, Ser) in Protease Trypsin cleaves the peptide bond.
- More active with Lys and Arg containing substrates because of a favorable interaction with an additional Asp residues in the enzyme.







https://shirleychemproject.weebly.com/

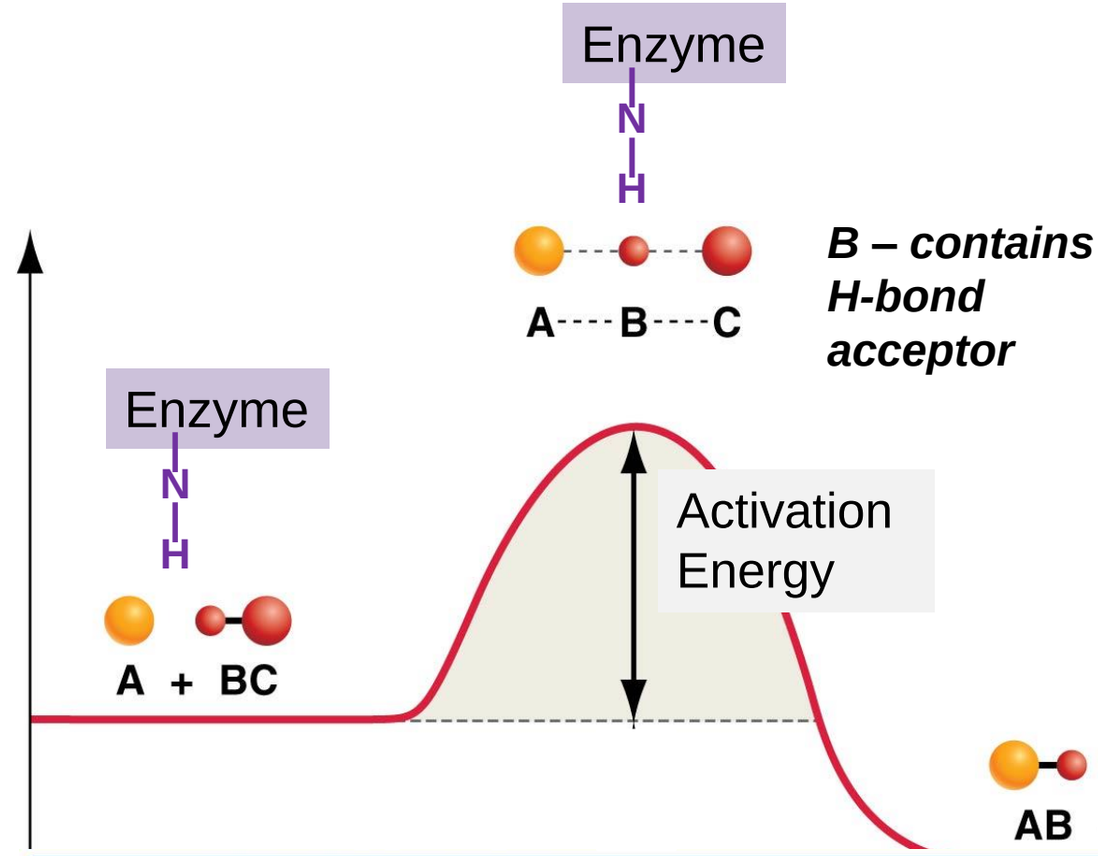
How Do Enzymes Increase Rates?

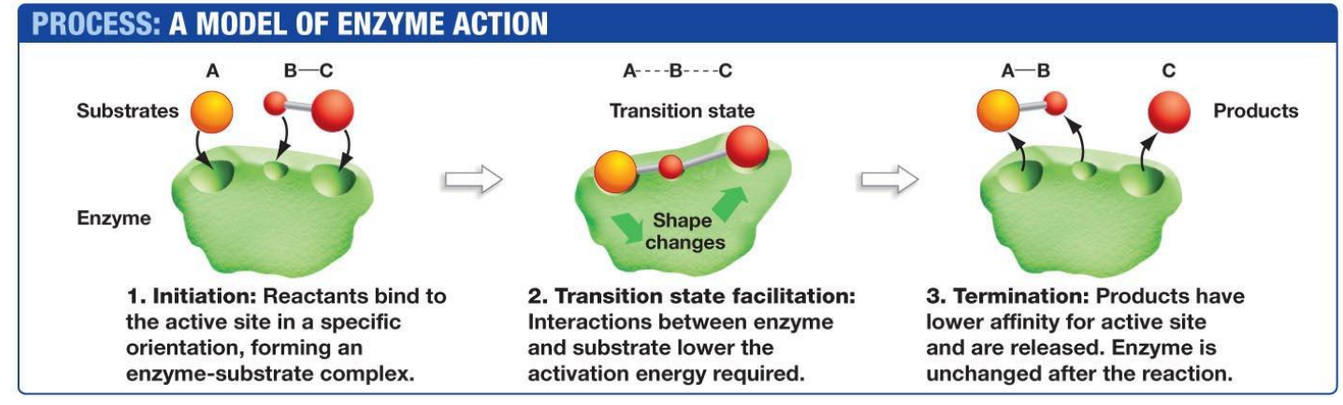
- **Transition state** = high energy intermediate that occurs during the reaction.
- Energy barrier is called the activation energy.
- Rate of product formation depends on the concentration of the transition state.

Low [X] = Slow reaction

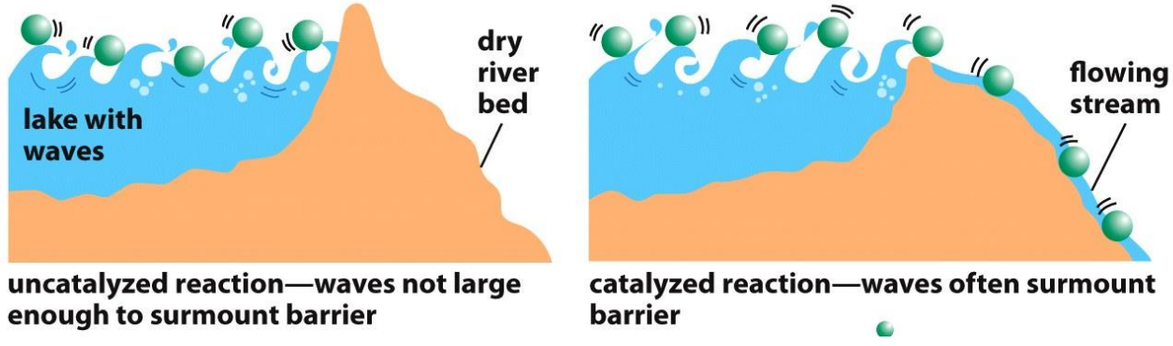
Higher [EX] = Faster reaction

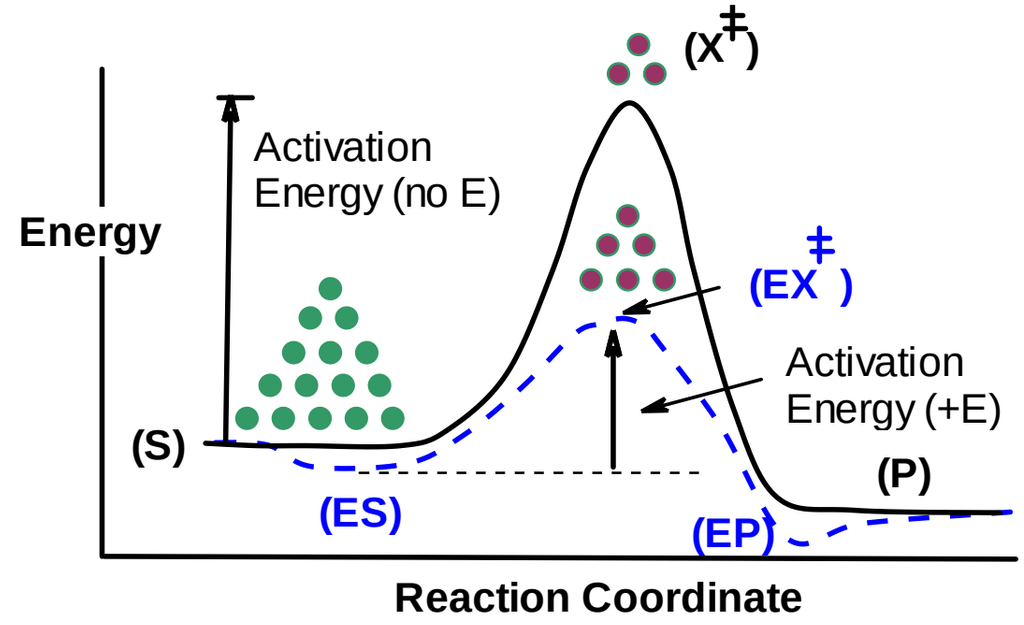
- Interactions between the enzyme and the substrate stabilize the **transition state** (X) and lower the activation energy required for the reaction to proceed.
- Stabilization can include:
  - Pre- alignment of key groups in the active site, reducing entropy cost of organizing groups.
  - Direct interactions with just the transition state, e.g. formation of new H-bonds.





A model of transition state stabilization.





Lower energy of transition state allows more substrates to reach transition state due to their thermal energy.

[S] = 15

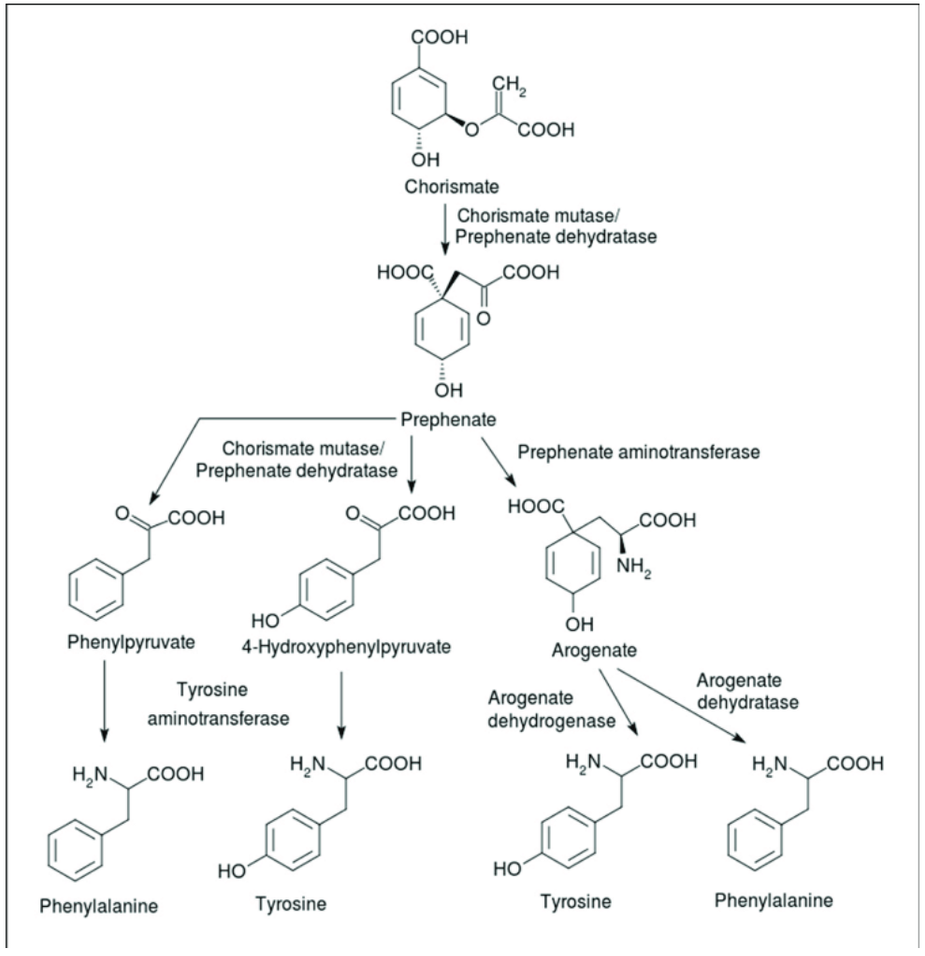
[X] = 3
[EX] = 6

How much faster will the rate be when the enzyme is present?

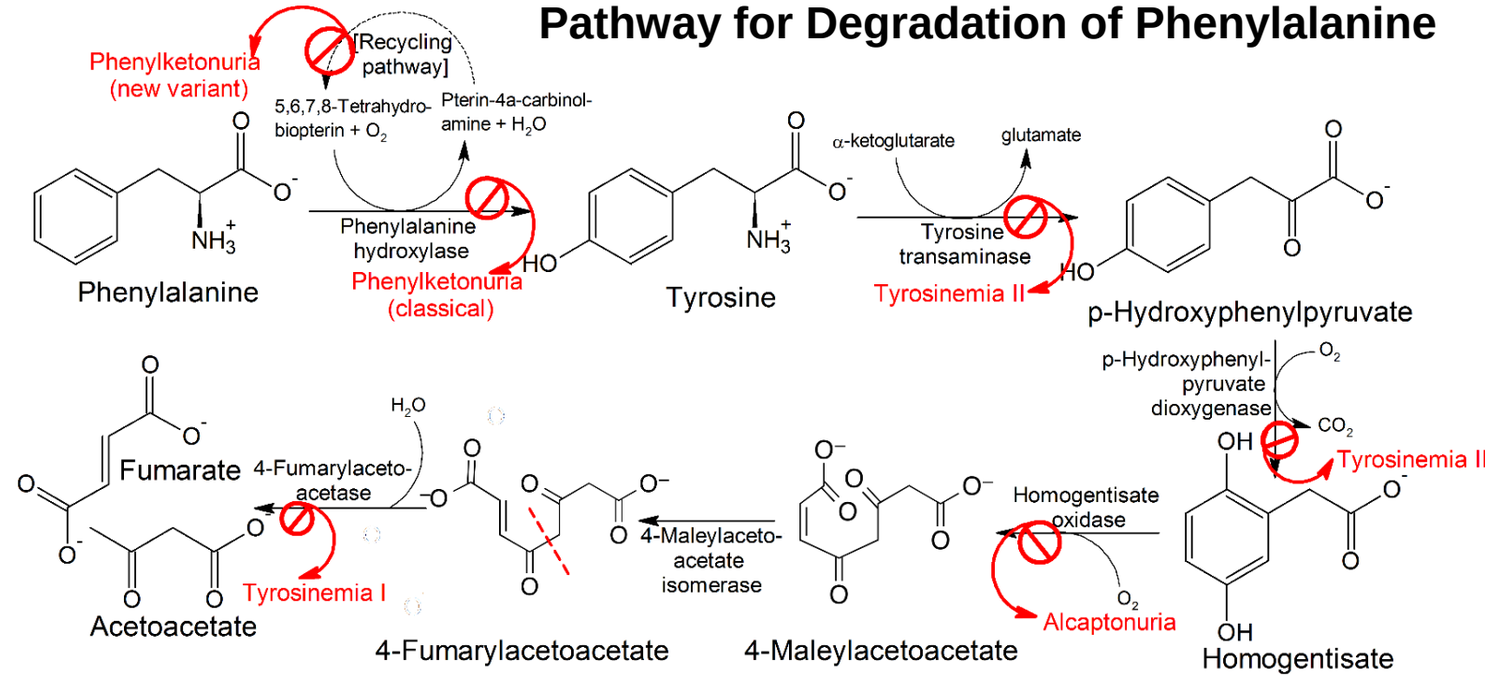
Enzymes, Metabolic Pathways, and Diseases

Synthetic Pathway for Phe, Tyr
(beginning with chorismite)

- Each step catalyzed by an enzyme



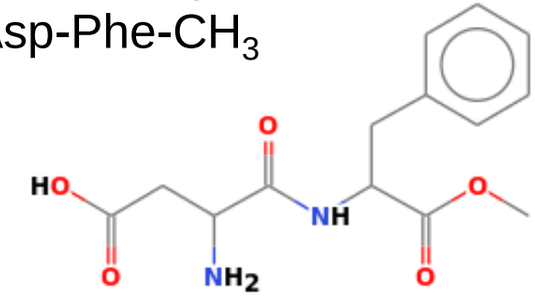
Pathway for Degradation of Phenylalanine



PKU Disease:

- Inactive phenylalanine hydroxylase
- Phe levels become toxic:
- Neurological problems
  - Intellectual disability
  - Developmental delays
  - Mental health disorders.

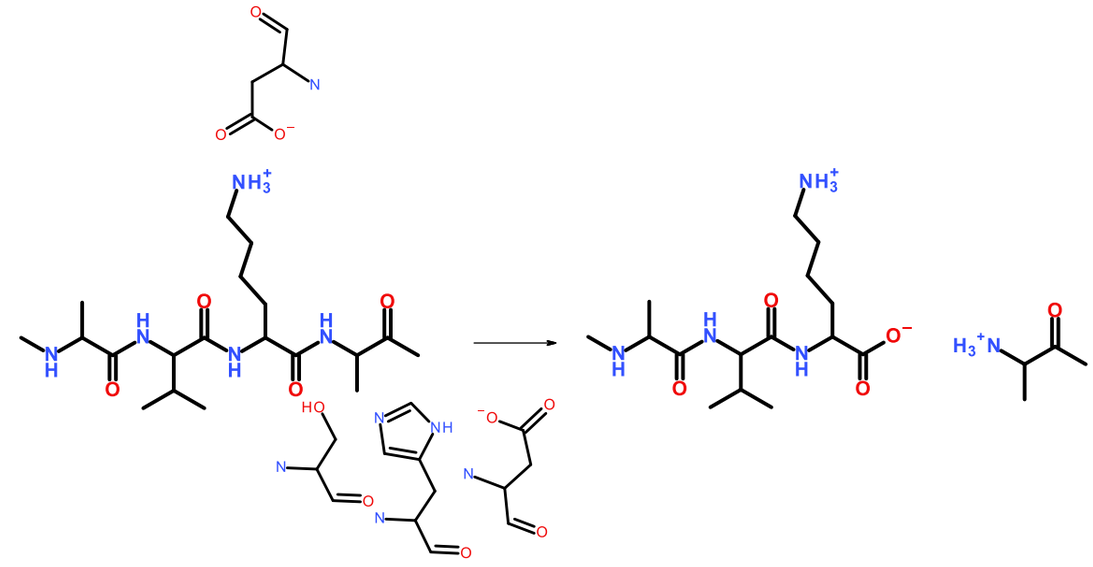
Aspartame (artificial sweetener)
Asp-Phe-CH_3

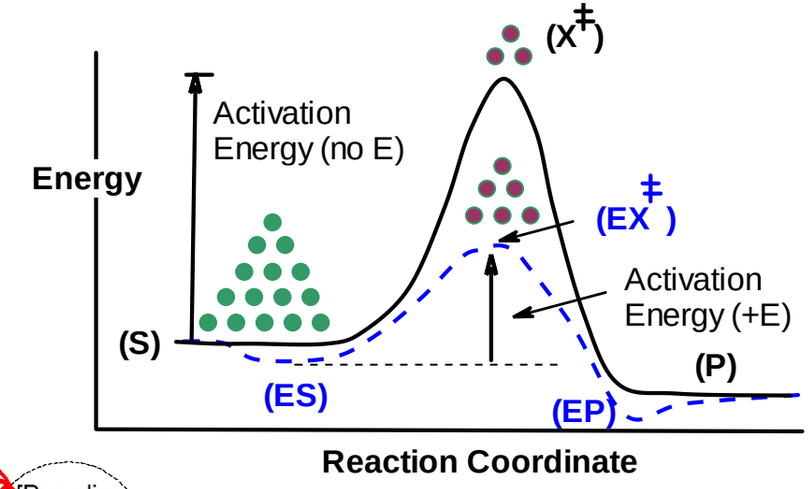


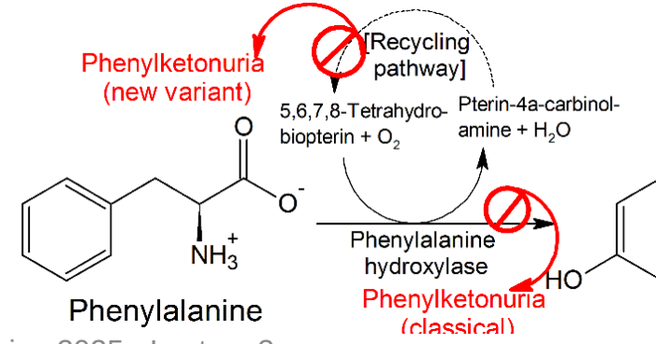
Key Points:

Enzymes:

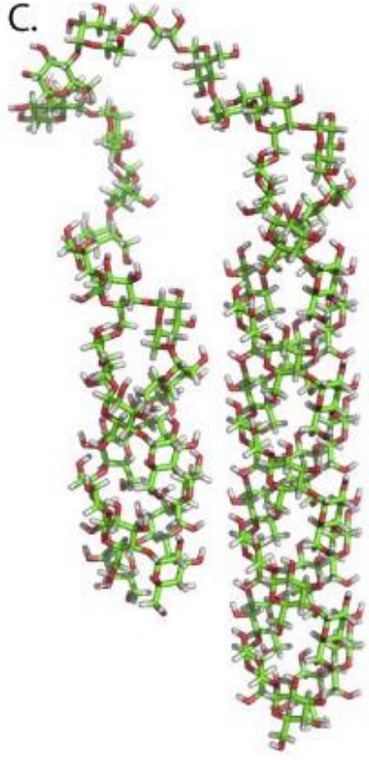
- Enzymes bind substrates (S), forming (ES) complex in active site, converting to P, releasing P.
- Rate enhancement since the transition state complex (EX) forms more readily with enzymes due to:
  - Bringing substrates and functional groups on the enzyme together by binding (less entropy change)
  - Directly lowering energy of transition state (X) through favorable interactions that are unique to the transition state, such as forming unique hydrogen bonds.
- Genetic diseases that lead to inactive metabolic enzymes can cause disease due to the build-up of toxic intermediates.



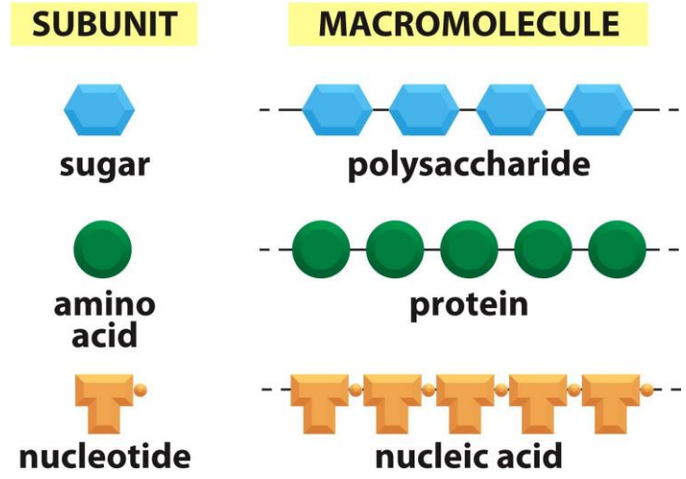




Carbohydrates



Polysaccharide



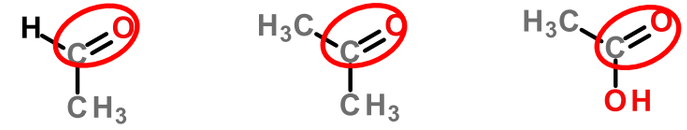
- General structure of monomers
- Disaccharides (e.g. lactose)
- Glycogen (glucose storage)
- Bacterial cell wall structure (antibiotic target)

- Lactose intolerance
- Glycogen storage disease

Carbohydrates

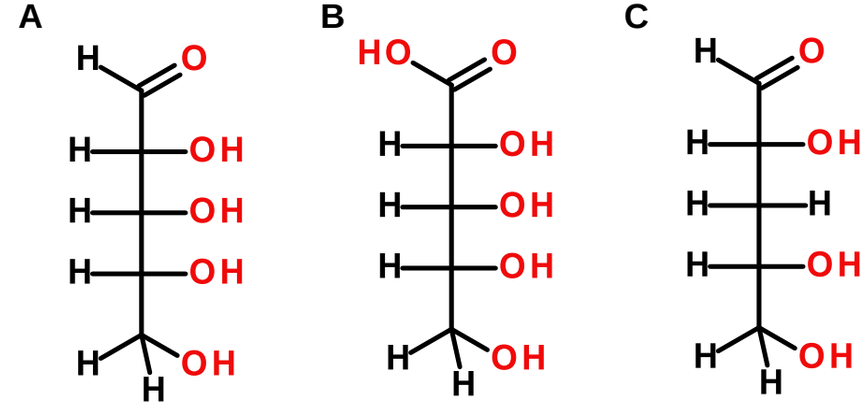
- Monosaccharides (one sugar),
- oligosaccharides (few sugars)
- polysaccharides (many sugars)
- Chemical formula is $(CH_2O)_n$ (e.g. hydrated carbon)
- They are molecules with:
  - one aldehyde or ketone group, on 1st or 2nd carbon
  - -OH group on *all* other carbons, leading to a chiral carbon for most carbons.

Functional groups:



aldehyde ketone carboxylic acid

Carbonyl group –> C=O



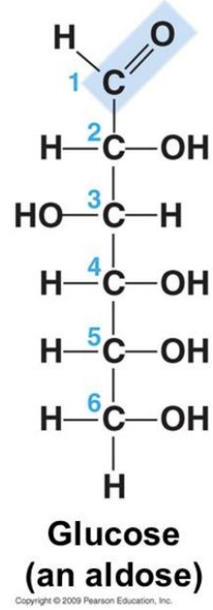
Only one of these is a carbohydrate, which one?

A B C

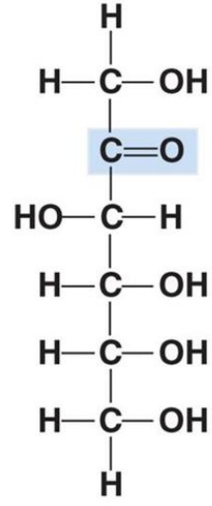
3 ways simple sugars (monosaccharides) differ from each other

1. Location of the carbonyl group
2. Number of carbons
3. Spatial arrangement of atoms (the position of the OH groups)

Aldose: Carbonyl group is located on C_1



Glucose (an aldose)

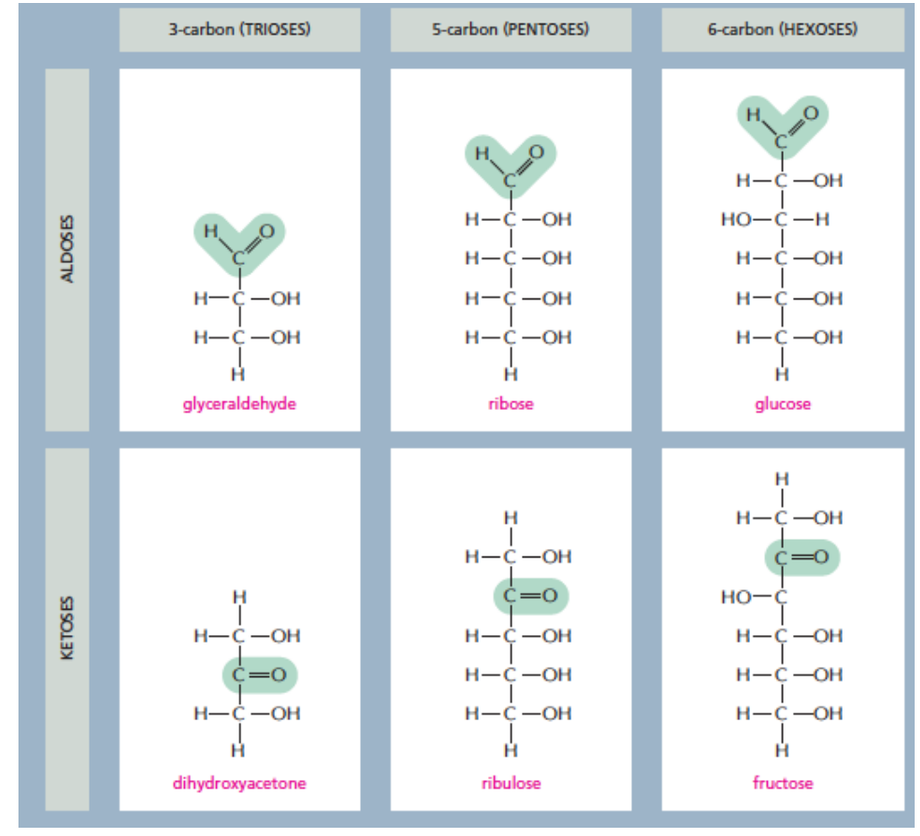


Fructose (a ketose)

What carbon is the carbonyl?

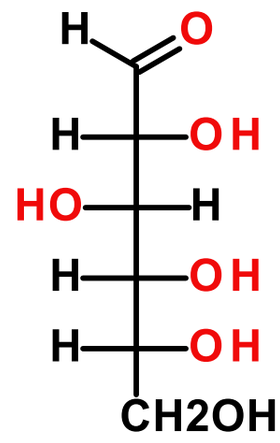
Numbering carbons: Carbon 1 is at the end closest to the C=O group.

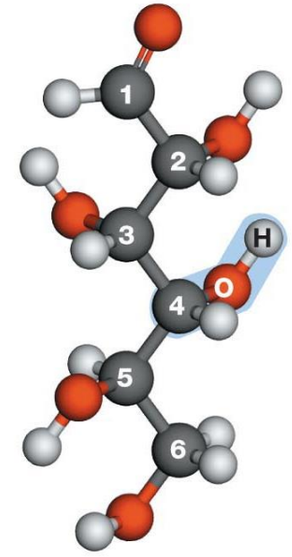
1. Location of the carbonyl group
2. Number of carbons
3. Spatial arrangement of atoms (the position of the OH groups)



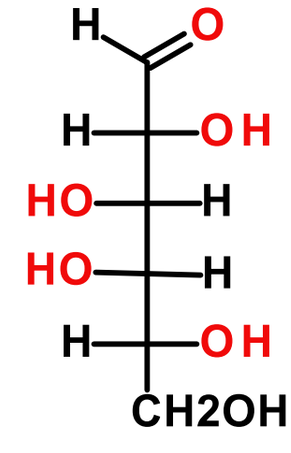
3 ways simple sugars (monosaccharides) differ from each other

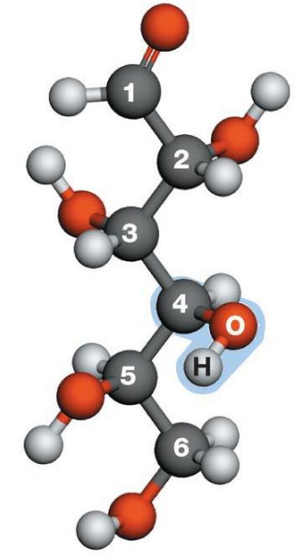
1. Location of the carbonyl group
2. Number of carbons
3. **Spatial arrangement of atoms (the position of the OH groups)**





Glucose





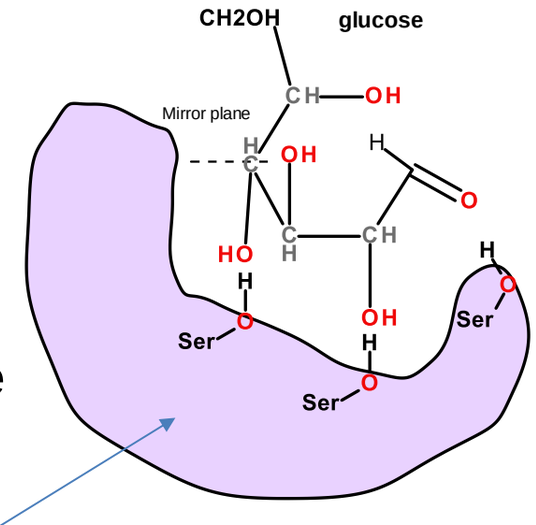
Galactose

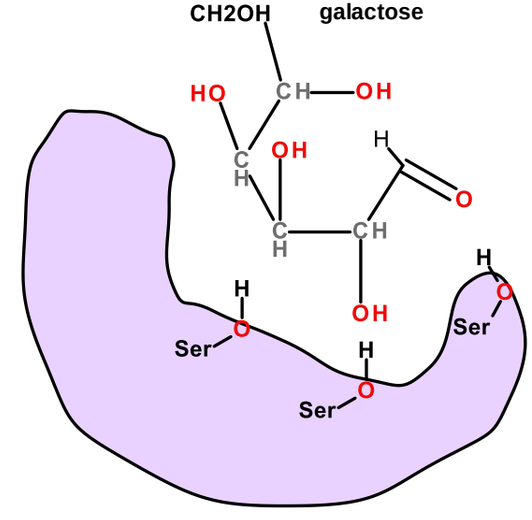
Both have the same chemical formula $C_6H_{12}O_6$. Both are aldose sugars with 6 carbons.
Yet their functions are different.

- Glucose can be used for energy immediately.
- Galactose has to be converted to glucose before it can be used for energy.

They have different interactions with enzymes due to the different chirality at carbon 4.

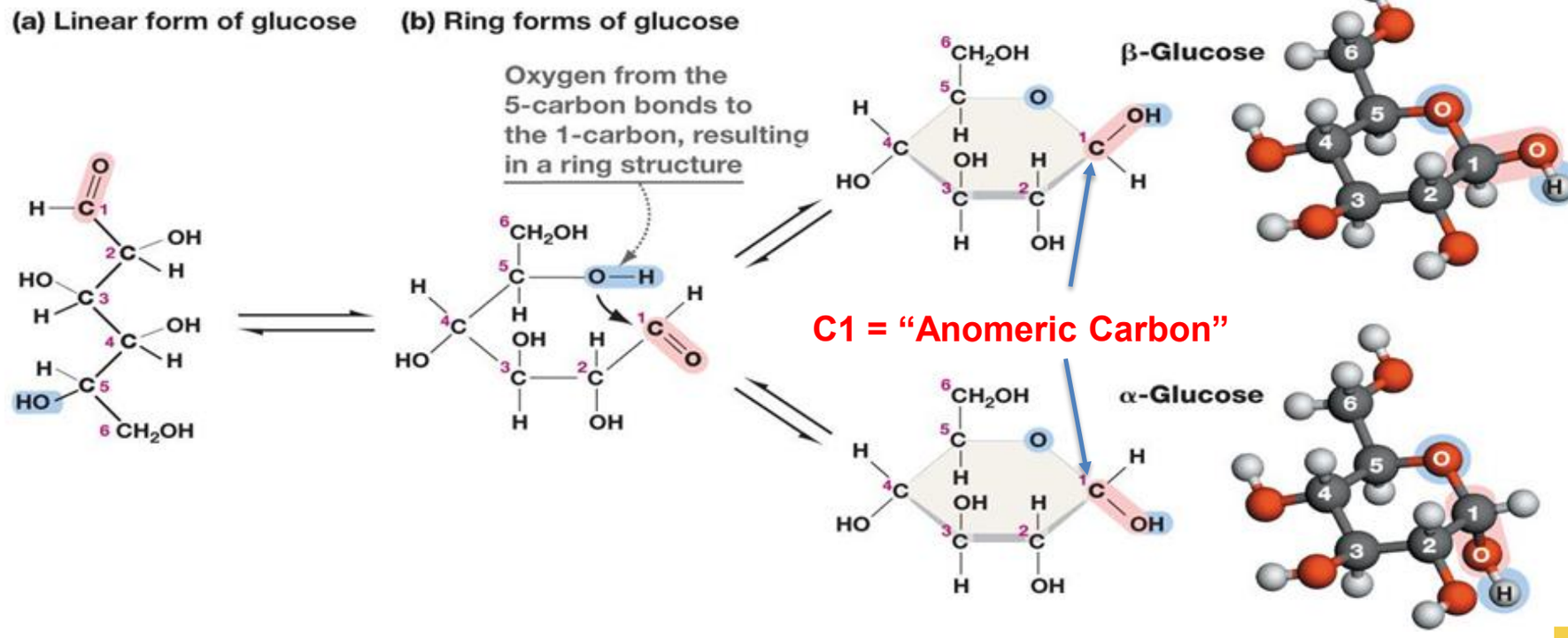
- OH is down in glucose
- OH is up in galactose





Enzyme specific for a-glucose

Ring formation in Monosaccharides:



- In aqueous solution, a hydroxyl group reacts with the aldehyde or ketone group on the same molecule, closing the molecule into a ring, with a bridging oxygen
- It is usually the 2nd to last -OH group, i.e. C5 in glucose, C4 in ribose.
- Stable ring sizes are 5 atoms or 6 atoms
- No atoms are lost or gained in this reaction.
- The carbonyl carbon becomes chiral and is called the ***anomeric carbon*****.**
- The rings with different chirality at C1 are different:

  α (new OH is down), β (new OH is up) *"(ants are down, birds are up)"*

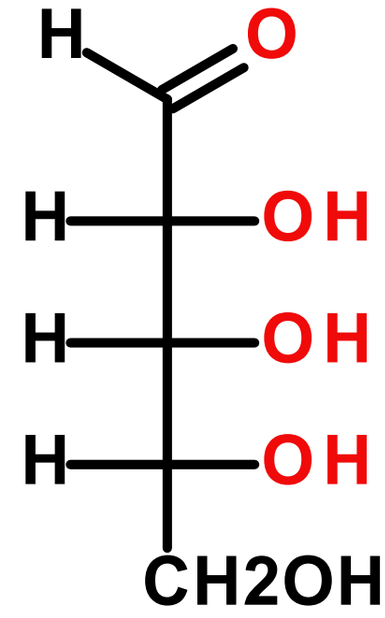
Example Problem:

The linear form of ribose, a 5 carbon aldose is shown on the right. This sugar is found in RNA (ribonucleic acid).

1. Number the carbons.

2. Which carbons are chiral? Mark them with a *.

3. Draw the cyclic form of α-ribose



Disaccharides

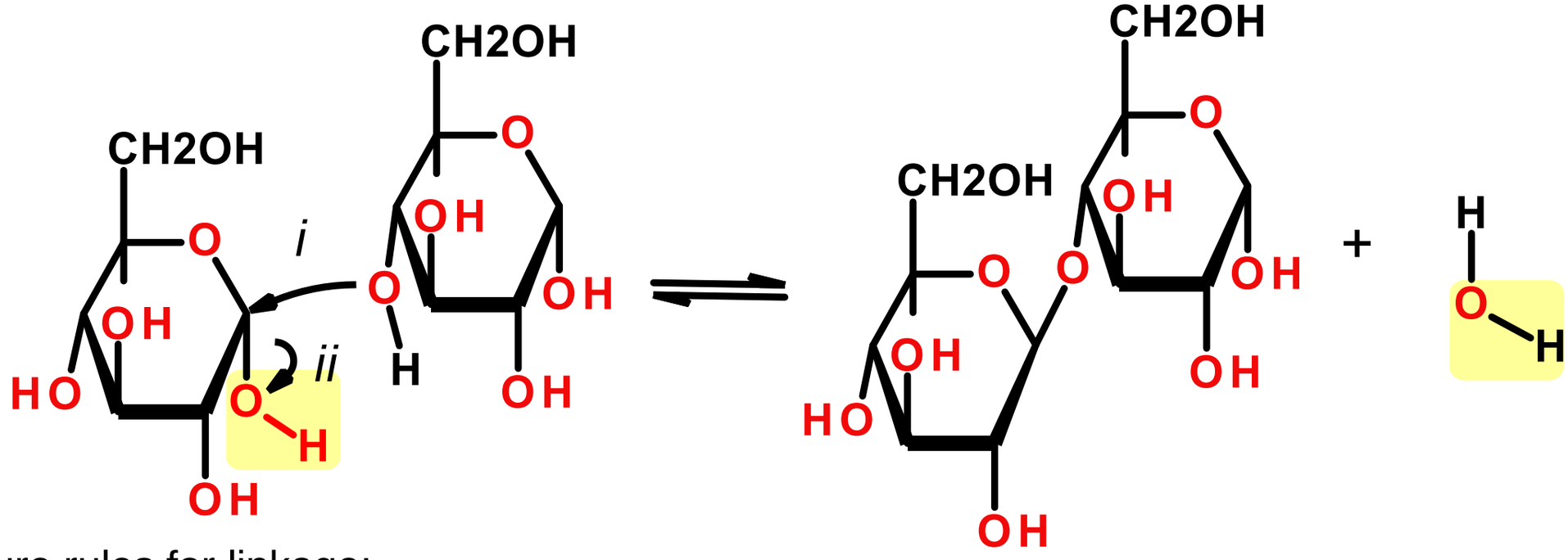
Linkage of the anomeric carbon of one monosaccharide to the OH of another monosaccharide via a *condensation* reaction.

The bond is termed a ***glycosidic bond:***

i) The anomeric carbon is the site of attack by another -OH group.

ii) A water is released

Why is the anomeric carbon the preferred site for nucleophilic attack?



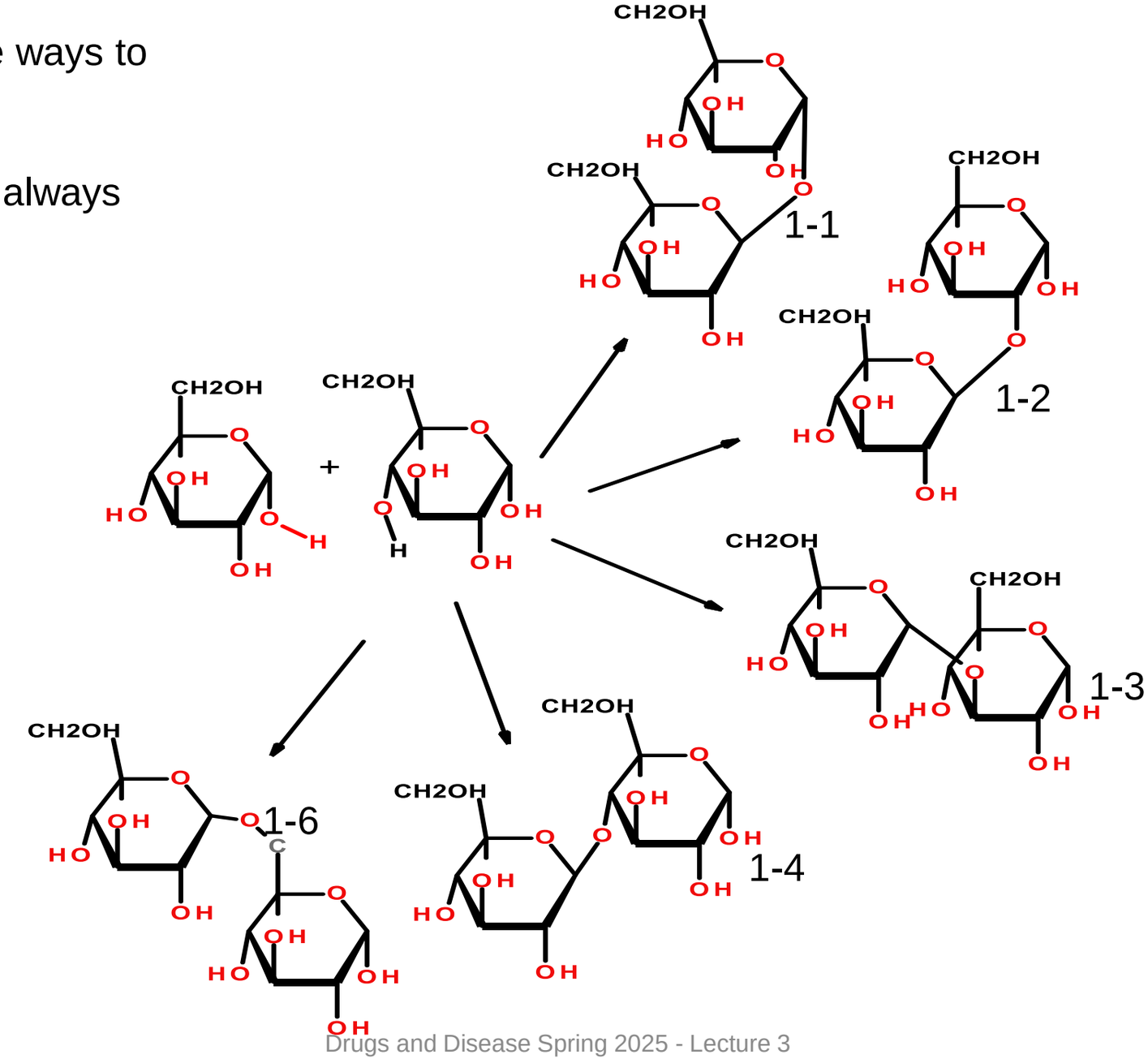
Nomenclature rules for linkage:

- Orientation of the **anomeric** involved in the linkage (α oxygen is down, β oxygen is up)
- Carbons involved in the linkage (e.g. 1-4)

Disaccharides

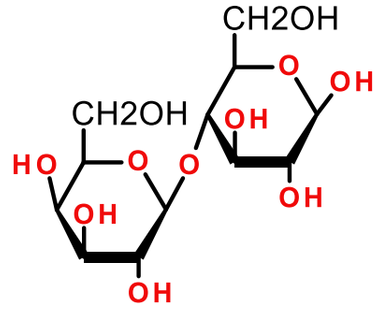
There are many possible ways to connect two sugars.

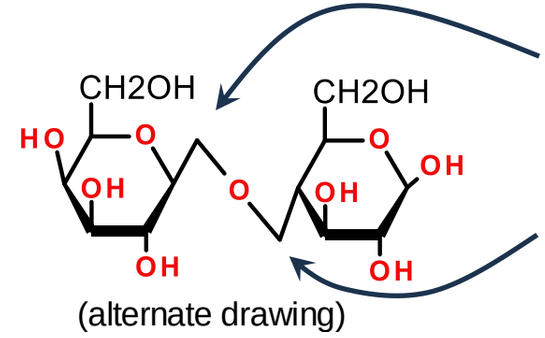
At least one anomeric is always involved.



Lactose (milk sugar)

Disaccharides





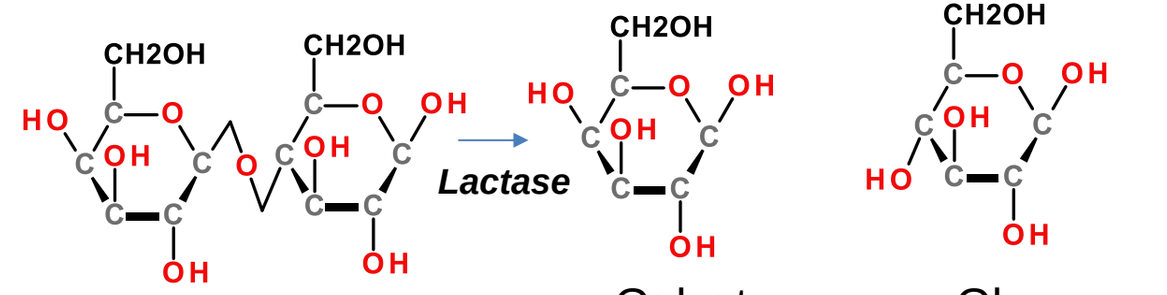
These kinks are not carbons but are drawn in this way to indicate that the chirality of the anomeric is beta (pointing up). The kinks allow the line to reach the downward pointing –OH on C4 in glucose.

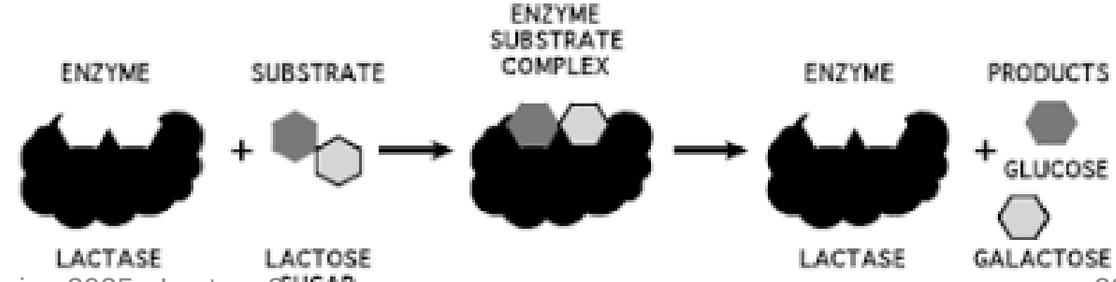
β-galactopyranosyl-(1→4)-β-glucopyranose

Lactose is the major sugar in mammalian milk.

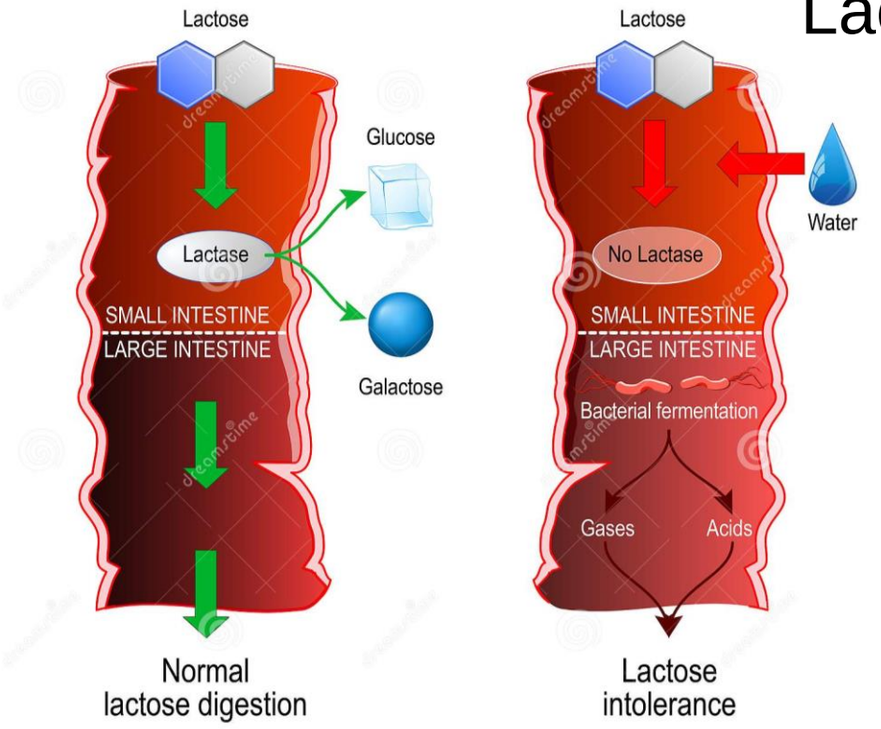
- Infants produce the enzyme ***lactase*** to hydrolyze the disaccharide to monosaccharides.
- Lactase expression is turned off in some adults, depending on their genetic background.

Metabolism of Lactose





Lactose Intolerance

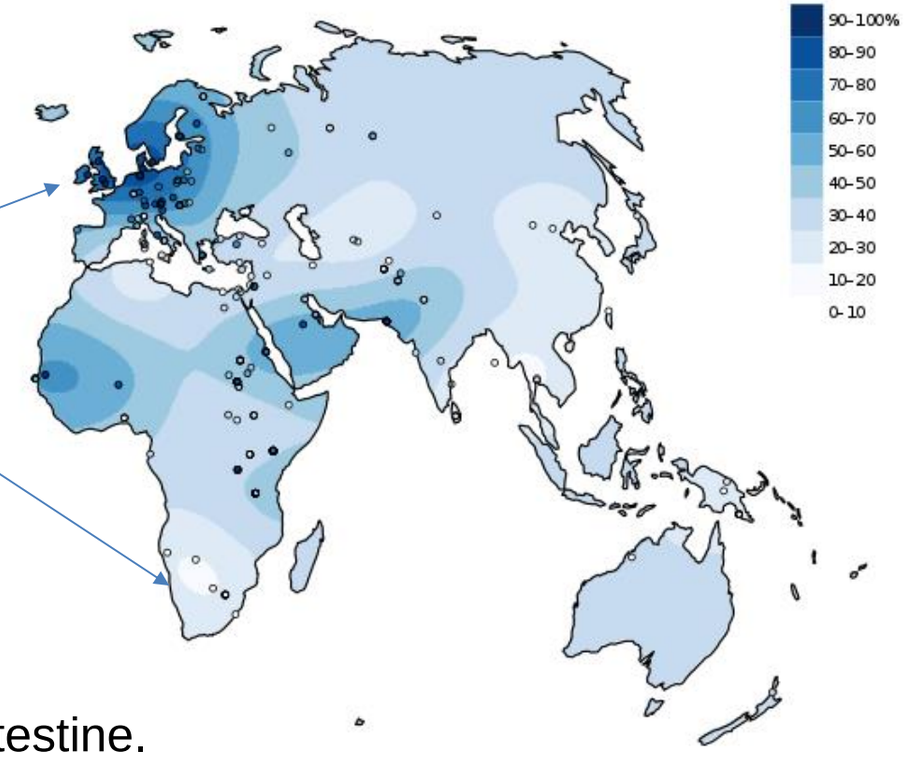


In which region do most of the people still produce lactase as adults?

A) UK

B) South Africa

Lactase Persistence



In an infant (lactase +):

- lactose is broken down to glucose and galactose in the small intestine.
- The two sugars are absorbed and used for energy

In a lactose intolerant individual (lactase -)

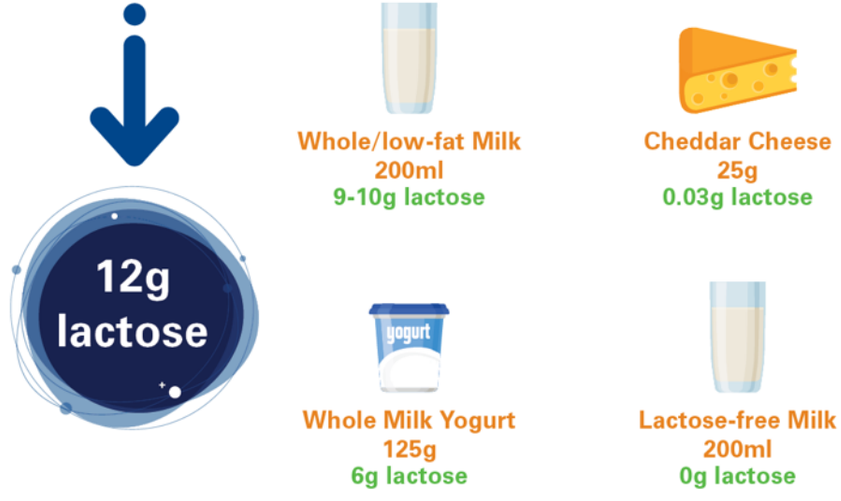
- The lactose is not absorbed in the small intestine, but instead draws water into the intestine due to osmosis – leading to bloating and diarrhea.
- Lactose enters the large intestine where gut bacteria use it as a carbon source, generating gas.

Lactose Intolerance

What to do if you are lactose intolerant:

A. Consume less lactose

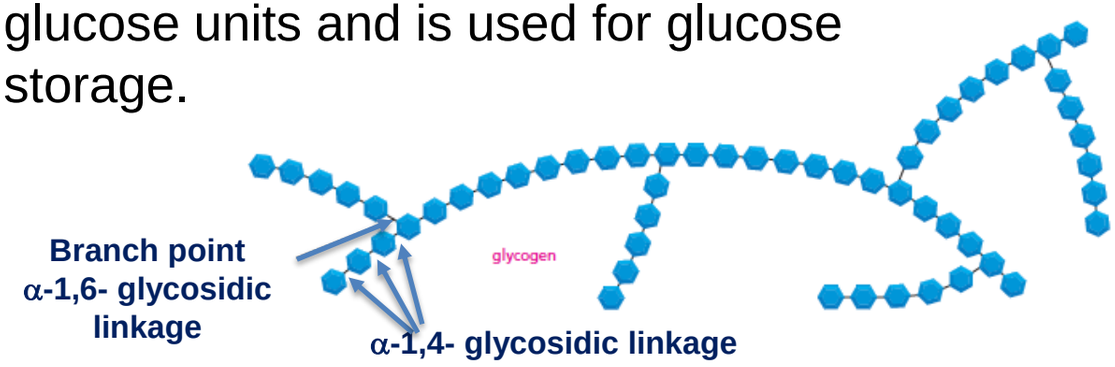
Most individuals with lactose maldigestion can tolerate up to 12g of lactose as a single dose with no, or minor, symptoms
The European Food Safety Authority (EFSA)

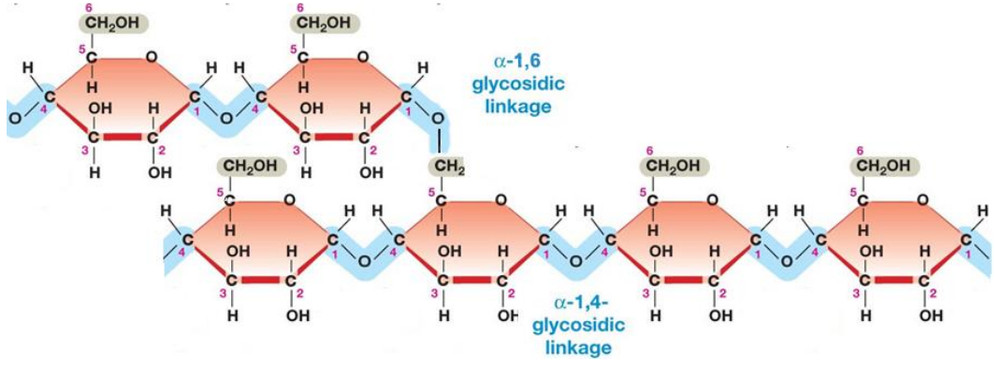


B. Hydrolyze the lactose to glucose and galactose before consumption.

Polysaccharides as Energy Storage – Glycogen Storage Disease

Glycogen and is made entirely of glucose units and is used for glucose storage.



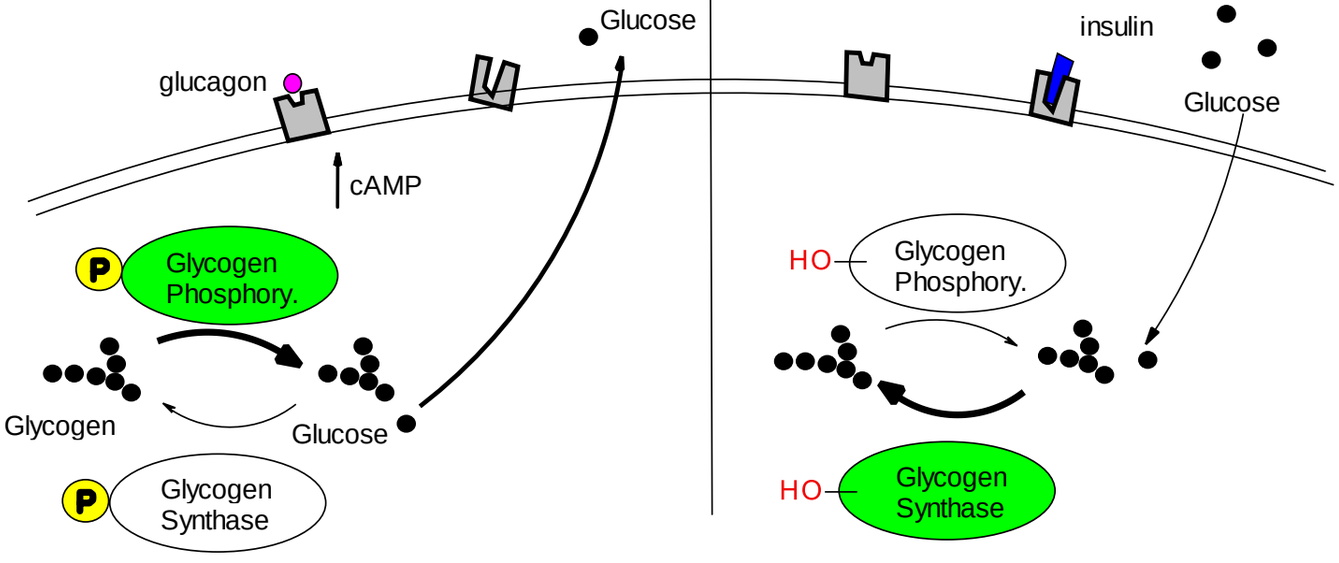


Glycogen Levels are regulated by hormones secreted due to blood glucose levels.

- Glucagon – low blood sugar
- Insulin – high blood sugar

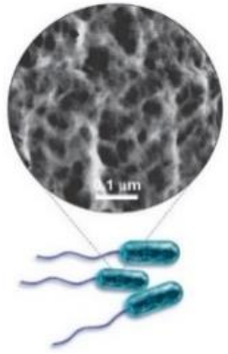
Two enzymes degrade or synthesize glycogen

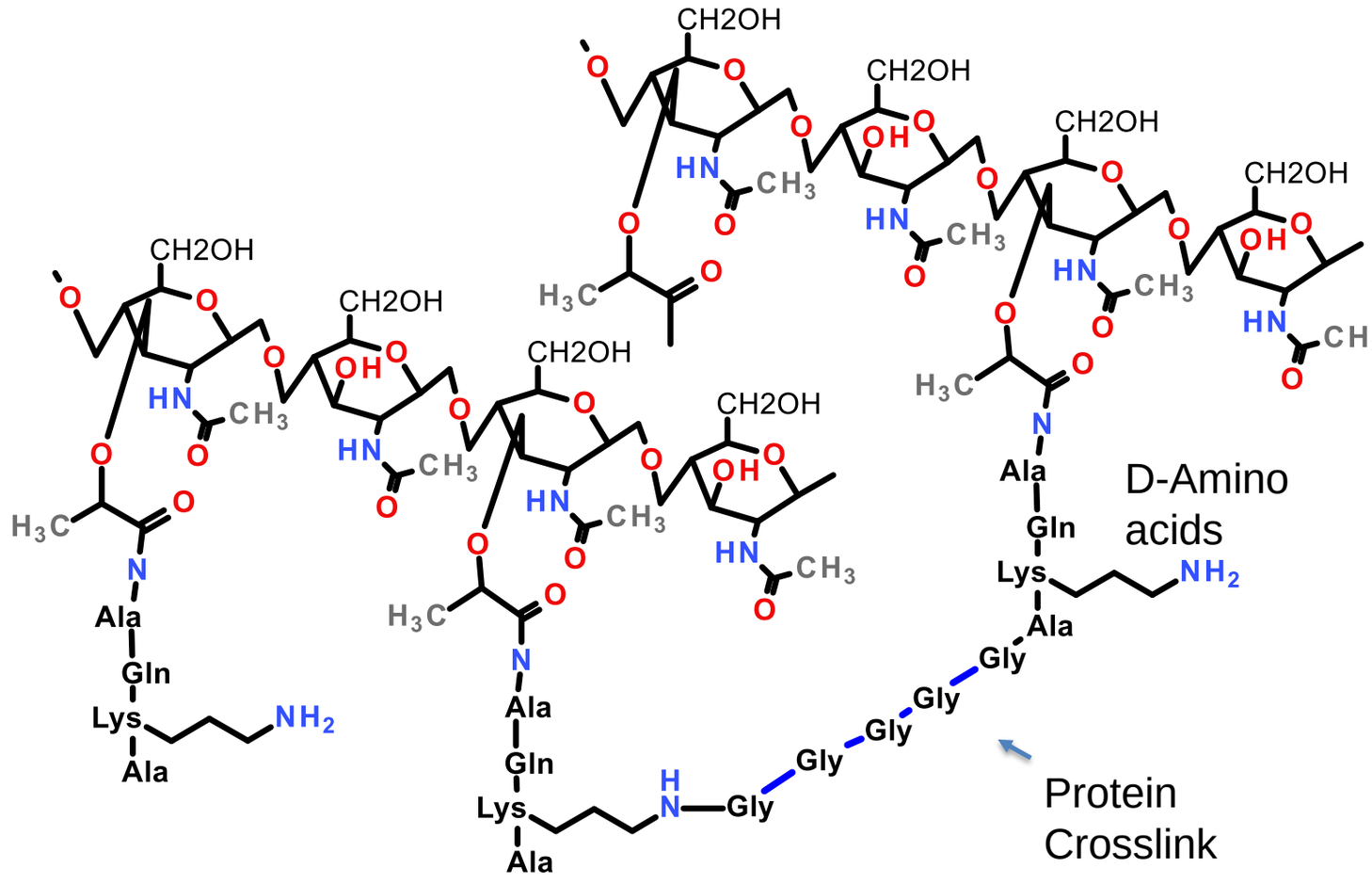
- Glycogen phosphorylase – releases glucose from glycogen
- Glycogen synthase – stores glucose in glycogen



Polysaccharides as Structural Molecules

Peptidoglycan (protein + sugar) in bacterial cell wall





Peptidoglycan (Bacterial Cell Wall)

Many antibiotics interfere with cell wall synthesis (e.g. penicillin)

Summary and Expectations for Carbohydrates

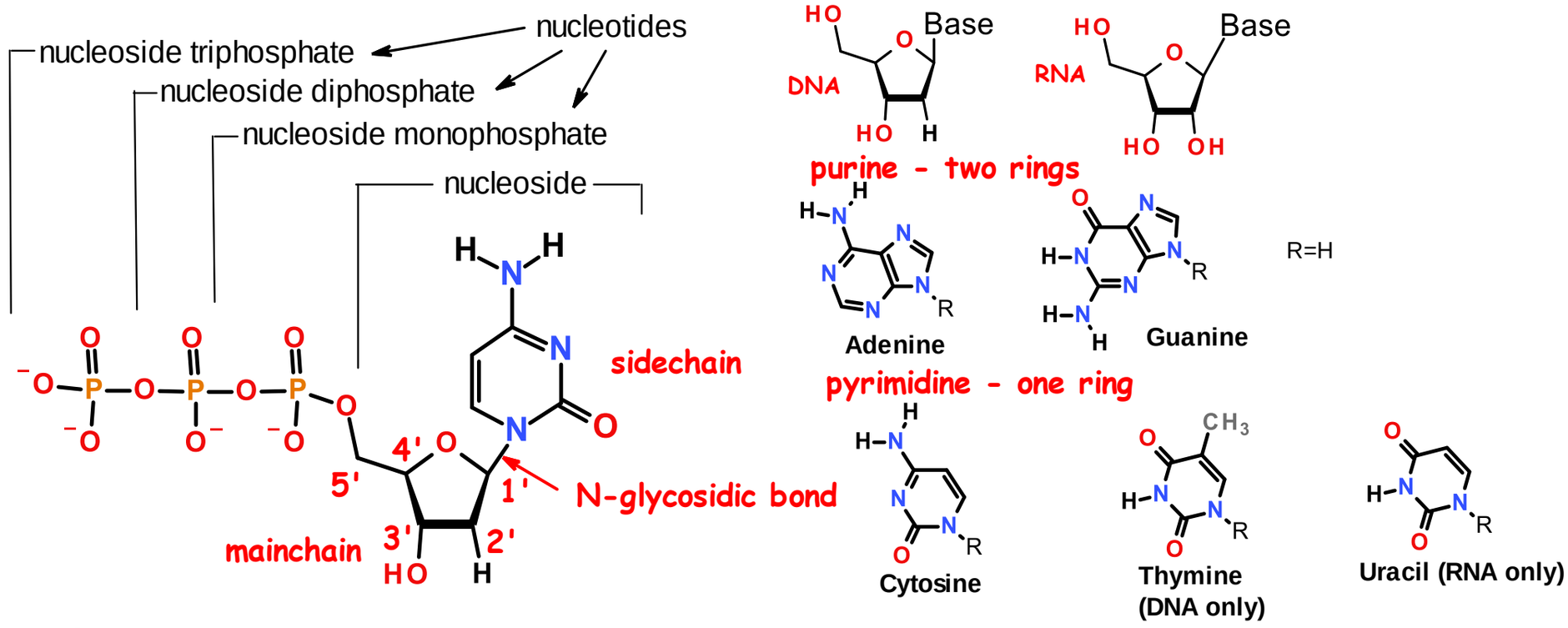
Key Points:

- General structure of monosaccharides - be able to distinguish between aldose and ketose (and identify compounds that are not sugars).
- Know how to number carbons on aldoses and ketoses
- Be able to describe the linkage between two monosaccharides (configuration at the anomeric carbon, atoms linked)
- Treatments for lactose intolerance
- Be able to describe the linkage between glucose molecules in:
  - Glycogen (glucose storage)
- Be able to describe the overall structure of the peptidoglycan in bacterial cell walls.

Nucleic Acid Technologies

- Review of DNA Structure
- Review of DNA Polymerase activity
- Nucleic Acid Technologies – PCR & Sequencing

Nucleic Acid Structure



Monomeric Units

a) Nucleoside triphosphates are the building blocks of nucleic acids (**dNTP = dATP, dGTP, dCTP, dTTP**)

b) The base ("sidechain") is connected to the C1' of the sugar ("mainchain") by an **N-linked glycosidic** bond.
   Base + sugar = **nucleoside**.
   Base + sugar + n-phosphates = **nucleotide**

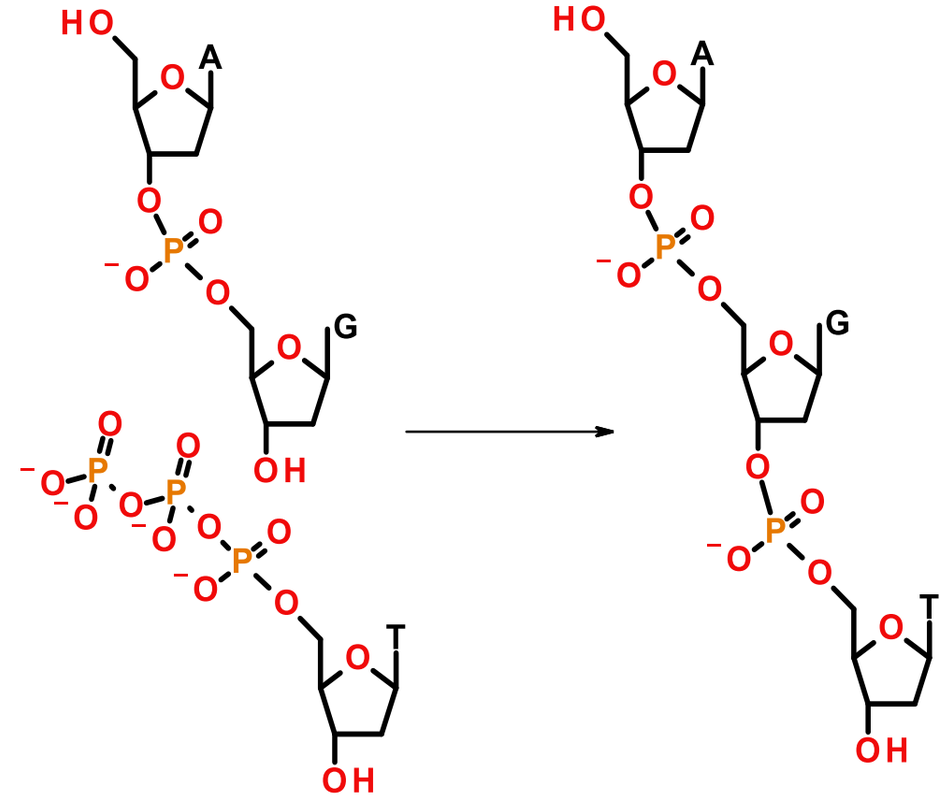
c) The carbon atoms on the sugar are numbered 1' to 5'. The primes distinguish the atoms on the sugar from those on the base.

d) DNA differs from RNA in the sugar (deoxyribose versus ribose) and one base.

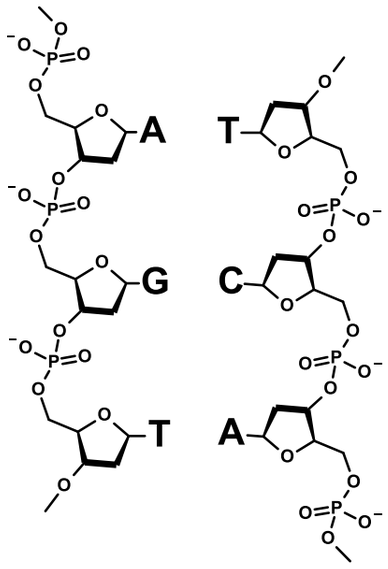
e) Four different monomers, A, G, C, T in DNA. U replaces T in RNA.

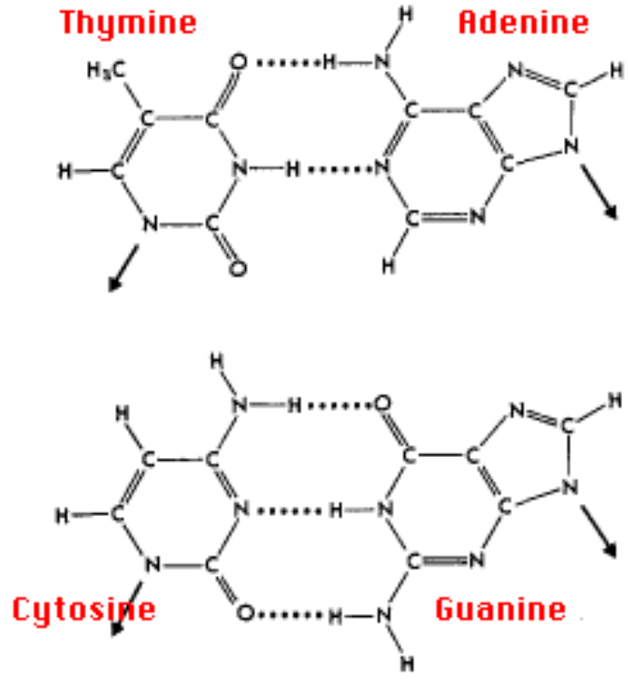
DNA and RNA are Polynucleotides:

- Two phosphates are lost during polymer formation.
- The **phosphodiester** backbone is comprised of deoxyribose (DNA) or ribose (RNA) sugars bridged by one phosphate between the **3' and 5'** positions of the sugars. *Be able to draw this structure.*
- The phosphates are always ionized (pK_a~1), nucleic acids are **polyanions**. The negative charge is important for protein interactions (and electrophoresis).
- Note the polarity: 5' → 3'. *Be able to identify the 5' and 3' ends:*
  - Start at the end atom and move down the chain. The first carbon you find defines the end.



Sequence of nucleotide bases is written in the 5'-3' direction.

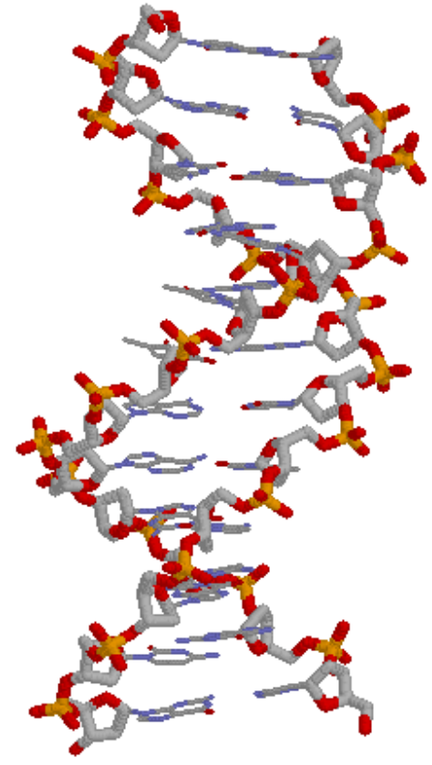




Nomenclature

5’ AGT 3’
3’ TCA 5’

= AGT
TCA



https://www.andrew.cmu.edu/user/rule/jsmol/nucleic.html

Double Helical Structures: B-DNA

a) The helix is right-handed; the chains are **antiparallel**.
b) **10 bp/turn**.
c) The helix interior is filled with stacked base, phosphates and deoxyriboses on the outside.
d) T pairs with A via two "Watson-Crick H-bonds"
e) C pairs with G via three "Watson-Crick hydrogen bonds"
f) Opposite strand termed “complimentary strand”. Top strand is always written 5’->3’, lower strand 3’ -> 5’.

Introduction to Central Dogma

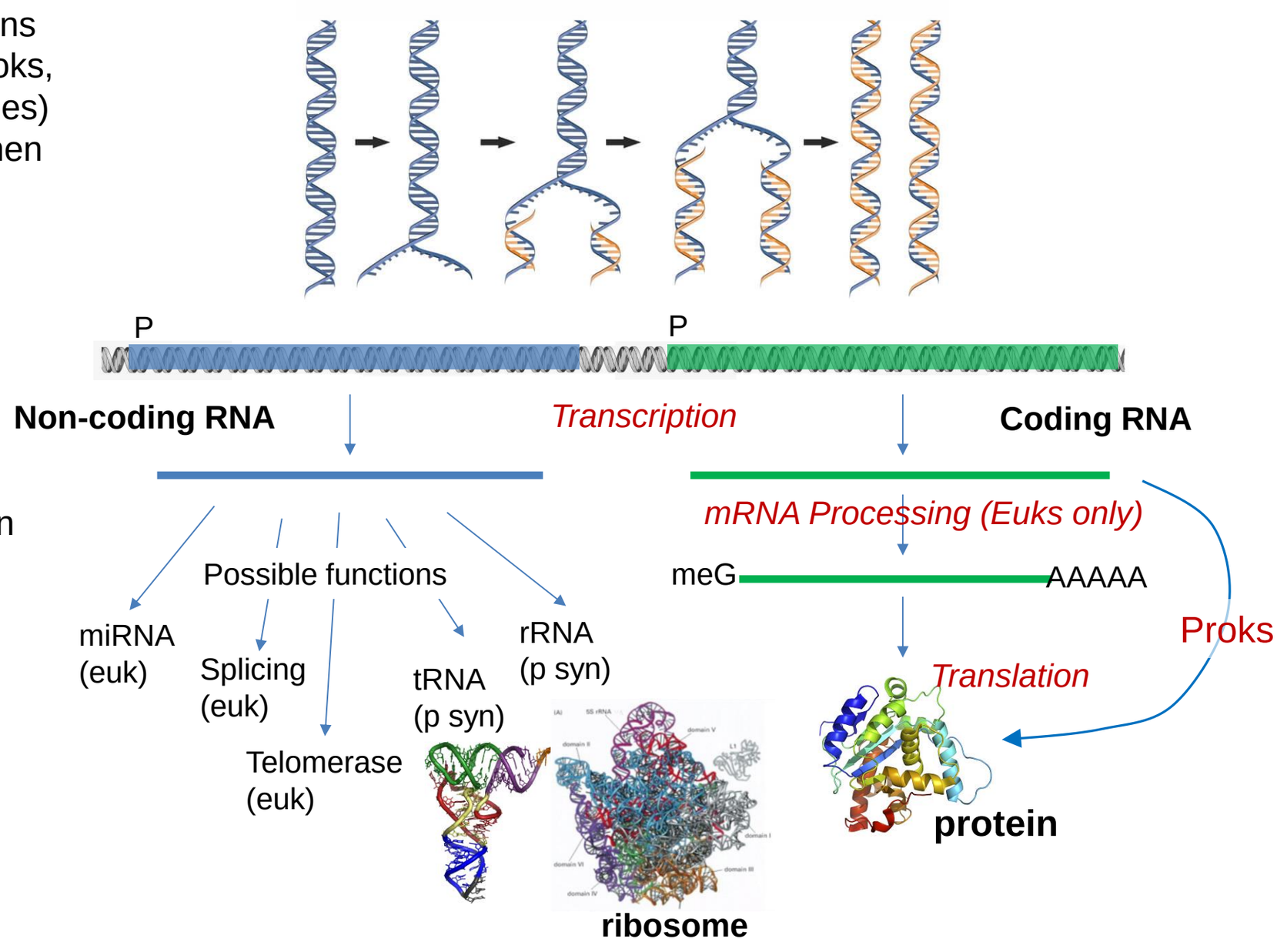
Genome: Entire DNA content of an organism, contains all of the instructions for life. Single circular molecule in Proks, multiple linear molecules (chromosomes) in Euks. The genome is *replicated* when cells divide.

Gene – a segment of DNA that is converted (*transcribed*) to RNA. A *promoter* (P) sequence on the DNA is the minimal requirement for the production of RNA.

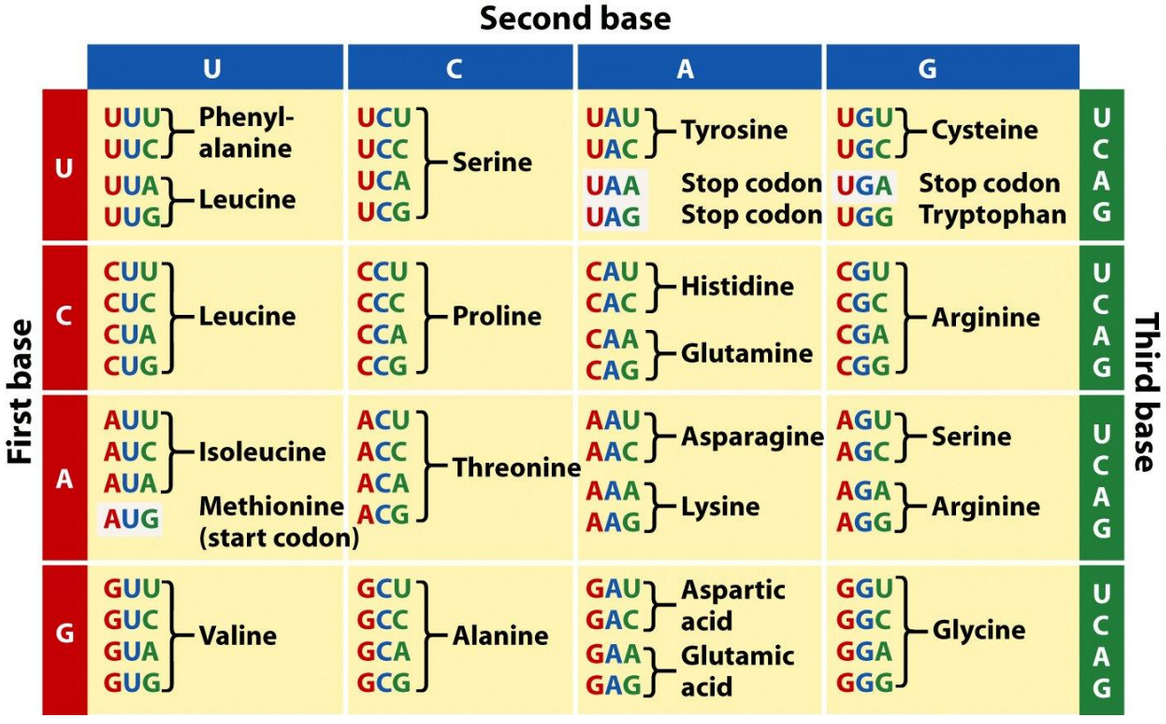
RNA molecules are often processed in **Eukaryotic cells** before they are functional

Many RNAs are functional on their own

mRNA are *translated* to a protein.



The Genetic Code – Converting a DNA/RNA Sequence to a Protein



Codon = 3 bases that code for an amino acid

....ATATGCCCATGTGGTAA..

(DNA Sequence)

...AUAUGCCCAUGUGGUAA..

(mRNA Sequence)

..U–**AUG**–CCC–AUG–UGG–**UAA**

(Punctuated RNA sequence – how the ribosome interprets the sequence)

_____-_____-_____-_____

(Protein Sequence)

- Each codon codes for one amino acid.
- Many amino acids are coded by more than one codon.
- Most organisms use the same codon table – some codons have different meanings in some organisms.

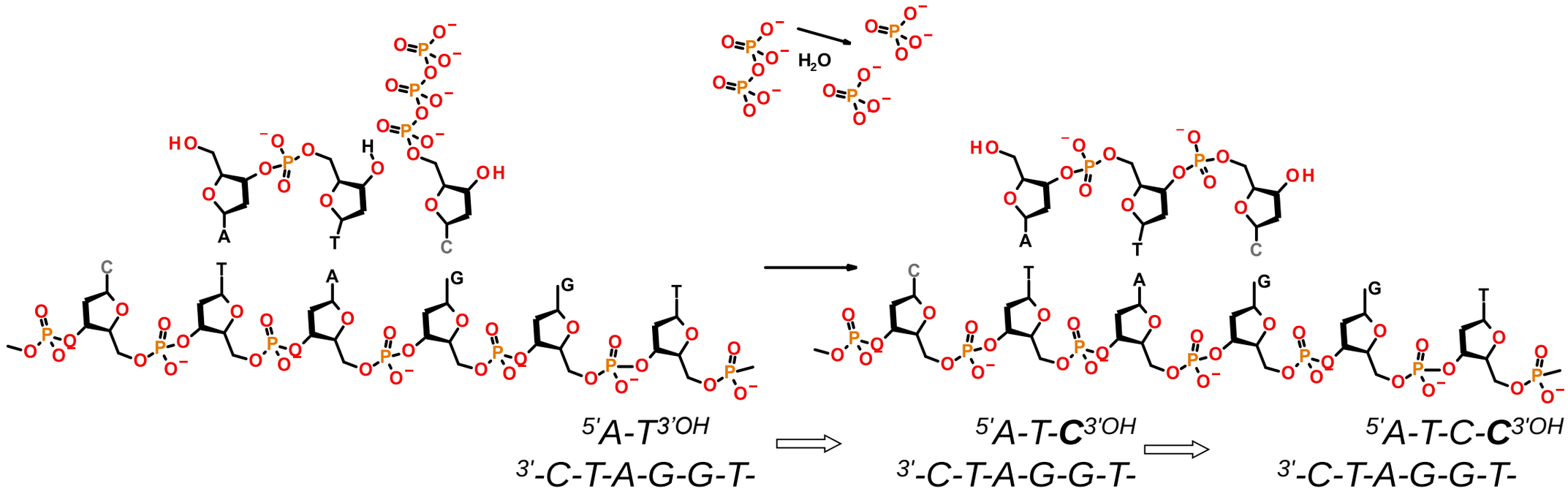
Special Codons:

AUG = Is used to begin almost all proteins that are synthesized on the ribosome, codes for methionine when found internally.

UAA, UAG, UGA = stop codons, terminate synthesis

DNA Polymerases – Used in DNA Sequencing and PCR

- **DNA polymerases** utilize a **template** to direct the order of added bases,
- The enzyme will continue to the end of the template.
- Require a basepaired **primer** with a 3'OH. Primer can be DNA or RNA, DNA is used for laboratory work, RNA is used by the cell during replication
- New dNTP added to the 3' hydroxyl of the existing polymer, elongation in the **5' to 3' direction.**
- Pyrophosphate (PP) is released and hydrolyzed to two inorganic phosphates.



Expectations: Know the features of this reaction.

DNA Polymerase – Fundamental Activity.

```
                    5′ A-T-C-A

3′A-T-G-C-C-G-T-A-G-T-C-G-T-A-C-A-G-T-A-C-G-T-G-C-A
  1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
                    1                   2
```

A short 4 base primer (ATCA) is added to a template, and the temperature is lowered to allow annealing (basepairing) of the primer to the template.

1. *Where (what position) will this primer anneal?*
2. *What is the first base added by the polymerase? A G C T*
3. *What is the last base added by the polymerase? A G C T*

Repeat Expansion Diseases – Errors in DNA Replication

Handbook of Clinical Neurology
Volume 147, 2018, Pages 105-123

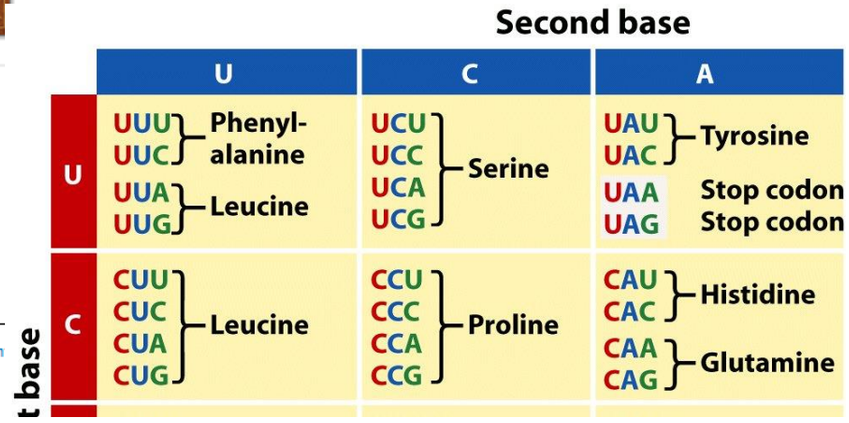
Chapter 9 - Repeat expansion diseases

Henry Paulson

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/B978-0-444-63233-3.00009-9



- CAG – at least 10 diseases (Huntington disease, spinal and bulbar muscular atrophy, dentatorubral-pallidoluysian atrophy and seven SCAs)
- CGG – fragile X, fragile X tremor ataxia syndrome, other fragile sites (GCC, CCG)
- CTG – myotonic dystrophy type 1, Huntington disease-like 2, spinocerebellar ataxia type 8, Fuchs corneal dystrophy
- GAA – Friedreich ataxia
- GCC – FRAXE mental retardation
- GCG – oculopharyngeal muscular dystrophy
- CCTG – myotonic dystrophy type 1
- ATTCT – spinocerebellar ataxia type 10
- TGGAA – spinocerebellar ataxia type 31
- GGCCTG – spinocerebellar ataxia type 36
- GGGGCC – C9ORF72 frontotemporal dementia/amyotrophic lateral sclerosis
- CCCCGCCCCGCG – EPM1 (myoclonic epilepsy)

- Repeats in coding regions of genes will generate long stretches of the same amino acid.

  CAGCAGCAG = GluGluGlu
- Repeats outside of coding regions can affect gene expression by changing binding of transcription factors.
- These repeats can grow due to slippage of primer during replication
- More repeats = more chance of developing disease.

- The number of repeats can be detected by:
  - DNA sequencing
  - PCR

Repeat Expansions – How Do They Grow?

Original Sequence - 3 repeats (CAG)

-TATATCCAGCAGCAGAGTATA-
-ATATAGGTCGTCGTCTCATAT-

During Replication in the cell

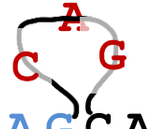
-TA
-ATATAGGTCGTCGTCTCATAT-

-TATATCCAGCAGCAG
-ATATAGGTCGTCGTCTCATAT-

3’ end comes loose (primer slippage)

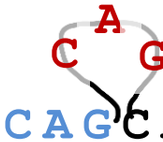
-TATATCCAGCAGCAG
-ATATAGGTCGTCGTCTCATAT-

Looped out DNA



-TATATCCAGCAG
-ATATAGGTCGTCGTCTCATAT-

Replication continues



-TATATCCAGCAGCAGAGTATA-
-ATATAGGTCGTCGTCTCATAT-

Next replication (upper strand as the template)

-TATATCCAGCAGCAGCAGAGTATA-
TCATAT-

4 repeats

-TATATCCAGCAGCAGCAGAGTATA-
-ATATAGCTCGTCGTCGTCTCATAT-

DNA Sequencing – Sanger (dideoxy) Sequencing

DNA Sequencing - Determining the Order of Bases in the DNA.

Sanger Sequencing:

- Second method to generate long (~1000 base) sequence information (an earlier chemical method developed by Gilbert proved to be impractical for most laboratories (hydrazine = rocket fuel was required)
- Sanger was awarded his 2nd Nobel prize for this work in 1980, shared with Gilbert.

Determine the position of all four bases in a DNA strand = Sequence (video)

Sanger Sequencing:

Primer $^{5'}$**C-A-T-A-T-G**^{OH}

Template 3′G-A-A-G-T-C-G-A-A-G-G-T-A-T-A-C-**C**-A-T-T-A-G-G-C-C-A-T-G-C-A-C-G-T----

Known Seq (plasmid) Unknown sequence (insert)

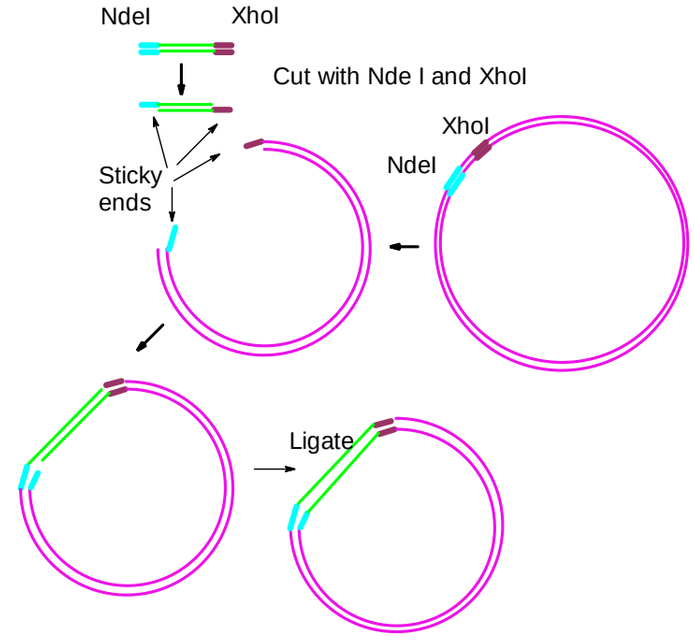
Sequenced region (~1000 bases)

5′C-A-T-A-T-G-G-T-A-A-T-C-C-G-G-T-A-C-G-T-G-C-A----

Template 3′G-A-A-G-T-C-G-A-A-G-G-T-A-T-A-C-**C**-A-T-T-A-G-G-C-C-A-T-G-C-A-C-G-T----

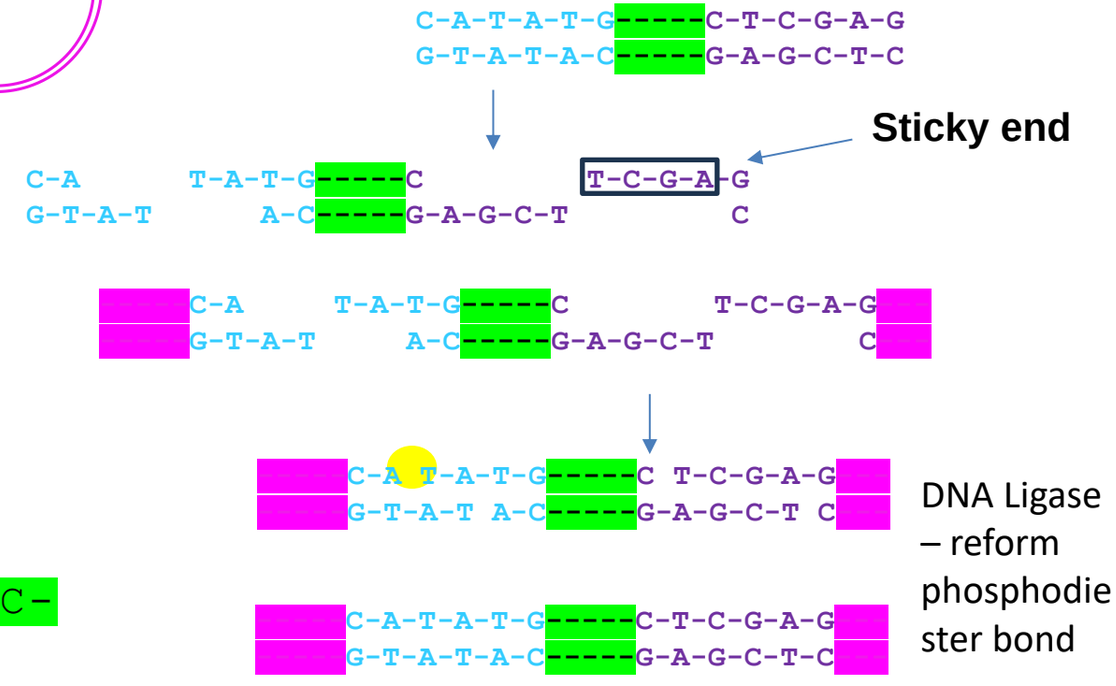
DNA Sequencing - Determining the Order of Bases Added by DNA Polymerase

- The DNA to be sequenced is inserted into a circular piece of double stranded DNA called a plasmid. The DNA sequence of the plasmid is known.
- The insertion is often accomplished using restriction enzymes that generate single stranded overhangs that allow DNA molecules to be efficiently joined.
- Restriction sites can be added to any DNA fragment using a number of techniques:
  - Addition of a short linker (same site on both ends)
  - PCR (different sites on each end)

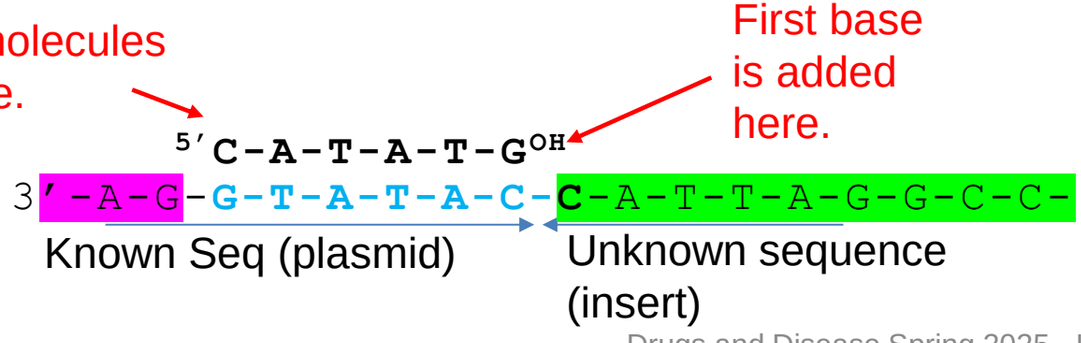


Restriction Enzymes

- Recognize a specific sequence in the DNA
- Sequence has 2-fold symmetry – same on the top and bottom strand
- Cuts both strands, most generate single-stranded DNA (sticky ends).
- Complementary sticky ends can bind to each other.

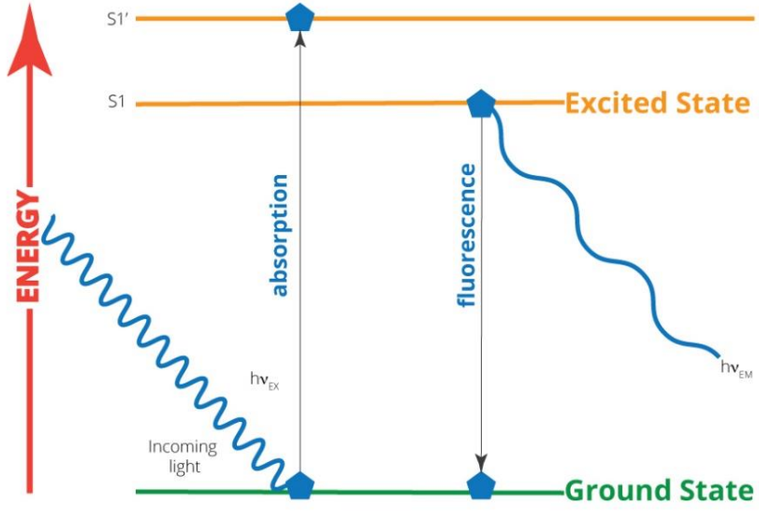


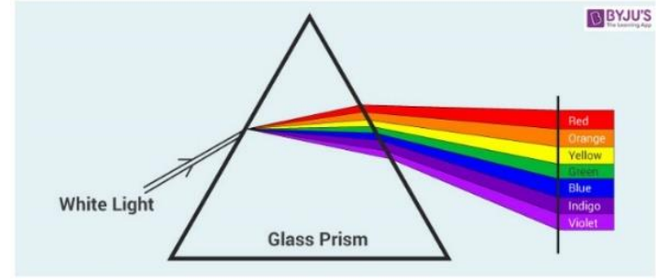
Key Point 1. Start sequencing at known location with primer that anneals at a ***unique*** location on the plasmid, "upstream" from the region to be sequenced.



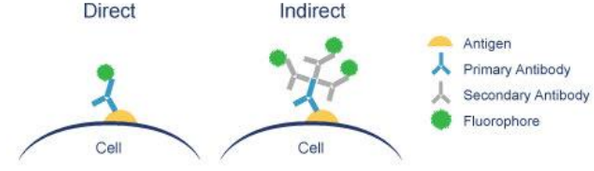
DNA Sequencing Methods Use Fluorescent Bases - What is Fluorescence?

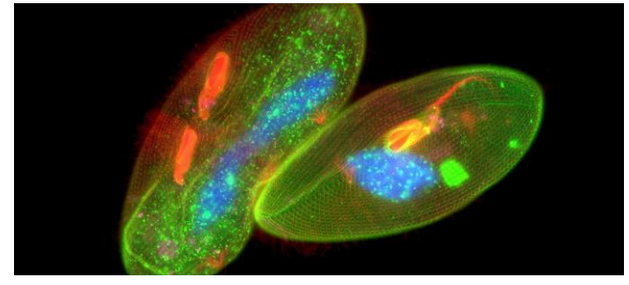
- When molecules absorb light an electron goes from a lower shell to a higher shell. This is where the energy from the light goes.
- In most molecules the electron goes back down to its original shell with the release of heat.
- Fluorescent molecules emit the energy as light of a longer wavelength (different color).
- The color that is emitted depends on the molecule.





Fluorescently tagged antibodies can be used to stain components of cell with fluorophores.

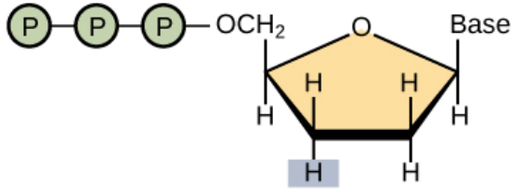




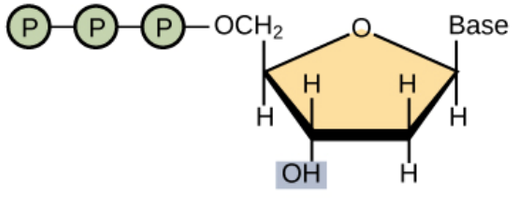
DNA Sequencing - Determining the Order of Bases Added by DNA Polymerase

Key point 2. Use a mixture of normal bases (dNTPs) and dideoxy bases (ddNTP) for polymerization. Ratio of dNTP to ddNTP is (100:1), ***most of the time elongation occurs.***

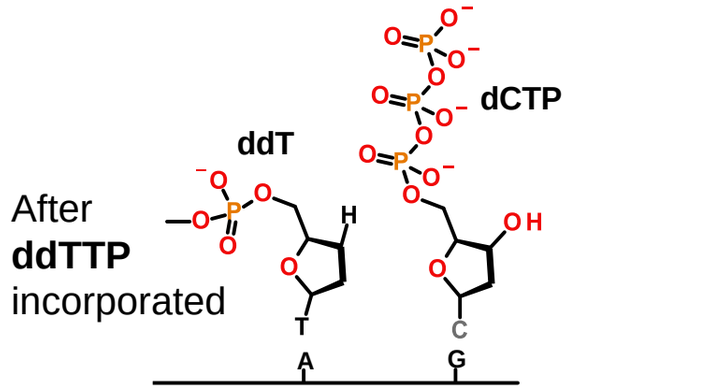
- ddNTPs can be added to the DNA since they have a 5'-triphosphate but ***terminate*** the chain due to the lack of a 3'-OH. ~ 1 in 100 chains terminate at each base addition

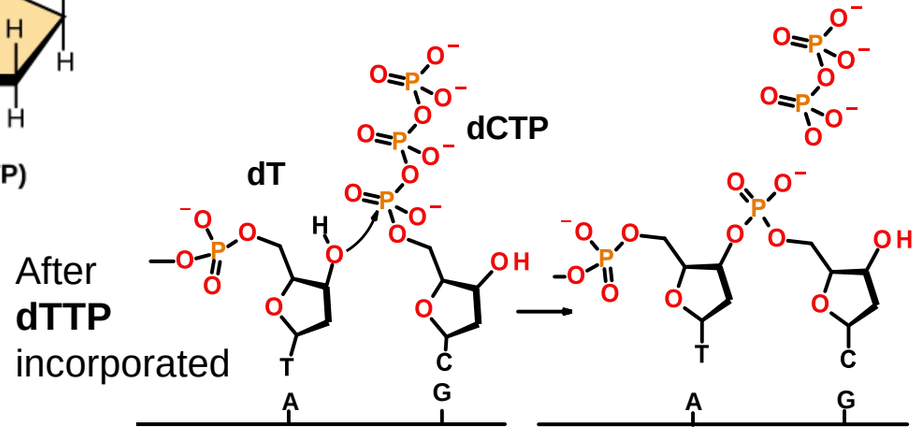


Dideoxynucleotide (ddNTP)



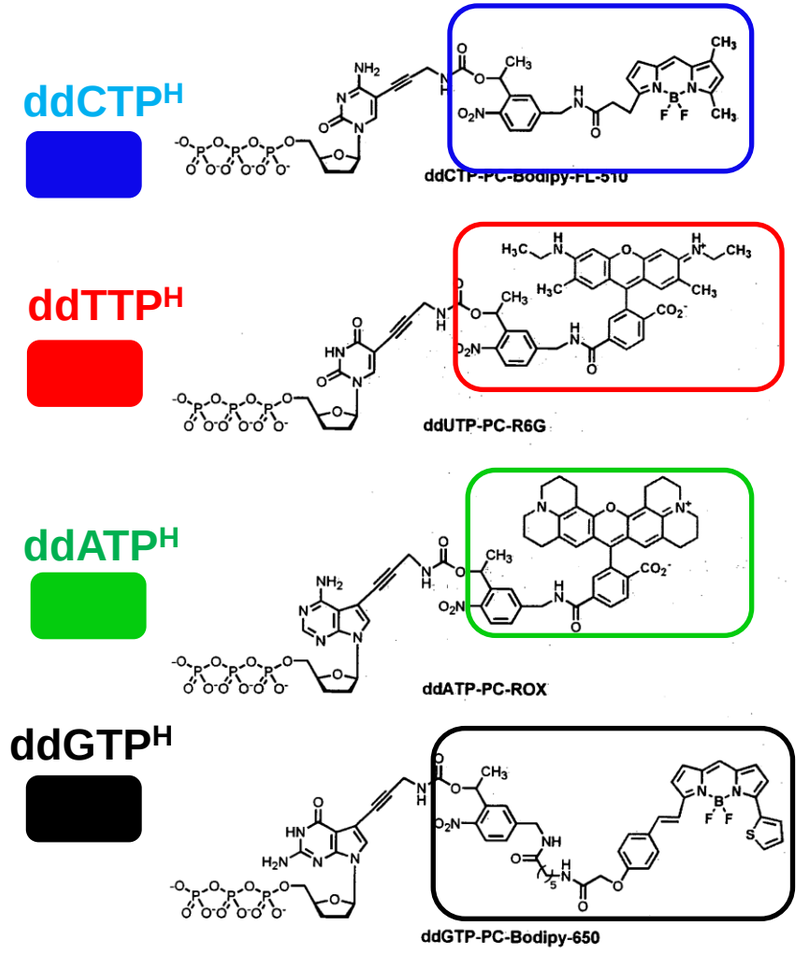
Deoxynucleotide (dNTP)



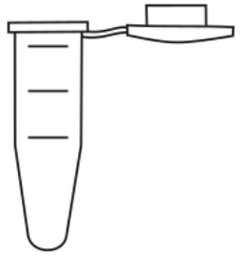


Key point 3. The ddNTPs are color coded by different fluorescent emission wavelengths.

The ddNTP that terminated the chain is known from its fluorescent color.



DNA Sequencing – Generation of Fluorescent Fragments



Template
Primer
DNA Pol
dTNP, ddNTP

5′-C-A-T-A-T-G
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

(1000 molecules)

Length=7, Black fluor.

(10) 5′-C-A-T-A-T-G-G^{H}
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

(990) 5′-C-A-T-A-T-G-G^{OH}
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

5′-C-A-T-A-T-G-G^{OH}
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

(990 molecules)

Length=8, Red fluor.

(10) 5′-C-A-T-A-T-G-G-T^{H}
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

(980) 5′-C-A-T-A-T-G-G-T^{OH}
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

5′-C-A-T-A-T-G-G-T^{OH}
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

(980 molecules)

Length=9, Green fluor

(10) 5′-C-A-T-A-T-G-G-T-A^{H}
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

(970) 5′-C-A-T-A-T-G-G-T-A^{OH}
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

5′-C-A-T-A-T-G-G-T-A^{OH}
3′-A-G-G-T-A-T-A-C-C-A-T-T-A-G-G-C

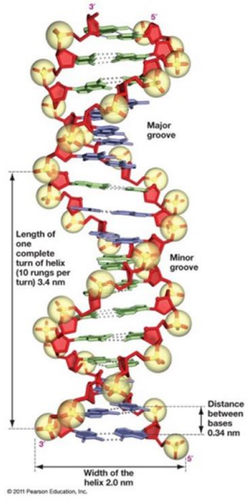
***All* Possible Fragments are Made:**

1. Each begins with the primer
2. Each ends with a *known* ddNTP, based on the color of the fluorescence.
3. Each is one longer than the previous.

C-A-T-A-T-G-**G**
C-A-T-A-T-G-G-**T**
C-A-T-A-T-G-G-T-**A**
C-A-T-A-T-G-G-T-A-**A**
C-A-T-A-T-G-G-T-A-A-**T**
C-A-T-A-T-G-G-T-A-A-T-**C**
C-A-T-A-T-G-G-T-A-A-T-C-**C**
C-A-T-A-T-G-G-T-A-A-T-C-C-**G**

Primer | Added by Pol.

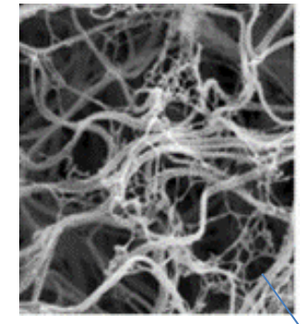
Size Determination of Fragments from DNA Sequencing Capillary Electrophoresis

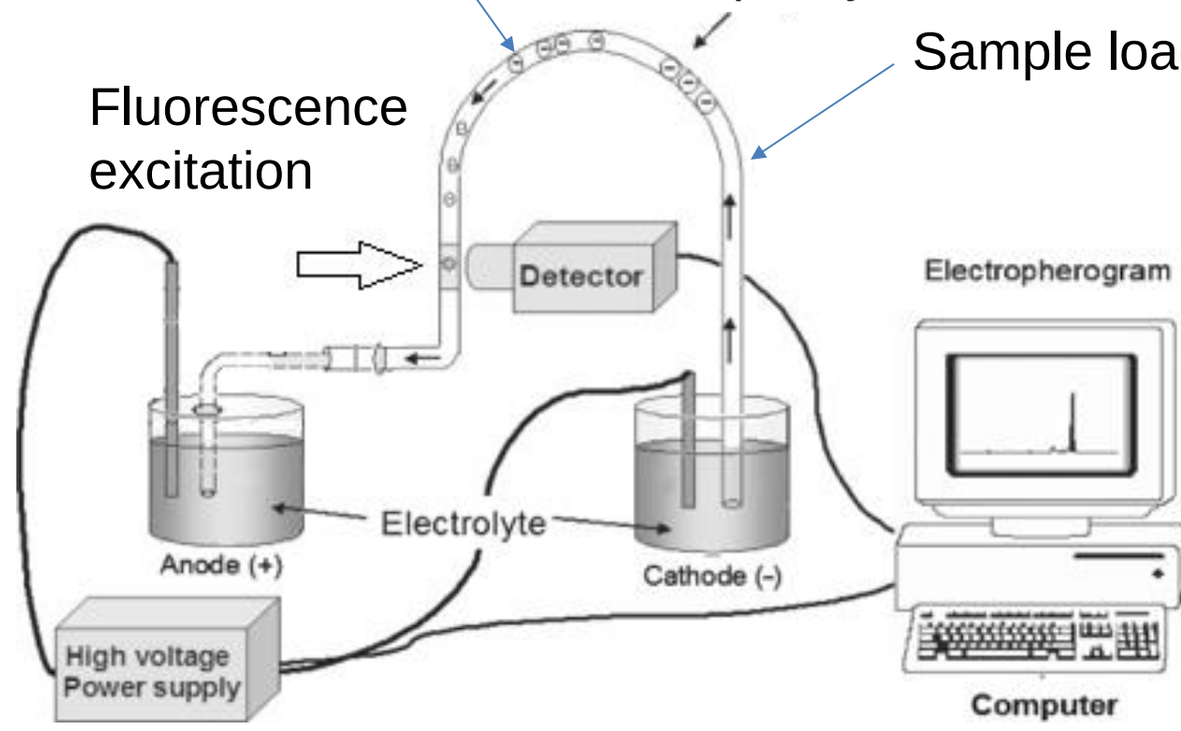


DNA has a negative charge.
It will migrate towards the anode.

Capillary is filled with a gel that causes separation by size.

DNA molecules that are smaller migrate _____.



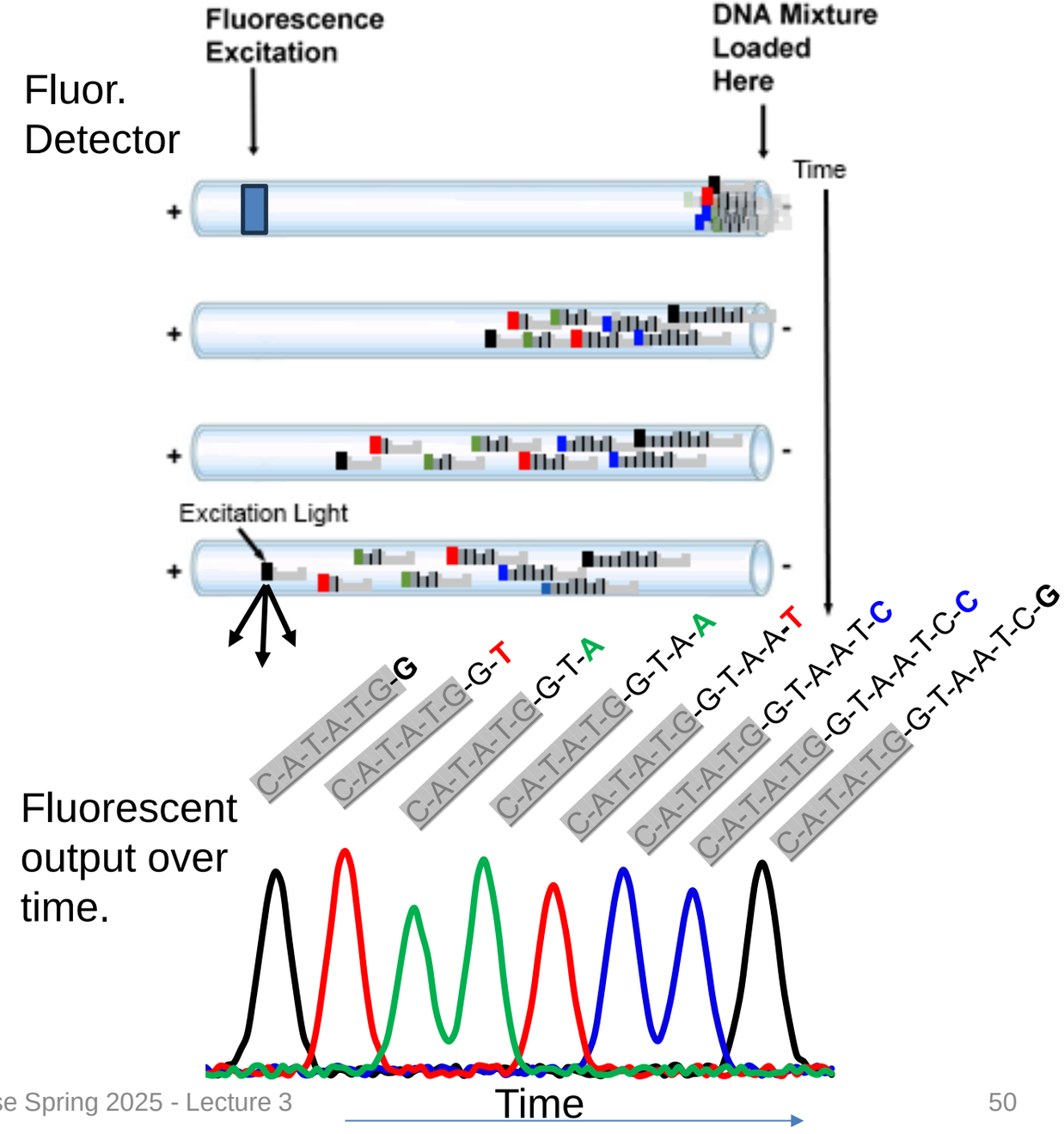


DNA Sequencing – Analysis of Fragments to Determine Order of Addition

4. Capillary Gel Electrophoresis

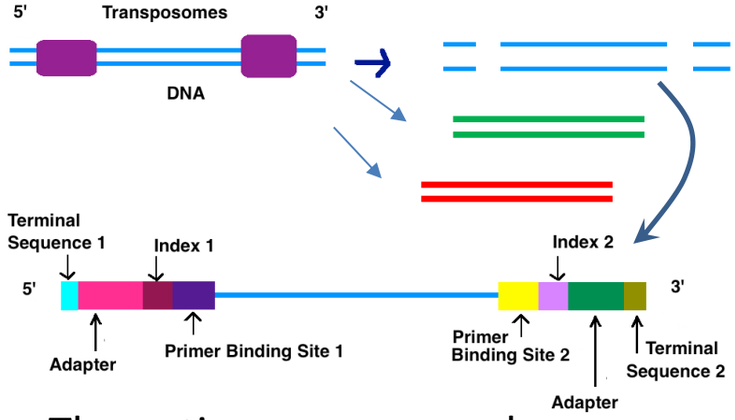
- Migration due to the voltage because of the neg. charge on DNA phosphates
- Separation of DNA molecules by size, smaller travel through gel faster.
- Fragments reach the detector in the order of their length: primer+1 first, primer+2 second, etc.
- At the detector, a laser excites the fluorescence.
- Only fluorescent DNA molecules (terminated with ddNTP) give a signal.
- The color of the emitted fluorescence gives the dideoxy base at the 3' end of the DNA fragment.
- The order of peaks gives the sequence that is complementary to the template (= strand with primer).

5'-C-A-T-A-T-G G-T-A-A-T-C-C-G
3'-A-G-G-C-T-A-T-A-C-C-A-T-T-A-G-G-C

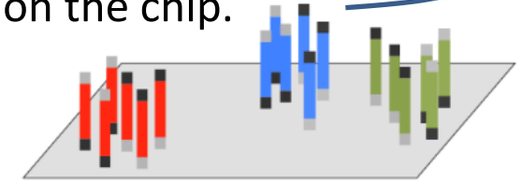


Newer Sequencing Methods-Illumina Dye Sequencing – Next Generation High Throughput

A. Obtaining the DNA

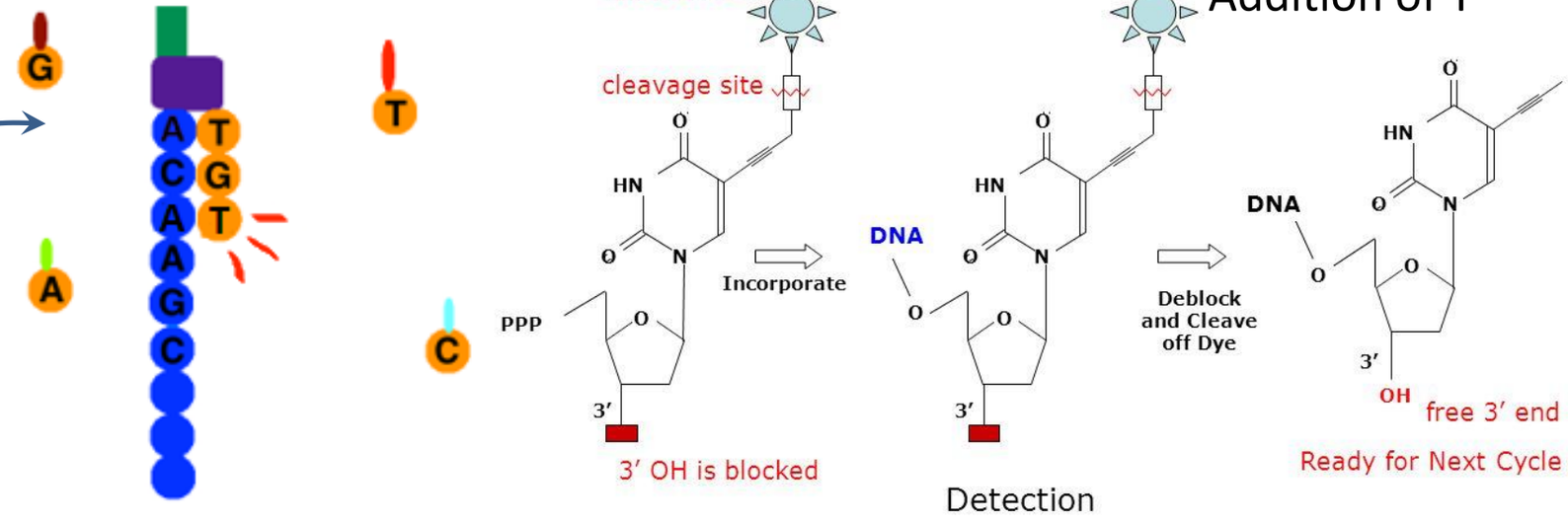


- The entire genome can be sequenced.
- The DNA is fragmented into small 100 base pieces.
- Synthetic DNA is added to the ends (sites for primers for sequencing)
- Different fragments are bound to different location on a solid surface (chip).
- All fragments are sequenced at the same time on the chip.



Cluster formation

B. Sequencing by synthesis – Fluorescent labeling & *reversible* 3'-OH blocking

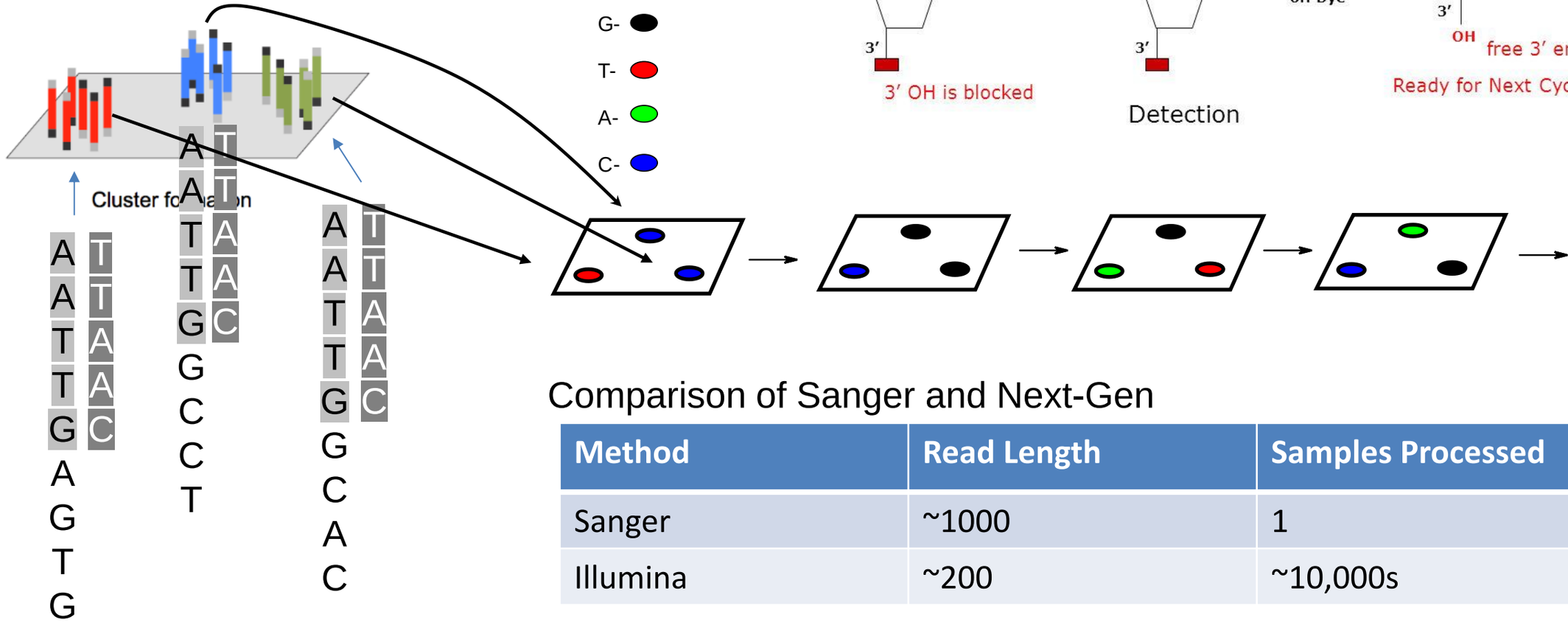


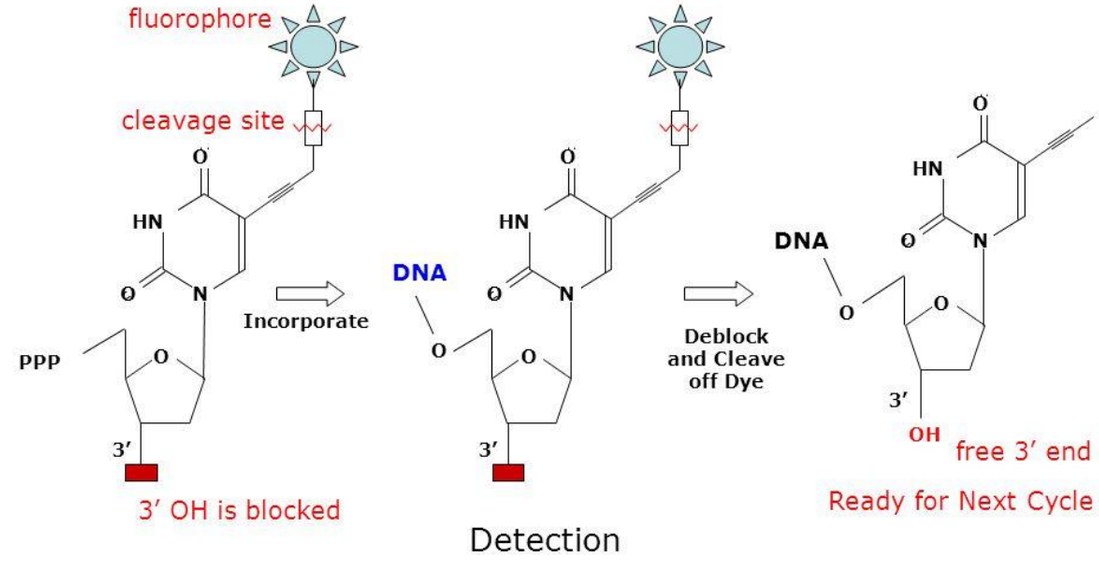
1. Only one base is added at a time (3'-OH is blocked)
2. The base that is added is determined by the color of the fluorescent base.
3. 3'-OH blocking group and the fluorescent group are removed prior to the next addition. ~100 cycles can be performed.

Next Generation - Data

1. Primer anneals
2. Add dNTPs (3'blocked) + Polymerase
3. Wash to remove unincorporated dNTPs
4. Detect base added to each cluster using fluorescent emission
5. Unblock 3'-OH

Repeat.



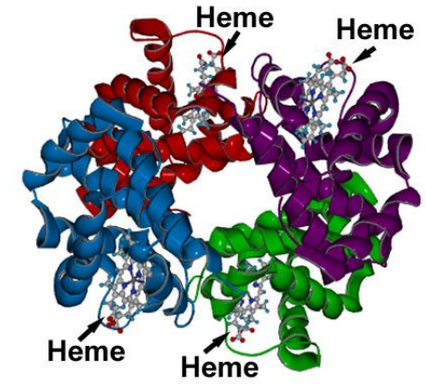


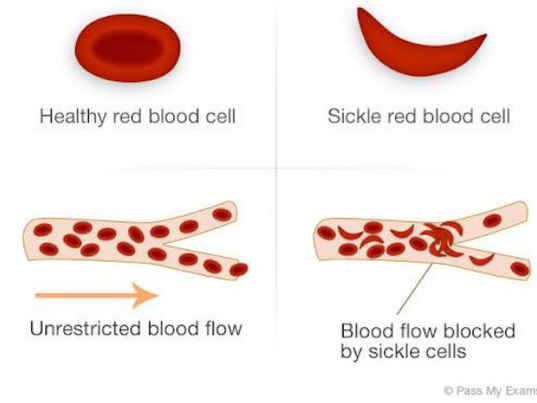
Comparison of Sanger and Next-Gen

Method	Read Length	Samples Processed
Sanger	~1000	1
Illumina	~200	~10,000s

Genotyping at the Molecular Level with DNA Sequencing.

- Sickle cell anemia is caused by a single mutation in the beta chain of hemoglobin
- This mutation causes the hemoglobin to form long polymers that distort the shape of the red blood cell.
- Determining whether someone has the mutation can be useful for treatment.



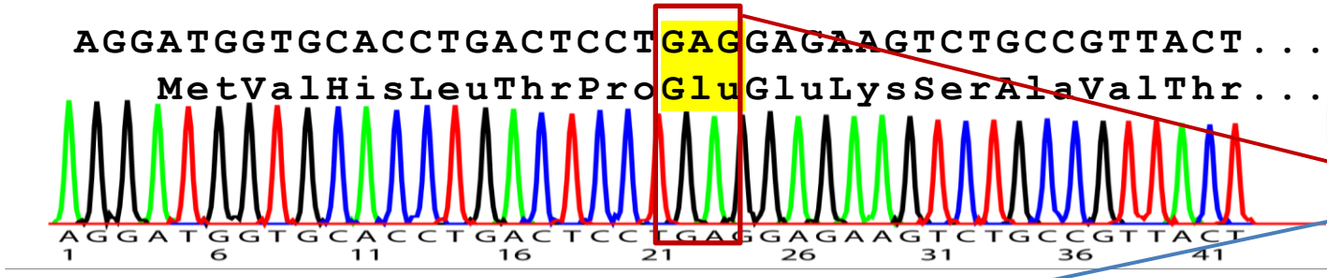


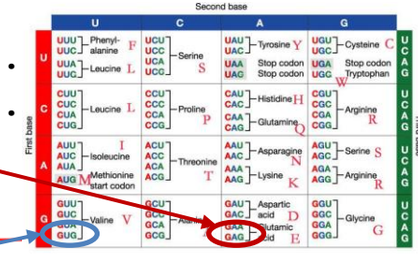
The 5' end of the Hb gene is shown on the right (ATG=start). Using GGTGCCAG as a sequencing primer gives the following sequences for the normal and mutant genes:

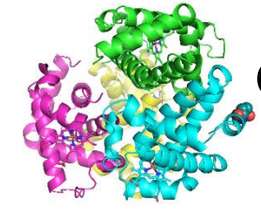
First dd-base added by polymerase

GGTGCCAGAGGATGGTGCACCTGACTCCTGAGGAGAAGTC..
CCACGGTCTCCTACCACGTGGACTGAGGACTCCTCTTCAG..

Sequencing data for the normal beta chain is:

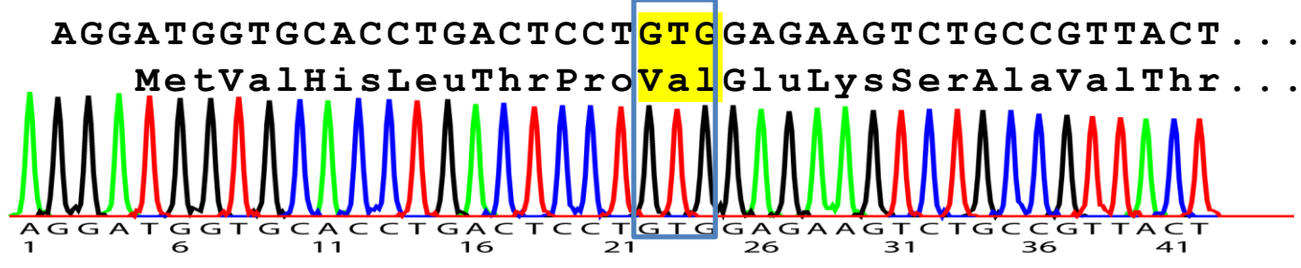






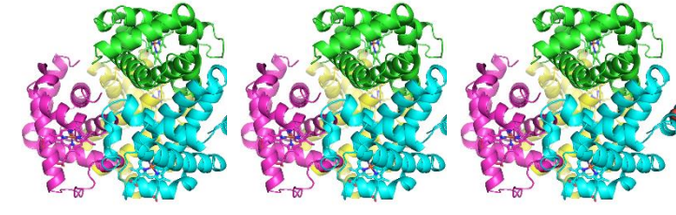
Glu6

Sequencing data for the mutation:



False color code:

A=Green
G=Black
T=Red
C=Blue



Val6

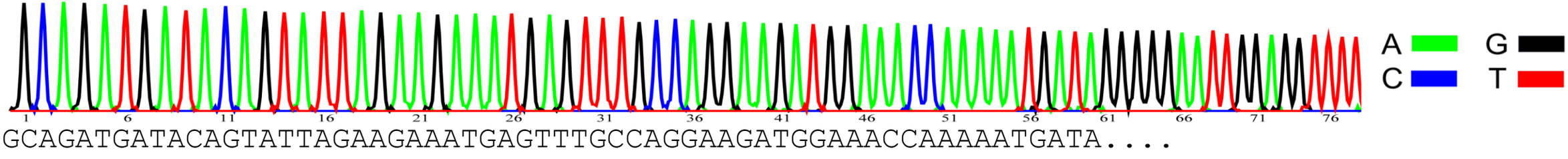
Sequencing Summary & Expectations

Sanger Sequencing:

- Gives the sequence that is complementary to the template strand = “top” strand, same strand at the primer.
- The start of the sequencing information is defined by a primer that anneals to the template (therefore some of the sequence has to be known, how this is done will be described later)
- Dideoxy sequencing is carried out by adding both dideoxynucleotide triphosphates (ddNTPs) and deoxyribonucleotide triphosphates (dNTPs) to the synthesis reactions, at a ratio of 1:100. Most growing chains do not terminate.
- ddNTPs are identical to dNTPs except that they lack the 3' hydroxyl group. Because of the missing 3’-OH, DNA polymerization stops once one ddNTP is added to a growing strand.
- The type of the added base is determined by “color coding” each base.
- The location of added bases is determined by measuring the size of the DNA fragment that was terminated by the ddNTP.
- It is possible to sequence approximately 1000 bases by this method.

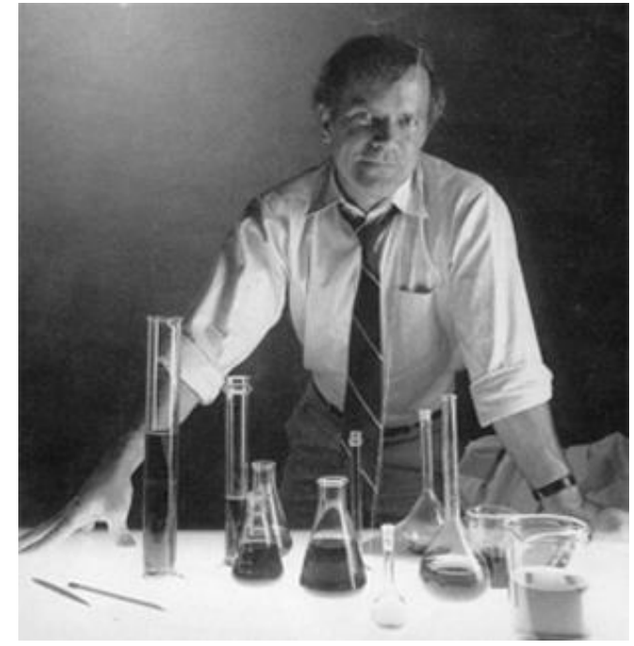
Next Gen-Sequencing:

- Simultaneous sequencing of a large number of fragments
- Shorter “reads” 100 versus 1000 bases/template



Polymerase Chain Reaction - PCR

- In 1983, Kary Mullis developed the molecular biology technique that has since revolutionized genetic research, earning him the Nobel Prize in 1993.
- PCR had an impact on four main areas of biotechnology: gene mapping, cloning, DNA sequencing, and gene detection (e.g. coronavirus).
- PCR is now used as a medical diagnostic tool to detect specific mutations that may cause genetic disease, in criminal investigations and courts of law to identify suspects on a molecular level, and in the sequencing of the human genome.

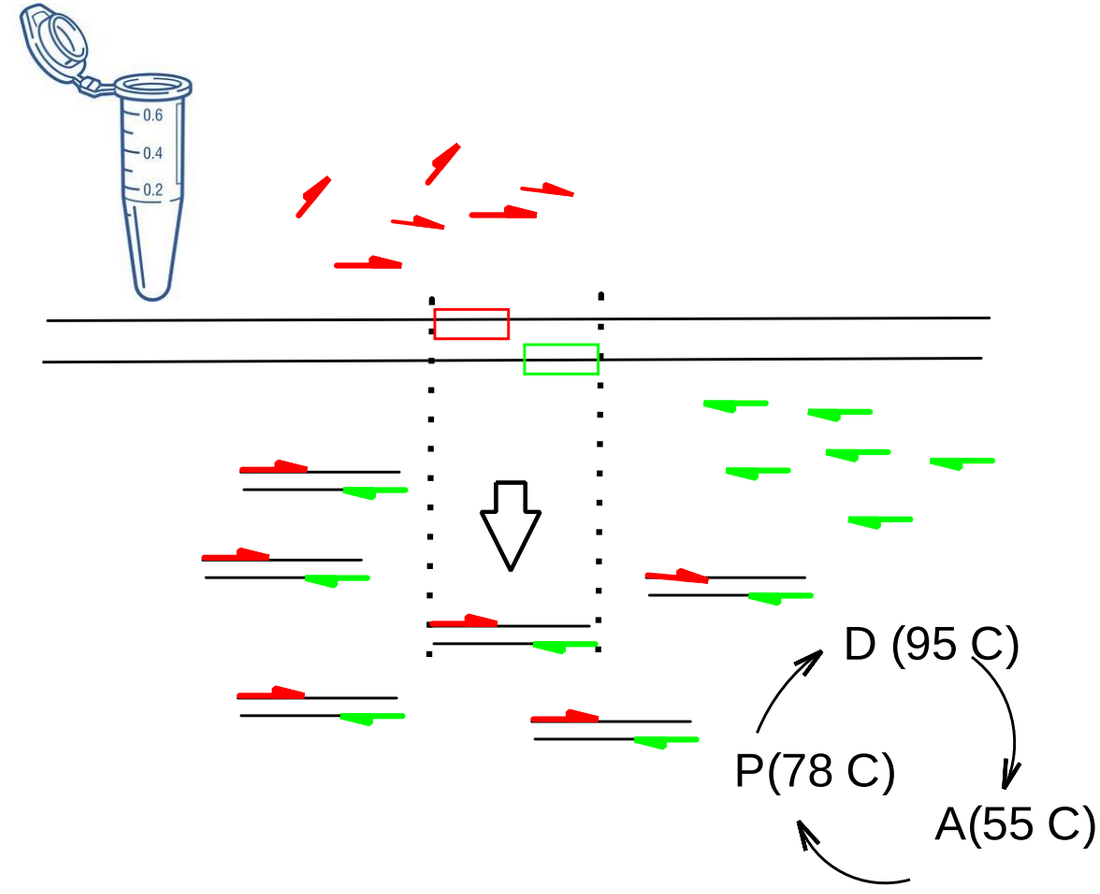


Expectations:

1. Be able to explain how PCR works to amplify a segment of DNA.
2. Be able to give the left and right primers.
3. Apply PCR approaches to determine genotype and detection of viruses.

Polymerase Chain Reaction

- PCR is an *in vitro* DNA synthesis reaction in which a specific section of DNA is replicated over and over generating exponentially large amounts of a specific piece of DNA from trace amounts of starting material (template).
- Template can be trace amounts of DNA from a drop of blood, a single hair follicle, or a cheek cell.
- The region of DNA that is copied is specified by the sequence of two primers, which are short ssDNA that initiate polymerase activity. The primers are in vast excess over the DNA.
- The location of the amplified segment is *defined* by two primers (left = upstream, right = downstream):
  - they anneal to their templates according to Watson-Crick pairing rules (A-T, G-C),
  - initiate polymerization from those sites,
  - they are incorporated into the final PCR product.
  - Left primer = sequence of top strand at left boundary
  - Right primer = sequence of bottom strand at right boundary
- The primers are DNA and are synthesized chemically, they can be any desired sequence.
- If there is no homology between the primers and the input DNA, then no PCR product will be formed.

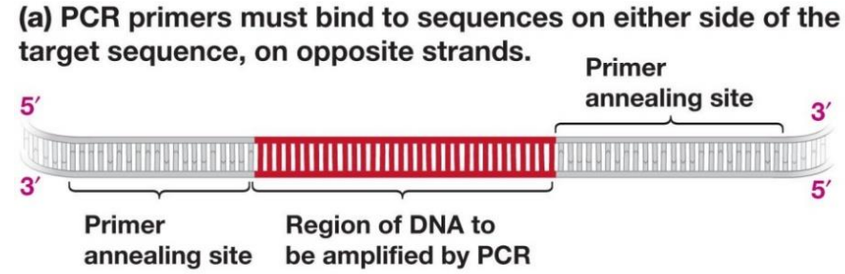


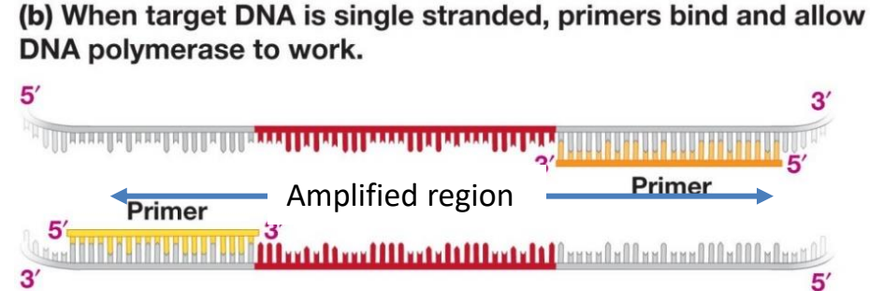
Each PCR cycle consists of three steps:

1. Denaturation of the DNA to make it single stranded (2 min at 98 C)
2. Lowering of temperature to let the primers form double-stranded DNA (1 min at 55 C)
3. Elongation by DNA polymerase (1 min/kb at 78 C)

PCR – Primer Design

- Before a region of DNA can be amplified, one must identify and determine the sequence of a piece of DNA upstream and downstream of the region of interest.
- These areas are then used to determine the sequence of oligonucleotide primers that will be synthesized and used as starting points for DNA replication.
- Primers are complimentary to the up- and down-stream regions of the sequence to be amplified, so they stick, or anneal, to those regions.
  - Left primer = sequence of top strand on the left. This primer will anneal to the bottom strand.
  - Right primer = sequence of bottom strand on the right. This primer will anneal to the top strand.
- Primers are in large excess over the template DNA, they are never used up.
- The primers are incorporated into the final PCR product.





Amplified region

5′--AAG**CTGAC**TAGTCGATGCGAATGTGCGGTGC—3′
3′--TTCGACTGATCAGCTACGCTT**ACACG**CCACG—5′

5′CTGAC

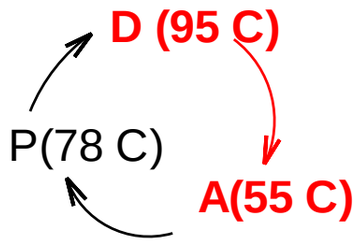
ACACG5′ = 5′GCACA

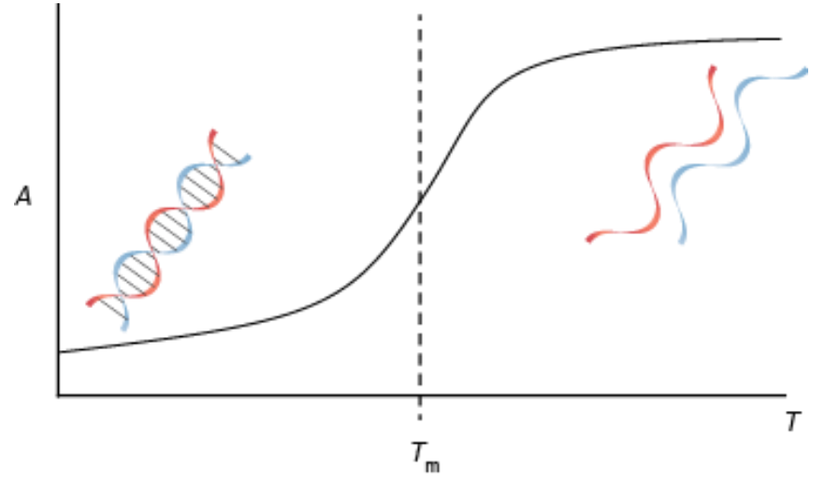
Know these rules!

CTGACTAGTCGATGCGAATGTGC
GACTGATCAGCTACGCTT**ACACG**

Note: Actual primer lengths are 20-30 bases, in the illustrations here and on problem sets, much shorter primers are used.

PCR Step 1 - Thermal Stability of Double Stranded DNA (dsDNA)





5′ A-T-C-A
3′A-T-G-C-C-G-T-A-G-T-C-G-T-A-C-A-G-T-A-C-G-T-G-C-A

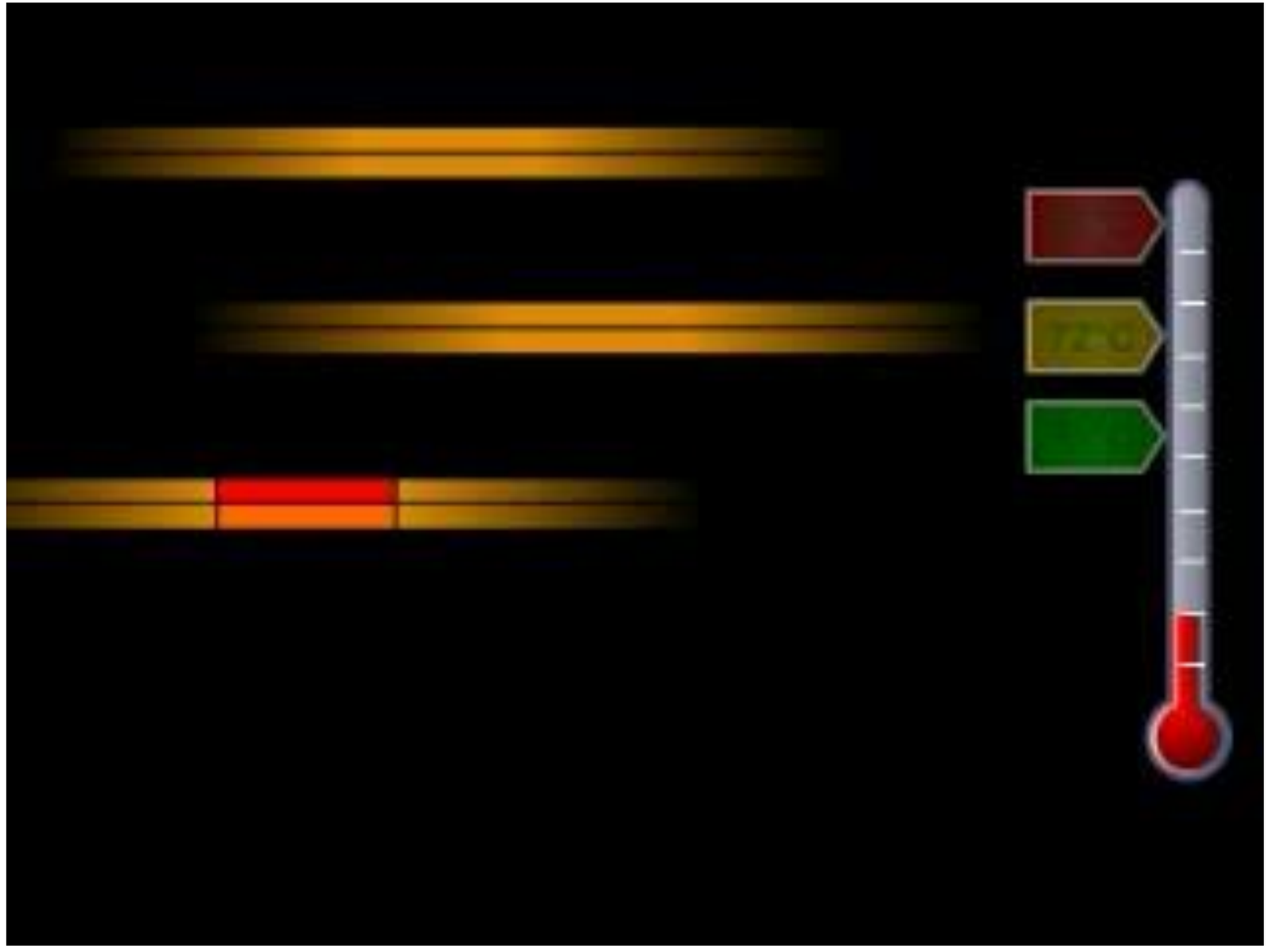
5′ A-T-C-A
3′A-T-G-C-C-G-T-A-G-T-C-G-T-A-C-A-G-T-A-C-G-T-G-C-A

Polymerase Characteristics

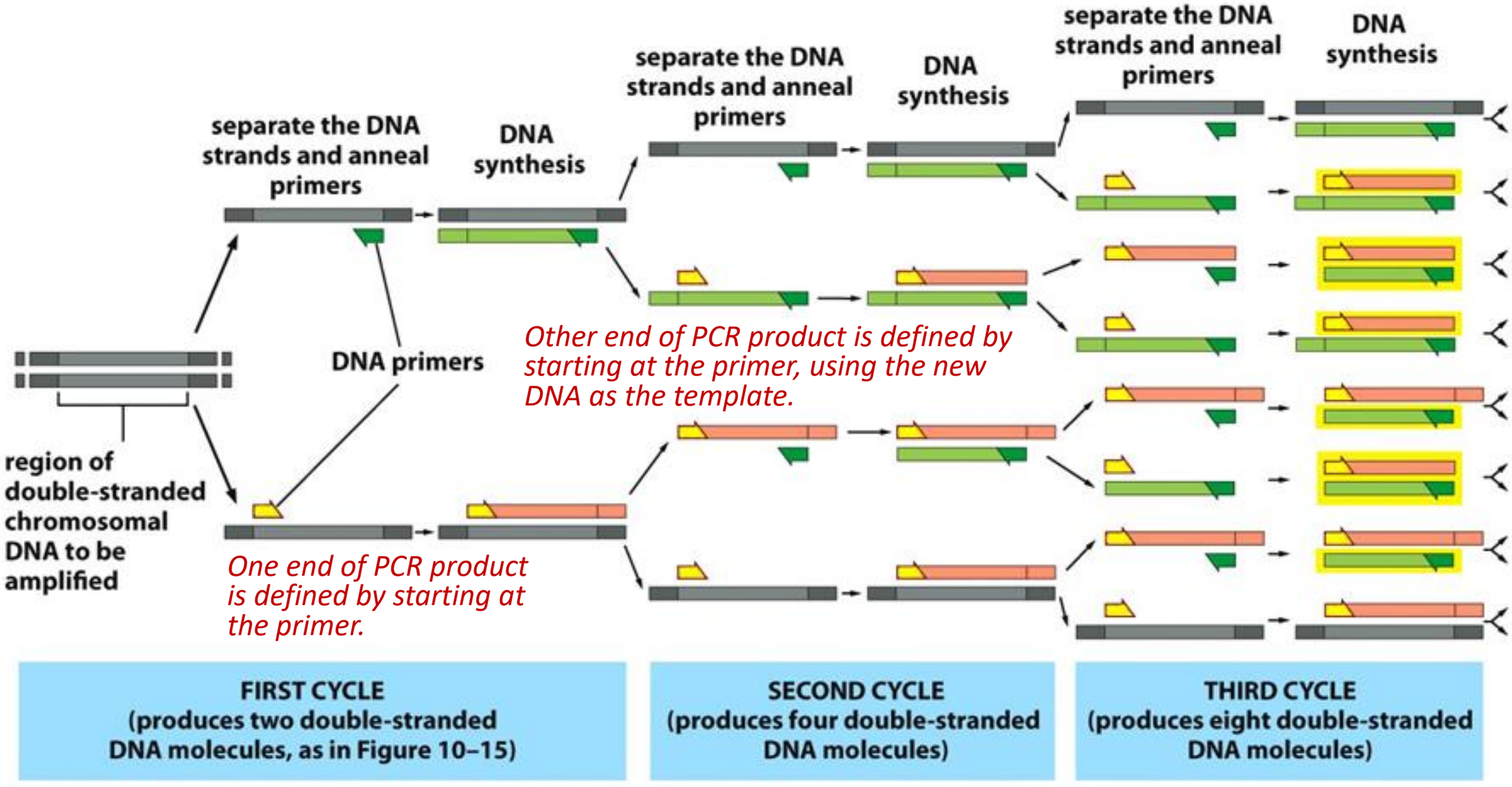
- Since the first step of each cycle (D) requires heating to high temperature, a thermostable polymerase is required.
- The first polymerase, Taq, was isolated from *Thermus Aquaticus,* a bacterial living in hot springs (Yellowstone National Park)
- A number of different polymerases with improved properties have been developed.

Watch Me!

PCR Animation



Three PCR Cycles



After 30 cycles there will be 2^{30}, or over 1 billion times more copies than at the beginning!!!

Detailed Events During First Three PCR Cycles

Cycle I

--AAG**CTGAC**TAGTCGATGCGAATGTGCGGTGC--U
--TTCGACTGATCAGCTACGCTT**ACACG**CCACG--L

Denature & Anneal

--AAG**CTGAC**TAGTCGATGCGAATGTGCGGTGC-U
ACACG5′

5′CTGAC
--TTCGACTGATCAGCTACGCTT**ACACG**CCACG-L

Polymerase

--AAG**CTGAC**TAGTCGATGCGAATGTGCGGTGC-U
<-**TTCGACTGATCAGCTACGCTTACACG5′**

5′**CTGACTAGTCGATGCGAATGTGCGGTGC->**
--TTCGACTGATCAGCTACGCTTACACGCCACG-L

So far - defined one end of the product

Note:

- Polymerization starts at the primer (add to 3'-OH)
- Polymerase always goes to the end of the template.

Final Product

CTGACTAGTCGATGCGAATGTGC
GACTGATCAGCTACGCTT**ACACG**

Cycle II

--AAG**CTGAC**TAGTCGATGCGAATGTGCGGTGC-U
--TTCGACTGATCAGCTACGCTTACACG5′

5′CTGACTAGTCGATGCGAATGTGCGGTGC--
--TTCGACTGATCAGCTACGCTT**ACACG**CCACG-L

Denature & Anneal

--AAG**CTGAC**TAGTCGATGCGAATGTGCGGTGC--

5′CTGAC
--TTCGACTGATCAGCTACGCTTACACG5′ **L strand**

5′CTGAC**TAGTCGATGCGAATGTGCGGTGC- U Strand**

Polymerase **ACACG5′**

--TTCGACTGATCAGCTACGCTT**ACACG**CCACG--

--AAG**CTGAC**TAGTCGATGCGAATGTGCGGTGC--

5′CTGAC**TAGTCGATGCGAATGTGC**
--TTCGACTGATCAGCTACGCTTACACG5′

5′CTGAC**TAGTCGATGCGAATGTGCGGTGC--**
GACTGATCAGCTACGCTTACACG5′

--TTCGACTGATCAGCTACGCTT**ACACG**CCACG--

Now have one strand of the product

Detailed Events during first Three PCR Cycles

Cycle 3

```
    5'CTGACTAGTCGATGCGAATGTGC
--TTCGACTGATCAGCTACGCTTACACG5'

     5'CTGACAGTCGATGCGAATGTGCGGTGC--
       GACTGTCAGCTACGCTTACACG5'
```

Denature & Anneal

```
     5'CTGACTAGTCGATGCGAATGTGC
                       ACACG5'

  --TTCGACTGATCAGCTACGCTTACACG5'

     5'CTGACAGTCGATGCGAATGTGCGGTGC--

     5'CTGAC
       GACTGTCAGCTACGCTTACACG5'
```

Polymerase

```
   5'CTGACTAGTCGATGCGAATGTGC
     GACTGATCAGCTACGCTTACACG5'

--TTCGACTGATCAGCTACGCTTACACG5'

   5'CTGACAGTCGATGCGAATGTGCGGTGC--

   5'CTGACAGTCGATGCGAATGTGC
     GACTGTCAGCTACGCTTACACG5'
```

Now have complete PCR product.
The product will double in each of the following cycles.

Note that the primers are the first bases at the ends of each strand of the PCR product.

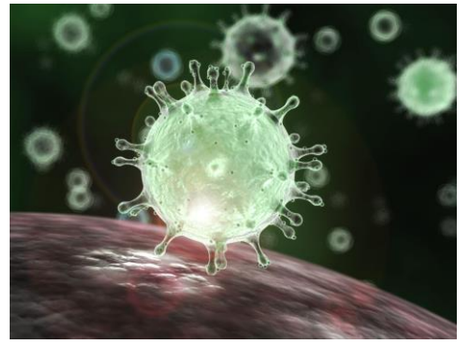
Example – follow the PCR cycles for the following template with primers 5' AATT (left) and 5' GGCC (right)

```
----AATT-------------------------GGCC----
----TTAA-------------------------CCGG----
```

```
AATT-------------------------GGCC
TTAA-------------------------CCGG
```

PCR Applications - Detection of Viruses

Coronavirus



Sequence of Covid-19 (top strand only)

```
1     attaaaggtt tataccttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct
61    gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact
121   cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc
181   ttctgcaggc tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt

28261 cgaacaaact aaaatgtctg ataatggacc ccaaaatcag cgaaatgcac cccgcattac
28321 gtttggtgga ccctcagatt caactggcag taaccagaat ggagaacgca gtggggcgcg
28381 atcaaaacaa cgtcggcccc aaggtttacc caataatact gcgtcttggt tcaccgctct
28441 cactcaacat ggcaaggaag accttaaatt ccctcgagga caaggcgttc caattaacac

29701 gggaggactt gaaagagcca ccacattttc accgaggcca cgcggagtac gatcgagtgt
29761 acagtgaaca atgctaggga gagctgccta tatggaagag ccctaatgtg taaaattaat
29821 tttagtagtg ctatccccat gtgattttaa tagcttctta ggagaatgac aaaaaaaaaa
29881 aaaaaaaaaa aaaaaaaaaa aaa
```

CDC Recommended PCR Primers

2019-Novel Coronavirus (2019-nCoV) Real-time rRT-PCR Panel Primers and Probes

Name	Description	Oligonucleotide Sequence (5'>3')	Label[1]	Working Conc.
2019-nCoV_N1-F	2019-nCoV_N1 Forward Primer	5'-GAC CCC AAA ATC AGC GAA AT-3'	None	20 µM
2019-nCoV_N1-R	2019-nCoV_N1 Reverse Primer	5'-TCT GGT TAC TGC CAG TTG AAT CTG-3'	None	20 µM

dsSeq of above bold & circled region

```
28271 aaaatgtctgataatgGACCCCAAAATCAGCGAAATgcaccccgcattacgtttggtggaccctcagattcaactggcagtaaccagaatggagaacgca
      ttttacagactattacctggggttttagtcgctttacgtggggcgtaatgcaaaccacctgggaGTCTAAGTTGACCGTCATTGGTCTtacctcttgcgt
```

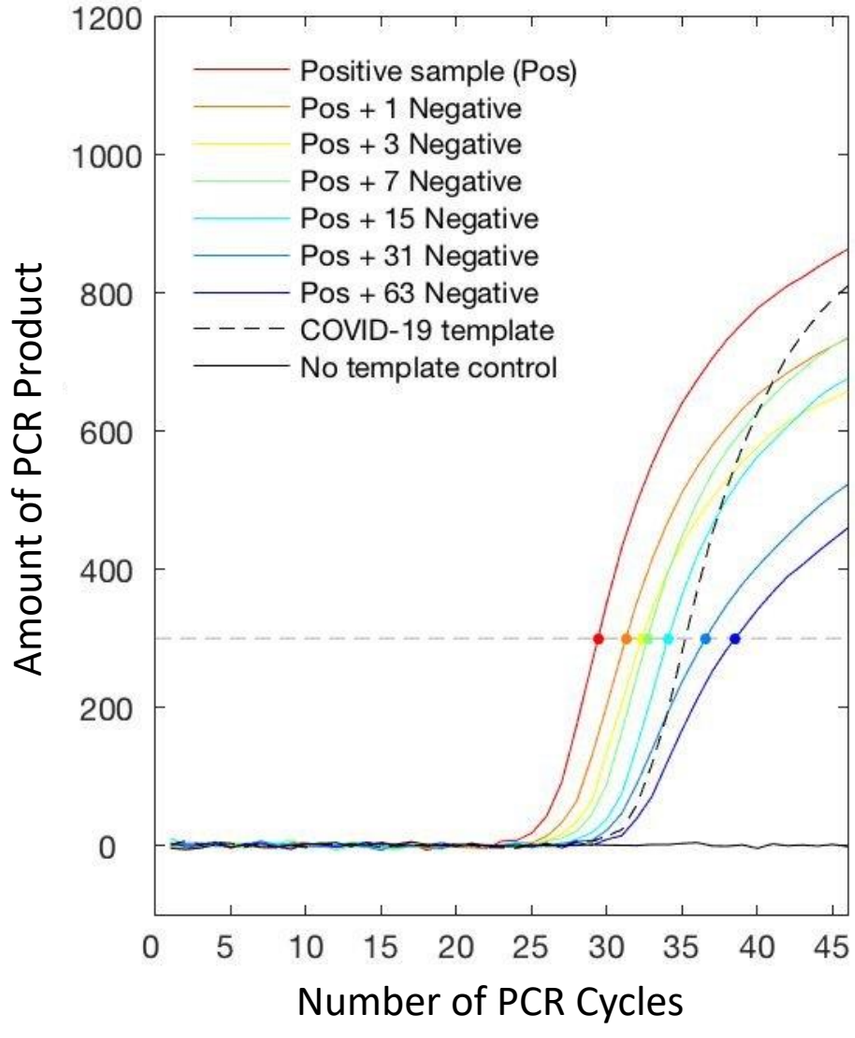
PCR Product

```
GACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAACCAGA
CTGGGGTTTTAGTCGCTTTACGTGGGGCGTAATGCAAACCACCTGGGAGTCTAAGTTGACCGTCATTGGTCT
```

Will PCR generate products if the viral DNA is not present?

Covid 19 PCR Test: Detection of the PCR Product.

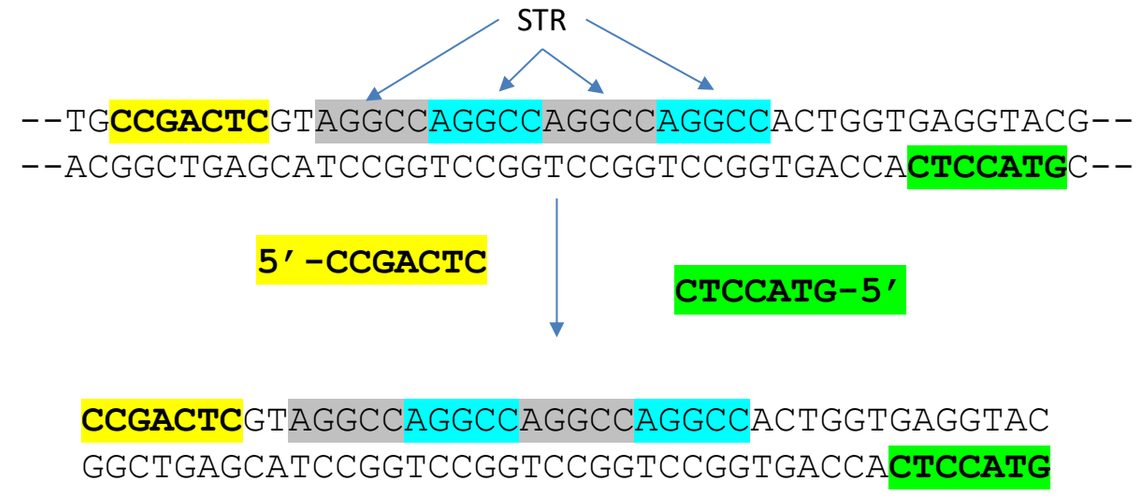
https://www.medrxiv.org/content/10.1101/2020.03.26.20039438v1



- Production of PCR products (double stranded DNA) causes an increase in signal (fluorescence)
- Signal above 300 considered to be positive (dashed gray line)
- Dots represent when a sample crosses the fluorescence threshold.
- Red curve (Positive sample) shows a threshold level of PCR product after 27 cycles.
- Next 6 samples are the positive sample mixed with up to 63 negative samples, showing that it is possible to test pooled samples.
- - - - is a ***positive control*** amount of Covid template. It shows that you can detect a PCR product if the covid genome is present.
- Solid black line is a ***negative control***, no Covid DNA. It shows that addition of covid template will lead to a signal.

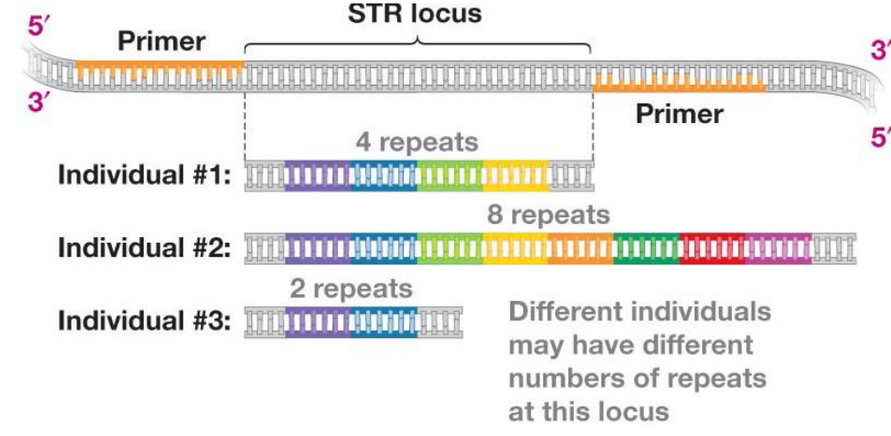
PCR Applications – Identification of Individuals

- Regions of DNA have variable numbers of repeated DNA sequences (Short tandem repeats, STR). The number of STR can differ from one person to the next.
- Individuals will inherit one copy of the repeat from each parent. The length of the inherited DNA can be the same or different.
- PCR Primers are designed to be outside the repeated region, so that they will anneal to a single location on the chromosome and then amplify the region containing the STR
- PCR Product length = primer lengths + number of tandem repeats (+ any DNA between the primers and the repeats). *Individuals can be differentiated by the length of the PCR product if they have different numbers of STR*



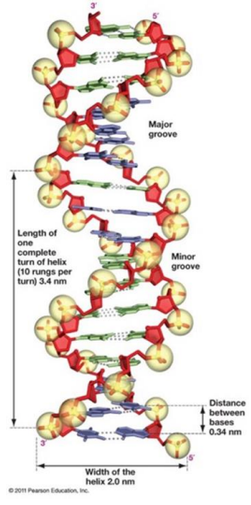
Which individual has the shortest PCR product?

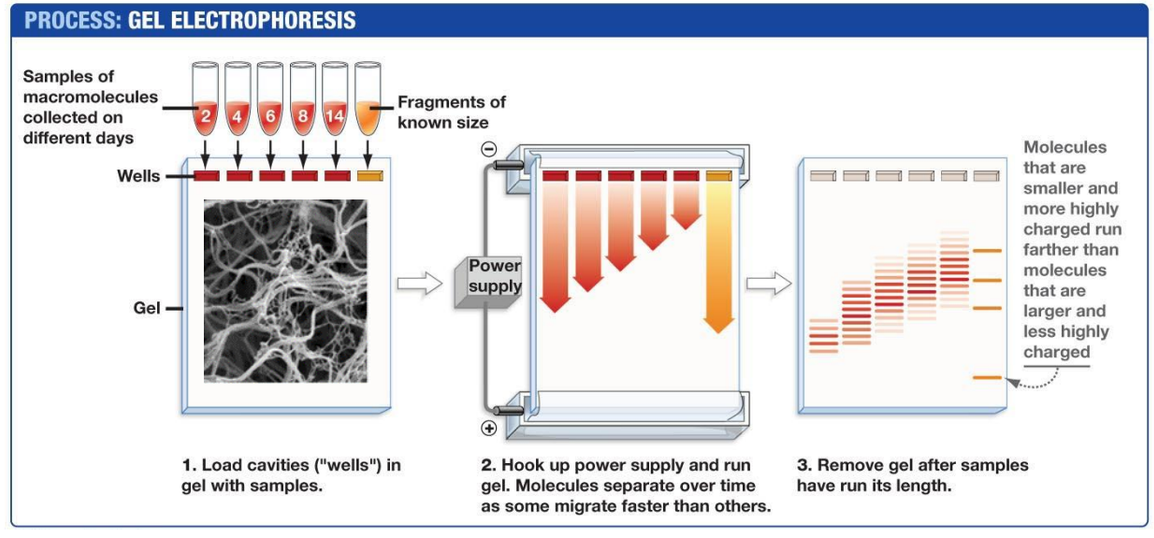
Which has the longest?

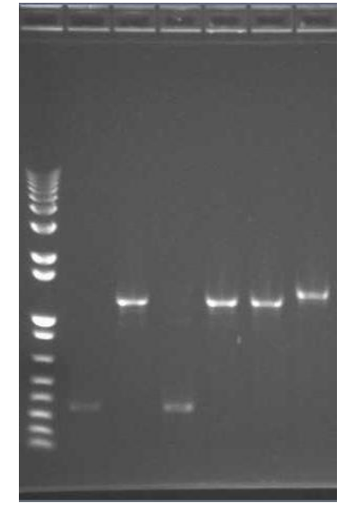


Size Determination of PCR products - Agarose Gel Electrophoresis.

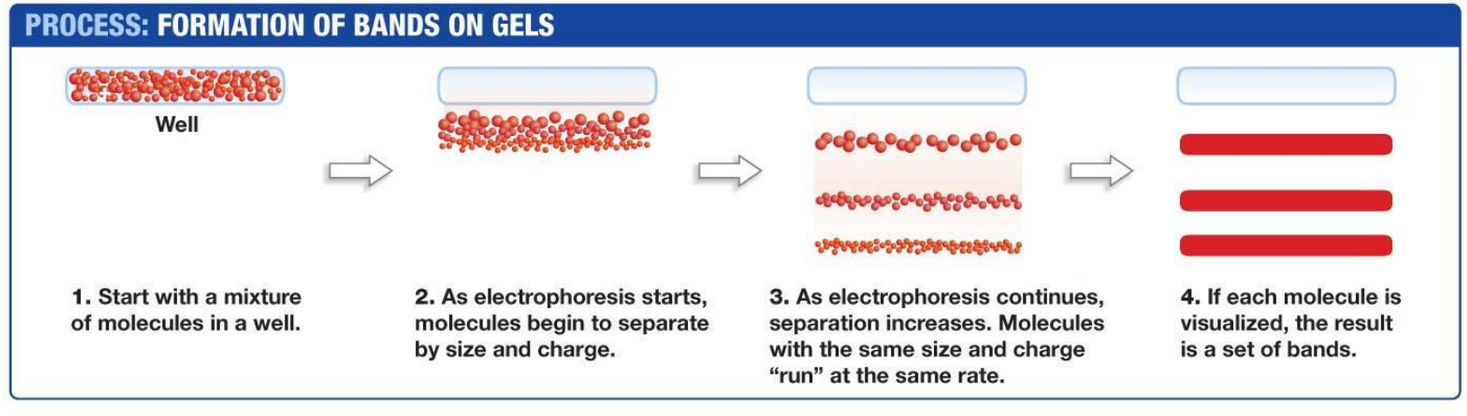
https://dnalc.cshl.edu/resources/animations/gelelectrophoresis.html





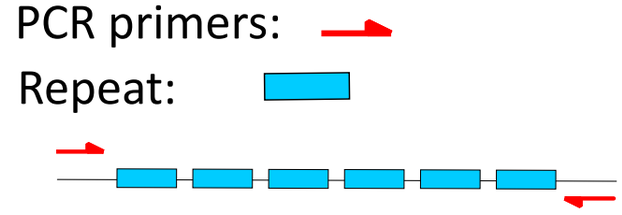


Which are the smallest PCR fragments?



Short Tandem Repeats to Test Paternity

1. DNA samples (blood, cheek cells) would be obtained from:
   - Mother
   - Child
   - Candidate fathers.
2. PCR would be preformed using primers that amplify a segment of the chromosome containing repeats.
3. Each individual would show 2 bands on the gel, corresponding to the PCR product from each chromosome (we have two copies of each chromosome).
4. The child would inherit one copy from the mother and the other from the father:
   - One of the child's PCR product would match one of the mothers.
   - The other PCR product from the child would match one of the PCR products from the father.



Lane 1: Child
Lane 2: Mother

Lanes A, B, C: Possible Fathers

1. *Which PCR product is from the mother? From the father?*

2. *Who is **not** the father?*

3. *Who **may** be the father?*