



Algorithms in Bioinformatics: A Practical Introduction

Introduction to Molecular Biology



Outline

- Cell
- DNA, RNA, Protein
- Genome, Chromosome, and Gene
- Central Dogma (from DNA to Protein)
- Mutation
- List of biotechnology tools
- Brief History of Bioinformatics



Our body

- Our **body** consists of a number of organs
- Each **organ** composes of a number of tissues
- Each **tissue** composes of **cells** of the same type.



Cell

- Cell performs two type of functions:
 - Perform chemical reactions necessary to maintain our life
 - Pass the information for maintaining life to the next generation
- Actors:
 - **Protein** performs chemical reactions
 - **DNA** stores and passes information
 - **RNA** is the intermediate between DNA and proteins

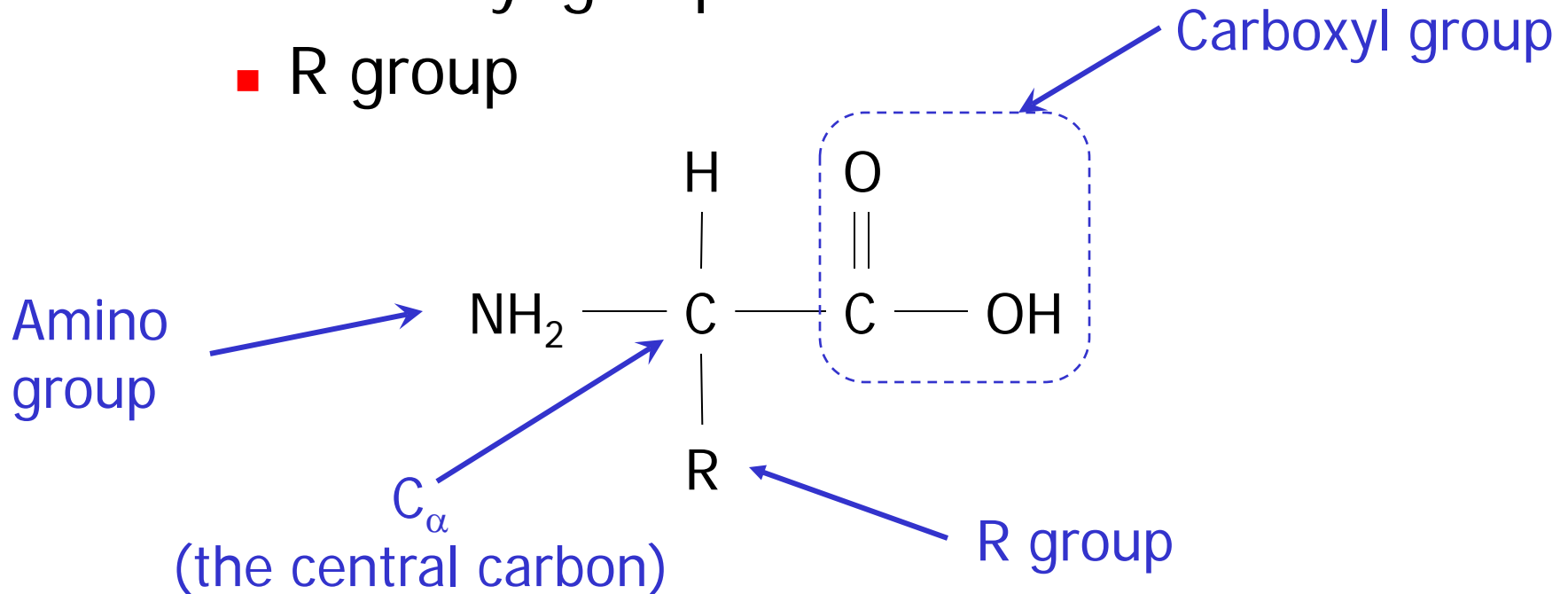


Protein

- Protein is a sequence composed of an alphabet of 20 amino acids.
 - The length is in the range of 20 to more than 5000 amino acids.
 - In average, protein contains around 350 amino acids.
- Protein folds into three-dimensional shape, which form the building blocks and perform most of the chemical reactions within a cell.

Amino acid

- Each amino acid consist of
 - Amino group
 - Carboxyl group
 - R group





Classification of amino acids (I)

- 20 common amino acids can be classified into 4 types.
- Positively charged (basic) amino acids:
 - Arginine (Arg, R)
 - Histidine (His, H)
 - Lysine (Lys, K)
- Negatively charged (acidic) amino acids:
 - Aspartic acid (Asp, D)
 - Glutamic acid (Glu, E)

Classification of amino acids

(II)

- Polar amino acids:
 - Overall uncharged, but uneven charge distribution. Can form hydrogen bonds with water. They are called **hydrophilic**. Often found on the outer surface of a folded protein.
 - Asparagine (Asn, N)
 - Cysteine (Cys, C)
 - Glutamine (Gln, Q)
 - Glycine (Gly, G)
 - Serine (Ser, S)
 - Threonine (Thr, T)
 - Tyrosine (Tyr, Y)

Classification of amino acids (III)

- non-polar amino acids:
 - Overall uncharged and uniform charge distribution. Cannot form hydrogen bonds with water. They are called **hydrophobic**. Tend to appear on the inside surface of a folded protein.
 - Alanine (Ala, A)
 - Isoleucine (Ile, I)
 - Leucine (Leu, L)
 - Methionine (Met, M)
 - Phenylalanine (Phe, F)
 - Proline (Pro, P)
 - Tryptophan (Trp, W)
 - Valine (Val, V)

Summary of the amino acid properties

Amino Acid	1-Letter	3-Letter	Avg. Mass (Da)	volume	Side chain polarity	Side chain acidity or basicity	Hydropathy index
Alanine	A	Ala	89.09404	67	non-polar	Neutral	1.8
Cysteine	C	Cys	121.15404	86	polar	basic (strongly)	-4.5
Aspartic acid	D	Asp	133.10384	91	polar	Neutral	-3.5
Glutamic acid	E	Glu	147.13074	109	polar	acidic	-3.5
Phenylalanine	F	Phe	165.19184	135	polar	neutral	2.5
Glycine	G	Gly	75.06714	48	polar	acidic	-3.5
Histidine	H	His	155.15634	118	polar	neutral	-3.5
Isoleucine	I	Ile	131.17464	124	non-polar	neutral	-0.4
Lysine	K	Lys	146.18934	135	polar	basic (weakly)	-3.2
Leucine	L	Leu	131.17464	124	non-polar	neutral	4.5
Methionine	M	Met	149.20784	124	non-polar	neutral	3.8
Asparagine	N	Asn	132.11904	96	polar	basic	-3.9
Proline	P	Pro	115.13194	90	non-polar	neutral	1.9
Glutamine	Q	Gln	146.14594	114	non-polar	neutral	2.8
Arginine	R	Arg	174.20274	148	non-polar	neutral	-1.6
Serine	S	Ser	105.09344	73	polar	neutral	-0.8
Threonine	T	Thr	119.12034	93	polar	neutral	-0.7
Valine	V	Val	117.14784	105	non-polar	neutral	-0.9
Tryptophan	W	Trp	204.22844	163	polar	neutral	-1.3
Tyrosine	Y	Tyr	181.19124	141	non-polar	neutral	4.2

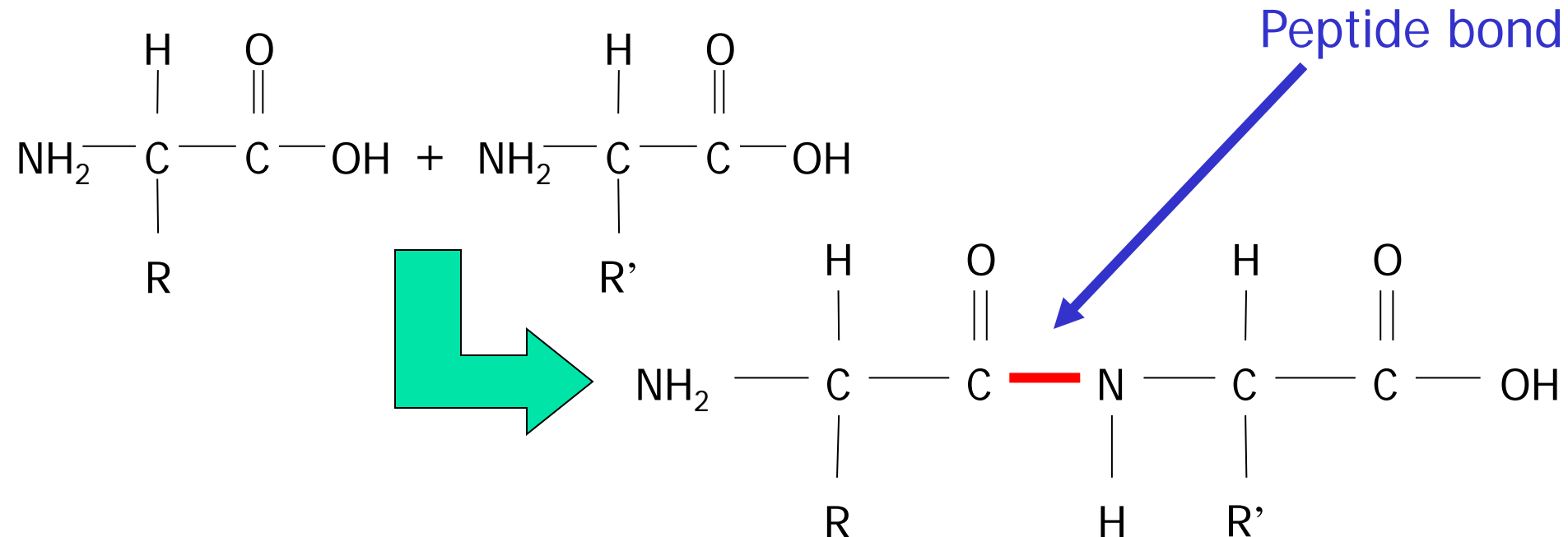


Nonstandard amino acids

- Two non-standard amino acids which can be specified by genetic code:
 - Selenocysteine is incorporated into some proteins at a UGA codon, which is normally a stop codon.
 - Pyrrolysine is used by some methanogenic archaea in enzymes that they use to produce methane. It is coded for with the codon UAG.
- Non-standard amino acids which do not appear in protein:
 - E.g. lanthionine, 2-aminoisobutyric acid, and dehydroalanine
 - They often occur as intermediates in the metabolic pathways for standard amino acids
- Non-standard amino acids which are formed through modification to the R-groups of standard amino acids:
 - E.g. hydroxyproline is made by a posttranslational modification of proline.

Polypeptide

- Protein or polypeptide chain is formed by joining the amino acids together via a peptide bond.
- One end of the polypeptide is the amino group, which is called N-terminus. The other end of the polypeptide is the carboxyl group, which is called C-terminus.





Protein structure

- Primary structure
 - The amino acid sequence
- Secondary structure
 - The local structure formed by hydrogen bonding: α -helices and β -sheets.
- Tertiary structure
 - The interaction of α -helices and β -sheets due to hydrophobic effect
- Quaternary structure
 - The interaction of more than one protein to form protein complex

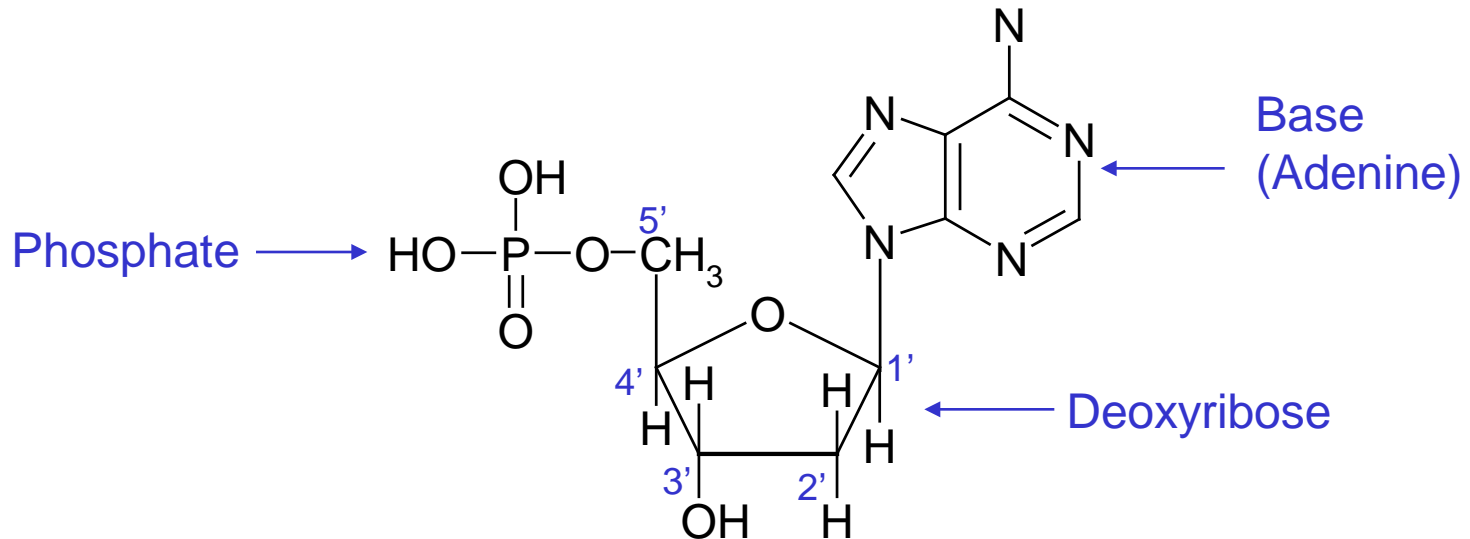


DNA

- DNA stores the instruction needed by the cell to perform daily life function.
- It consists of two strands which interwoven together and form a double helix.
- Each strand is a chain of some small molecules called nucleotides.

Nucleotide for DNA

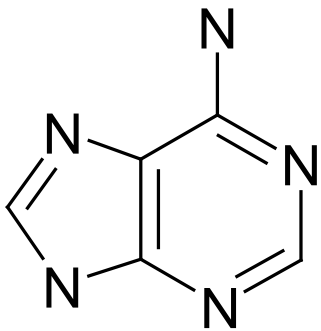
- Nucleotide consists of three parts:
 - Deoxyribose
 - Phosphate (bound to the 5' carbon)
 - Base (bound to the 1' carbon)



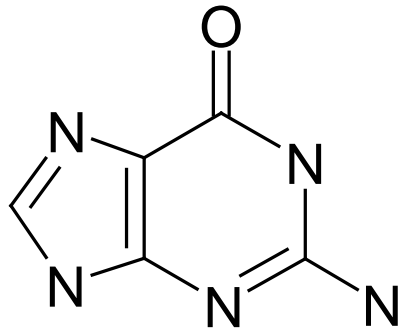


More on bases

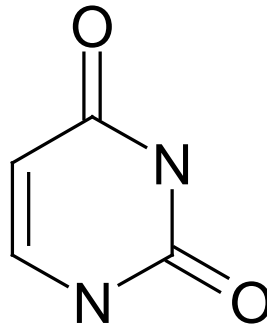
- There are 5 different nucleotides: adenine(A), cytosine(C), guanine(G), thymine(T), and uracil(U).
- A, G are called **purines**. They have a 2-ring structure.
- C, T, U are called **pyrimidines**. They have a 1-ring structure.
- DNA only uses A, C, G, and T.



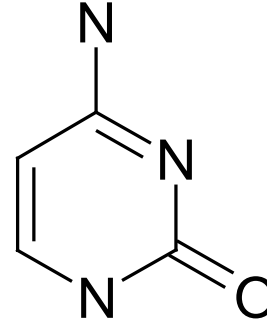
Adenine



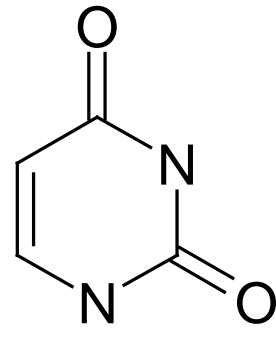
Guanine



Thymine



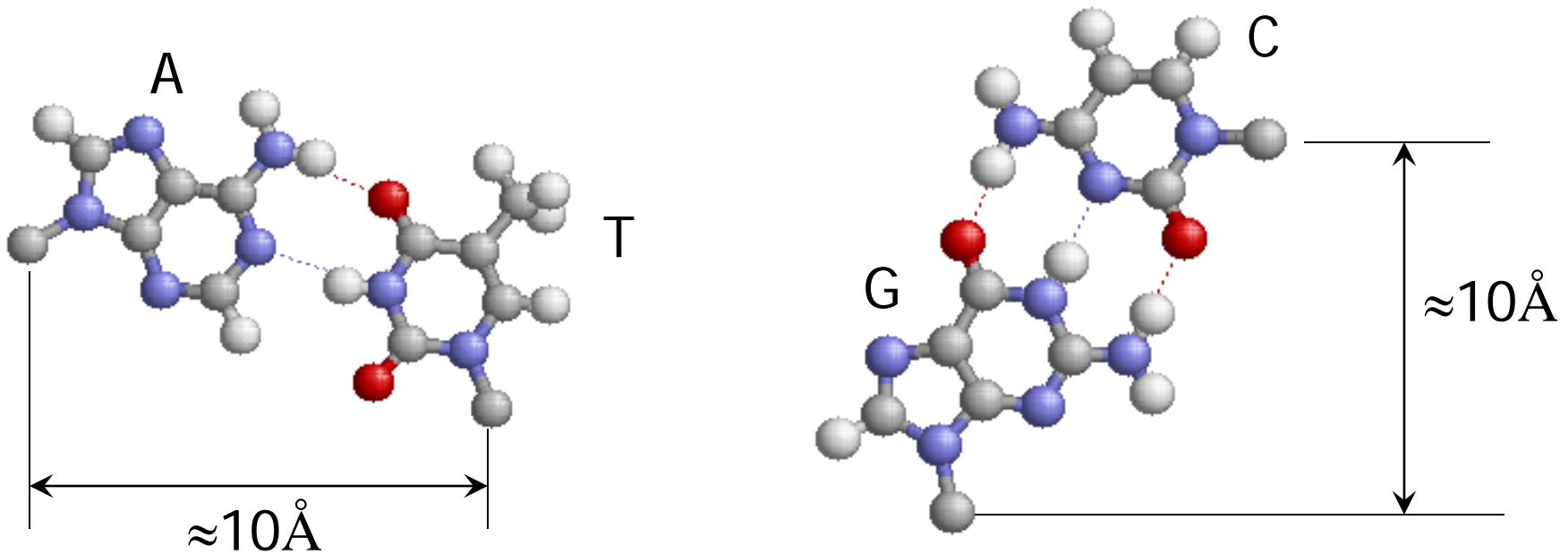
Cytosine

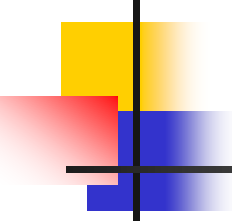


Uracil

Watson-Crick rules

- Complementary bases:
 - A with T (two hydrogen-bonds)
 - C with G (three hydrogen-bonds)



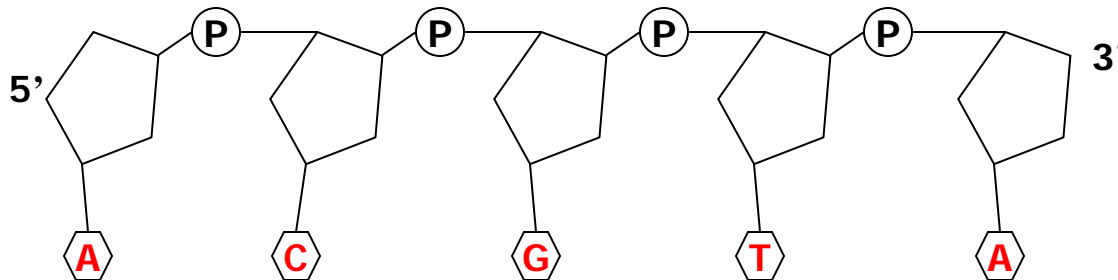


Reasons behind the complementary bases

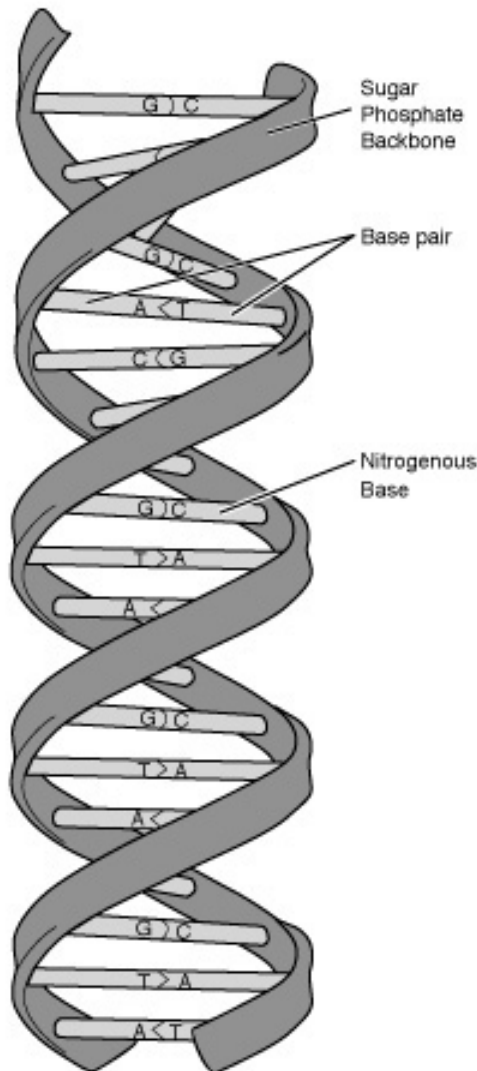
- Purines (A or G) cannot pair up because they are too big
- Pyrimidines (C or T) cannot pair up because they are too small
- G and T (or A and C) cannot pair up because they are chemically incompatible

Orientation of a DNA

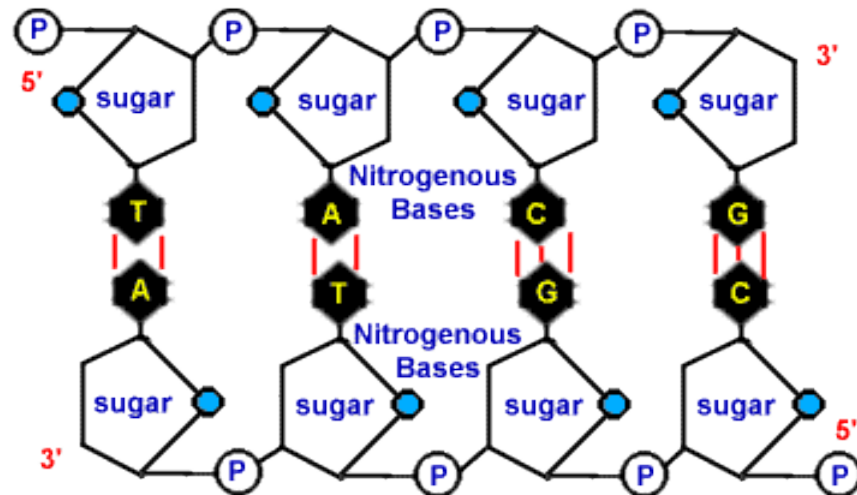
- One strand of DNA is generated by chaining together nucleotides.
- It forms a phosphate-sugar backbone.
- It has direction: from 5' to 3'. (Because DNA always extends from 3' end.)
- Upstream: from 5' to 3'
- Downstream: from 3' to 5'



Double stranded DNA



- Normally, DNA is double stranded within a cell. The two strands are antiparallel. One strand is the **reverse complement** of another one.
- The double strands are interwoven together and form a double helix.
- One reason for double stranded is that it eases DNA replicate.





Circular form of DNA

- DNA usually exists in linear form
 - E.g. in human, yeast, exists in linear form
- In some simple organism, DNA exists in circular form.
 - E.g. in E. coli, exists in circular form



What is the locations of DNAs in a cell?

- Two types of organisms: Prokaryotes and Eukaryotes.
- In **Prokaryotes**: single celled organisms with no nuclei (e.g. bacteria)
 - DNA swims within the cell
- In **Eukaryotes**: organisms with single or multiple cells. Their cells have nuclei. (e.g. plant and animal)
 - DNA locates within the nucleus.



Some terms related to DNA

- Genome
- Chromosome
- Gene



Chromosome

- Usually, a DNA is tightly wound around **histone** proteins and forms a **chromosome**.
- The total information stored in all chromosomes constitute a **genome**.
- In most multi-cell organisms, every cell contains the same complete set of genome.
 - May have some small different due to mutation
- Example:
 - Human Genome: has 3G base pairs, organized in 23 pairs of chromosomes



Gene

- A **gene** is a sequence of DNA that encodes a protein or an RNA molecule.
- In human genome, it is expected there are 30,000 – 35,000 genes.
- For gene that encodes protein,
 - In Prokaryotic genome, one gene corresponds to one protein
 - In Eukaryotic genome, one gene can correspond to more than one protein because of the process “alternative splicing” (discuss later!)

Complexity of the organism vs. genome size

- Human Genome: 3G base pairs
- Amoeba dubia (a single cell organism): 670G base pairs



- Thus, genome size has no relationship with the complexity of the organism

Number of genes vs. genome size

- Prokaryotic genome: E.g. E. coli
 - Number of base pairs: 5M
 - Number of genes: 4k
 - Average length of a gene: 1000 bp
- Eukaryotic genome: E.g. Human
 - Number of base pairs: 3G
 - Estimated number of genes: 20k – 30k
 - Estimated average length of a gene: 1000-2000 bp
- Note that 90% of the E. coli genome consists of coding regions.
- Less than 3% of the human genome is believed to be coding regions. The rest is called **junk DNA**.
- Thus, for Eukaryotic genome, the genome size has no relationship with the number of genes!

Note that before 2001, the people think we have 100000 genes

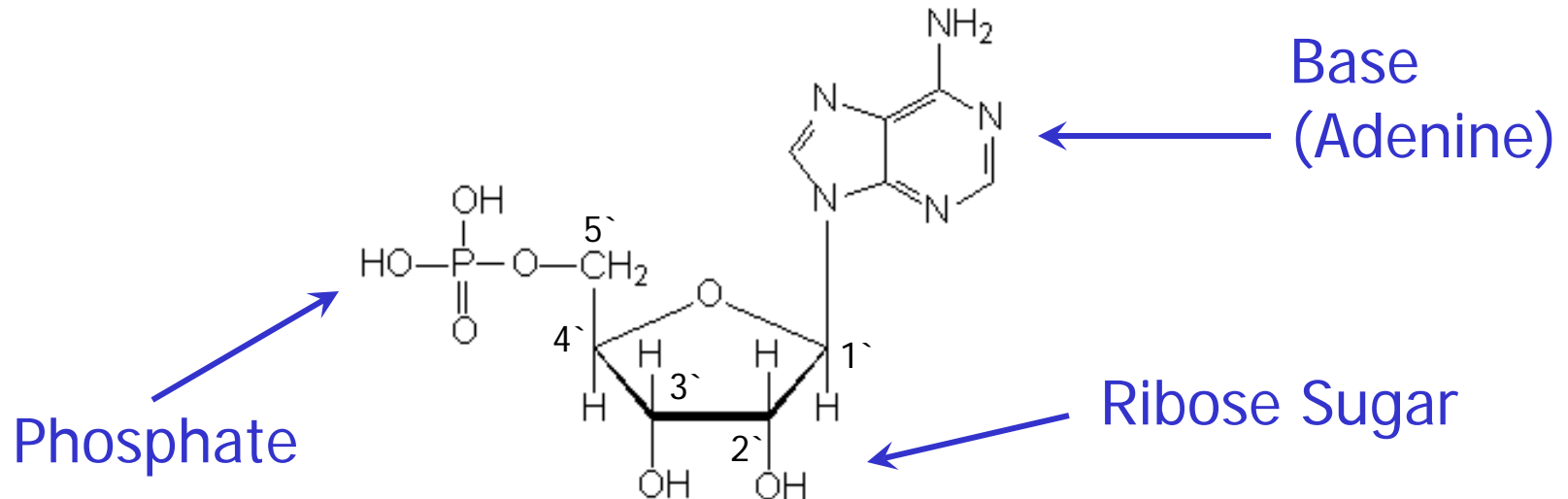


RNA

- RNA has both the properties of DNA and protein
 - Similar to DNA, it can store and transfer information
 - Similar to protein, it can form complex 3-dimensional structure and perform some functions.

Nucleotide for RNA

- Nucleotide consists of three parts:
 - Ribose Sugar (has an extra OH group at 2')
 - Phosphate (bound to the 5' carbon)
 - Base (bound to the 1' carbon)





RNA vs DNA

- RNA is single stranded.
- The nucleotides of RNA are quite similar to that of DNA, except that it has an extra OH at position 2'. (see previous slide!)
 - Due to this extra OH, it can form more hydrogen bonds than DNA. Thus, RNA can form complexity 3-dimensional structure.
- RNA use the base U instead of T.
 - U is chemically similar to T. In particular, U is also complementary to A.

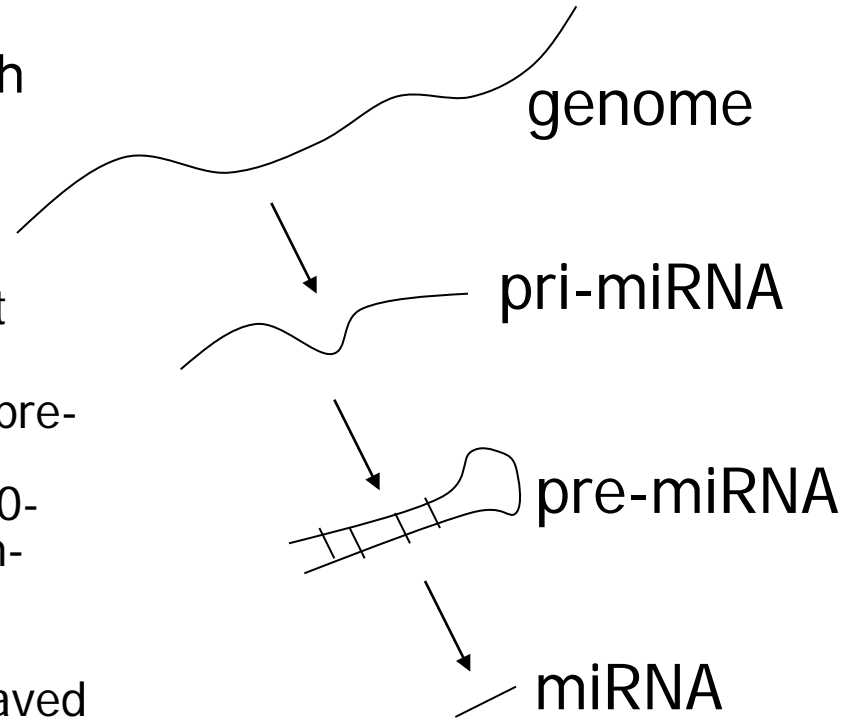


Non-coding RNA

- transfer RNA (tRNA)
- ribosomal RNA (rRNA)
- small RNAs including
 - snoRNAs
 - microRNAs
 - siRNAs
 - piRNAs
- long ncRNAs
 - Examples: Xist, Evf, Air, CTN and PINK
 - People expected there are over 30k long ncRNAs.

microRNA (miRNA) (I)

- miRNA is a single-stranded RNA of length ~22.
- Its formation is as follows:
 - miRNA is encoded as a non-coding RNA.
 - It first transcribed as a primary transcript called primary miRNA (pri-miRNA).
 - It then cleaved into a precursor miRNA (pre-miRNA) with the help of the nuclease Drosha. Precursor miRNA is of length ~60-80 nt and can potentially fold into a stem-loop structure.
 - The pre-miRNA is transported into the cytoplasm by Exportin 5. It is further cleaved into a mature miRNA by the endonuclease Dicer.





RNA interference

- Suppose an miRNA is partially complementary to an mRNAs.
- When miRNA is integrated with the RNA-induced silencing complex (RISC),
 - It down-regulate the mRNA by either translational repression or mRNA cleavage.
- Naturally, RNA interference are used
 - as a cell defense mechanism that represses the expression of viral genes.
 - to regulate development
- We now apply it to knockdown our gene targets.
- In 2006, Andrew Fire and Craig C. Mello shared the Nobel Prize in Physiology or Medicine for their work on RNA interference in *C. elegans*.



Replicate or Repair of DNA

- DNA is double stranded.
- When the cells divide,
 - DNA needs to be duplicated and passes to the two daughter cells.
 - With the help of DNA polymerase, the two strands of DNA serve as template for the synthesis of another complementary strands, generating two identical double stranded DNAs for the two daughter cells.
- When one strand is damaged,
 - it is repaired with the information of another strand.

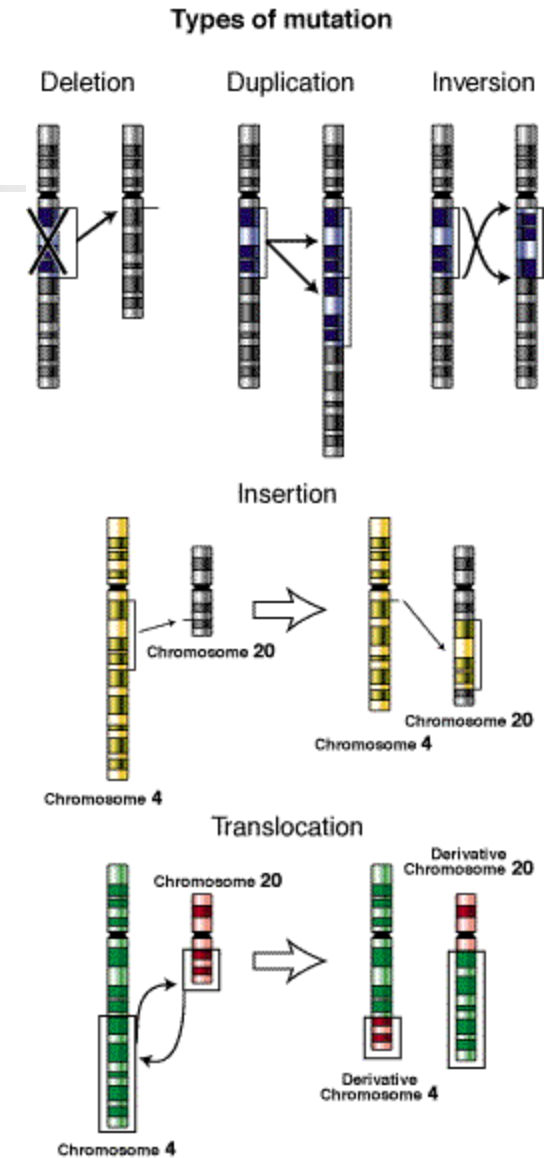


Mutation

- Despite the near-perfect replication, infrequent unrepaired mistakes are still possible.
 - Those mistakes are called **mutations**.
- Occasionally, some mutations make the cells or organisms survive better in the environment.
 - The selection of the fittest individuals to survive is called **natural selection**.
- Mutation and natural selection have resulted in the evolution of a diversified organisms.

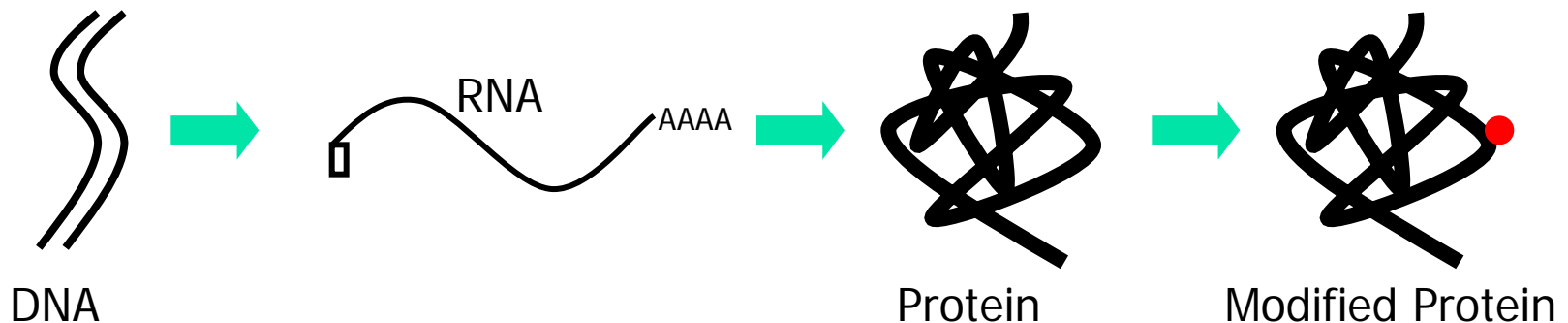
Mutation

- **Mutation** is the change of genome by sudden
- It is the basis of evolution
- It is also the cause of cancer
- Note: mutation can occur in DNA, RNA, and Protein

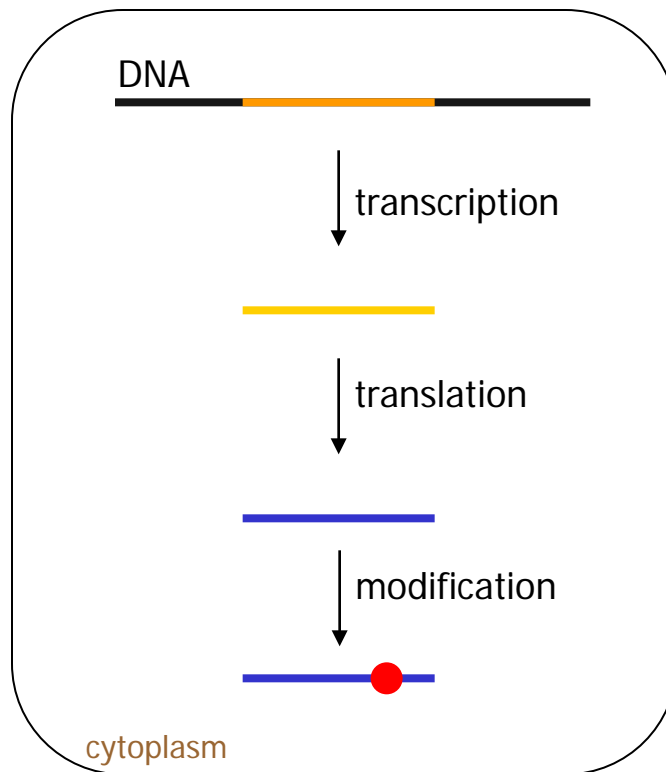


Central Dogma

- Central Dogma tells us how we get the protein from the gene. This process is called **gene expression**.
- The expression of gene consists of two steps
 - **Transcription**: DNA → mRNA
 - **Translation**: mRNA → Protein
 - **Post-translation Modification**: Protein → Modified protein



Central Dogma for Prokaryotes





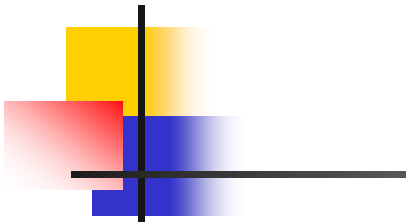
Transcription (Prokaryotes)

- Synthesize a piece of RNA (**messenger RNA, mRNA**) from one strand of the DNA gene.
 1. An enzyme RNA polymerase temporarily separates the double-stranded DNA
 2. It begins the transcription at the transcription start site.
 3. $A \rightarrow A$, $C \rightarrow C$, $G \rightarrow G$, and $T \rightarrow U$
 4. Once the RNA polymerase reaches the transcription start site, transcription stop.



Translation

- Translation synthesizes a protein from a mRNA.
- In fact, each amino acids are encoded by consecutive sequences of 3 nucleotides, called **codon**.
- The decoding table from codon to amino acid is called **genetic code**.
- Note:
 - There are $4^3=64$ different codons. Thus, the codons are not one-to-one correspondence to the 20 amino acids.
 - All organisms use the same decoding table!
 - The codons that encode the same amino acid tend to have the same first and second nucleotide.
 - Recall that amino acids can be classified into 4 groups. A single base change in a codon is usually not sufficient to cause a codon to code for an amino acid in different group.



Genetic code

- Start codon: ATG (also code for M)
- Stop codon: TAA, TAG, TGA

	T	C	A	G	
T	TTT Phe [F] TTC Phe [F] TTA Leu [L] TTG Leu [L]	TCT Ser [S] TCC Ser [S] TCA Ser [S] TCG Ser [S]	TAT Tyr [Y] TAC Tyr [Y] TAA Ter [end] TAG Ter [end]	TGT Cys [C] TGC Cys [C] TGA Ter [end] TGG Trp [W]	T C A G
C	CTT Leu [L] CTC Leu [L] CTA Leu [L] CTG Leu [L]	CCT Pro [P] CCC Pro [P] CCA Pro [P] CCG Pro [P]	CAT His [H] CAC His [H] CAA Gln [Q] CAG Gln [Q]	CGT Arg [R] CGC Arg [R] CGA Arg [R] CGG Arg [R]	T C A G
A	ATT Ile [I] ATC Ile [I] ATA Ile [I] ATG Met [M]	ACT Thr [T] ACC Thr [T] ACA Thr [T] ACG Thr [T]	AAT Asn [N] AAC Asn [N] AAA Lys [K] AAG Lys [K]	AGT Ser [S] AGC Ser [S] AGA Arg [R] AGG Arg [R]	T C A G
G	GTT Val [V] GTC Val [V] GTA Val [V] GTG Val [V]	GCT Ala [A] GCC Ala [A] GCA Ala [A] GCG Ala [A]	GAT Asp [D] GAC Asp [D] GAA Glu [E] GAG Glu [E]	GGT Gly [G] GGC Gly [G] GGA Gly [G] GGG Gly [G]	T C A G



Codon usage

- All but 2 amino acids (W and M) are coded by more than one codon.
- S is coded by 6 different codons.
- Different organisms often prefers one particular codon to encode a particular amino acid.
- For *S. pombe*, *C. elegans*, *D. melanogaster*, and many unicellular organisms,
 - highly expressed genes, such as those encoding ribosomal proteins, have biased patterns of codon usage.
 - People expected that such biase is to enhance the efficiency of translation.



More on Gene Structure



- Gene has 4 regions
 - **Coding region** contains the codons for protein. It is also called open reading frame. Its length is a multiple of 3. It must begin with start codon, end with end codon, and the rest of its codons are not a end codon.
 - **mRNA transcript** contains 5' untranslated region + coding region + 3' untranslated region
 - **Regulatory region** contains promoter, which regulate the transcription process.

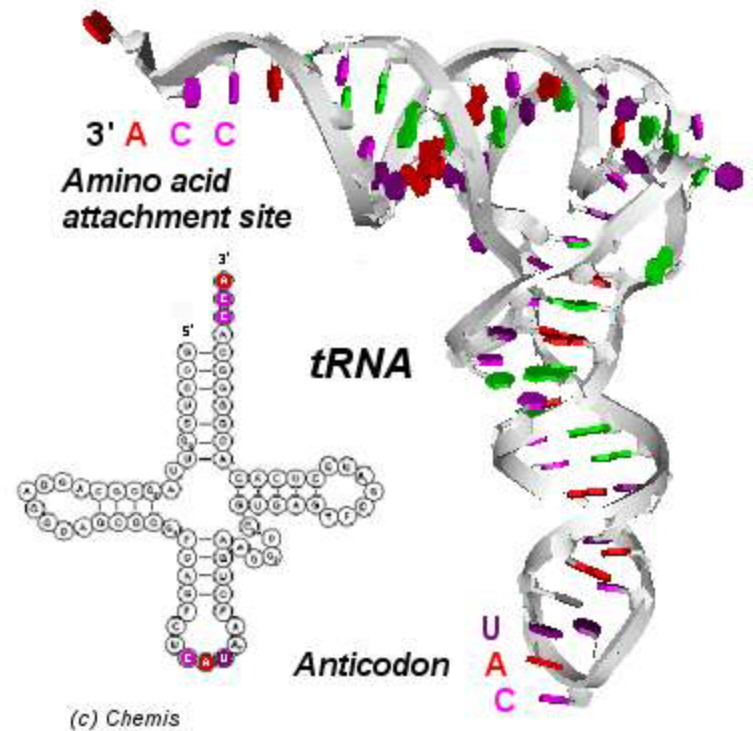


The translation process

- The translation process is handled by a molecular complex **ribosome** which consists of both proteins and ribosomal RNA (rRNA)
 1. Ribosome read mRNA and the translation starts around start codon (translation start site)
 2. With the help of **tRNA**, each codon is translated to an amino acid
 3. The translation stop once ribosome read the stop codon (translation stop site)

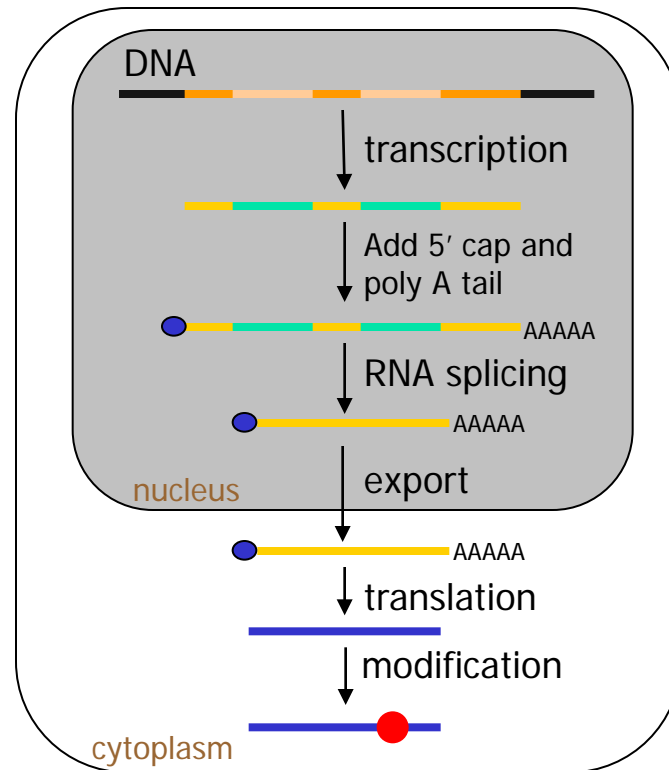
More on tRNA

- tRNA --- transfer RNA
- There are 61 different tRNAs, each correspond to a nontermination codon
- Each tRNA folds to form a cloverleaf-shaped structure
 - One side holds an **anticodon**
 - The other side holds the appropriate amino acid



Central Dogma for Eucaryotes

- Transcription is done within nucleus
- Translation is done outside nucleus



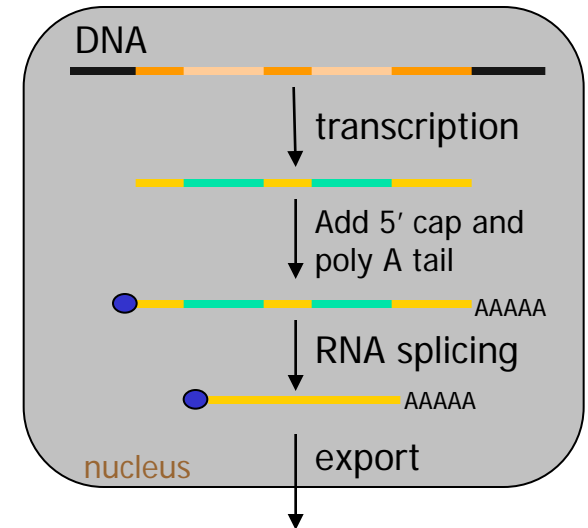


Introns and exons

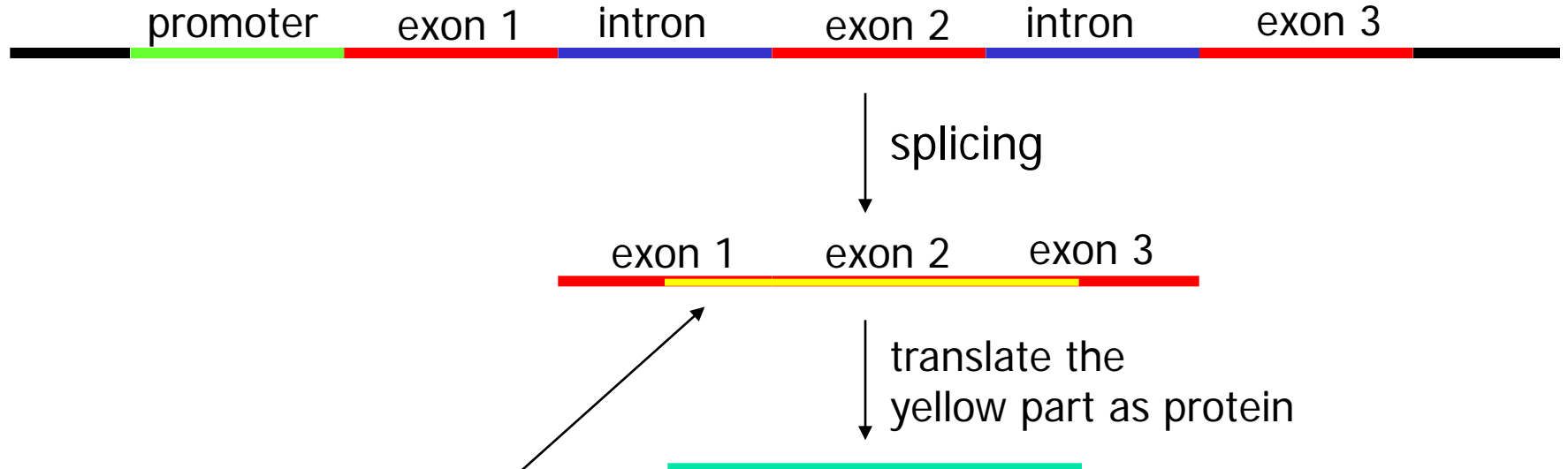
- Eukaryotic genes contain **introns** and **exons**.
 - Introns are sequences that ultimately will be spliced out of the mRNA
 - Introns normally satisfies the GT-AG rule, that is, intron begins with GT and end with AG.
 - Each gene can have many introns and each intron may has thousands bases.
- Introns can be very long. An extreme example (gene that associated with the disease cystic fibrosis in humans):
 - With 24 introns of total length $\approx 1\text{M}$
 - The total length of exons $\approx 1\text{k}$

Transcription (Eukaryotes)

1. Transcription produces the **pre-mRNA** which contains both introns and exons
2. **5' cap** and **poly-A tail** are added to pre-mRNA
3. RNA splicing removes the introns and mRNA is produced.
4. mRNA are transported out of the nucleus



Gene structure (Eukaryotes)



The length of the yellow part must be multiple of 3!



Post-translation modification (PTM)

- Post-translation modification is the chemical modification of a protein after its translation. It involves
 - Addition of functional groups
 - E.g acylation, methylation, phosphorylation
 - Addition of other peptides
 - E.g. ubiquitination, the covalent linkage to the protein ubiquitin.
 - Structural changes
 - E.g. disulfide bridges, the covalent linkage of two cysteine amino acids.



Examples of PTM (Kinase and Phosphatases)

- Phosphorylation is a process to add a phosphate (PO_4) group to a protein.
- Kinase and Phosphatases can phosphorylate and dephosphorylate a protein.
- This process changes the conformation of proteins and causes them to become activated or deactivated.
- For example, phosphorylation of p53 (tumor suppressor protein) causes apoptotic cell death.
- Phosphorylation is used to dynamically turn on or off many signaling pathways.



Example of PTM (tRNA)

- Aminoacylation is the process of adding an aminoacyl group to a protein.
- tRNA applies aminoacylation to covalently link its 3' end CCA to an amino acid.
- This process is known as an aminoacyl tRNA synthetase.



Population genetic

- Given the genome of two individuals of the same species, if there exists a position (called loci) where the single nucleotides between the two individuals are different, we call it a single nucleotide polymorphism (SNP).
- For human, we expect SNPs are responsible for over 80% of the variation between two individuals.
- Hence, understanding SNPs can help us to understand the different within a population.
- For example, in human, SNPs control the color of hair, the blood type, etc of different individual. Also, many diseases like cancer are related to SNPs.



Basic Biotechnological Tools

- Cutting and breaking DNA
 - Restriction Enzymes
 - Shotgun method
- Copying DNA
 - Cloning
 - Polymerase Chain Reaction – PCR
- Measuring length of DNA
 - Gel Electrophoresis



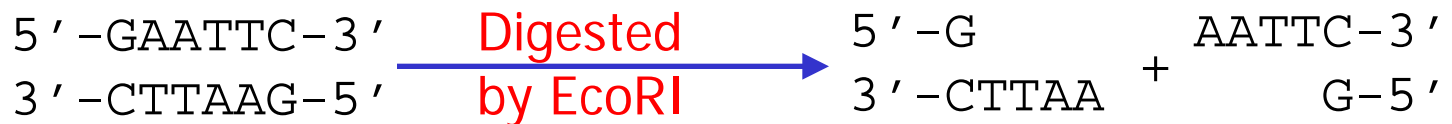
Restriction Enzymes

- Restriction enzyme recognizes certain point, called restriction site, in the DNA with a particular pattern and break it. Such process is called **digestion**.
- Naturally, restriction enzymes are used to break foreign DNA to avoid infection.
- Example:
 - EcoRI is the first restriction enzyme discovered that cuts DNA wherever the sequence GAATTC is found.
 - Similar to most of the other restriction enzymes, GAATTC is a **palindrome**, that is, GAATTC is its own reverse complement.
- Currently, more than 300 known restriction enzymes have been discovered.



EcoRI

- EcoRI is the first discovered restriction enzyme.



- It cut between G and A. **Sticky ends** are created.
- Note that some restriction enzymes give rise to **blunt ends** instead of sticky ends.



Shotgun method

- Break the DNA molecule into small pieces randomly
- Method:
 - Have a solution having a large amount of purified DNA
 - By applying high vibration, each molecule is broken randomly into small fragments.



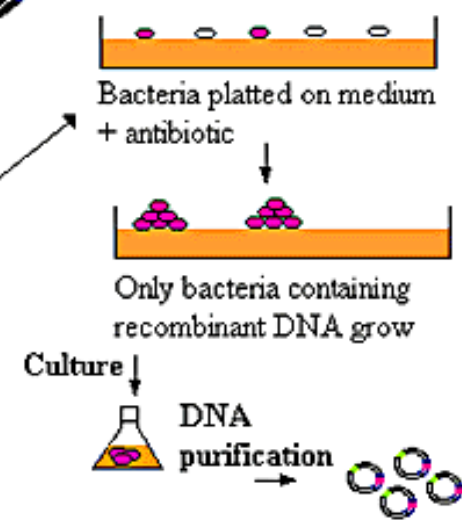
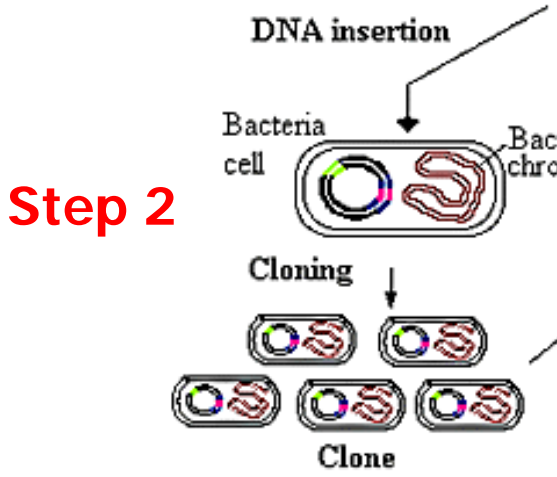
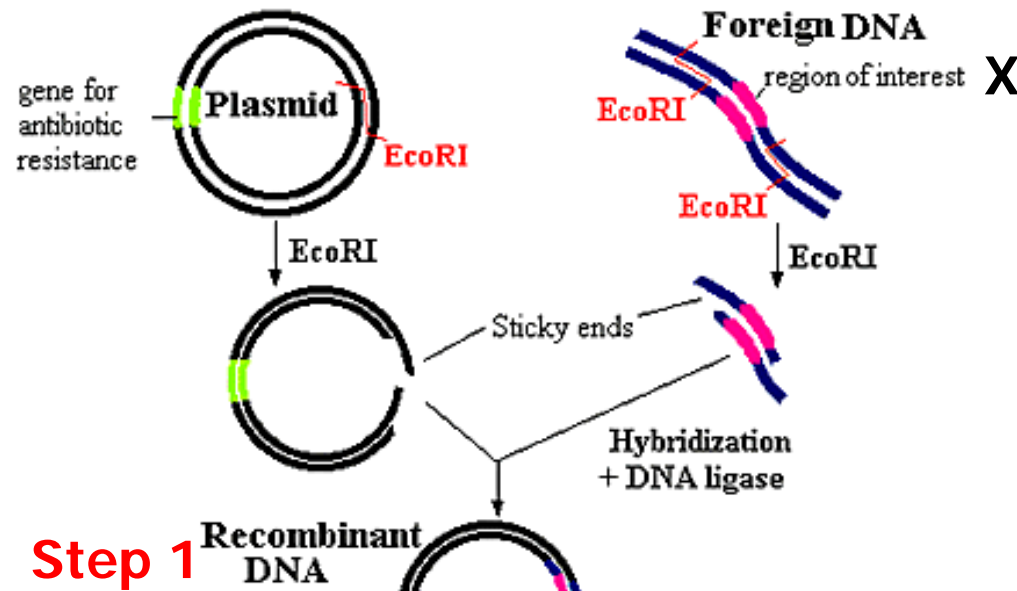
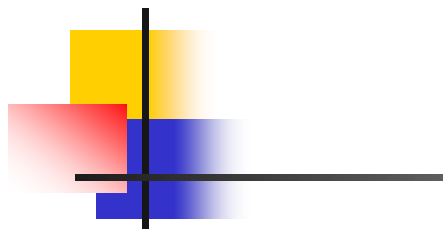
Cloning

- For many experiments, small amounts of DNAs are not enough.
- Cloning is one way to replicate DNAs.



Cloning by plasmid vector

- Given a piece of DNA X, the cloning process is as follows.
 1. Insert X into a **plasmid vector** with **antibiotic-resistance** gene and a **recombinant** DNA molecule is formed
 2. Insert the recombinant into the **host** cell (usually, E. coli).
 3. Grow the host cells in the presence of **antibiotic**.
 - Note that only cells with antibiotic-resistance gene can grow
 - When we duplicate the host cell, X is also duplicated.
 4. Select those cells with antibiotic-resistance genes.
 5. Kill them and extract X
- Note: cloning requires several days.



Cloning into a plasmid



More on cloning

- Cloning using **plasmid vector** is easy to manipulate in the laboratory. However, it can only replicate short DNA fragments (< 25k)
- To replicate long DNA fragments (10k-100k), we can use **yeast vector**.



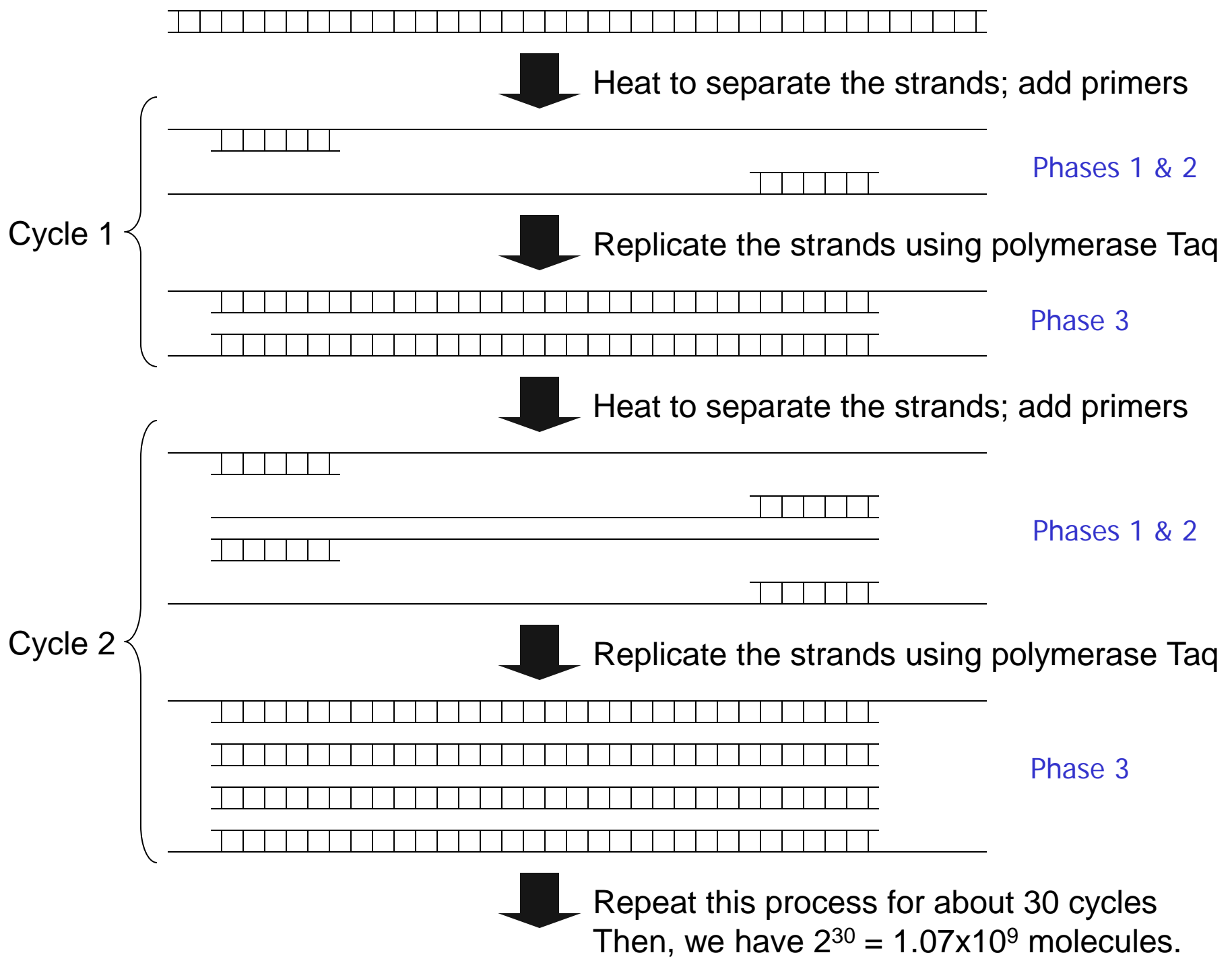
Polymerase Chain Reaction (PCR)

- PCR is invented by Kary B. Mullis in 1984
- PCR allows rapid replication of a selected region of a DNA without the need for a living cell.
 - Automated! Time required: a few hours
- Inputs for PCR:
 - Two oligonucleotides are synthesized, each complementary to the two ends of the region. They are used as primers.
 - Thermostable DNA polymerase TaqI
 - Taq stands for the bacterium *Thermos aquaticus* that grows in the yellowstone hot springs.



Polymerase Chain Reaction (PCR)

- PCR consists of repeating a cycle with three phases 25-30 times. Each cycle takes about 5 minutes
 - Phase 1: separate double stranded DNA by heat
 - Phase 2: cool; add synthesis primers
 - Phase 3: Add DNA polymerase TaqI to catalyze 5' to 3' DNA synthesis
- Then, the selected region has been amplified exponentially.





Example applications of PCR

- PCR method is used to amplify DNA segments to the point where it can be readily isolated for use.
- Example applications:
 - Clone DNA fragments from mummies
 - Detection of viral infections



Gel electrophoresis

- Developed by Frederick Sanger in 1977
- A technique used to separate a mixture of DNA fragments of different lengths.
- We apply an electrical field to the mixture of DNA.
- Note that DNA is negative charged. Due to friction, small molecules travel faster than large molecules.
- The mixture is separated into bands, each containing DNA molecules of the same length.



Applications

- Separating DNA sequences from a mixture
 - For example, after a genome is digested by a restriction enzyme, hundreds or thousands of DNA fragments are yielded. By Gel Electrophoresis, the fragments can be separated.
- Sequence Reconstruction
 - See next slide

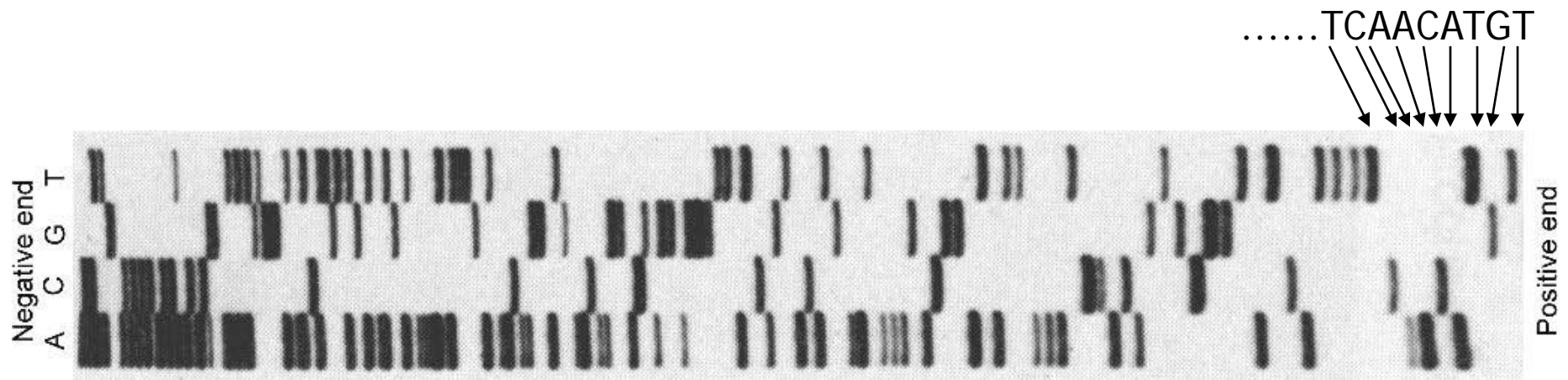


Sequencing by Gel electrophoresis

- An application of gel electrophoresis is to reconstruct DNA sequence of length 500-800 within a few hours
- Idea:
 - Generating all sequences end with A
 - Using gel electrophoresis, the sequences end with A are separated into different bands. Such information tells us the positions of A's in the sequence.
 - Similar for C, G, and T

Read the sequence

- We have four groups of fragments: A, C, G, and T.
- All fragments are placed in negative end.
- The fragments move to the positive end.
- From the relative distances of the fragments, we can reconstruct the sequence.
- The sequence is **TGTACAACT...**





Hybridization

- Among thousands of DNA fragments, Biologists routinely need to find a DNA fragment which contains a particular DNA subsequence.
- This can be done based on hybridization.
 1. Suppose we need to find a DNA fragments which contains ACCGAT.
 2. Create probes which is inversely complementary to ACCGAT.
 3. Mix the probes with the DNA fragments.
 4. Due to the hybridization rule (A=T, C≡G), DNA fragments which contain ACCGAT will hybridize with the probes.

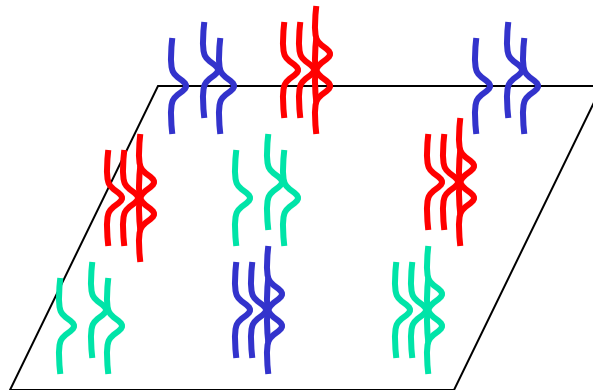


DNA array

- The idea of hybridization leads to the DNA array technology.
- In the past, “one gene in one experiment”
- Hard to get the whole picture
- DNA array is a technology which allows researchers to do experiment on a set of genes or even the whole genome.

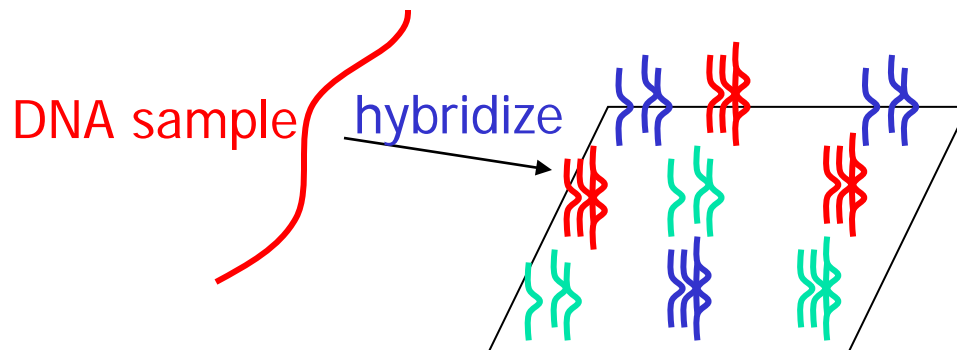
DNA array's idea (I)

- An orderly arrangement of thousands of spots.
- Each spot contains many copies of the same DNA fragment.



DNA array's idea (II)

- When the array is exposed to the target solution, DNA fragments in both array and target solution will match based on hybridization rule:
 - A=T, C≡G (hydrogen bond)
- Such idea allows us to do thousands of hybridization experiments at the same time.





Applications of DNA arrays

- **Sequencing by hybridization**
 - A promising alternative to sequencing by gel electrophoresis
 - It may be able to reconstruct longer DNA sequences in shorter time
- **Expression profile of a cell**
 - DNA arrays allow us to monitor the activities within a cell
 - Each spot contains the complement of a particular gene
 - Due to hybridization, we can measure the concentration of different mRNAs within a cell
- **SNP detection**
 - Using probes with different alleles to detect the single nucleotide variation.
- **Many many other applications!**



More advance tools

- Mass Spectrometry
- SAGE, PET technology
- ...



History of bioinformatics (I)

- 1866: Gregor Mendel discover genetics
 - Mendel's experiments on peas unveil some biological elements called genes which pass information from generation to generation
 - At that time, people think genetic information is carried by some “chromosomal” protein
- 1869: DNA was discovered
- 1944: Avery and McCarty demonstrate DNA is the major carrier of genetic information
- 1953: James Watson and Francis Crick deduced the three dimensional structure of DNA



History of bioinformatics (II)

- 1961: Elucidation of the genetic code, mapping DNA to peptides (proteins) [Marshall Nirenberg]
- 1968: Discovery of Restriction Enzyme
- 1970's: Development of DNA sequencing techniques: sequence segmentation and electrophoresis
- 1985: Development of Polymerase-Chain-Reaction (PCR): By exploiting natural replication, amplify DNA samples so that they are enough for doing experiment
- 1986: Discovery of RNA Splicing



History of bioinformatics (III)

- 1980-1990: Complete sequencing of the genomes of various organisms
- 1990: Launch of Human Genome Project (HGP)
- 1998: The discovery of post-transcription control called RNA interference [Fire and Mello]
- 2000: By shotgun sequencing, Craig Venter and Francis Collins jointly announced the publication of the first draft of the human genome.
- In the future 10 to 20 years:
 - Genomes to Life (GTL) Project
 - Understanding the detail mechanism of the cell
 - ENCODE Project
 - Annotating the whole genome
 - HAPMAP Project
 - Studying the variation of DNA for individuals