For written notes on this lecture, please read chapter 1 of *The Practical Bioinformatician*

# CS2220: Introduction to Computational Biology
# Lecture 1: Essence of Bioinformatics

**Limsoon Wong**

**12 January 2006**

NUS
National University
of Singapore

# Plan

**An very brief overview of …**

- **Molecular biology**
- **Tools and instruments for molecular biology**
- **Themes and applications of bioinformatics**
- **Commonly used data sources**

**Tools and instruments for molecular biology will be covered in a distributed manner in later lectures as and when needed**

# Molecular Biology Overview

National University of Singapore

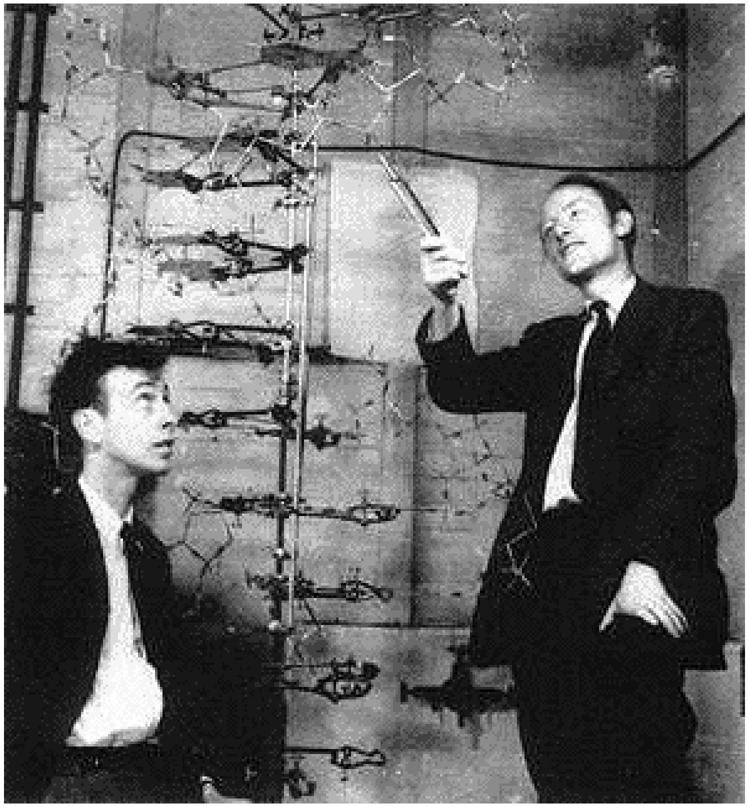
# Body and Cell

- **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**

- **Cells perform two types of function**
  - Chemical reactions needed to maintain our life
  - Pass info for maintaining life to next generation
- **In particular**
  - Protein performs chemical reactions
  - DNA stores & passes info
  - RNA is intermediate between DNA & proteins

# DNA

- **Stores instructions needed by the cell to perform daily life function**
- **Consists of two strands interwoven together and form a double helix**
- **Each strand is a chain of some small molecules called nucleotides**



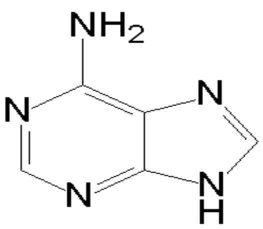
Francis Crick shows James Watson the model of DNA in their room number 103 of the Austin Wing at the Cavendish Laboratories, Cambridge

# Classification of Nucleotides

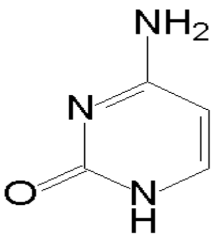
- **5 diff nucleotides: adenine(A), cytosine(C), guanine(G), thymine(T), & uracil(U)**

- **A, G are purines. They have a 2-ring structure**
- **C, T, U are pyrimidines. They have a 1-ring structure**

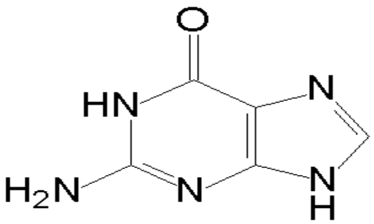
- **DNA only uses A, C, G, & T**



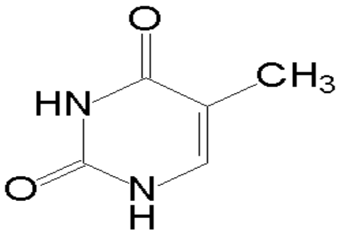
A



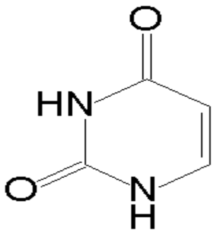
C



G



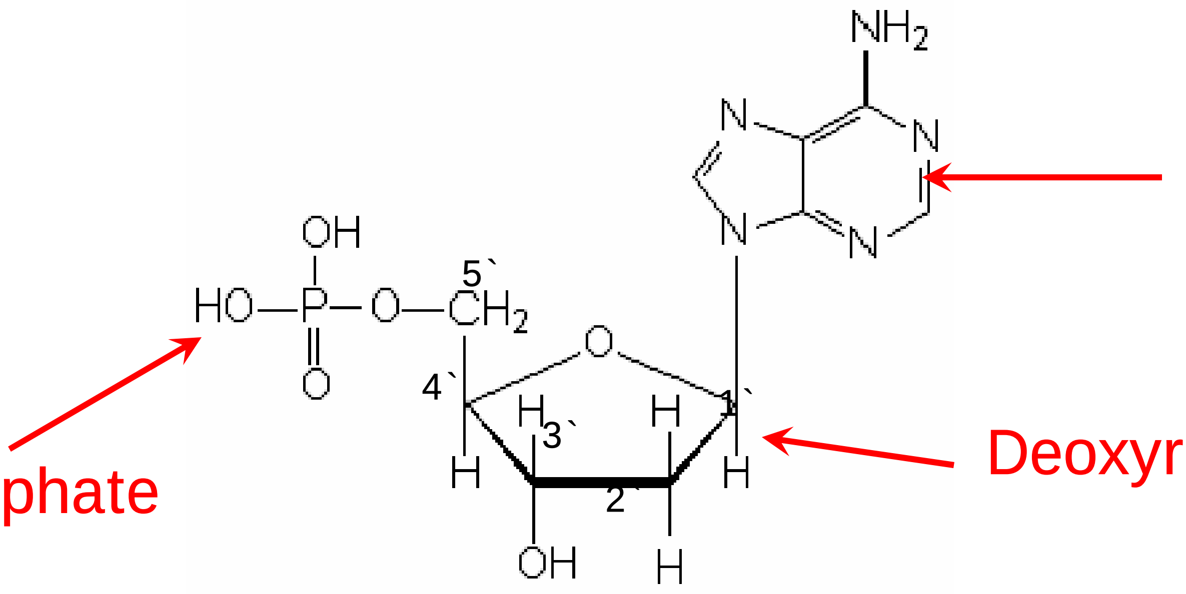
T



U

# Nucleotide

- **Consists of three parts:**
  - Deoxyribose
  - Phosphate (bound to the 5’ carbon)
  - Base (bound to the 1’ carbon)



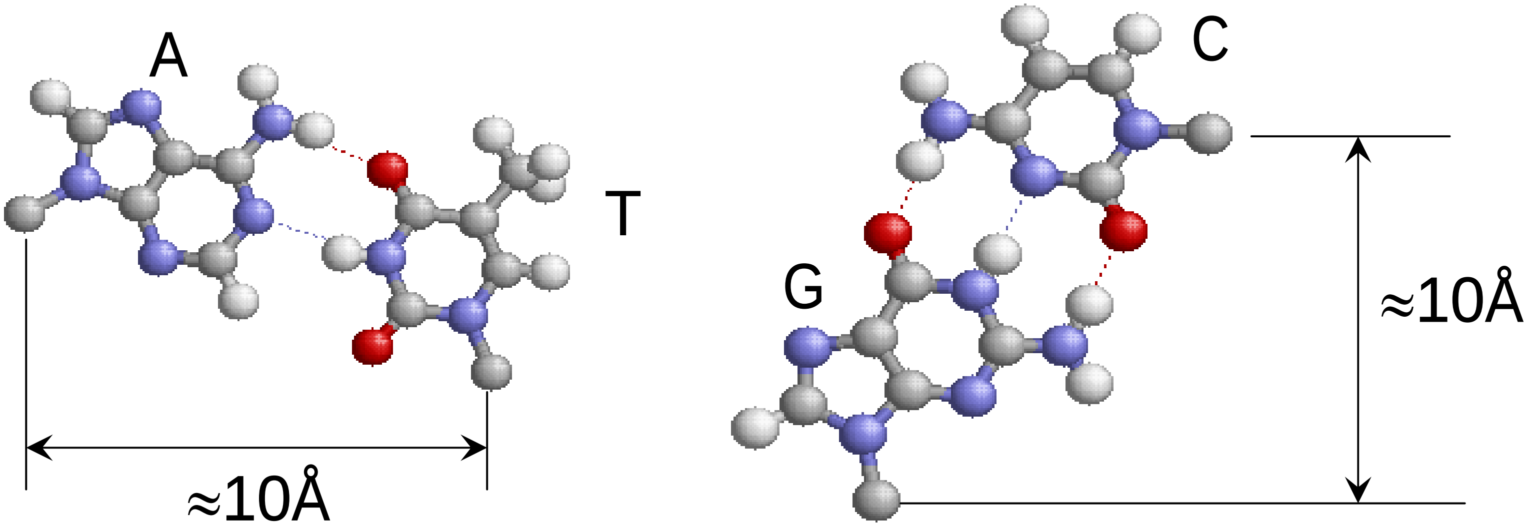
Base (Adenine)

Phosphate

Deoxyribose

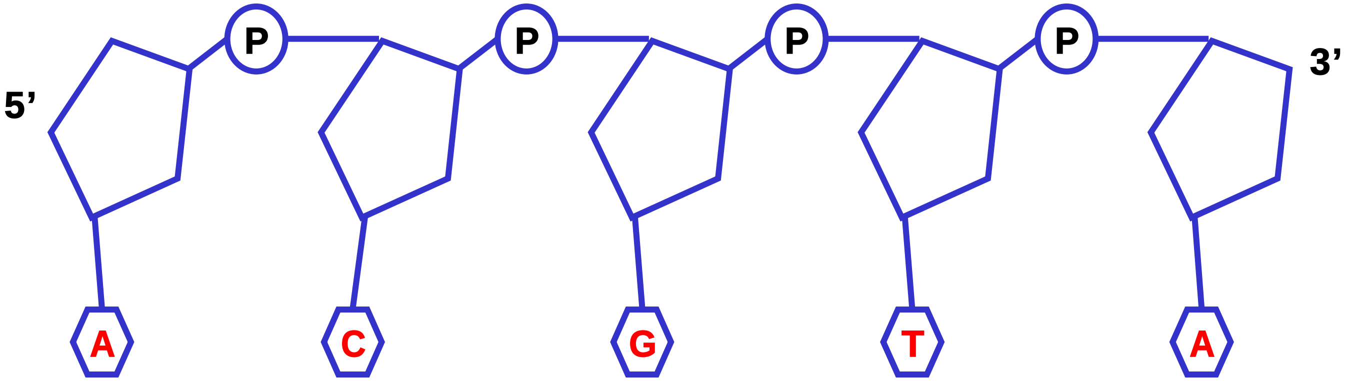
# Watson-Crick Rule

- **DNA is double stranded in a cell**
- **One strand is reverse complement of the other**
- **Complementary bases:**
  - – A with T (two hydrogen-bonds)
  - – C with G (three hydrogen-bonds)

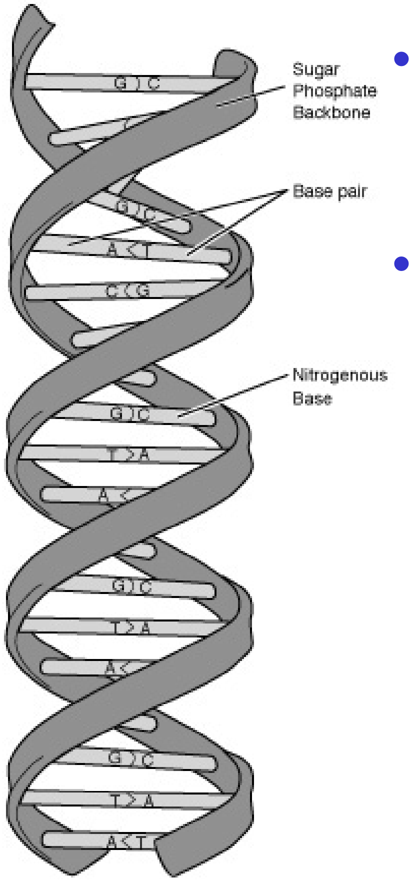


# Orientation of a DNA

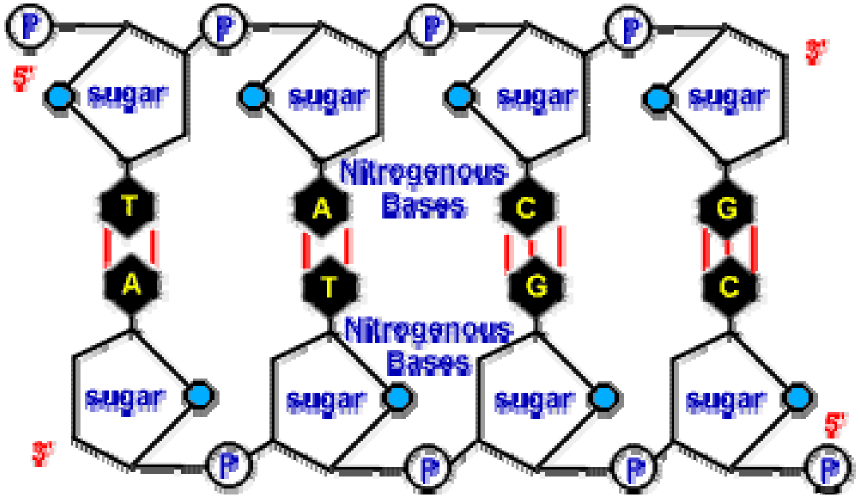
- **One strand of DNA is generated by chaining together nucleotides, forming a phosphate-sugar backbone**
- **It has direction: from 5’ to 3’, because DNA always extends from 3’ end:**
  - Upstream, from 5’ to 3’
  - Downstream, from 3’ to 5’



# Double Stranded DNA



- **DNA is double stranded in a cell. The two strands are anti-parallel. One strand is reverse complement of the other**
- **The double strands are interwoven to form a double helix**



# Locations of DNAs in a Cell?

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

# Chromosome

- **DNA is usually tightly wound around histone proteins and forms a chromosome**
- **The total info stored in all chromosomes constitutes a genome**
- **In most multi-cell organisms, every cell contains the same complete set of chromosomes**
  - May have some small diff due to mutation
- **Human genome has 3G bases, organized in 23 pairs of chromosomes**

# Gene

- **The physical and functional unit of heredity that carries info from one generation to the next**
- **A sequence of DNA that encodes a protein or an RNA molecule**
- **About 30,000 – 35,000 (protein-coding) genes in human genome**
- **For gene that encodes protein**
  - In Prokaryotic genome, one gene corresponds to one protein
  - In Eukaryotic genome, one gene may correspond to more than one protein because of the process "alternative splicing"

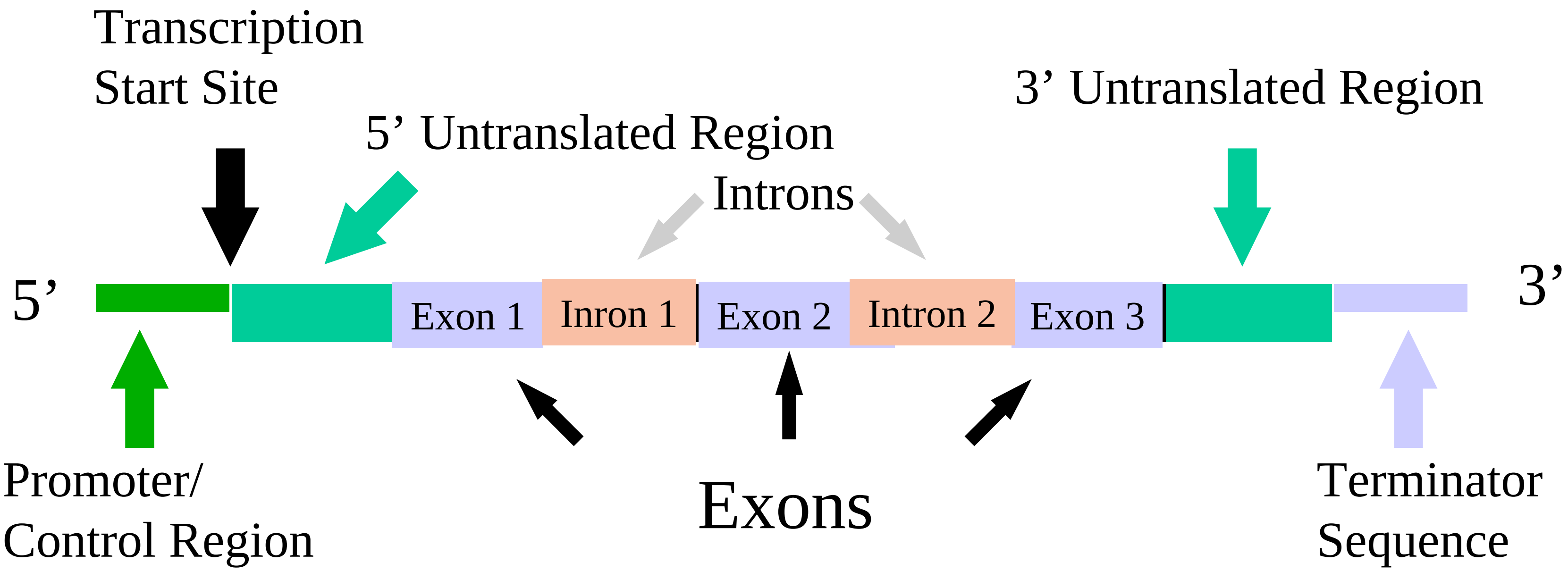
# Introns and Exons

- **Eukaryotic genes contain introns & exons**
  - Introns are seq that are ultimately spliced out of mRNA
  - Introns normally satisfy GT-AG rule, viz. begin w/ GT & end w/ AG
  - Each gene can have many introns & each intron can have thousands bases

- **Introns can be very long**

- **An extreme example is a gene associated with cystic fibrosis in human:**
  - Length of 24 introns ~1Mb
  - Length of exons ~1kb

# A “Simple” Gene



Transcription
Start Site

5’ Untranslated Region

Introns

3’ Untranslated Region

5’

Exon 1

Inron 1

Exon 2

Intron 2

Exon 3

3’

Promoter/
Control Region

Exons

Terminator
Sequence

# Complexity of Organism vs. Genome Size

- **Human Genome: 3G base pairs**
- **Amoeba dubia (a single cell organism): 600G base pairs**

⇒ **Genome size has no relationship with the complexity of the organism**

# Number of Genes vs. Genome Size

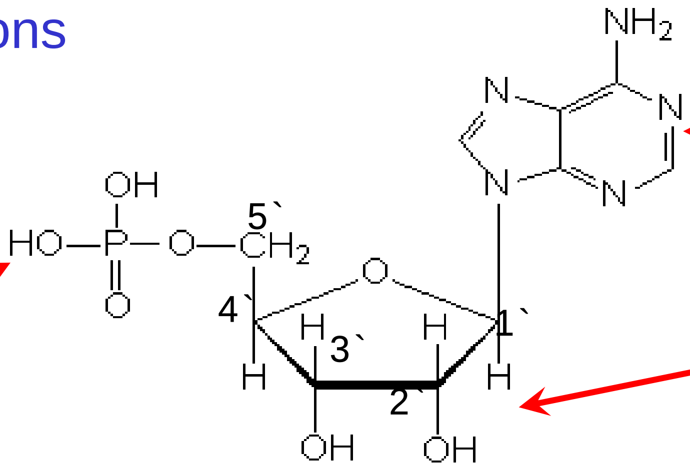
- **Prokaryotic genome (e.g., *E. coli*)**
  - Number of base pairs: 5M
  - Number of genes: 4k
  - Average length of a gene: 1000 bp
- **Eukaryotic genome (e.g., human)**
  - Number of base pairs: 3G
  - Estimated number of genes: 30k – 35k
  - Estimated average length of a gene: 1000-2000 bp
- **~ 90% of *E. coli* genome are coding regions**
- **< 3% of human genome are coding regions**

⇒ **Genome size has no relationship w/ number of genes**

# RNA

- **RNA has both the properties of DNA & protein**
  - Similar to DNA, it can store & transfer info
  - Similar to protein, it can form complex 3D structure & perform some functions

- **Nucleotide for RNA has of three parts:**
  - Ribose Sugar (has an extra OH group at 2’)
  - Phosphate (bound to 5’ carbon)
  - Base (bound to 1’ carbon)



Base (Adenine)

Phosphate

Ribose Sugar

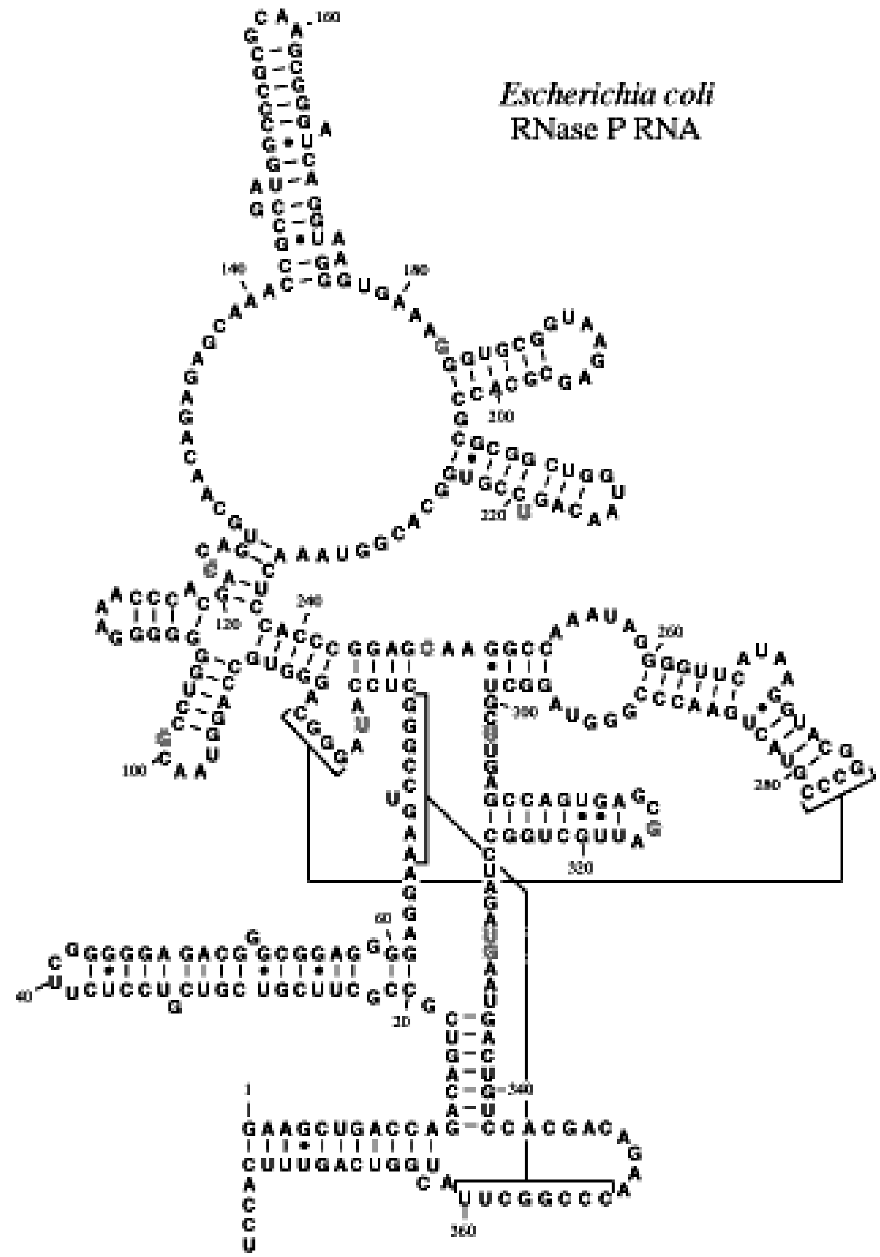
# RNA vs DNA

- **RNA is single stranded**

- **Nucleotides of RNA are similar to that of DNA, except that have an extra OH at position 2’**
  - Due to this extra OH, it can form more hydrogen bonds than DNA
  - So RNA can form complex 3D structure

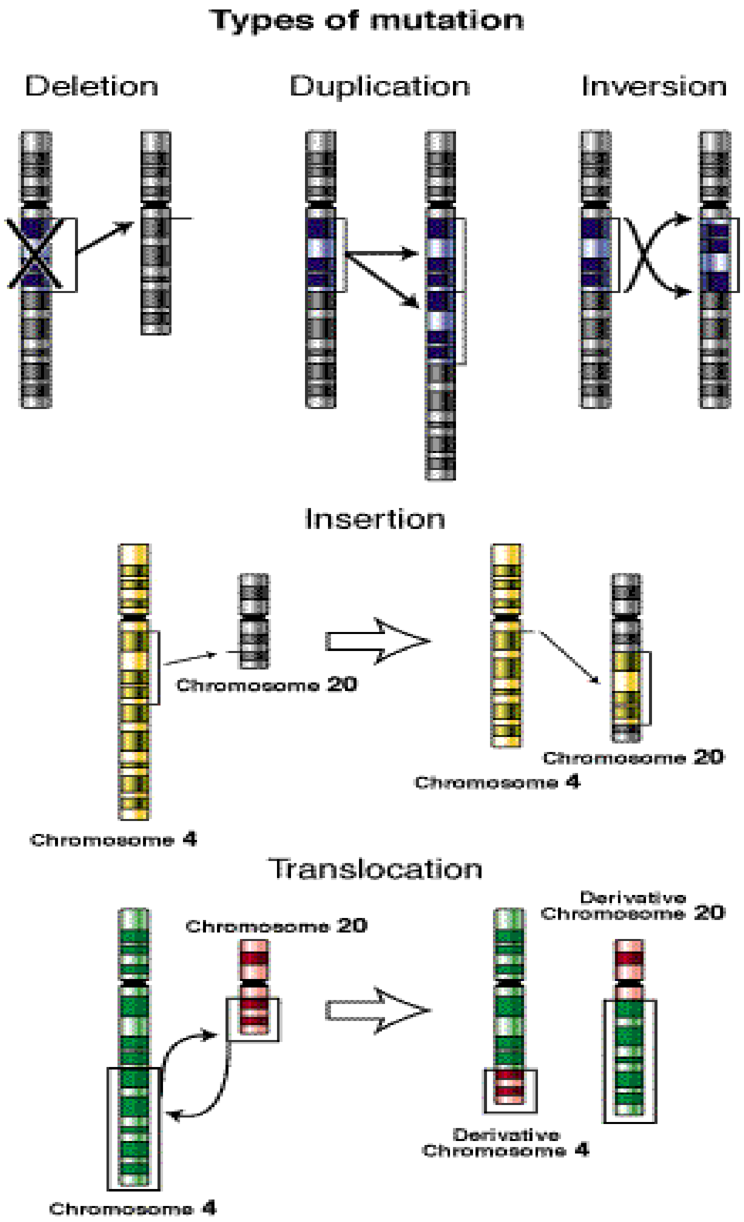
- **RNA use the base U instead of T**
  - U is chemically similar to T
  - In particular, U is also complementary to A

# RNA Secondary Structure



- **E. coli Rnase P RNA secondary structure**

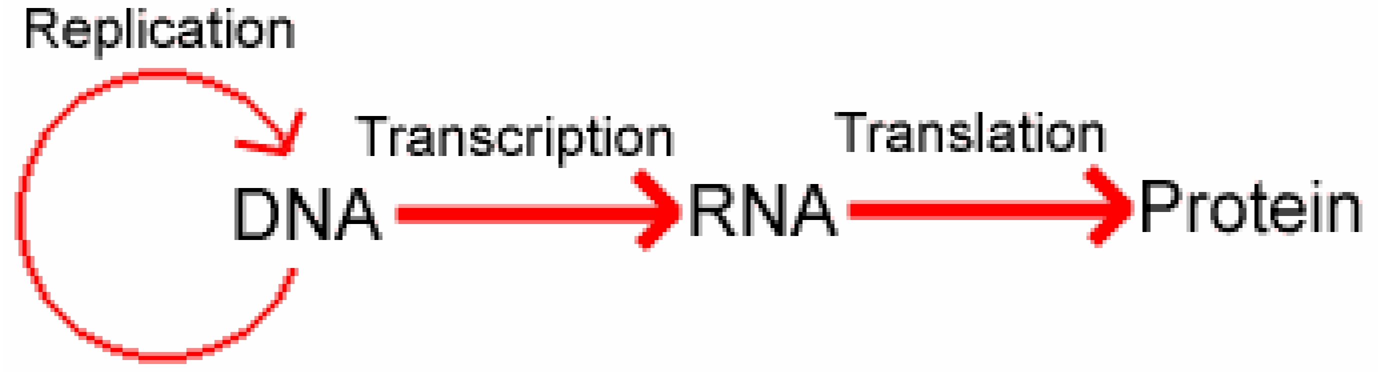
Image source: www.mbio.ncsu.edu/JWB/MB409/lecture/ lecture05/lecture05.htm

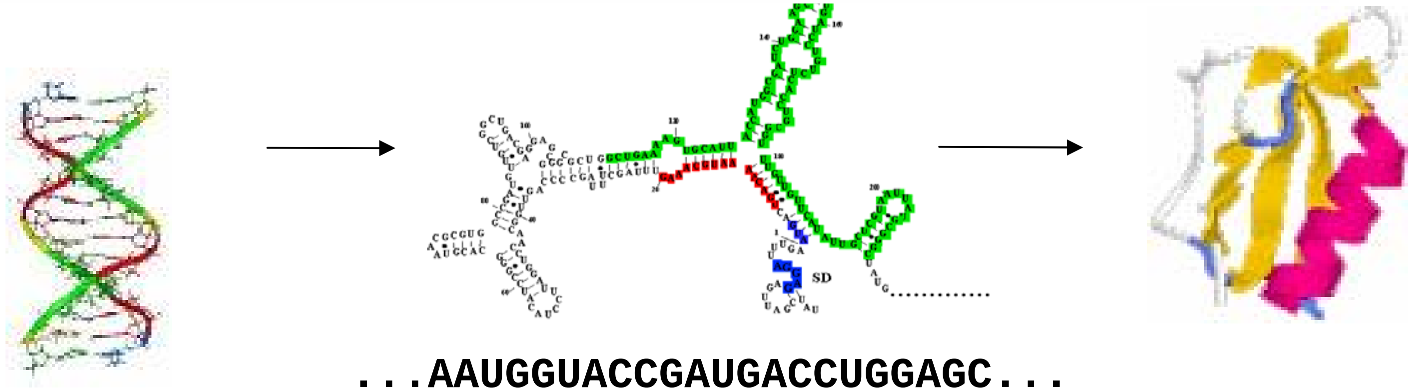


# Mutation

- **Mutation is a sudden change of genome**
- **Basis of evolution**
- **Cause of cancer**
- **Can occur in DNA, RNA, & Protein**

# Central Dogma



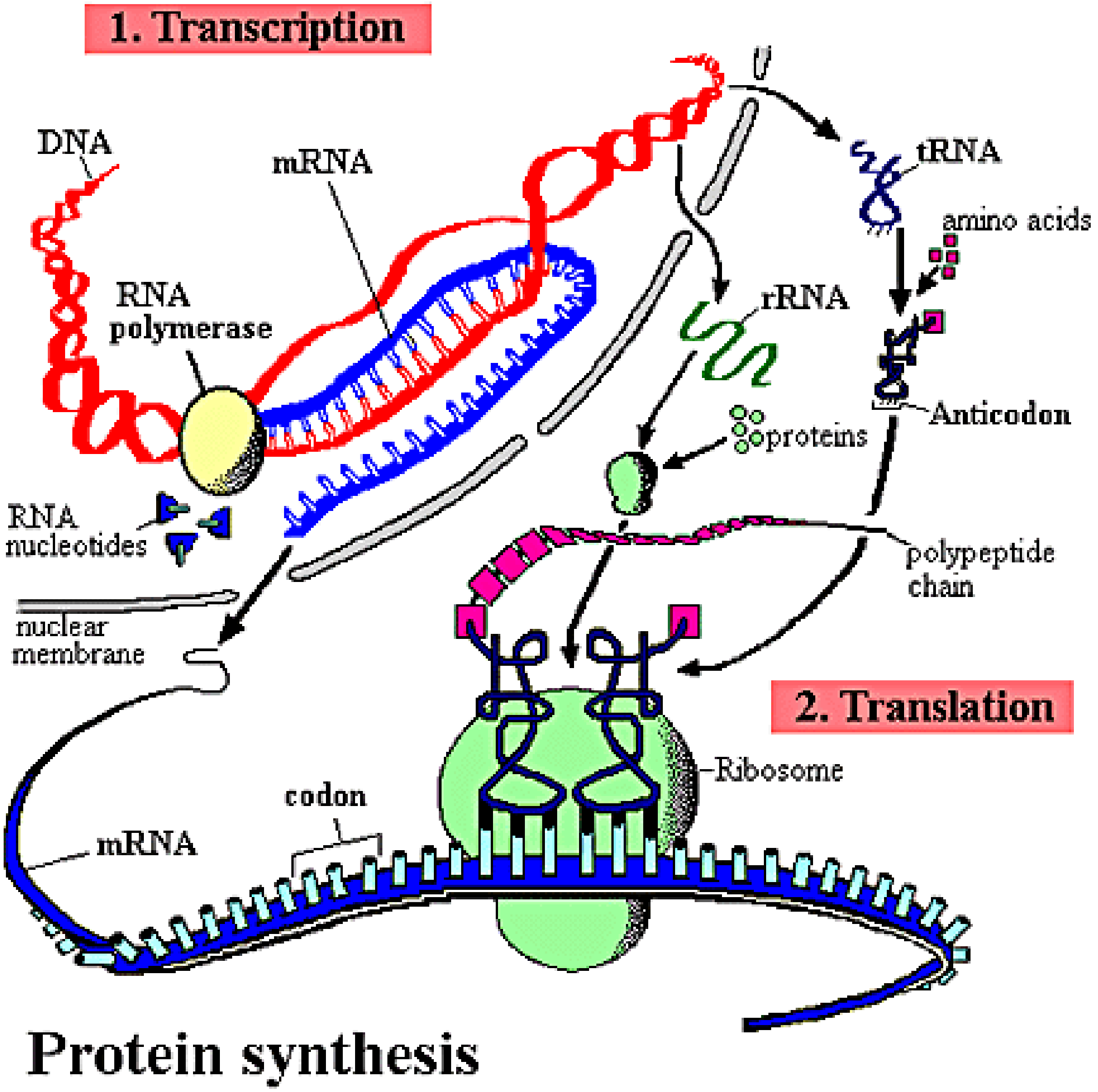


**...AAUGGUACCGAUGACCUGGAGC...**

**...AATGGTACCGATGACCTG...**

**...TRLRPLLALLALWP...**

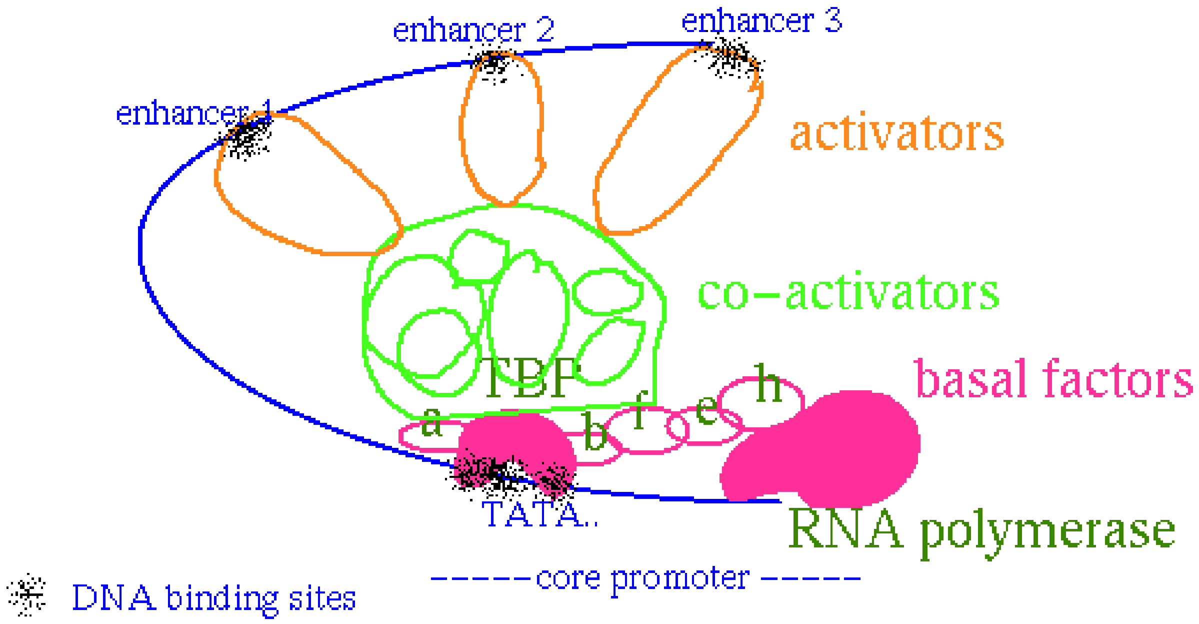
# Players in Protein Synthesis



# Transcription

- **Synthesize mRNA from one strand of DNA**
  - An enzyme RNA polymerase temporarily separates double-stranded DNA
  - It begins transcription at transcription start site
  - A → A, C→C, G→G, & T→U
  - Once RNA polymerase reaches transcription stop site, transcription stops
- **Additional "steps" for Eukaryotes**
  - Transcription produces pre-mRNA that contains both introns & exons
  - 5' cap & poly-A tail are added to pre-mRNA
  - RNA splicing removes introns & mRNA is made
  - mRNA are transported out of nucleus

# Promoter and Enhancers



- **Promoter necessary to start transcription**
- **Enhancer can affect transcription from afar**

# Translation

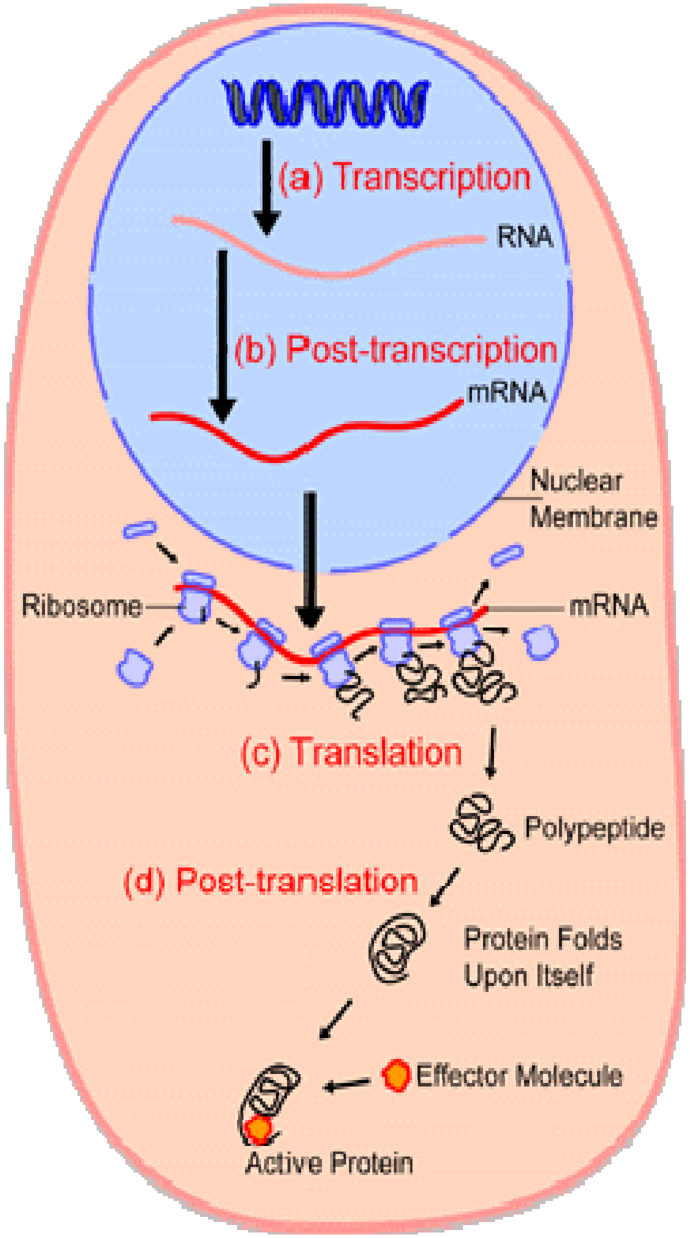
- Synthesize protein from mRNA
- Each amino acid is encoded by consecutive seq of 3 nucleotides, called a codon
- The decoding table from codon to amino acid is called genetic code

- $4^3$=64 diff codons

$\Rightarrow$ Codons are not 1-to-1 corr to 20 amino acids

- All organisms use the same decoding table
- Recall that amino acids can be classified into 4 groups. A single-base change in a codon is usu insufficient to cause a codon to code for an amino acid in diff group

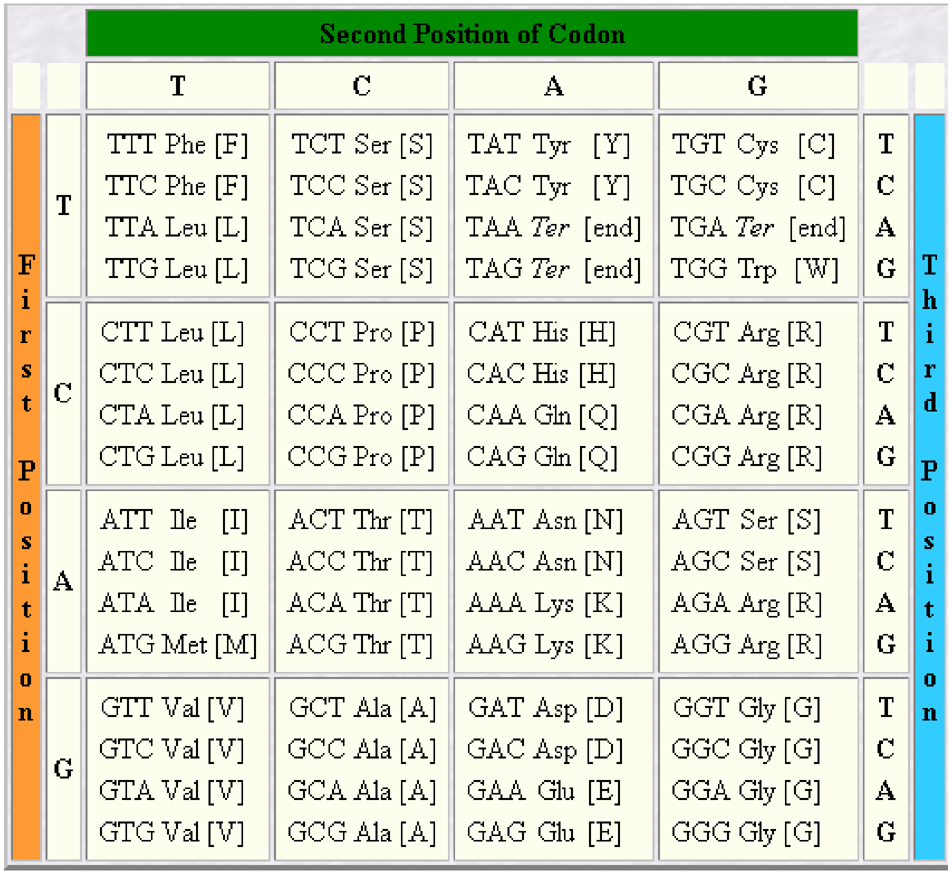
# Protein Synthesis



- **Within nucleus (light blue), genes (dark blue) are transcribed to RNA**
- **Post-transcriptional modification and control, results in a mature mRNA (red)**
- **mRNA translocated to cytoplasm (peach)**
- **mRNA translated by ribosomes (purple) that match codons of mRNA to anti-codons of tRNA**
- **Newly synthesized proteins (black) are further modified, such as by binding to an effector molecule (orange), to become fully active**

# Genetic Code

- **Start codon**
  - ATG (code for M)
- **Stop codon**
  - TAA
  - TAG
  - TGA



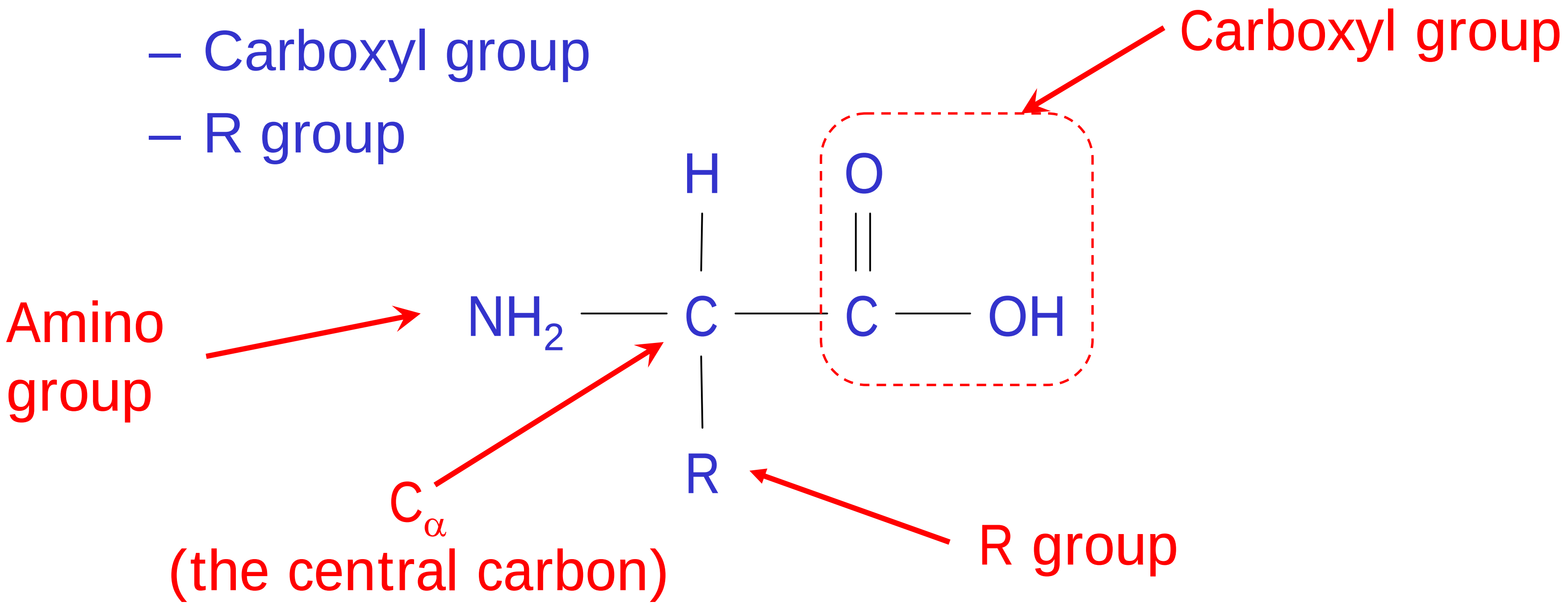
| | | Second Position of Codon | | | | | |
|---|---|---|---|---|---|---|---|
| | | T | C | A | G | | |
| First Position | T | TTT Phe [F] | TCT Ser [S] | TAT Tyr [Y] | TGT Cys [C] | T | Third Position |
| | | TTC Phe [F] | TCC Ser [S] | TAC Tyr [Y] | TGC Cys [C] | C | |
| | | TTA Leu [L] | TCA Ser [S]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

# Protein

- **A sequence composed from an alphabet of 20 amino acids**
    - Length is usually 20 to 5000 amino acids
    - Average around 350 amino acids
- **Folds into 3D shape, forming the building block & performing most of the chemical reactions within a cell**

# Amino Acid

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



$\mathrm{NH_2-\overset{\displaystyle H}{\underset{\displaystyle R}{C}}-\overset{\displaystyle O}{\overset{\|}{C}}-OH}$

Carboxyl group

Amino group

$C_\alpha$ (the central carbon)

R group

# Classification of Amino Acids

- **Amino acids can be classified into 4 types**

- **Positively charged (basic)**
  - Arginine (Arg, R)
  - Histidine (His, H)
  - Lysine (Lys, K)

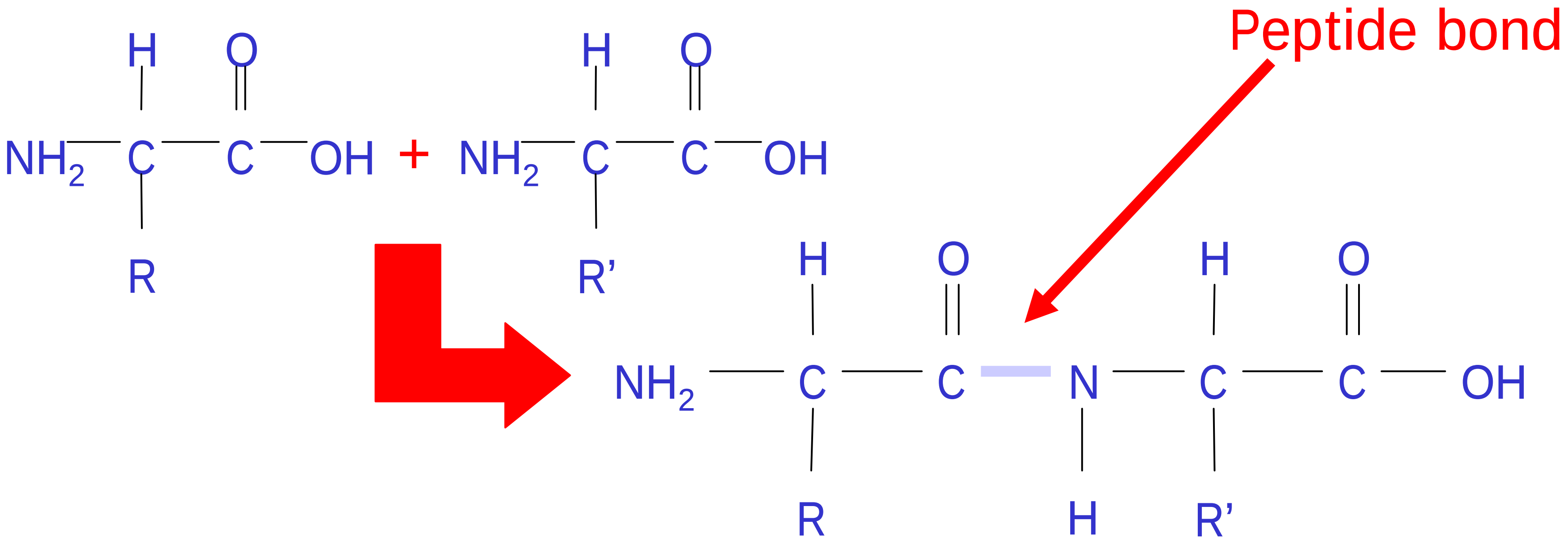
- **Negatively charged (acidic)**
  - Aspartic acid (Asp, D)
  - Glutamic acid (Glu, E)

# Classification of Amino Acids

- **Polar (overall uncharged, but uneven charge distribution. can form hydrogen bonds with water. they are called hydrophilic)**
  - Asparagine (Asn, N)
  - Cysteine (Cys, C)
  - Glutamine (Gln, Q)
  - Glycine (Gly, G)
  - Serine (Ser, S)
  - Threonine (Thr, T)
  - Tyrosine (Tyr, Y)
- **Nonpolar (overall uncharged and uniform charge distribution. cant form hydrogen bonds with water. they are called hydrophobic)**
  - Alanine (Ala, A)
  - Isoleucine (Ile, I)
  - Leucine (Leu, L)
  - Methionine (Met, M)
  - Phenylalanine (Phe, F)
  - Proline (Pro, P)
  - Tryptophan (Trp, W)
  - Valine (Val, V)

# Protein & Polypeptide Chain

- **Formed by joining amino acids via peptide bond**
- **One end the amino group, called N-terminus**
- **The other end is the carboxyl group, called C-terminus**



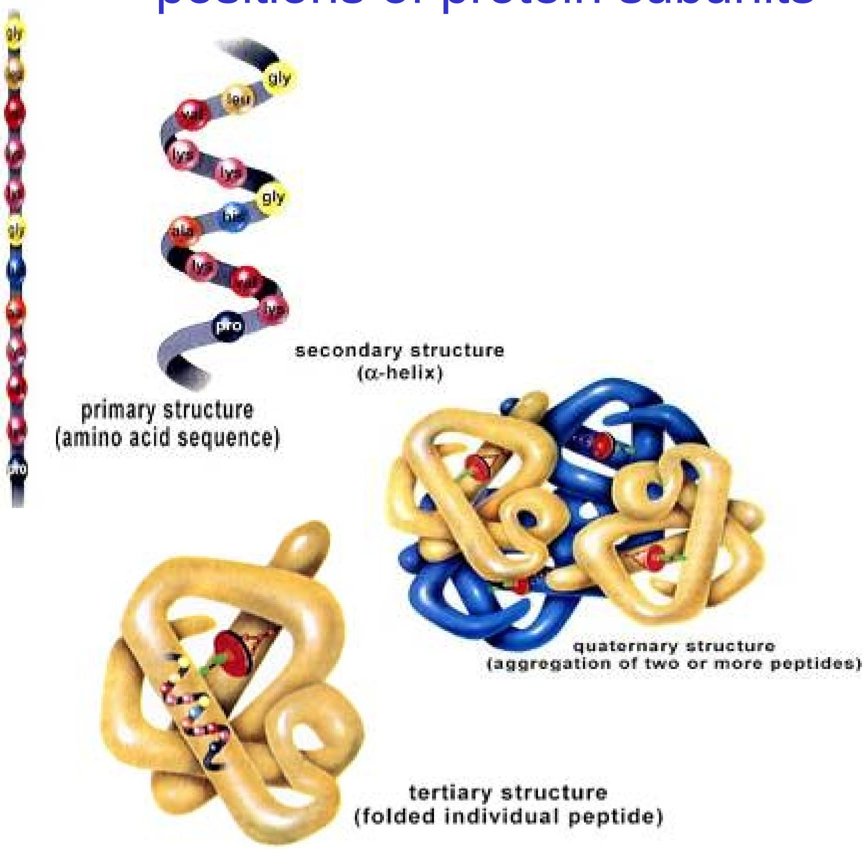
$$\mathrm{NH_2}-\overset{\mathrm{H}}{\underset{\mathrm{R}}{\mathrm{C}}}-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-\mathrm{OH} + \mathrm{NH_2}-\overset{\mathrm{H}}{\underset{\mathrm{R'}}{\mathrm{C}}}-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-\mathrm{OH}$$

$$\rightarrow \mathrm{NH_2}-\overset{\mathrm{H}}{\underset{\mathrm{R}}{\mathrm{C}}}-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-\underset{\mathrm{H}}{\mathrm{N}}-\overset{\mathrm{H}}{\underset{\mathrm{R'}}{\mathrm{C}}}-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-\mathrm{OH}$$

Peptide bond

# Proteins Structure

- **Primary**
  - Seq of amino acids forming a polypeptide chain
- **Secondary**
  - Local organization into sec structures such as α helices and β sheets
- **Tertiary**
  - 3D arrangements of amino acids as they react to one another due to the polarity and resulting interactions betw their side chains
- **Quaternary**
  - Number and relative positions of protein subunits



# Eukaryote Cell Structure

- **Cell membrane---a cell's protective coat**
  - Separate and protect cell from env
  - Made from double layer of lipids and proteins

- **Genetic material**
  - DNA and RNA

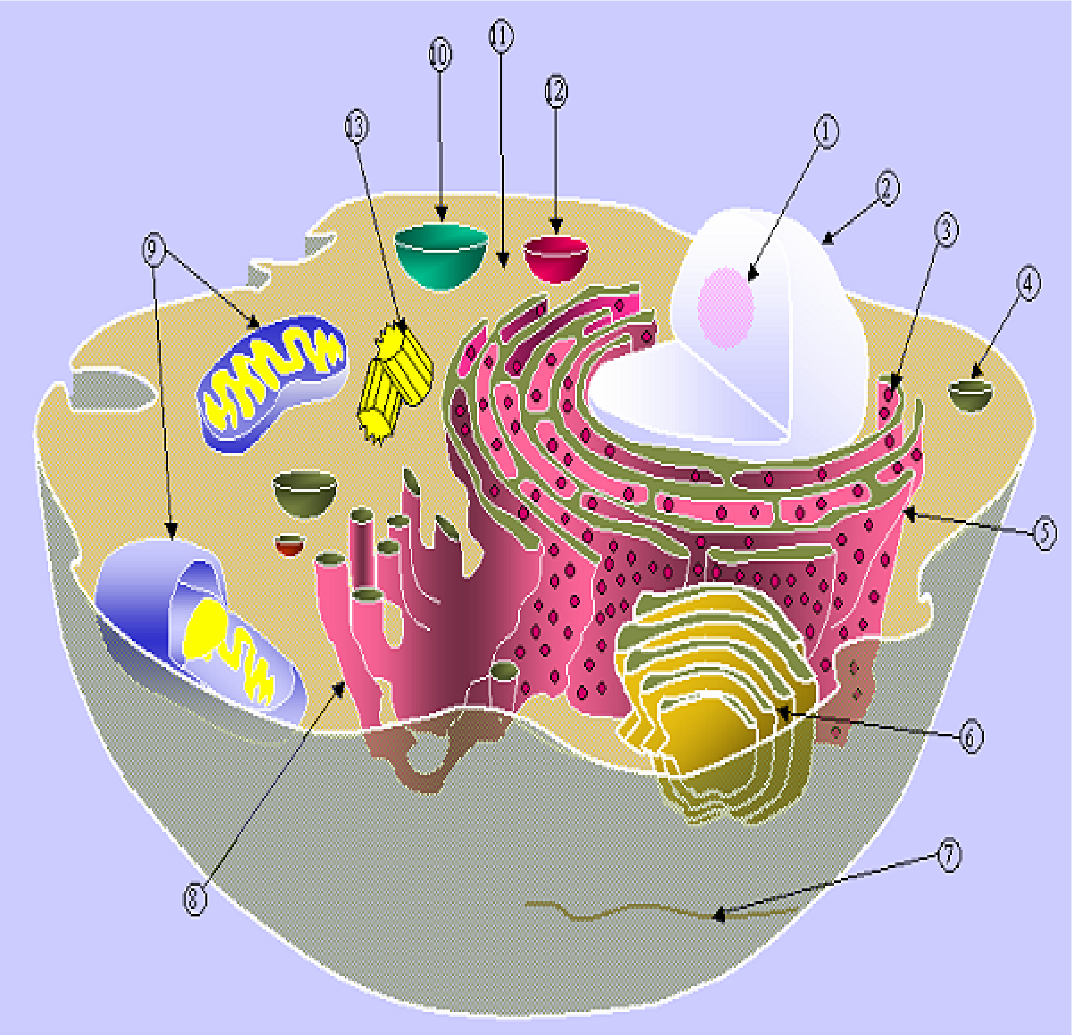
- **Organelles--- a cell's “little organs”**

- **Cytoskeleton---a cell's scaffold**
  - organize and maintain the cell's shape
  - anchor organelles in place
  - Help uptake of external materials by a cell
  - move parts of the cell during growth and motility

# Organelles

- **Cell nucleus---a cell's info ctr**
  - House a cell's chromosomes
  - DNA replication and RNA synthesis occur here
- **Ribosomes---the protein production machine**
  - Process genetic instructions carried by mRNA into protein
- **Mitochondria & chloroplasts---the power generators**
  - Self-replicating organelles in cytoplasm, w/ own genome
  - Generate energy, process involves metabolic pathways
- **Endoplasmic reticulum**
  - rough ER help to export proteins from cell after mRNA translation
  - smooth ER is important in lipid synthesis, detoxification etc.
- **Golgi apparatus---central delivery system for the cell**
  - Site for protein processing, packaging, and transport

# Eukaryote Cell Structure



1. Nucleolus
2. Nucleus
3. Ribosome
4. Vesicle
5. Rough ER
6. Golgi apparatus
7. Cytoskeleton
8. Smooth ER
9. Mitochondrion
10. Vacuole
11. Cytoplasm
12. Lysosome
13. Centriole

# Processes In/Out of the Cells

- **Biological pathway: Molecular interaction network in biological processes**

- **Regulatory pathway**
  - Genetic information processing
  - Environmental information processing
  - Cellular processes

- **Metabolic pathway**
  - Enzymatic processes creating energy and other parts of the cell

# Regulatory Pathways

- **Genetic information processing**
  - – Transcription, Translation, Sorting and Degradation. Replication and Repair
- **Environmental information processing**
  - – Membrane transport, Signal transduction, Ligand receptor interaction
- **Cellular processes**
  - – Cell motility, Cell growth and death, Cell communication, Development, Behavior

# Signal Transduction Pathways

- **Signal transduction is a process by which a cell converts one kind of signal/stimulus into another**
- **Stimuli/Responses**
  - Stimuli: factors from env of a cell, e.g., kinds of molecules buffeting its surface, temperature, …
  - Responses: how cell react to stimuli, e.g., activate of a gene, produce metabolic energy, …
- **Types of signals**
  - Extracellular: binding of “extracellular” signaling molecules to receptors that face out from membrane
  - Intracellular: trigger by extracellular signal
  - Intercellular: between cells

# Type of Intercellular Signaling

- **Endocrine**
  - Broad effect, specific receptor, travel thru blood
  - Hormones
- **Paracrine**
  - Within local tissue, enzyme/extracellular matrix
- **Autocrine**
  - Affect only cells of the same type
- **Juxtacrine**
  - Transmitted along cell membranes
  - Capable of affecting either the emitting cell or cells immediately adjacent

# Type of Signaling Proteins

- **Signal molecule**
  - – Bring signal to outside the cell
- **Receptors**
  - – Bring signal from outside to inside the cell
  - – One end outside membrane, the other end inside
  - – Applied to cell membrane and nucleus membrane
- **Intracellular signaling protein**
  - – Second messengers inside cells
  - – Pass message from receptors to target protein within a cell including the nucleus
- **Target protein**
  - – Final recipient of signal. Might be many

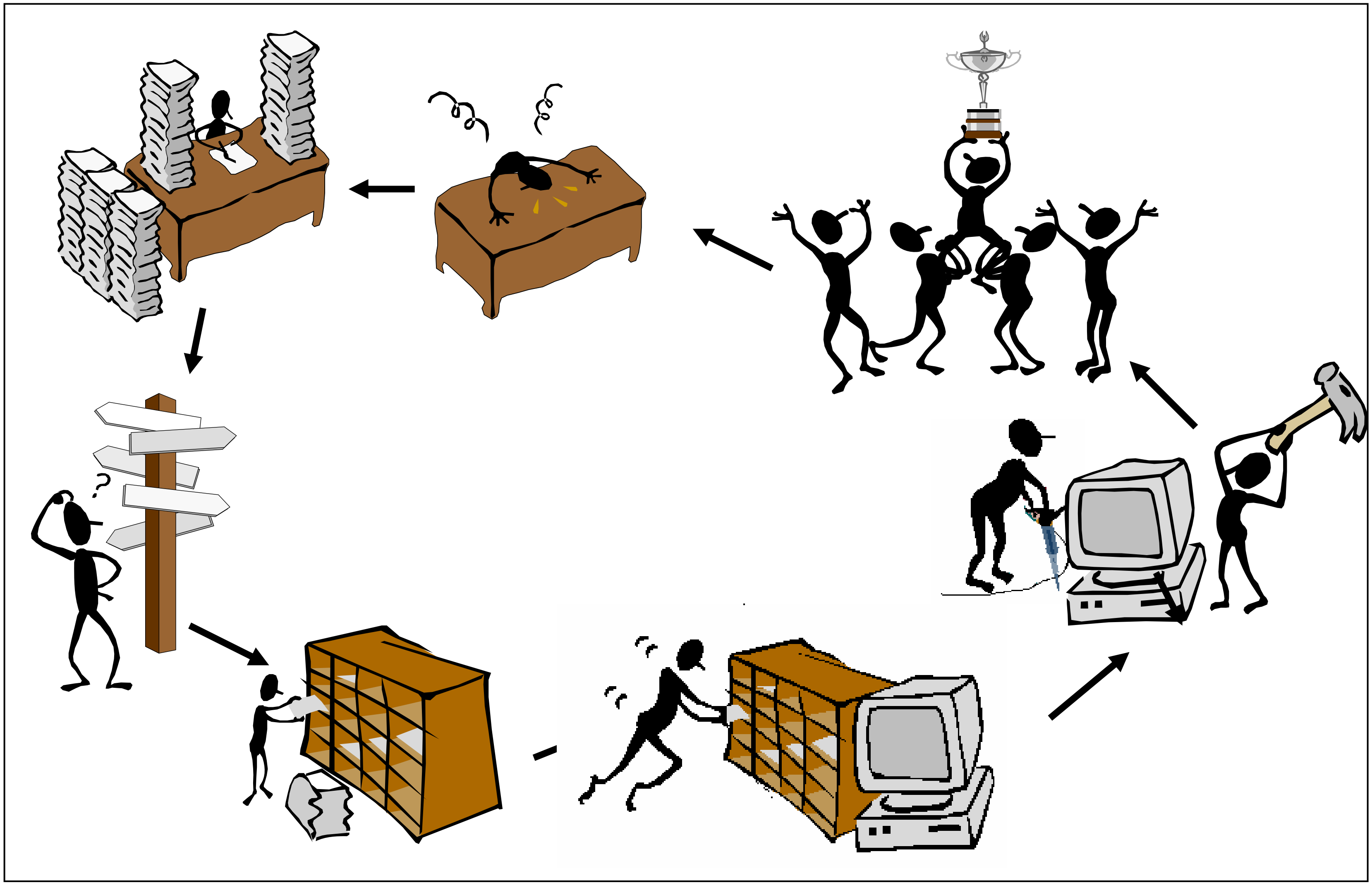
# Metabolic Pathways

- **Cell metabolism is the sum of many ongoing individual processes by which living cells process nutrient molecules to maintain a living state**
- **Anabolism**
  - Energy is consumed to make or combine simpler substances---e.g., amino acids---into more complex compounds, such as enzymes and nucleic acids
- **Catabolism**
  - Complex molecules are broken down to produce energy and reducing power
  - Carbohydrate catabolism, Fat catabolism, Protein catabolism

# Themes and Applications of Bioinformatics

National University of Singapore

# What is Bioinformatics?



# Themes of Bioinformatics

Bioinformatics =

Data Mgmt +

Knowledge Discovery +

Sequence Analysis +

Physical Modeling + ….

Knowledge Discovery =

Statistics + Algorithms + Databases

# Benefits of Bioinformatics

To the patient:

Better drug, better treatment

To the pharma:

Save time, save cost, make more $

To the scientist:

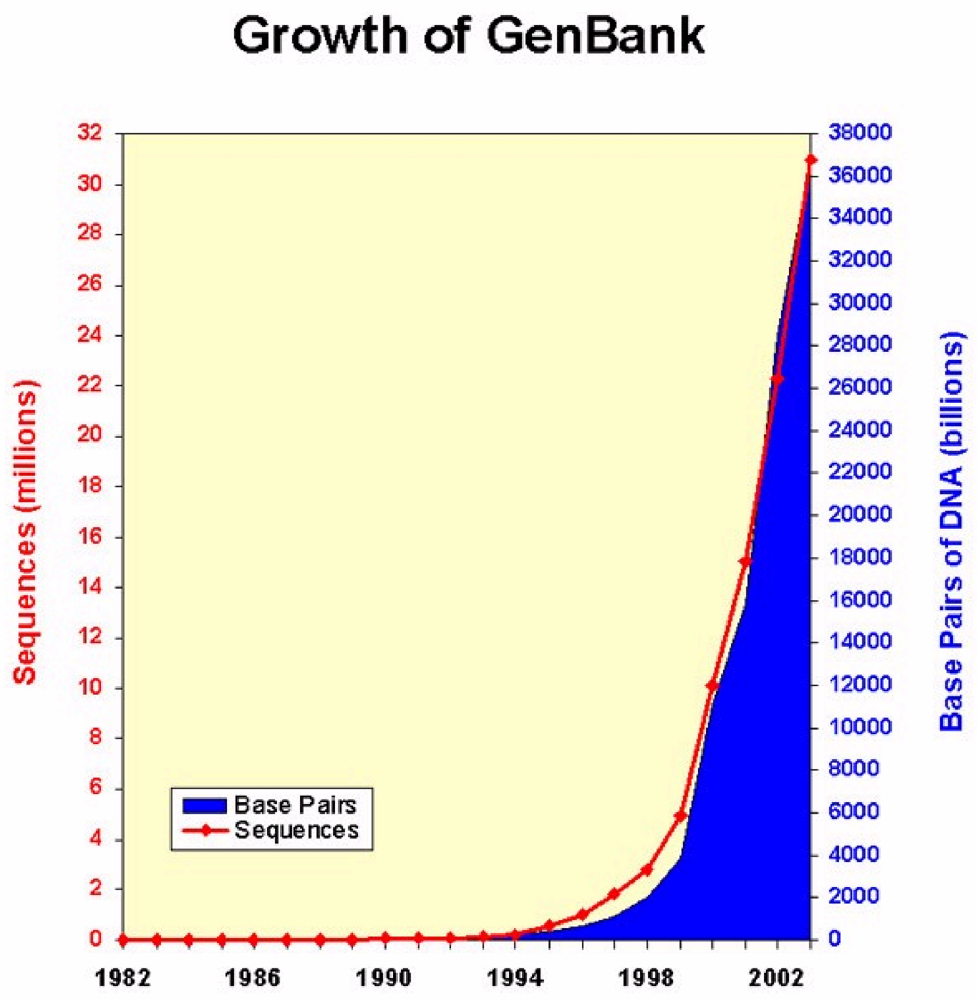
Better science

# Some Bioinformatics Problems

- **Biological Data Searching**
- **Gene/Promoter finding**
- **Cis-regulatory DNA**
- **Gene/Protein Network**
- **Protein/RNA Structure Prediction**
- **Evolutionary Tree reconstruction**
- **Infer Protein Function**
- **Disease Diagnosis**
- **Disease Prognosis**
- **Disease Treatment Optimization, ...**

# Biological Data Searching

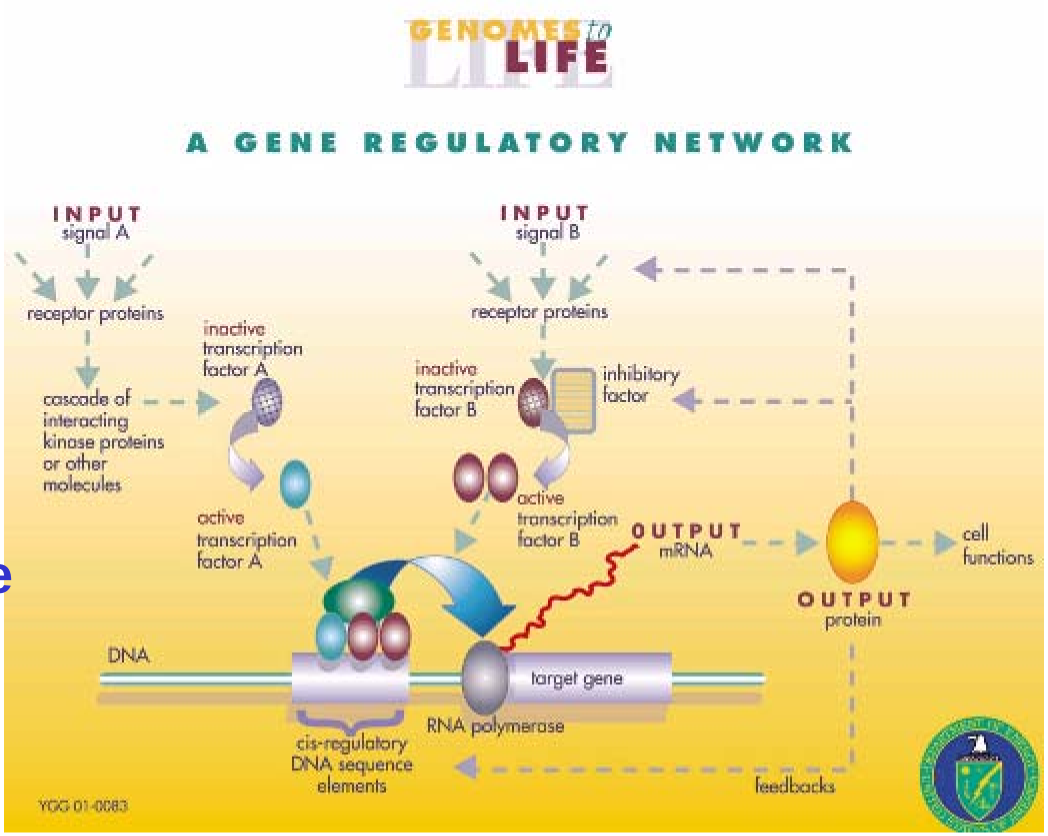
- **Biological Data is increasing rapidly**
- **Biologists need to locate required info**
- **Difficulties:**
  - Too much
  - Too heterogeneous
  - Too distributed
  - Too many errors
  - Due to mutation, need approximate search



# Cis-Regulatory DNAs

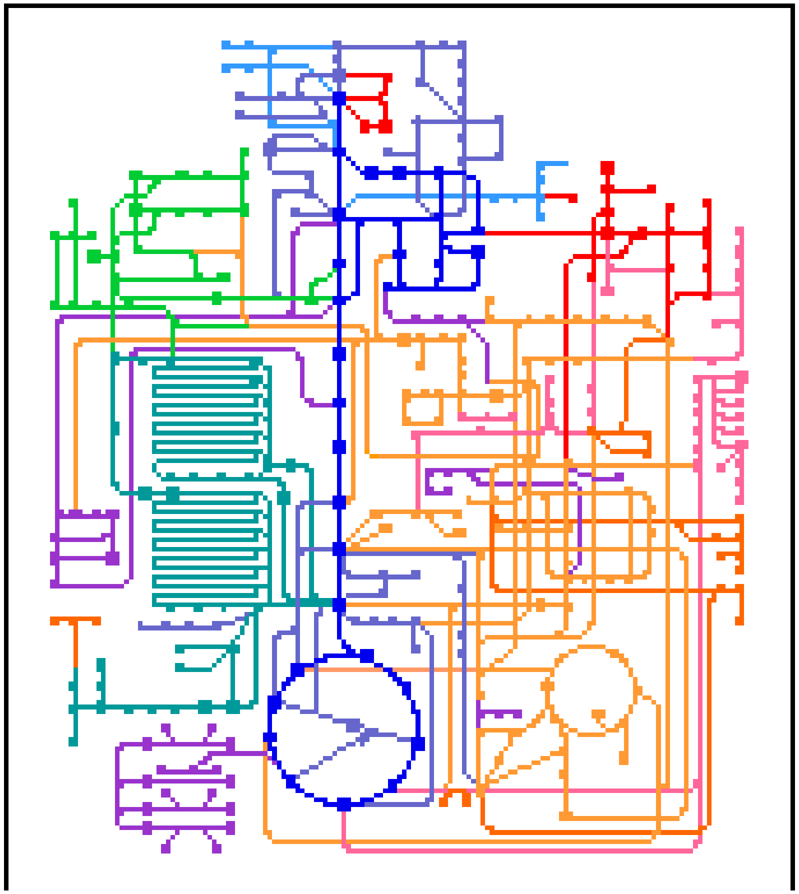
- **Cis-regulatory DNAs control whether genes should express or not**
- **Cis-regulatory may locate in promoter region, intron, or exon**
- **Finding and understanding cis-regulatory DNAs is one of the key problem in coming years**

Image credit: US DOE



# Gene Networks

- Inside a cell is a complex system
- Expression of one gene depends on expression of another gene
- Such interactions can be represented using gene network
- Understanding such networks helps identify association betw genes & diseases



# Protein/RNA structure prediction

- **Structure of Protein/RNA is essential to its functionality**
- **Important to have some ways to predict the structure of a protein/RNA given its sequence**
- **This problem is important & it is always considered as a “grand challenge” problem in bioinformatics**

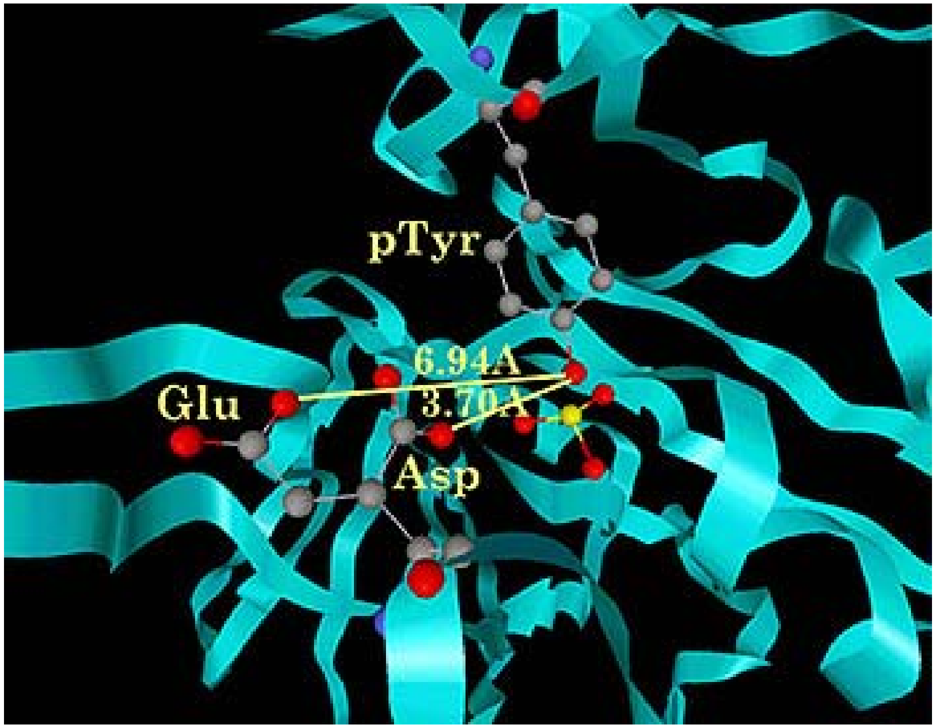


Image credit: Kolatkar

# Evolutionary Tree Reconstruction

- **Protein/RNA/DNA mutates**
- **Evolutionary Tree studies evolutionary relationship among set of protein/RNA/DNAs**
- **Figures out origin of species**

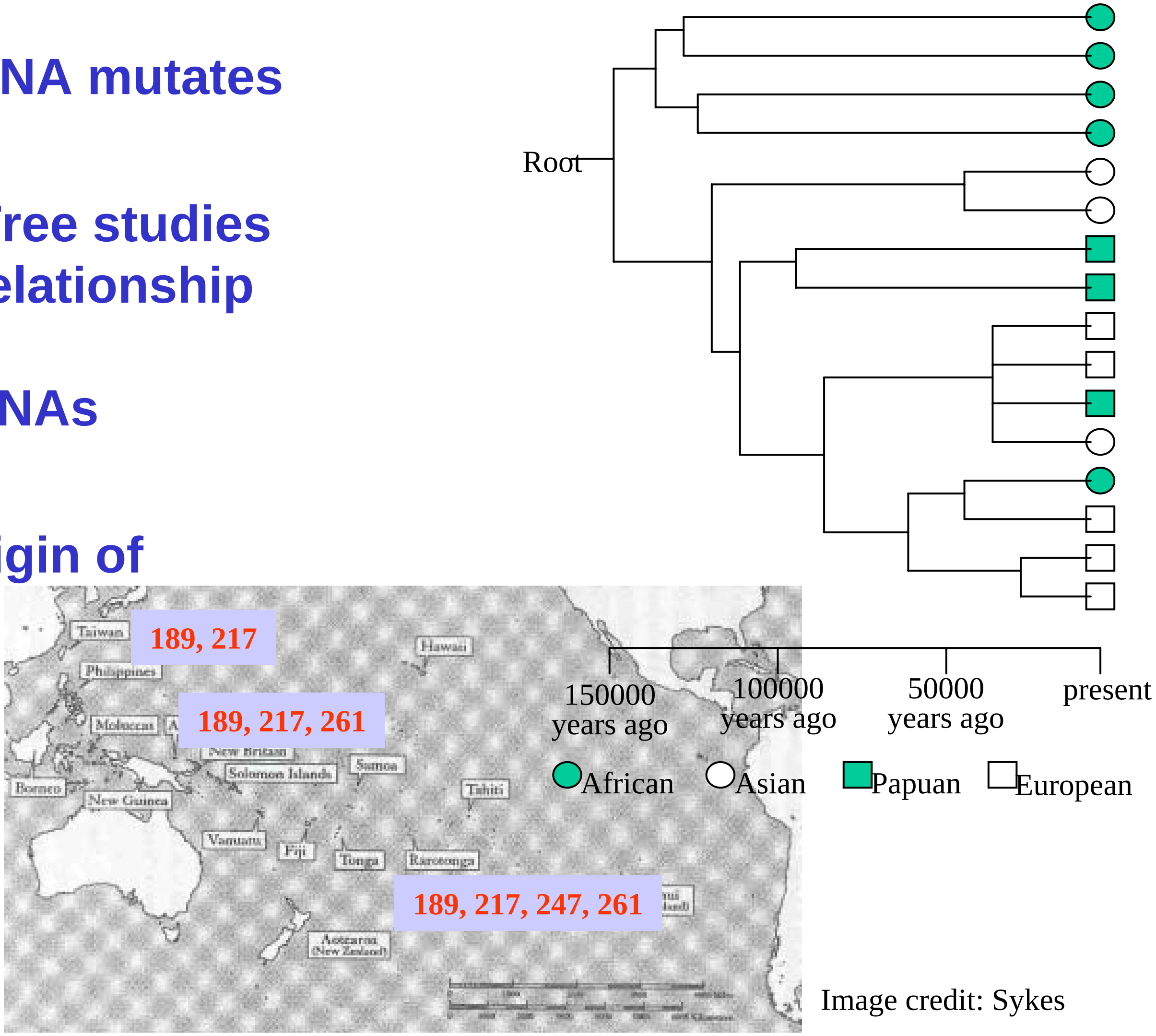


Image credit: Sykes

# Breast Cancer Outcome Prediction

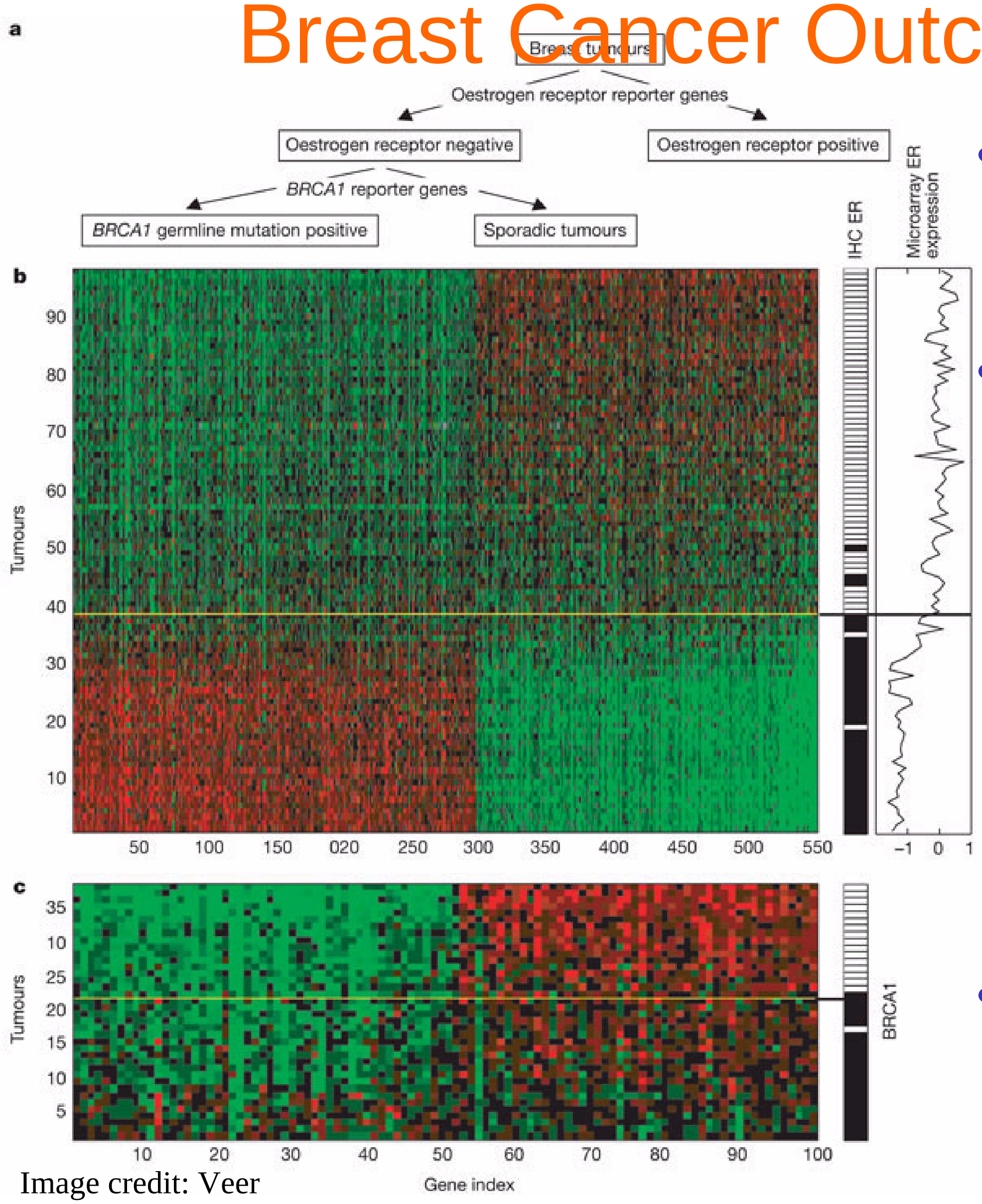


Image credit: Veer

- **Van't Veer et al., *Nature* 415:530-536, 2002**

- **Training set contains 78 patient samples**
  - 34 patients develop distance metastases in 5 yrs
  - 44 patients remain healthy from the disease after initial diagnosis for >5 yrs

- **Testing set contains 12 relapse & 7 non-relapse samples**

# Commonly Used Data Sources

# Type of Biological Databases

- **Micro Level**
  - Contain info on the composition of DNA, RNA, Protein Sequences
- **Macro Level**
  - Contain info on interactions
    - **Gene Expression**
    - **Metabolites**
    - **Protein-Protein Interaction**
    - **Biological Network**
- **Metadata**
  - Ontology
  - Literature

Exercise: Name a protein sequence database and a DNA sequence data

# Transcriptome Database

- **Complete collection of all possible mRNAs (including splice variants) of an organism**
- **Regions of an organism's genome that get transcribed into messenger RNA**
- **Transcriptome can be extended to include all transcribed elements, including non-coding RNAs used for structural and regulatory purposes**

Exercise: Name a transcriptome database

# Gene Expression Databases

- **Detect what genes are being expressed or found in a cell of a tissue sample**
- **Single-gene analysis**
  - Northern Blot
  - In Situ Hybridization
  - RT-PCR
- **Many Genes: High Throughput Arrays**
  - cDNA Microarray
  - Affymetrix GeneChip® Microarray

Exercise: Name a gene expression database

# Metabolites Database

- **A metabolite is an organic compound that is a starting material in, an intermediate in, or an end product of metabolism**

- **Metabolites dataset are also generated from mass spectrometry which measure the mass the these simple molecules, thus allowing us to estimate what are the metabolites in a tissue**

- **Starting metabolites:**
  - Small, of simple structure, absorbed by the organism as food
  - E.g., vitamins and amino acids
- **Intermediary metabolites:**
  - The most common metabolites
  - May be synthesized from other metabolites, or broken down into simpler compounds, often with the release of chemical energy
  - E.g., glucose
- **End products of metabolism**
  - Final result of the breakdown of other metabolites
  - Excreted from the organism without further change
  - E.g., urea, carbon dioxide

# Protein-Protein Interaction Databases

- **Proteins are true workhorses**
  - Lots of the cell's activities are performed thru PPI including message passing, gene regulation, etc.
- **Function of a protein also depends on proteins it interact with**
- **Methods for generating PPI database include:**
  - biochemical purifications, yeast-two hydrid, synthetic lethals, in silico predictions, mRNA-co-expression
- **Contain many false positives & false negatives**

Exercise: Name a PPI database

Any Question?

NUS
National University
of Singapore

# Acknowledgements

- **Most of the slides used in this lecture are based on original slides created by**
  - Ken Sung
  - Anthony Tung
- **Inaccuracies and errors are mine**

# References

- **S.K.Ng, "Molecular Biology for the Practical Bioinformatician", *The Practical Bioinformatician*, Chapter 1, pages 1—30, WSPC, 2004**
- **DOE HGP Primer**