Lecture 1
Introduction to Micorarrays and Concepts of Molecular Biology

M. Saleet Jafri

Program in Bioinformatics and Computational Biology

George Mason University

Lecture 1

Overview of Molecular and Cellular Biology

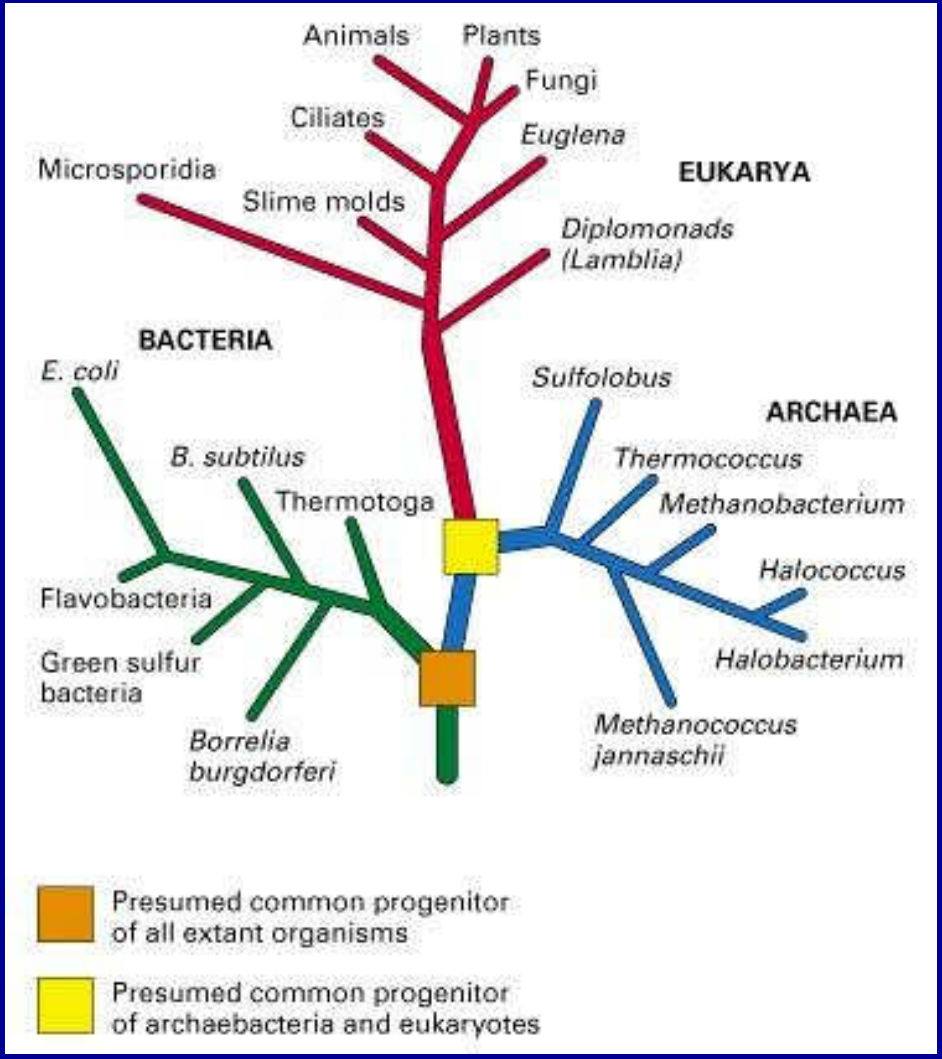
Biological References

Part I: Molecular Biology Review

Where do biological sequences come from?

Life

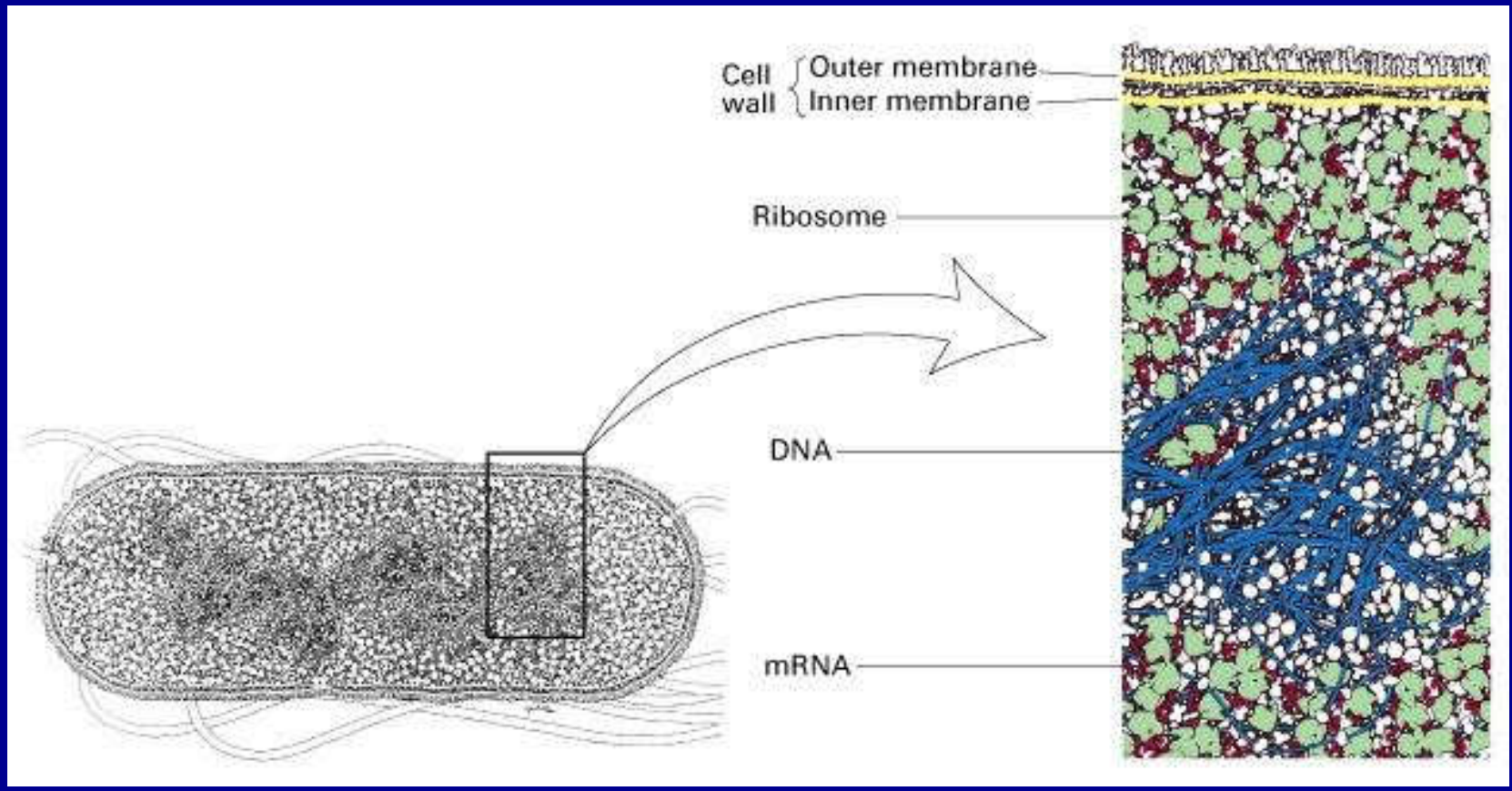
Terrestrial Life



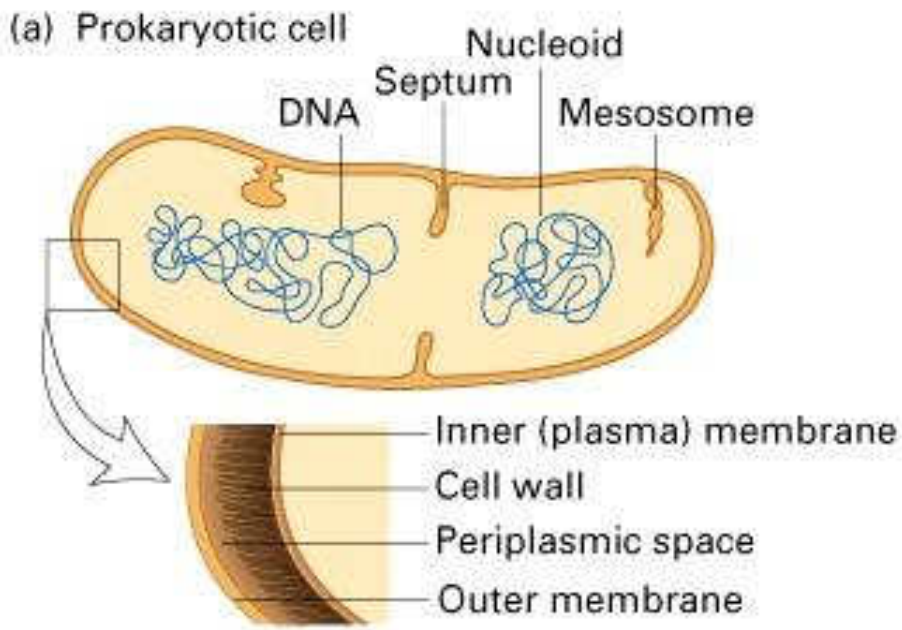
Cell types

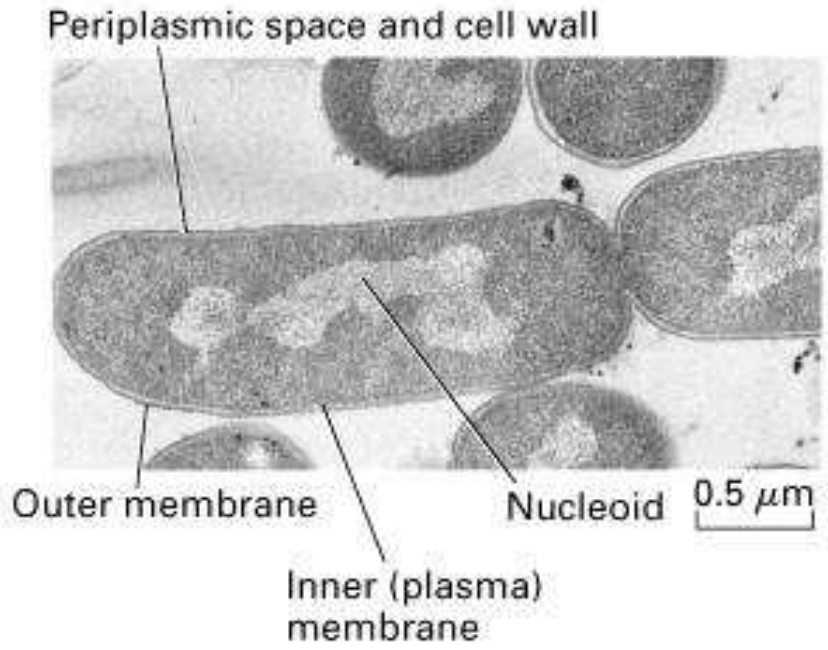
Prokaryotes

Prokaryotic Cell



Prokaryotic Cell

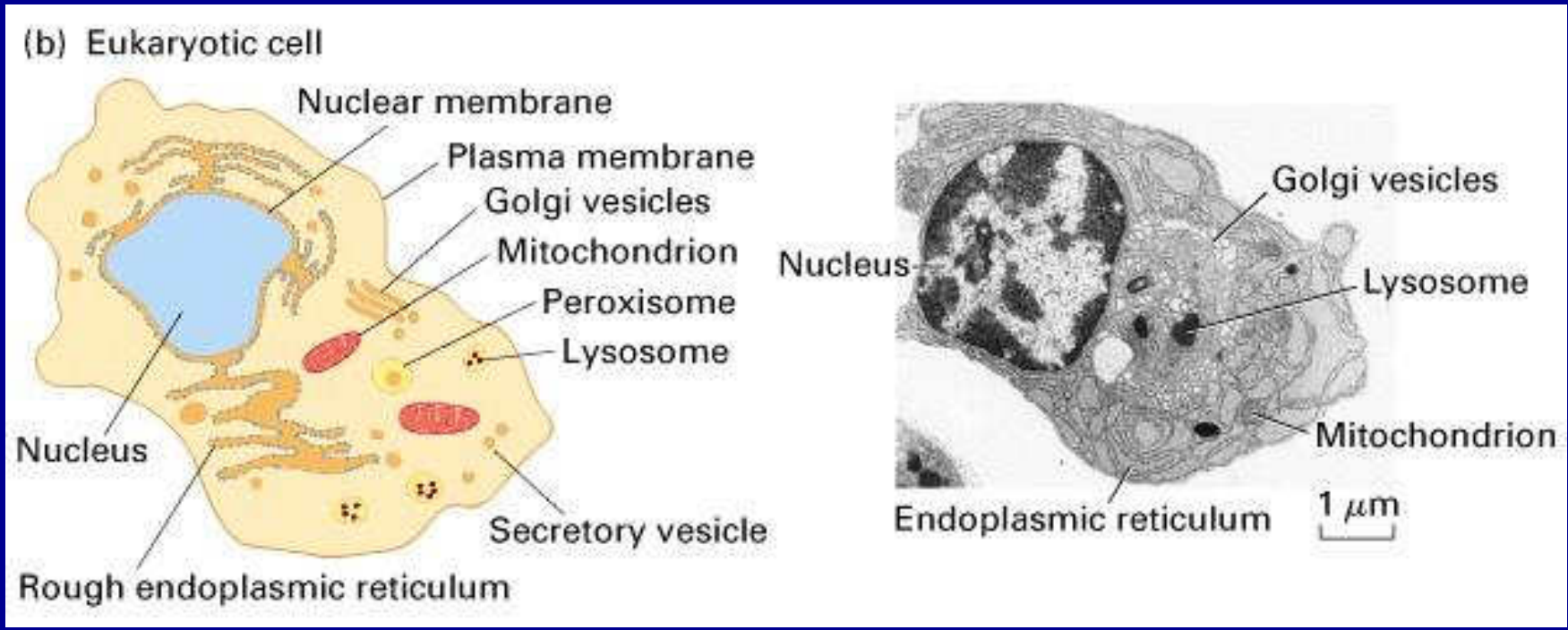




Eukaryotes

Eukaryotic Cell

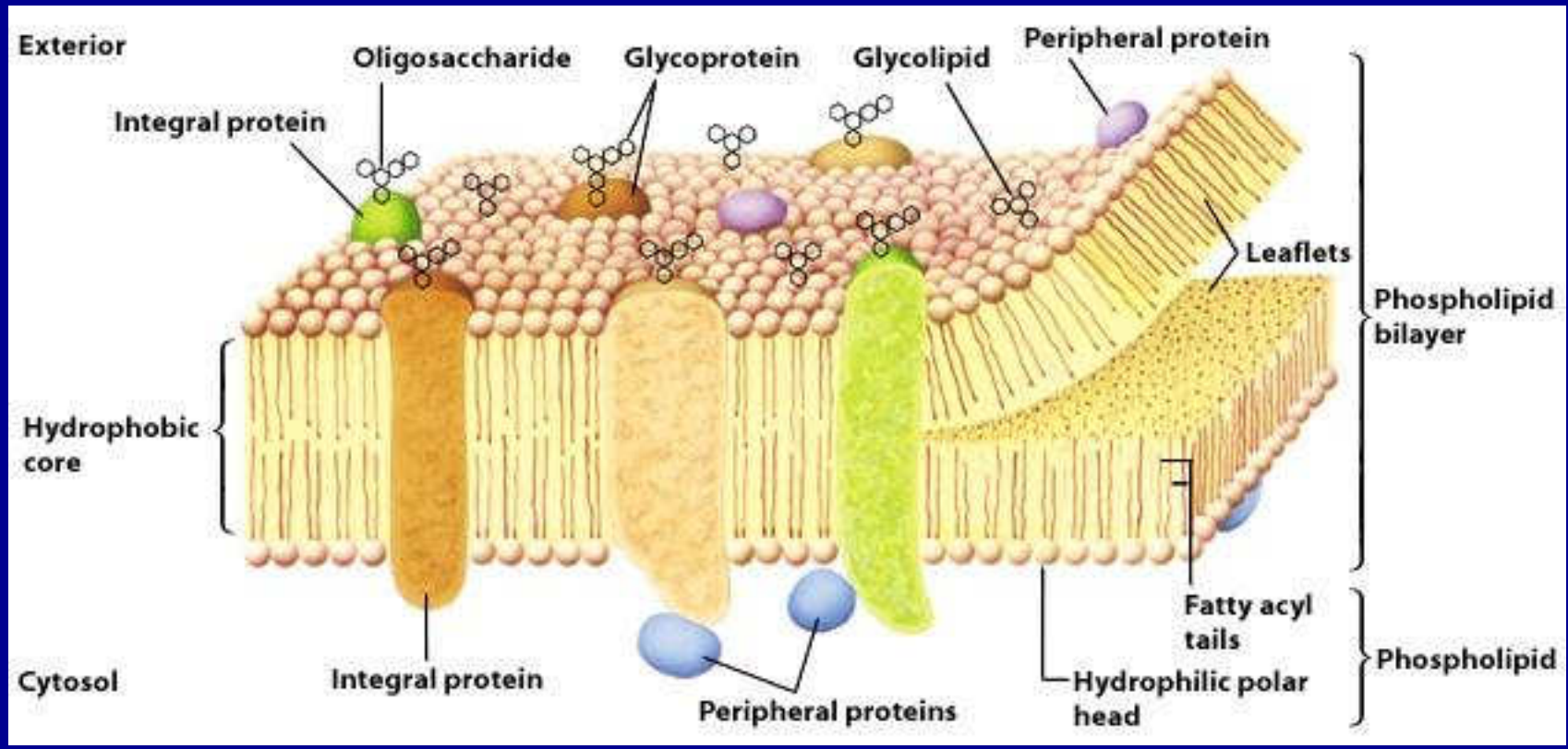
Eukaryotic Cell



Eukaryotic Cell Organelles

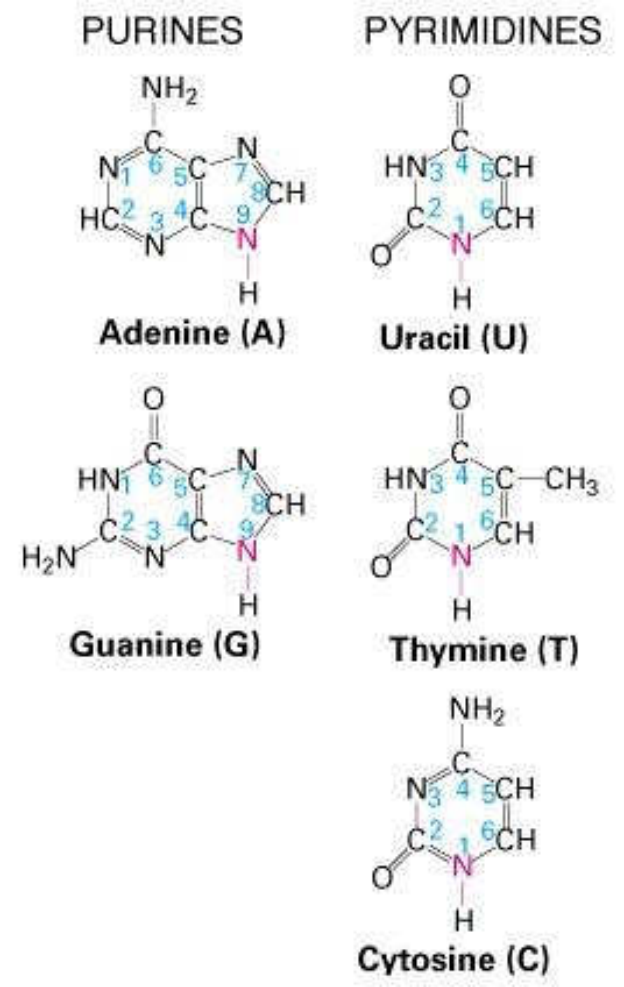
Eukaryotic Cell Organelles

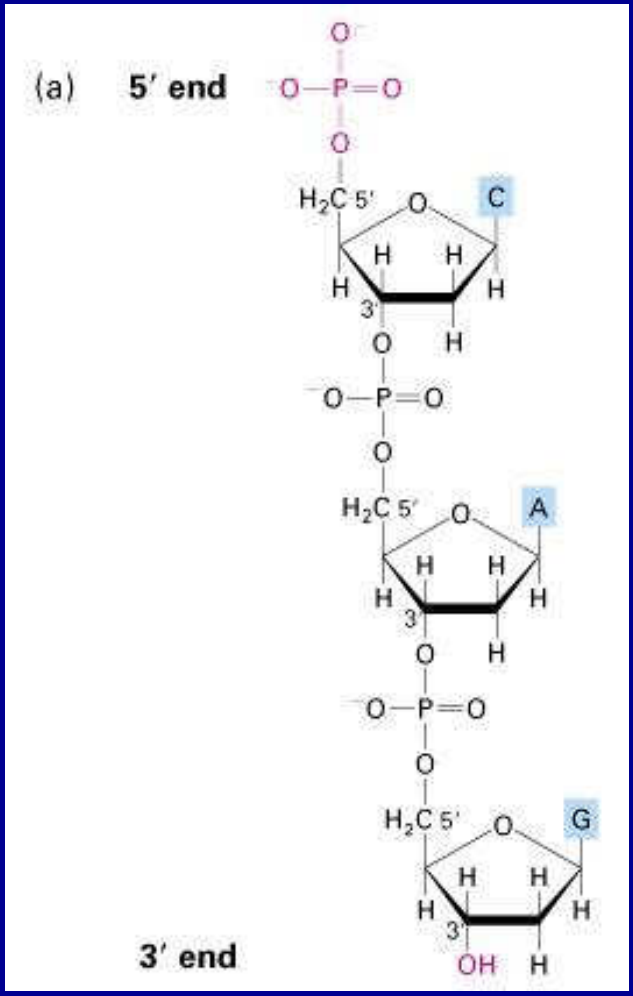
Eukaryotic Membrane



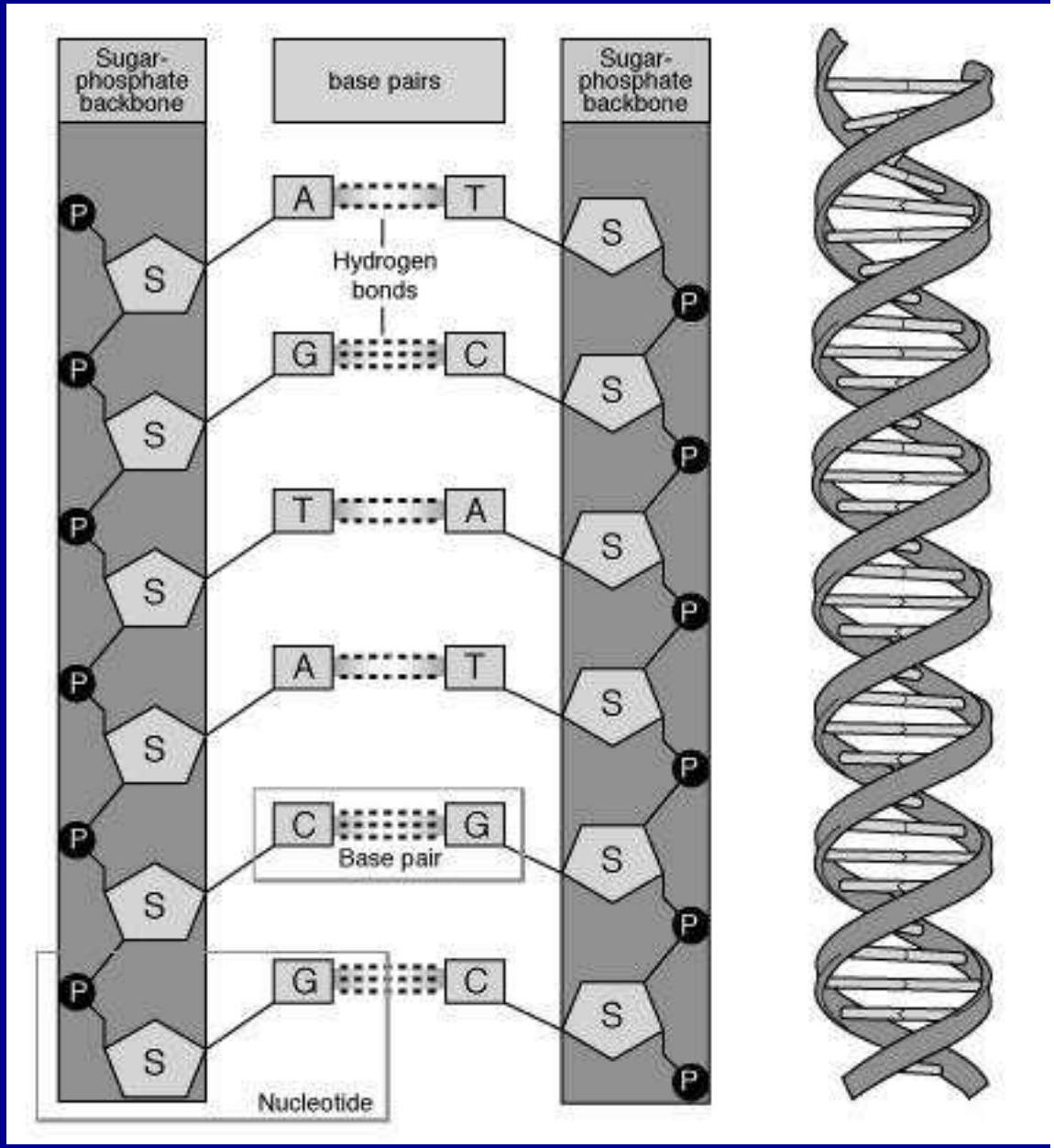
Nucleic Acids

Nucleic Acid Structure





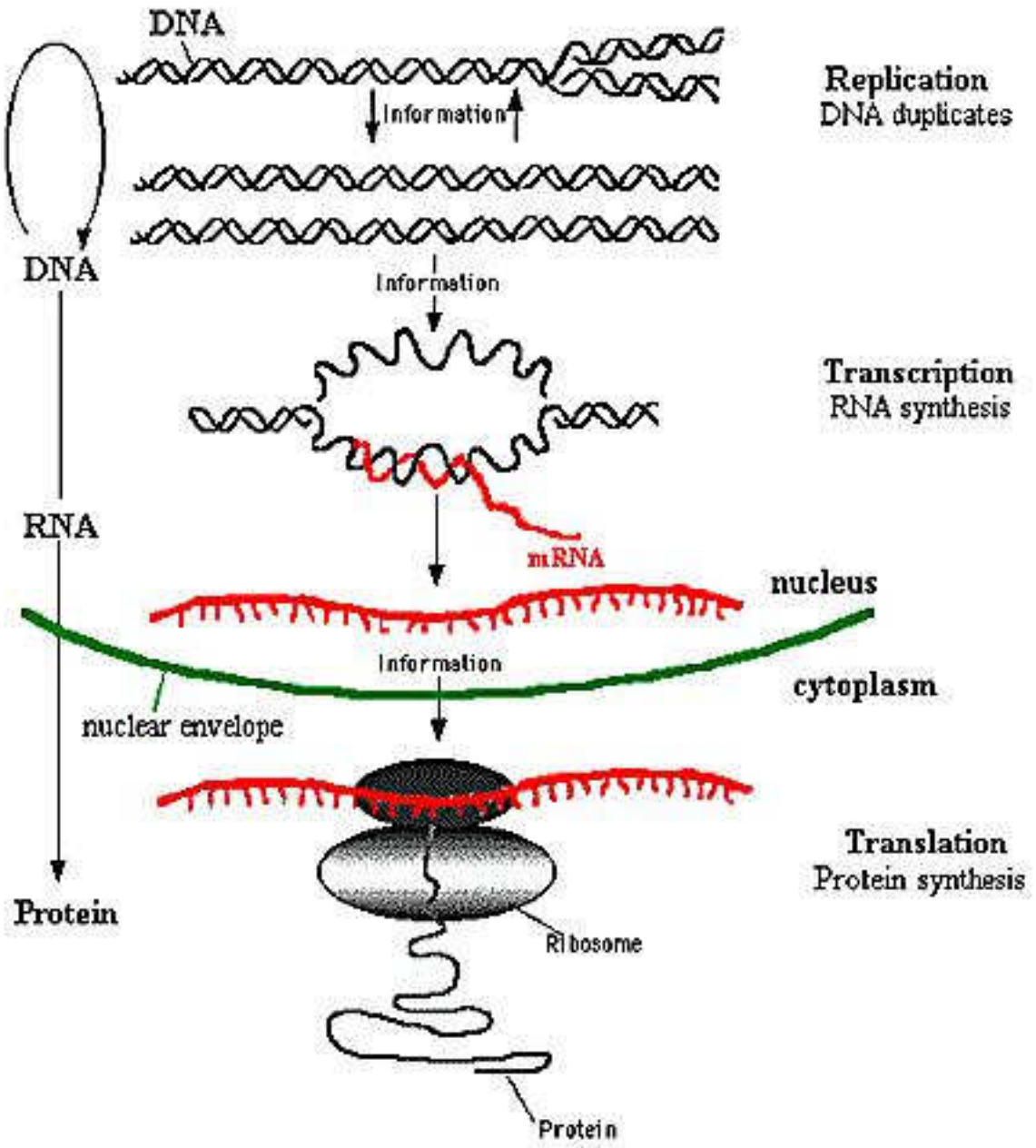
DNA


Sugar-phosphate backbone
base pairs
Sugar-phosphate backbone
A
T
Hydrogen bonds
G
C
T
A
A
T
C
G
Base pair
G
C
Nucleotide
P
S

RNA

Central Dogma

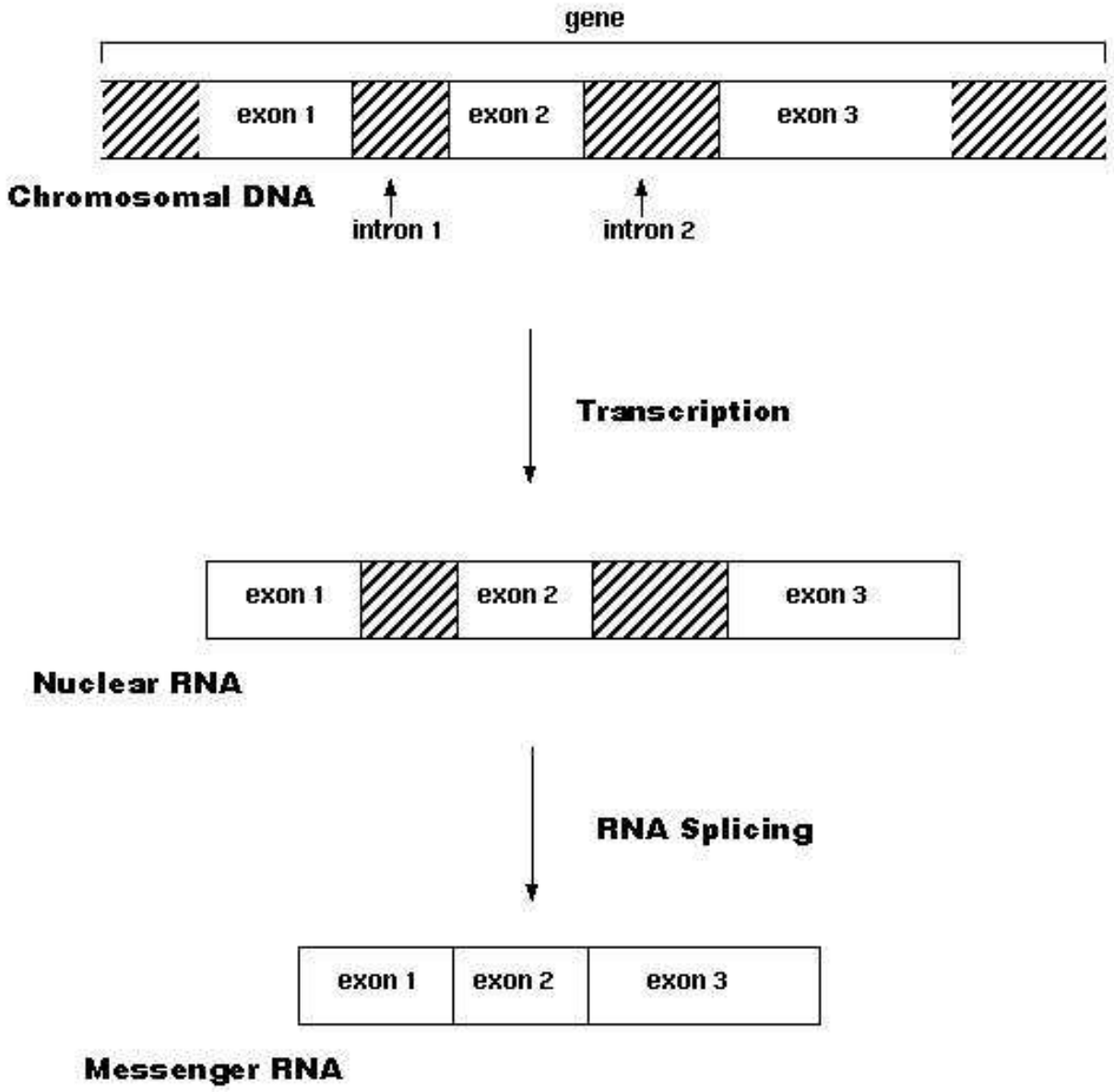
DNA



The Central Dogma of Molecular Biology

Gene Transcription or DNA Transcription

Transcription of DNA to Messenger RNA

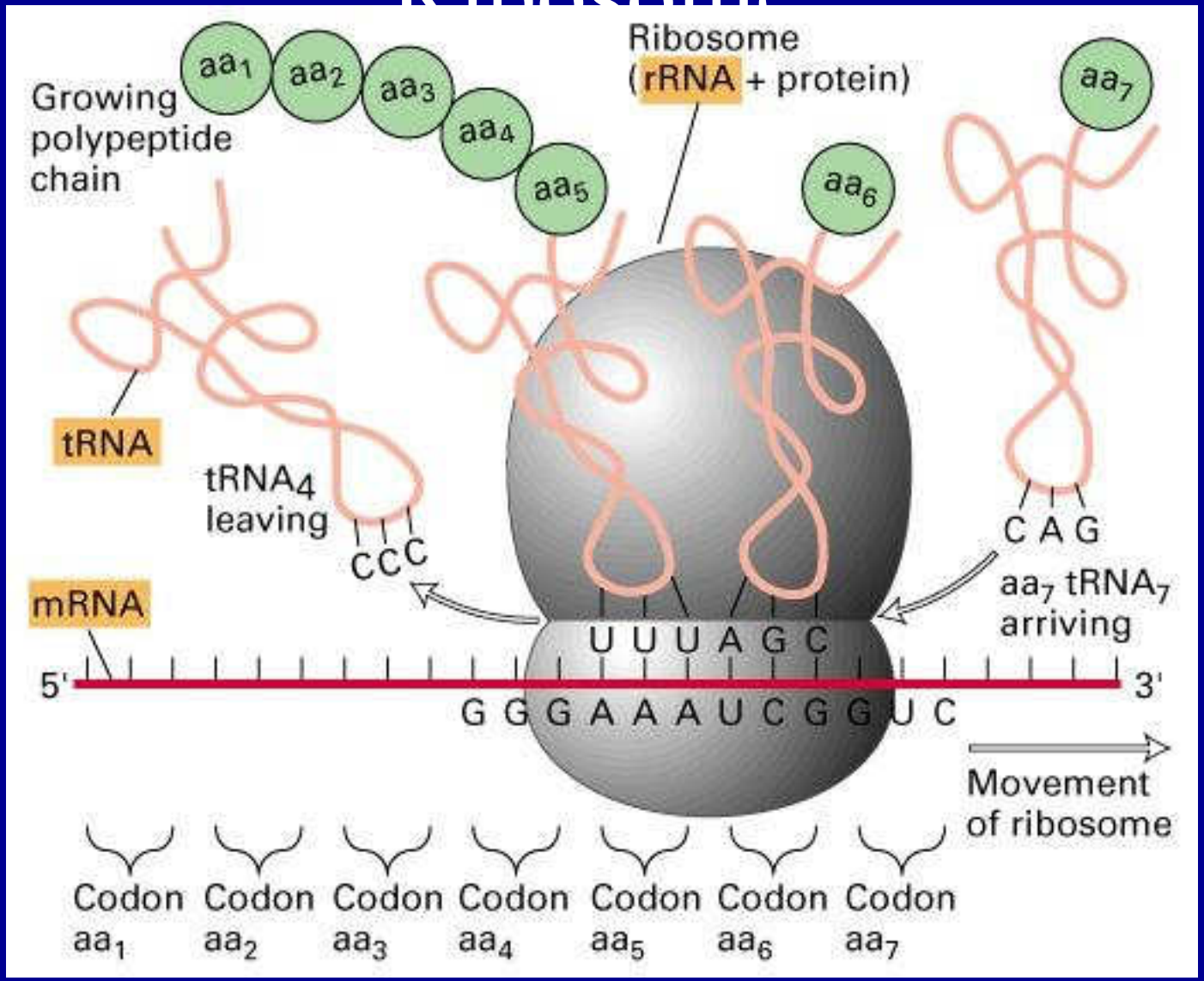


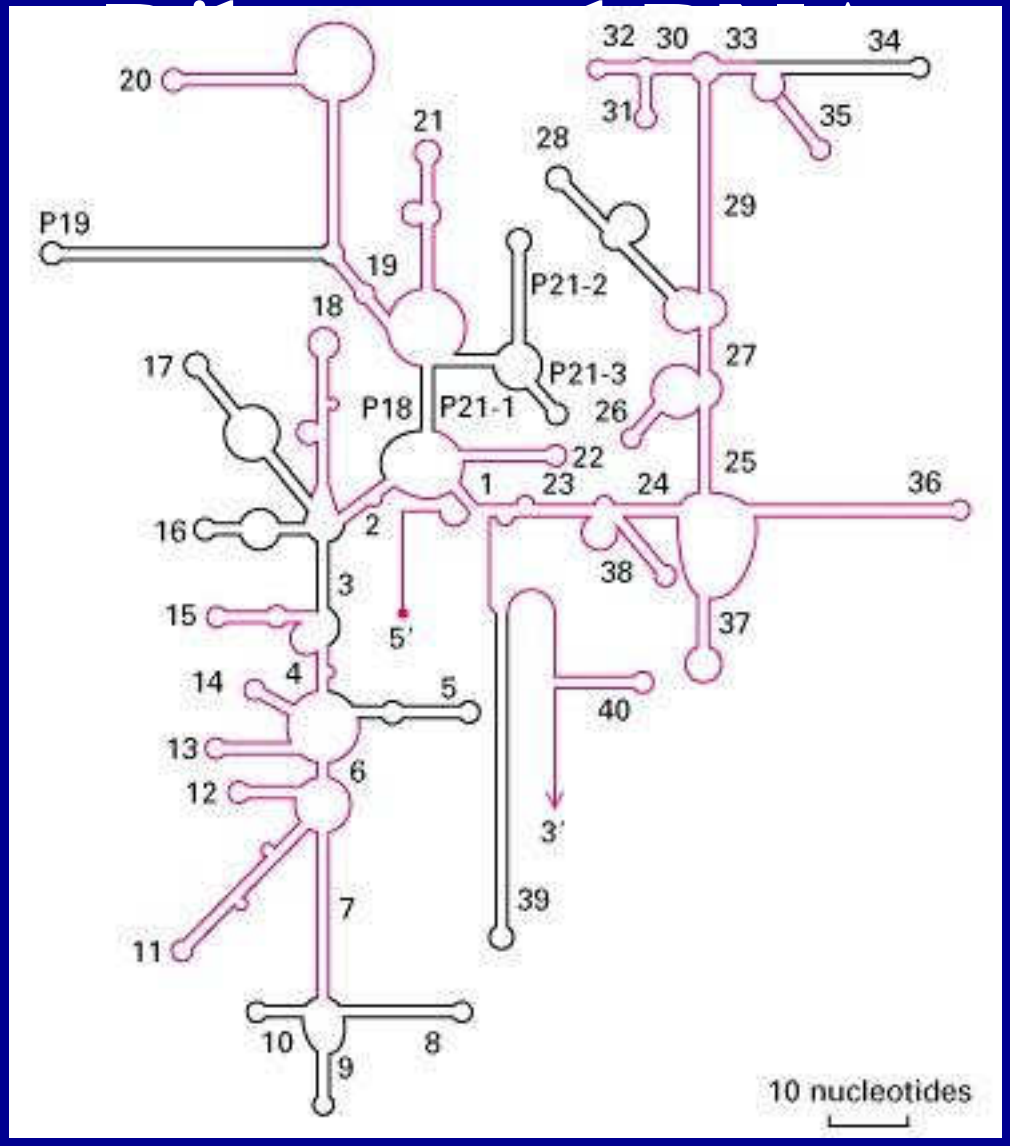
Translation

Gene Translation

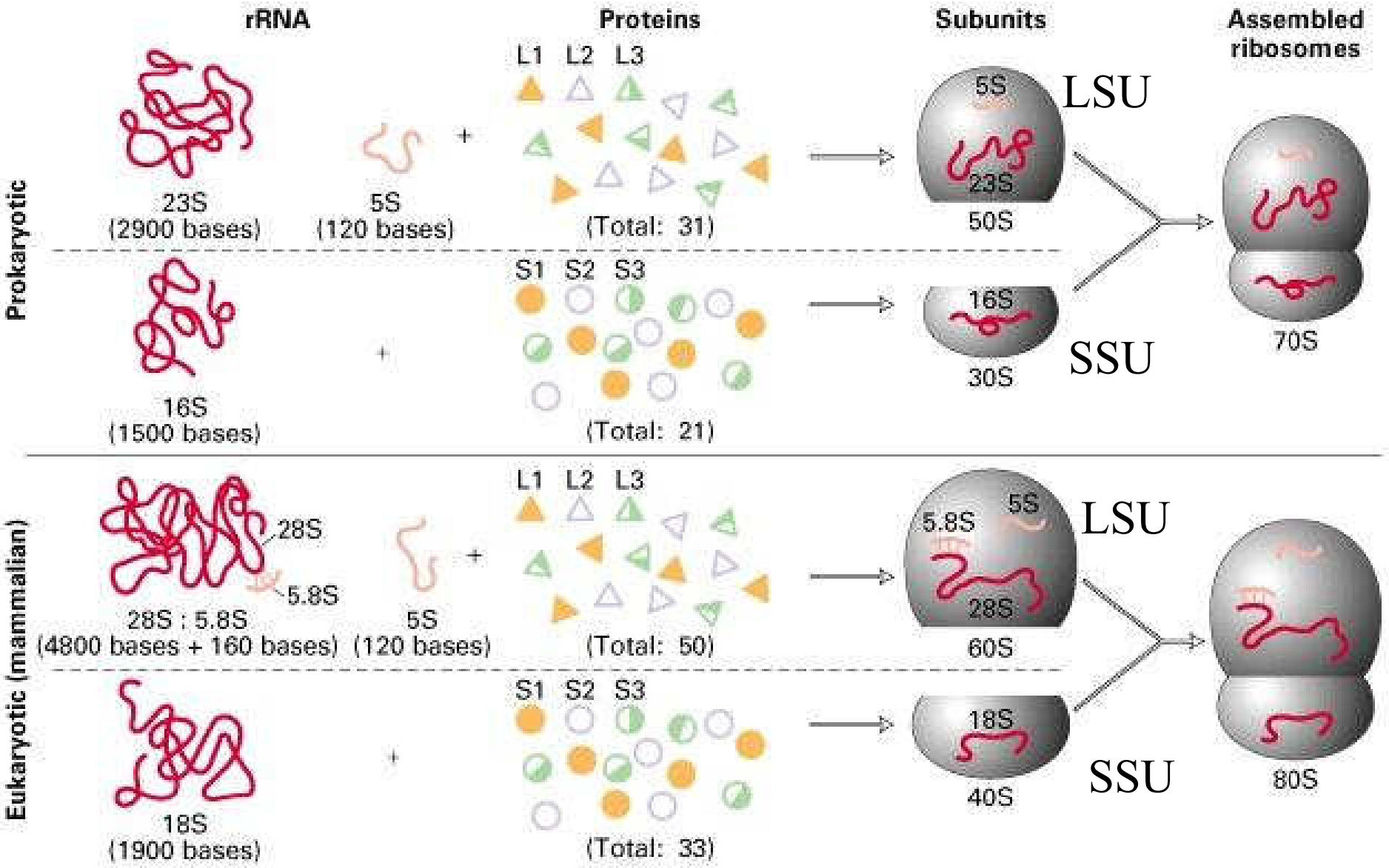
Translation

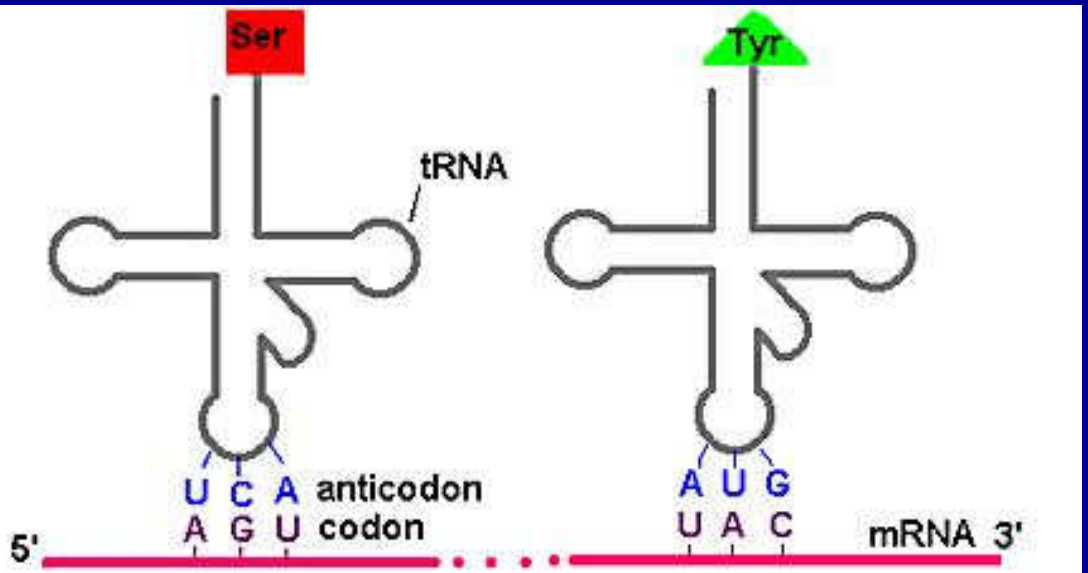
Ribosome


Growing polypeptide chain
aa1
aa2
aa3
aa4
aa5
Ribosome (rRNA + protein)
aa6
aa7
tRNA
tRNA4 leaving
C C C
C A G
aa7 tRNA7 arriving
mRNA
5'
U U U A G C
3'
G G G A A A U C G G U C
Movement of ribosome
Codon aa1
Codon aa2
Codon aa3
Codon aa4
Codon aa5
Codon aa6
Codon aa7


20
21
32
30
33
34
31
35
28
29
P19
19
P21-2
18
27
17
P21-3
P18
P21-1
26
22
25
1
23
24
36
16
2
38
3
15
5'
37
4
14
5
40
13
6
12
3'
39
7
11
10
8
9
10 nucleotides

Eukaryotic and Prokaryotic

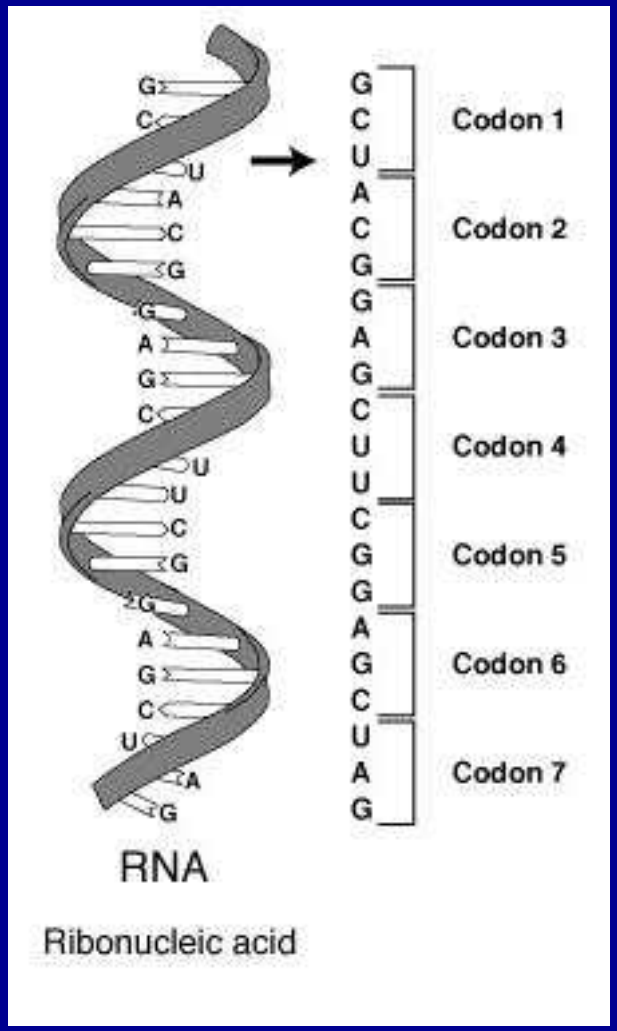




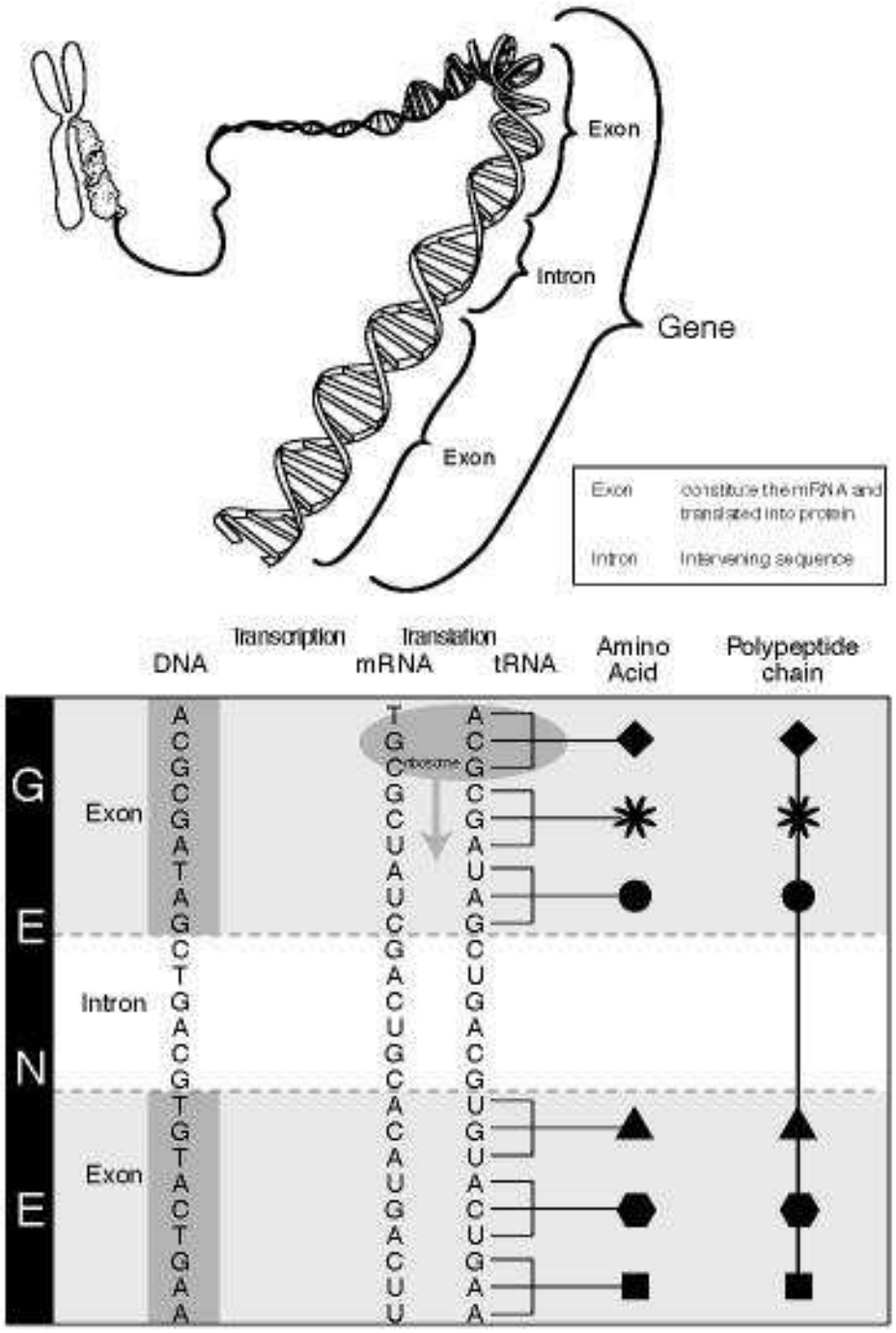
2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

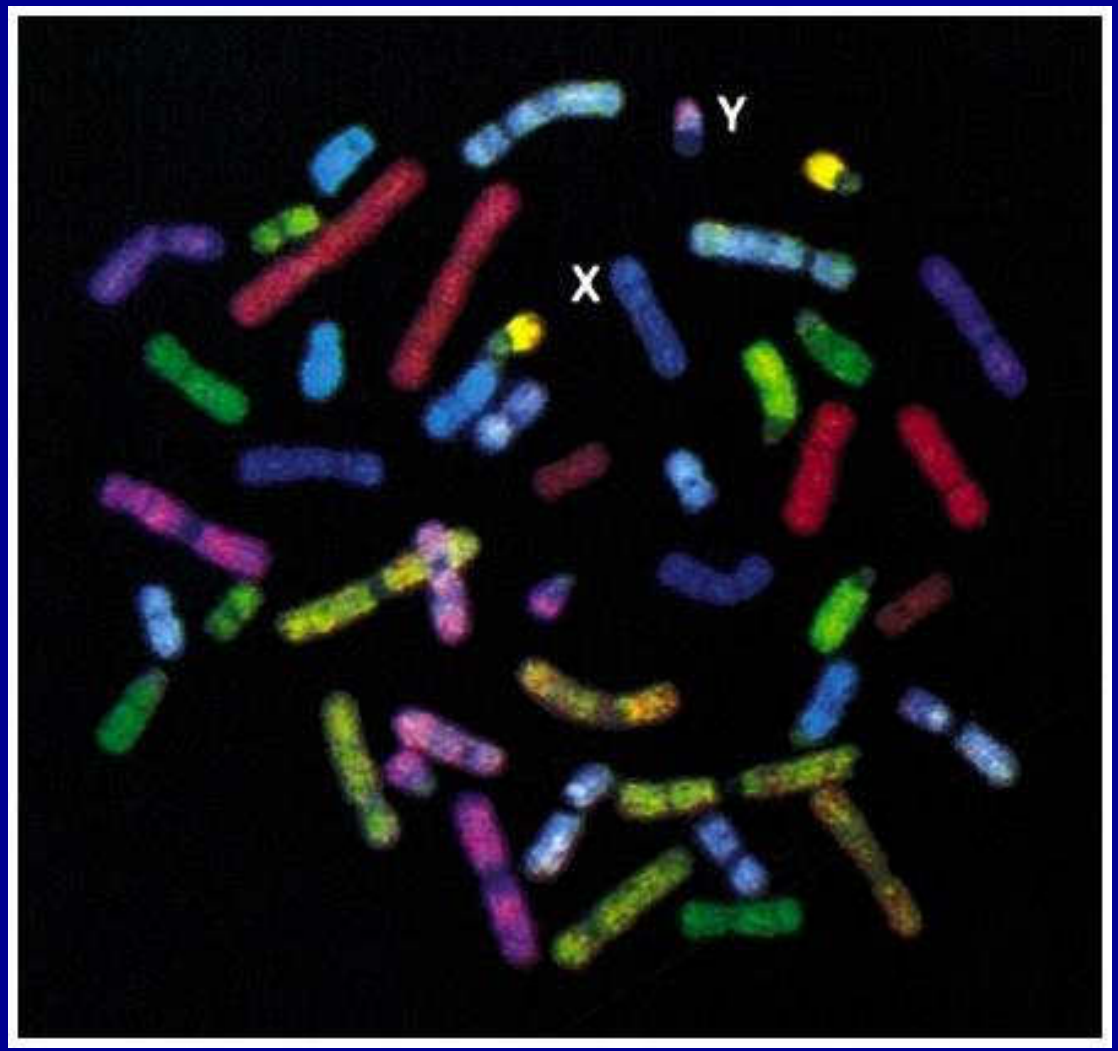

G
C
U
A
C
G
G
A
G
C
U
U
C
G
G
A
G
C
U
A
G
Codon 1
Codon 2
Codon 3
Codon 4
Codon 5
Codon 6
Codon 7
RNA
Ribonucleic acid

DNA Structure

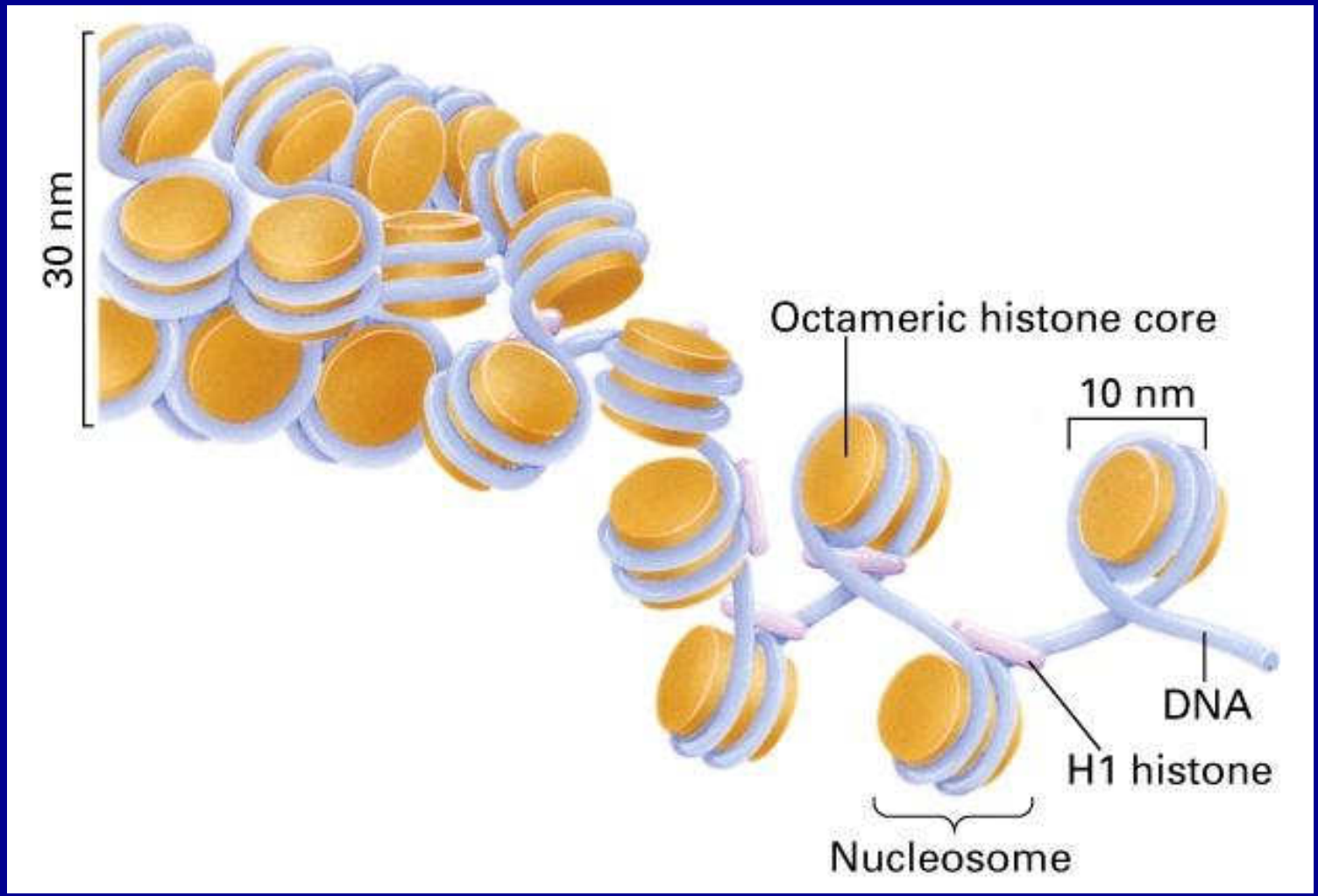

Exon
Intron
Gene
Exon
Exon constitute the mRNA and translated into protein
Intron Intervening sequence
DNA
Transcription
mRNA
Translation
tRNA
Amino Acid
Polypeptide chain
GENE
Exon
Intron
Exon
ribosome

Chromosomes

Chromosomes



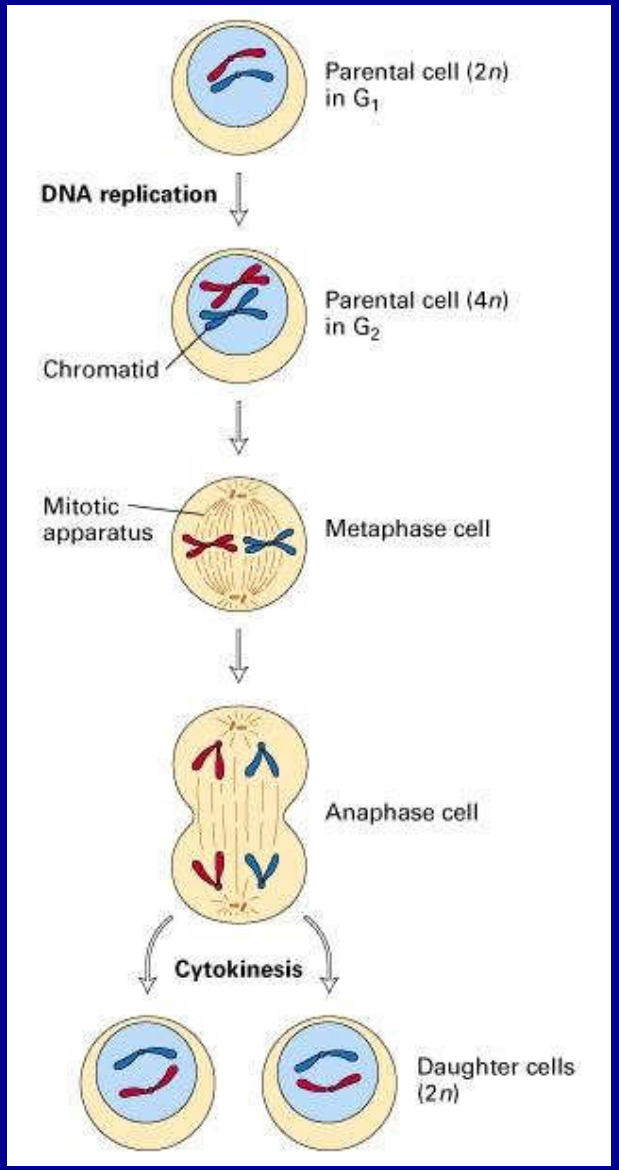
Chromatin Structure



Gene Coding and Replication

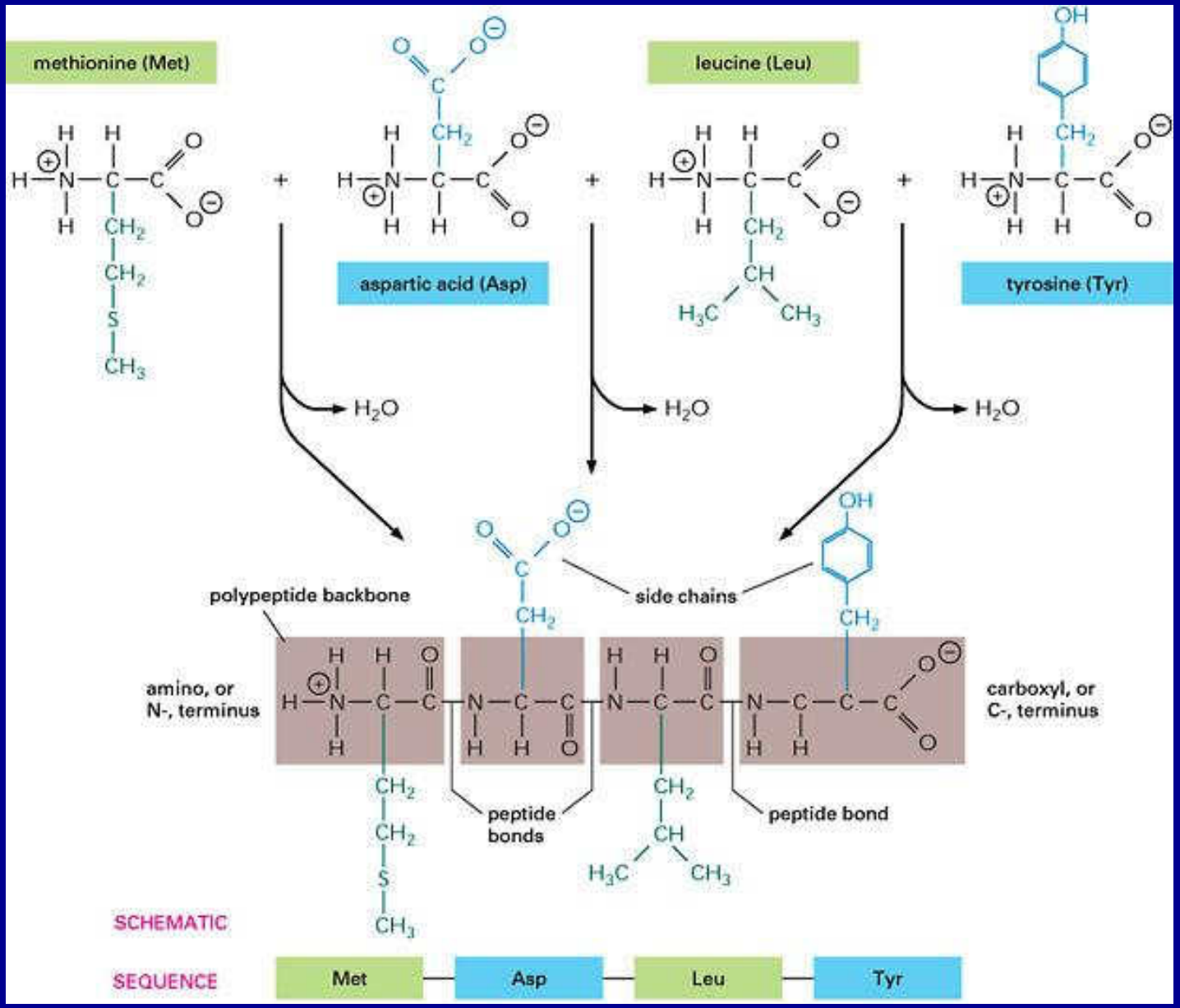
Gene Coding and Replication

Mitosis



QuickTime

Proteins


methionine (Met)
aspartic acid (Asp)
leucine (Leu)
tyrosine (Tyr)
H_2O
H_2O
H_2O
polypeptide backbone
side chains
amino, or
N-, terminus
carboxyl, or
C-, terminus
peptide
bonds
peptide bond
SCHEMATIC
SEQUENCE
Met
Asp
Leu
Tyr

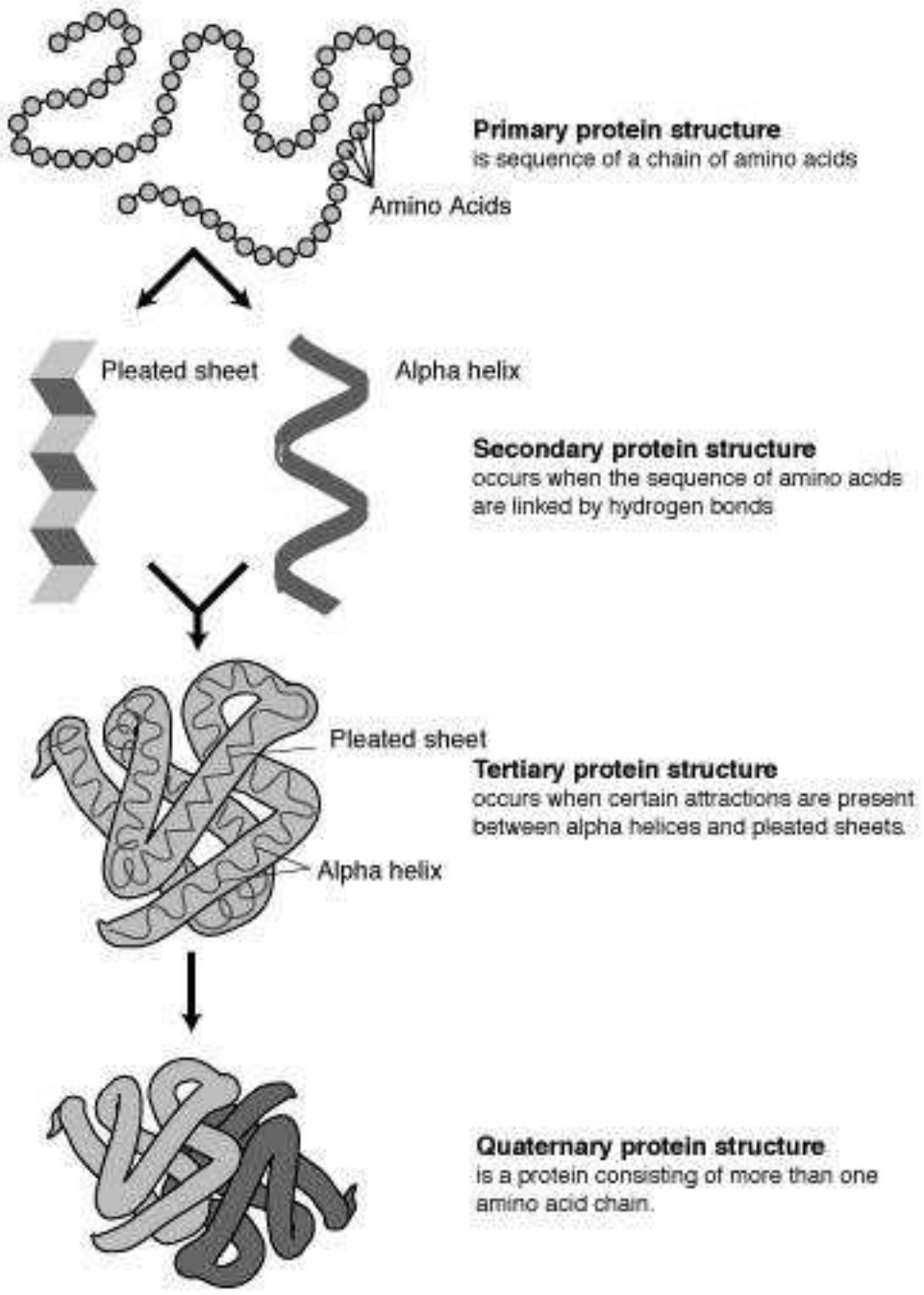
Protein Folding

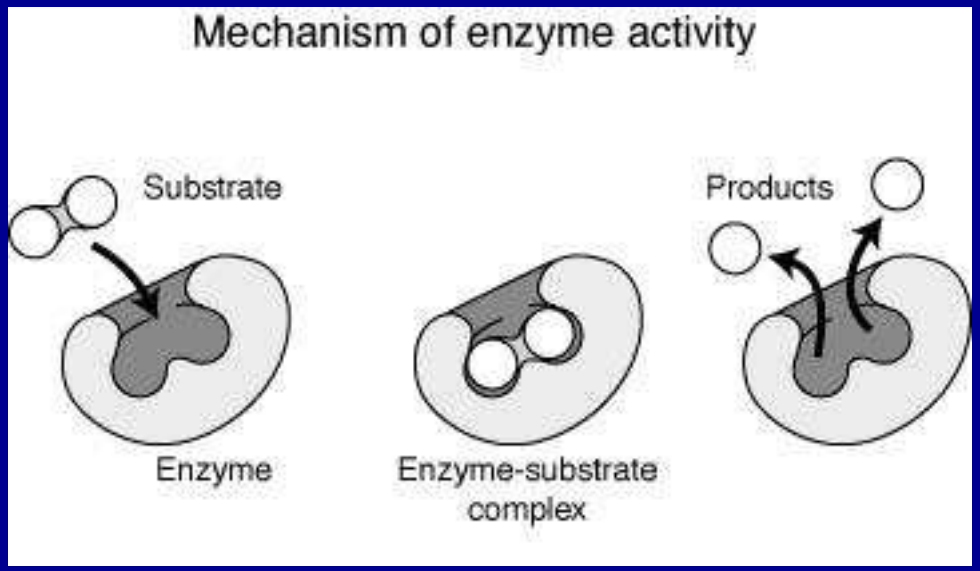
Primary structure: residue sequence

Secondary structure:
local structures
(Helices, sheets, loops)

Tertiary structure:
position of each atom

Quaternary structure:
how groups of proteins pack together




Mechanism of enzyme activity
Substrate
Products
Enzyme
Enzyme-substrate complex

The protein

Cell Signaling and Biochemical Pathways

Molecular Biology Summary

Polymerase Chain Reaction

In order to make enough RNA for microarray experiments, PCR is used to amplify the amount of RNA

Most PCR methods typically amplify nulceic fragments of up to 10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size.

Polymerase Chain Reaction

A basic PCR set up requires several components and reagents. These components include:

Polymerase Chain Reaction

The PCR is commonly carried out in a reaction volume of 10-200

Polymerase Chain Reaction

The PCR usually consists of a series of 20 to 40 repeated temperature changes called cycles; each cycle typically consists of 2-3 discrete temperature steps. Most commonly PCR is carried out with cycles that have three temperature steps (Fig. 2). The cycling is often preceded by a single temperature step (called *hold*) at a high temperature (>90°C), and followed by one hold at the end for final product extension or brief storage. The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters. These include the enzyme used for DNA synthesis, the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (Tm) of the primers

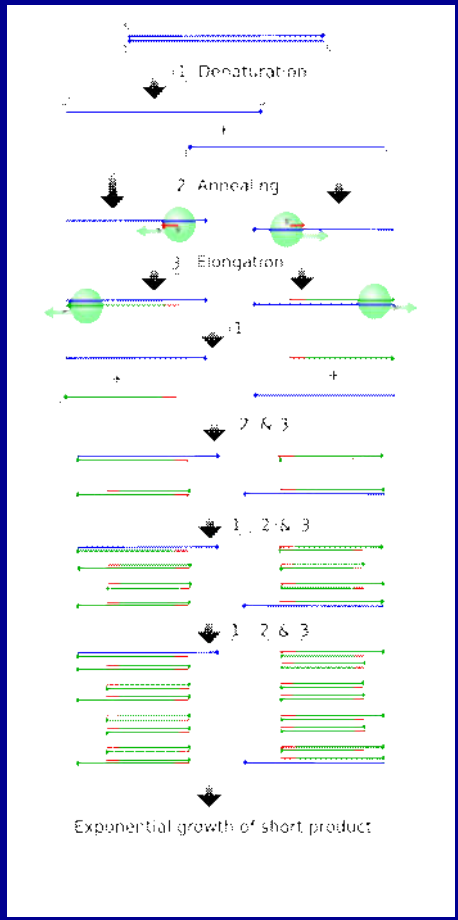
From Wikipedia

Polymerase Chain Reaction

Denaturation step: This step is the first regular cycling event and consists of heating the reaction to 94-98°C for 20-30 seconds. It causes melting of DNA template and primers by disrupting the hydrogen bonds between complementary bases of the DNA strands, yielding single strands of DNA.

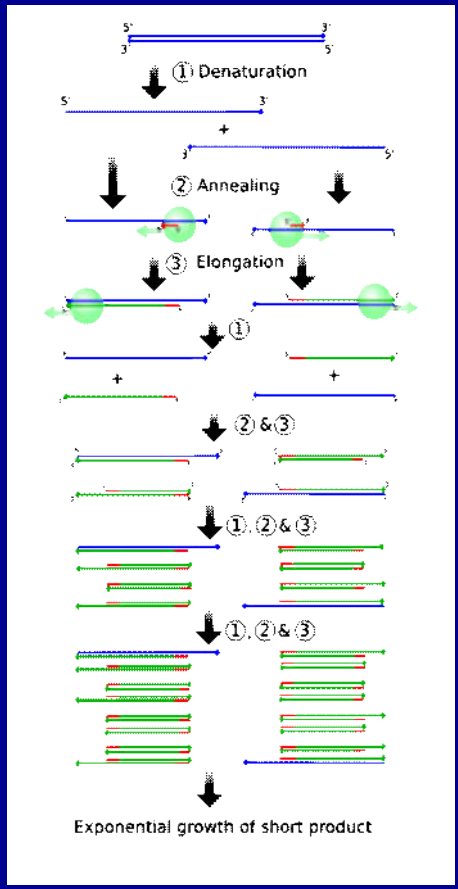
Annealing step: The reaction temperature is lowered to 50-65°C for 20-40 seconds allowing annealing of the primers to the single-stranded DNA template. Typically the annealing temperature is about 3-5 degrees Celsius below the Tm of the primers used. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA synthesis.

From Wikipedia



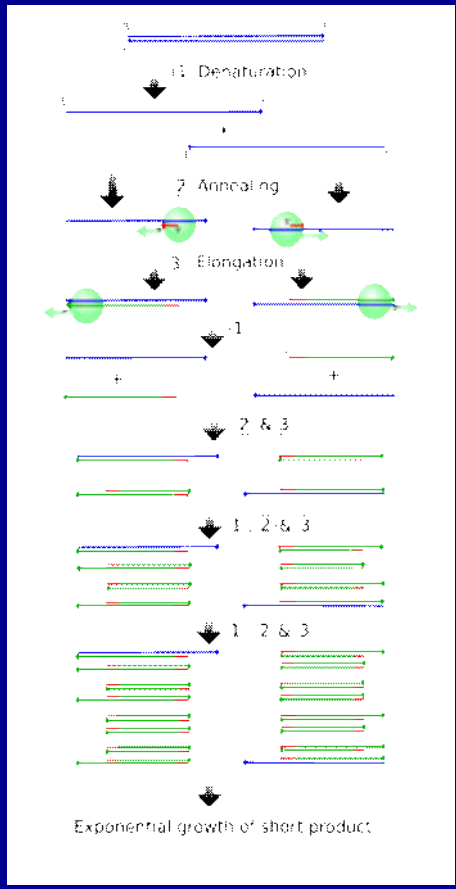
Polymerase Chain Reaction

Extension/elongation step: The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75-80°C,[10][11] and commonly a temperature of 72°C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the DNA polymerase used and on the length of the DNA fragment to be amplified. As a rule-of-thumb, at its optimum temperature, the DNA polymerase will polymerize a thousand bases per minute. Under optimum conditions, i.e., if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential (geometric) amplification of the specific DNA fragment.



Polymerase Chain Reaction

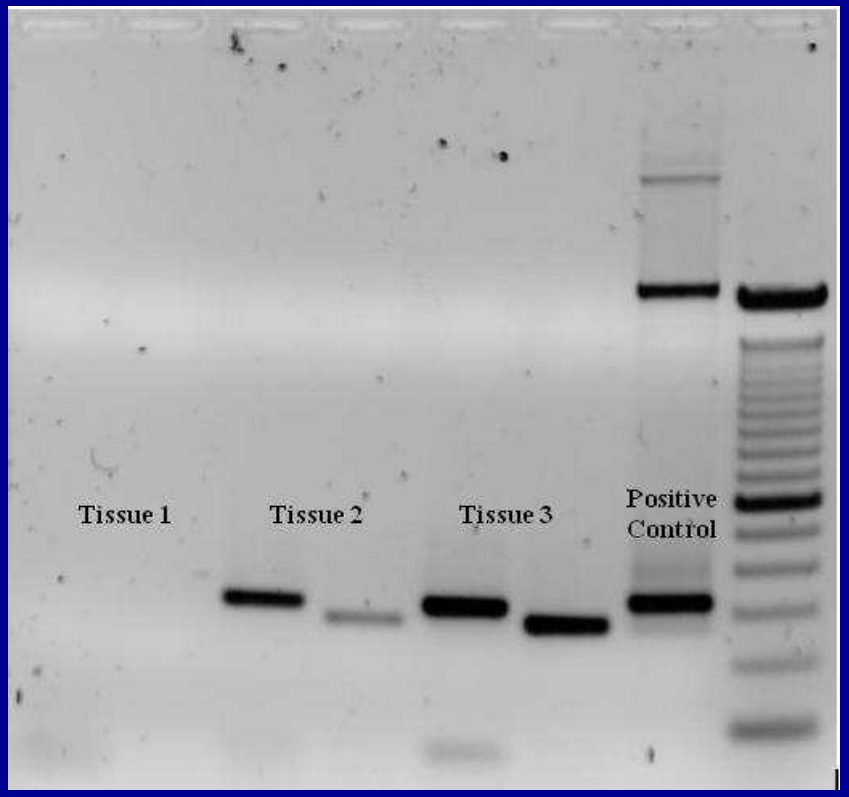
Final hold: This step at 4-15°C for an indefinite time may be employed for short-term storage of the reaction.



Polymerase Chain Reaction

Ethidium bromide-stained PCR products after gel electrophoresis. Two sets of primers were used to amplify a target sequence from three different tissue samples. No amplification is present in sample #1; DNA bands in sample #2 and #3 indicate successful amplification of the target sequence. The gel also shows a positive control, and a DNA ladder containing DNA fragments of defined length for sizing the bands in the experimental PCRs.

To check whether the PCR generated the anticipated DNA fragment (also sometimes referred to as the amplimer or amplicon), agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size, run on the gel alongside the PCR products.



Lecture 1

Introduction to Microarrays

Microarray Data Analysis

Gene chips allow the simultaneous monitoring of the expression level of thousands of genes. Many statistical and computational methods are used to analyze this data. These include:

What is Microarray Data?

In spite of the ability to allow us to simultaneously monitor the expression of thousands of genes, there are some liabilities with micorarray data. Each micorarray is very expensive, the statistical reproducibility of the data is relatively poor, and there are a lot of genes and complex interactions in the genome.

Microarray data is often arranged in an n x m matrix **M** with rows for the n genes and columns for the m biological samples in which gene expression has been monitored. Hence, m_{ij} is the expression level of gene i in sample j. A row $\mathbf{e}_i$ is the *gene expression pattern* of gene i over all the samples. A column $\mathbf{s}_j$ is the expression level of all genes in a sample j and is called the *sample expression pattern*.

Types of Microarrays

cDNA Microarray

Nylon Membrane and Plastic Arrays (by Clontech)

Oligonucleotide Silicon Chips (by Affymetrix)

For What Do We Use Microarray Data?

For What Do We Use Microarray Data?

For What Do We Use Microarray Data?

Statistical Methods Can Help

Statistical Methods Can Help

Statistical Methods Can Help

Preprocessing Microarray Data

Preprocessing Microarray Data

Ratioing the data

Log-tranforming ratioed data

Alternative to ratioing the data

An alternative that eliminated both of the outlier problems above is

$$\frac{r_{ij}}{r_{ij} + g_i}$$

This gives a value in [0,1] and can be interpreted at the probability of gene i is higher in sample j than in control.

Differencing the data

Scaling data across chips to account for chip-to-chip difference

Zero-centering a gene on a sample expression pattern

This in effect the same as subtracting the mean expression pattern.

Suppose that **x** is an expression pattern for a particular gene g_i whose components are log-ratios. Let where is the

Weighting the components of a gene or sample expression pattern differently

If we have a matrix of weights $\mathbf{W}=\mathrm{diag}(w_1,$

Handling missing data

Variation filtering expression patterns

Discretizing expression data

Sometimes we might want to convert gene or sample expression pattern into discrete values. For example, if we have log-ratio, we may want to simply look at whether something is up- or down-regulated. To do this, we can do the following:

In this case +1 would indicate up-regulation, 0 would indicate no change and

Measuring Dissimilarity of Expression Data

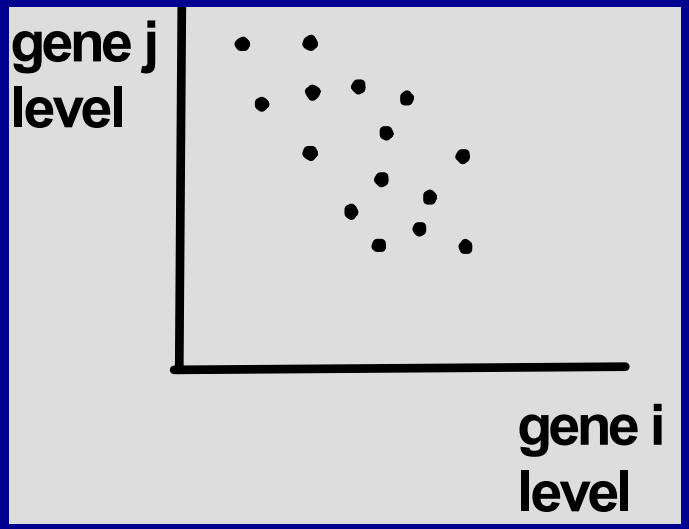
Example Distance Metric

Euclidean Distance

Example Dissimilarity Measures

Visualizing Micorarray Data

It is usually easiest to understand data if it can be represented in 2 or 3 dimensions. For example, a 2-D scatter plot of the expression levels of genes *i* and *j* over a number of samples can show the relationship between these two genes.



Principal Components Analysis

Principal Components Analysis

Example

PCA and Microarrays

Sample Application 1

PCA and Microarrays

Sample Application 2

PCA and Microarrays

PCA allows us to see how good the correlation among these cells is. To use PCA, we would hybridize k different samples on the same chip. For each sample, the expression levels of a gene x in the n cells is an n-dimensional vector. Hence, there are k points in n-dimensional space. Using PCA, if most of the variance is explained by the first principal component, the effective dimensionality of the data is 1 and there cells are highly correlated.

PCA Limitations

Cluster Analysis of Microarray Data

Recall that microarray data can be thought of as gene expression patterns or sample expression patterns. These can be each considered to be vectors. The first thing we have to do before applying cluster analysis is to find a distance between the various expression pattern vectors. This is done using similarity/dissimilarity measures such as Euclidean distance, Mahalonobis distance, or linear correlation coefficients. Once a distance matrix is computed, the following clustering algorithms can be used. The clusters formed can differ significantly depending upon the distance measure used.

Cluster Analysis of Microarray Data

Hierarchical Clustering

Cluster Analysis of Microarray Data

k-Means Clustering

Cluster Analysis of Microarray Data

Self-organizing Maps

Hidden Markov Models and Microarray Data

We can use Hidden Markov models for pattern recognition in the study of micorarray data. Suppose that we want to consider gene expression data from a tissue sample and want to know if it is control or different from the control (diseased, experimentally altered, responding to drug, etc.). Consider the gene expression data vector as a set of emissions, one for each vector coordinate. Each emission has a value that is defined by some probability distribution function. This can be continuous, or can even discrete. To make it discrete, the data should be preprocessed to indicate, up-regulation, down-regulation, or no significant change.

Finding Genes Expressed Unusually Different in a Population

The following section addresses the question: Is gene g expressed unusually in the sample?

The first thing to do is to come up with a formal mathematical definition for what unusual is. Assume that the microarray data is log transformed ratio data. If a histogram is constructed of the data, it should yield roughly a normal distribution. Anything that is out near either tail can be considered to be unusually expressed. Note that this can be either a high or low expression level.

Finding Genes Expressed Unusually Different in a Population

Calculate the Z-score for the data point considered

$$Z = \frac{e_g - \mu}{\sigma}$$

where e_g is the expression level, μ is the mean and σ is the standard deviation. The Z value will give an indication of the how far the data is toward the tail (α - level).

Use statistical inference (hypothesis testing).