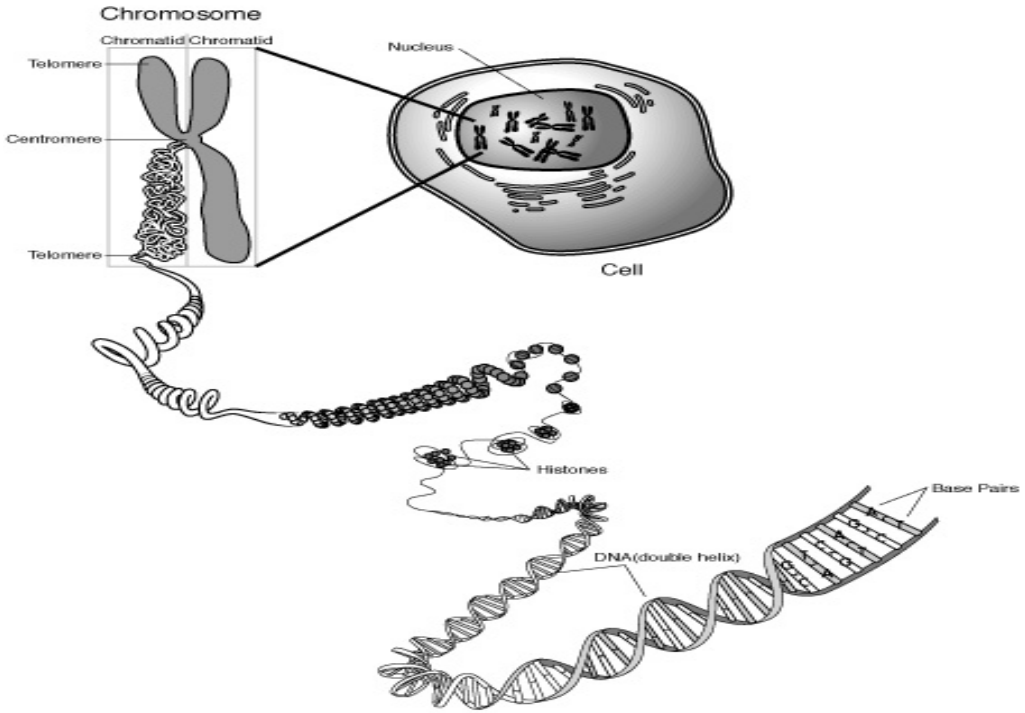
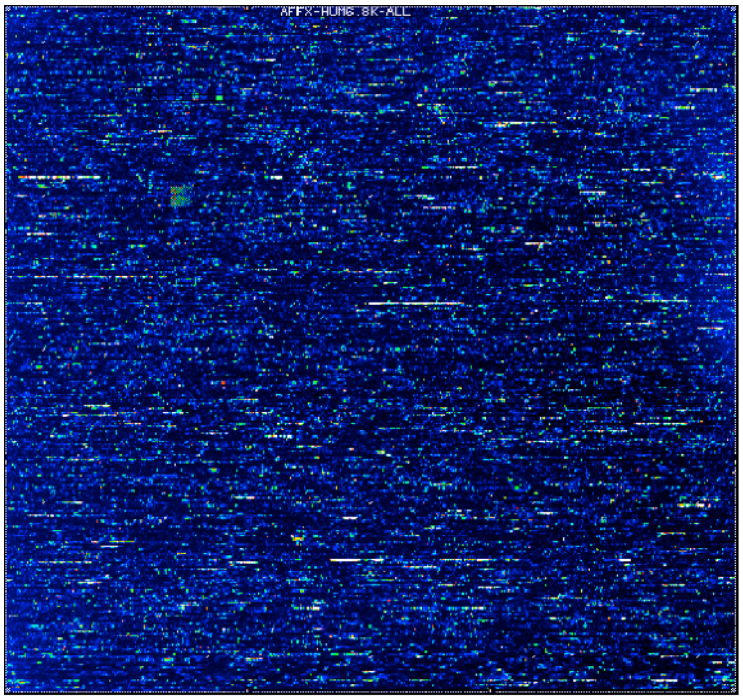
Statistics for Genomic Data Analysis

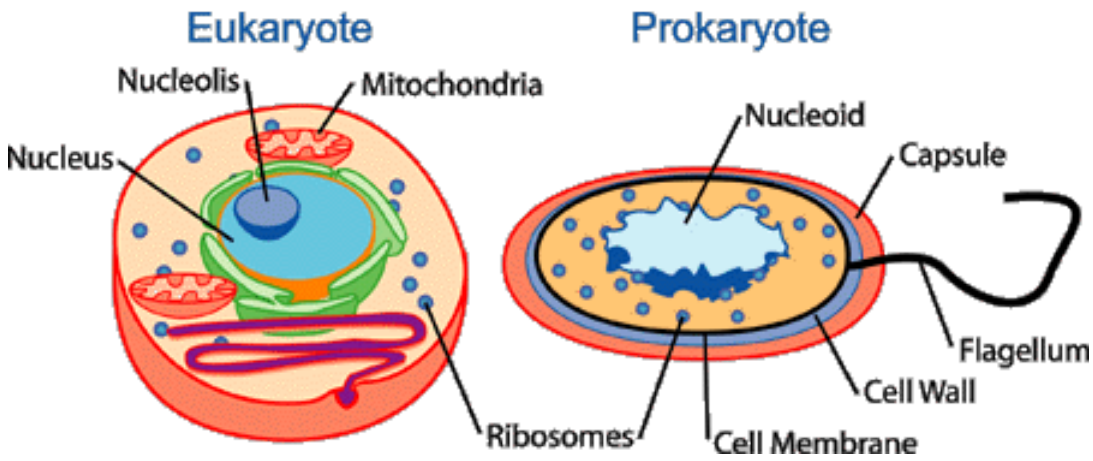
Molecular Biology Background; Affymetrix chips





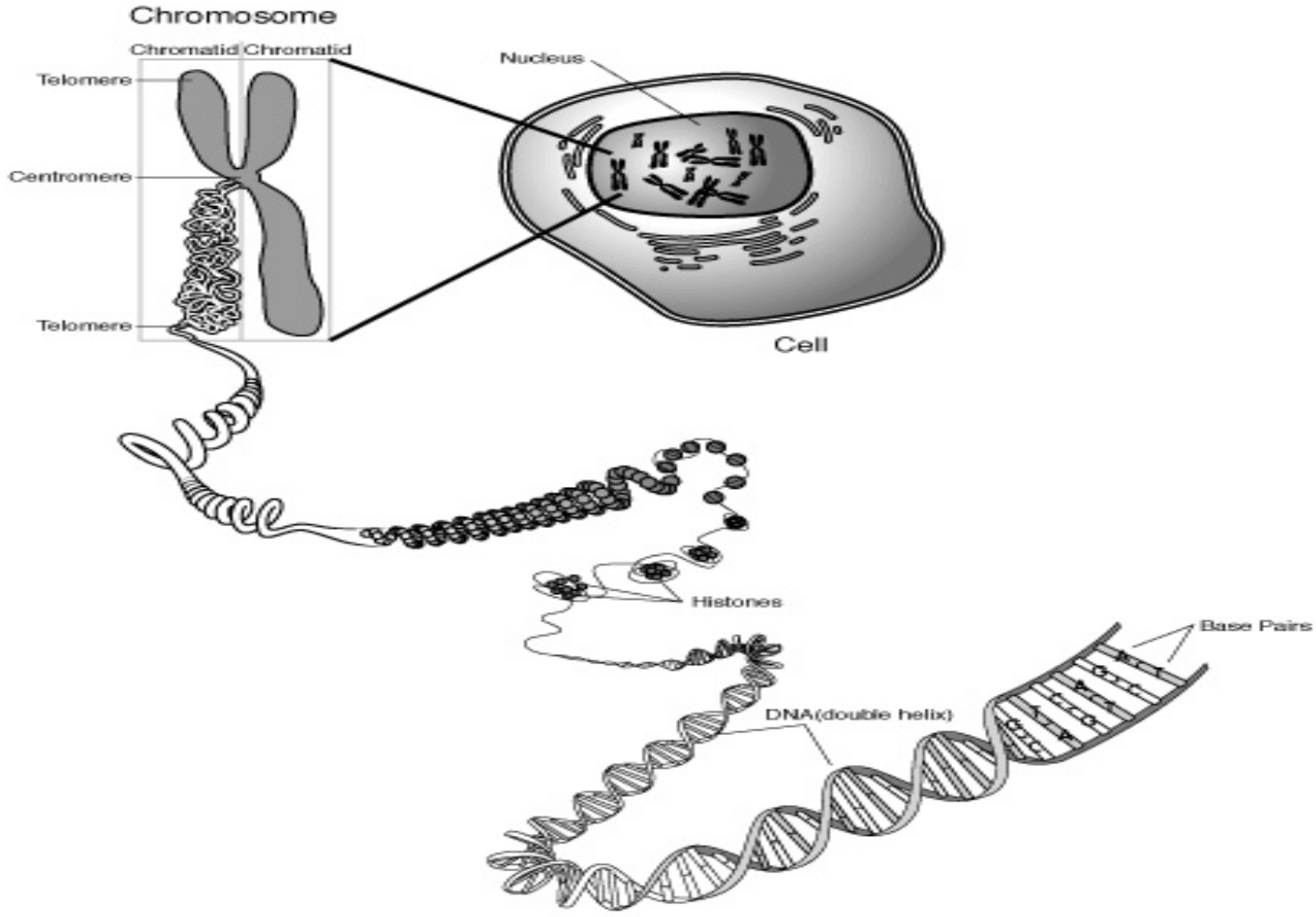
http://moodle.epfl.ch/course/view.php?id=15271

Types of organisms*

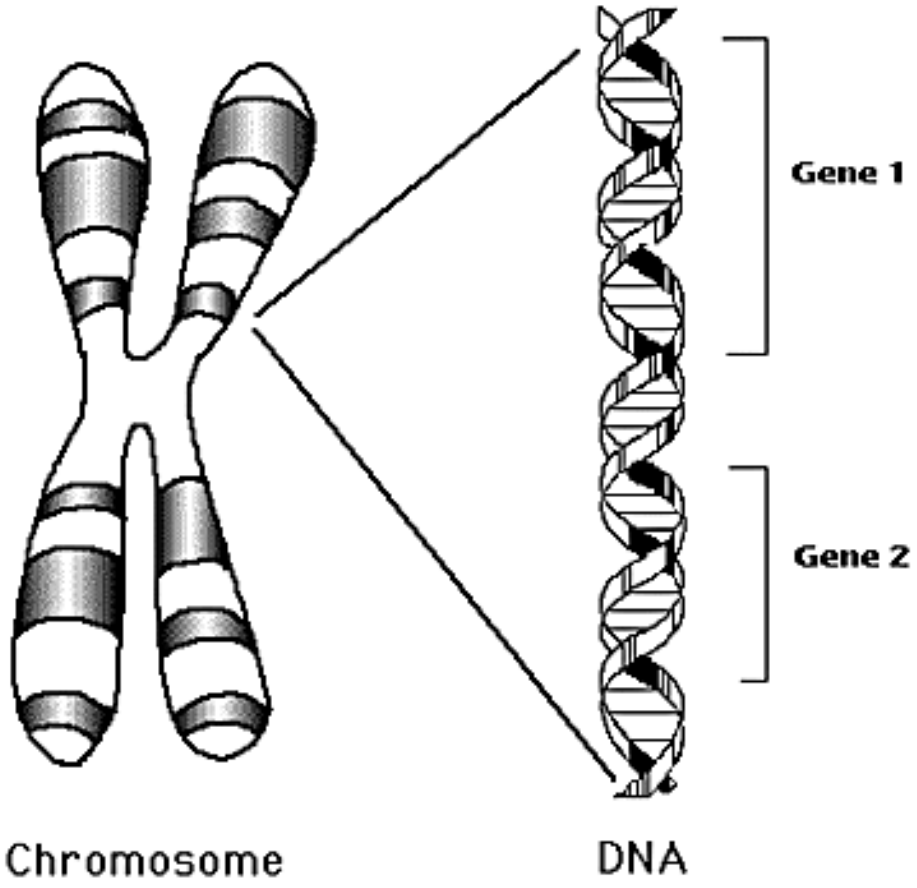


** Every biological 'rule' has exceptions!*

Chromosomes and DNA



Genes are linearly arranged along chromosomes



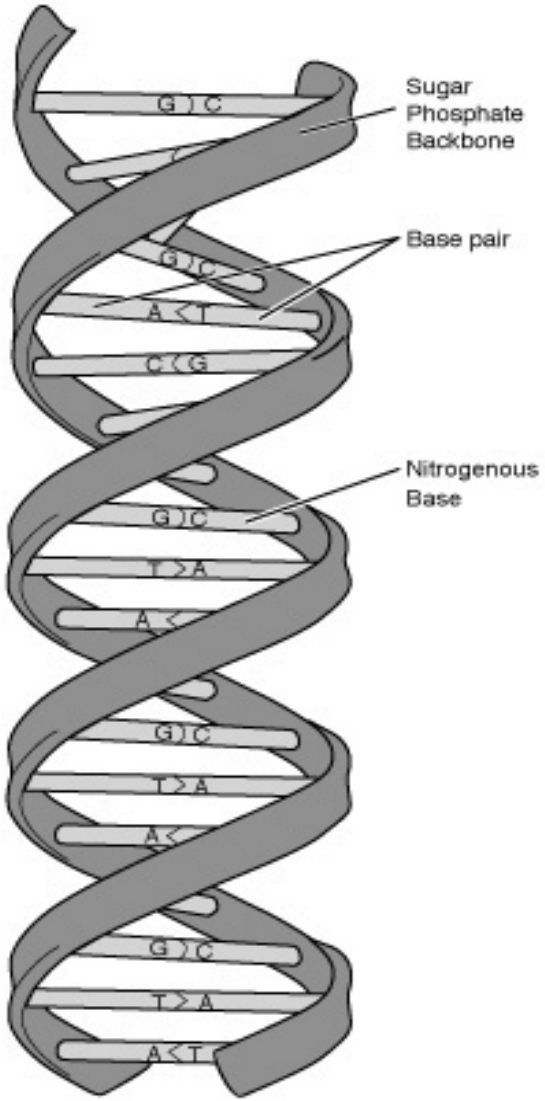
DNA Structure Discovery

- "We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest." *Nature (1953), 171:737*

DNA

- *Double-stranded* linear polymer
- Composed of four molecular subunits (*nucleotides*)
- Each nucleotide comprises a *phosphate group*, a *deoxyribose sugar*, and one of four nitrogen *bases*: adenine (A), guanine (G), cytosine (C), or thymine (T)
- Strands held together by weak hydrogen bonds between *complementary bases*
- Base-pairing: G pairs with C; A pairs with T

DNA Structure (overview)



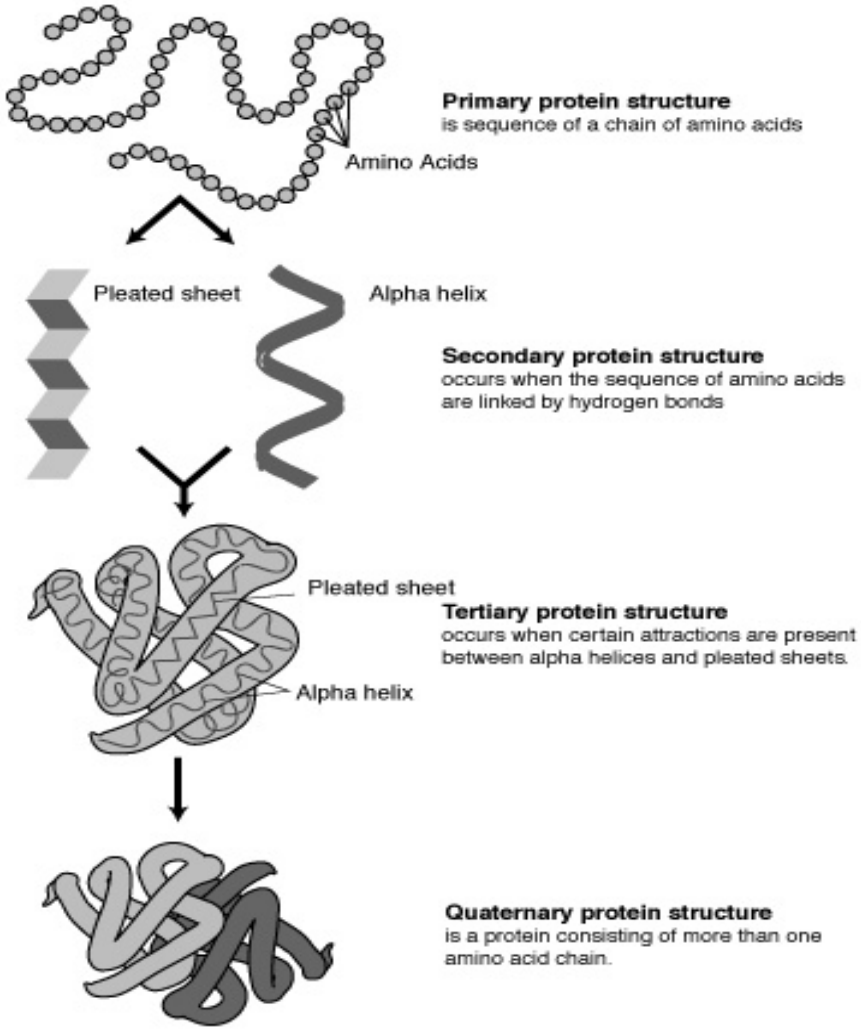
Proteins

- *Proteins*: Macromolecules composed of one or more chains of *amino acids*
- *Amino acids*: class of 20 different organic compounds containing a basic amino group (-NH2) and an acidic carboxyl group (-COOH)
- *Amino acid order* is determined by the *base sequence of nucleotides* in the gene coding for the protein
- Proteins function as enzymes, antibodies, structures, *etc.*

Amino Acid Codes

Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine
Asx	B	Asn or Asp
Glx	Z	Gln or Glu
Sec	U	Selenocysteine
Unk	X	Unknown

Multiple Levels of Protein Strucure

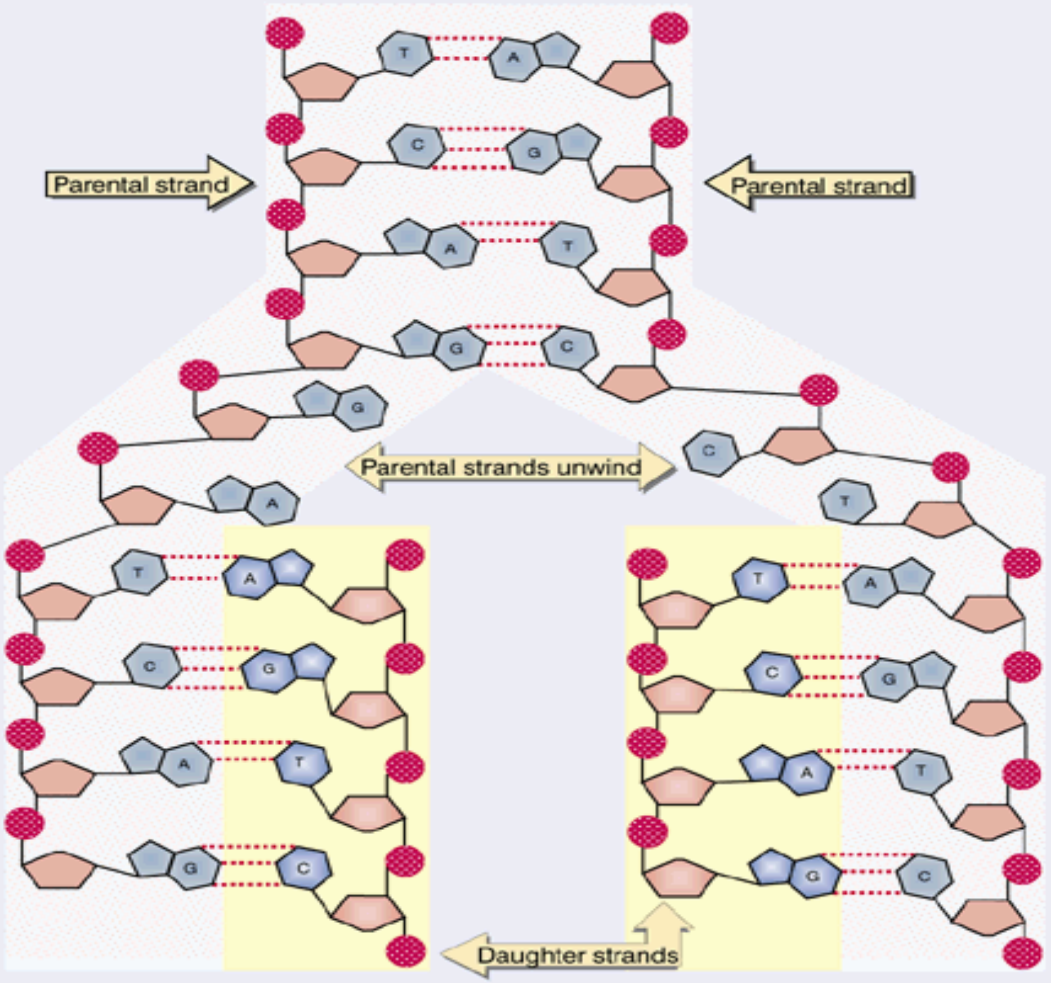


(← Protein folding)

DNA Replication

- "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." *Nature (1953), 171:737*

Figure 1.9 Base pairing provides the mechanism for replicating DNA.



DNA replication overview

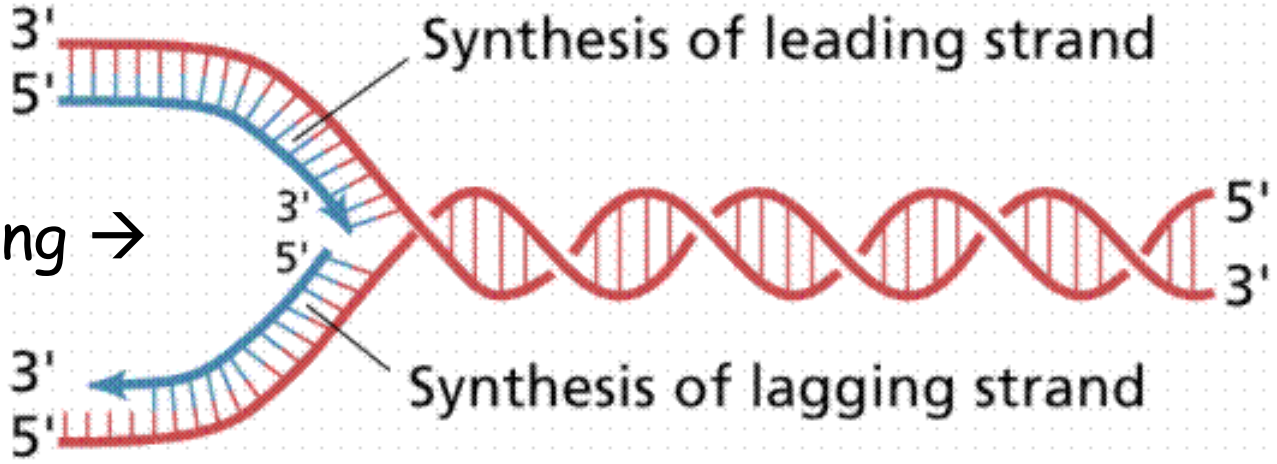
- (Also called DNA *synthesis*)
- The DNA strand that is copied to form a new strand is called a *template*
- *Both* original DNA strands are copied
- Two new duplexes each consist of one of the original strands plus its copy (semiconservative replication)

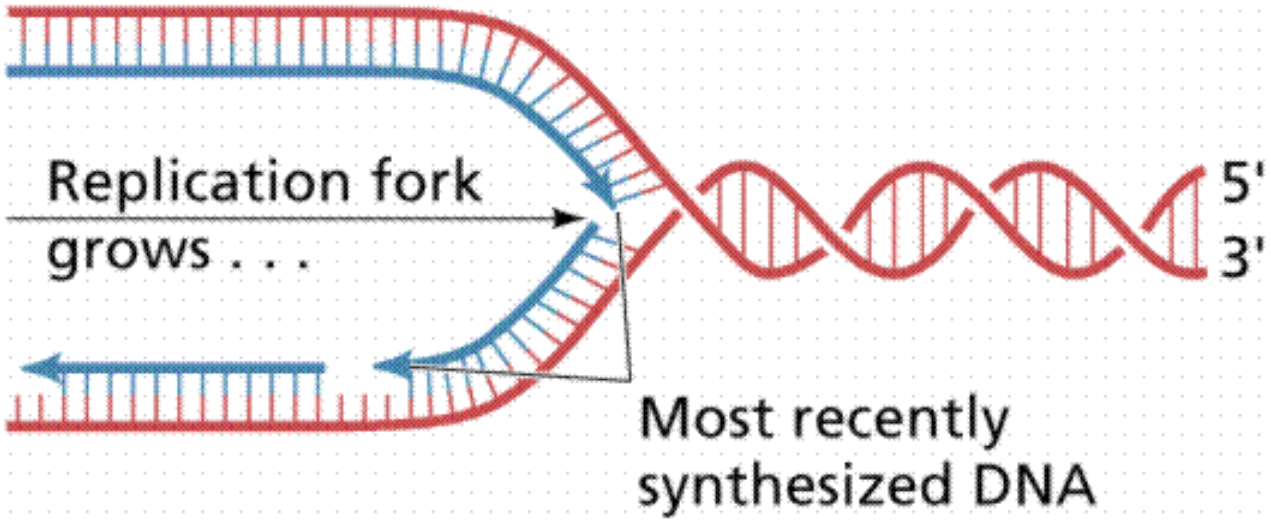
DNA Replication (in more detail)

- DNA synthesis occurs in the *chemical direction 5'→3'*
- Nucleic acid chains are assembled from 5' triphosphates of deoxyribonucleosides (triphosphates supply energy)
- DNA polymerases are enzymes that copy (replicate) DNA
- DNA polymerases require a short preexisting DNA strand (primer) to begin chain growth; with a primer base-paired to the template strand, a DNA polymerase adds nucleotides to the free hydroxyl group at the 3' end of the primer
- DNA replication requires assembly of many (> 30) proteins at a growing replication fork: helicases to unwind, primases to prime, ligases to ligate (join), topisomerases to remove supercoils, RNA polymerase, *etc.*

DNA Synthesis

DNA is unwinding →





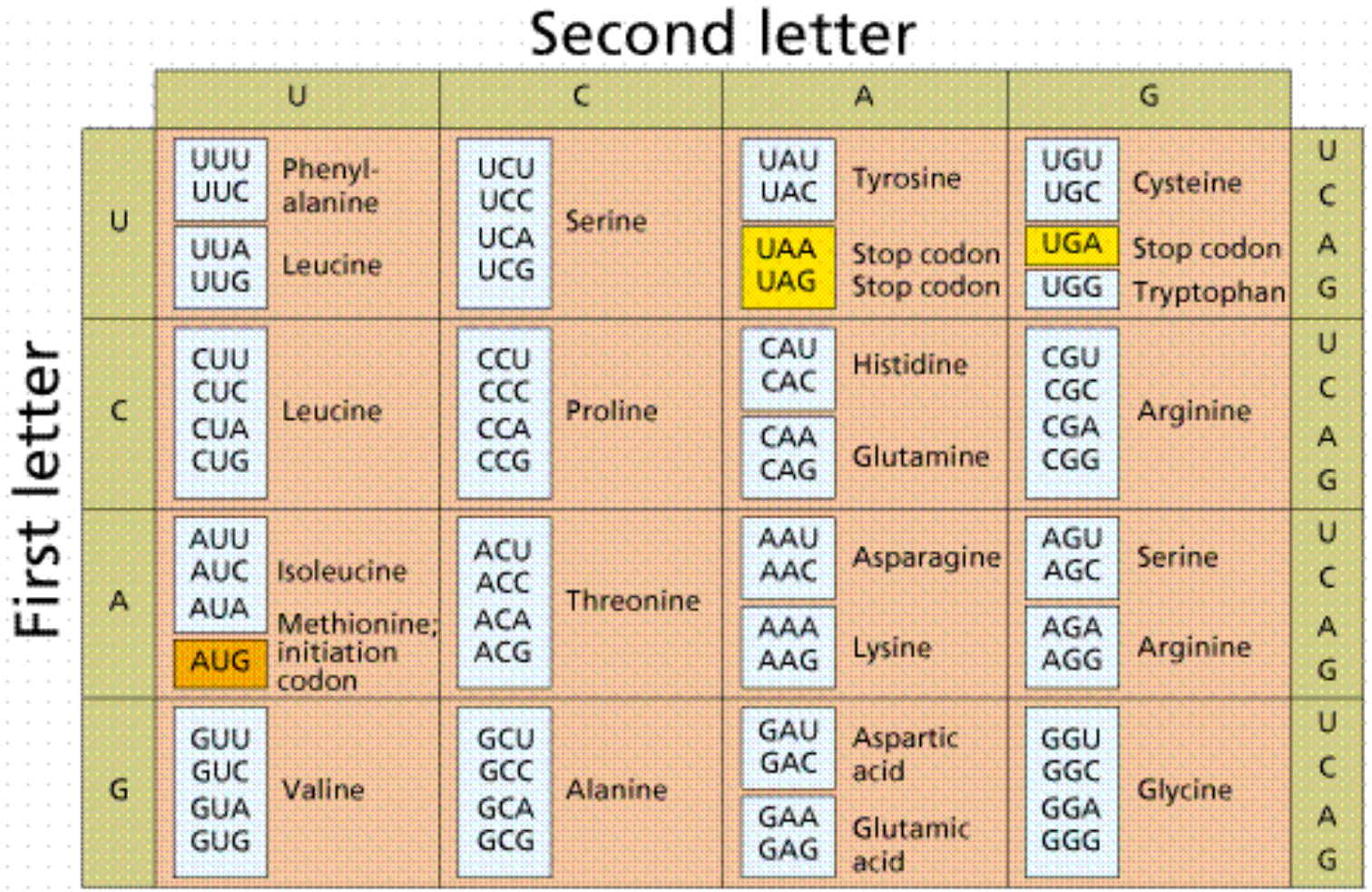
RNA

- RNA (*ribonucleic acid*) is similar to DNA, but
  - RNA is *single-stranded*
  - the sugar is ribose (not deoxyribose)
  - *uracil* (U) is used instead of thymine
- Important for *protein synthesis*, other cell activities
- There are *several classes* of RNA molecules, including *messenger RNA (mRNA)*, transfer RNA (tRNA), ribosomal RNA (rRNA), and other small RNAs

The Genetic Code

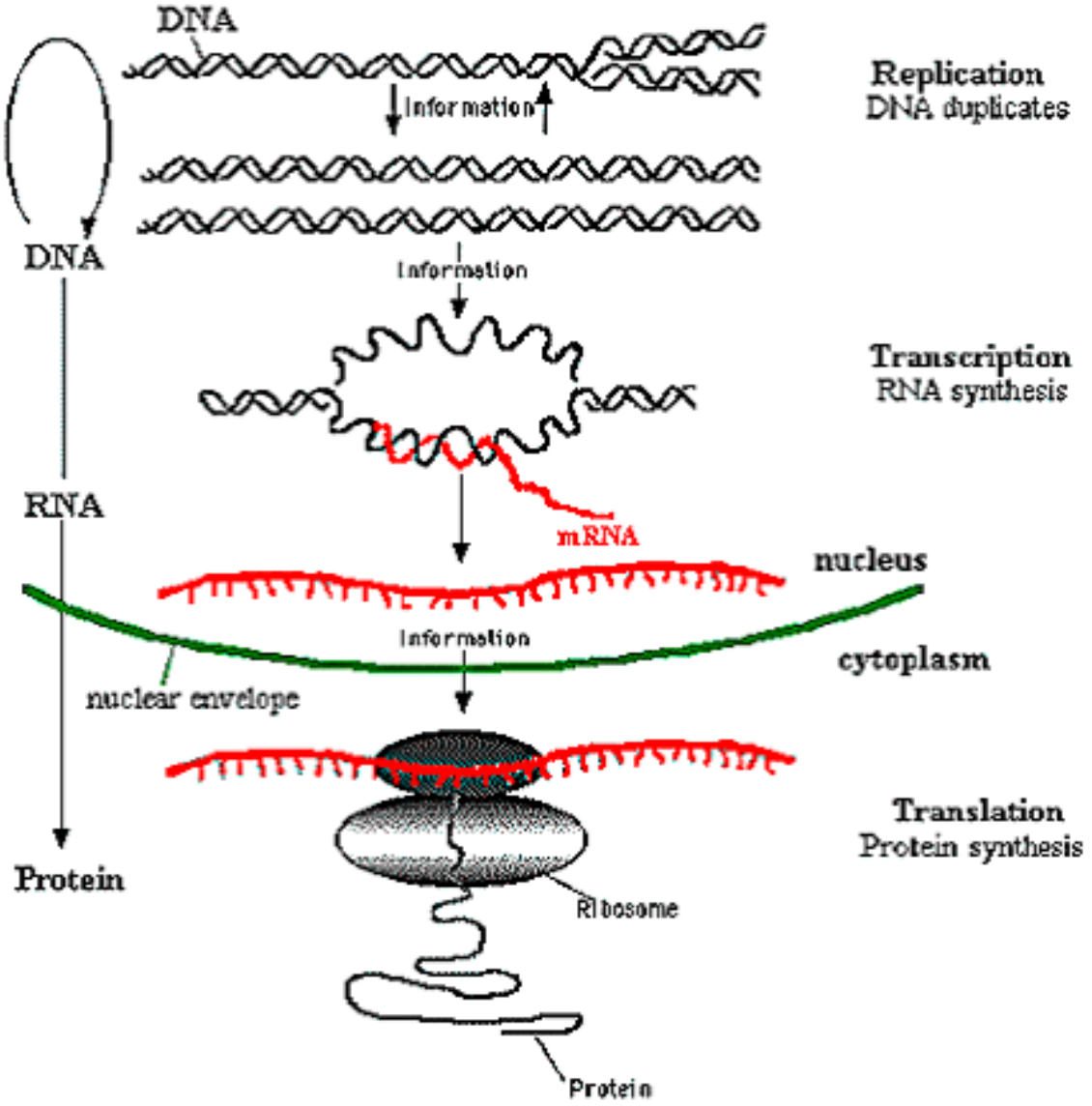
- DNA: sequence of four different nucleotides
- Protein: sequence of twenty different amino acids
- The correspondence between the four-letter DNA alphabet and the twenty-letter protein alphabet is specified by the *genetic code*, which relates *nucleotide triplets*, or *codons*, to amino acids

Standard Genetic Code

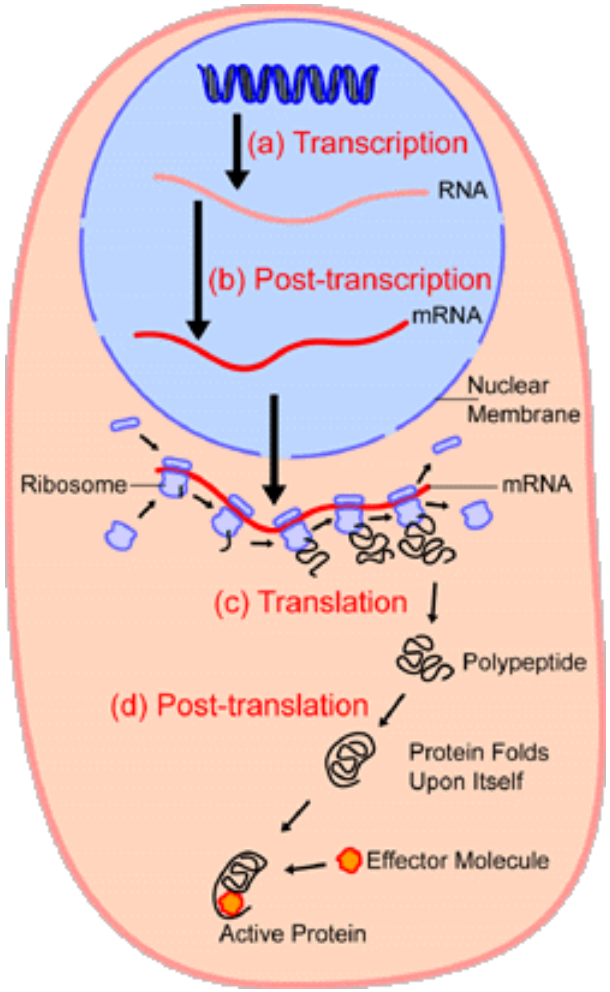


Central Dogma of Molecular Biology

(RT)



Protein Synthesis in the Cell



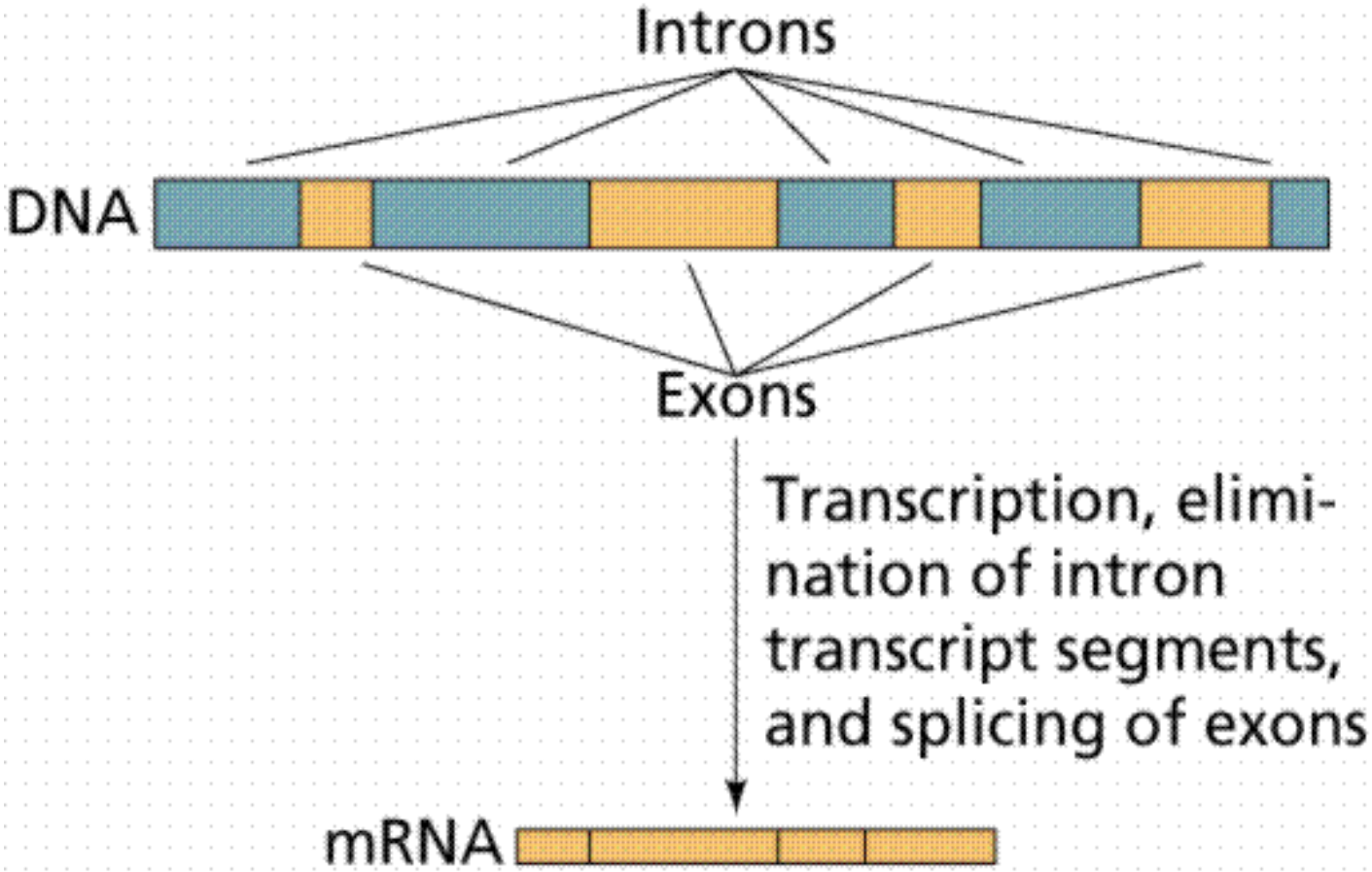
Transcription

- A complex process involving several steps and many proteins (enzymes)
- RNA polymerase synthesizes a single strand of RNA against the DNA template strand (*anti-sense strand*), adding nucleotides to the 3' end of the RNA chain
- *Initiation* is regulated by *transcription factors*, including promoters, usually an initiator element and TATA box, usually lying just upstream (at the 5' end) of the coding region
- 3' end cleaved at AAUAAA, poly-A tail added

Exons and Introns

- Most of the genome is *non-coding regions*
- Some non-coding regions (centromeres and telomeres) may have specific chomosomal functions
- Others have *regulatory* purposes
- Non-coding, non-functional DNA used to be called 'junk DNA', but has recently been found to have regulatory (and other) functions (ENCODE PROJECT)
- The terms *exon* and *intron* refer to *(protein) coding* and *non-coding* DNA, respectively

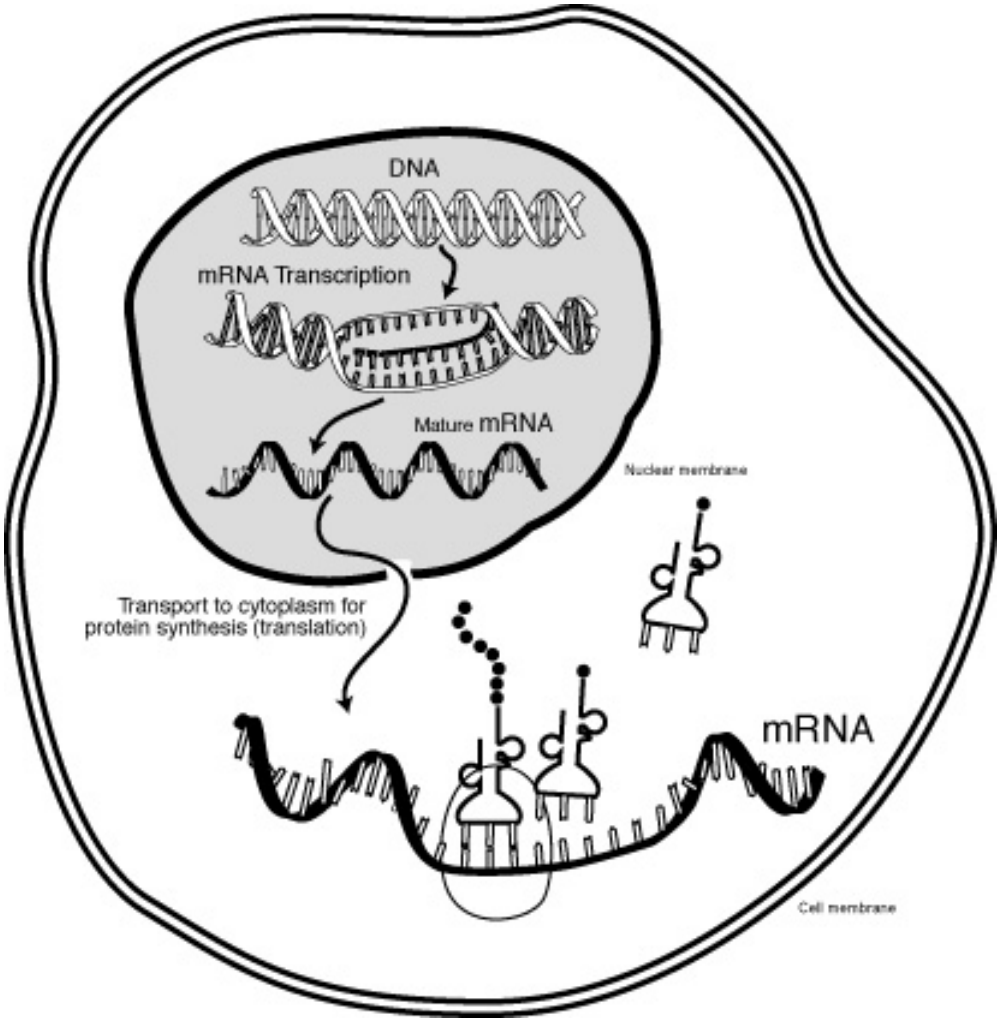
Intron Splicing



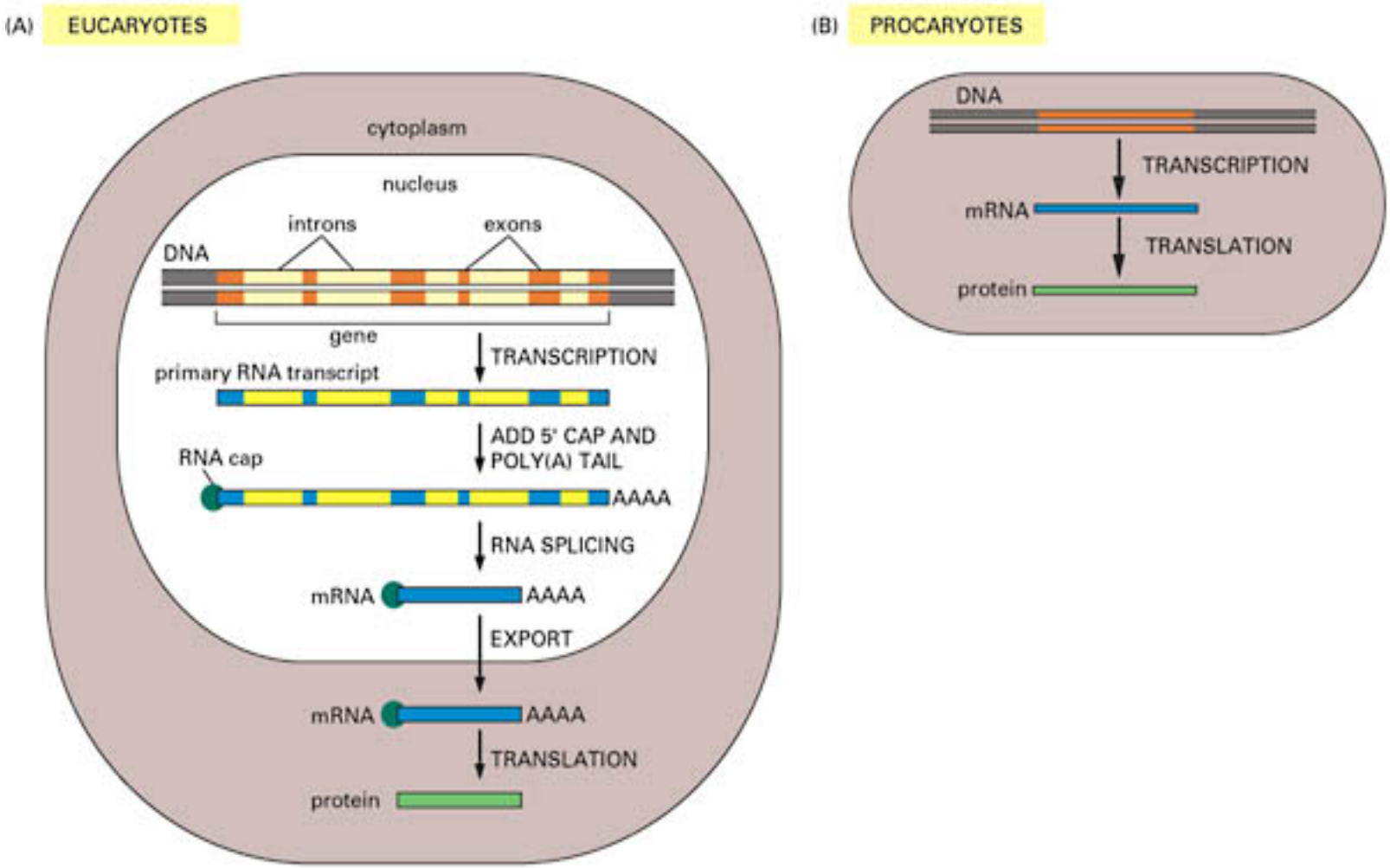
Translation

- The *AUG start codon* is recognized by methionyl-$tRNA_i^{Met}$
- Once the start codon has been identified, the ribosome incorporates amino acids into a polypeptide chain
- RNA is decoded by tRNA (transfer RNA) molecules, which each transport specific amino acids to the growing chain
- Translation ends when a *stop codon* (UAA, UAG, UGA) is reached

Translation Illustrated



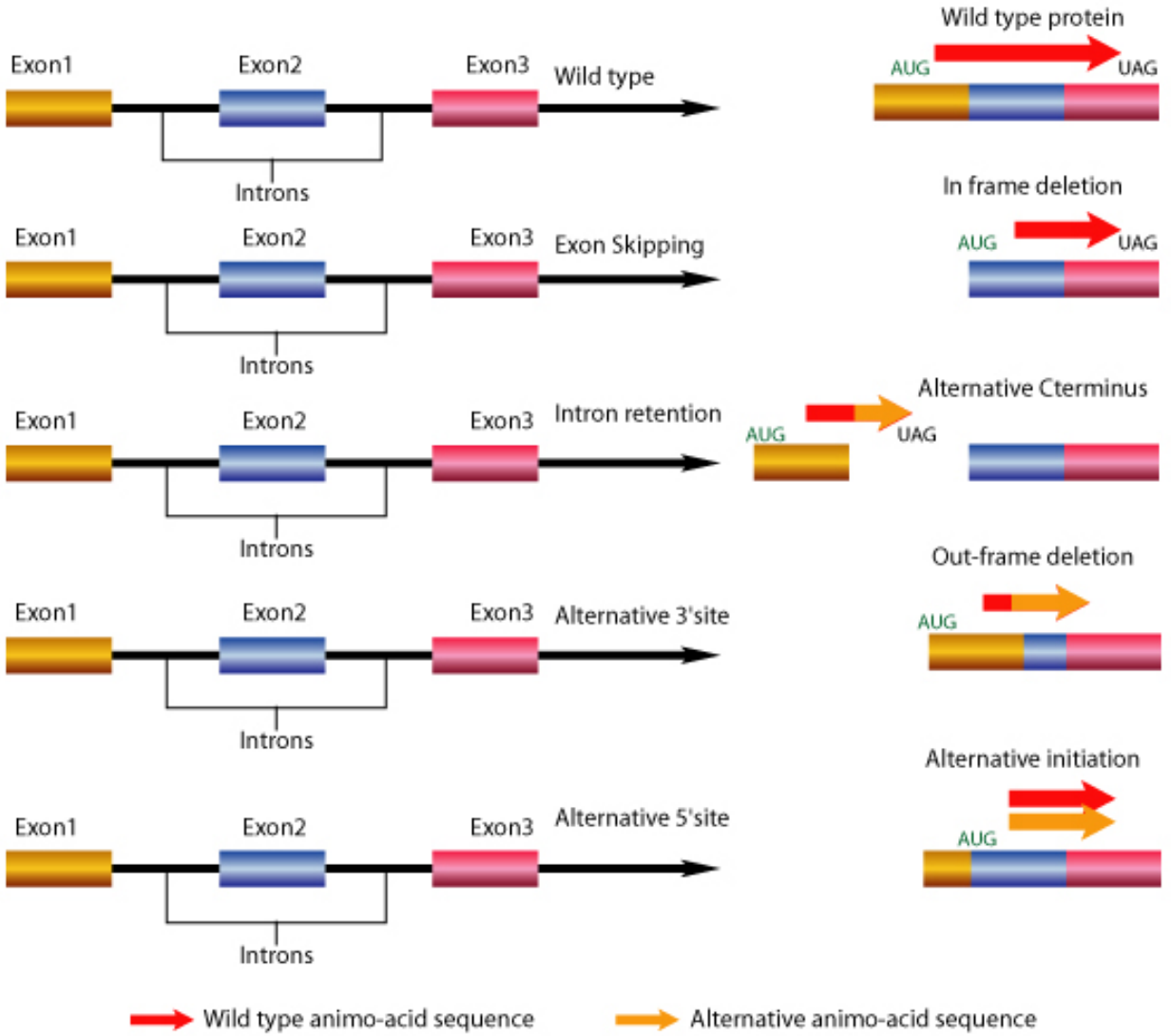
From Primary Transcript to Protein



Alternative Splicing (of Exons)

- How is it possible that there are over 1,000,000 human antibodies when there are only about 20,000 - 30,000 genes?
- *Alternative splicing* refers to the different ways the exons of a gene may be combined, producing different forms of proteins *within the same gene-coding region*
- Alternative pre-mRNA splicing is an important mechanism for regulating gene expression in higher eukaryotes

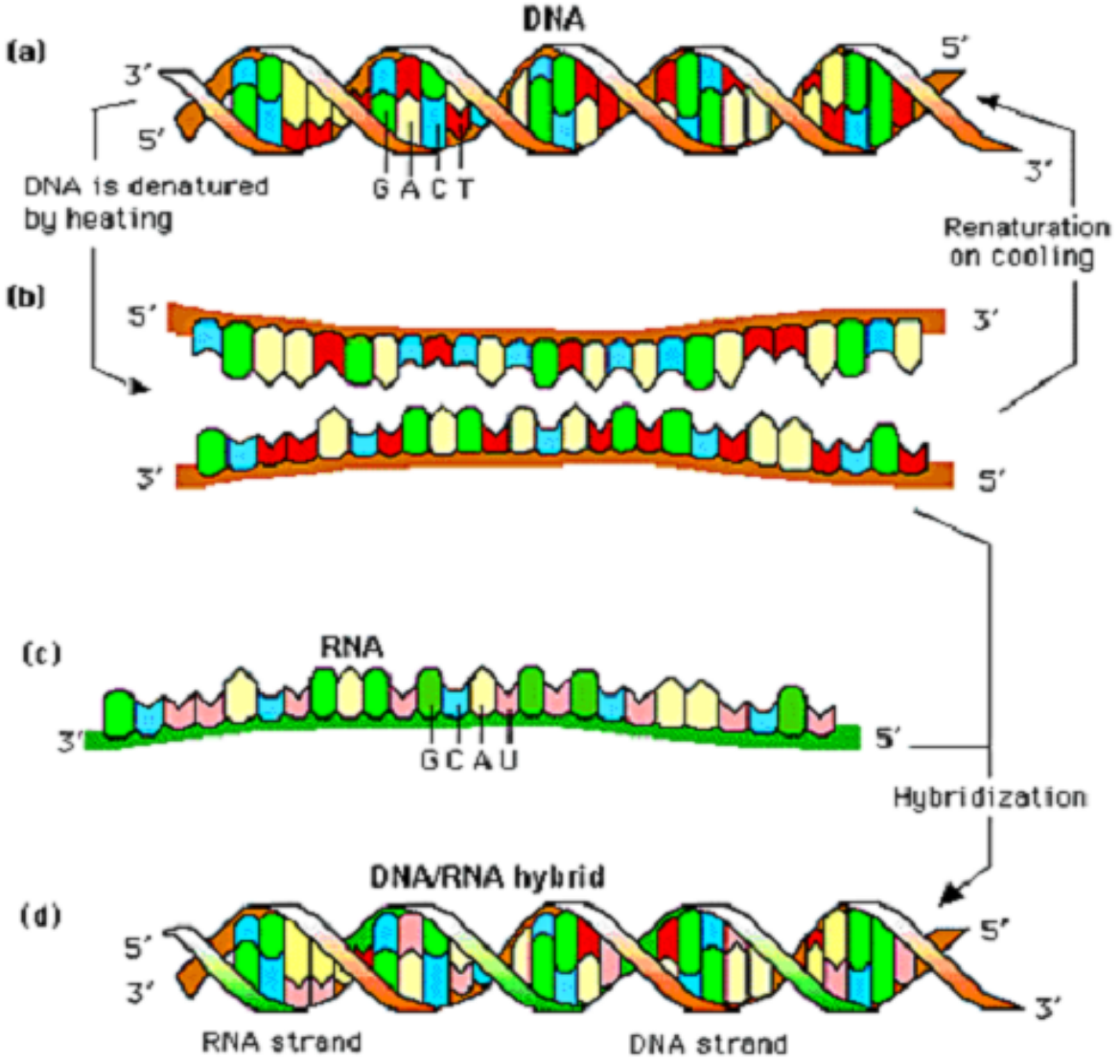
Alternative Splicing



Hybridization

- *Hybridization* exploits a potent feature of the DNA duplex – the sequence complementarity of the two strands
- Remarkably, DNA can reassemble with (nearly) perfect fidelity from separated strands
- Strands can be separated (denatured) by heating

Nucleic Acid Hybridization



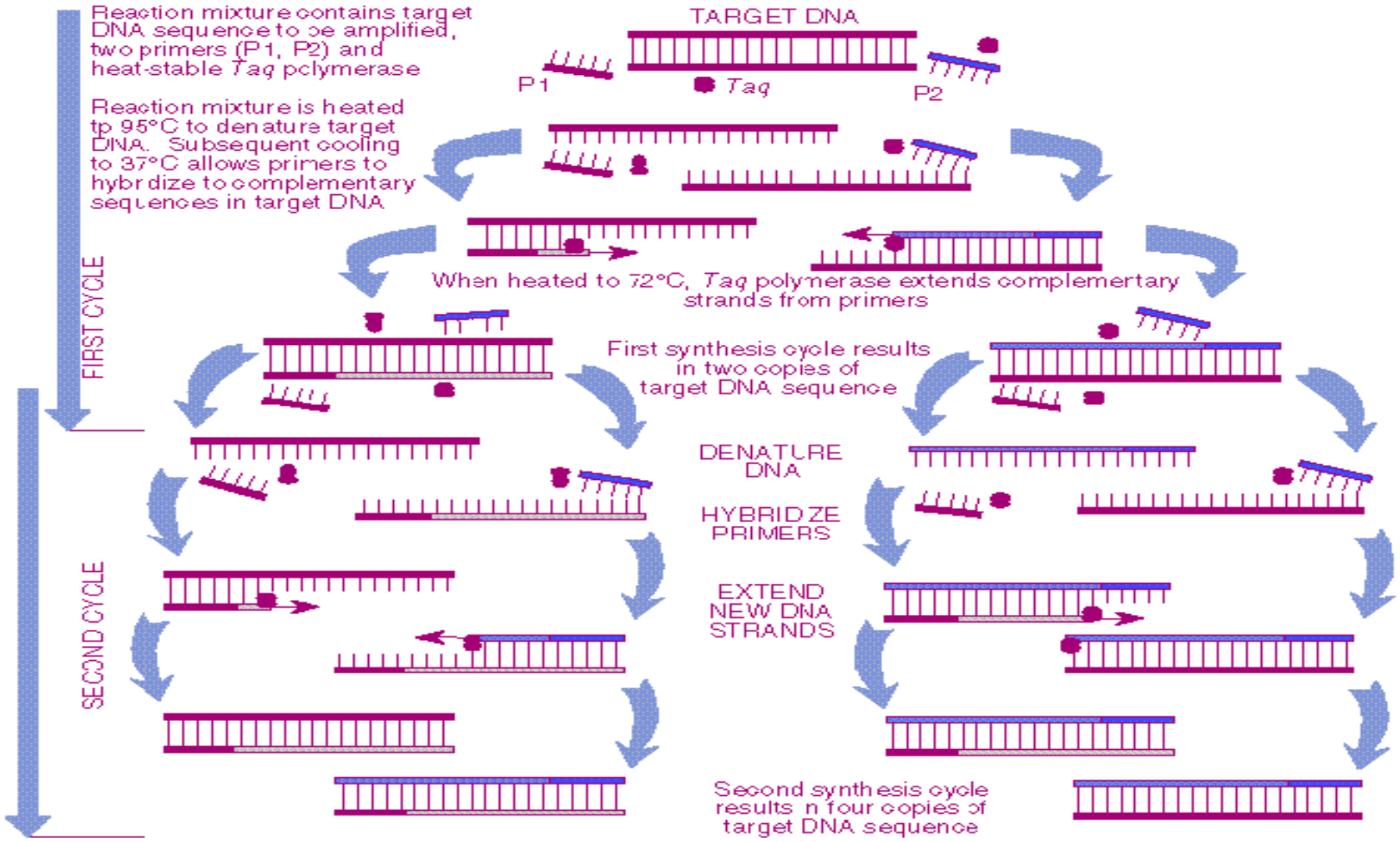
Polymerase Chain Reaction (PCR)

- *PCR* is used to *amplify* (copy) specific DNA sequences in a complex mixture when the ends of the sequence are known
- Source DNA is denatured into single strands
- Two synthetic oligonucleotides complementary to the 3' ends of the segment of interest are added in great excess to the denatured DNA, then the temperature is lowered
- The genomic DNA remains denatured, because the complementary strands are at too low a concentration to encounter each other during the period of incubation, but the specific oligos hybridize with their complementary sequences in the genomic DNA

PCR, ctd

- The hybridized oligos then serve as *primers* for DNA synthesis, which begins upon addition of a supply of nucleotides and a temperature resistant polymerase such as *Taq* polymerase, from Thermus aquaticus (a bacterium that lives in hot springs)
- Taq polymerase extends the primers at temperatures up to 72°C
- When synthesis is complete, the whole mixture is heated further (to 95°C) to melt the newly formed duplexes
- *Repeated cycles* (25—30) of synthesis (cooling) and melting (heating) quickly provide many DNA copies

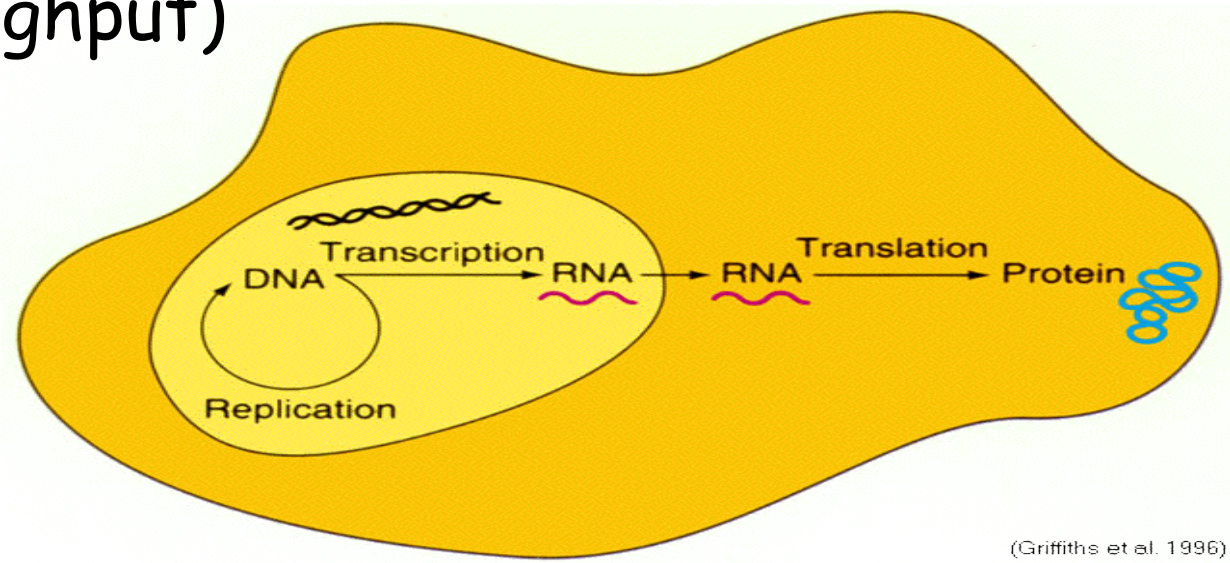
DNA Amplification Using Polymerase Chain Reaction



Source: *DNA Science*, see Fig. 13.

Measuring Gene Expression

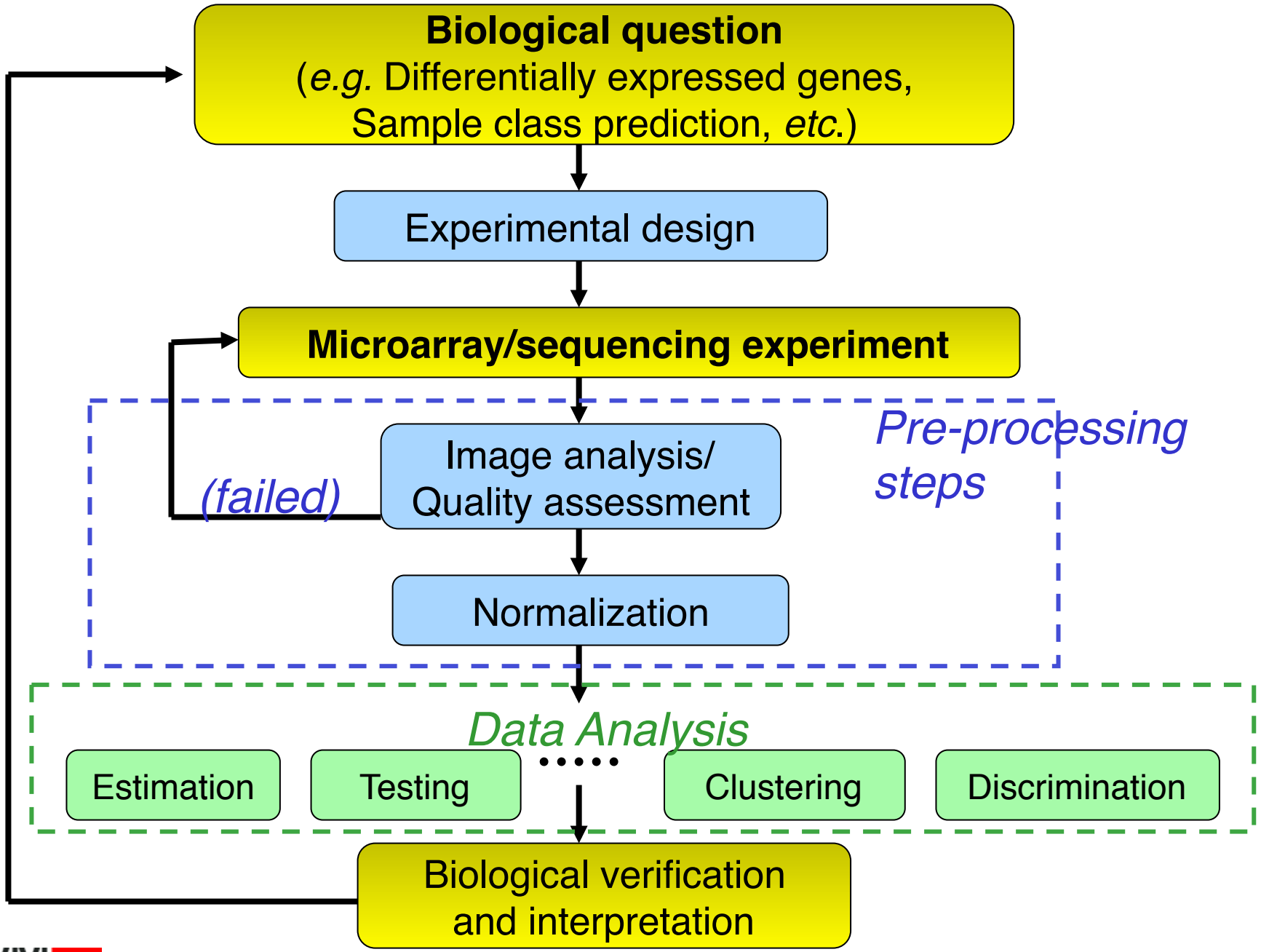
- *Idea* : measure the amount of mRNA to see which genes are being *expressed* in (used by) the cell
- Measuring *protein* would be more direct, but is currently harder (and not as high-throughput)



(Griffiths et al. 1996)

Areas Being Studied

- *Differential gene expression* between two (or more) sample types
- *Similar gene expression* across treatments
- Tumor *sub-class identification* using gene expression profiles
- Tumor *classification* (into known classes)
- *Identification of "marker" genes* that characterize different tumor classes
- Identification of *genes associated with clinical outcomes* (e.g. survival)

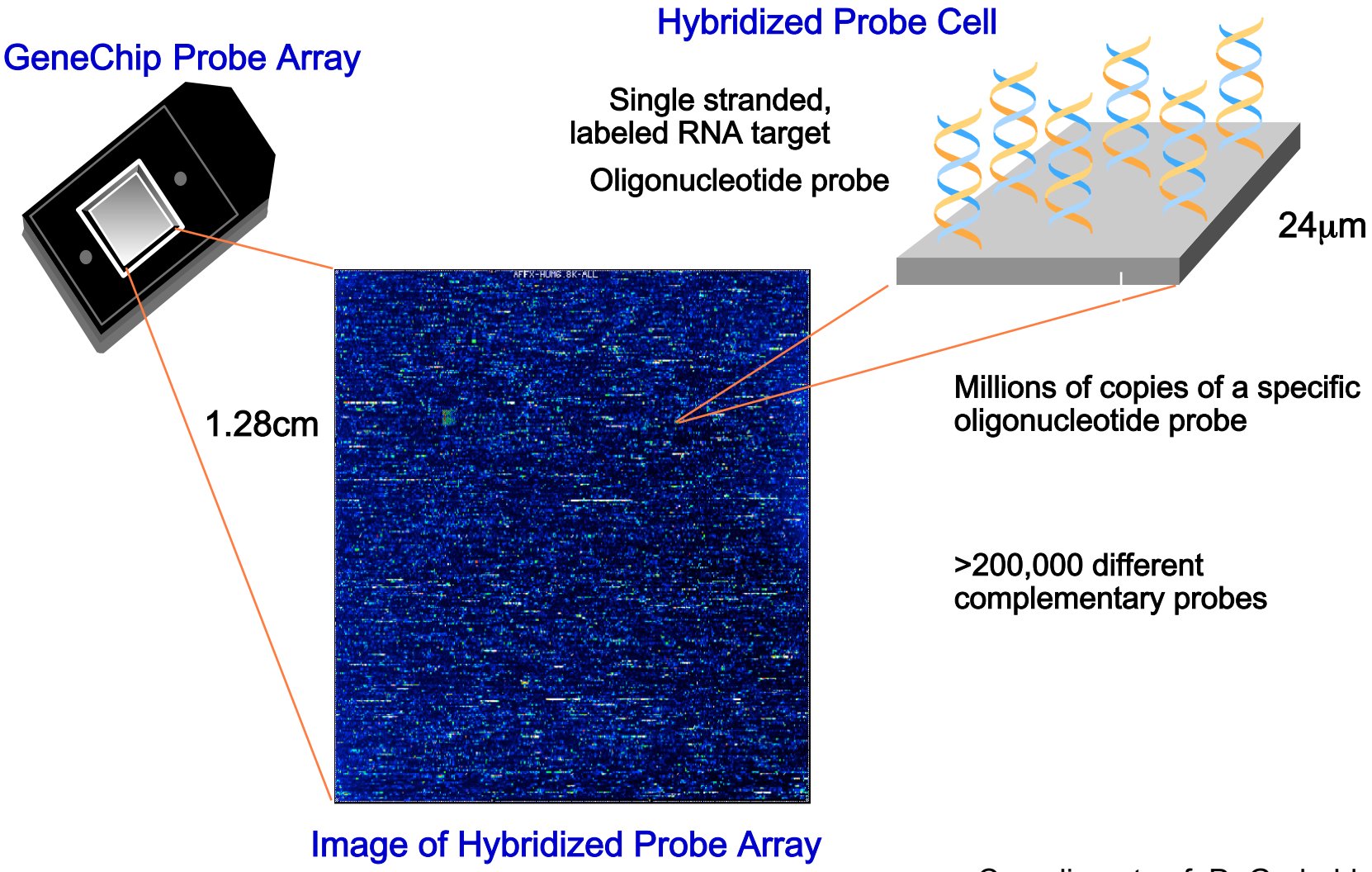

Biological question
(e.g. Differentially expressed genes,
Sample class prediction, etc.)
Experimental design
Microarray/sequencing experiment
Pre-processing steps
(failed)
Image analysis/
Quality assessment
Normalization
Data Analysis
Estimation
Testing
Clustering
Discrimination
Biological verification
and interpretation

Major Technologies

- cDNA probes (> 200 nt), usually produced by PCR, attached to either nylon or glass supports
- (Long) oligonucleotides (25-80 nt) attached to glass support
- *(Short) oligonucleotides (25-30 nt) synthesized in situ on silica wafers (Affymetrix)*
- Probes attached to tagged beads
- *Sequencing technologies*

Affymetrix GeneChip

Affymetrix GeneChip Probe Arrays



Compliments of D. Gerhold

Array design

probe pairs

PM

MM

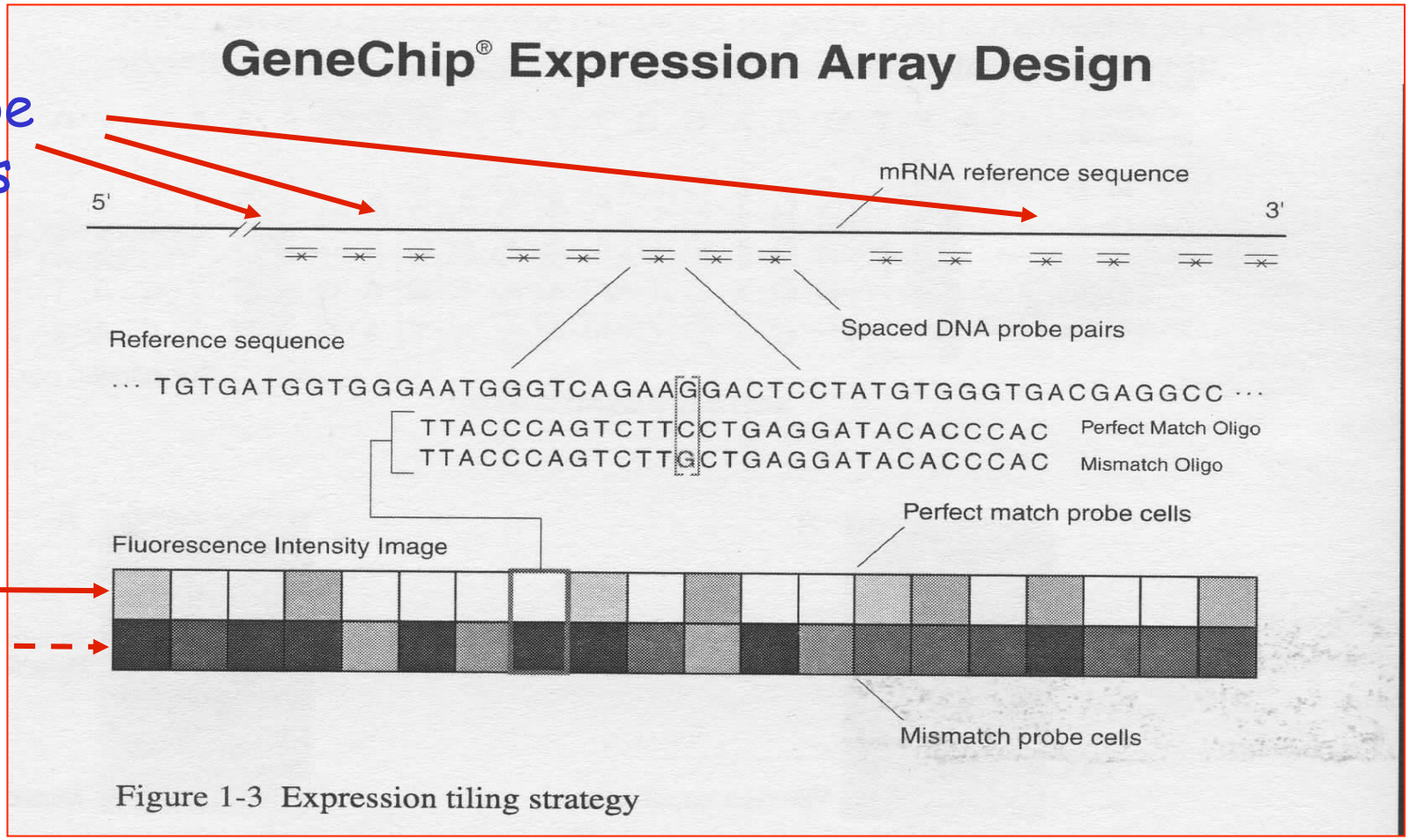


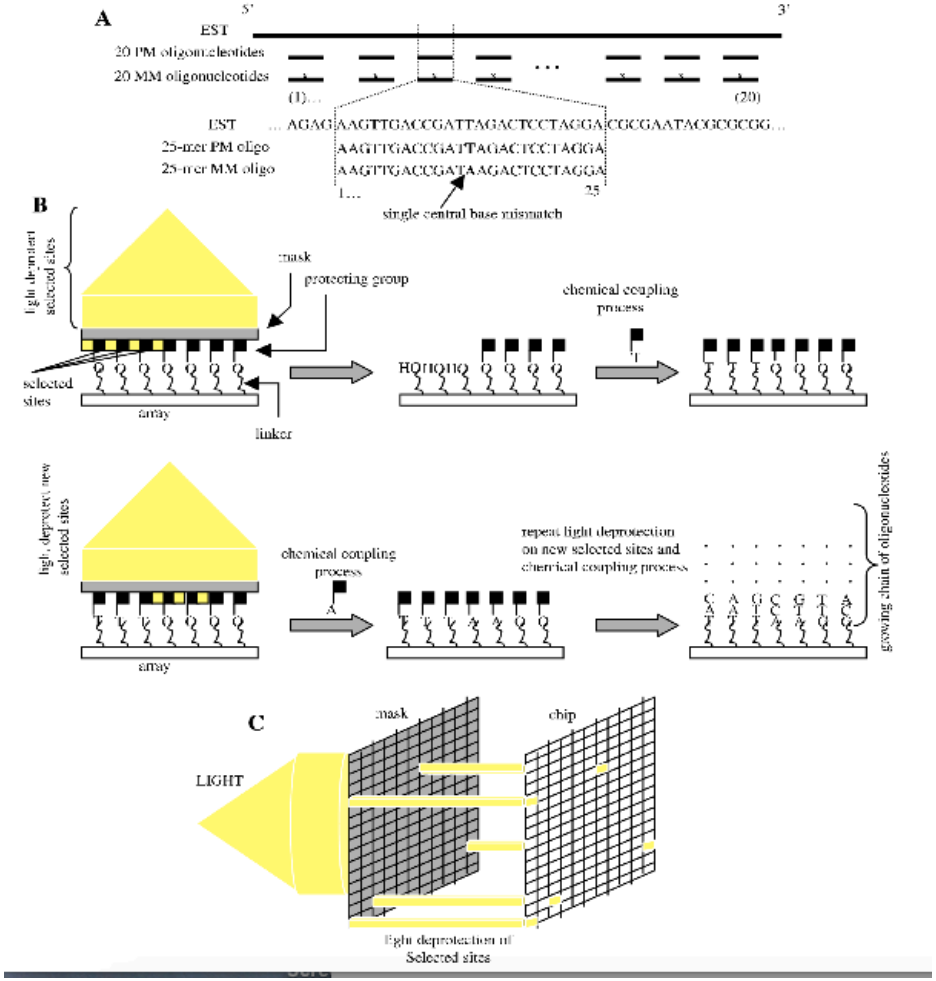
Figure 1-3 Expression tiling strategy

probe set = collection of probe pairs;
There are tens of thousands of probe sets per chip

Array manufacture

- **A.** A gene sequence is represented by (~20) subsequences of the gene, each of length 25 base pairs (oligonucleotides) => *PM probes*
- Another 20 subsequences with the same bases as the PMs, *except for one mismatch* (MM) at the central base (arrow), is used.
- **B.** The light-directed process of synthesizing the oligonucleotides on the chip (array)
- **C.** The schematics of the light, mask, and array in the oligonucleotide synthesis process (photolithography)

Array manufacture – graphically (LEGOS)



Experimental steps (I)

- Total RNA isolated from cells and processed
  - introns removed, exons spliced, poly-A tail
- RNA turned into double stranded DNA copy (cDNA) by reverse transcription
  - RNA not very stable - cDNA is a way to store the RNA for a longer period of time
- When it is time to run the array, the cDNA goes through *in vitro* transcription back to RNA (now known as cRNA)

Experimental steps (II) - biotin

- *Biotin* is also known as vitamin H or vitamin B7
- In Affy chip experiments, there are no fluorescent labels during the hybridization step
  - This procedure differs from the 2-channel glass slide microarrays, where the samples are tagged with fluorescent dyes before hybridization
- Instead, the protein *streptavidin*, which binds to biotin, is tagged with fluorescent dye and added afterwards

Experimental steps (III)

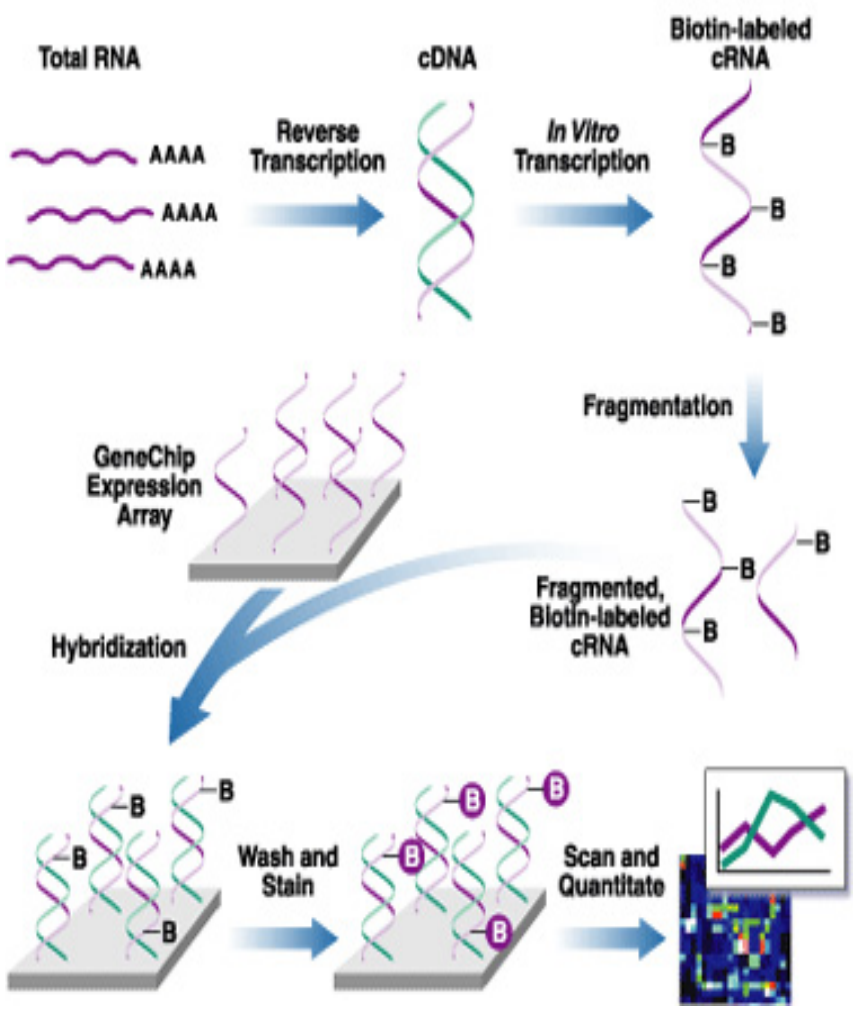
- RNA is labelled with *Biotin* by tagging U's
- This labelled cRNA is then randomly *fragmented* (30 to 400 bp lengths)
- The fragmented, biotinylated cRNA is added to the array
- Anywhere on the array where an RNA fragment and a probe are *complementary*, the RNA *sticks to the probes* in the feature (hybridizes)
- (*millions* of identical probes in each feature)

Experimental steps (III)

- Array is then *washed* to remove any RNA that is not stuck to it (i.e., no match was made) and then *stained* with the fluorescent molecule that sticks to Biotin
- The entire array is then *scanned* with a laser
- The images are processed and the information is stored in files
- Quantitative analysis of what genes were expressed and at what (approximate) level – measure *expression*

Steps of the expression array

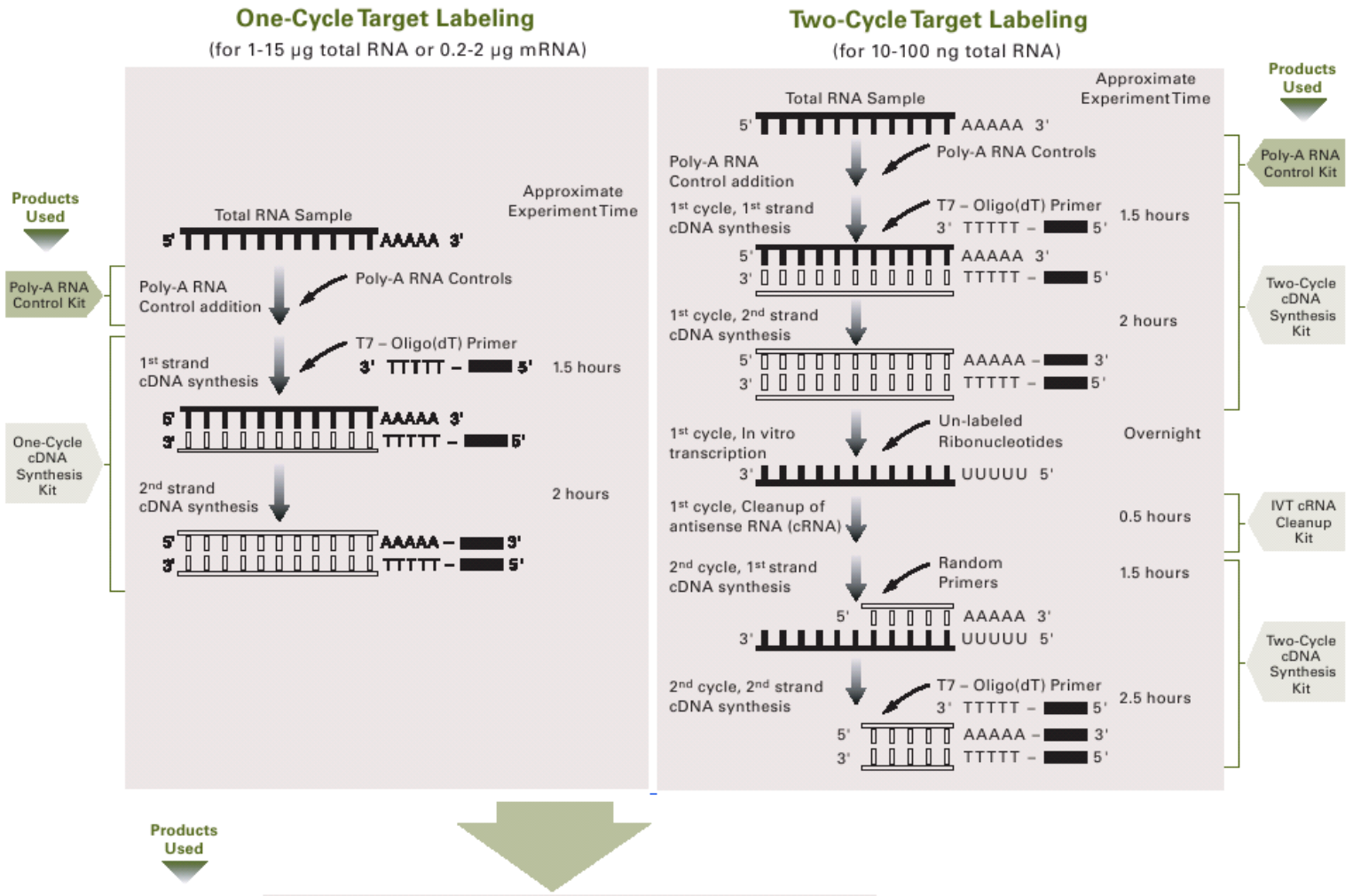
- Isolate total RNA
- Sample amplification and labeling
- Sample injected into microarray
- Probe array hybridization, washing
- Probe array scanning and intensity quantification
- Intensity translated into nucleic acid abundance (*expression measure*)



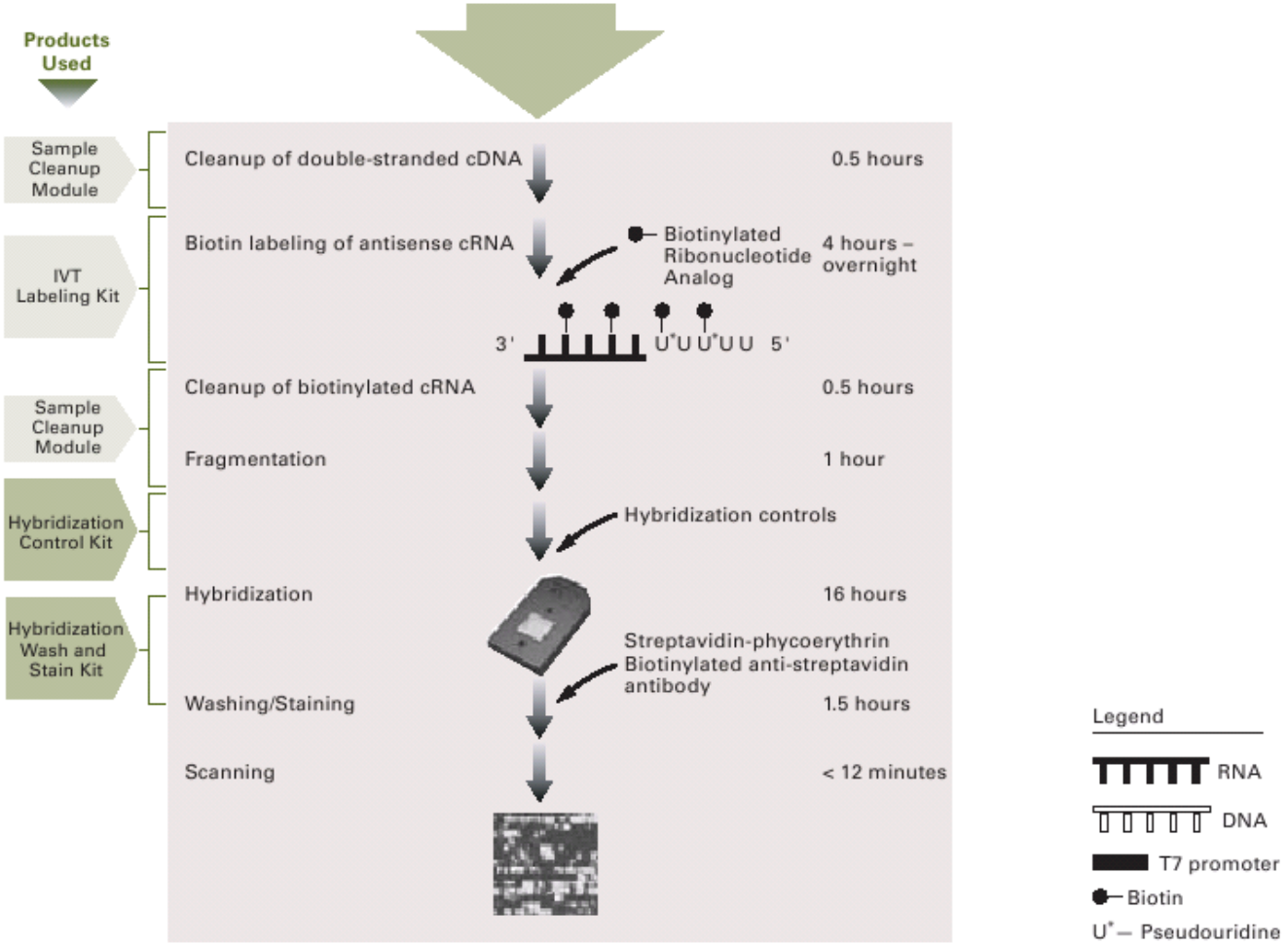
RNA-DNA hybridization complex

- Like DNA, the RNA backbone is also *negatively charged*
- RNA is *less stable* than DNA because it is more prone to hydrolysis
- Hybridisation of RNA to DNA
  - both molecules negatively charged, BUT:
  - hydrogen bonds between complementary bases are sufficient to bind RNA to DNA
  - => spiral structure (like DNA double helix)

cDNA synthesis



Biotin labeling, hybridization, scanning



RNA fragments hybridize to DNA probes

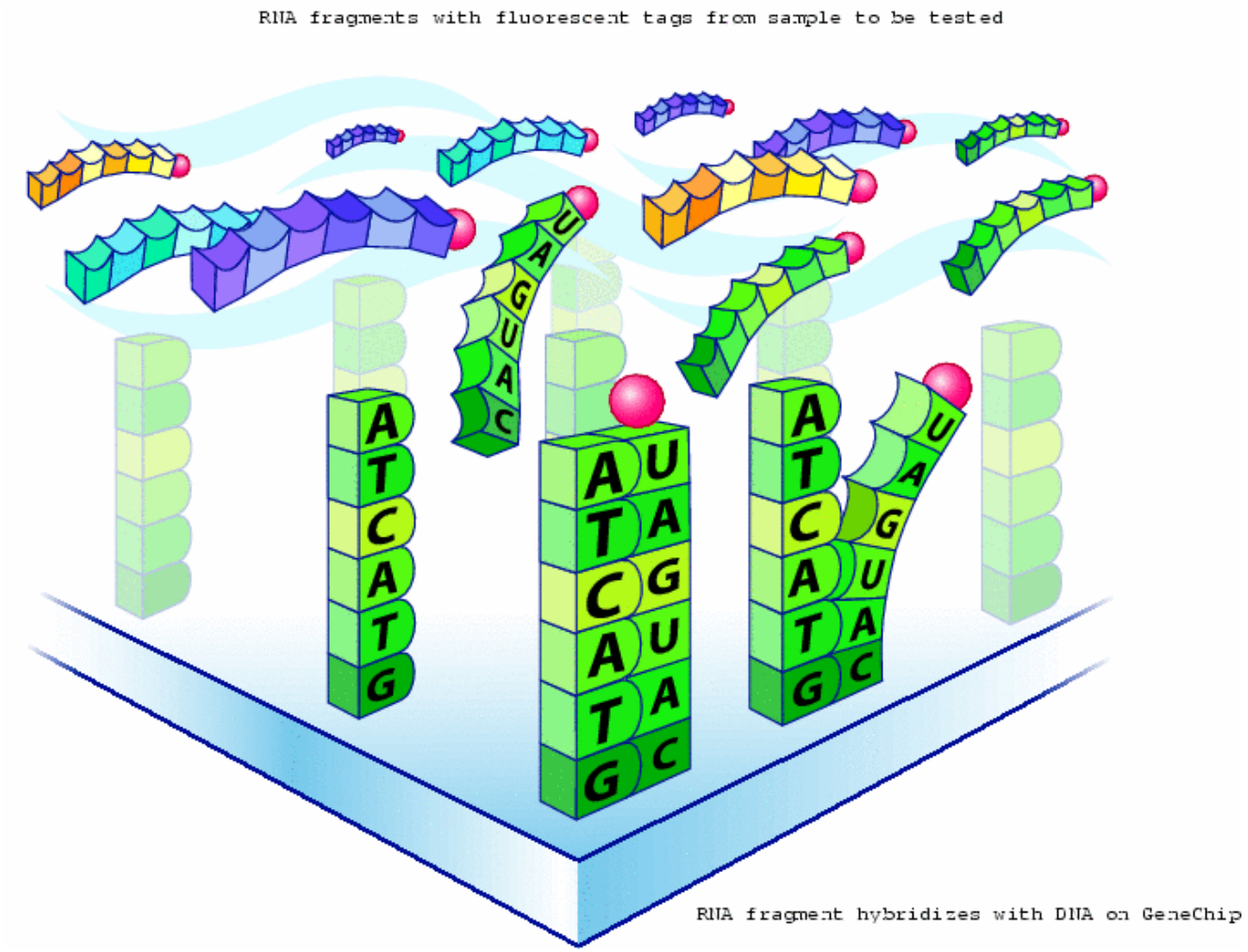
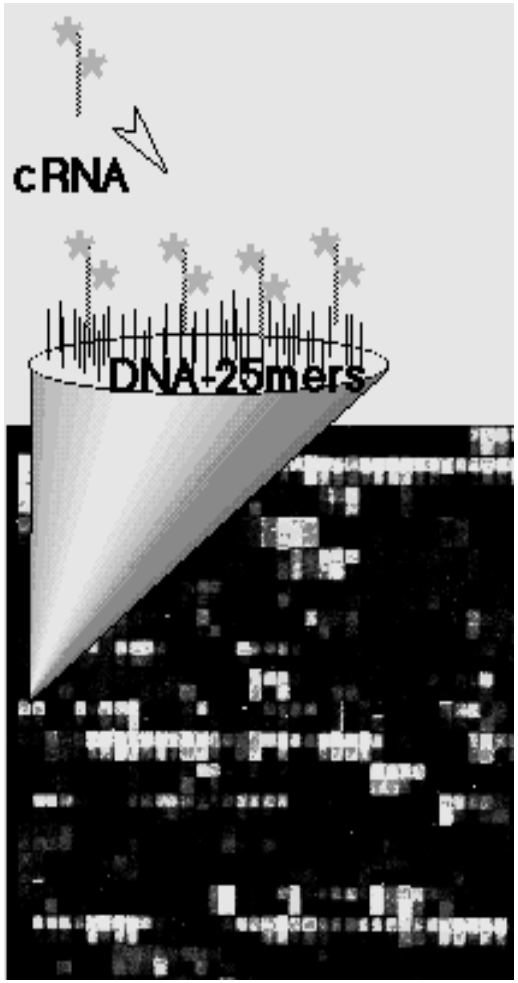


Image analysis



- About 100 pixels per probe cell
- These intensities are combined to form one number representing expression for the probe cell oligo

Measuring expression

- Summarize fluorescence intensities from ~11-20 PM,MM pairs (probe level data) into *one number* for each probe set ('gene')
- Call this number a *measure of expression (ME)*

Expression Measures (examples)

- *MAS 5.0/GCOS* – older Affymetrix
- *PLIER* - (Hubbell, newer Affymetrix)
- *Model Based Expression Index* (MBEI)
  - Li-Wong method, implemented in **dChip** (windows executable)
- *Robust Multichip Analysis* (RMA)
  - Irizarry *et al.*, Bolstad *et al.*; implemented in R package **affy**
  - gcrma (Wu *et al.*)
  - other variants
- *VSN* (Huber *et al.*, Rocke)