

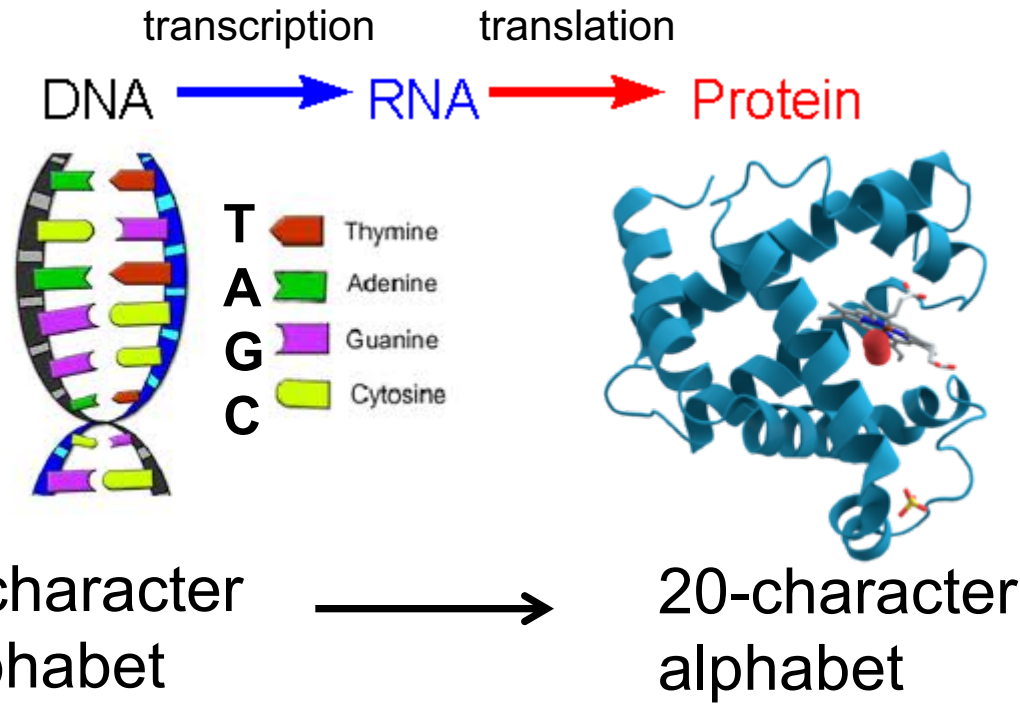
Gene Expression and Microarrays

Garrett M. Dancik, Ph.D.

Note: All images from slide 3 on are from Campbell Biology, 9th edition,
© 2011 Pearson Education, Inc.

Overview of gene expression

Central Dogma of
Molecular Biology:



- A *gene* is a unit of hereditary (DNA) that makes a functional RNA or protein
- The human genome is 3 billion characters long
- The human genome contains ~ 20,000 genes

Overview of gene expression: DNA → RNA → Protein

- Genes are made of DNA, a **nucleic acid** made of monomers called nucleotides
- A gene is a unit of inheritance that codes for the amino acid sequence of a polypeptide

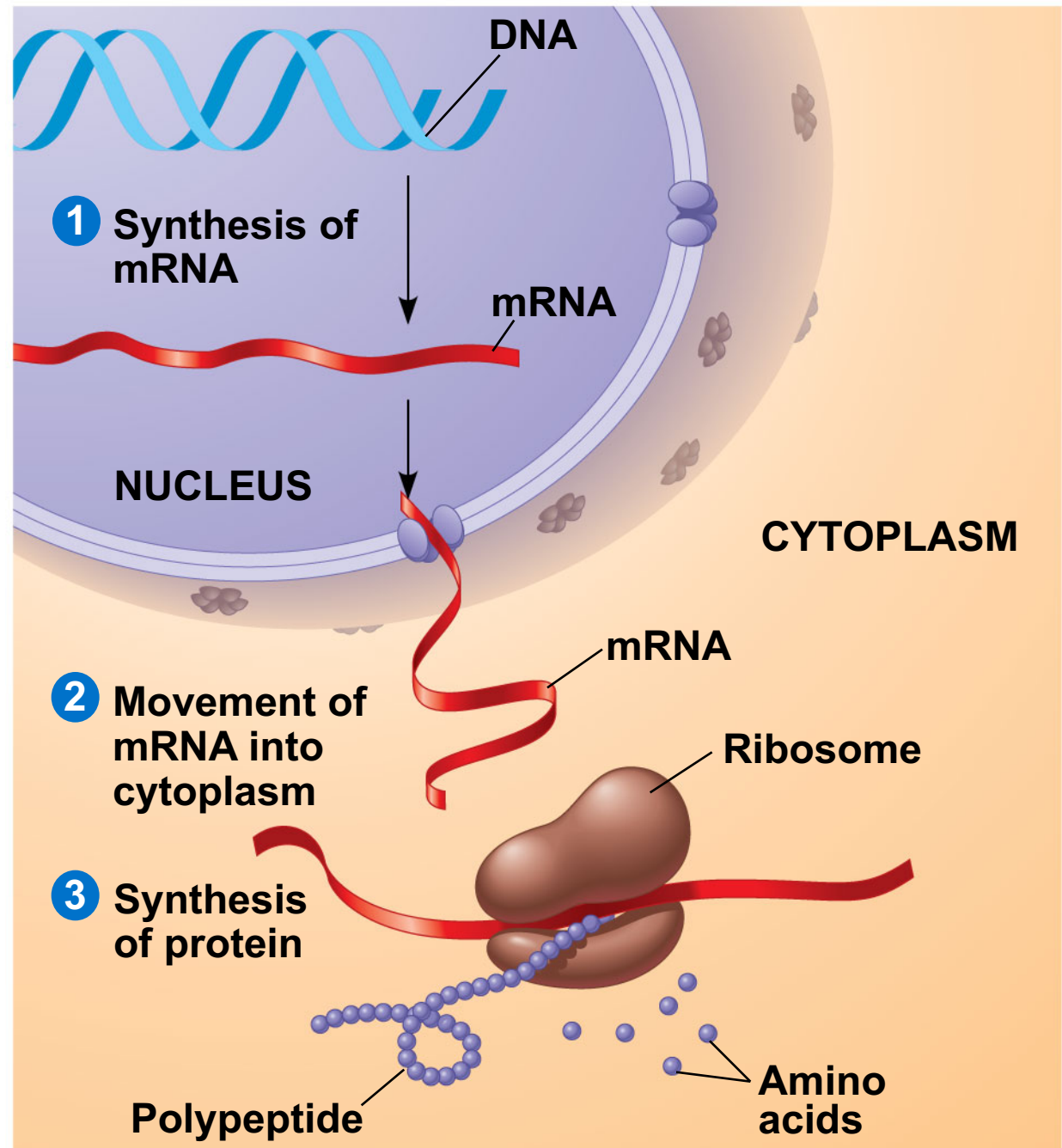
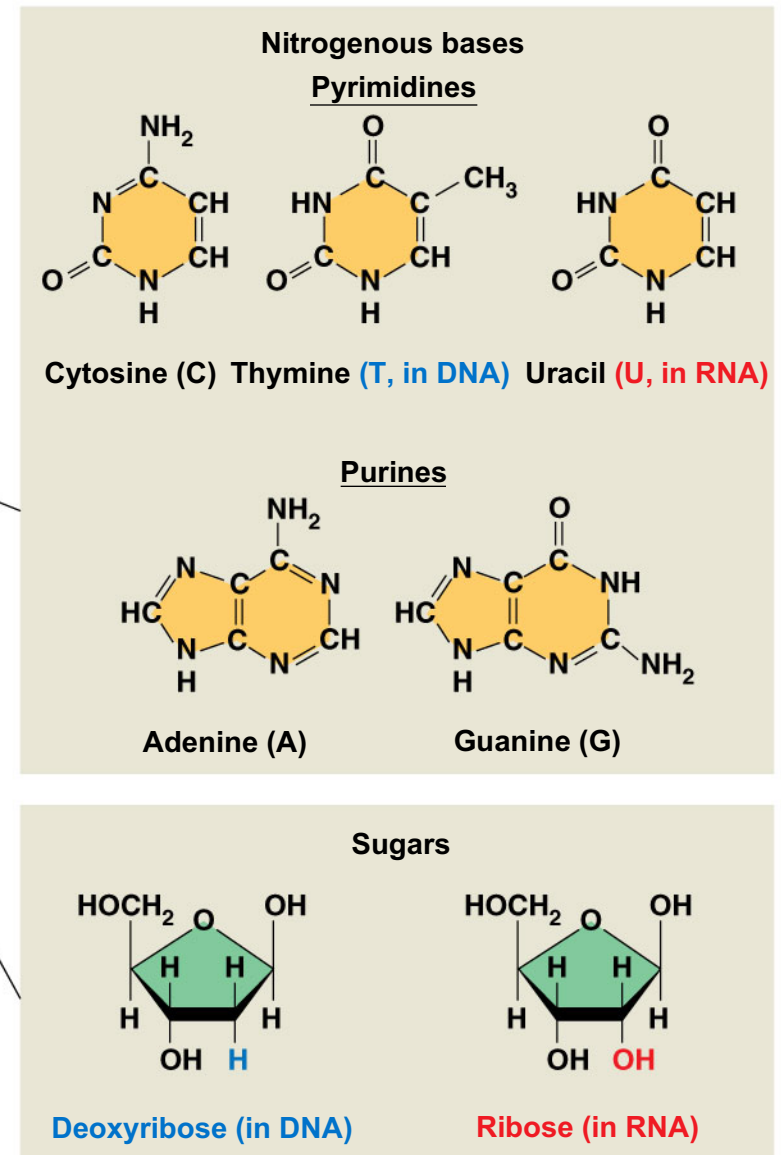
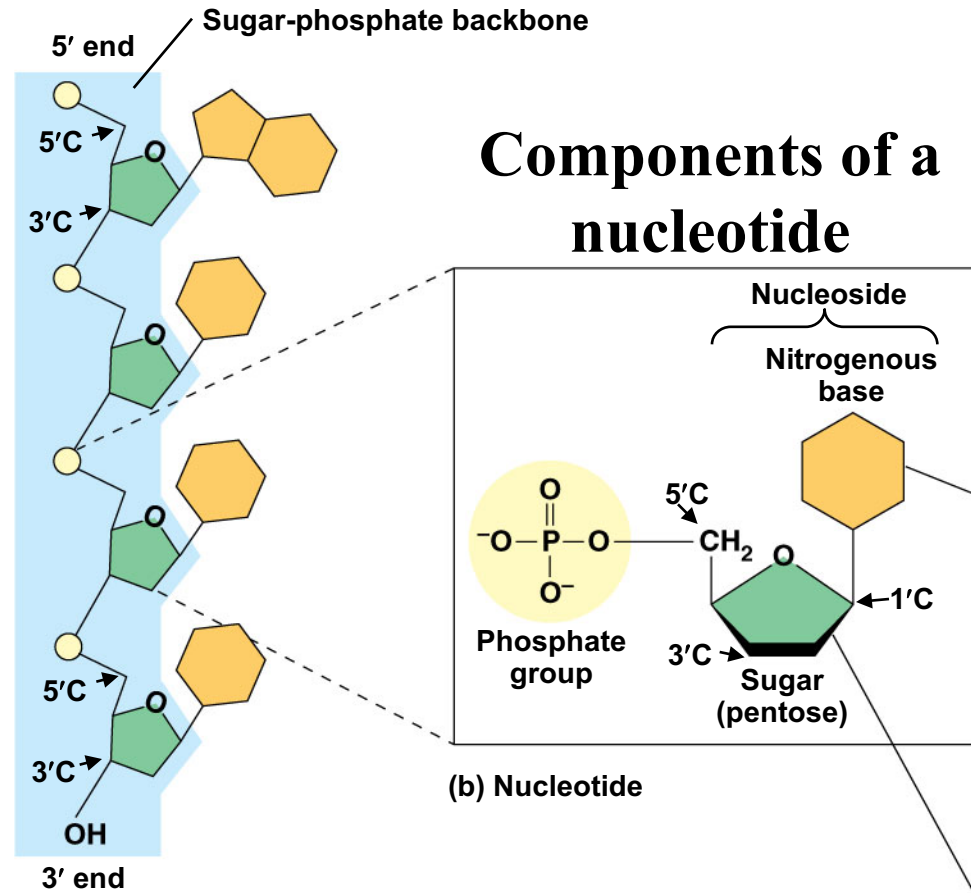


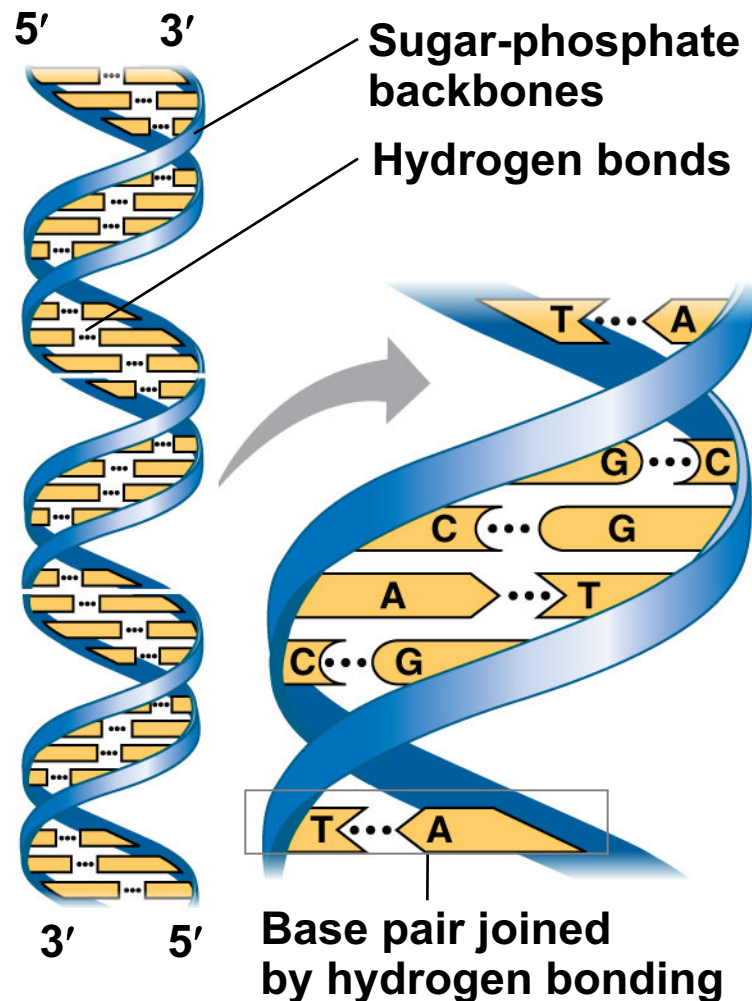
Figure 5.26



(c) Nucleoside components

In DNA, the sugar is **deoxyribose**;
in RNA, the sugar is **ribose**

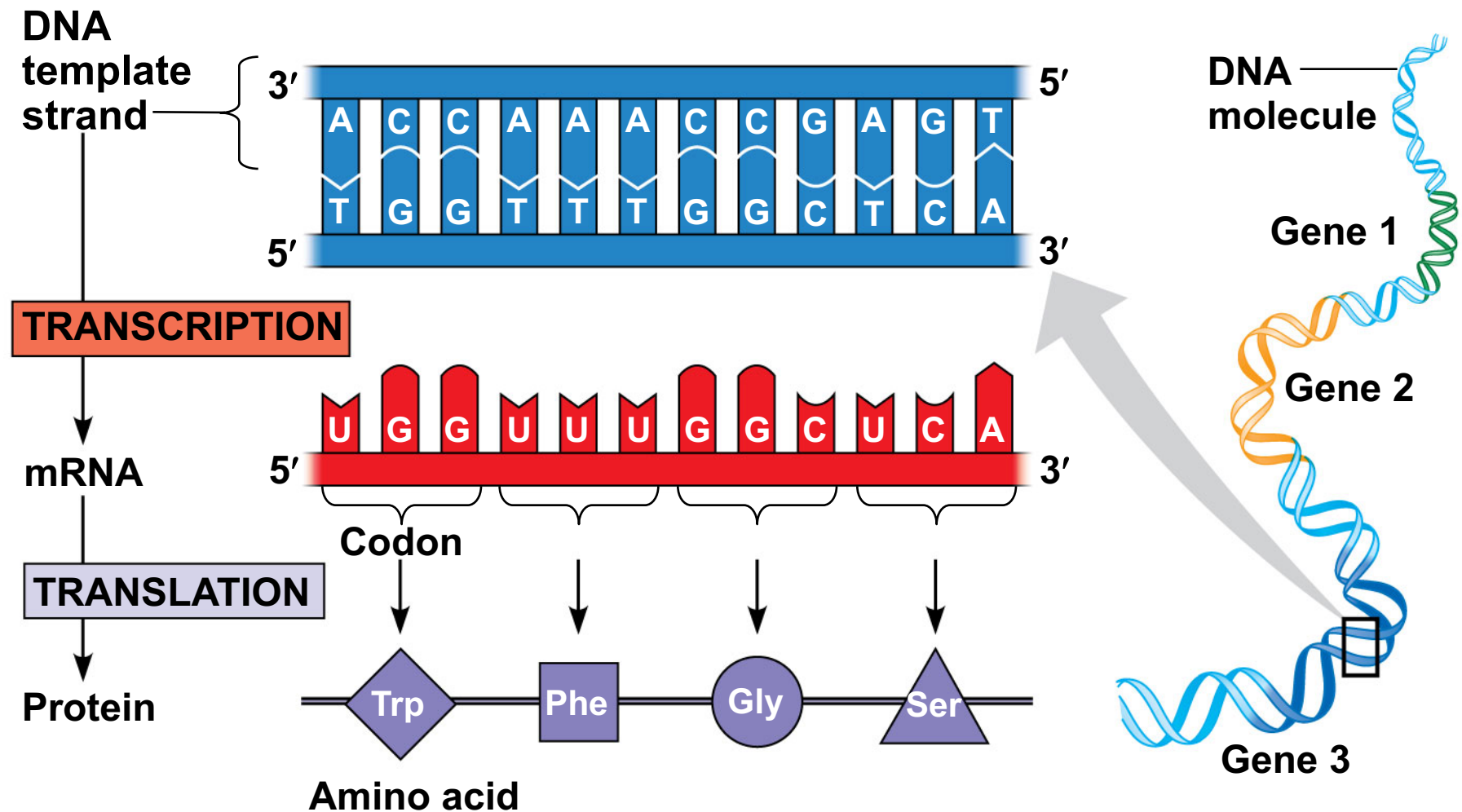
Nucleic Acids are made up of nucleotides



(a) DNA

© 2011 Pearson Education, Inc.

- Complementary base pairing
 - The nitrogenous bases in DNA pair up and form hydrogen bonds: adenine (A) always with thymine (T), and guanine (G) always with cytosine (C)
 - Complementary pairing can also occur between two RNA molecules or between parts of the same molecule
- In RNA, thymine is replaced by uracil (U) so A and U pair



© 2011 Pearson Education, Inc.

- The **genetic code** is a triplet code where a 3-nucleotide DNA word codes for a 3-nucleotide mRNA word (a **codon**) which codes for an amino acid

Mutations of one or a few nucleotides can affect protein structure and function

- **Mutations** are changes in the genetic material of a cell or virus
- **Point mutations** are chemical changes in just one base pair of a gene
 - May or may not change the protein
- **Insertions/deletions** may cause **frameshift** mutations that have a disastrous effect on the protein

Sickle-Cell Disease: A Change in Primary Structure

- A slight change in the amino acid (primary structure) can affect a protein's structure and ability to function
 - What causes a change in the primary structure?
- **Sickle-cell disease**, an inherited blood disorder, results from a single amino acid substitution in the protein hemoglobin

Point mutation that causes sickle cell disease

Wild-type hemoglobin

Wild-type hemoglobin DNA



mRNA



Normal hemoglobin



Sickle-cell hemoglobin

Mutant hemoglobin DNA



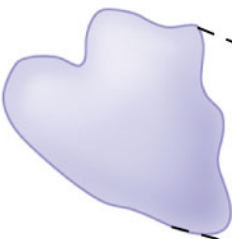
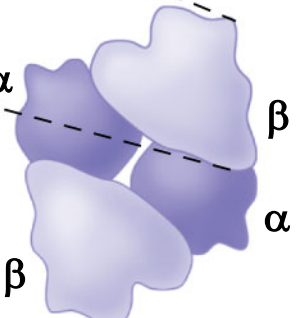
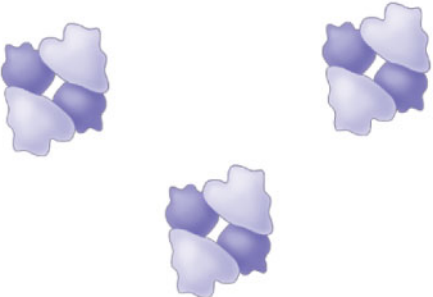
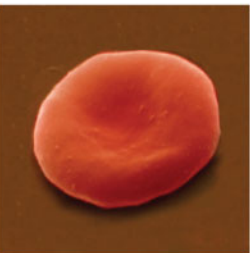

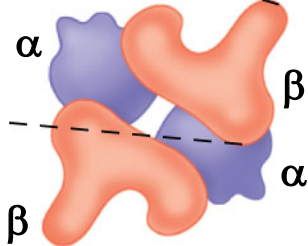
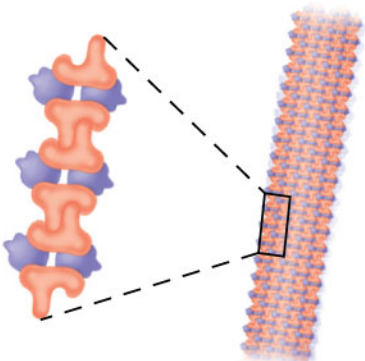

mRNA



Sickle-cell hemoglobin

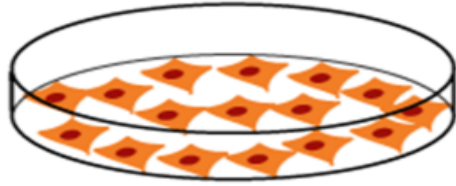


Figure 5.21

	Primary Structure	Secondary and Tertiary Structures	Quaternary Structure	Function	Red Blood Cell Shape
Normal hemoglobin	1 Val 2 His 3 Leu 4 Thr 5 Pro 6 Glu 7 Glu	 <p>β subunit</p>	<p>Normal hemoglobin</p> 	<p>Molecules do not associate with one another; each carries oxygen.</p> 	 <p>10 μm</p>
Sickle-cell hemoglobin	1 Val 2 His 3 Leu 4 Thr 5 Pro 6 Val 7 Glu	<p>Exposed hydrophobic region</p>  <p>β subunit</p>	<p>Sickle-cell hemoglobin</p> 	<p>Molecules crystallize into a fiber; capacity to carry oxygen is reduced.</p> 	 <p>10 μm</p>

Oligonucleotide microarray

Cells of person 1/condition 1



RNA isolation

mRNA

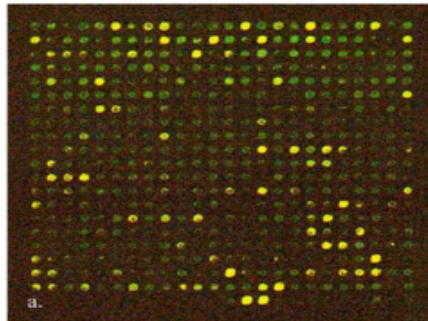
Reverse
transcriptase
labeling

cDNA

"Green Fluorescent" Targets

Hybridize to
microarray

Microarray with
short ssDNA
spanning the entire
genome



----ACTGA----

----ACTGA----

----GAGAT----

Probe 1: TGACT

Probe 2: CTCTA

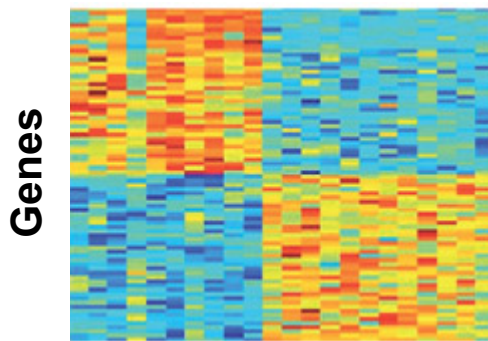
...

Probe 20000: TTTAG

Biomarkers and personalized medicine

Gene expression profiles

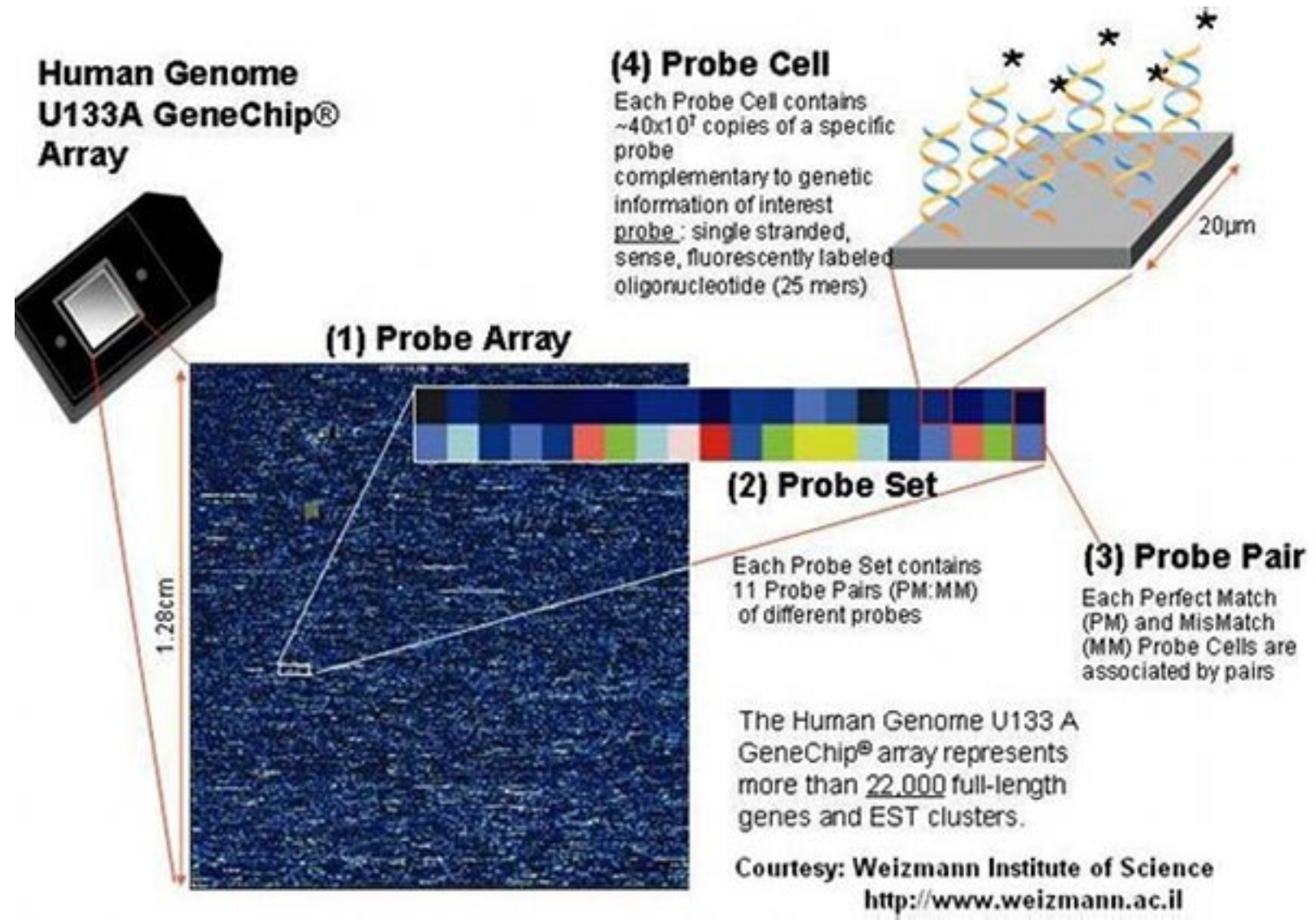
Samples



- Bioinformatics challenges
 - Identification of genes or gene signature
 - Choice of classification method or gene model

Possible comparisons	A	B	Biomarker identification (gene or gene signature)
	Tumor	Normal	<i>Diagnostic:</i> predictive of a clinical variable
	High risk	Low risk	<i>Prognostic:</i> predictive of disease outcome
	Responder	Non-responder	<i>Predictive:</i> predictive of therapeutic response

Microarrays in more detail



Microarray Analysis

- Analysis will be performed using several Bioconductor packages (<http://bioconductor.org>)
- Data is available from the Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>)
 - We will look at how to download raw and processed data from GEO

Gene Expression Omnibus (GEO)

- GEO (<http://www.ncbi.nlm.nih.gov/geo/>) is a public functional genomics data repository for gene expression (microarray) and sequence-based data.
- There are four kinds of records on GEO (<http://www.ncbi.nlm.nih.gov/geo/info/overview.html>)

Gene Expression Omnibus (GEO)

- A **GEO sample** (GSM*) describes an individual sample, including the experimental conditions in which it was collected, and the gene expression value for each element on the array.
- A **GEO platform** (GPL*) is a summary of the array used, and links the array probe to a gene
- A **GEO series** (GSE*) links together a collection of samples with one or more platforms for a particular experiment or study (such as profiling gene expression from 100 patients with lung cancer)
- A **GEO dataset** is a curated collection of samples that allows for user-friendly analysis. Not all series exist as datasets.