BE 183 Applied Genomic Technologies

Lecture 2

Genome Organization and Structure

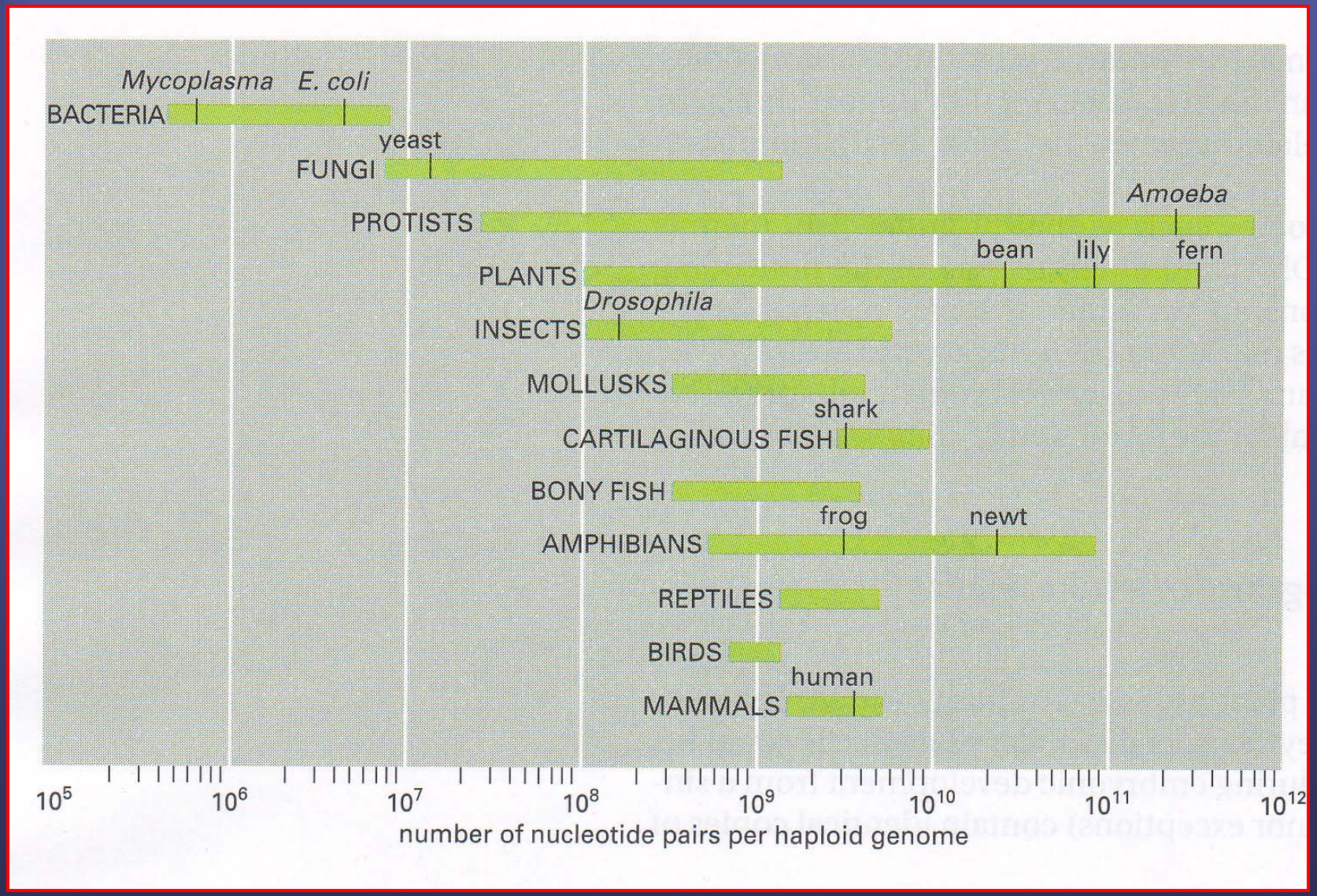
Trey Ideker

Departments of Bioengineering and Medicine
University of California San Diego

Genome Organization

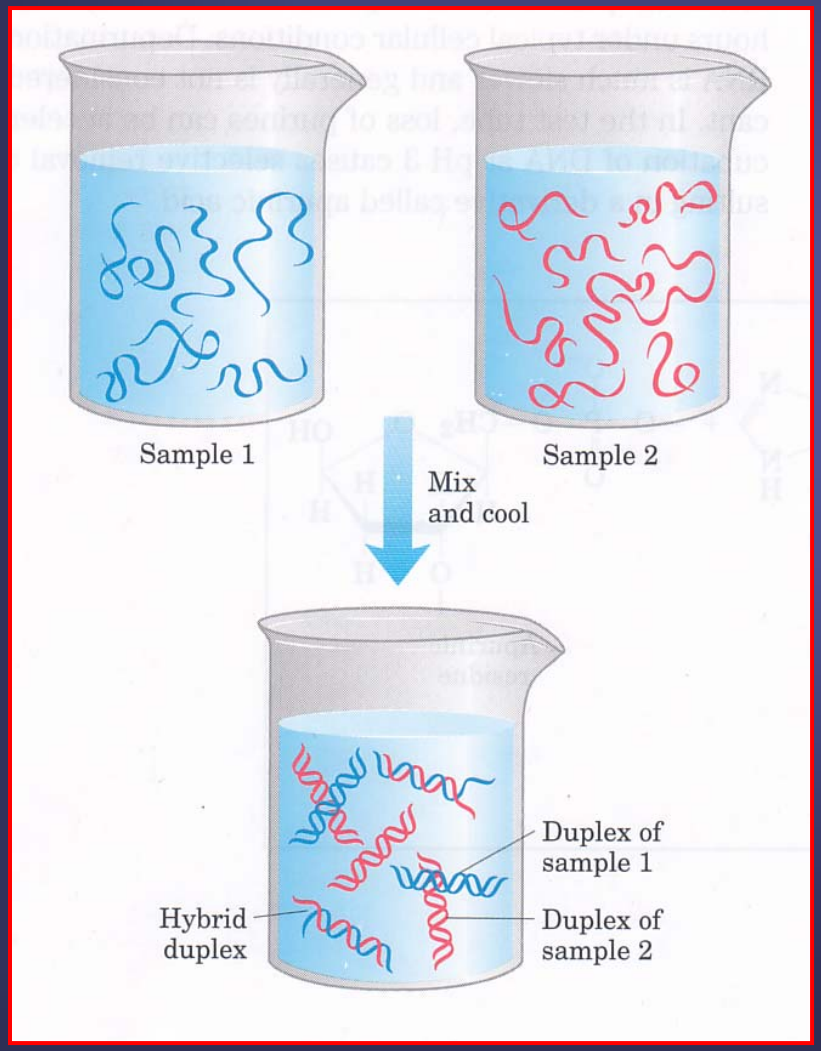
- Genome sizes and the C-value paradox
- DNA Hybridization: A basic technology
- $C_o t$ curves and genome complexity
- Repeated sequences
- Introns and exons
- Genome structure

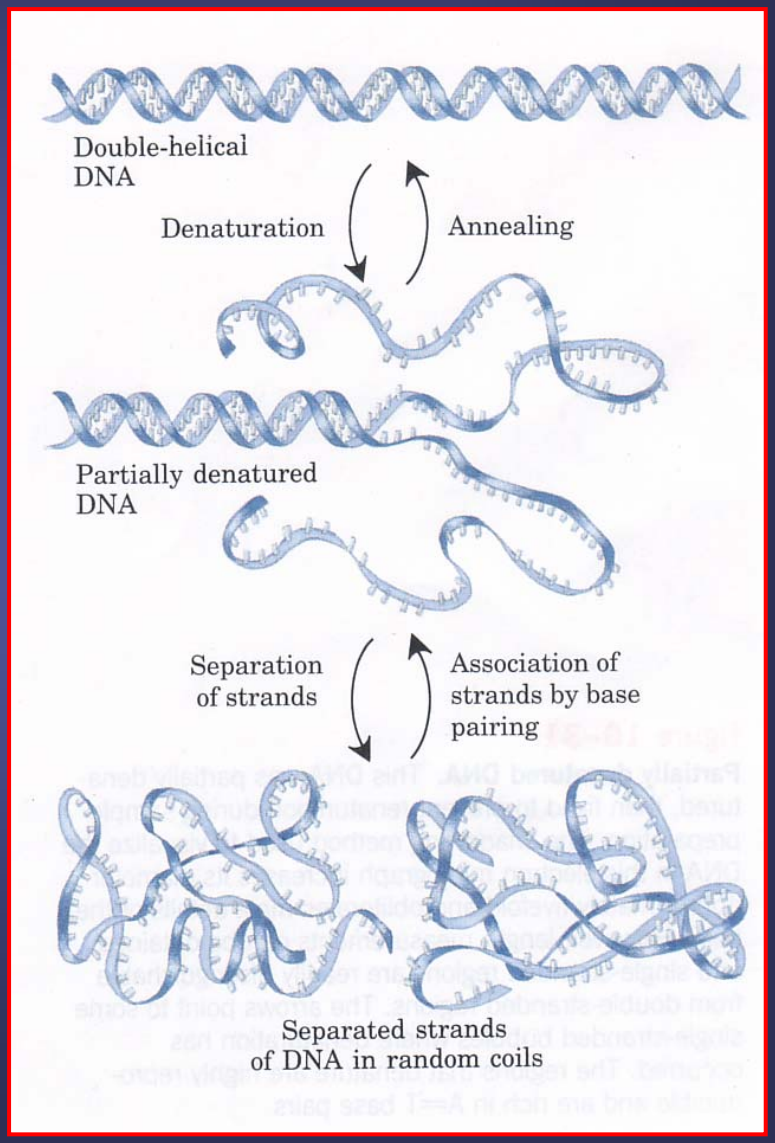
Genome sizes and the *C*-value paradox



Mol Bio Cell 4ed pp. 33

The most basic building block of DNA technology: DNA Hybridization, Denaturation and Annealing





DNA Hybridization Scheme

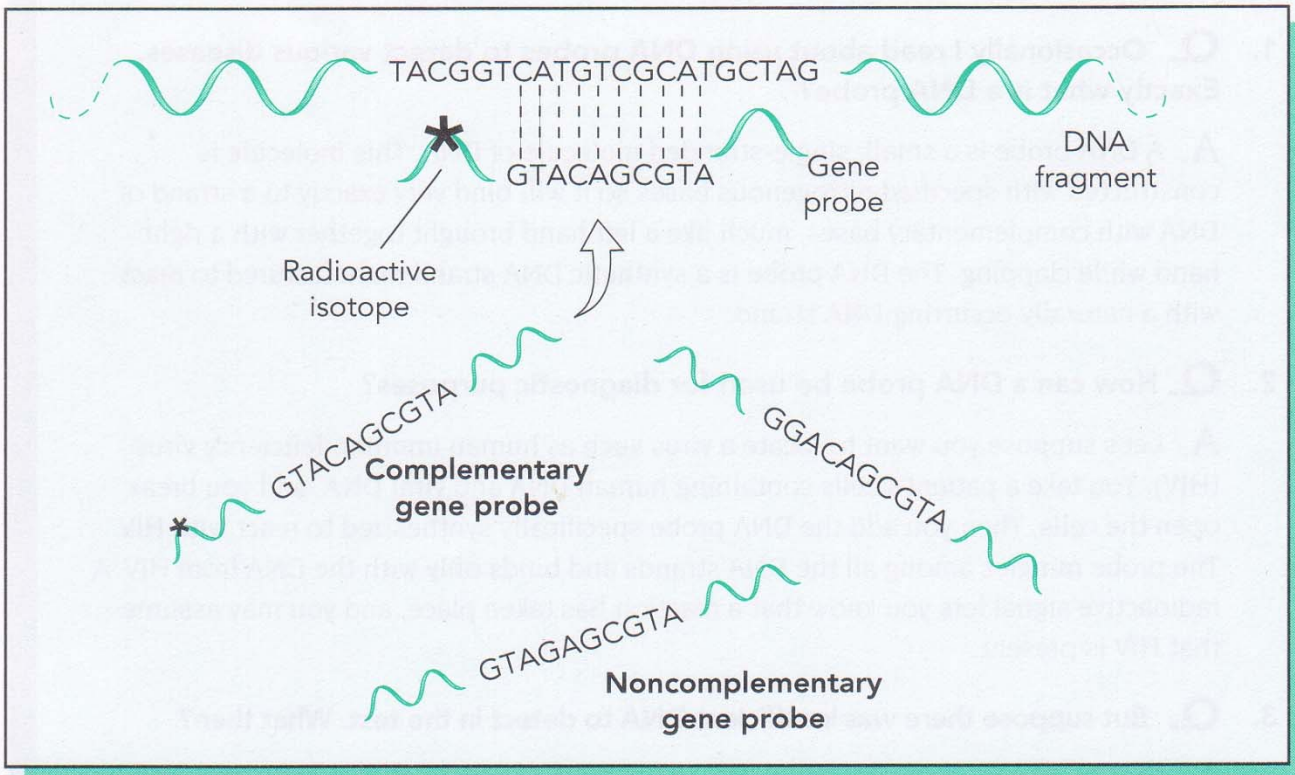
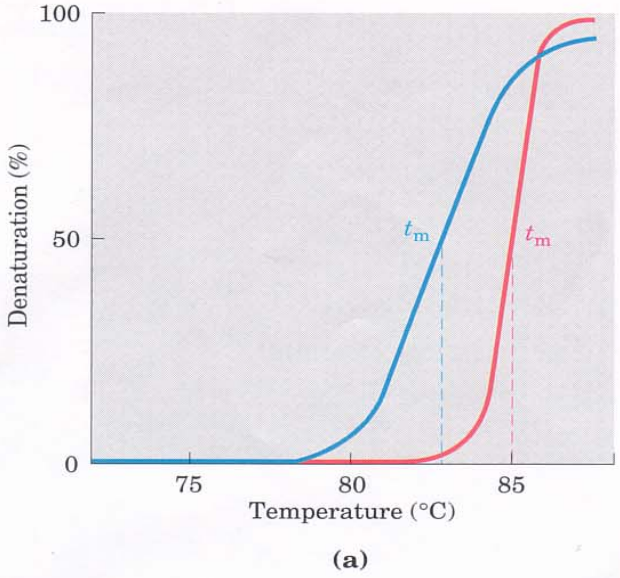


FIGURE 7.1

How a gene probe works. A gene probe is a single-stranded segment of DNA. When combined with a DNA molecule containing a complementary site, the gene probe seeks out the site and binds with it. If a radioactive molecule or atom is attached to the probe, the radioactivity accumulates at the binding site and signals that a reaction has taken place. Note in the diagram how the bases of only one probe complement the bases of the DNA fragment.

Hybridization - Heat denaturation, melting temperature (tm), other factors



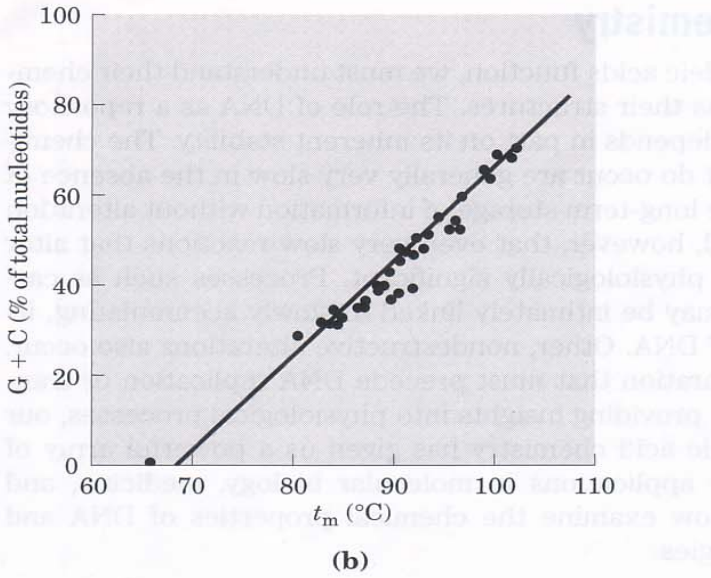
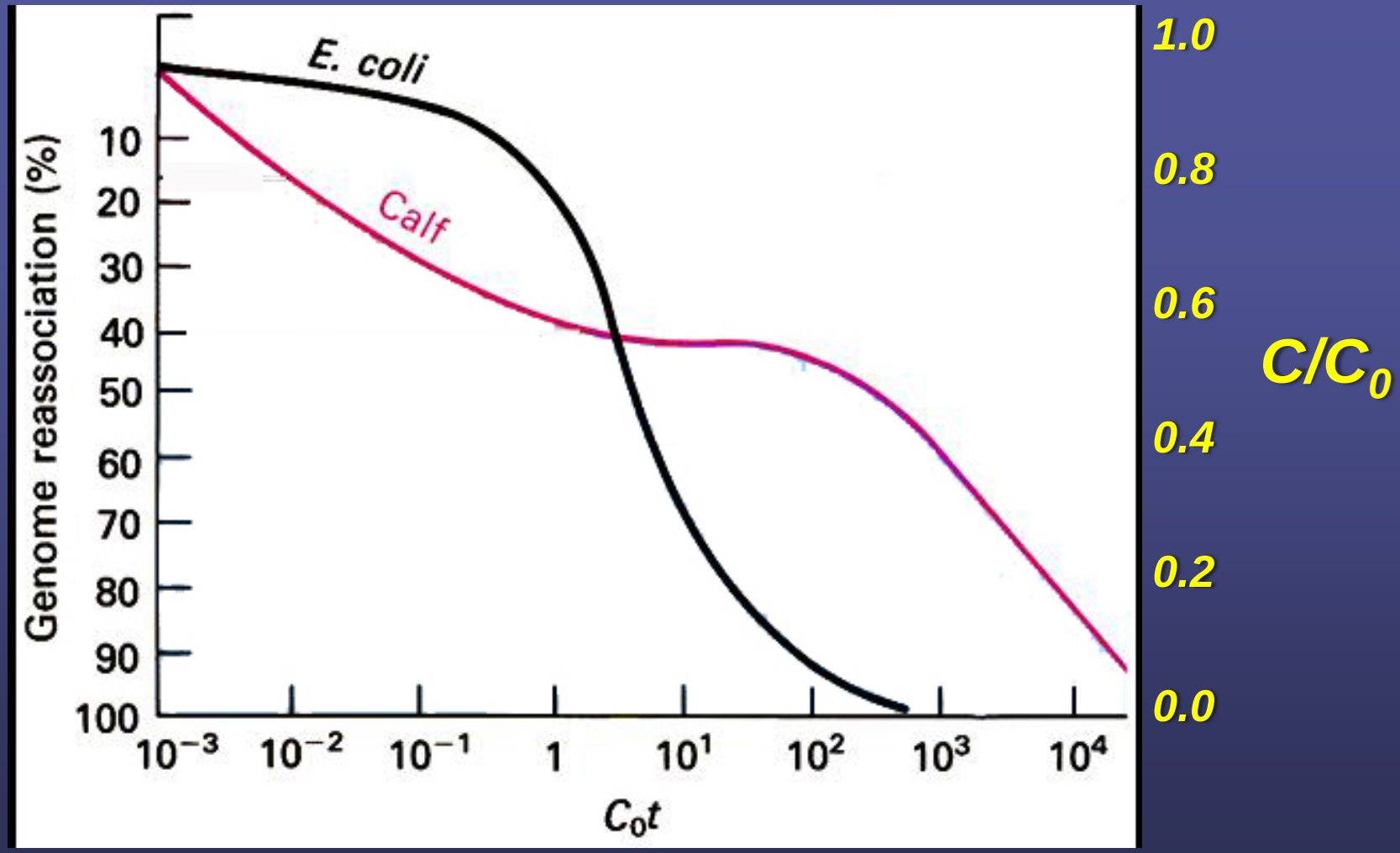


figure 10–30

Heat denaturation of DNA. **(a)** The denaturation or melting curves of two DNA specimens. The temperature at the midpoint of the transition (t_m) is the melting point; it depends on pH and ionic strength and on the size and base composition of the DNA. **(b)** Relationship between t_m and the G≡C content of a DNA.

Temperature, pH, size (# bp's),
G/C to A/T ratio, ionic strength,
chem denaturants, detergents, chaotropics
Stringency (high and low)

Reassociation – the opposite of denaturation



Reassociation kinetics- The $C_o t$ curve

C = Concentration of ssDNA

C_0 = Initial ssDNA conc.

k = reassociation rate const.

$t_{1/2}$ = reassociation half time

Big $C_0 t_{1/2}$ = Slow reassociation

This value is proportional to the number of different types of DNA fragments

$$\frac{dC}{dt} = -kC^2$$

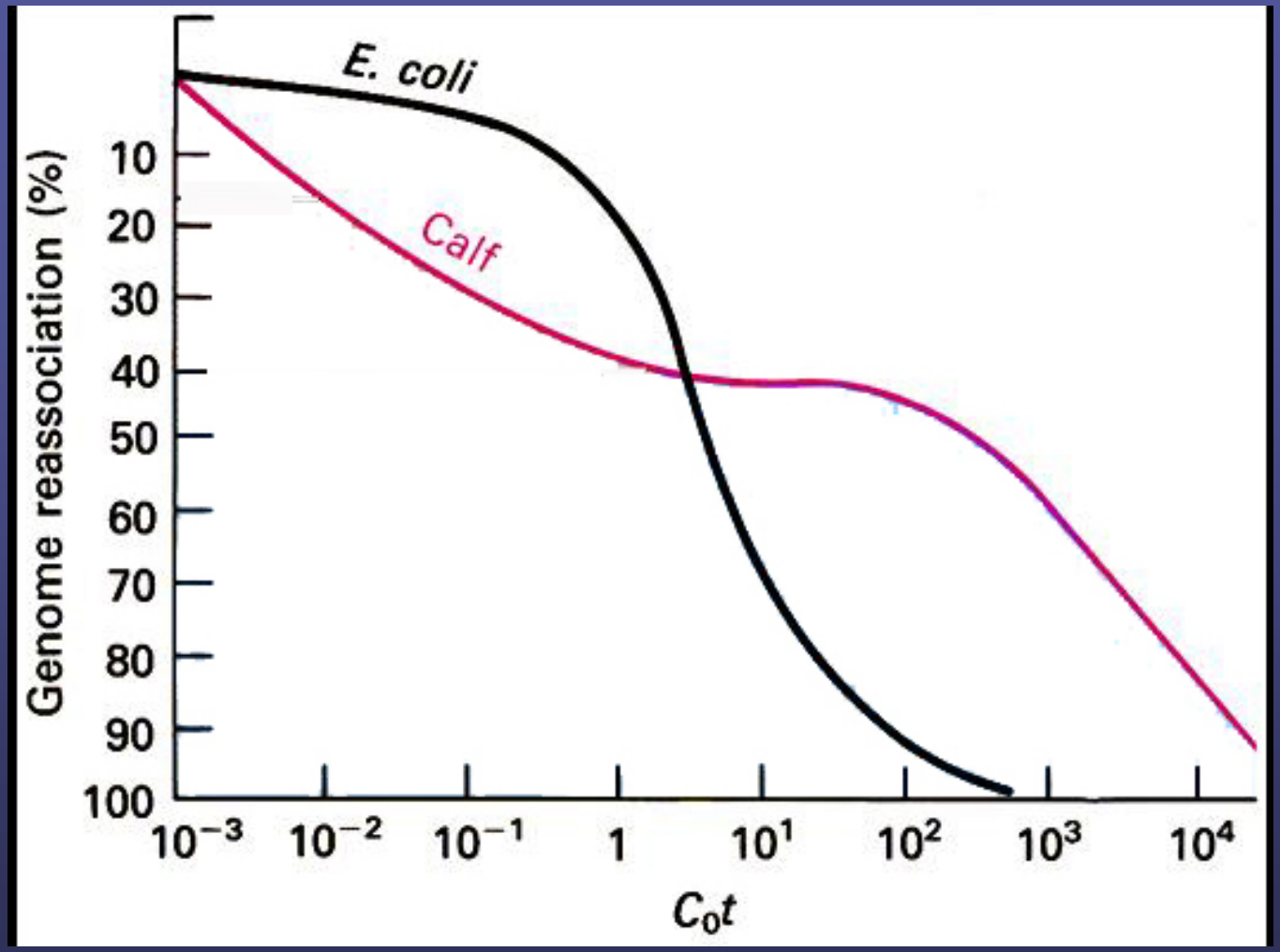
$$\frac{C}{C_0} = \frac{1}{1 + kC_0 t}$$

Comparison of sequence copy number for two organisms with different genome sizes

	Organism A	Organism B
Starting DNA concentration	10 pg/ml	10 pg/ml
Genome size	0.01 pg	2 pg
# genome copies/ml	1000	5
Relative concentration	200	1

Table 2.1 Primrose and Twyman

So why the striking difference in species? How do we interpret the curve for cow?



The $C_o t$ curve– many apparently "large" genomes are filled with repetitive sequences (resolution of the *C*-value paradox)

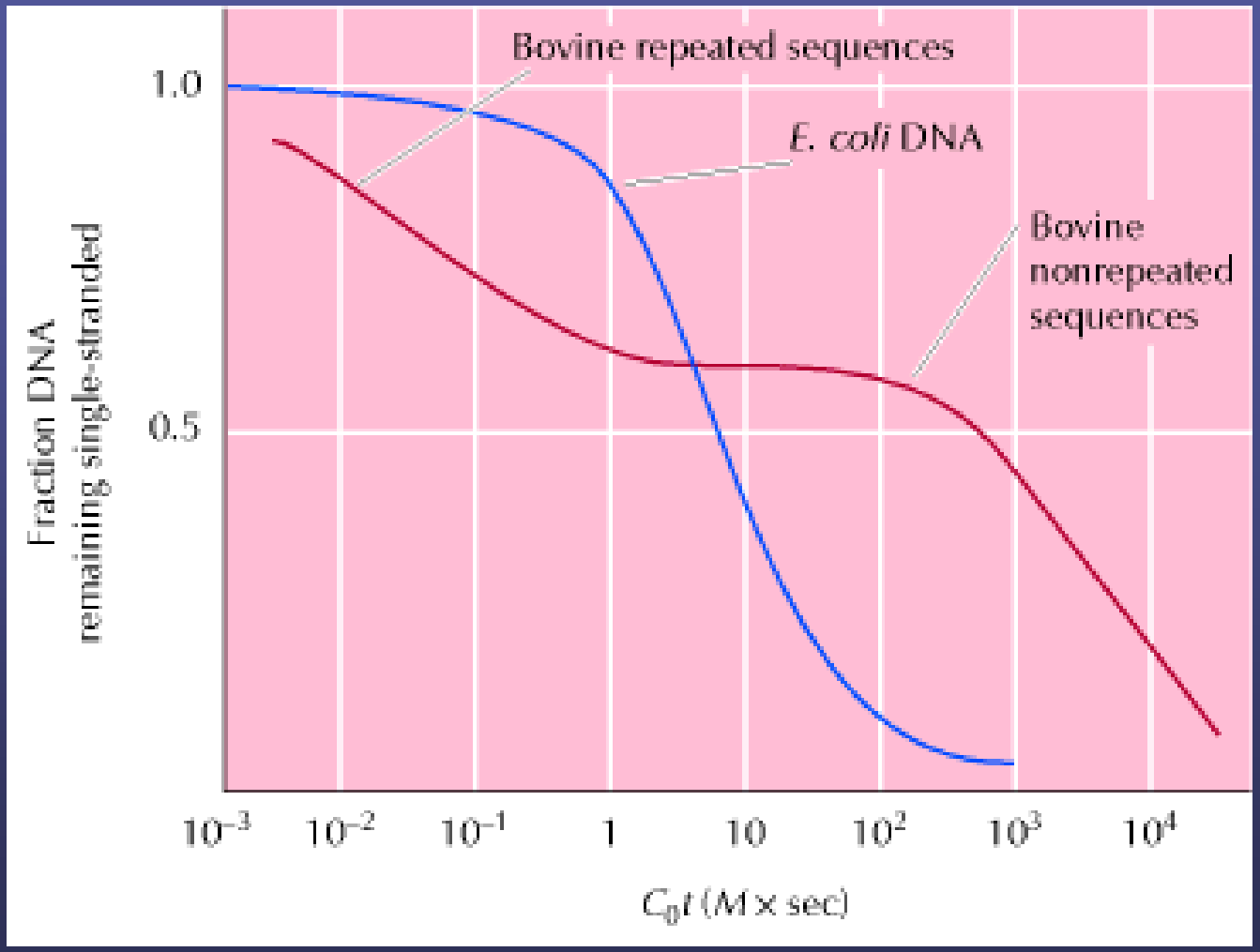


Fig. 4.6
The Cell: A Molecular Approach

Genome Organization

- Genome sizes and the C-value paradox
- DNA Hybridization: A basic technology
- $C_o t$ curves and genome complexity
- **Repeated sequences**
- **Introns and exons**
- **Genome structure**

Satellite DNA
CsCl density gradient column

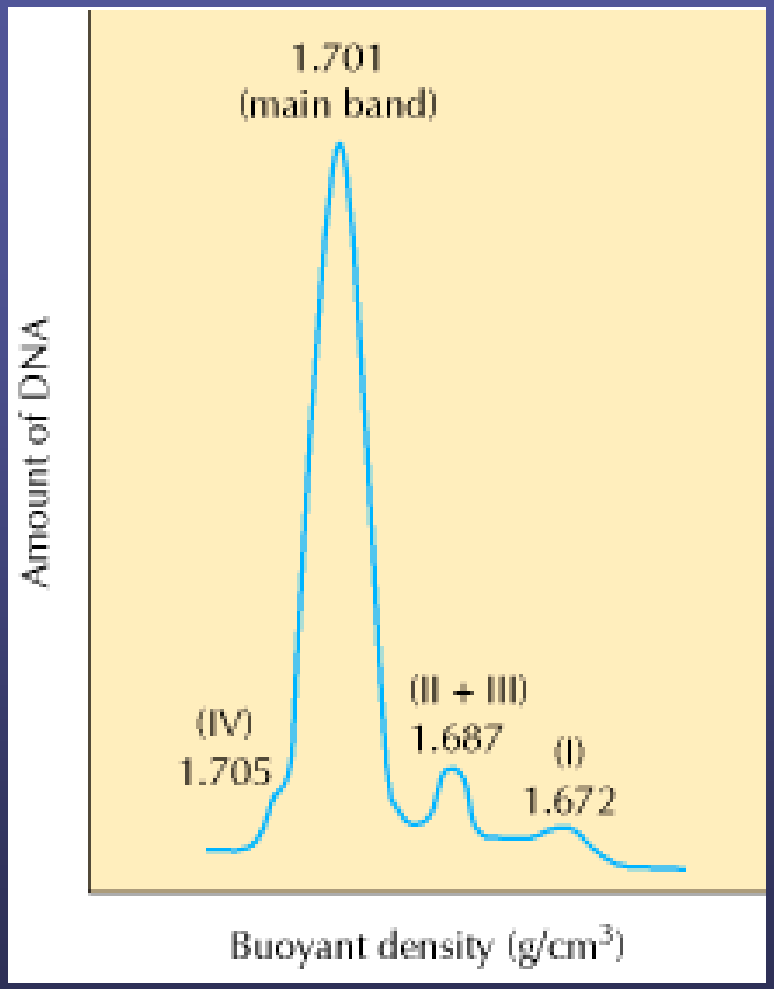
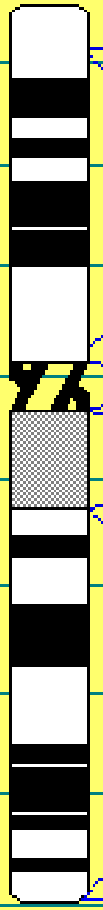


Fig. 4.7
The Cell: A Molecular Approach

Chromosomal organization of repeated sequences

chr 16



subtelomeric repeats

interspersed and tandem repeats — retroposons, minisatelites, microsatellites

tandem repeat — alpha satellite (340 bp dimer; 1-2 Mb)

tandem repeat — satellite II (5 bp monomer; 2-7 Mb)

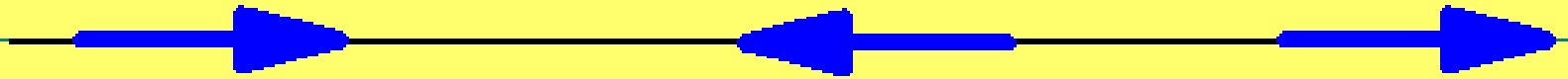
interspersed and tandem repeats — retroposons, minisatelites, microsatellites

tandem repeat — telomer $(TTAGGG)_n$ 5-12 kb

Organization of repeated sequences in the genome

- tandemly arrayed

- interspersed



Tandem vs. Interspersed repeats

- Tandem
  - Satellites, mini and microsatellites (VNTRs)

- Interspersed
  - Retrotransposons (class I)
    Autonomous
    LINE (10% of human genome)
  - Transposons (class II)
    Non-autonomous
    SINE (Alu- 10% of human genome)

Resources: Repbase, RepeatMasker

Genome Structure

- Linear/Circular/Segmented
- Centromere/Telomere (TTAGGG)
- Origin of replication
- Heterochromatin/Euchromatin
- GC content, GC isochores
- CpG islands
- Exons/Introns

G-banding patterns of human chromosomes

Mol Bio Cell 4ed
pp. 199

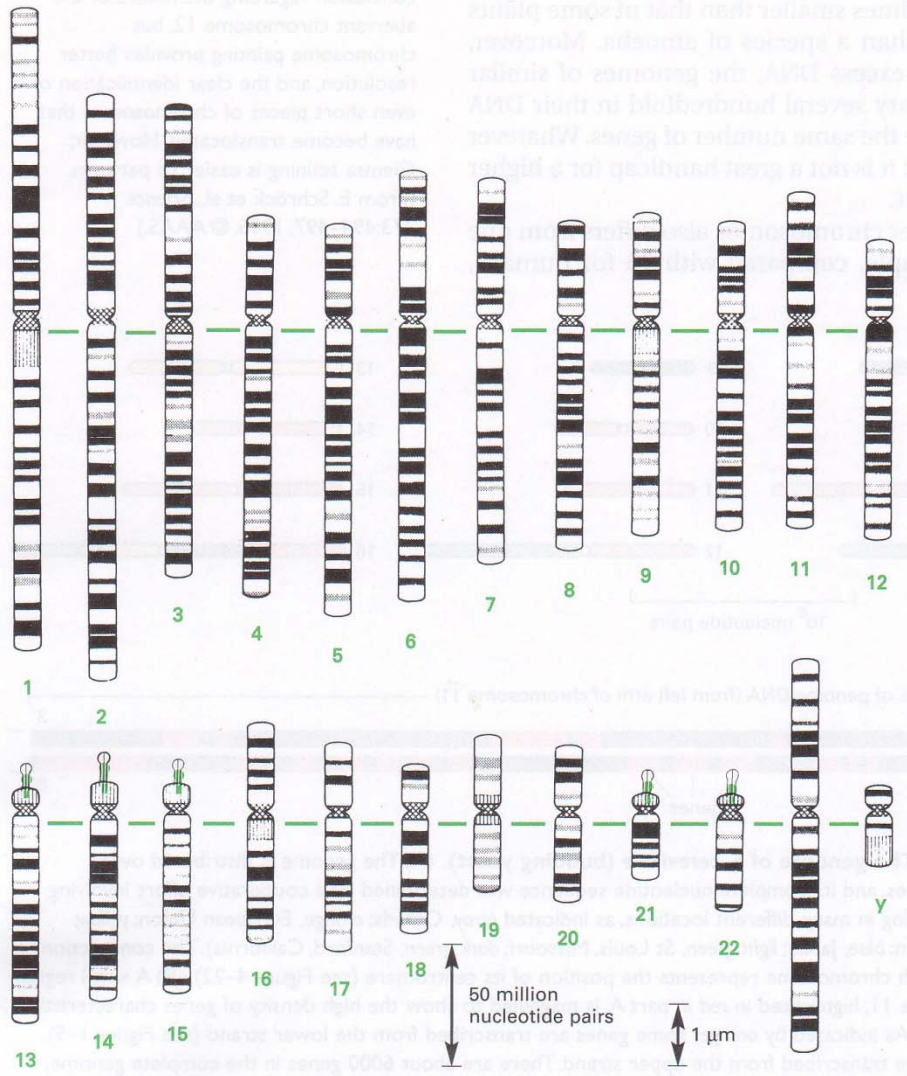
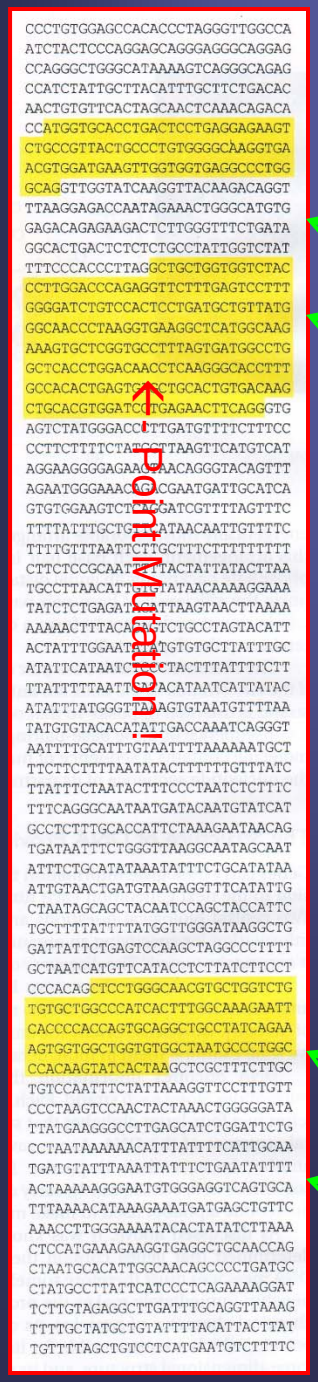


Figure 4–11 The banding patterns of human chromosomes. Chromosomes 1–22 are numbered in approximate order of size. A typical human somatic (non-germ line) cell contains two of each of these chromosomes, plus two sex chromosomes—two X chromosomes in a female, one X and one Y chromosome in a male. The chromosomes used to make these maps were stained at an early stage in mitosis, when the chromosomes are incompletely compacted. The *horizontal green line* represents the position of the centromere (see Figure 4–22), which appears as a constriction on mitotic chromosomes; the knobs on chromosomes 13, 14, 15, 21, and 22 indicate the positions of genes that code for the large ribosomal RNAs (discussed in Chapter 6). These patterns are obtained by staining chromosomes with Giemsa stain, and they can be observed under the light microscope. (Adapted from U. Franke, *Cytogenet. Cell Genet.* 31:24–32, 1981.)

Giemsa staining- AT rich

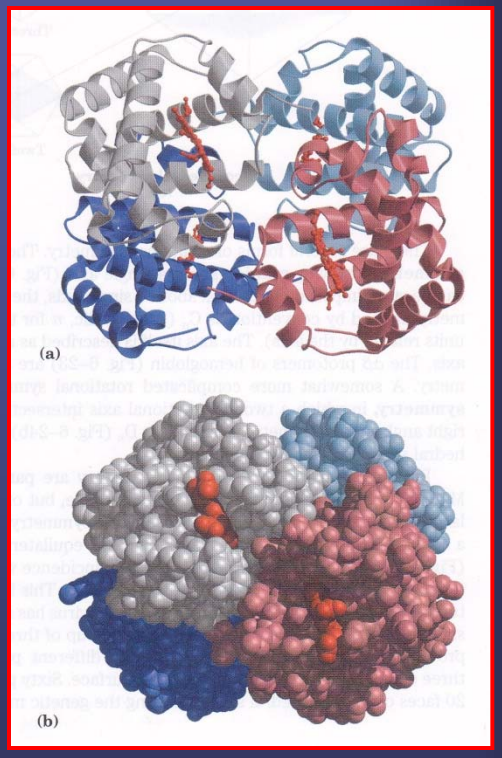
Naming: e.g., 2p11

Split genes: Introns and Exons



Sequence of Beta-Globin Gene

Figure 4–7 The nucleotide sequence of the human β-globin gene. This gene carries the information for the amino acid sequence of one of the two types of subunits of the hemoglobin molecule, which carries oxygen in the blood. A different gene, the α-globin gene, carries the information for the other type of hemoglobin subunit (a hemoglobin molecule has four subunits, two of each type). Only one of the two strands of the DNA double helix containing the β-globin gene is shown; the other strand has the exact complementary sequence. By convention, a nucleotide sequence is written from its 5′ end to its 3′ end, and it should be read from left to right in successive lines down the page as though it were normal English text. The DNA sequences highlighted in *yellow* show the three regions of the gene that specify the amino sequence for the β-globin protein. We see in Chapter 6 how the cell connects these three sequences together to synthesize a full-length β-globin protein.



Hemoglobin Protein Structure

Lehninger pp. 196 & 189

Distribution of exons in three species

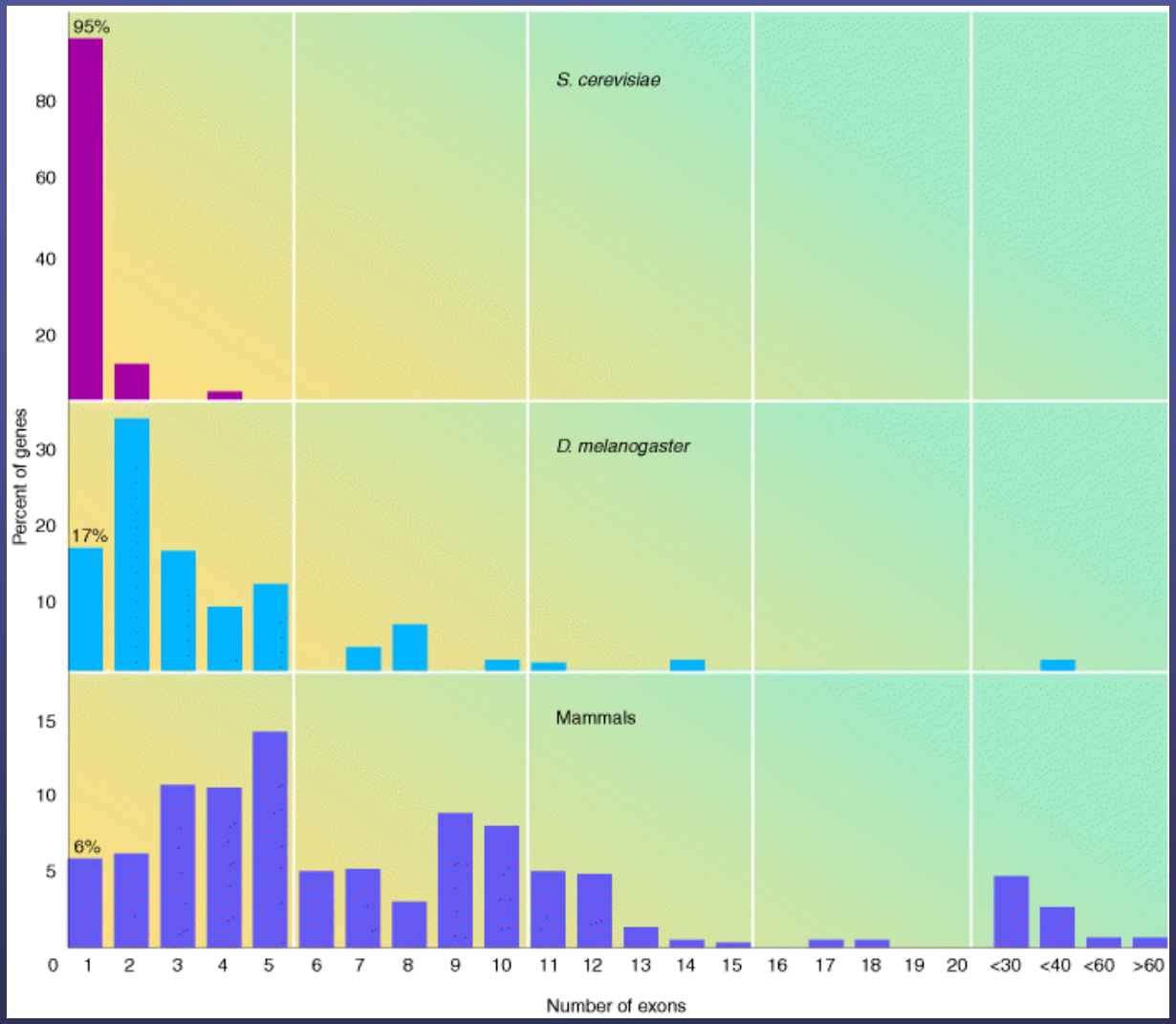


Figure 2.7 Primrose and Twyman

Given these features, how might one write a gene finder?

21

Towards writing a gene finding program: Characteristics of Open Reading Frames (ORFs)

Prokaryotes

- contiguous ORFs, no introns
- very little intergenic sequence
- with f(A,C,G,T) = 25%, ORF>300 bp every 36 kb on a single strand
- detecting large ORFs is a very good predictor for genes (with good specificity)

Eukaryotes

- typically 6 exons (150 bp) over ~30 kb
- Exceptions
  - 2.4 Mb (dystrophin gene)
  - 186 kb with 26 exons (69-3106 bp), 32.4 kb intron (blood coagulation factor VIII gene)
- ORFs >225 bp randomly every kb on a single strand
- detecting ORFs is NOT a good predictor for eukaryotic genes