

CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS08.html

Evaluation

<input type="checkbox"/> Semester Project	(45 %)
<input type="checkbox"/> Homework Assignments	(20 %)
<input type="checkbox"/> Exam	(15 %)
<input type="checkbox"/> Quizzes	(10 %)
<input type="checkbox"/> Summary Reports of Interest	(5 %)
<input type="checkbox"/> Class Participation	(5 %)

Course Homepage

<http://www.cis.fiu.edu/~giri/teach/BioinfS11.html>

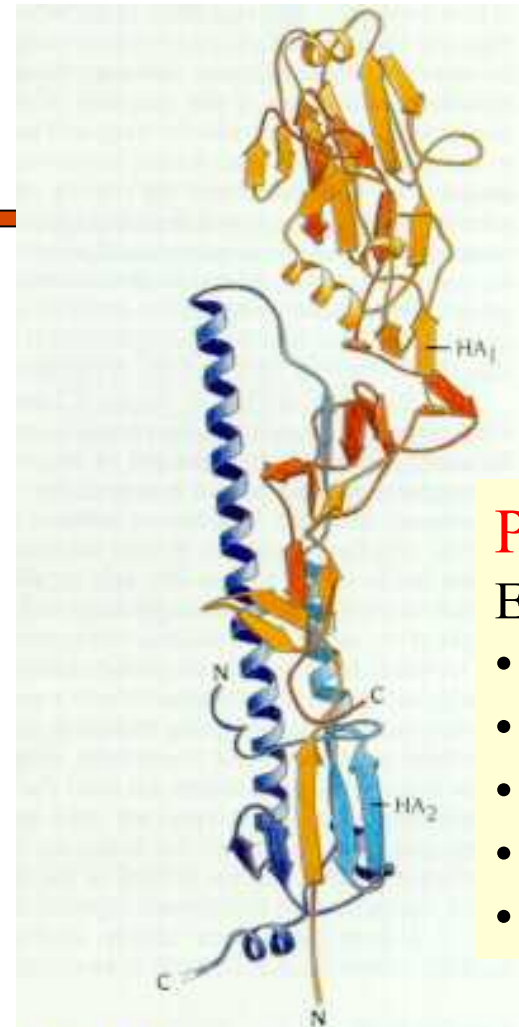
- Lecture notes, required reading material, homework, announcements, etc.*

Molecular Biology Background

2 star molecular players

Also Starring:
RNA

DNA



Protein

Examples:

- Hemoglobin
- Melanin
- Insulin
- Keratin
- RNA Polymerase

Figure S.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA₁ (red) and HA₂ (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75Å long. The globular head is formed by residues only from HA₁. (Courtesy of Don Wiley, Harvard University.)

The Polymeric Players

DNA

String with alphabet {A, C, G, T} **Nucleotides/Bases**

RNA

String with alphabet {A, C, G, U} **Nucleotides/Bases**

Protein

String with 20-letter alphabet **Amino acids/Residues**

Typical DNA Sequence

```
1  gggagaacac  ccggagaagg  aggaggaggc  gaagaaaagc  aacagaagcc  cagttgctgc
61  tccaggtccc  tcggacagag  ctttttccat  gtggagactc  tctcaatgga  cgtgccccct
121 agtgcttctt  agacggactg  cggctctccta  aaggctcgacc  atgggtggccg  ggacccgctg
181 tcttctagtg  ttgctgcttc  cccaggctcct  cctgggcggc  gcggccggcc  tcattccaga
241 gctgggccgc  aagaagttcg  ccgcggcatc  cagccgacce  ttgtcccggc  cttcggaaga
301 cgtcctcagc  gaatttgagt  tgaggctgct  cagcatgttt  ggcctgaagc  agagaccac
361 cccagcaag  gacgtcgtgg  tgcccccta  tatgctagat  ctgtaccgca  ggcactcagg
421 ccagccagga  gcgcccggcc  cagaccaccg  gctggagagg  gcagccagcc  gcgccaacac
481 cgtgcgcagc  ttccatcacg  aagaagccgt  ggaggaactt  ccagagatga  gtgggaaaac
541 ggcccggcgc  ttcttcttca  atttaagttc  tgtccccagt  gacgagtttc  tcacatctgc
601 agaactccag  atcttccggg  aacagataca  ggaagctttg  ggaaacagta  gtttccagca
661 ccgaattaat  atttatgaaa  ttataaagcc  tgcagcagcc  aacttgaaat  ttctgtgac
721 cagactattg  gacaccaggt  tagtgaatca  gaacacaagt  cagtgggaga  gcttcgacgt
781 caccagct  gtgatgcggt  ggaccacaca  gggacacacc  aaccatgggt  ttgtggtgga
841 agtggcccat  ttagaggaga  acccaggtgt  ctccaagaga  catgtgagga  ttagcaggtc
901 tttgcaccaa  gatgaacaca  gctggtcaca  gataaggcca  ttgctagtga  cttttggaca
961 tgatggaaaa  ggacatccgc  tccacaaacg  agaaaagcgt  caagccaaac  acaaacagcg
```

The building blocks of DNA & RNA

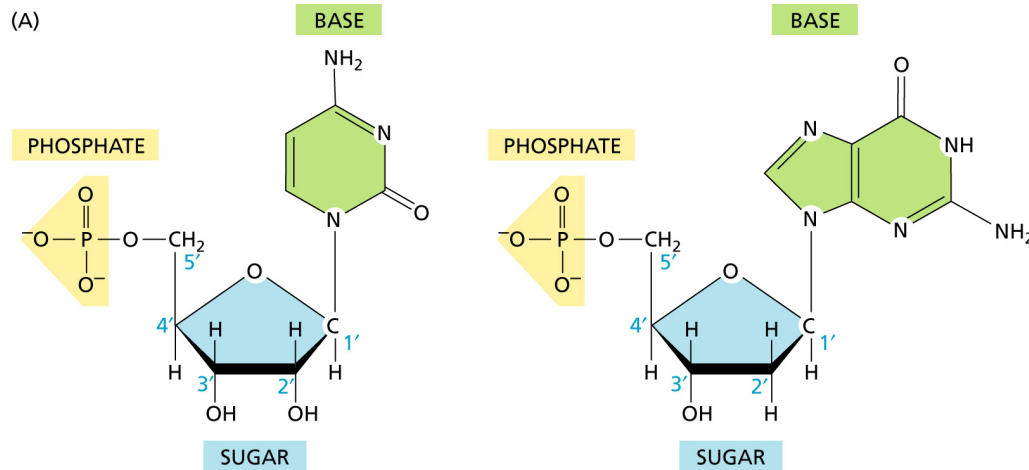
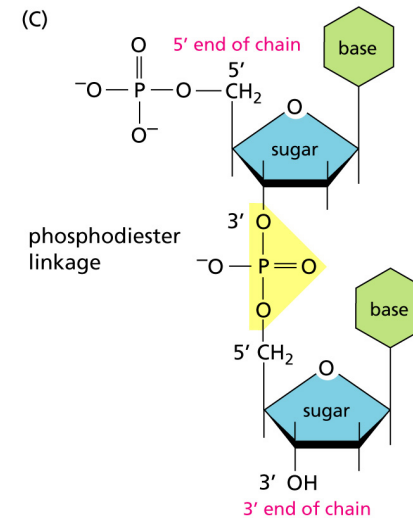
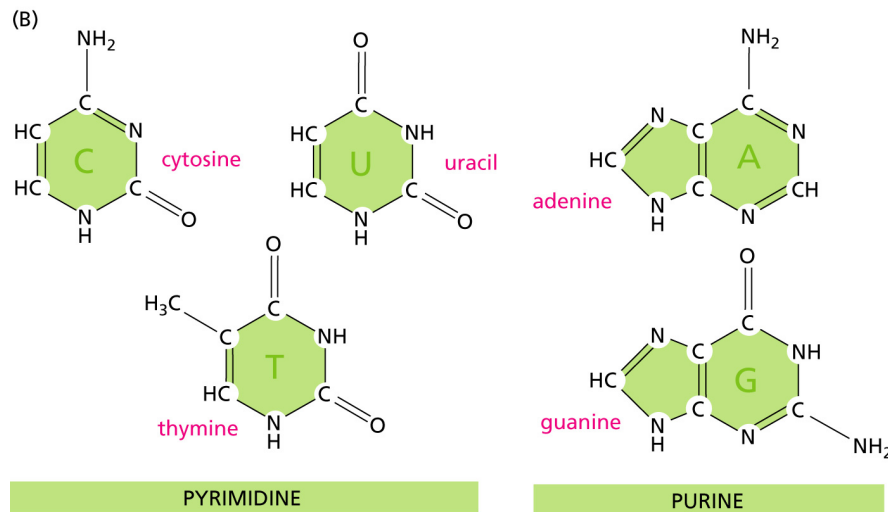
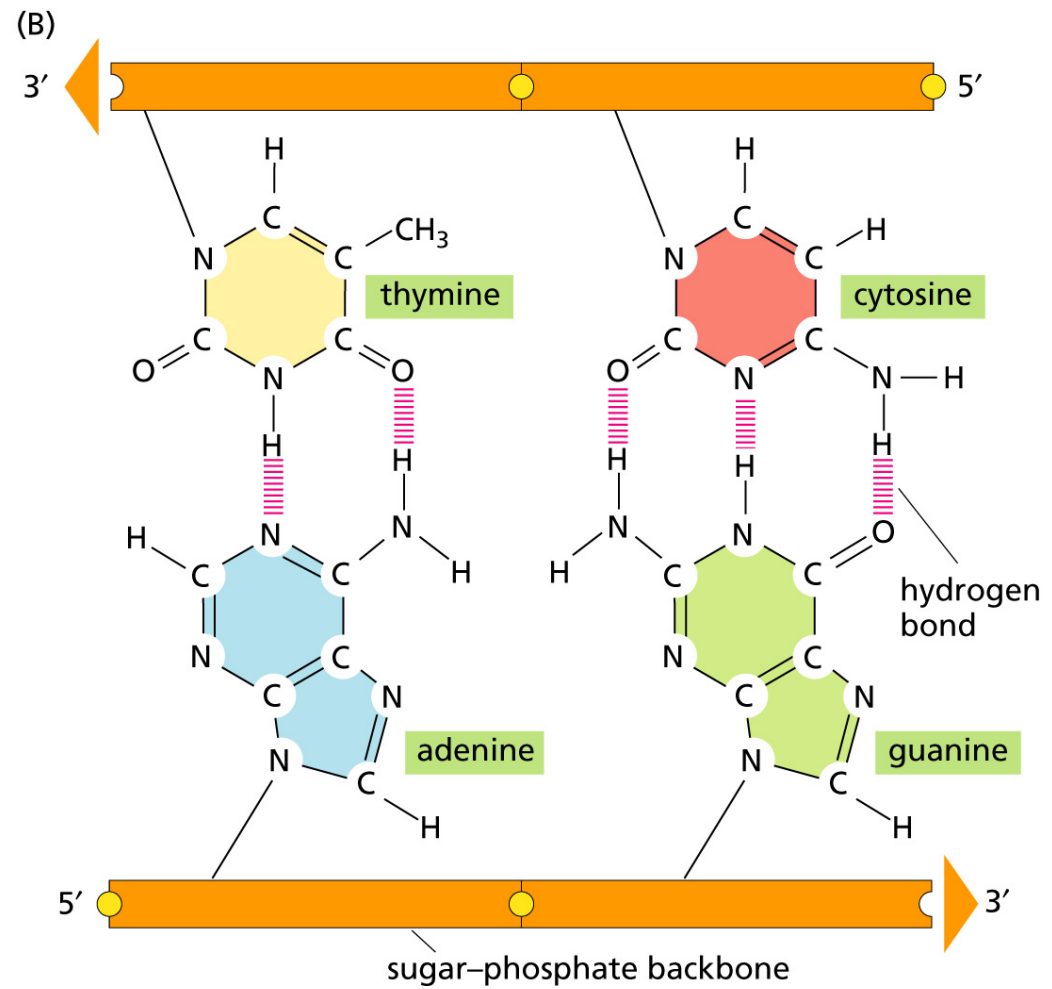
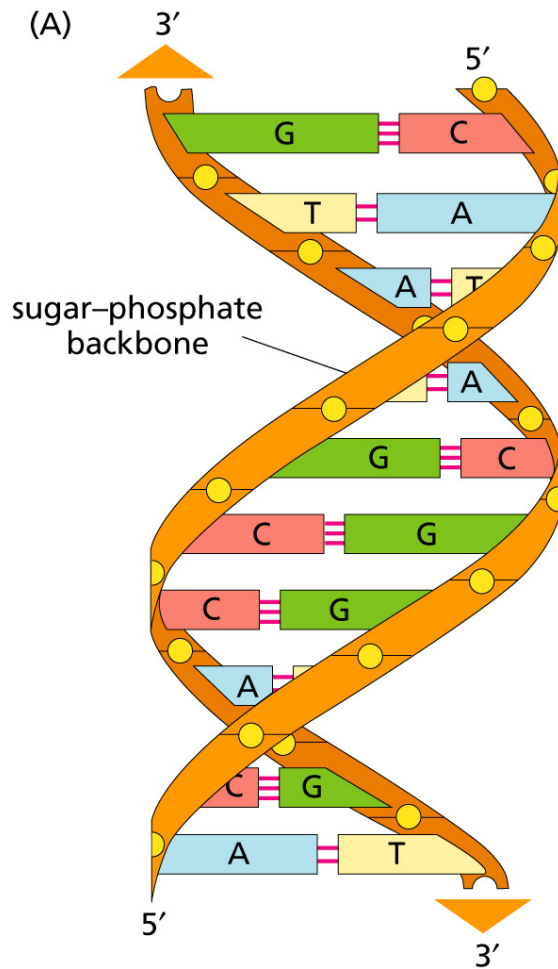


Fig 1.1, Zvelebil/Baum



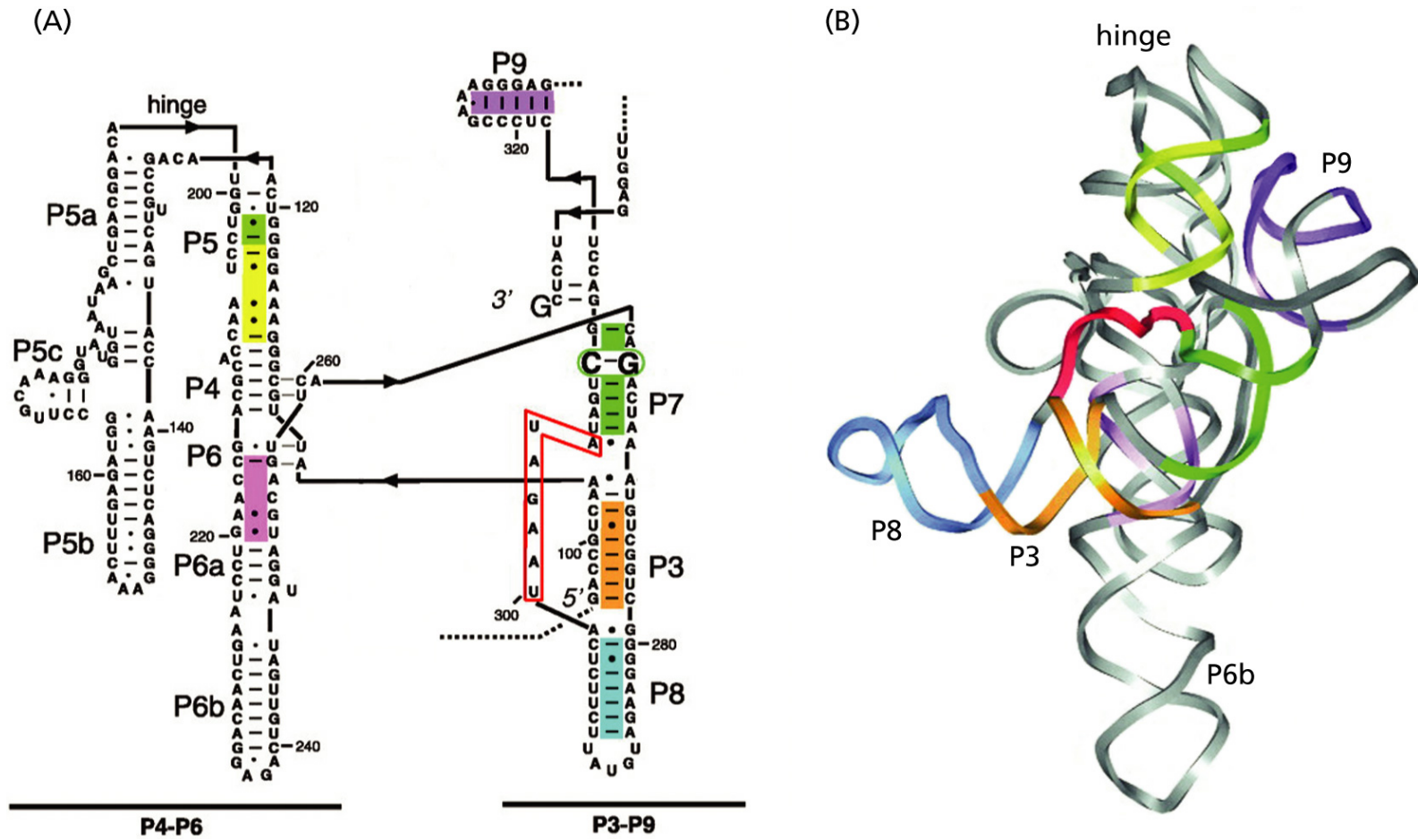
DNA double helix structure

Fig 1.3, Zvelebil/Baum



RNA molecule

Fig 1.5, Zvelebil/Baum



Proteins – Amino acids

amino acid	3 letter code	1 letter code
alanine	Ala	A
arginine	Arg	R
aspartic acid	Asp	D
asparagine	Asn	N
cysteine	Cys	C
glutamic acid	Glu	E
glutamine	Gln	Q
glycine	Gly	G
histine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
methionine	Met	M
phenylalanine	Phe	F
proline	Pro	P
serine	Ser	S
threonine	Thr	T
tryptophan	Trp	W
tyrosine	Tyr	Y
valine	Val	V

Table 1.1: *Amino acid abbreviations*

Typical protein sequence

```
/translation="MVAGTRCLLVLLLPQVLLGGAAGLIPELGRKKFAAASSRPLSRP  
SEDLSEFELRLLSMFGLKQRPTPSKDVVPPYMLDLYRRHSGQPGAPAPDHRLERAA  
SRANTVRSFHHEEAVEELPEMSGKTARRFFNLSSVPSDEFLLTSAELQIFREIQEAL  
GNSSFQHRINIYEI IKPAAANLKFVTRLLDTRLVNQNTSQWESFDVTPAVMRWTTQG  
HTNHGFVVEVAHLEENPGVSKRHVRI SRSLHQDEHSWSQIRPLLVTFGHDGKGHPLHK  
REKRQAKHKQRKRLKSSCKRHPLYVDFSDVGWNDWIVAPPGYHAFYCHGECPFPLADH  
LNSTNHAI VQTLVNSVNSKIPKACCVPTELSAISMLYLDENEKVVLKKNYQDMVVEGCG  
CR"
```

Missing letters of the alphabet: B, J, O, U, X, Z

Protein 3D Structure

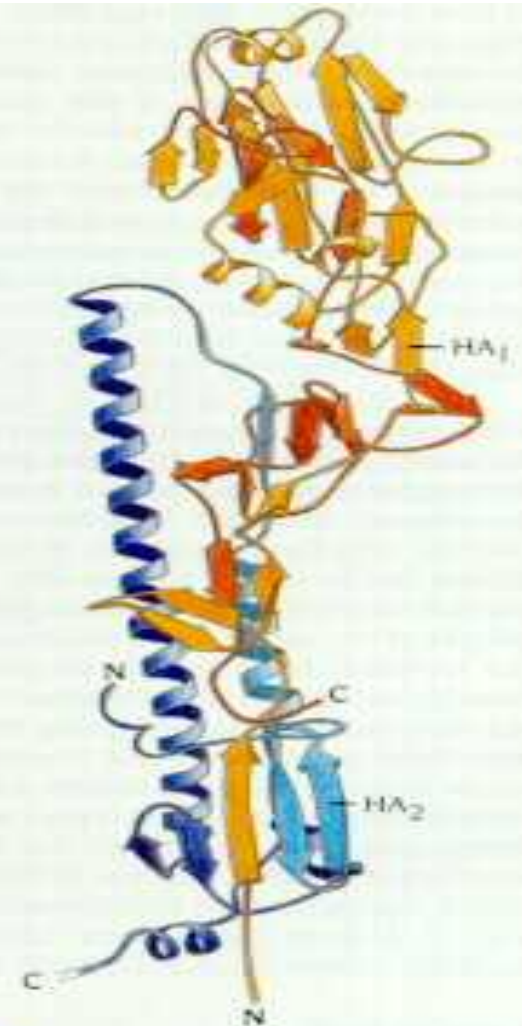
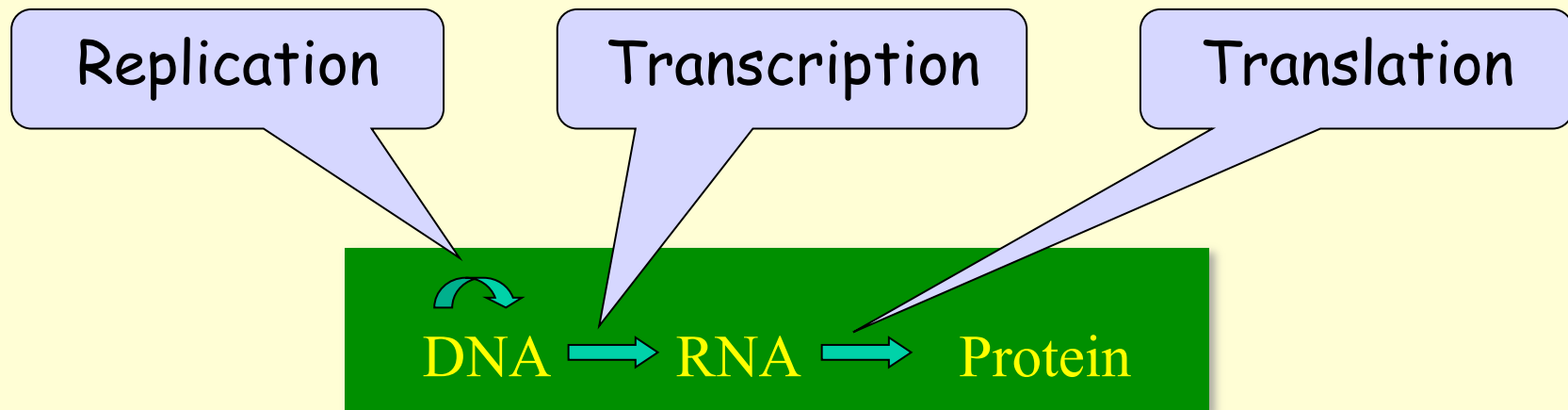


Figure S.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA₁ (red) and HA₂ (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75 Å long. The globular head is formed by residues only from HA₁. (Courtesy of Don Wiley, Harvard University.)

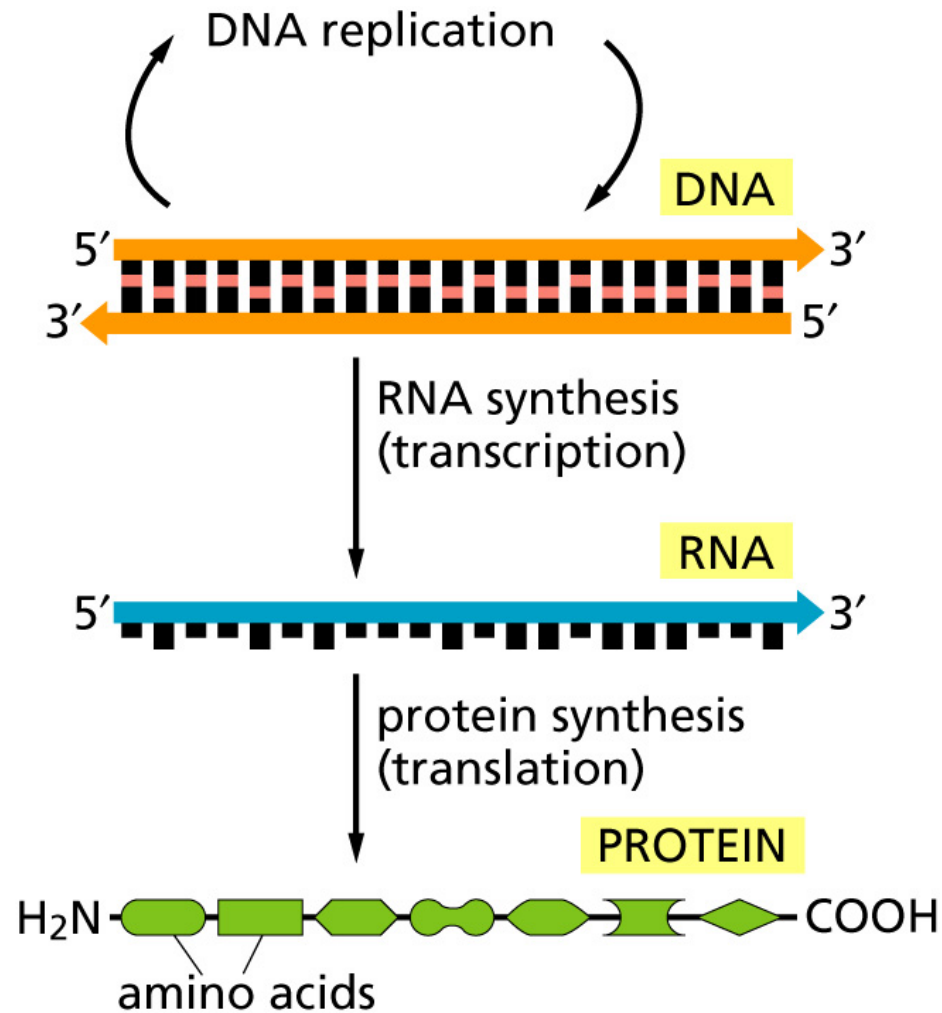
Central Dogma

- ❑ DNA acts as a template to replicate itself.
- ❑ DNA is transcribed into RNA.
- ❑ RNA is translated into **Protein**.



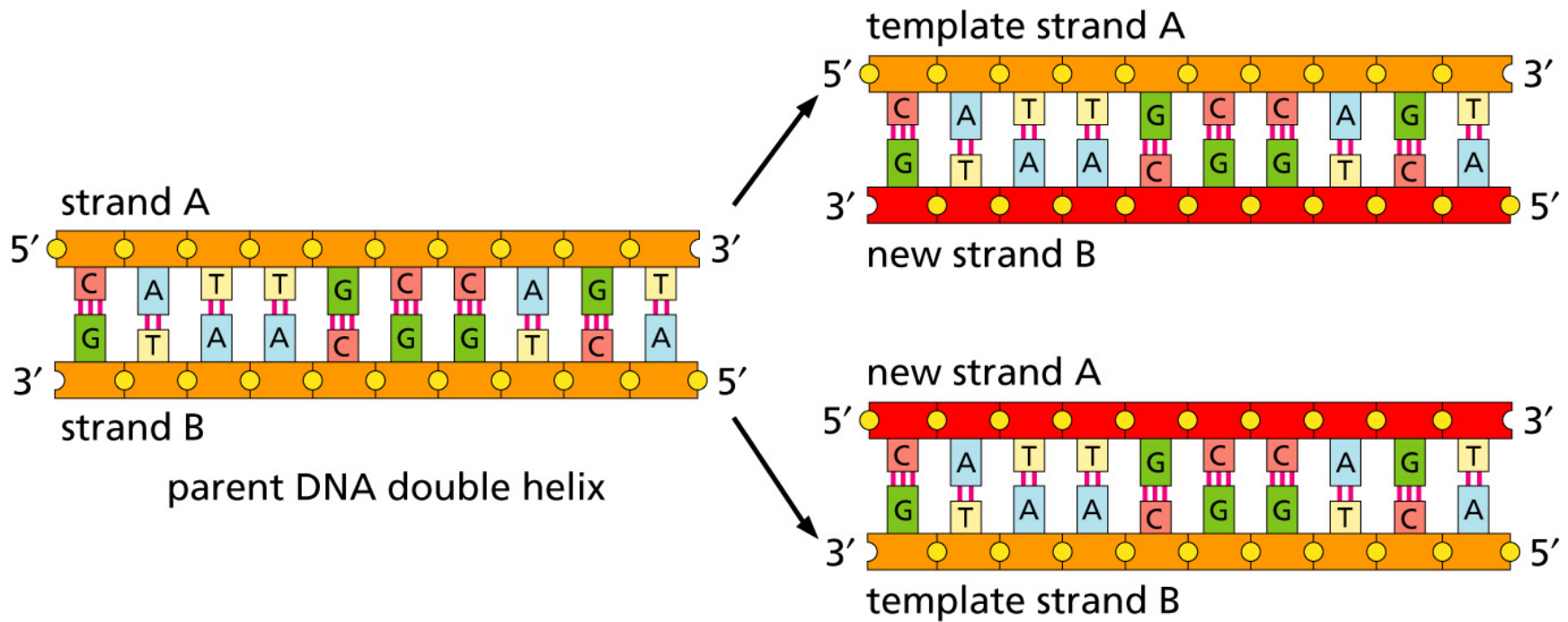
Central Dogma

Fig 1.6, Zvelebil/Baum

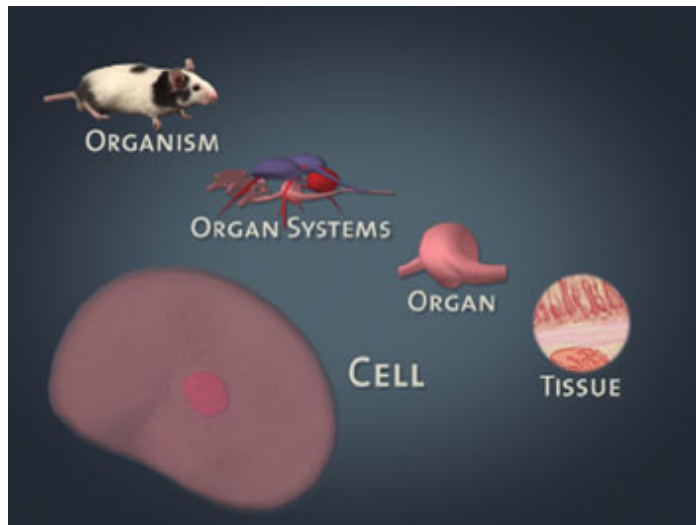


DNA Replication

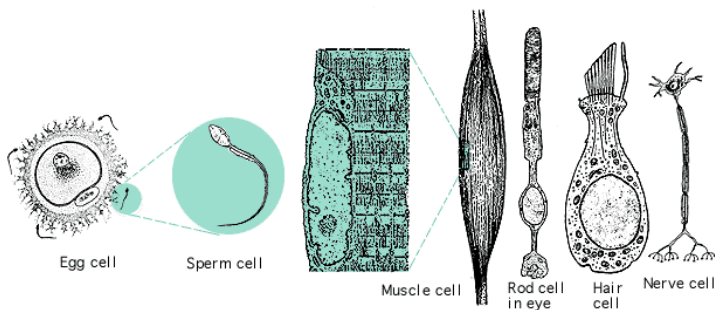
Fig 1.4, Zvelebil/Baum



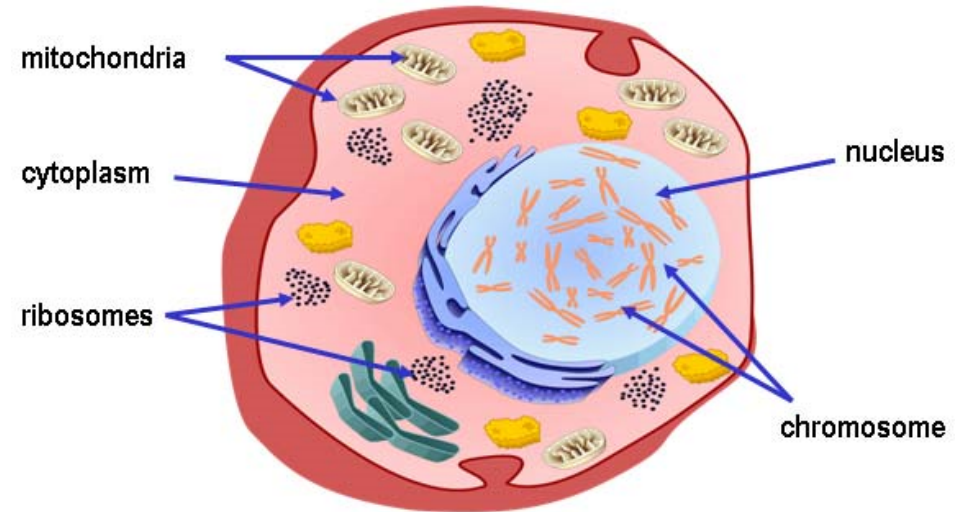
Cell



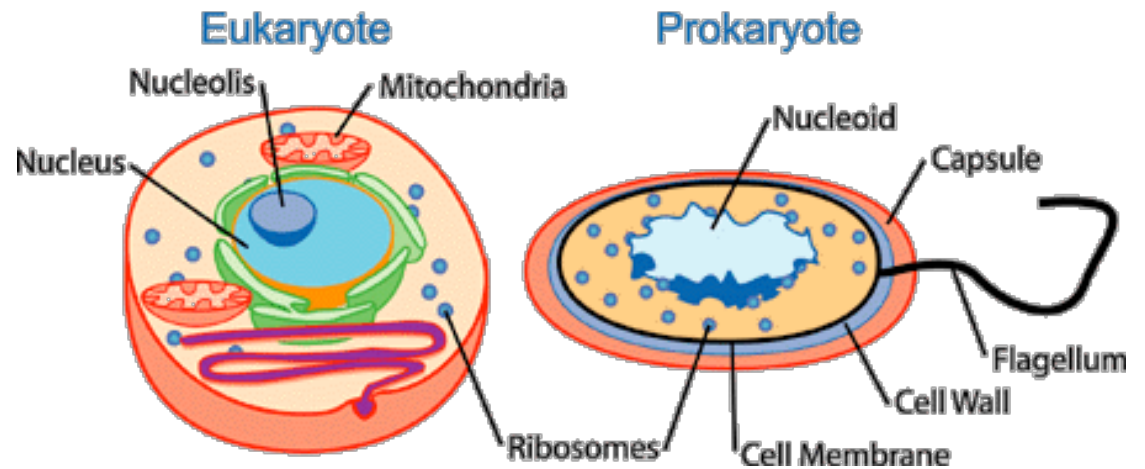
<http://www.learner.org/channel/courses/essential/life/session1/closer1.html>



<http://www.biology.eku.edu/RITCHISO/301notes1.htm>



http://www.biotechnologyonline.gov.au/popups/img_cellwithlabels.cfm

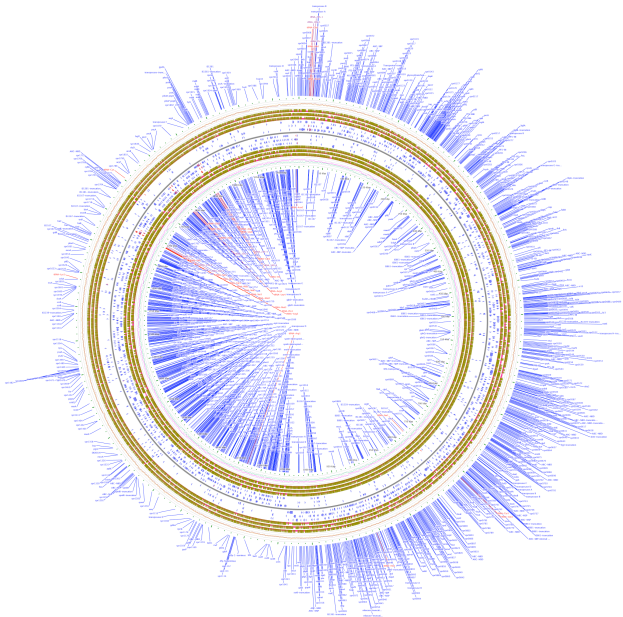


<http://en.wikipedia.org/wiki/File:Celltypes.png>

Chromosomes

Accession: NC_003098
2,038,615 bp

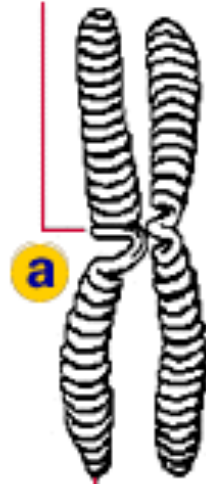
- Protein coding
- rRNA
- tRNA
- GC content
- AT content
- GC skew
- AT skew
- Start codon
- Stop codon



Streptococcus pneumoniae R6 complete genome

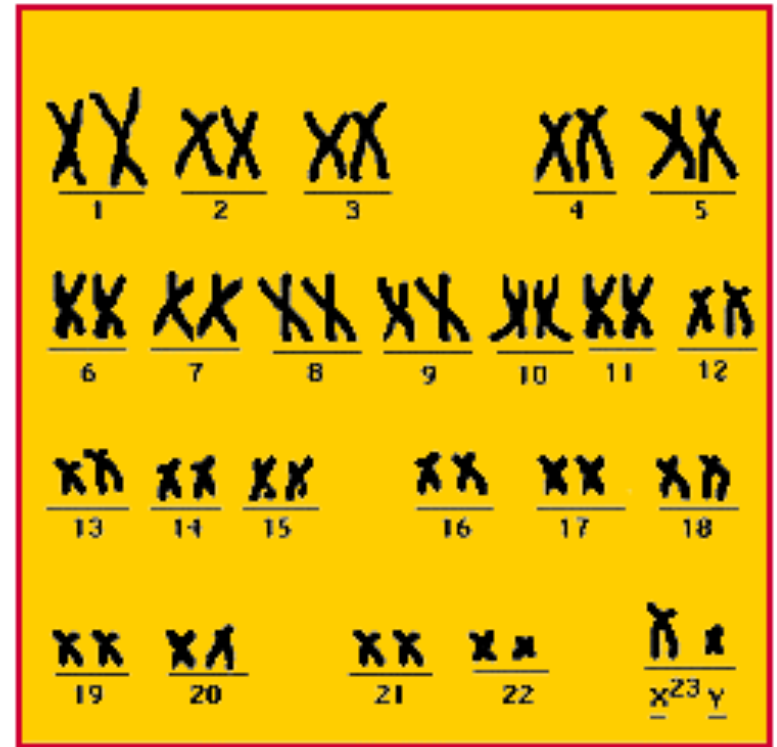
Human chromosomes!

centromere



a

chromatid

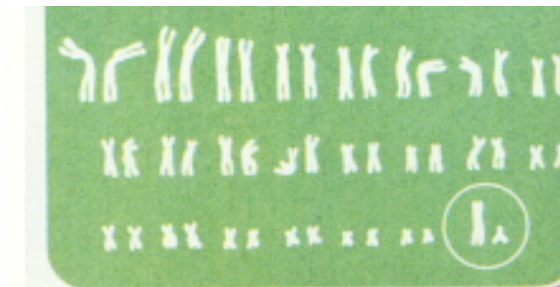
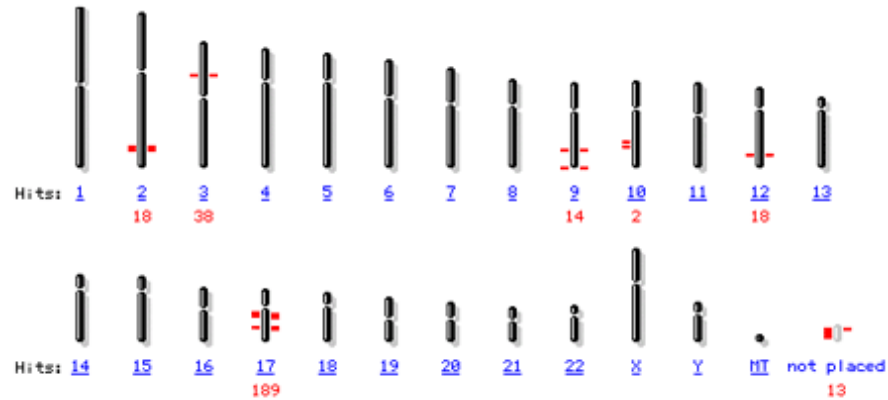


b

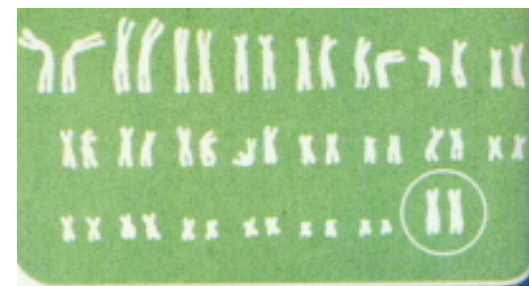
Chromosomes

Homo sapiens (human) genome view BLAST search the human genome

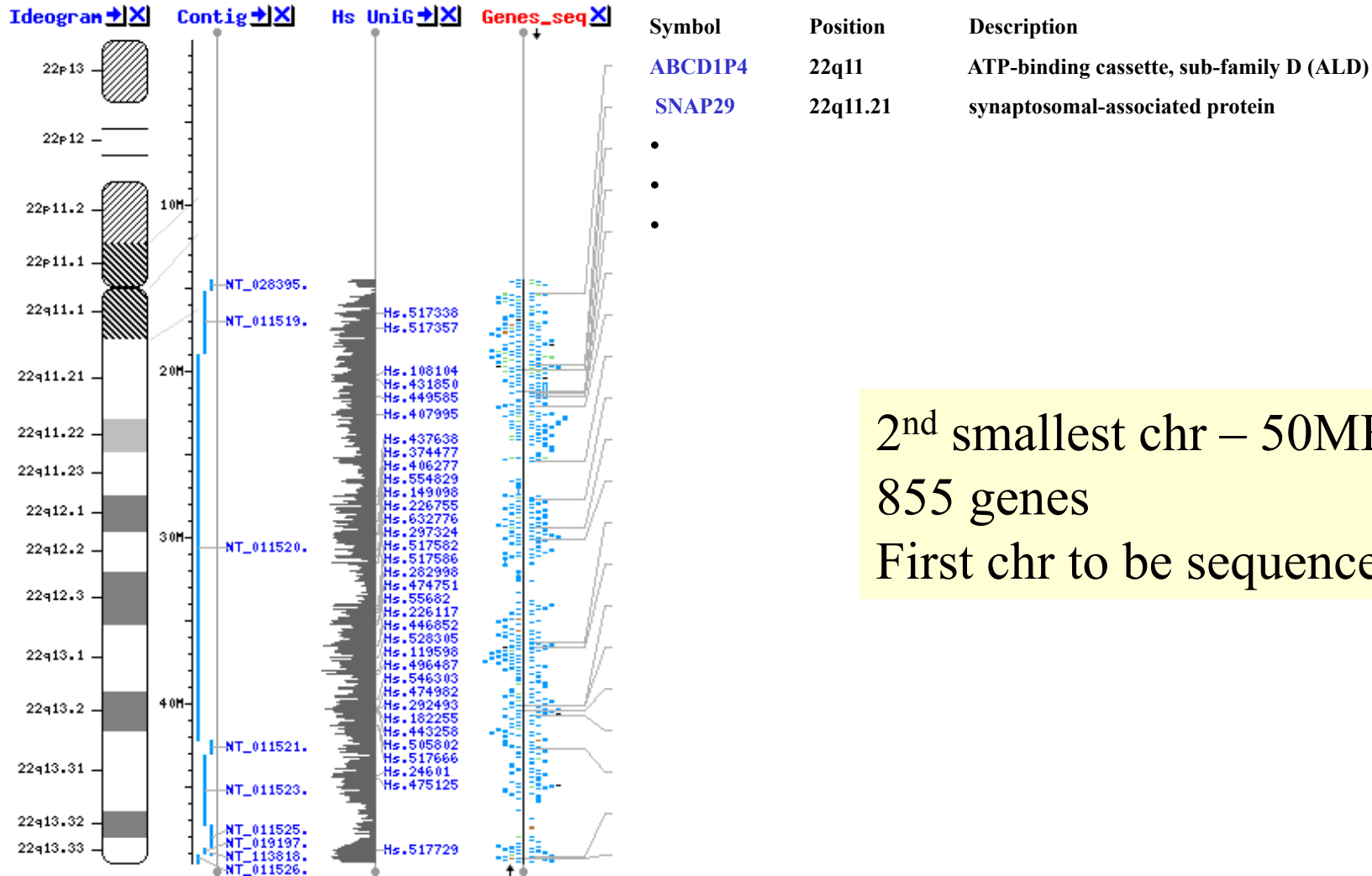
Build 36.2 statistics [Switch to previous build](#)



The chromosomal locations of several genes believed to be associated with the human BRCA1 gene implicated in breast cancer are highlighted.

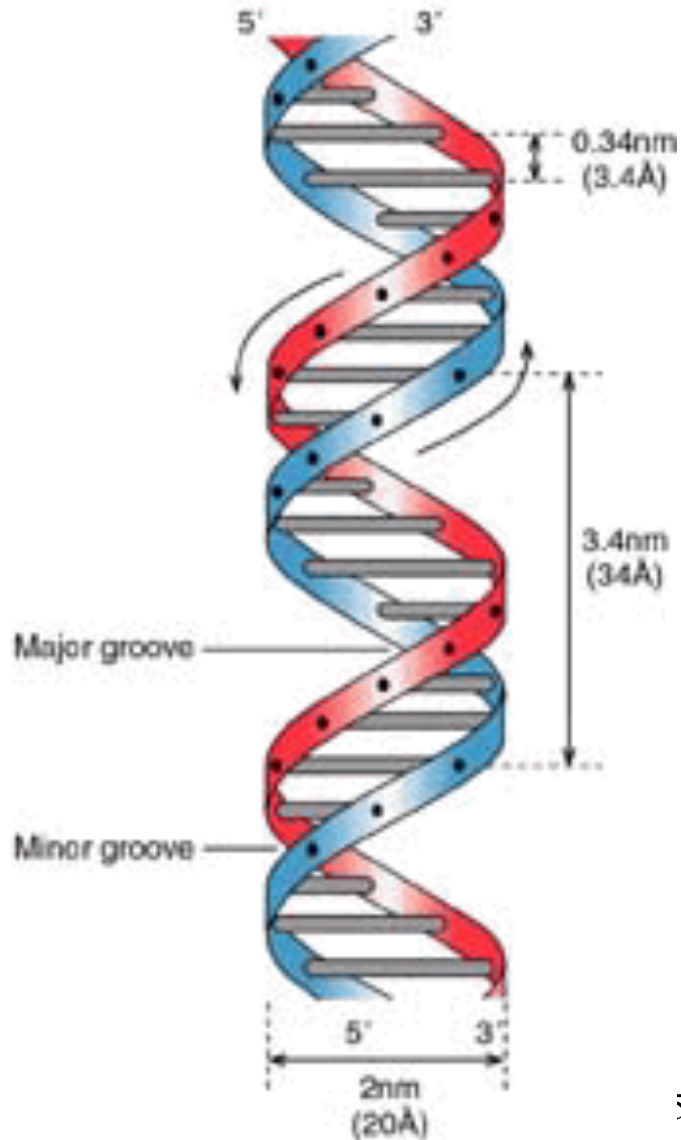


Human Chr 22

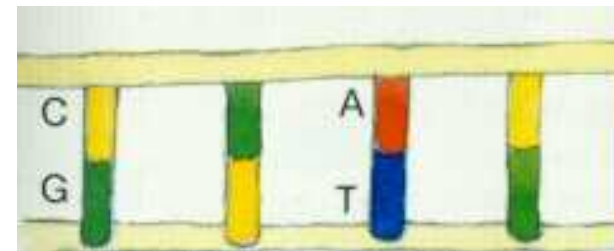


2nd smallest chr – 50MB
 855 genes
 First chr to be sequenced (1999)

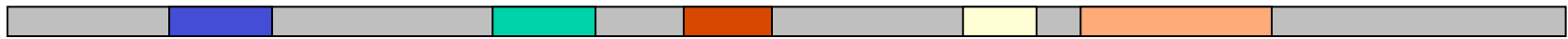
DNA Molecule



Complementary Bases



Genes



Basic Genetic Processes

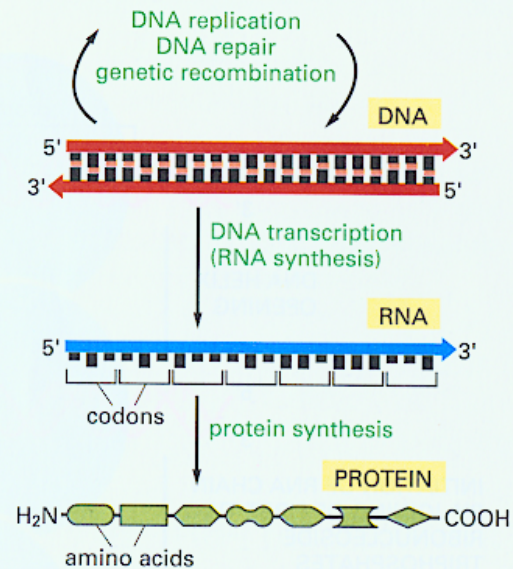
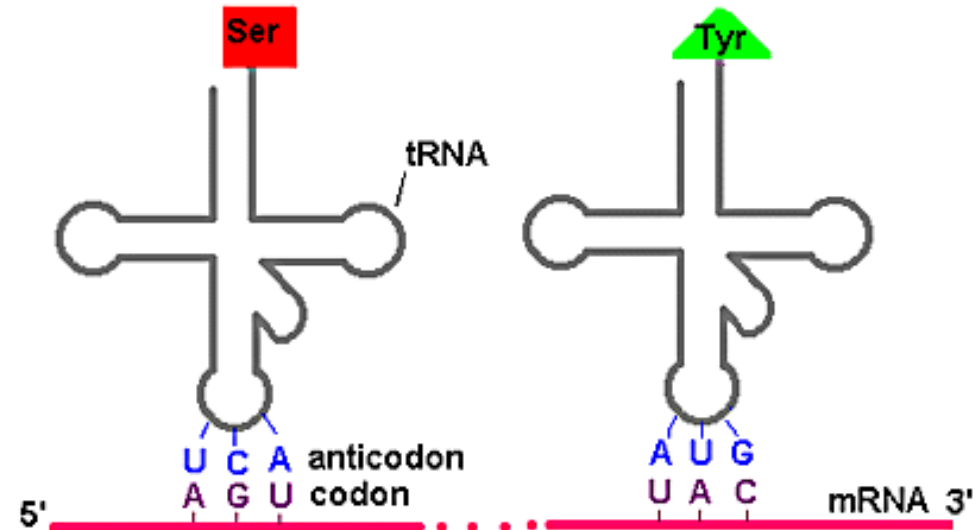


Figure 6-1 The basic genetic processes. The processes shown here are thought to occur in all present-day cells. Very early in the evolution of life, however, much simpler cells probably existed that lacked both DNA and proteins (see Figure 1-11). Note that a sequence of three nucleotides (a codon) in an RNA molecule codes for a specific amino acid in a protein.

The Genetic Code



		2nd base in codon				
		U	C	A	G	
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G
						3rd base in codon

The Genetic Code

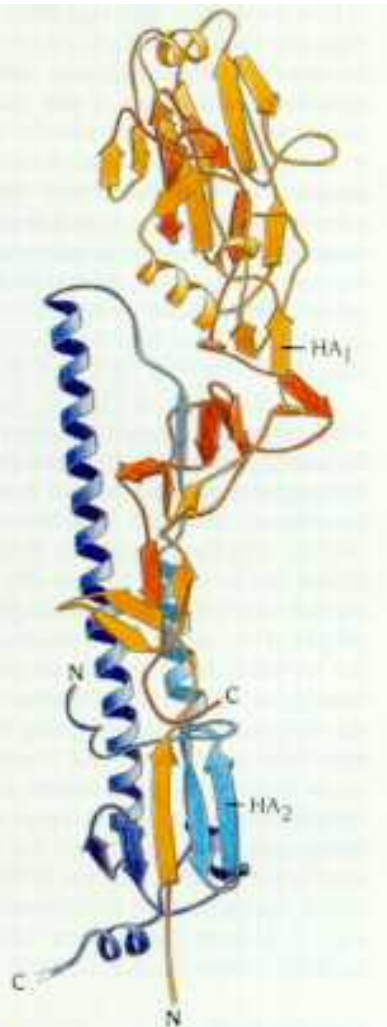
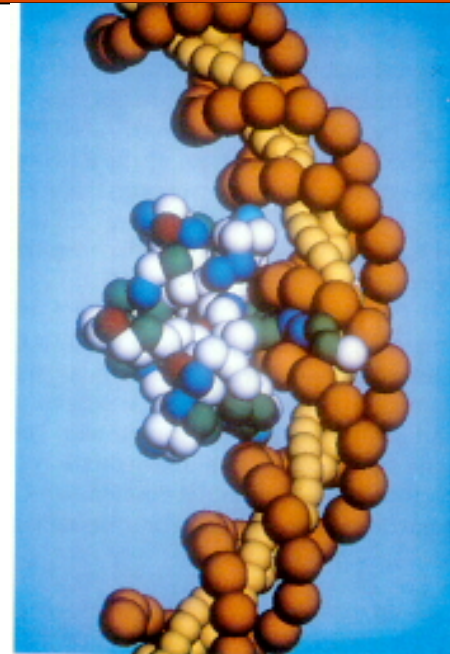
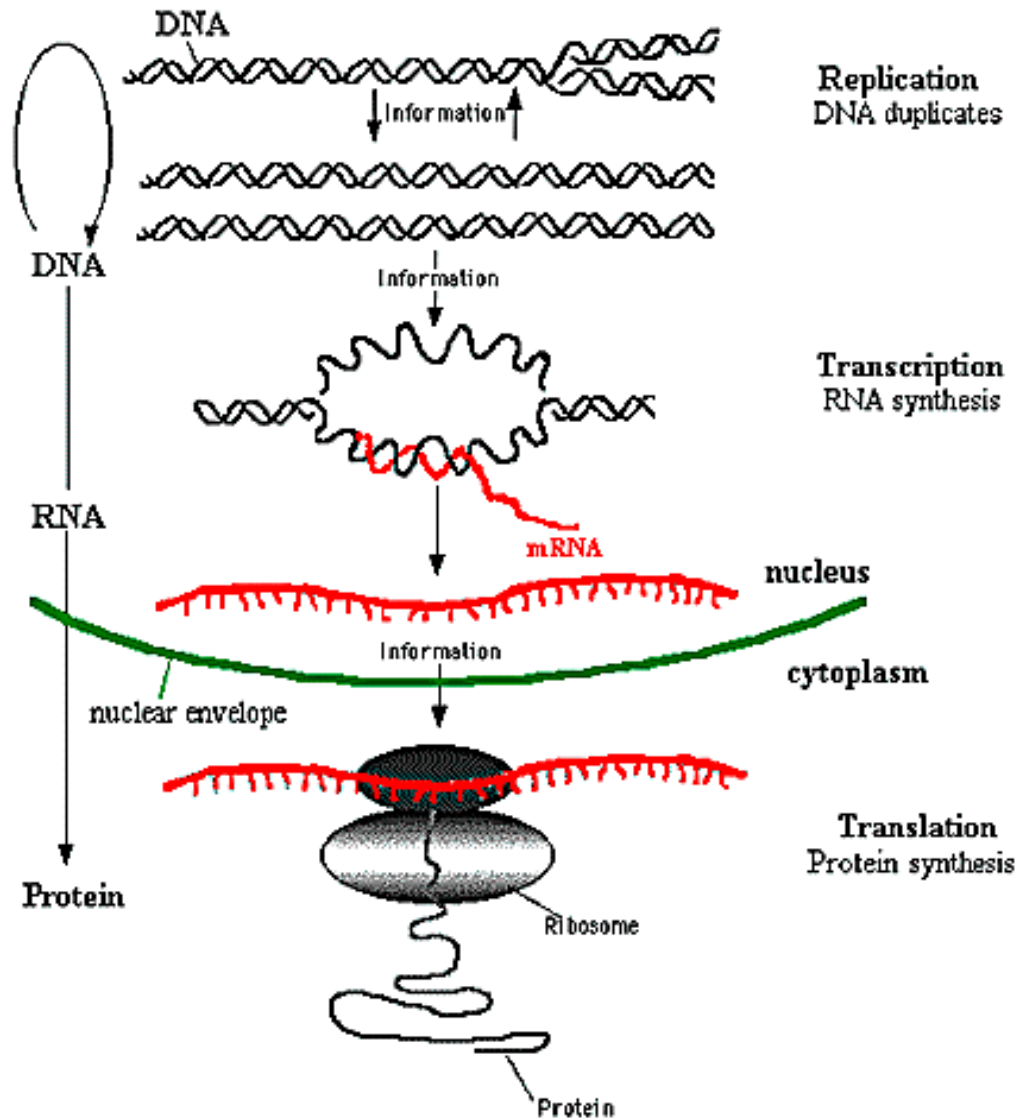


Figure S.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA₁ (red) and HA₂ (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75 Å long. The globular head is formed by residues only from HA₁. (Courtesy of Don Wiley, Harvard University.)

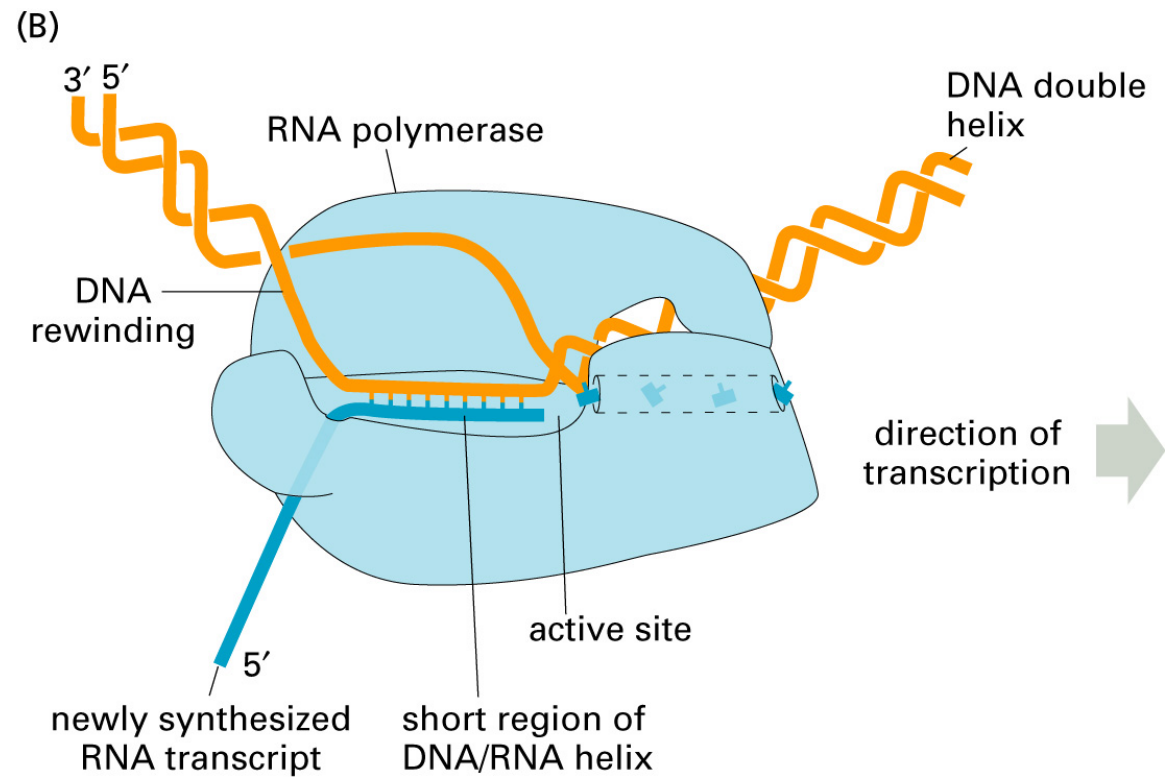
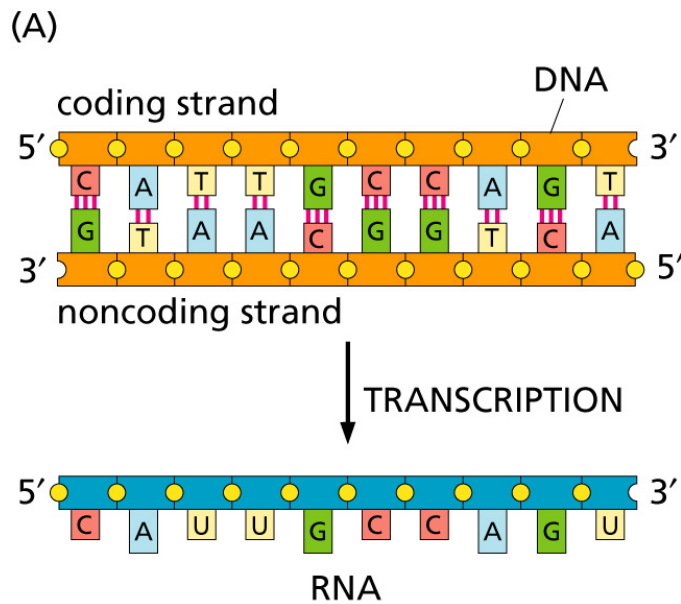


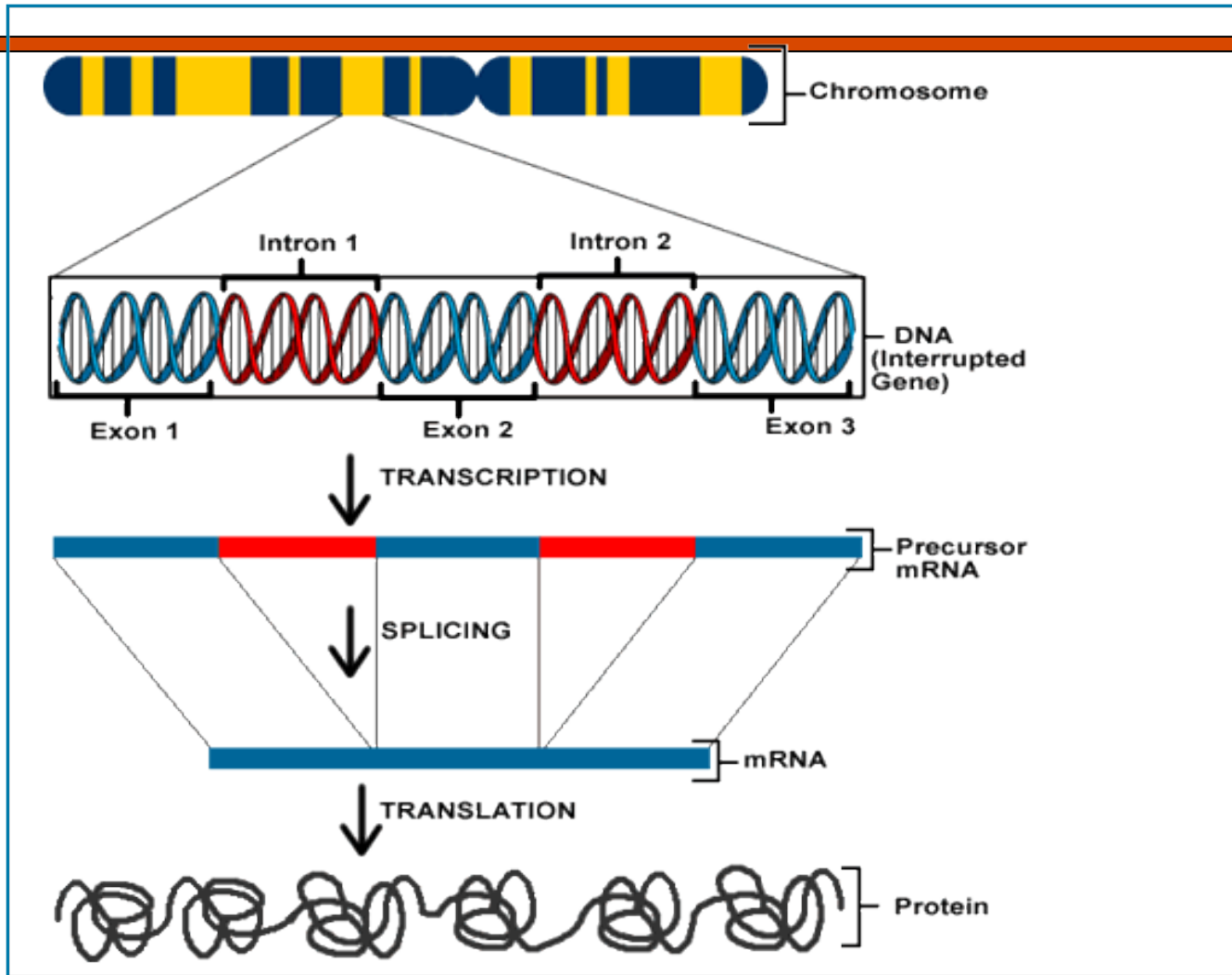


The Central Dogma of Molecular Biology

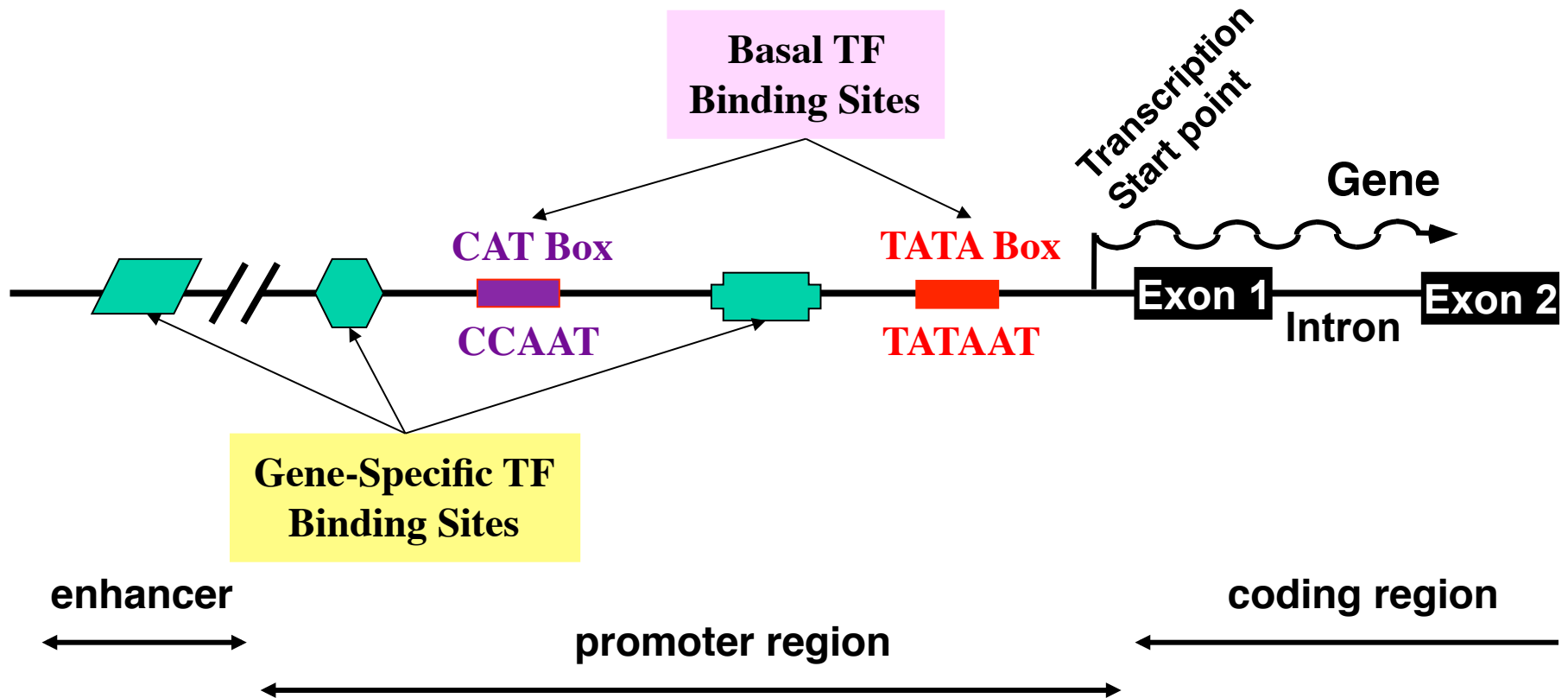
Transcription

Fig 1.7, Zvelebil/Baum

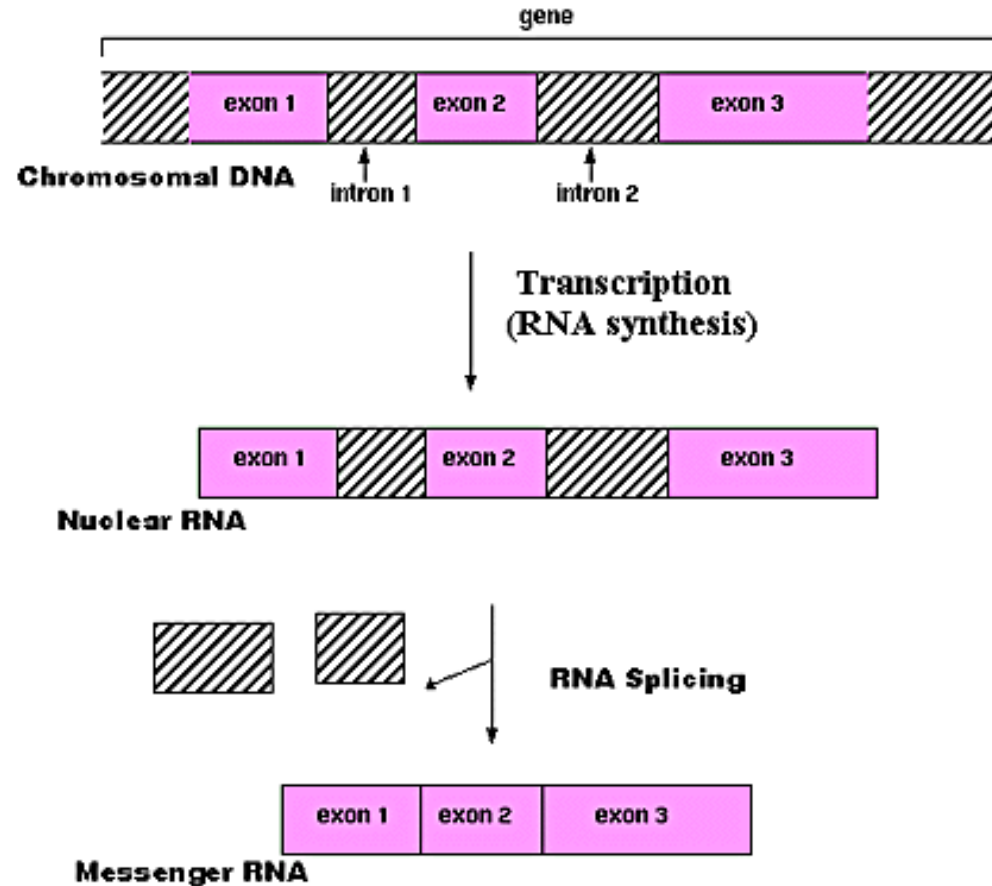




Transcription Regulation

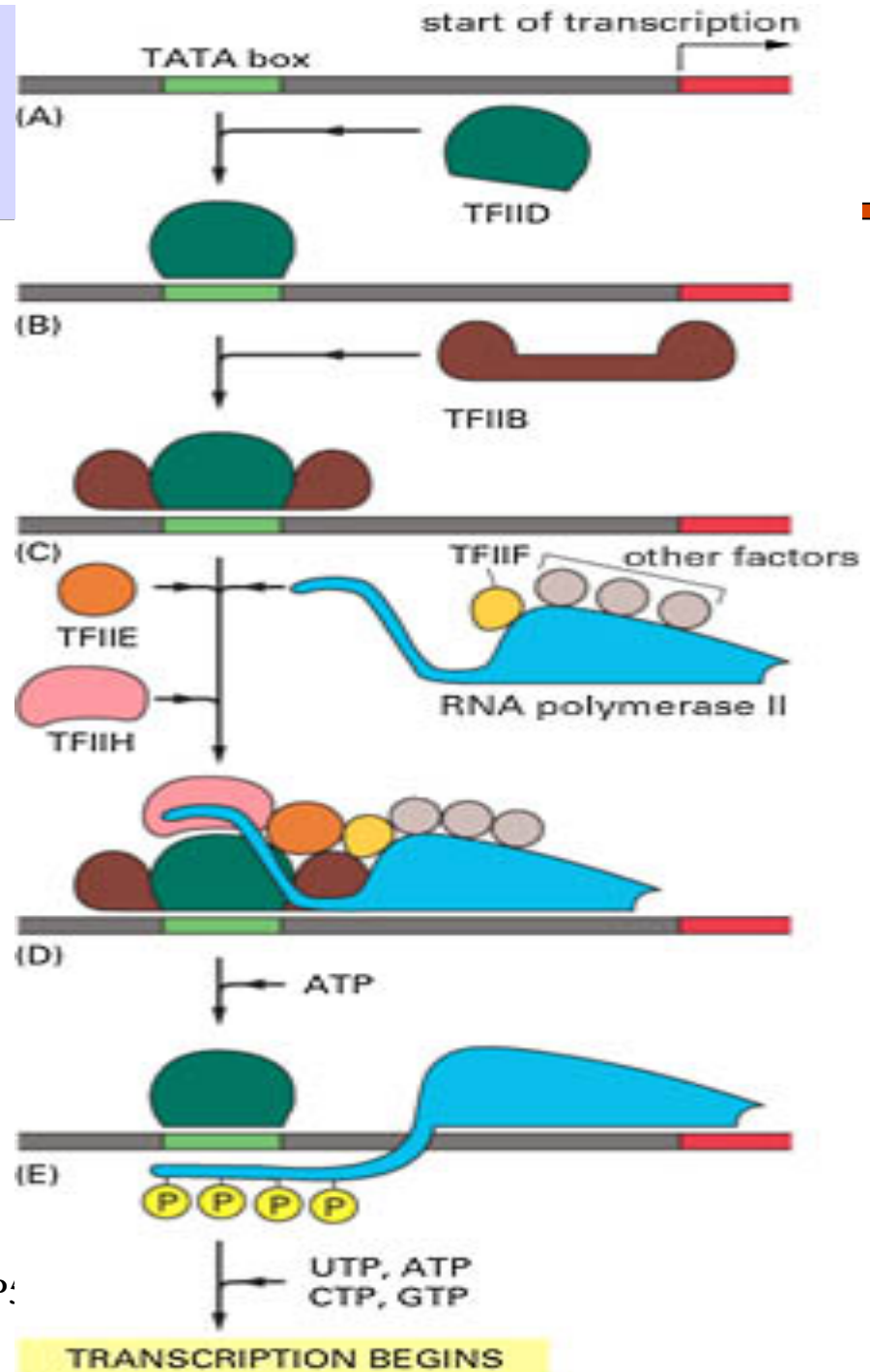


DNA Transcription



RNA synthesis and processing

Transcription Initiation



Transcription

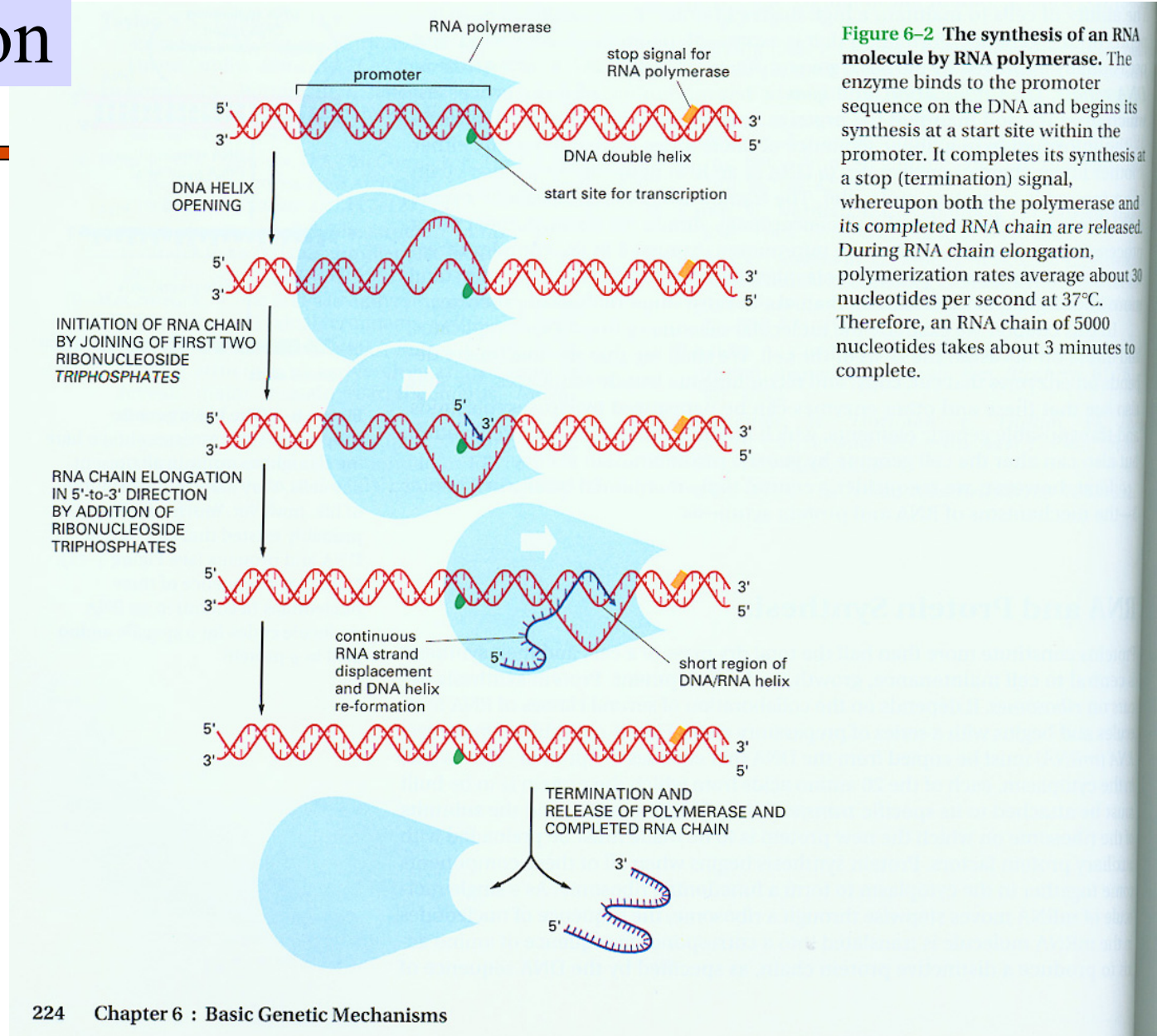
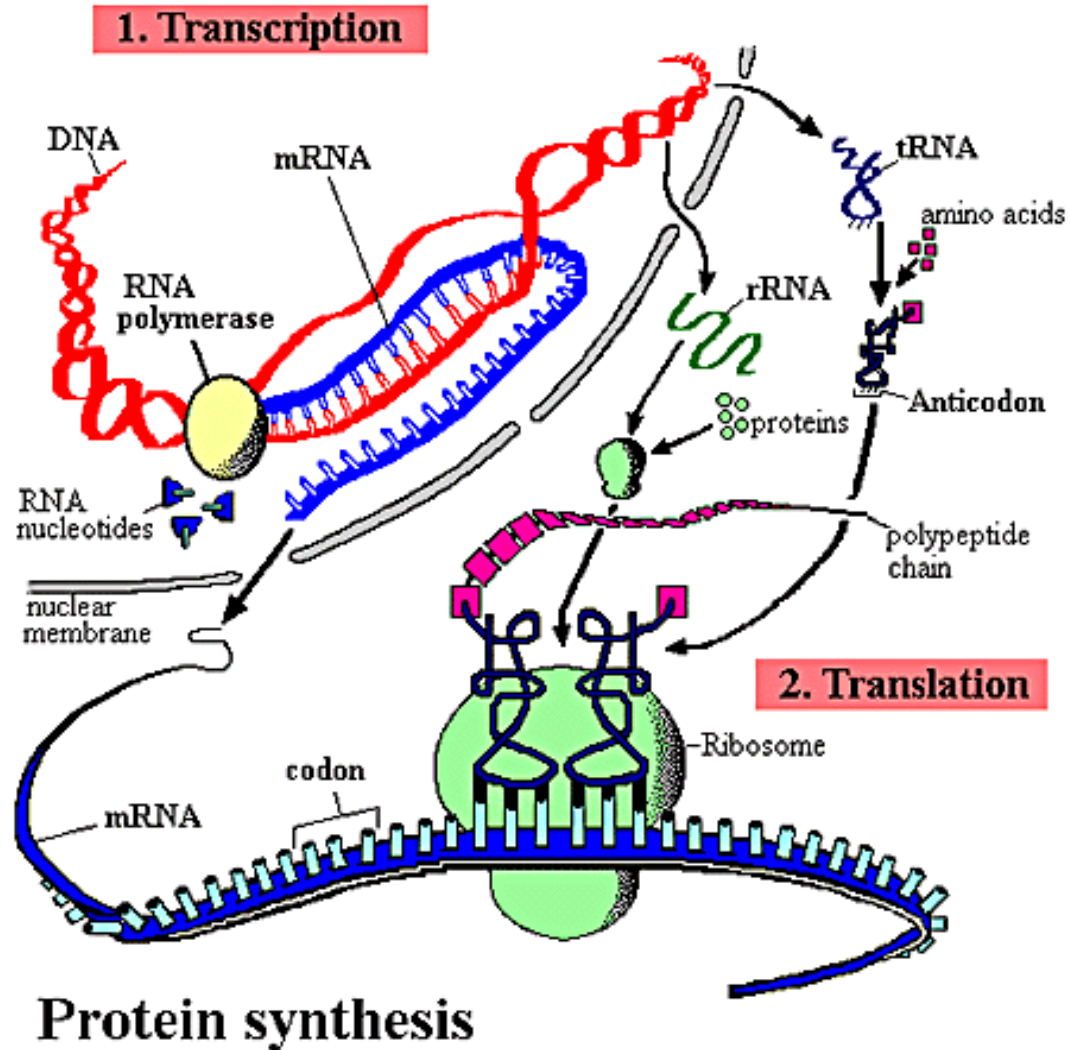


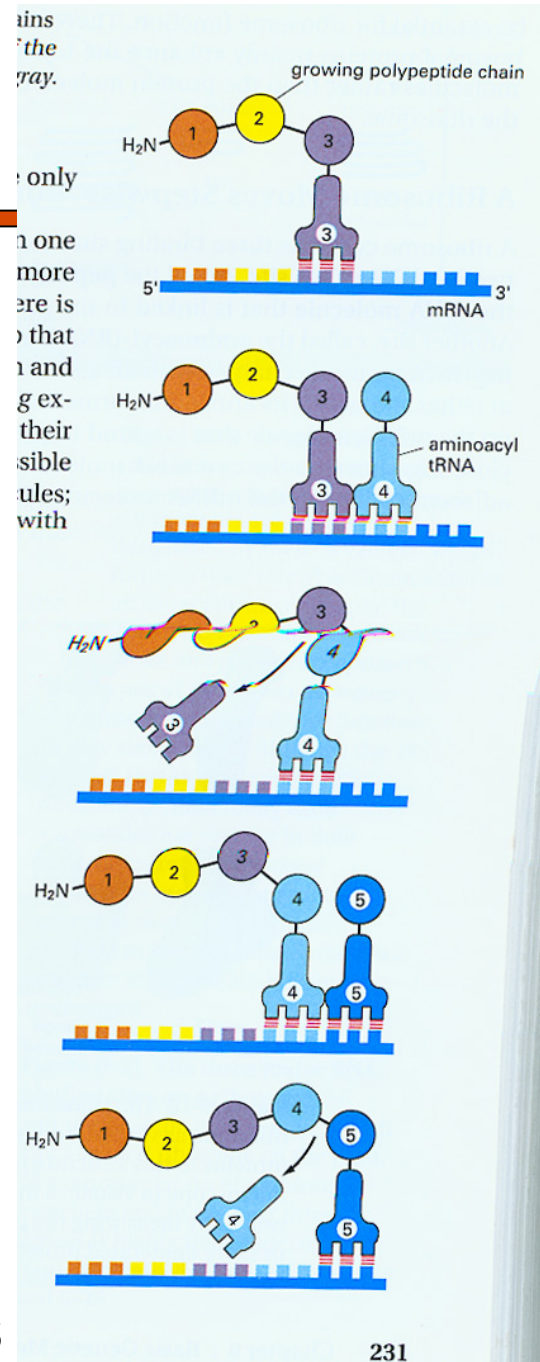
Figure 6-2 The synthesis of an RNA molecule by RNA polymerase. The enzyme binds to the promoter sequence on the DNA and begins its synthesis at a start site within the promoter. It completes its synthesis at a stop (termination) signal, whereupon both the polymerase and its completed RNA chain are released. During RNA chain elongation, polymerization rates average about 30 nucleotides per second at 37°C. Therefore, an RNA chain of 5000 nucleotides takes about 3 minutes to complete.

Protein Synthesis



Protein Synthesis:

Incorporation of amino acid into protein



Three major public DNA databases

□ GenBank

- NCBI (Natl Center for Biotechnology Information) www.ncbi.nlm.nih.gov

□ EMBL

- EBI (European Bioinformatics Inst)

□ DDBJ

- Japan's center

Entrez Portal @ NCBI

- PubMed
- DNA and Protein Sequence database
- Protein structure database
- Population study data sets
- Genome assemblies
- BLAST
- OMIM (Mendelian Inheritance in Man)
- TaxBrowser