CAP 5991 (3 Credits)
Introduction to Bioinformatics

CGS 5991 (2 Credits)
Bioinformatics Tools

Giri Narasimhan

Course Schedules

- CAP 5991 (3 credit) will meet every Tue from 11AM to 1:45PM.
- CGS 5991 (2 credit) will meet every Tue from 11AM to about 1PM. This course is not for CS students.
- Different exams and evaluation.

CAP/CGS 5991
Introduction to Bioinformatics

Overview of Course

- Preliminaries
- Sequence Alignment
- Multiple Sequence Alignment
- Phylogenetic Analysis
- Molecular Structure Analysis
- Gene Recognition
- Genomics, Functional Genomics
- Proteomics
- Pattern Discovery Techniques
- Programming Environments: BioPerl
- Databases and Software Packages
- Statistics for Bioinformatics
- Sequencing and Mapping
- Computational Learning Methods
  HMM, NN, SOM, SVM, GA
- Computational Predictive Methods
- Microarray Data Analysis
- Digital Image Analysis
- Protein Structure Analysis: SPDBV
- Emerging Biotechnologies

CAP/CGS 5991

Software Packages

- Databases and Software Packages (GenBank, SWISS-PROT)
- Programming Environments (BioPerl)
- Sequence Alignment & Multiple Sequence Alignment (BLAST, CLUSTALW, CLUSTALX)
- Phylogenetic Analysis (CLUSTALW, Phylip, PAUP, PAML)
- Learning Methods (HMMPro, GeneCluster, ASOM)
- Pattern Discovery Techniques (GYM, TEIRESIAS, APRIORI)
- Molecular Structure Analysis (DALI, RASMOL, SPDBV)
- Microarray Analysis (CLUSTER, SAM, GeneCluster, TreeView)
- Statistical Software Packages (SAS, R)

Evaluation

- Semester Project (50 %)
- Homework Assignments (20 %)
- Exams (20 %)
- Class Participation (10 %)

Reading List and Schedule

Watch that Course Home Page!

http://www.cs.fiu.edu/~giri/teach/5991F03.html

Introduction

• 1. What is Bioinformatics?

- Analysis of biological data with computing & statistical tools.

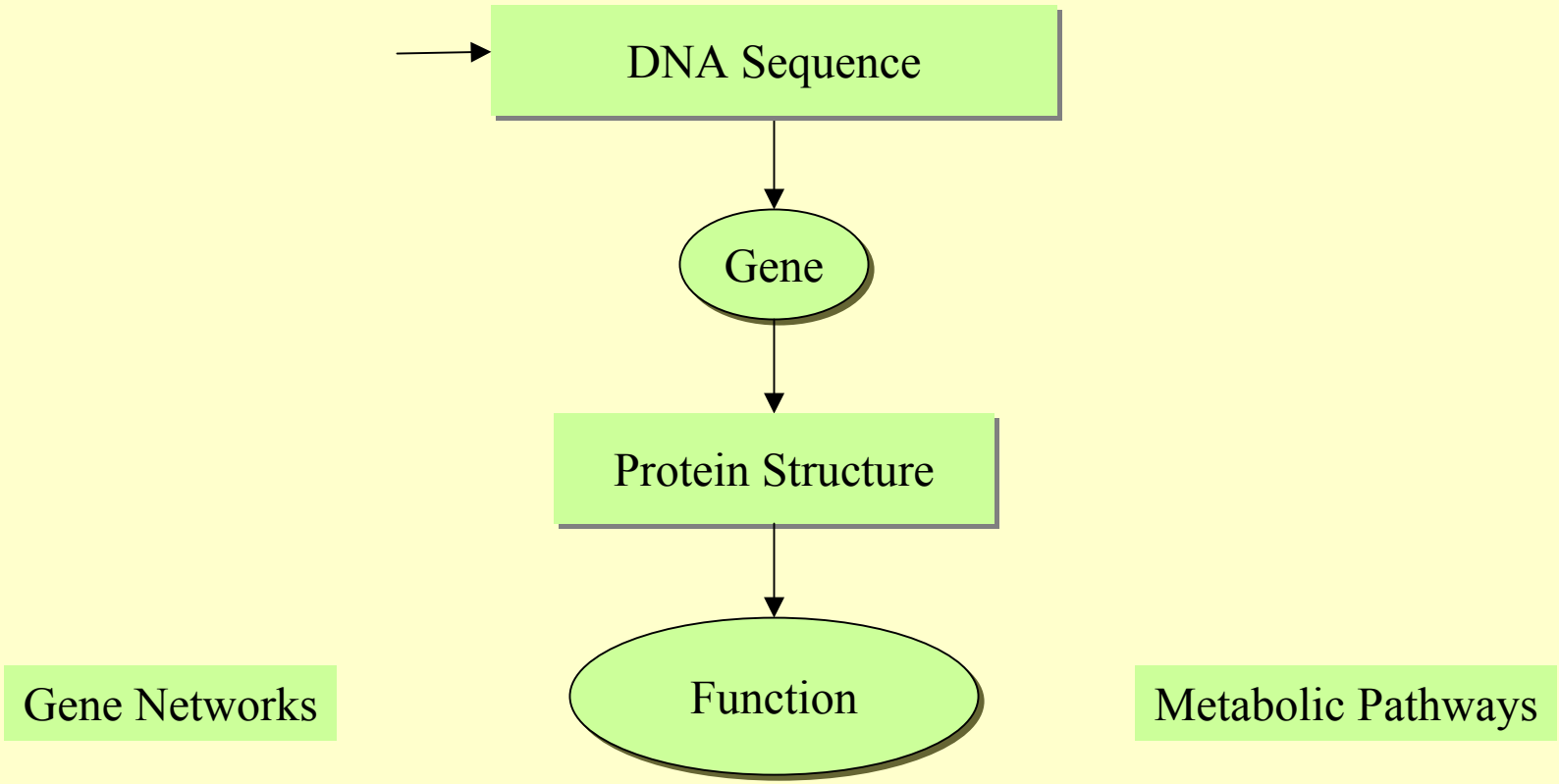
• 2. The different aspects of Informatics:

- Data Management (Database Technology, Internet Programming)
- Analysis/Interpretation of Data (Data Mining, Modeling, Statistical Tools)
- Development of Algorithms/ Data Structures
- Visualization and Interface Design (HCI, Graphics)

• 3. How to assist biological research?

- propose new models or correlations based on data from experiments
- verify a proposed model using known data
- propose new experiments based on model or analysis
- use predicted information to narrow down search in a biological investigation

Overall Goals



General Information

- **Over 20 billion bases in the NCBI database:**

  http://www.ncbi.nlm.nih.gov

- **Human Genome has ~3 billion bp with 30,000+ genes.**
- **Viruses have 300bp to 300Kb (1st one in 1978: Simian virus; 5Kb).**
- **86 complete microbial genomes sequenced.**
- **Number of whole genomes sequenced is over 100 (not including over 1000 viruses), including:**

  Caenorhabditis elegans, Arabidopsis thaliana, Drosophila melanogaster, Saccharomyces cerevisiae,

- **Chromosomal maps for many organisms including:**

  Mus musculus, Homo sapiens, Danio rerio, Zea mays, Oryza sativa

- **Swiss-Prot has over 132000 protein sequences.**

Genome Sizes

Organism	Size	Date	Est. # genes
HIV type 1	10 Kb		
H. influenzae	1.8 Mb	1995	1,740
E. coli	4.7 Mb	1997	4,000
S. cerevisiae	12.1 Mb	1996	6,034
C. elegans	97 Mb	1998	19,099
A. thaliana	100 Mb	2000	25,000
D. melanogaster	180 Mb	2000	13,061
M. musculus	3 Gb	2002	~30,000
H. sapiens	3 Gb	2001	30,000+

Caenorhabditis Elegans

- Entire genome - 1998
- 1st animal; 26th organism
- 8 year effort
- 97 million bases
- 19,099 genes
- 402 gene clusters
- 12,000 genes with known function
- thousands of mutants
- 7000 families of repeats
- Multicellular organism
- Nematode (phylum)
- Easy to experiment with
- Easily observable
- 959 cells
- 302 nerve cells
- 36% of proteins common w/ human

Homo sapiens

- 15 year effort, 3 billion bases, 100,000 gaps
- Variable density of:

  Genes, SNPs, CpG islands, recombination rates
- ~1.1% of the genome codes for proteins
- ~ 40-48 % of the genome consists of repeat sequences
- ~10 % of the genome consists of repeats called ALUs
- ~5 % of the genome consists of long repeats (>1 Kb)
- ~ 50 transposon-derived genes
- 223 genes common with bacteria that are missing from worm, fly or yeast.

(Approximate) String Matching

Input: Text **T** , Pattern **P**

Question(s):

- Does **P** occur in **T**?
- Find one occurrence of **P** in **T**.
- Find all occurrences of **P** in **T**.
- Count # of occurrences of **P** in **T**.
- Find longest substring of **P** in **T**.
- Find closest substring of **P** in **T**.
- Locate direct repeats of **P** in **T**.
- *Many More variants*

Applications:

- Is **P** already in the database **T**?
- Locate **P** in **T**.
- Can **P** be used as a primer for **T**?
- Is **P** homologous to anything in **T**?
- Has **P** been contaminated by **T**?
- Is *prefix*(**P**) = *suffix*(**T**)?
- Locate tandem repeats of **P** in **T**.

The Suffix Tree Data Structure

Borrelia burgdorferi:

- 1 million bases
- Shotgun Sequencing:
  - 4612 fragments
  - 2 million bases long totally
  - Using suffix trees - **15 min** for Fragment Assembly
  - Using Dynamic Programming - **10 days**

Repeats in DNA Sequences

Genomic Imprinting: Some genes are expressed only when inherited from one specific parent.

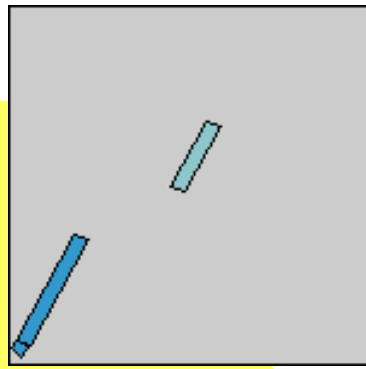
16 such genes are known; 5 inherited from mother; rest from father.

These 16 genes have a lot of **repeats**.

Repeats are of size 25 to 120 bp and of total length 1500.
The repeats are unique to these imprinted regions.
They have no obvious homolgy to each other
or to other highly repetitive mammalian sequences.

Repeats are also known to be responsible for several genetic diseases:
Fragile X, Huntington's disease, Kennedy's disease,
myotonic dystrophy, ataxia.

Drosophila Eyeless vs. Human Aniridia



```
24 IERLPSLEDMAHKGHSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSN 83
   I R P+   M +  HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSN
17 IPRPPARASMQNS-HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSN 75

84 GCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQEN 143
   GCVSKILGRYYETGSIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E
76 GCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEG 135

144 VCTNDNIPSVSSINRVLRNLAAQKEQ 169
    VCTNDNIPSVSSINRVLRNLA++K+Q
136 VCTNDNIPSVSSINRVLRNLASEKQQ 161

398 TEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQV 457
    +++ Q RL LKRKLQRNRTSFT +QI++LEKEFERTHYPDVFARERLA KI LPEARIQV
222 SDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQV 281

458 WFSNRRAKWRREEKLRNQRR 477
    WFSNRRAKWRREEKLRNQRR
282 WFSNRRAKWRREEKLRNQRR 301
```

Sequence Alignment

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++           +L+ L+++H+ K
LGB2_LUPLU  NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
            GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
  F11G11.2  GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

HBA_HUMAN: Human Alpha Globin
HBB_HUMAN: Human Beta Globin
F11G11.2 : Leghaemoglobin from yellow lupin

- **Needleman-Wunsch**
- **Smith-Waterman**

Sequence Alignment

Input: Sequence **A** , Sequence **B**, Database **D**

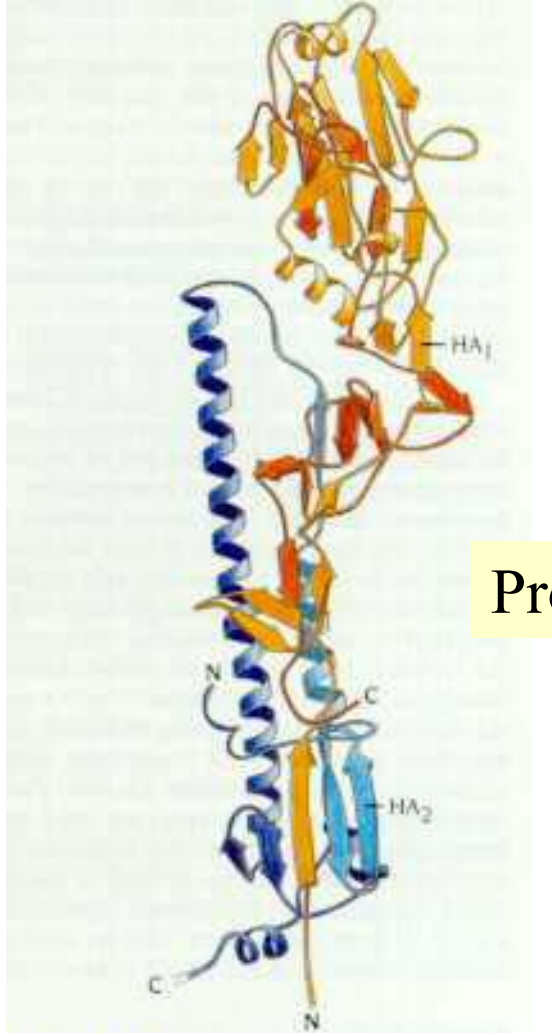
Question(s):

- Align **A** and **B.**
  - Determine *similarity* (**A** , **B**).
- Align **A** and **D.**
  - Find sequence in **D** with maximum similarity to **A**.
- *Many More variants*

Molecular Biology Background

The 2 star players

DNA



Protein

Figure 5.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (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 HA_1. (Courtesy of Don Wiley, Harvard University.)

The Players

DNA

String with alphabet {A, C, G, T}

Nucleotides/Bases

RNA

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

Protein

String with 20-letter alphabet

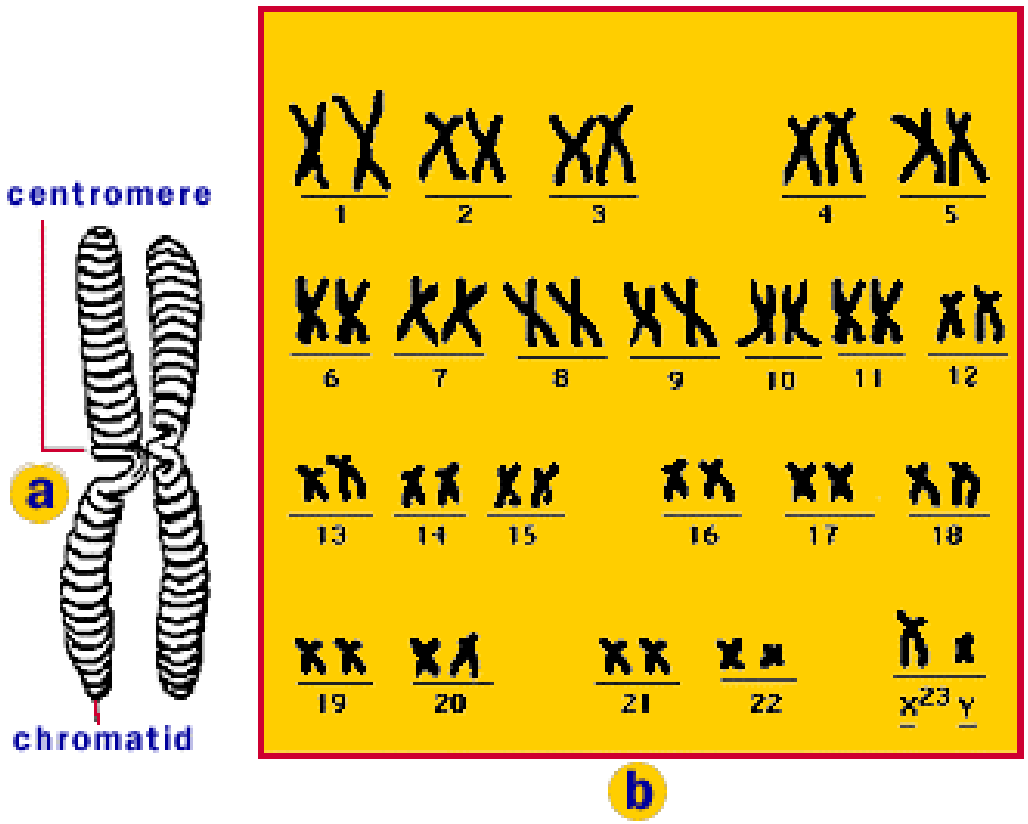
Amino acids/Residues

Central Dogma

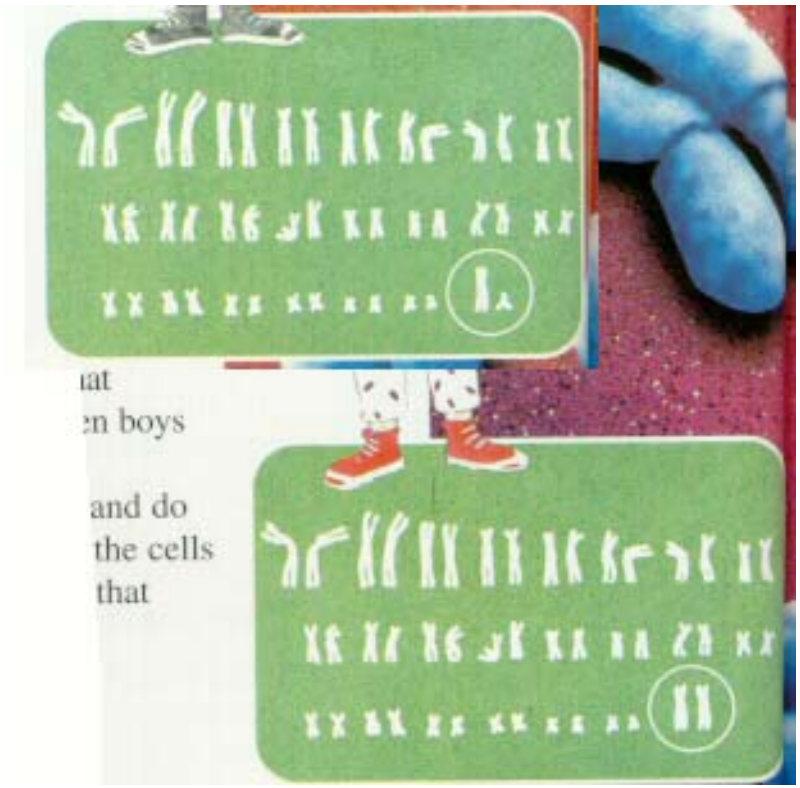
- DNA acts as a template to replicate itself.
- DNA is transscribed into RNA.
- RNA is translated into **Protein**.

Chromosomes

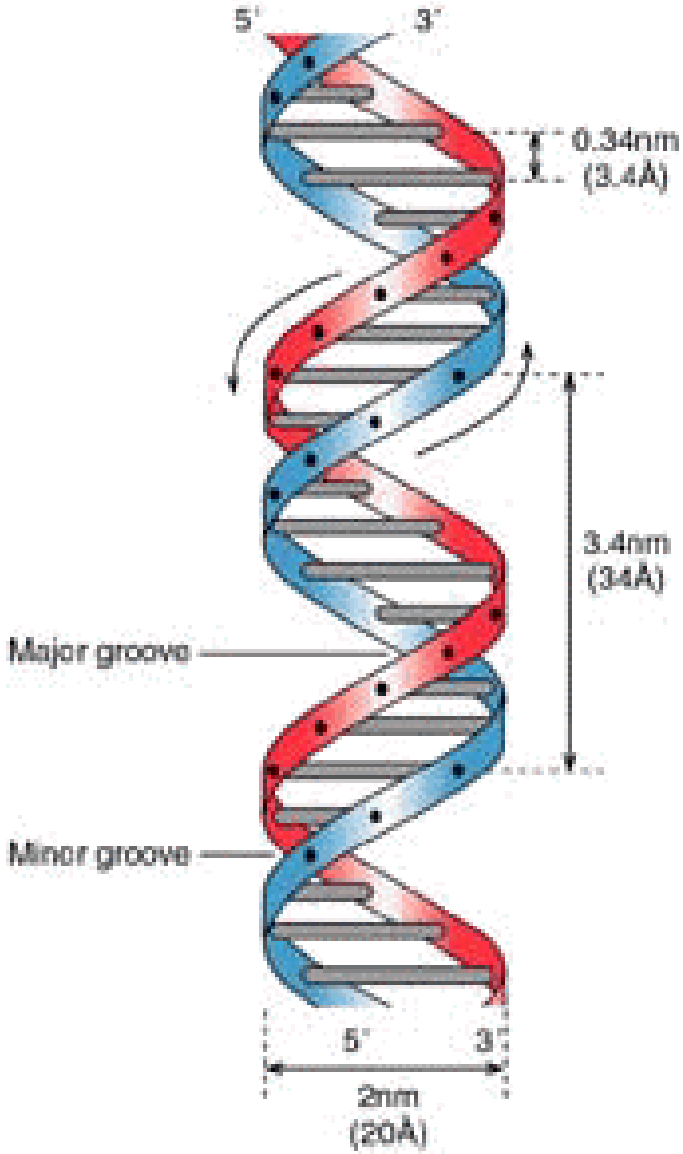
Human chromosomes!



Chromosomes

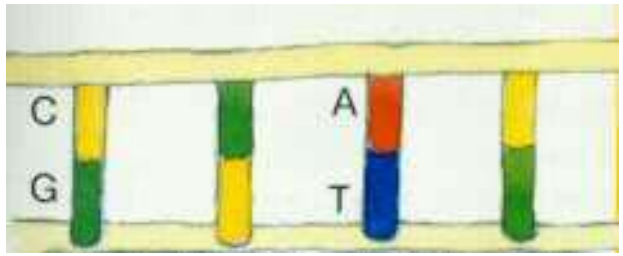


DNA Molecule



DNA

Complementary Bases



Proteins – Amino acids

amino acid	3 letter code	1 letter code
alanine	Ala	A
arginine	Arg	R
aspartic acid	Asp	D
asparginine	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*

RNA

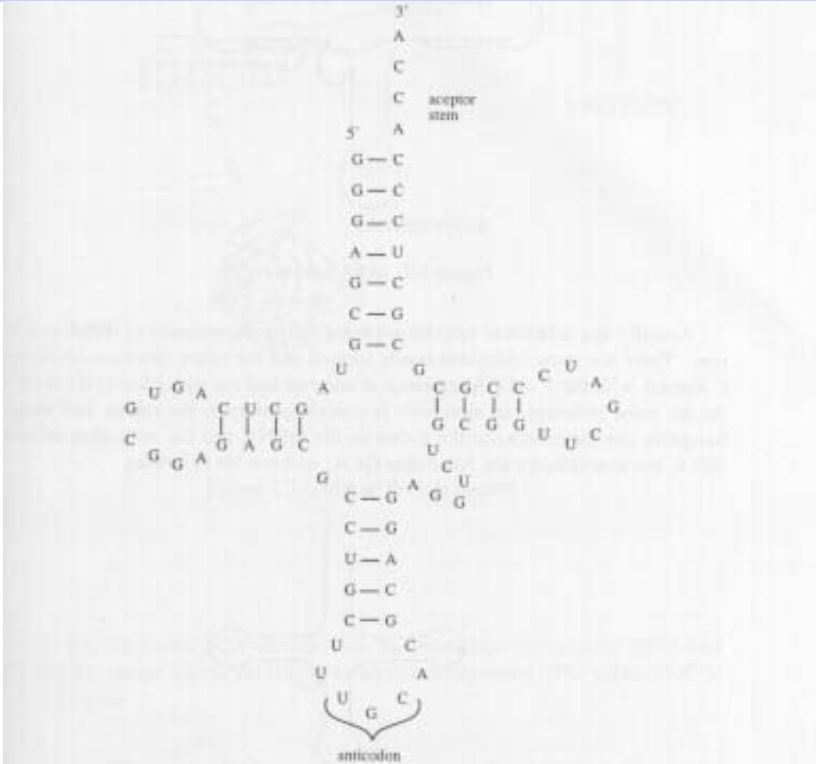
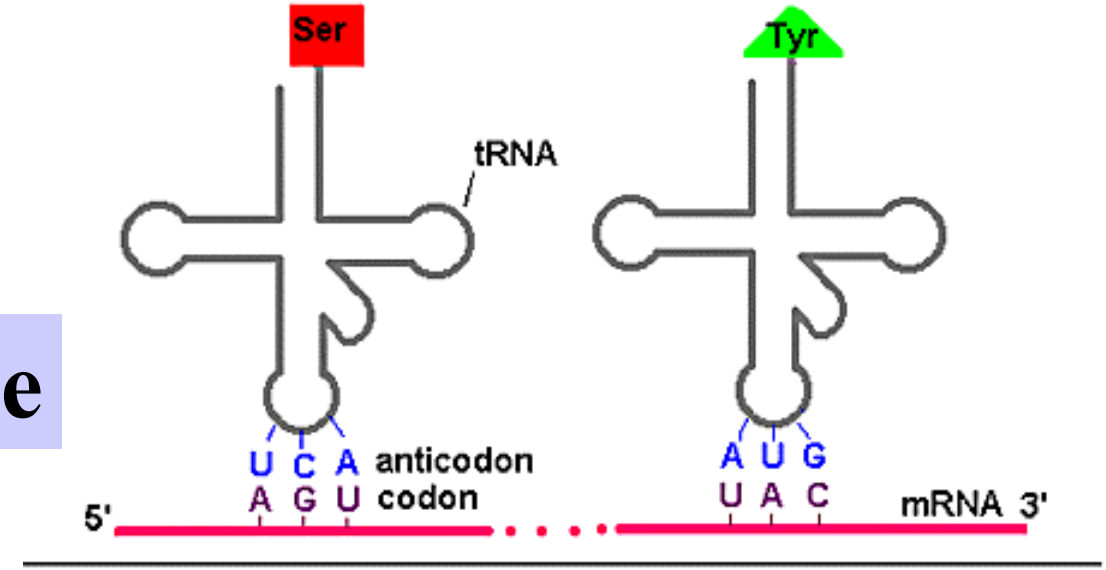


Figure 1.1: E. coli *Ala tRNA*

The Genetic Code



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

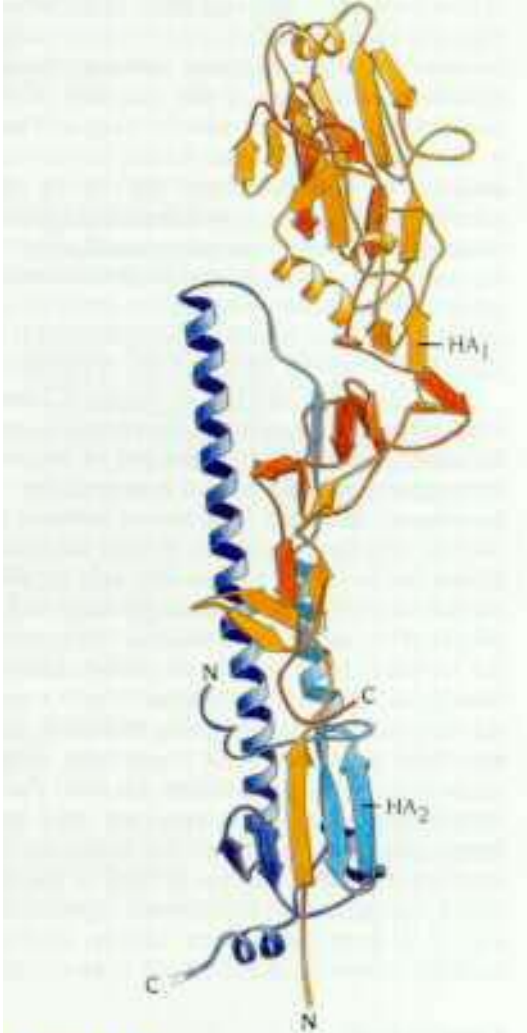
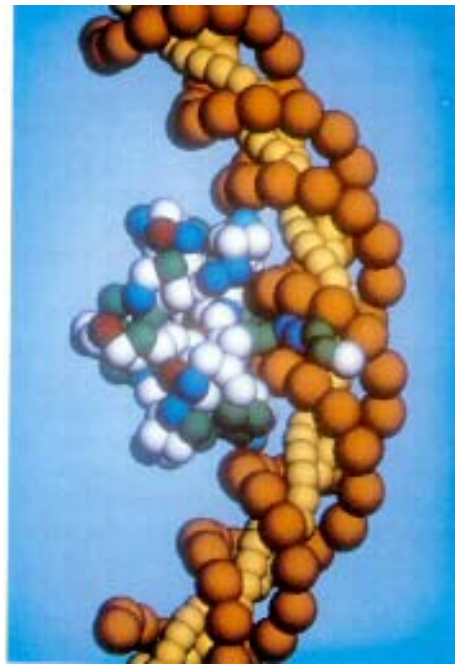
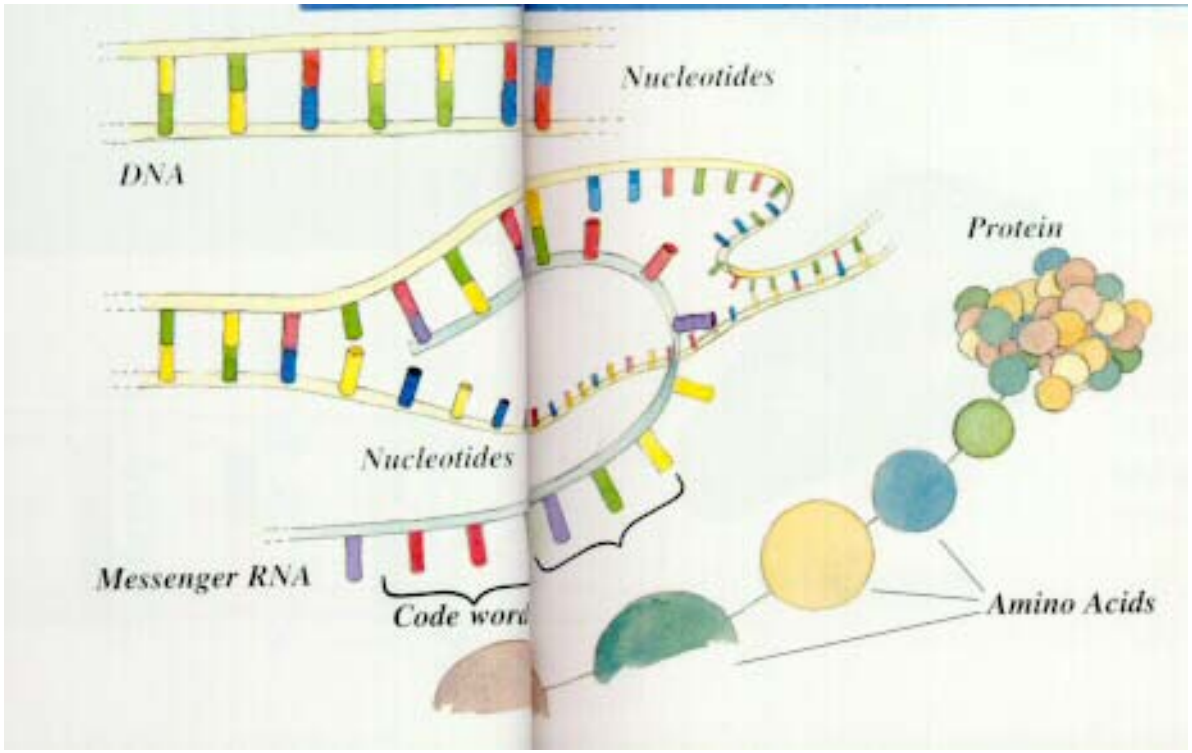
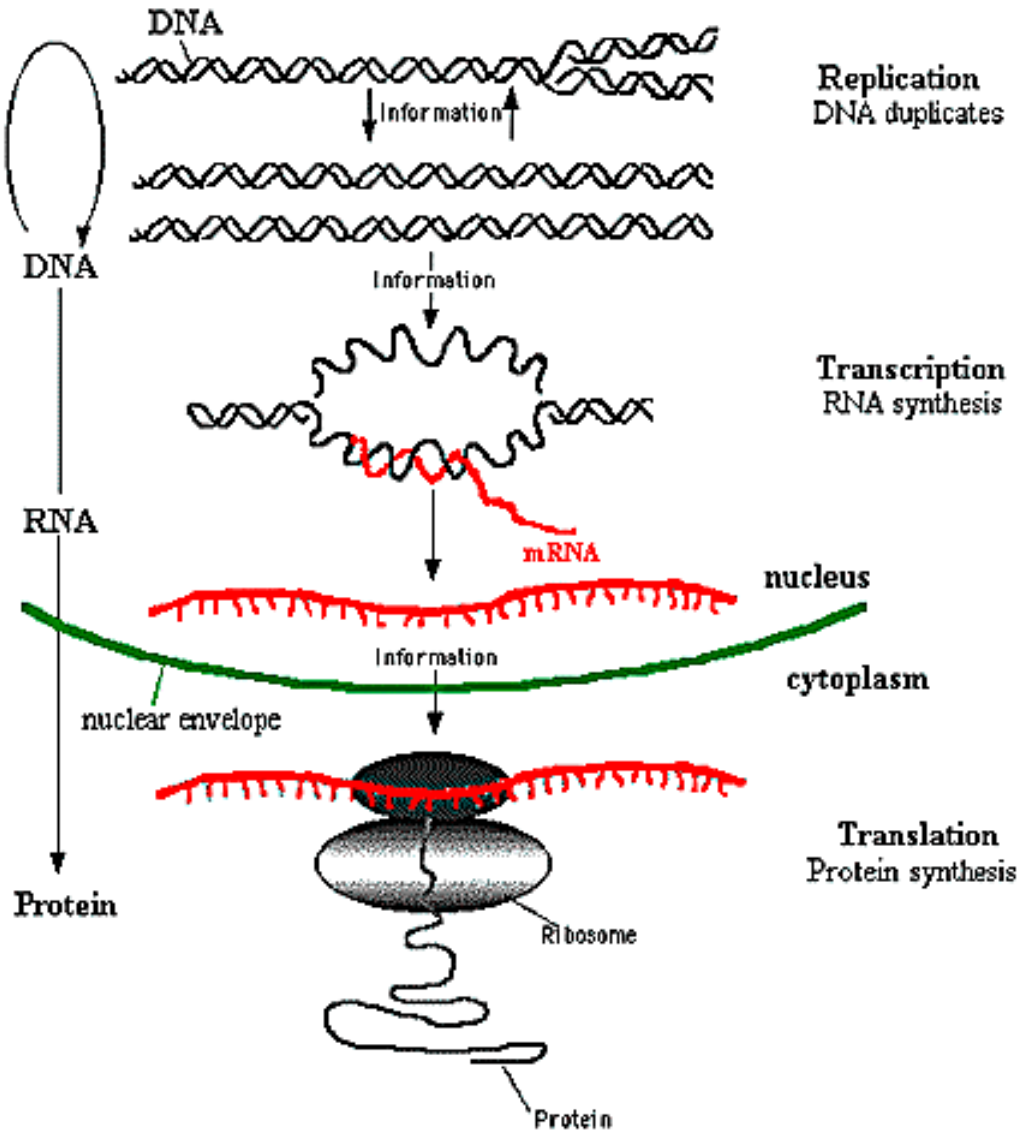


Figure 8.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike 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 HA_1. (Courtesy of Don Wiley, Harvard University.)

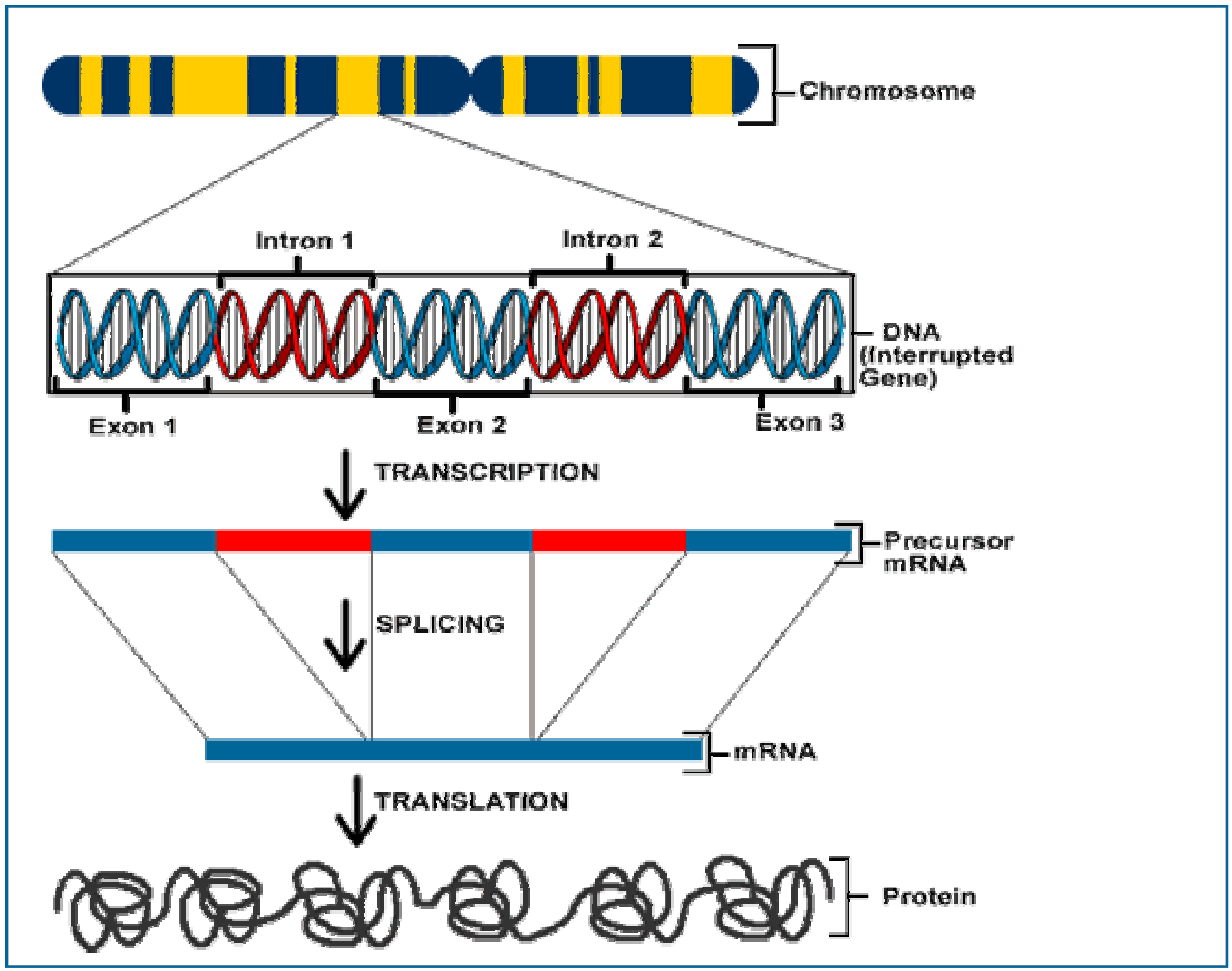


DNA → RNA → Protein

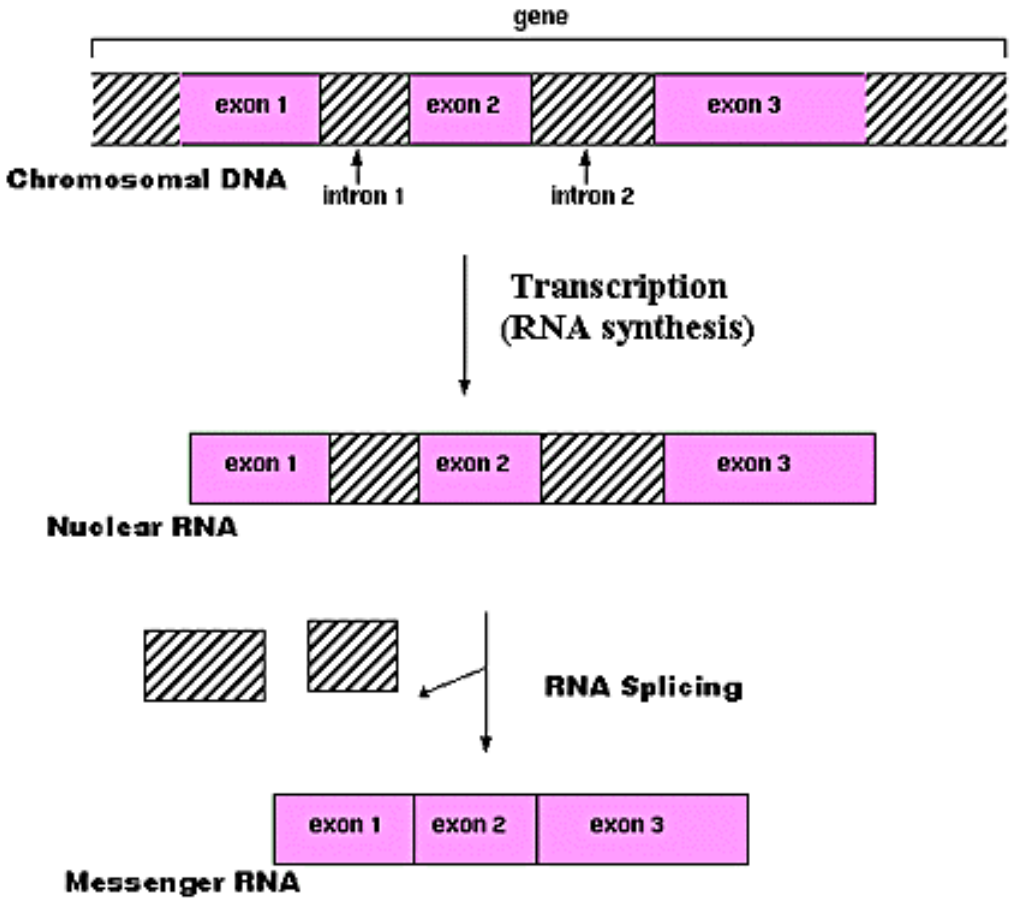




The Central Dogma of Molecular Biology

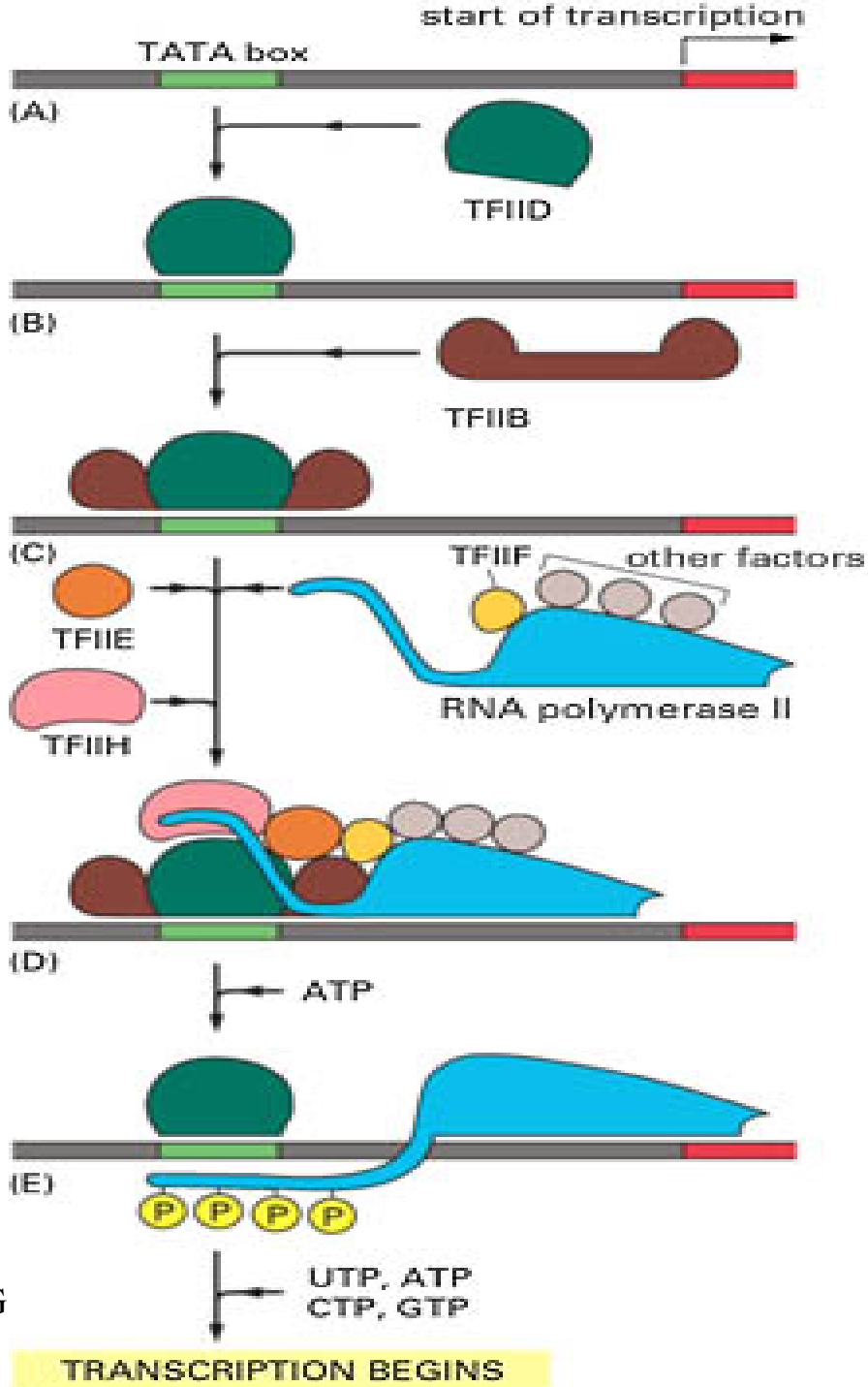

Chromosome
Intron 1
Intron 2
DNA
(Interrupted
Gene)
Exon 1
Exon 2
Exon 3
TRANSCRIPTION
Precursor
mRNA
SPLICING
mRNA
TRANSLATION
Protein

DNA Transcription



RNA synthesis and processing

Transcription Initiation



Transcription

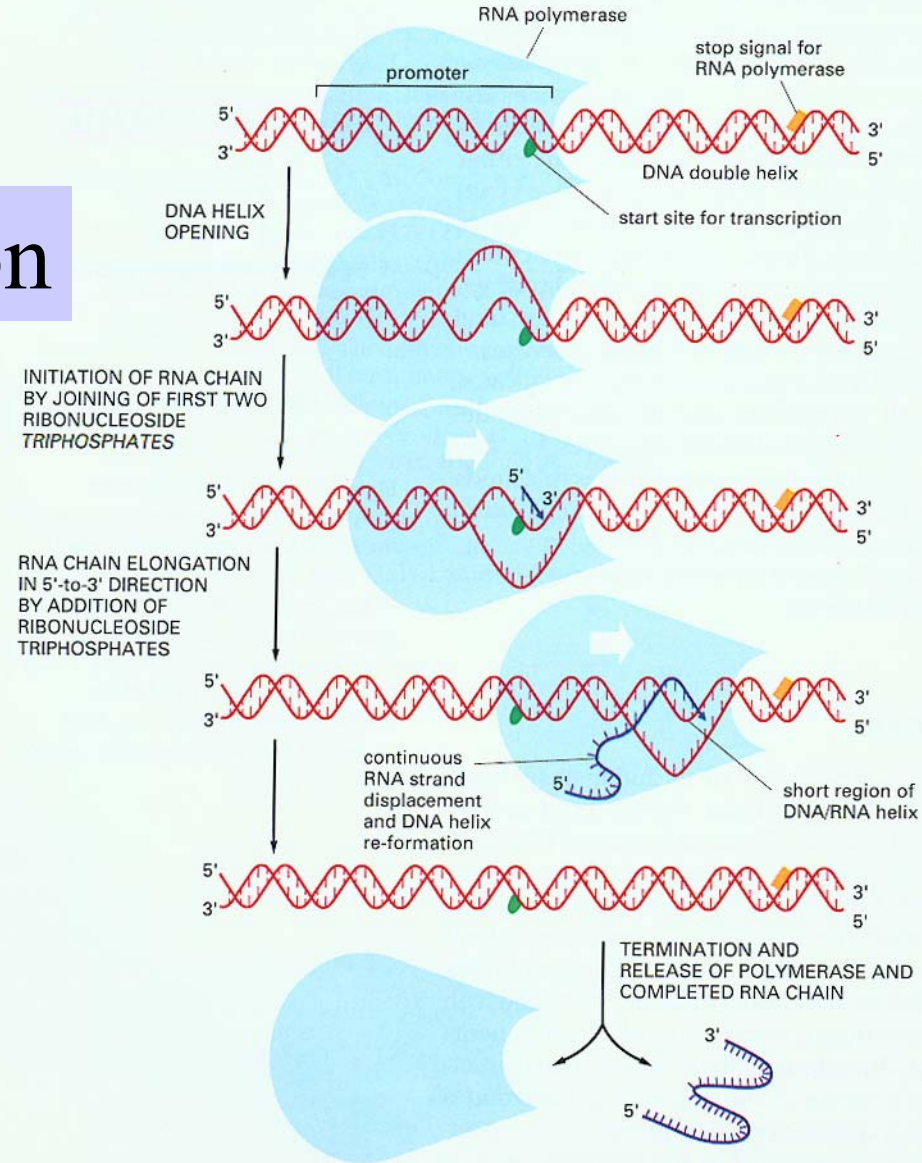


Figure 6–2 The synthesis of an RNA molecule by RNA polymerase. The enzyme binds to the promoter sequence on the DNA and begins its synthesis at a start site within the promoter. It completes its synthesis at a stop (termination) signal, whereupon both the polymerase and its completed RNA chain are released. 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.

224 Chapter 6 : Basic Genetic Mechanisms

Transcription Steps

RNA polymerase needs many transcription factors (TFIIA,TFIIB, etc.)

(A) The promoter sequence (TATA box) is located 25 nucleotides away from transcription initiation site.

(B) The TATA box is recognized and bound by transcription factor TFIID, which then enables the adjacent binding of TFIIB. DNA is somewhat distorted in the process.

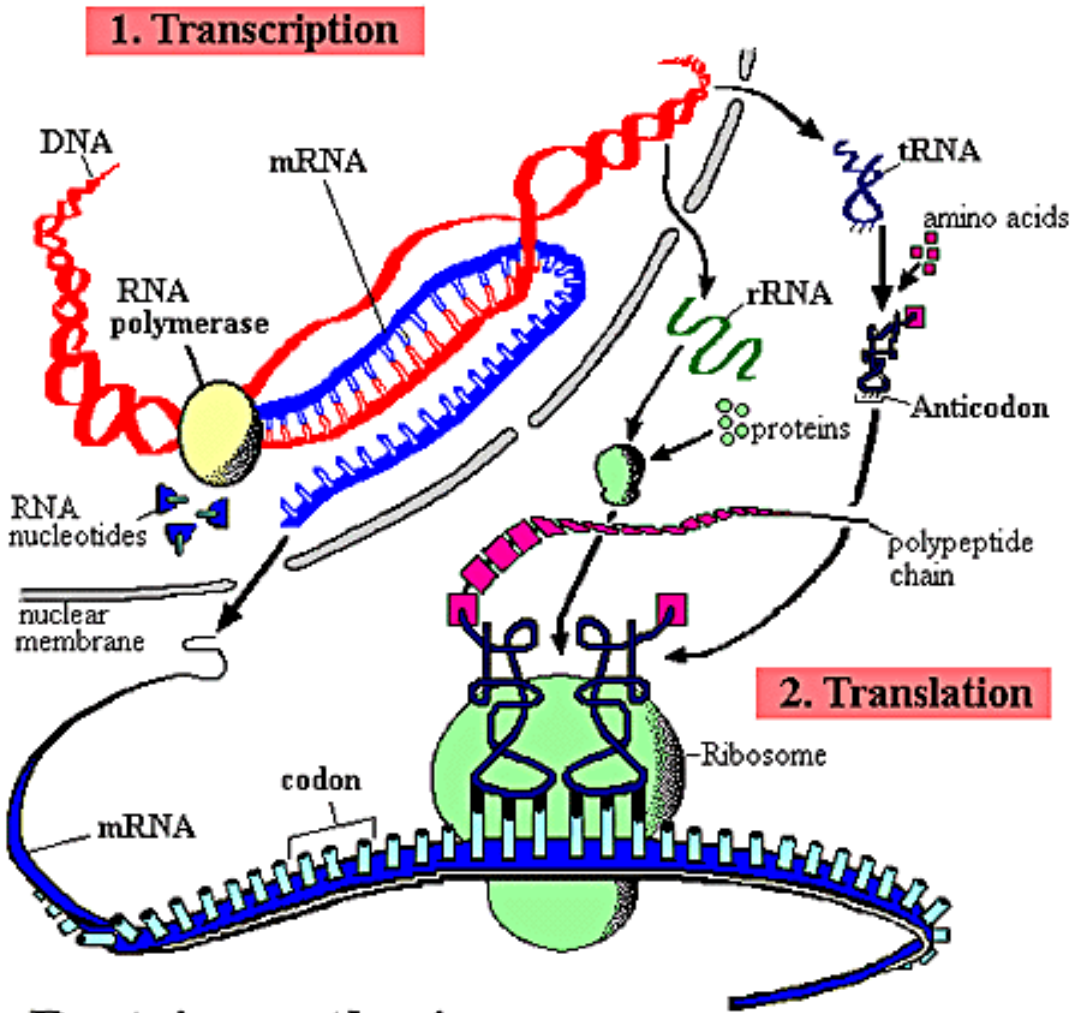
(D) The rest of the general transcription factors as well as the RNA polymerase itself assemble at the promoter. What order?

(E) TFIIH then uses ATP to phosphorylate RNA polymerase II, changing its conformation so that the polymerase is released from the complex and is able to start transcribing. As shown, the site of phosphorylation is a long polypeptide tail that extends from the polymerase molecule.

Transcription Factors

- The general transcription factors have been highly conserved in evolution; some of those from human cells can be replaced in biochemical experiments by the corresponding factors from simple yeasts.

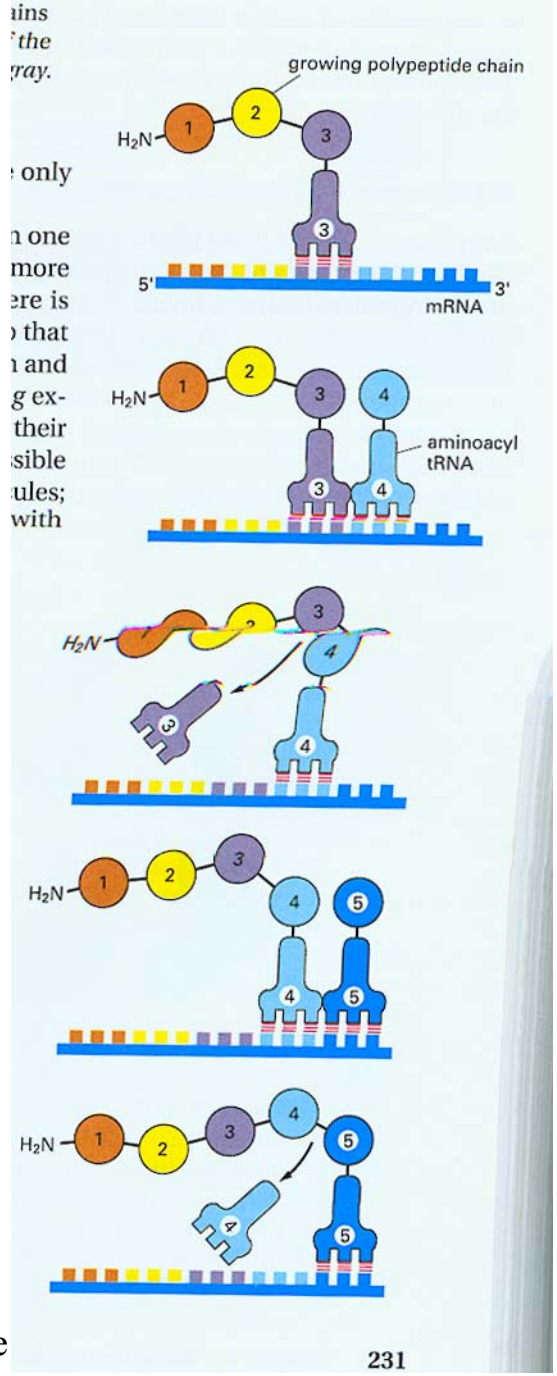
Protein Synthesis

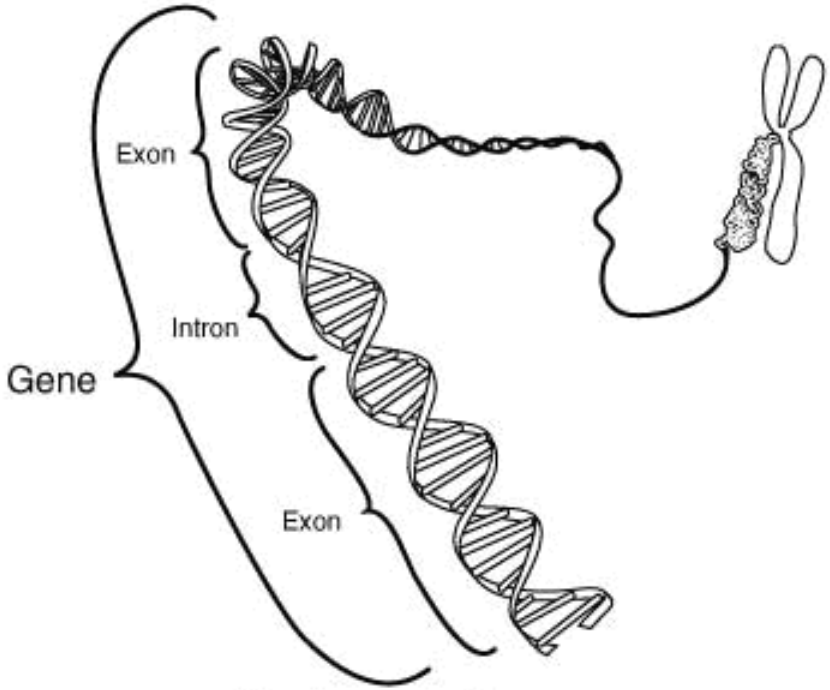


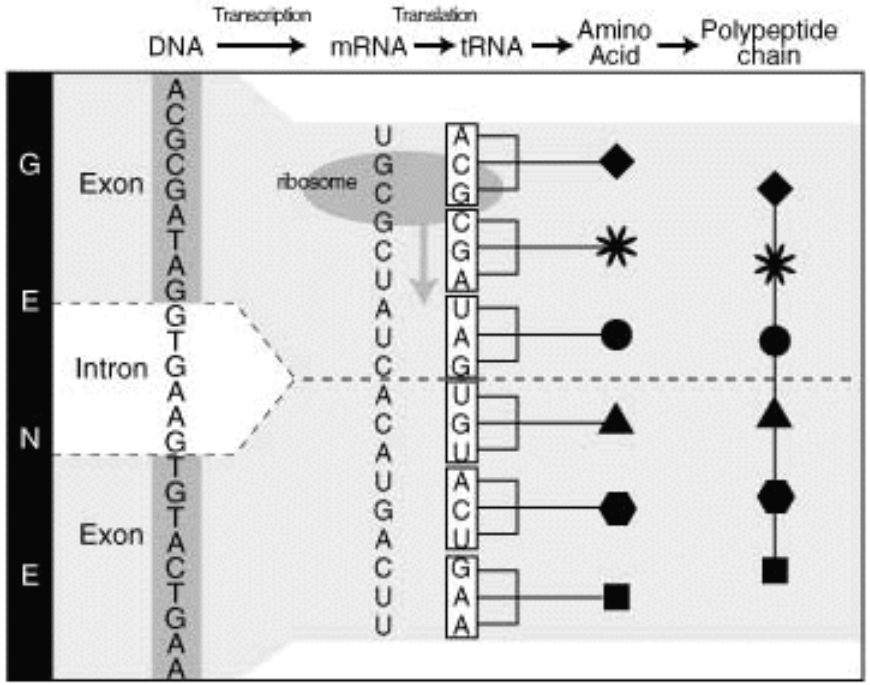
Protein synthesis

Protein Synthesis:

Incorporation of amino acid into protein




Exon
Intron
Gene
Exon


Transcription
Translation
DNA
mRNA
tRNA
Amino Acid
Polypeptide chain
G
E
N
E
Exon
Intron
Exon
ACGCGATAGGTGAAGTGTACTGAA
ribosome
UGCGCUAUCACACAUGACUU
ACG
CGA
UAG
UGU
ACU
GAA