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/BioinfS13.html

Evaluation

- Semester Project (45 %)
- Homework Assignments (20 %)
- Exam (15 %)
- Quizzes (10 %)
- Summary Reports of Interest (5 %)
- Class Participation (5 %)

Course Homepage

http://www.cis.fiu.edu/~giri/teach/BioinfS13.html

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

Molecular Biology Background

2 star molecular players

Also Starring:
RNA

DNA

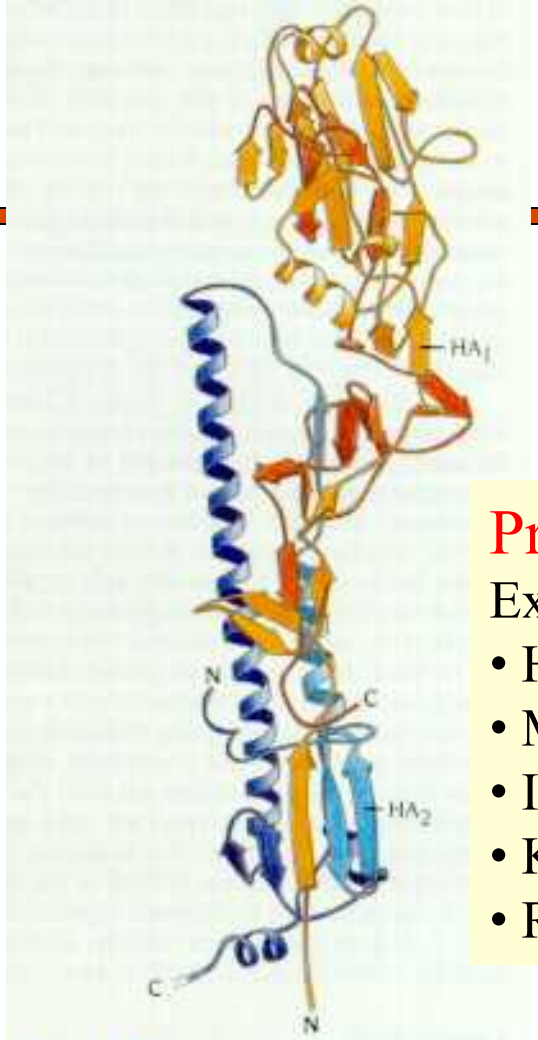


Figure 5.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75Å long. The globular head is formed by residues only from HA_1. (Courtesy of Don Wiley, Harvard University.)

Protein

Examples:

- Hemoglobin
- Melanin
- Insulin
- Keratin
- RNA Polymerase

The Polymeric Players

DNA

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

RNA

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

Protein

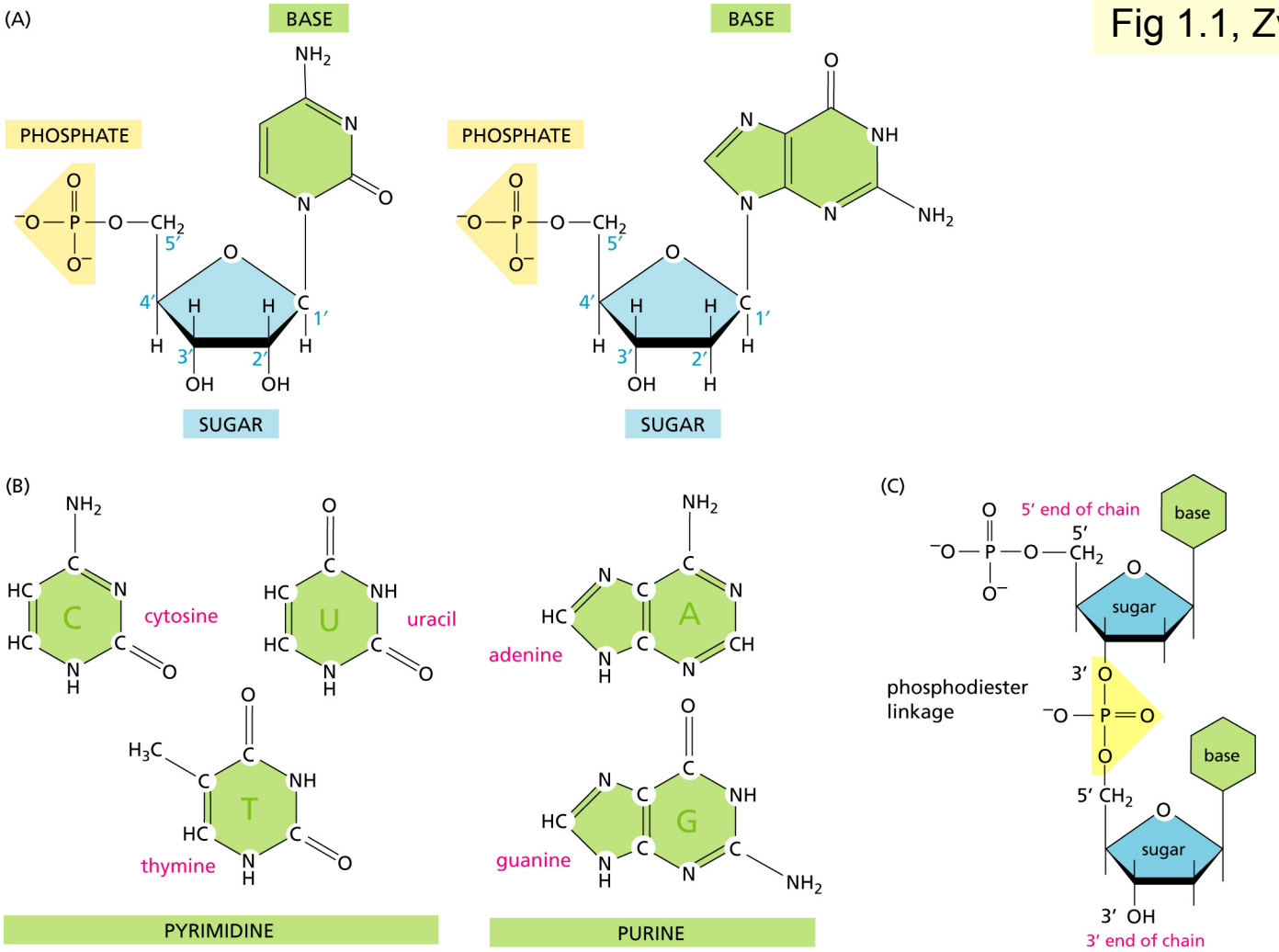
String with 20-letter alphabet **Amino acids/Residues**

Typical DNA Sequence

```
  1 gggagaacac ccggagaagg aggaggaggc gaagaaaagc aacagaagcc cagttgctgc
 61 tccaggtccc tcggacagag ctttttccat gtggagactc tctcaatgga cgtgccccct
121 agtgcttctt agacggactg cggtctccta aaggtcgacc atggtggccg ggacccgctg
181 tcttctagtg ttgctgcttc cccaggtcct cctgggcggc gcggccggcc tcattccaga
241 gctgggccgc aagaagttcg ccgcggcatc cagccgaccc ttgtcccggc cttcggaaga
301 cgtcctcagc gaatttgagt tgaggctgct cagcatgttt ggcctgaagc agagacccac
361 ccccagcaag gacgtcgtgg tgccccccta tatgctagat ctgtaccgca ggcactcagg
421 ccagccagga gcgcccgccc 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 ttcctgtgac
721 cagactattg gacaccaggt tagtgaatca gaacacaagt cagtgggaga gcttcgacgt
781 caccccagct 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
```

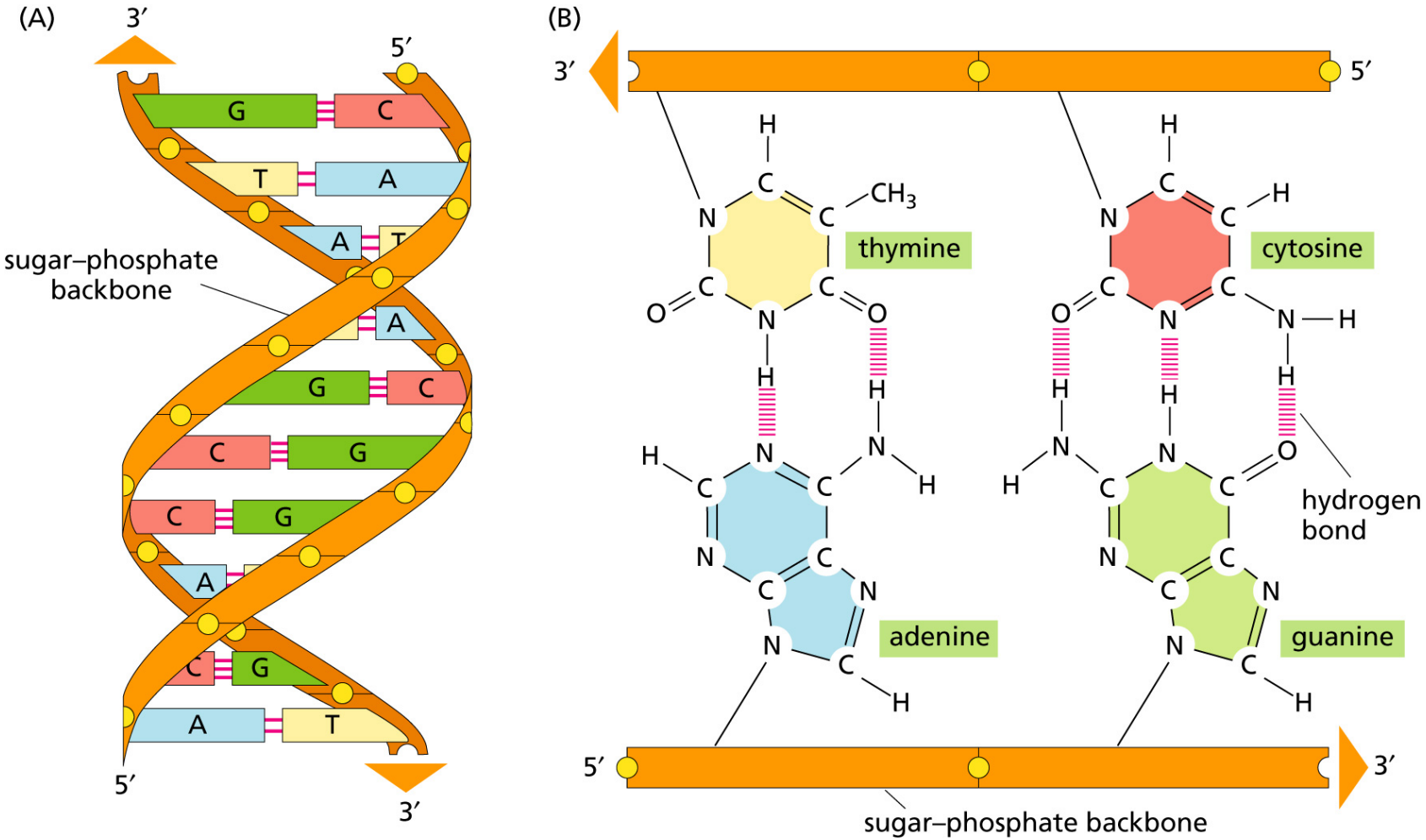
The building blocks of DNA & RNA

Fig 1.1, Zvelebil/Baum



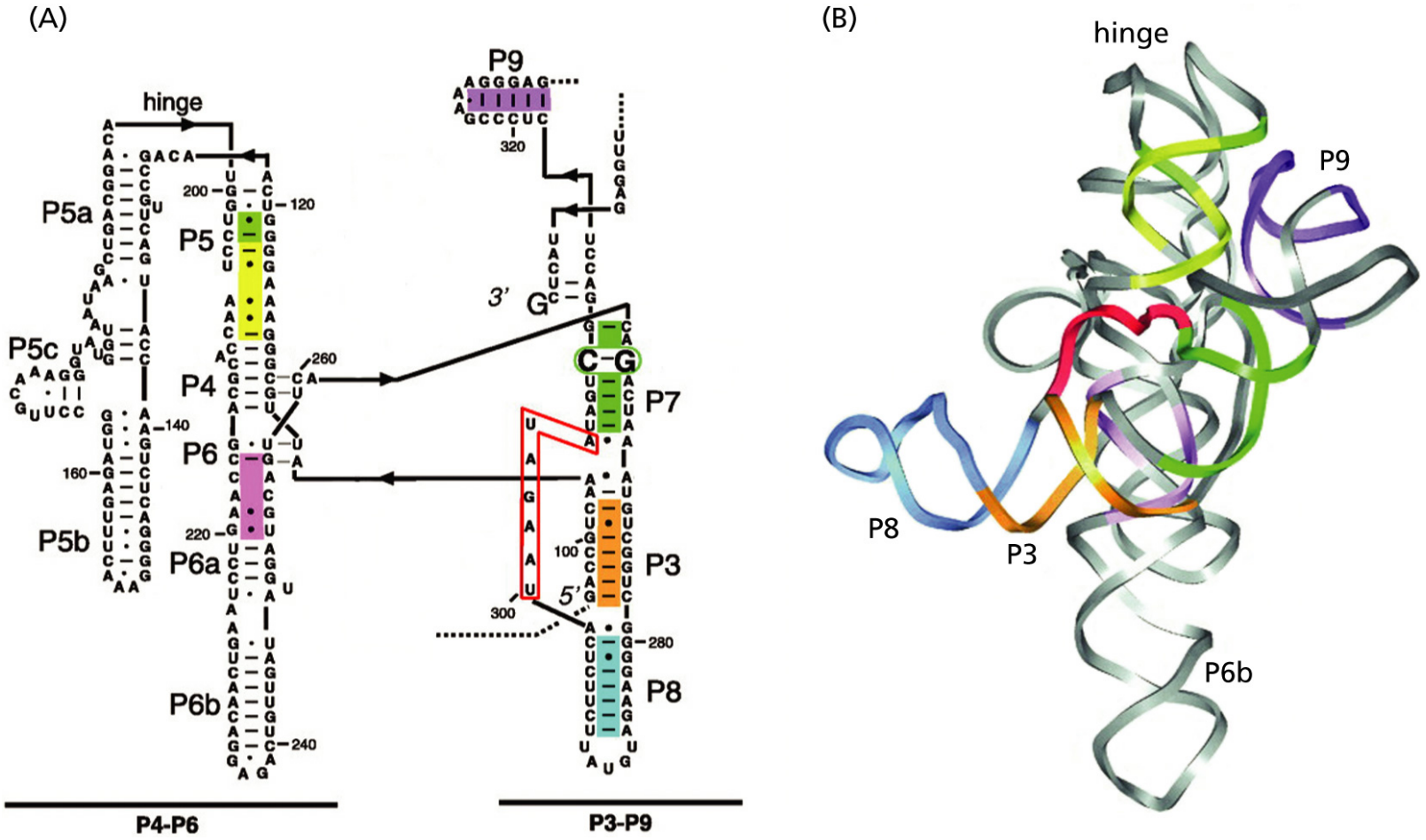
DNA double helix structure

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
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*

Typical protein sequence

/translation="MVAGTRCLLVLLLPQVLLGGAAGLIPELGRKKFAAASSRPLSRP
SEDVLSEFELRLLSMFGLKQRPTPSKDVVVPPYMLDLYRRHSGQPGAPAPDHRLERAA
SRANTVRSFHHEEAVEELPEMSGKTARRFFFNLSSVPSDEFLTSAELQIFREQIQEAL
GNSSFQHRINIYEIIKPAAANLKFPVTRLLDTRLVNQNTSQWESFDVTPAVMRWTTQG
HTNHGFVVEVAHLEENPGVSKRHVRISRSLHQDEHSWSQIRPLLVTFGHDGKGHPLHK
REKRQAKHKQRKRLKSSCKRHPLYVDFSDVGWNDWIVAPPGYHAFYCHGECPFPLADH
LNSTNHAIVQTLVNSVNSKIPKACCVPTELSAISMLYLDENEKVVLKNYQDMVVEGCG
CR"

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

Protein 3D Structure

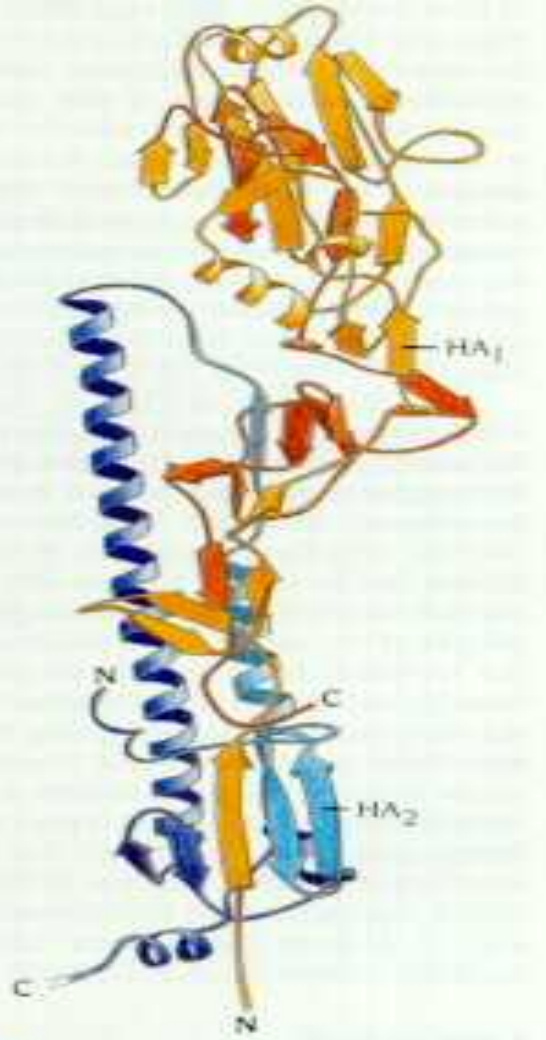
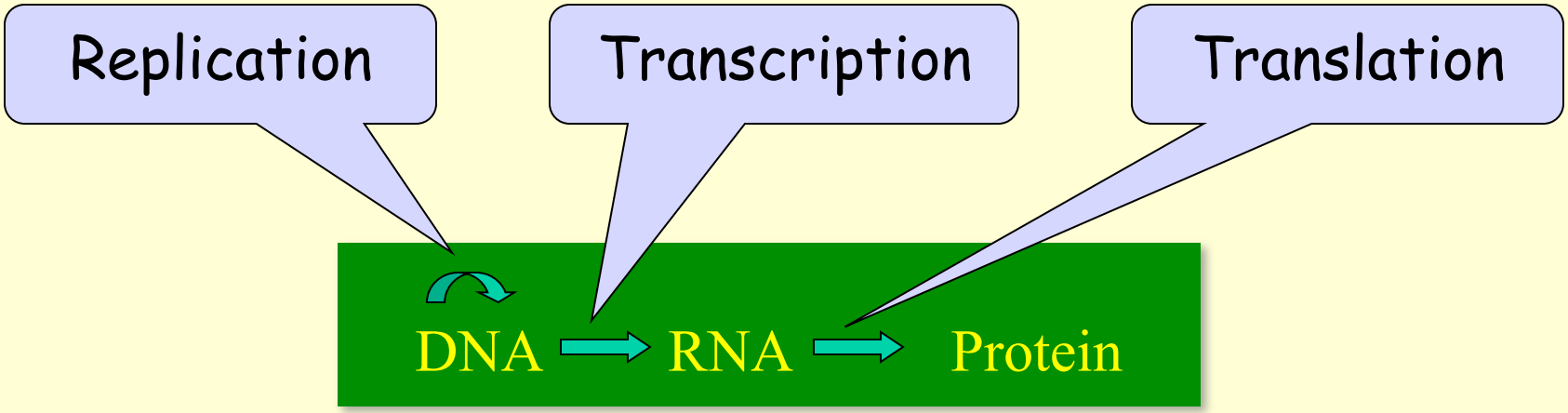


Figure 8.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike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

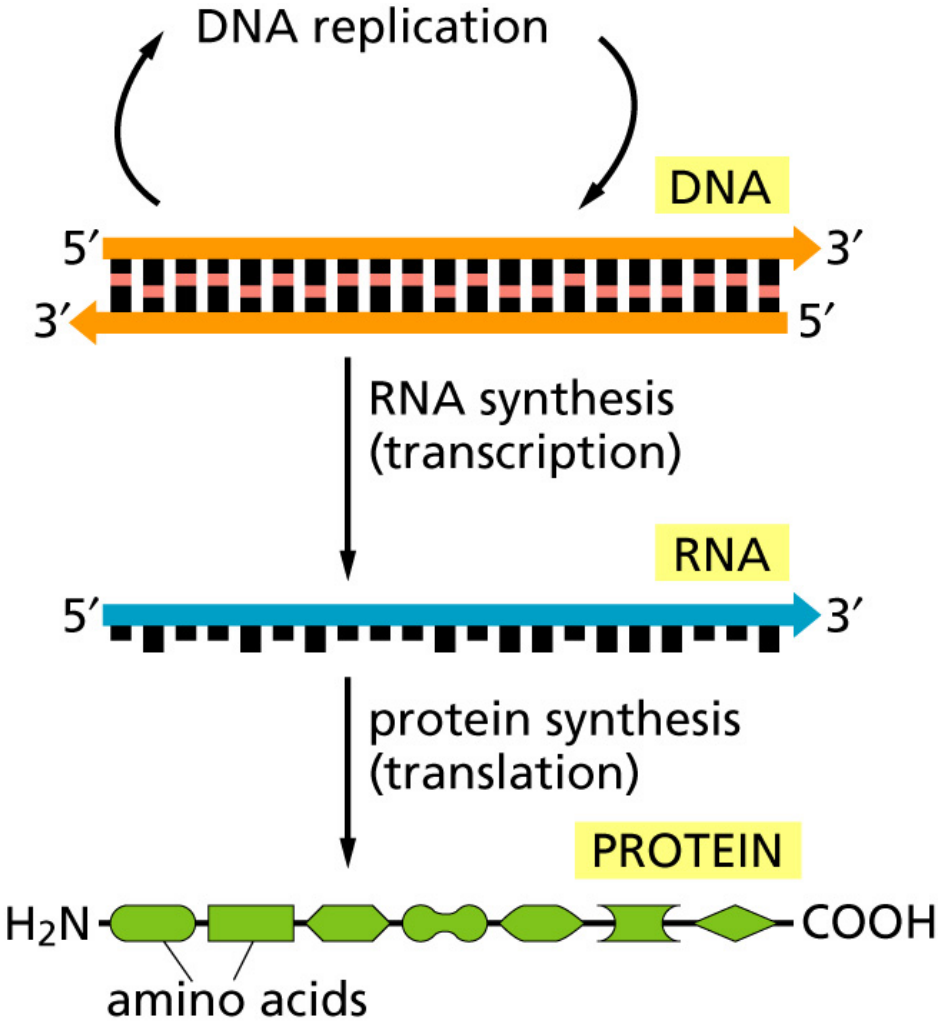
Central Dogma

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



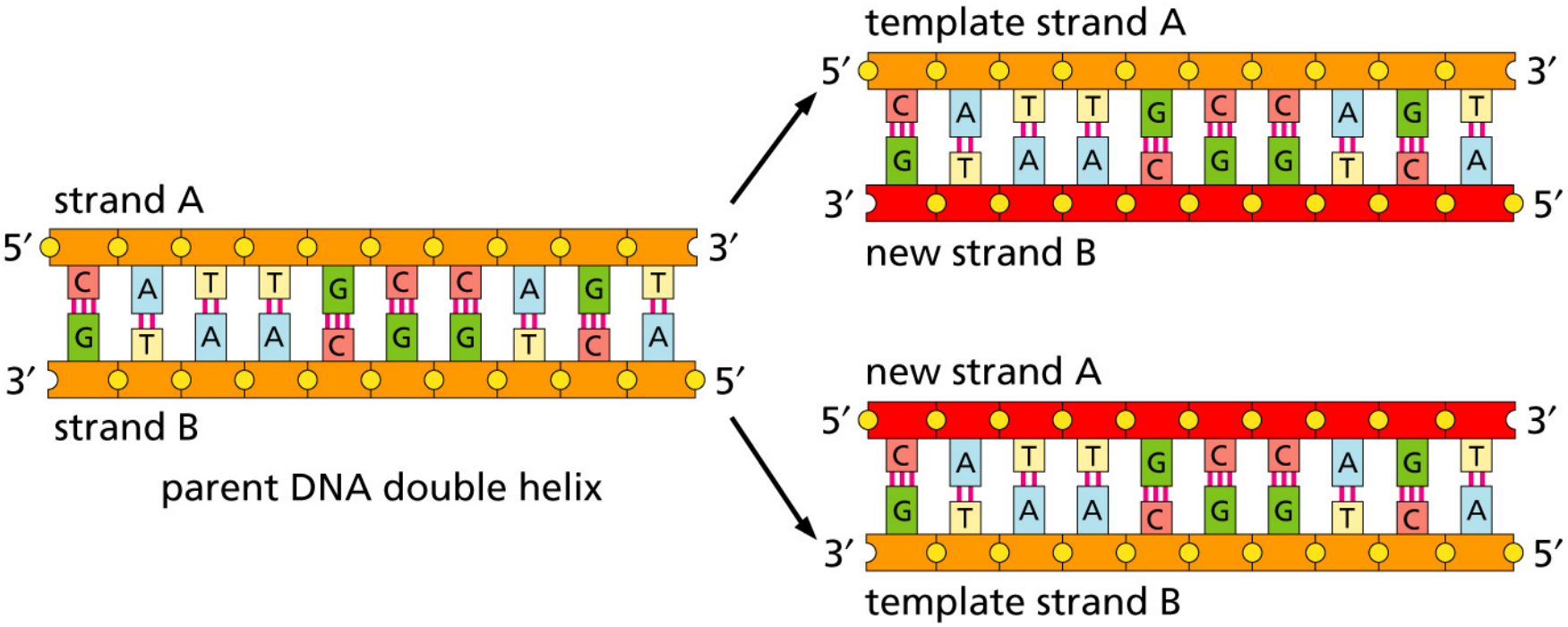
Central Dogma

Fig 1.6, Zvelebil/Baum

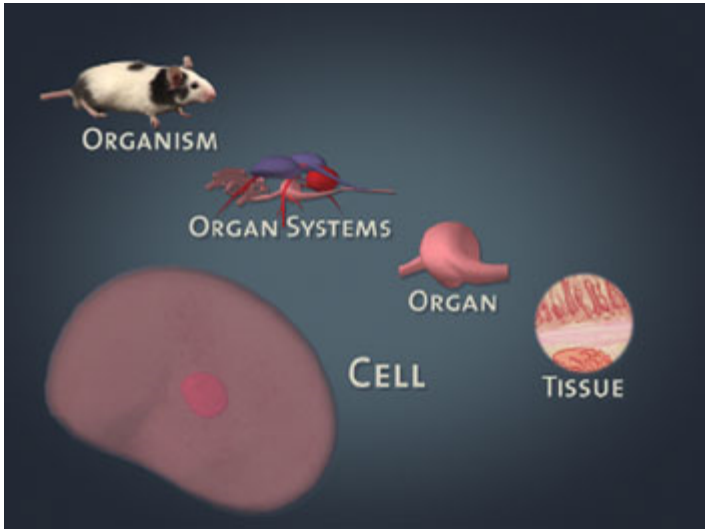


DNA Replication

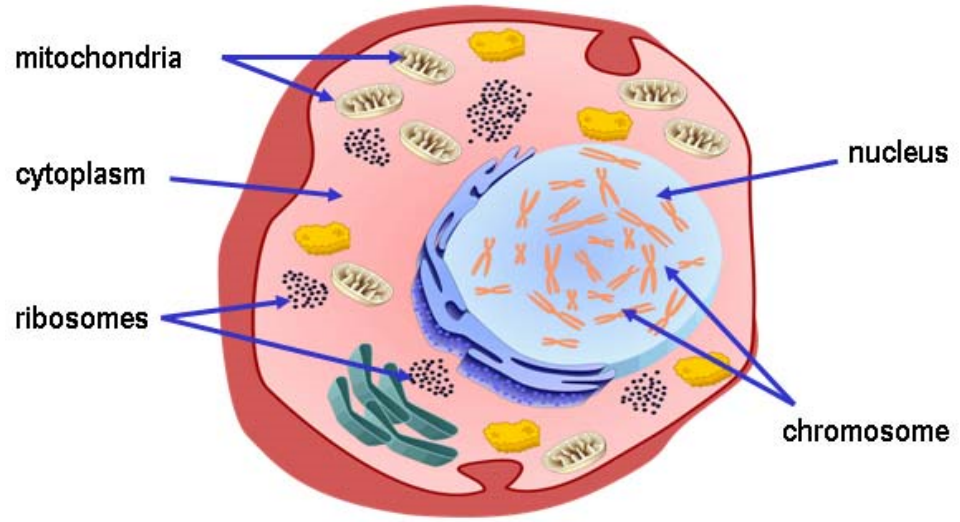
Fig 1.4, Zvelebil/Baum



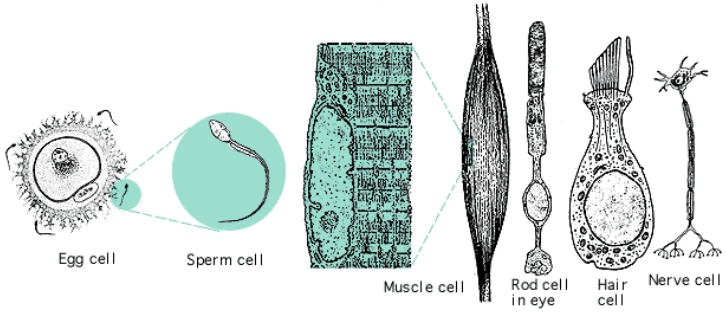
Cell



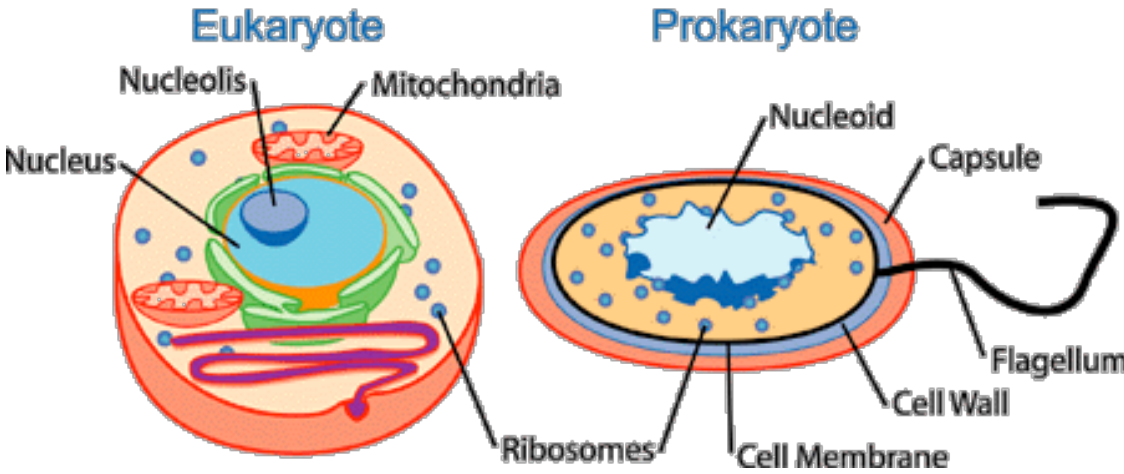
http://www.learner.org/channel/courses/essential/life/session1/closer1.html



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



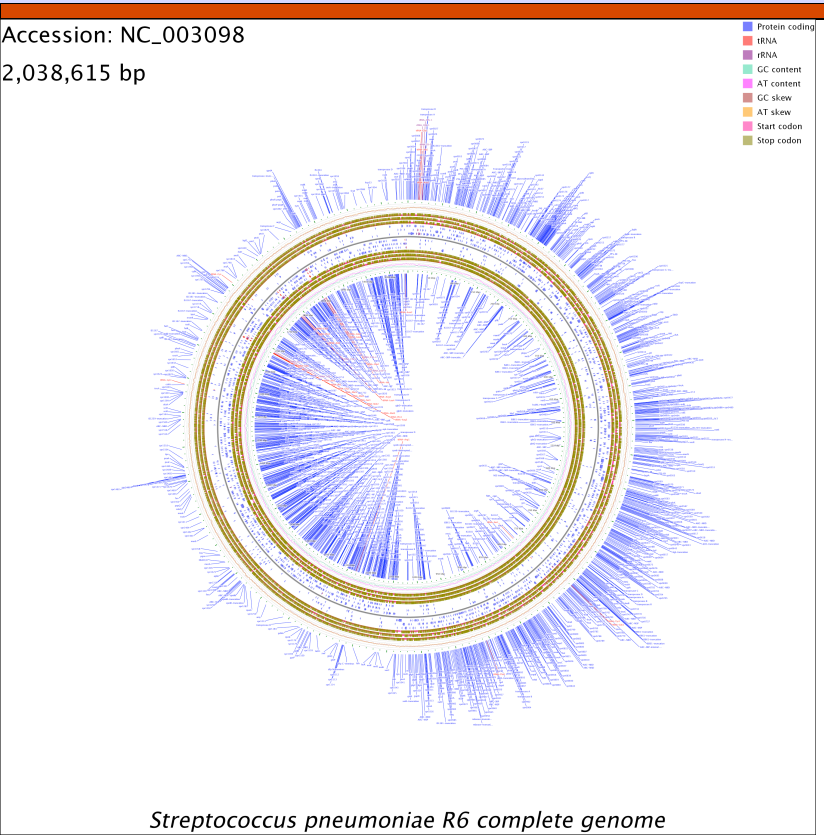
http://www.biology.eku.edu/RITCHISO/301notes1.htm



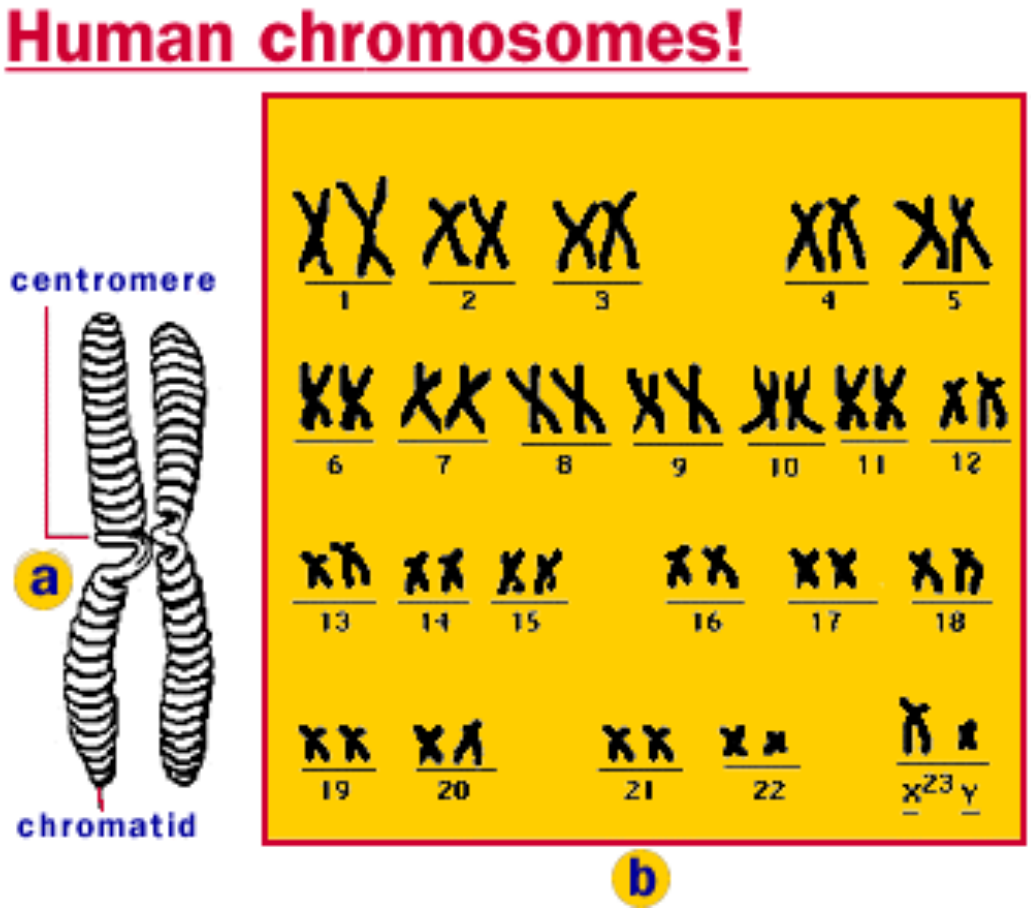
http://en.wikipedia.org/wiki/File:Celltypes.png

http://www.cellsalive.com/cells/cell_model.htm

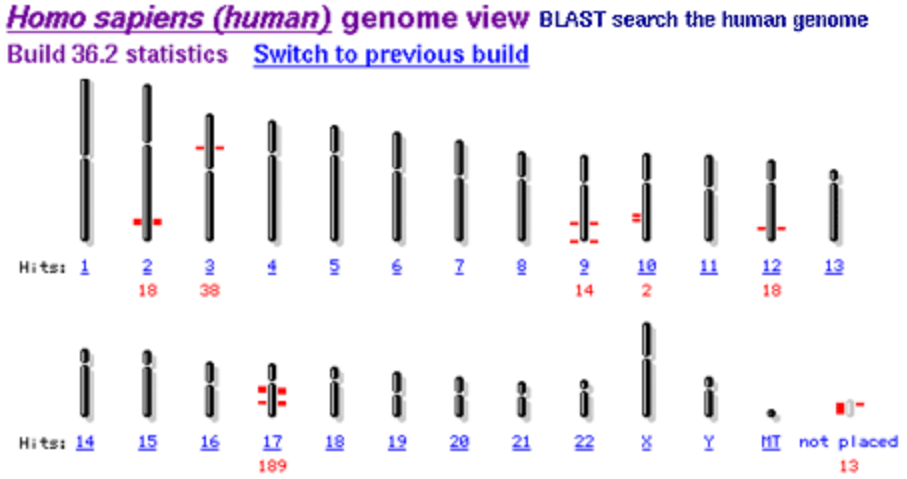
Chromosomes

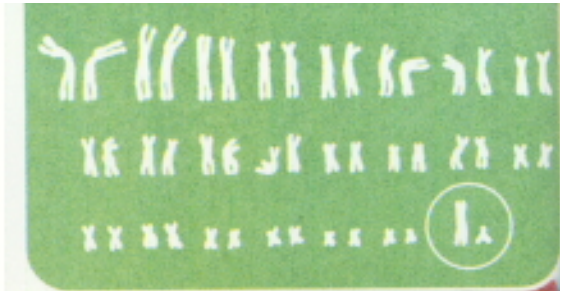


Streptococcus pneumoniae R6 complete genome

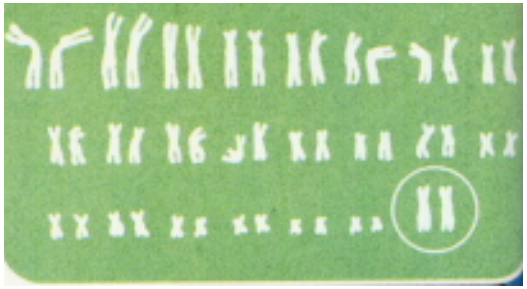


Chromosomes

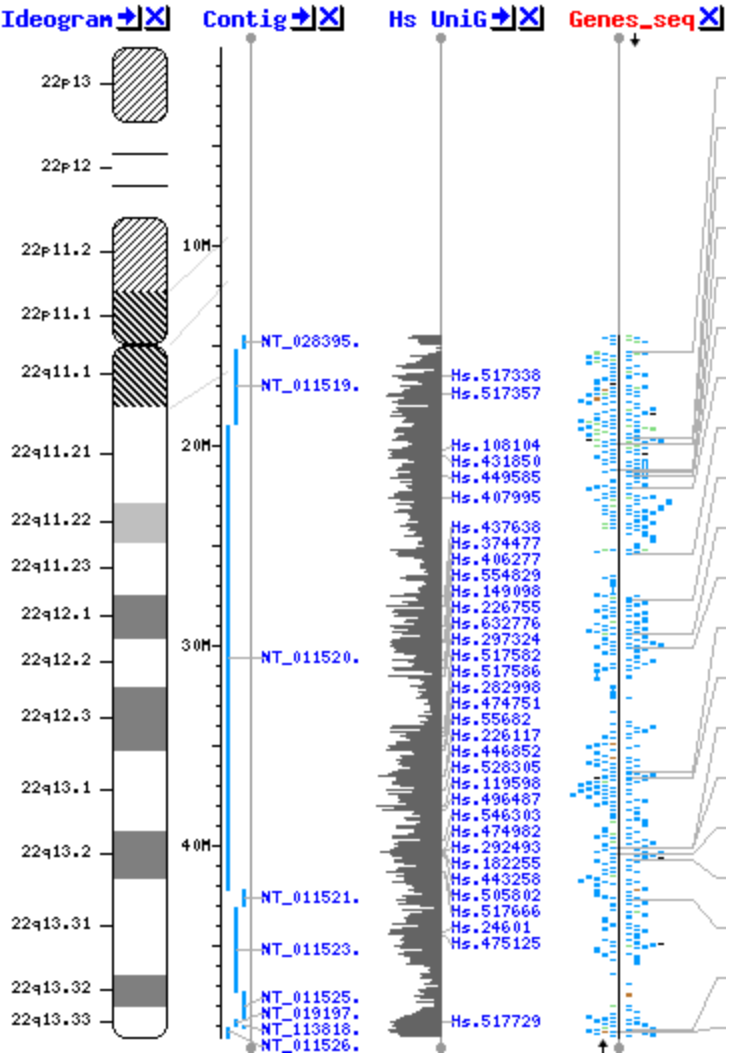




The chromosomal locations of several genes believed to be associated with the human BRCA1 gene implicated in breast cancer are highlighted.



Human Chr 22

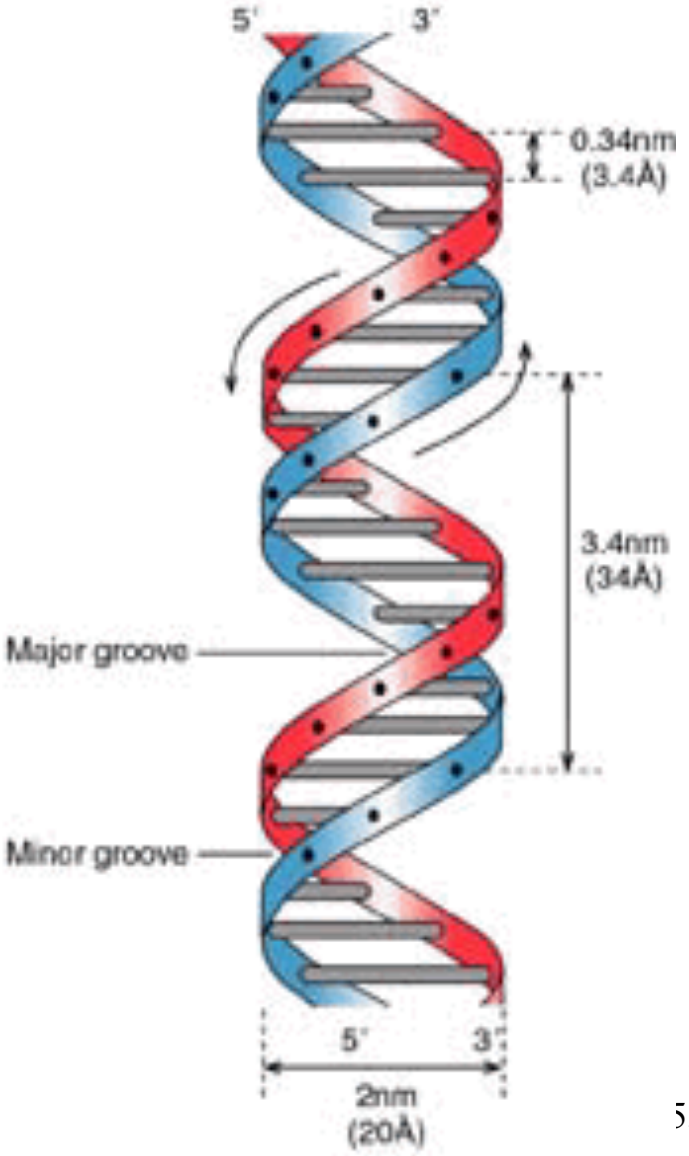


Symbol	Position	Description
ABCD1P4	22q11	ATP-binding cassette, sub-family D (ALD)
SNAP29	22q11.21	synaptosomal-associated protein

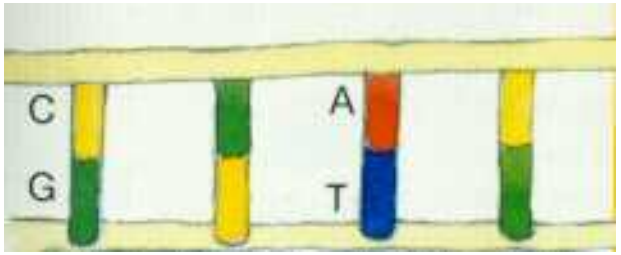
-
-
-

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

DNA Molecule



Complementary Bases



Genes

Basic Genetic Processes

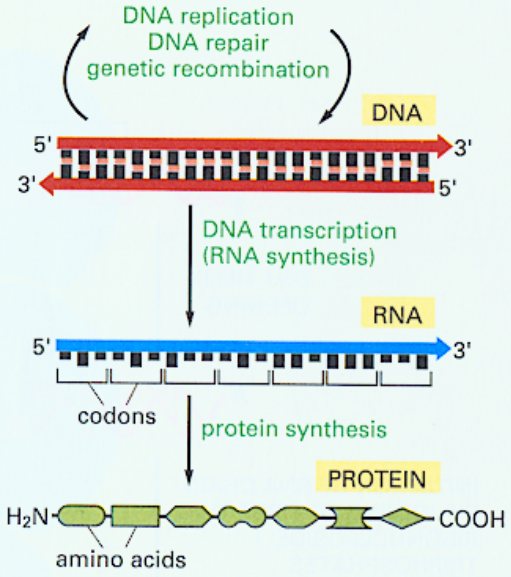
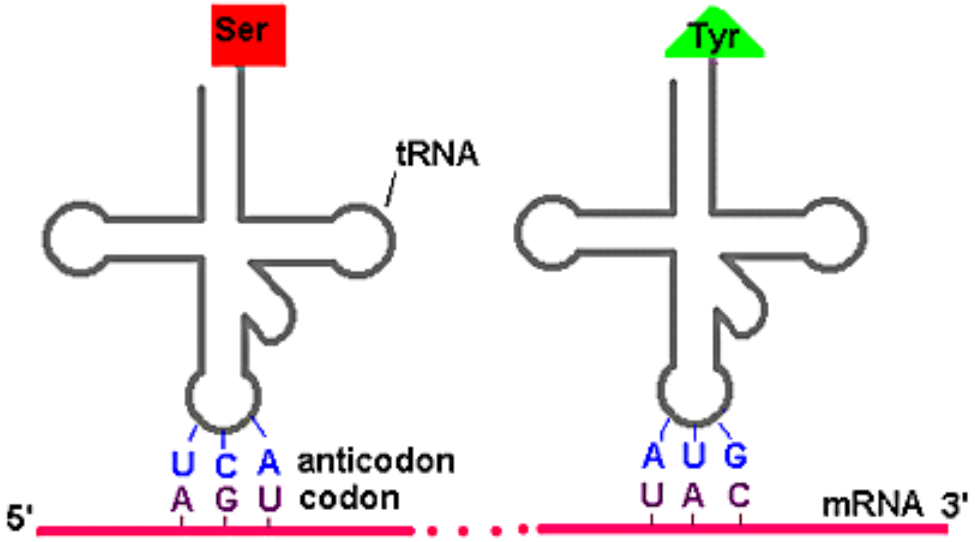


Figure 6–1 The basic genetic processes. The processes shown here are thought to occur in all present-day cells. Very early in the evolution of life, however, much simpler cells probably existed that lacked both DNA and proteins (see Figure 1–11). Note that a sequence of three nucleotides (a codon) in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

The Genetic Code



2nd base in codon

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

The Genetic Code

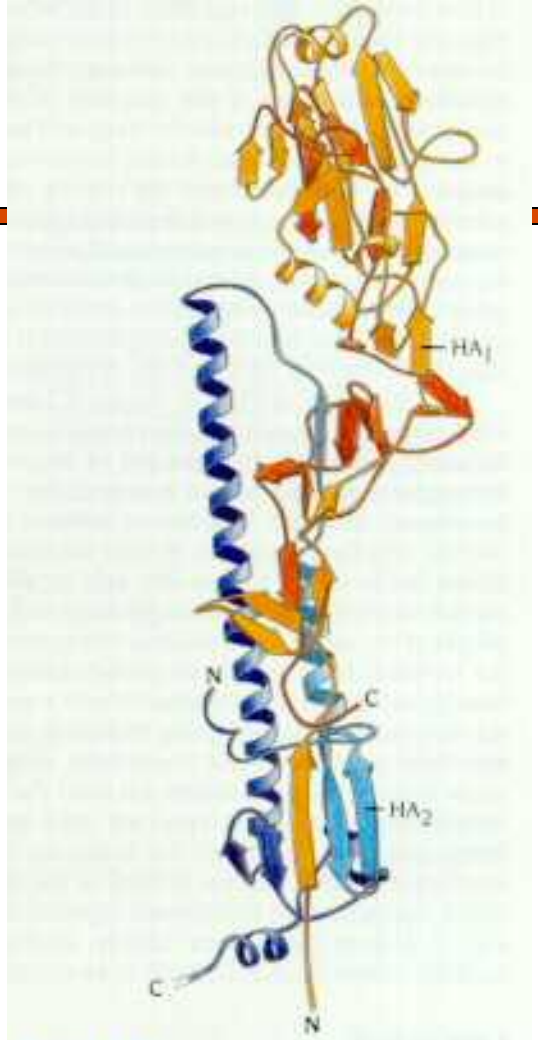
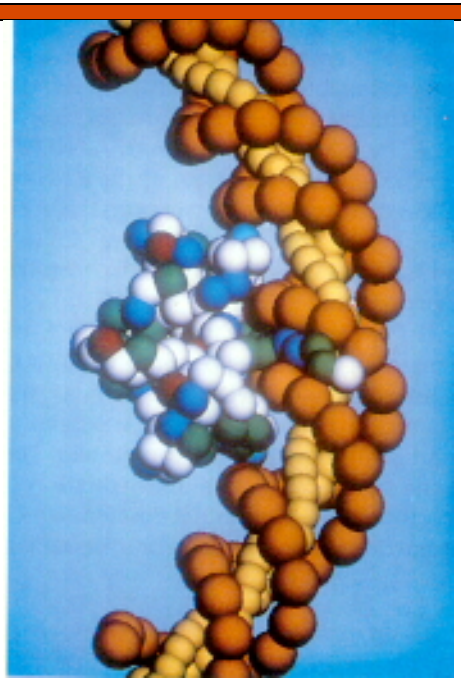
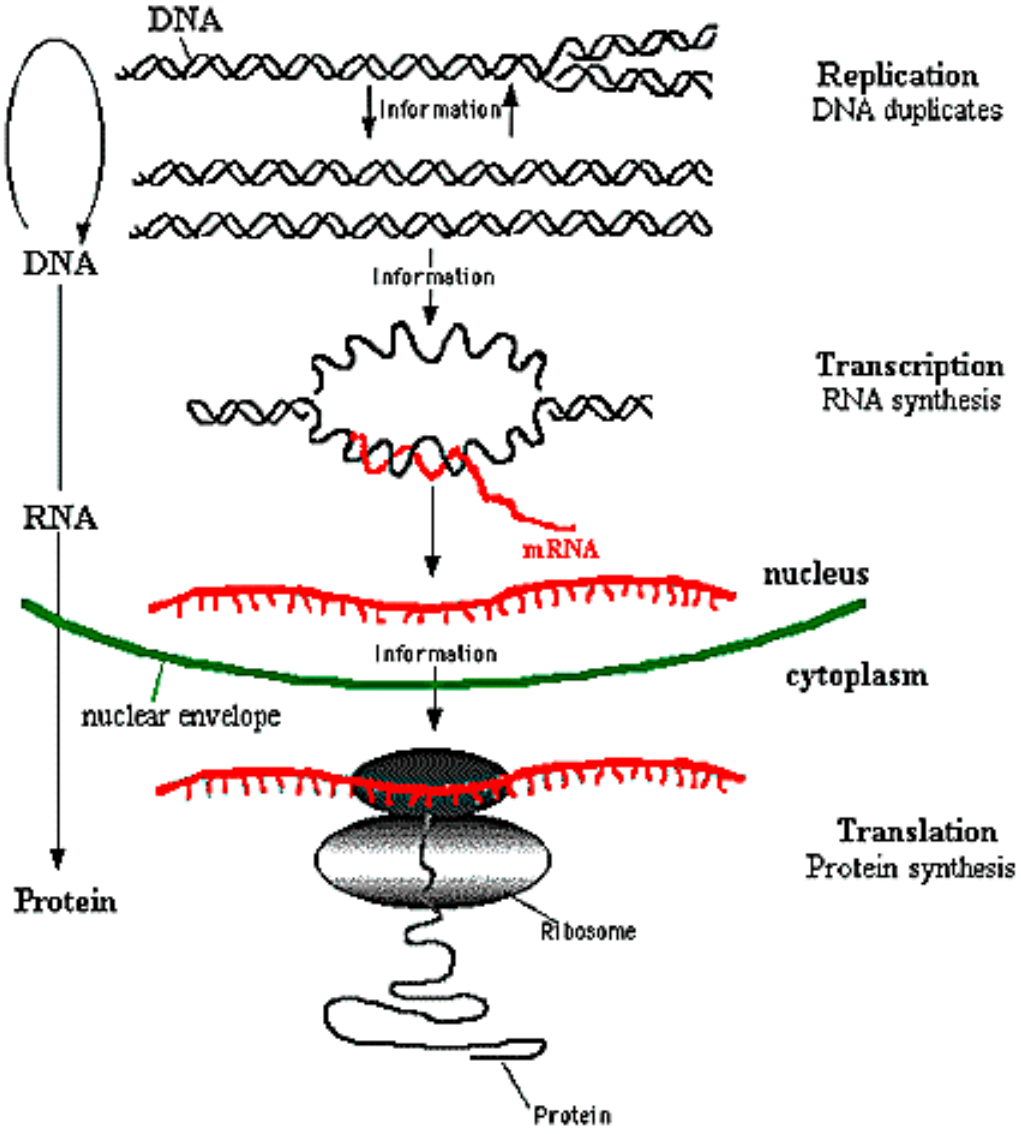


Figure 5.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75Å long. The globular head is formed by residues only from HA_1. (Courtesy of Don Wiley, Harvard University.)

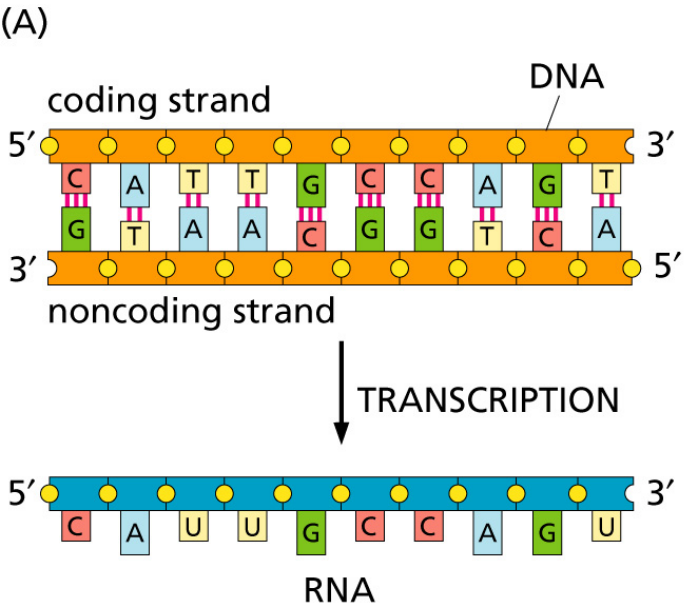


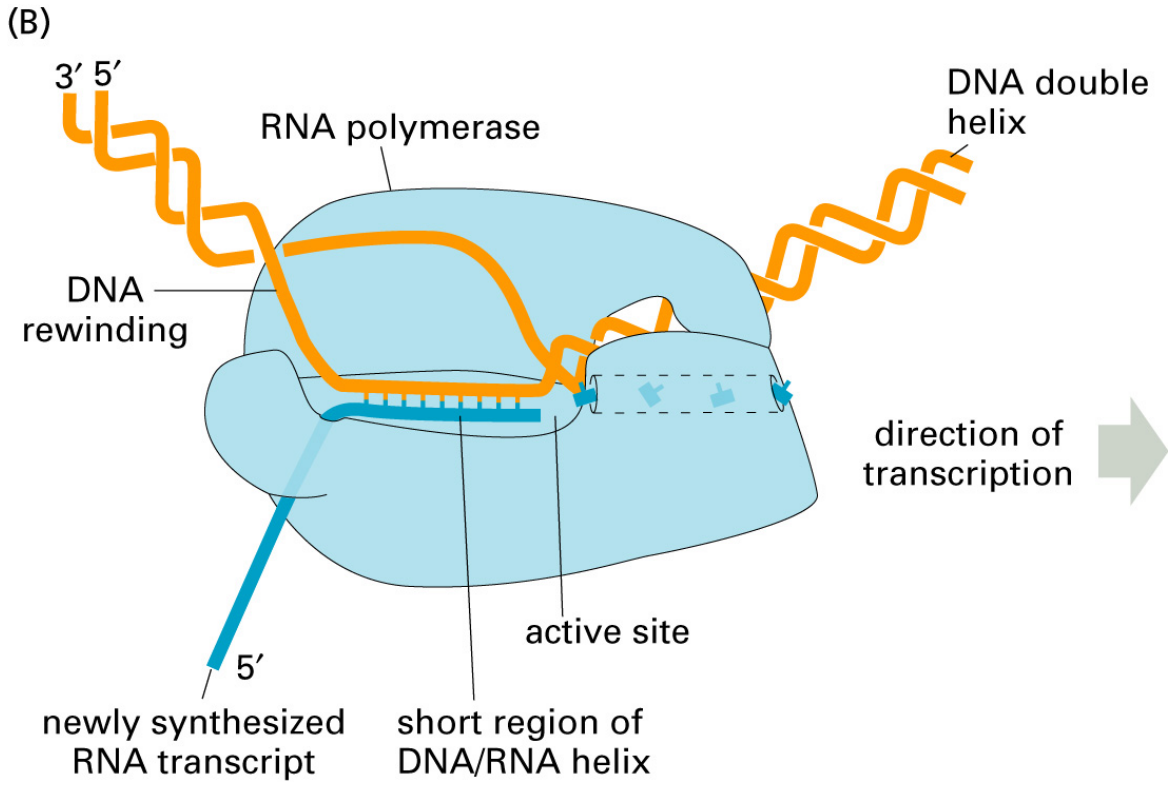


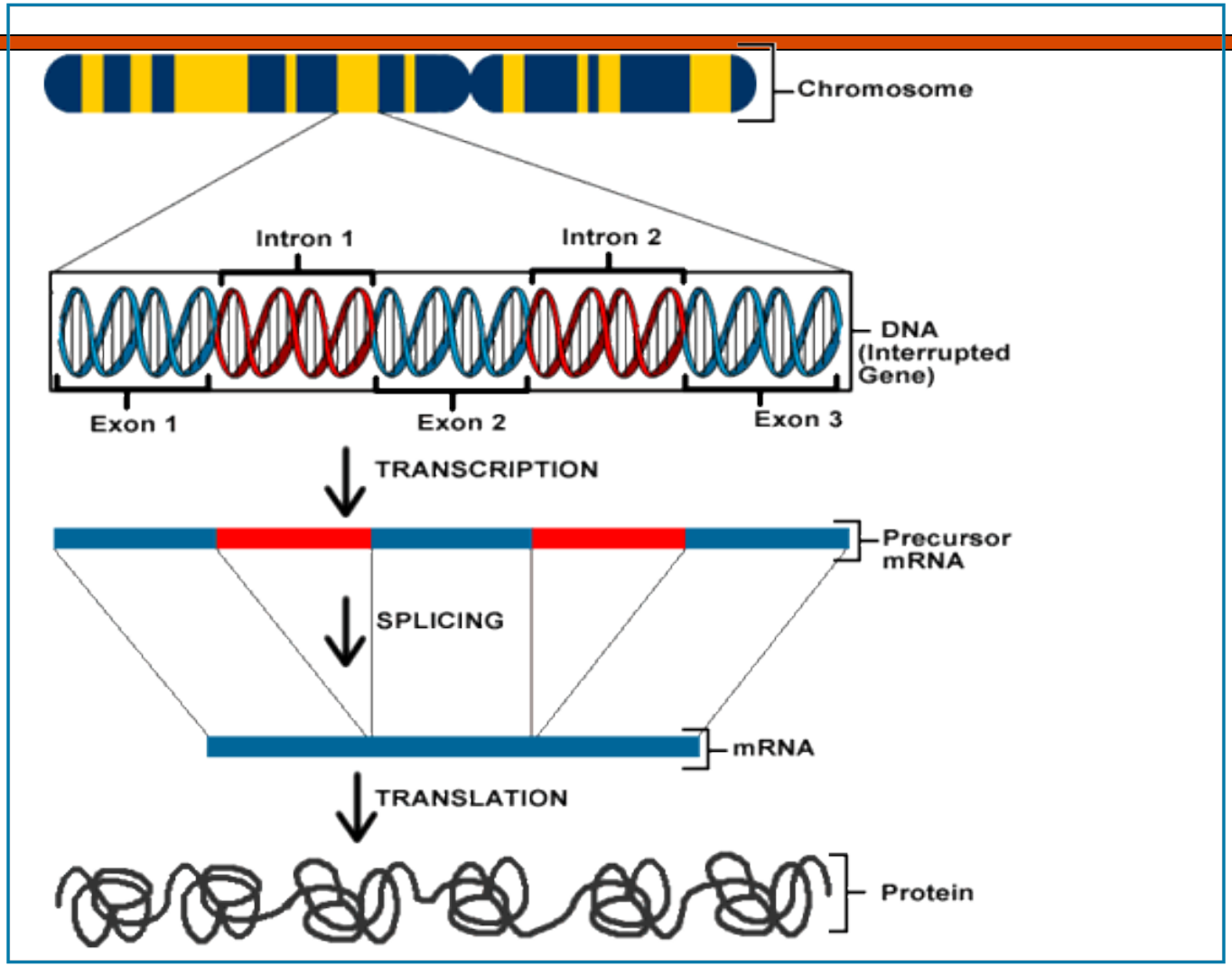
The Central Dogma of Molecular Biology

Transcription

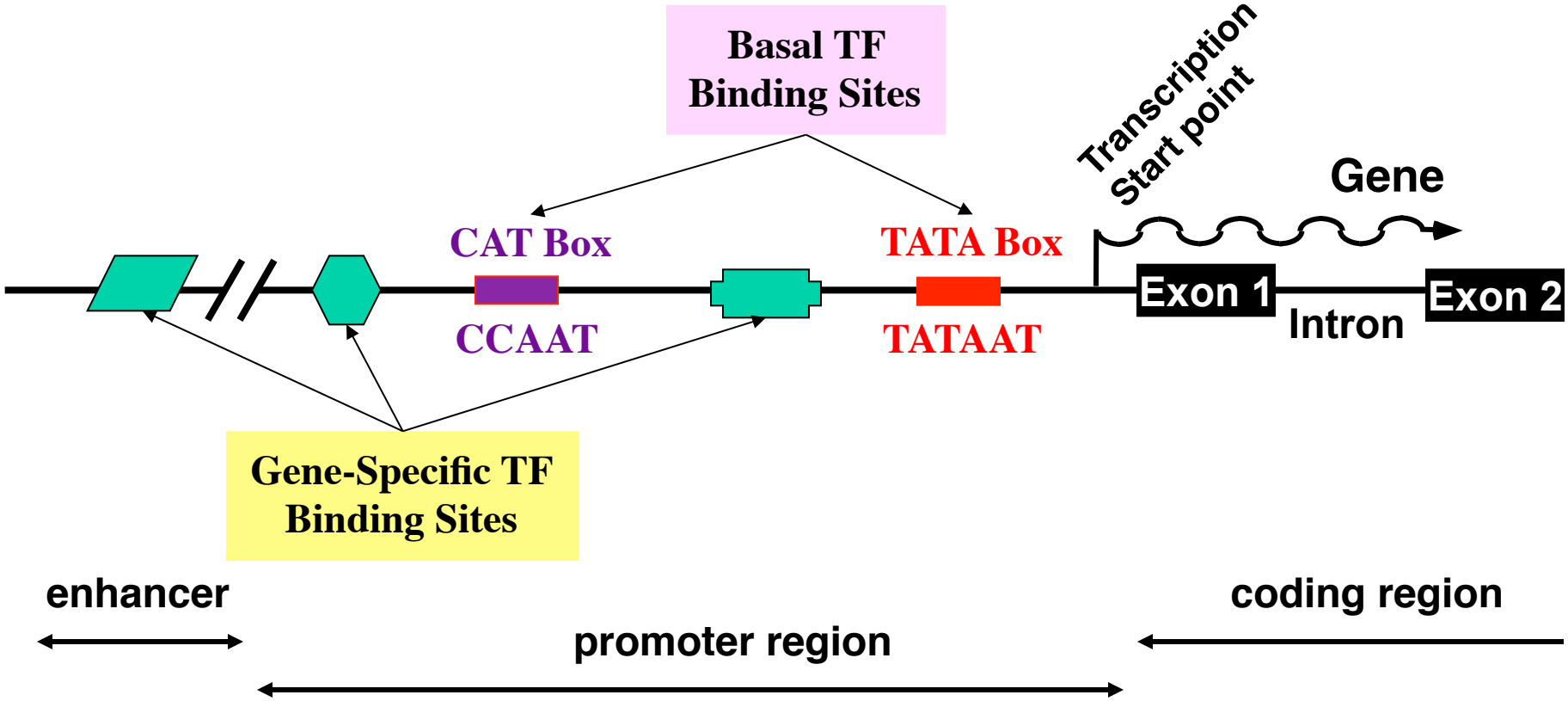
Fig 1.7, Zvelebil/Baum





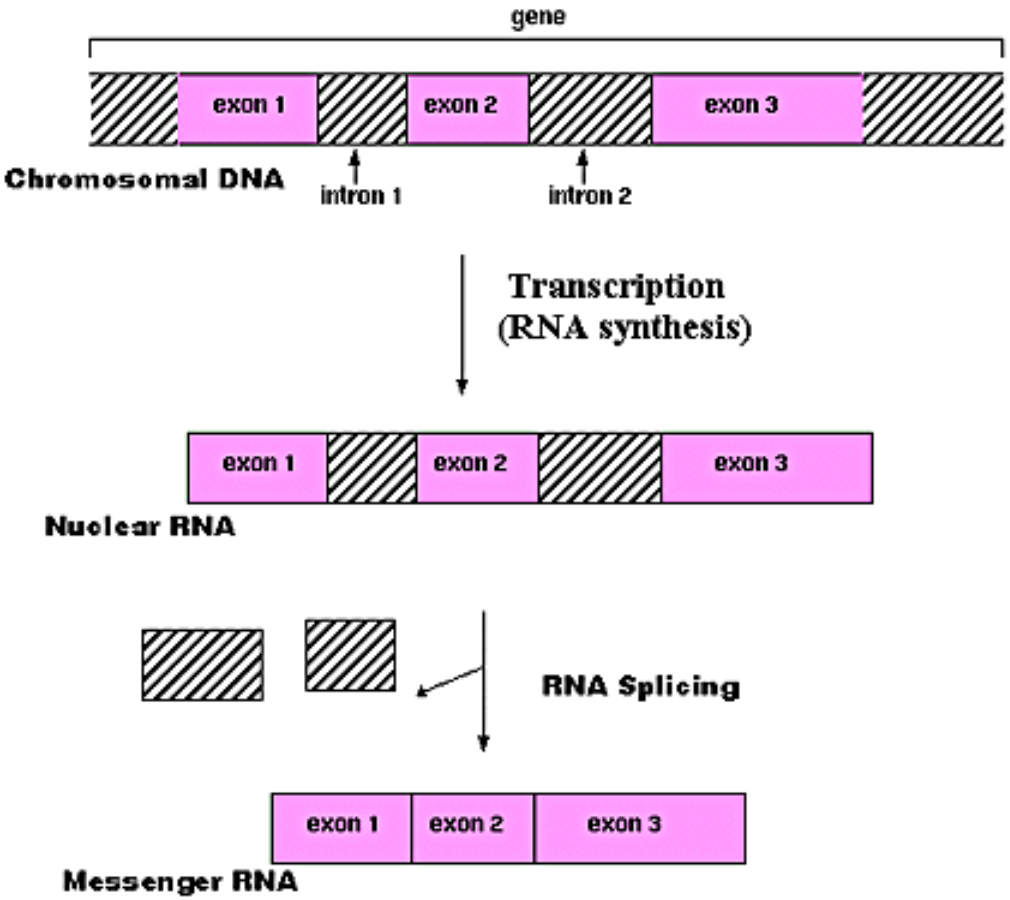

Chromosome
Intron 1
Intron 2
DNA
(Interrupted
Gene)
Exon 1
Exon 2
Exon 3
TRANSCRIPTION
Precursor
mRNA
SPLICING
mRNA
TRANSLATION
Protein

Transcription Regulation



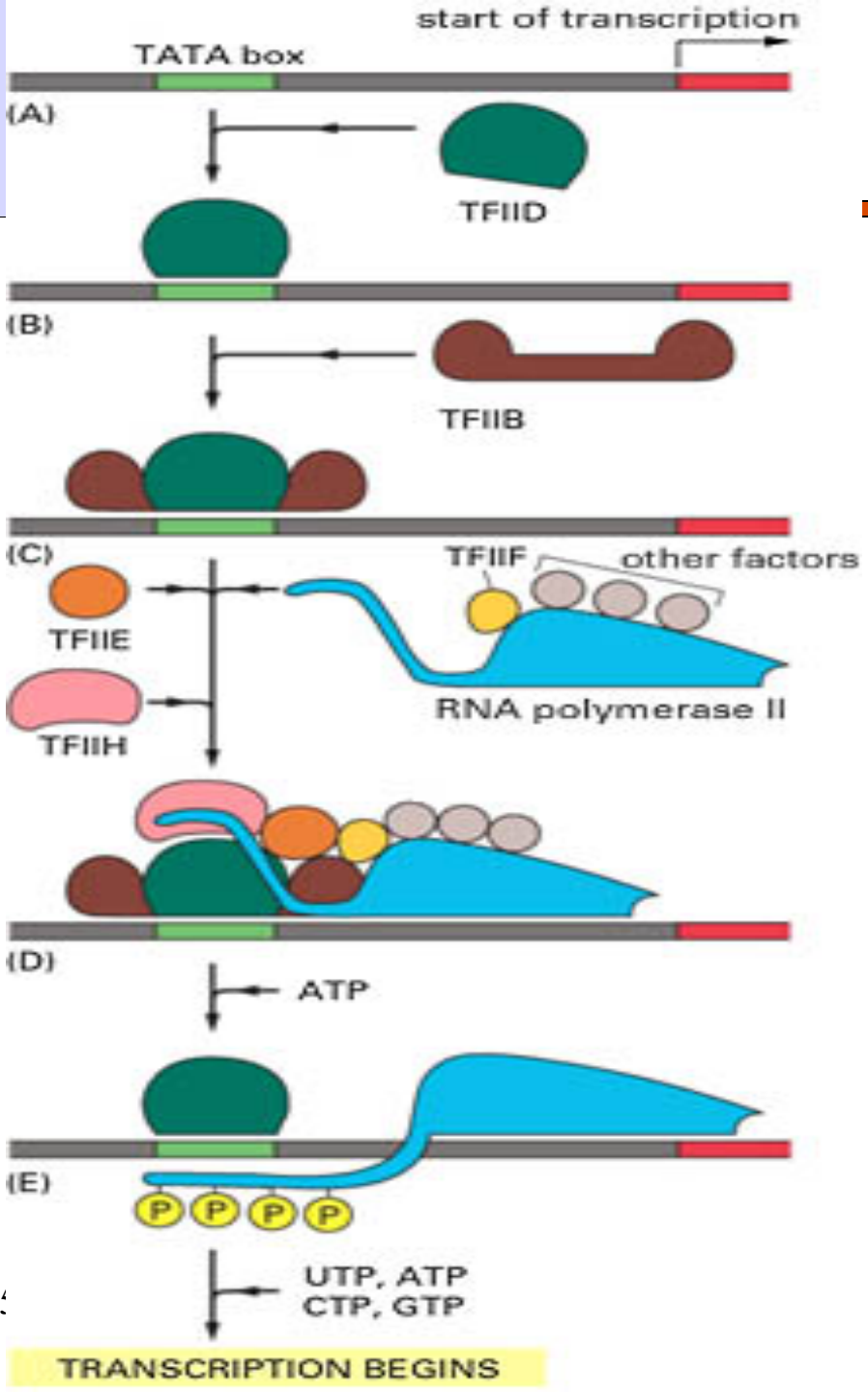
Courtesy: Kalai Mathee

DNA Transcription



RNA synthesis and processing

Transcription Initiation



Transcription

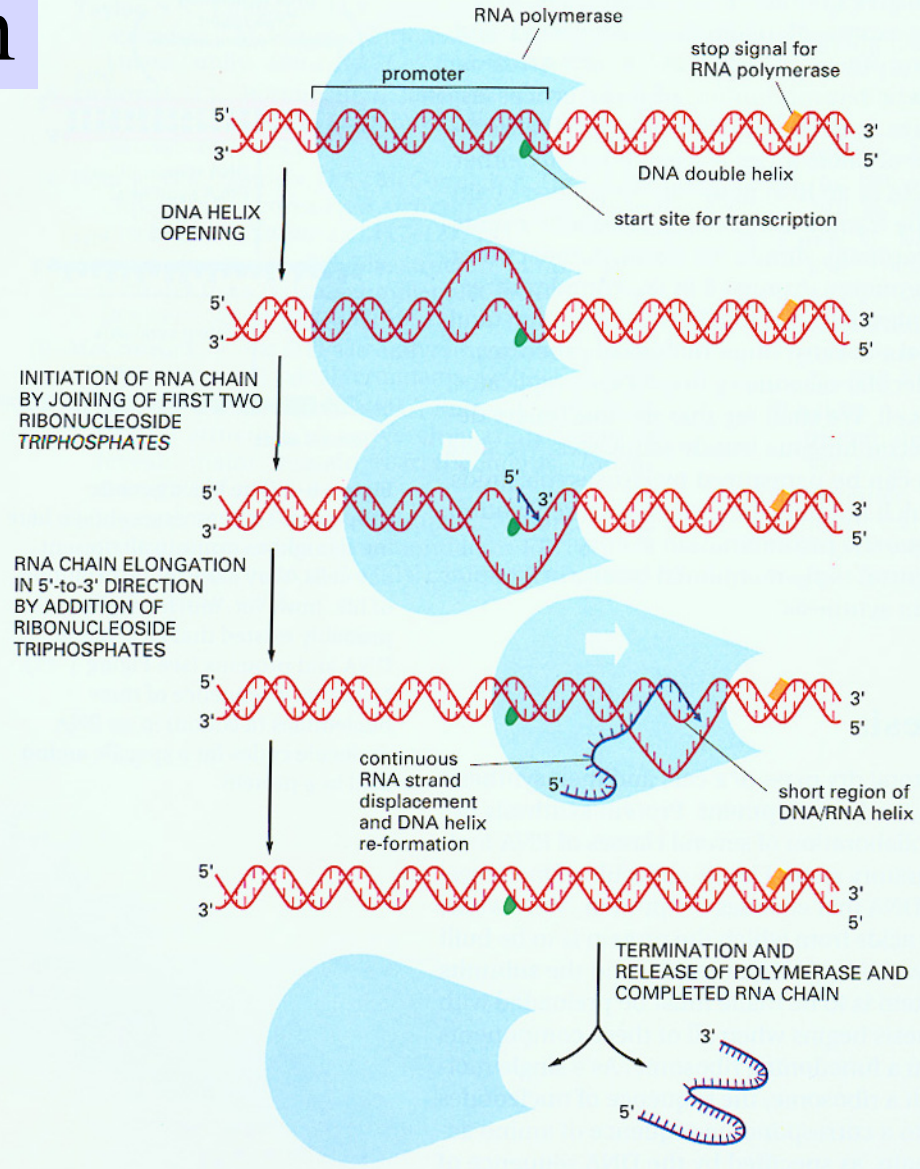
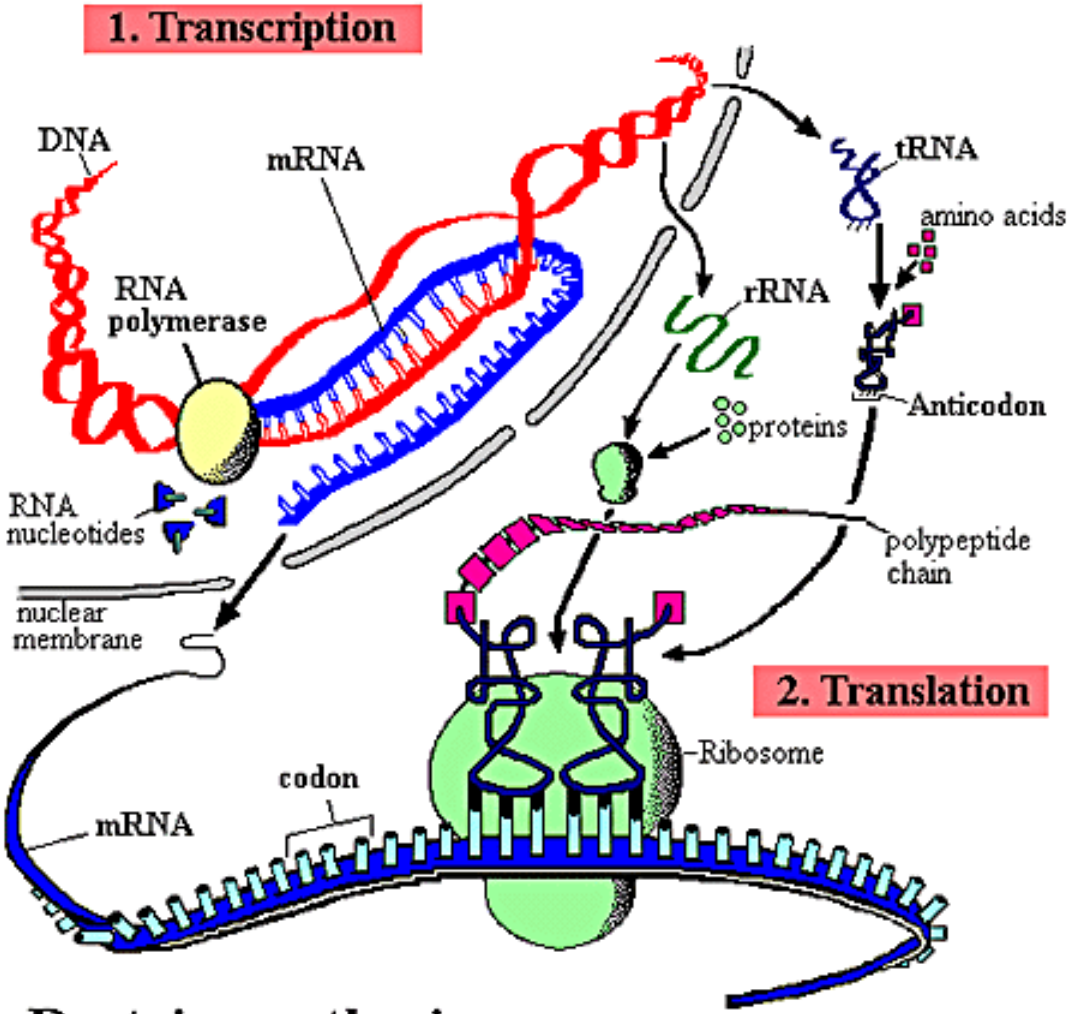


Figure 6–2 The synthesis of an RNA molecule by RNA polymerase. The enzyme binds to the promoter sequence on the DNA and begins its synthesis at a start site within the promoter. It completes its synthesis at a stop (termination) signal, whereupon both the polymerase and its completed RNA chain are released.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

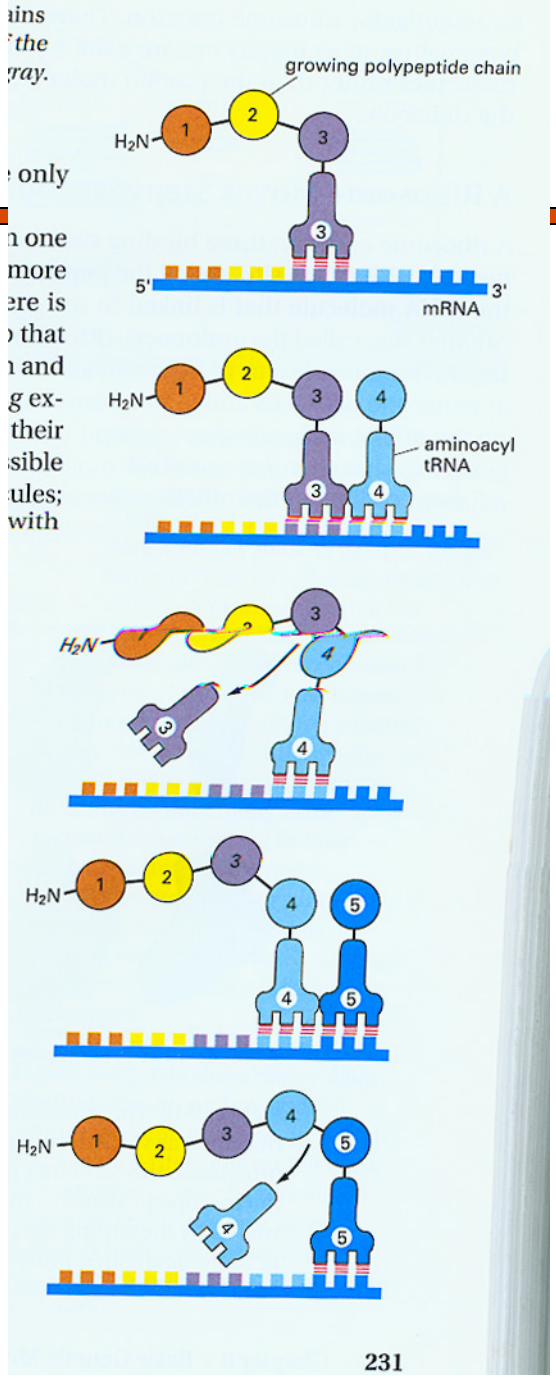
224 Chapter 6 : Basic Genetic Mechanisms

Protein Synthesis



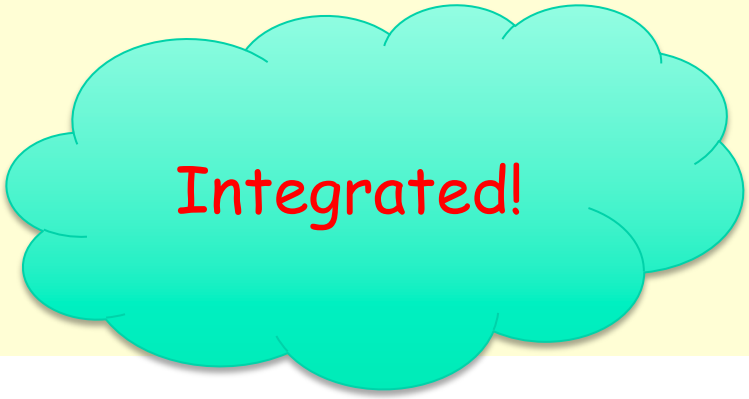
Protein Synthesis:

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; Bookshelf
- DNA and Protein Sequence database
- Protein structure database
- Genome assemblies
- BLAST
- SNP
- TaxBrowser
- Population study data sets
- PubChem (small mols)
- GEO (Gene Expression Omnibus)
- OMIM (Mendelian Inheritance in Man)

Youtube videos:
http://www.youtube.com/ncbinlm

Other critical databases

- PDB (http://www.wwpdb.org/)
- KEGG (http://www.genome.jp/kegg/)
- MetaCyc (http://metacyc.org)
- Reactome (http://www.reactome.org)
- ENCODE (http://encodeproject.org/ENCODE/ functional elements in human genome
- 1000 Genomes Project; Int'l HapMap Project
- Human Microbiome Project
- Human Epigenome Project
- Gene Ontology (GO)

Sequence Alignment

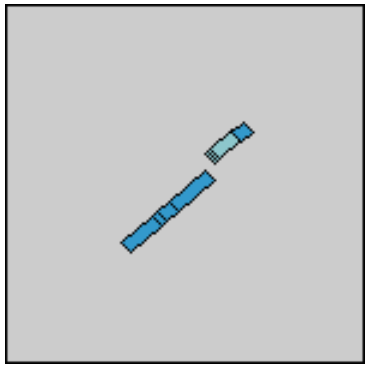
1. Can show sequences are close

rpoA [Pseudomonas aeruginosa] with rpoA [Pseudomonas fluorescence]

```
Query  1    MQISVNEFLTPRHIDVQVVSPTRAKITLEPLERGFGHTLGNALRRILLSSMPGCAVVEAE  60
            MQ SVNEFLTPRHIDVQVVS TRAKITLEPLERGFGHTLGNALRRILLSSMPGCAVVEAE
Sbjct  1    MQSSVNEFLTPRHIDVQVVSQTRAKITLEPLERGFGHTLGNALRRILLSSMPGCAVVEAE  60

Query  61   IDGVLHEYSAIEGVQEDVIEILLNLKGLAIKLHGRDEVTLTLSKKGSGVVTAADIQLDHD  120
            IDGVLHEYSAIEGVQEDVIEILLNLKGLAIKLHGRDEVTLTL+KKGSGVVTAADIQLDHD
Sbjct  61   IDGVLHEYSAIEGVQEDVIEILLNLKGLAIKLHGRDEVTLTLAKKGSGVVTAADIQLDHD  120

Query  121  VEIVNPDHVIANLASNGALNMKLTVARGRGYEPADSRQSDEDESRSIGRLQLDSSFSPVR  180
            VEI+N DHVIANLA NGALNMKL VARGRGYEPAD+RQSDEDESRSIGRLQLD+SFSPVR
Sbjct  121  VEIINGDHVIANLADNGALNMKLKVARGRGYEPADARQSDEDESRSIGRLQLDASFSPVR  180

Query  181  RIAYVVENARVEQRTNLDKLVIDLETNGTLDPEEAIRRAATILQQQLAAFVDLKGDSEPV  240
            R++YVVENARVEQRTNLDKLV+DLETNGTLDPEEAIRRAATILQQQLAAFVDLKGDSEPV
Sbjct  181  RVSYVVENARVEQRTNLDKLVLDLETNGTLDPEEAIRRAATILQQQLAAFVDLKGDSEPV  240

Query  241  VIEQEDEIDPILLRPVDDLELTVRSANCLKAENIYYIGDLIQRTEVELLKTPNLGKKSLT  300
            V EQEDEIDPILLRPVDDLELTVRSANCLKAENIYYIGDLIQRTEVELLKTPNLGKKSLT
Sbjct  241  VEEQEDEIDPILLRPVDDLELTVRSANCLKAENIYYIGDLIQRTEVELLKTPNLGKKSLT  300

Query  301  EIKDVLASRGLSLGMRLDNWPPASLKKDDKATA  333
            EIKDVLASRGLSLGMRLDNWPPASLKKDDKATA
Sbjct  301  EIKDVLASRGLSLGMRLDNWPPASLKKDDKATA  333
```

2. Can show sequences have similar parts

```
Sequence 1 gi 332624 Simian sarcoma virus v-sis transforming
protein p28 gene, complete cds; and 3' LTR long terminal repeat,
complete sequence. Length 2984 (1 .. 2984)
Sequence 2 gi 4505680 Homo sapiens platelet-derived growth factor
beta polypeptide (simian sarcoma viral (v-sis) oncogene homolog)
(PDGFB), transcript variant 1, mRNA Length 3373 (1 .. 3373)
```



3. Can identify similar sequences from DB

V-sis Oncogene – Homologies

Sequences producing significant alignments:	Score (bits)	E Value
gi|332623|gb|J02396.1|SEG_SSVPCS2 Simian sarcoma virus v-si...	4591	0.0
gi|61774|emb|V01201.1|RESSV1 Simian sarcoma virus proviral ...	4504	0.0
gi|332622|gb|J02395.1|SEG_SSVPCS1 Simian sarcoma virus LTR ...	1283	0.0
gi|885929|gb|U20589.1|GLU20589 Gibbon leukemia virus envelo...	1140	0.0
gi|4505680|ref|NM_002608.1| Homo sapiens platelet-derived g...	954	0.0
gi|20987438|gb|BC029822.1| Homo sapiens, platelet-derived g...	954	0.0
gi|338210|gb|M12783.1|HUMSISPDG Human c-sis/platelet-derive...	954	0.0

4. Can pinpoint mutations

870GTGGCTGCTTCTTTGGTTGTGCTGTGGCTCCTTGGAAA

X

870GTGGCTGCTTCTTTGGTTGTGCTGTAGCTCCTTGGAAA

5. Can be basis for discoveries

- Early 1970s: Simian sarcoma virus causes cancer in some species of monkeys.
- 1970s: infection by certain viruses cause some cells in culture (in vitro) to grow without bounds.
  - Hypothesis: Certain genes (oncogenes) in viruses encode cellular growth factors, which are proteins needed to stimulate growth of a cell colony. Thus uncontrolled quantities of growth factors produced by the infected cells cause cancer-like behavior.
- 1983:
  - The oncogene from SSV called v-sis was isolated and sequenced.
  - The partial amino-acid sequence for platelet-derived growth factor (PDGF) was sequenced and published. It stimulates the proliferation of normal cells.
  - R.F. Doolittle was maintaining one of the earliest home-grown databases of published amino-acid sequences.
  - Sequence Alignment of v-sis and PDGF showed something surprising.

PDGF and v-sis

- One region of 31 amino acids had 26 exact matches
- Another region of 39 residues had 35 exact matches.
- Conclusion:
  - The previously harmless virus incorporates the normal growth-related gene (proto-oncogene) of its host into its genome.
  - The gene gets mutated in the virus, or moves closer to a strong enhancer, or moves away from a repressor.
  - This causes an uncontrolled amount of the product (the growth factor, for example) when the virus infects a cell.
- Several other oncogenes known to be similar to growth-regulating proteins in normal cells.

Sequence Alignment

```
>gi|4505680|ref|NM_002608.1|    Homo sapiens platelet-derived growth
   factor beta polypeptide (simian sarcoma viral (v-sis) oncogene
   homolog) (PDGFB), transcript variant 1, mRNA Length = 3373 Score = 954
   bits (481), Expect = 0.0 Identities = 634/681 (93%), Gaps = 3/681 (0%)
   Strand = Plus / Plus
Query: 1015 agggggaccccattcctgaggagctctataagatgctgagtggccactcgattcgctcct 1074
           |||||||||||||||| |||||||| ||| |||||||||||| ||||||||| |||||||
Sbjct: 1084 agggggaccccattcccgaggagctttatgagatgctgagtgaccactcgatccgctcct 1143
       > 21 E  G  D  P  I  P  E  E  L  Y  E  M  L  S  D  H  S  I  R  S
Query: 1075 tcgatgacctccagcgcctgctgcagggagactccggaaaagaagatggggctgagctgg 1134
           | ||||| ||||| |||||||||| |||||| ||||| | |||||||||| ||| |||
Sbjct: 1144 ttgatgatctccaacgcctgctgcacggagaccccggagaggaagatggggccgagttgg 1203
      > 61 D  L  N  M  T  R  S  H  S  G  G  E  L  E  S  L  A  R  G  R
```

6. Can help describe motifs, domains, and families of sequences

- Family alignment for the ITAM domain (Immunoreceptor tyrosine-based activation motif)

```
CD3D_MOUSE/1-2   EQLYQPLRDR EDTQ-YSRLG GN
Q90768/1-21      DQLYQPLGER NDGQ-YSQLA TA
CD3G_SHEEP/1-2   DQLYQPLKER EDDQ-YSHLR KK
P79951/1-21      NDLYQPLGQR SEDT-YSHLN SR
FCEG_CAVPO/1-2   DGIYTGLSTR NQET-YETLK HE
CD3Z_HUMAN/3-0   DGLYQGLSTA TKDT-YDALH MQ
C79A_BOVIN/1-2   ENLYEGLNLD DCSM-YEDIS RG
C79B_MOUSE/1-2   DHTYEGLNID QTAT-YEDIV TL
CD3H_MOUSE/1-2   NQLYNELNLG RREE-YDVLE KK
CD3Z_SHEEP/1-2   NPVYNELNVG RREE-YAVLD RR
CD3E_HUMAN/1-2   NPDYEPIRKG QRDL-YSGLN QR
CD3H_MOUSE/2-0   EGVYNALQKD KMAEAYSEIG TK
Consensus/60%    -.lYpsLspc pcsp.YspLs pp
```

Simple Modular Architecture Research Tool

Implications of Sequence Alignment

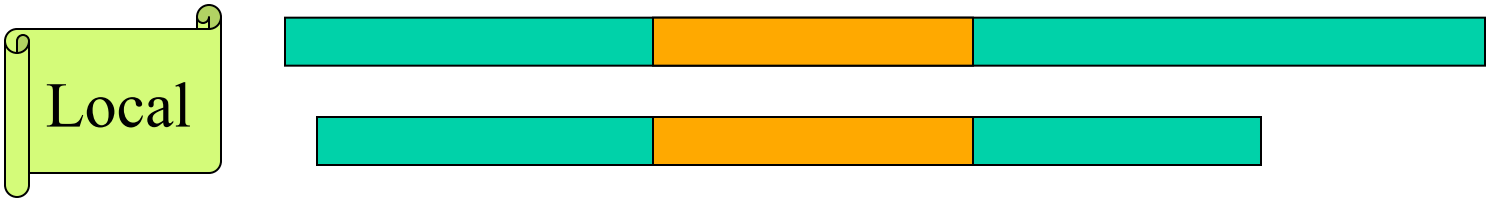
- Mutation in DNA is a natural evolutionary process. Thus sequence similarity may indicate common ancestry.
- In biomolecular sequences (DNA, RNA, protein), high sequence similarity implies significant structural and/or functional similarity.

Similarity vs. Homology

- Homologous sequences share common ancestry.
- Similar sequences are "near" to each other by some appropriately defined measurable criteria.

Types of Sequence Alignments - 1

- Global Alignment: similarity over entire length



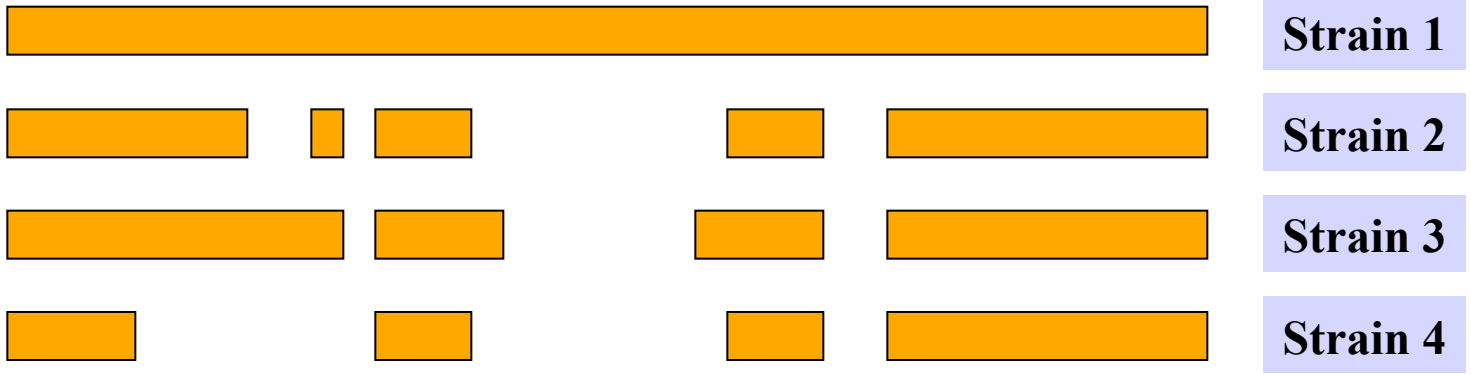
- Local Alignment: no overall similarity, but some segment(s) is/are similar

Types of Sequence Alignments - 2

Semi-Global

- Semi-global Alignment: end segments may not be similar

Multiple



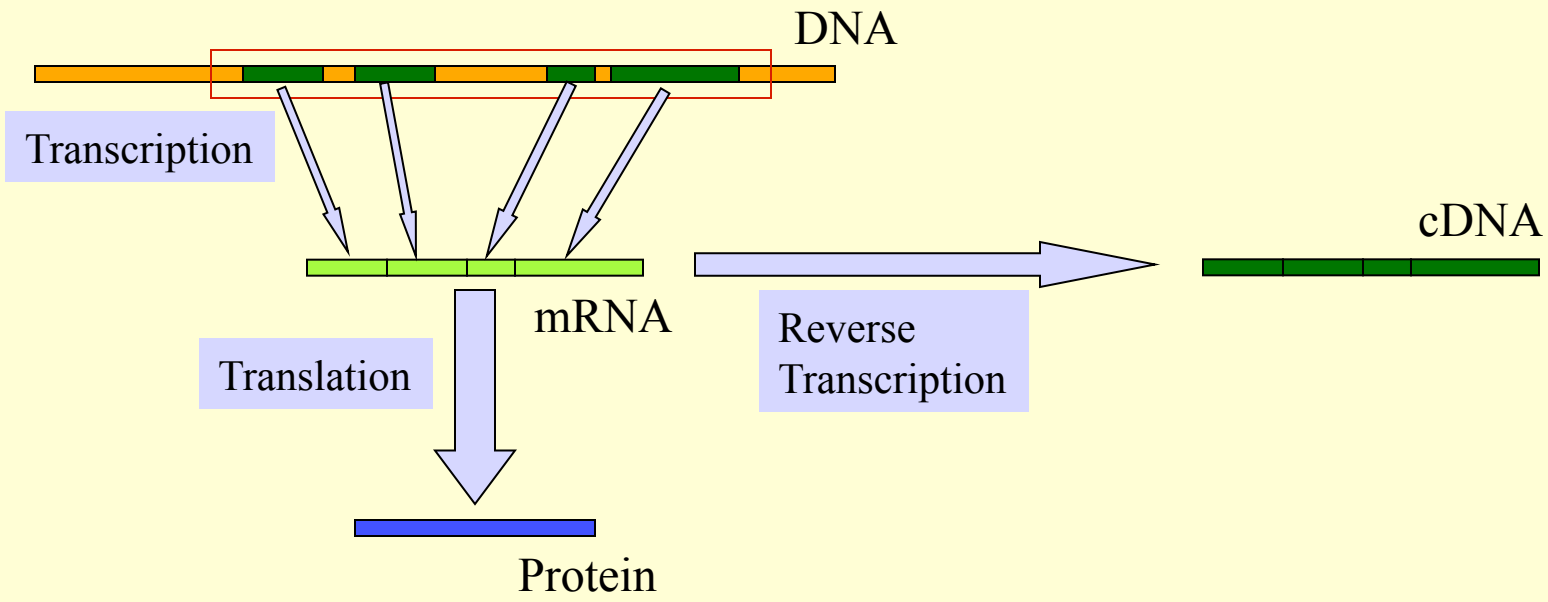
- Multiple Alignment: similarity between sets of sequences

Sequence Alignment

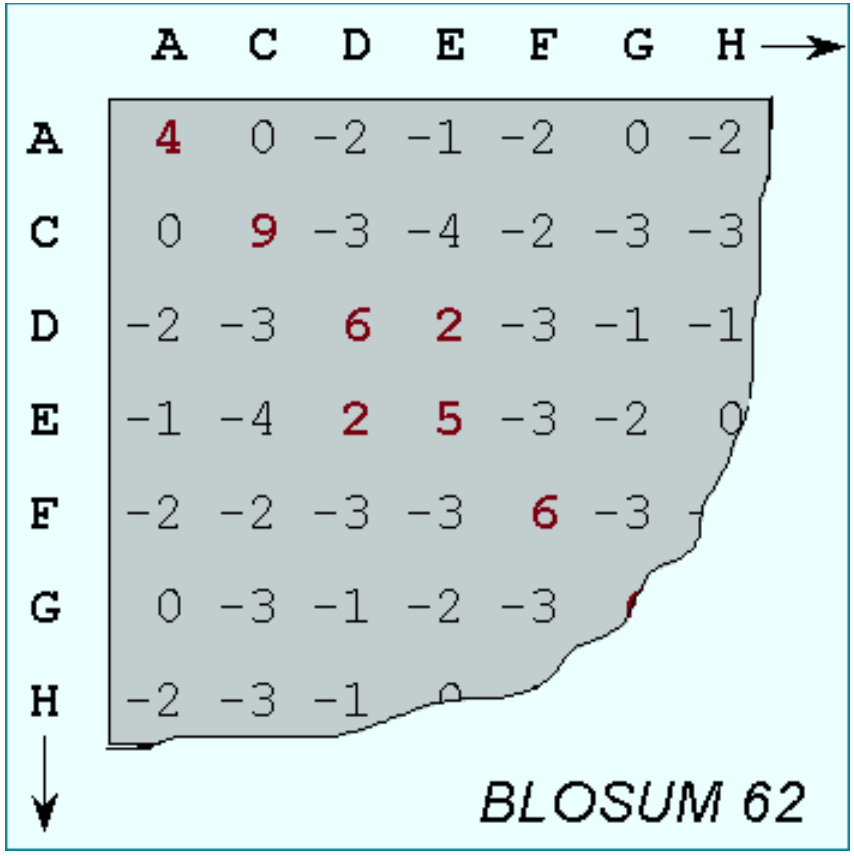
- Global:
  - Needleman-Wunsch-Sellers (1970).
- Local:
  - Smith-Waterman (1981)
  - Useful when commonality is small and global alignment is meaningless. Often unaligned portions "mask" short stretches of aligned portions. Example: comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.
- Dynamic Programming (DP) based.

Why gaps?

- Example: Finding the gene site for a given (eukaryotic) cDNA requires "gaps".
- What is cDNA? cDNA = Copy DNA



How to score mismatches?



BLAST & FASTA

- FASTA

  [Lipman, Pearson '85, '88]

- Basic Local Alignment Search Tool

  [Altschul, Gish, Miller, Myers, Lipman '90]

BLAST Overview

- Program(s) to search all sequence databases
- Tremendous Speed/Less Sensitive
- Statistical Significance reported
- WWWBLAST, QBLAST (send now, retrieve results later), Standalone BLAST, BLASTcl3 (Client version, TCP/IP connection to NCBI server), BLAST URLAPI (to access QBLAST, no local client)

BLAST

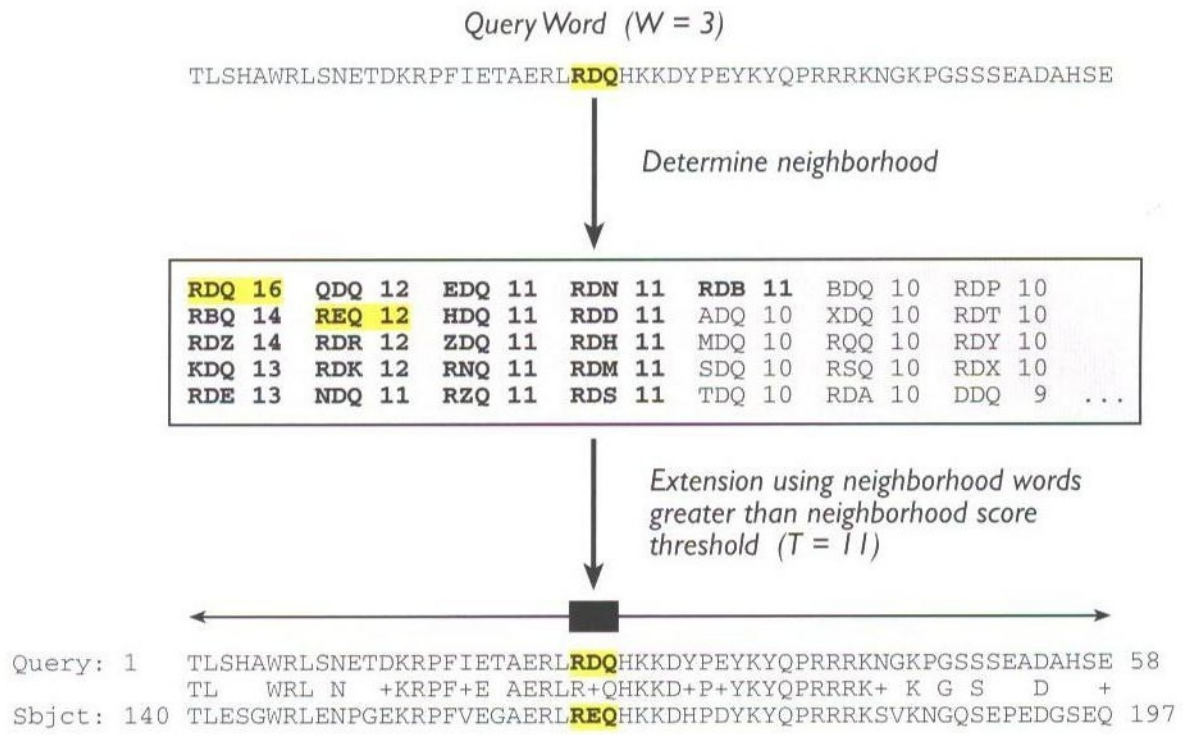


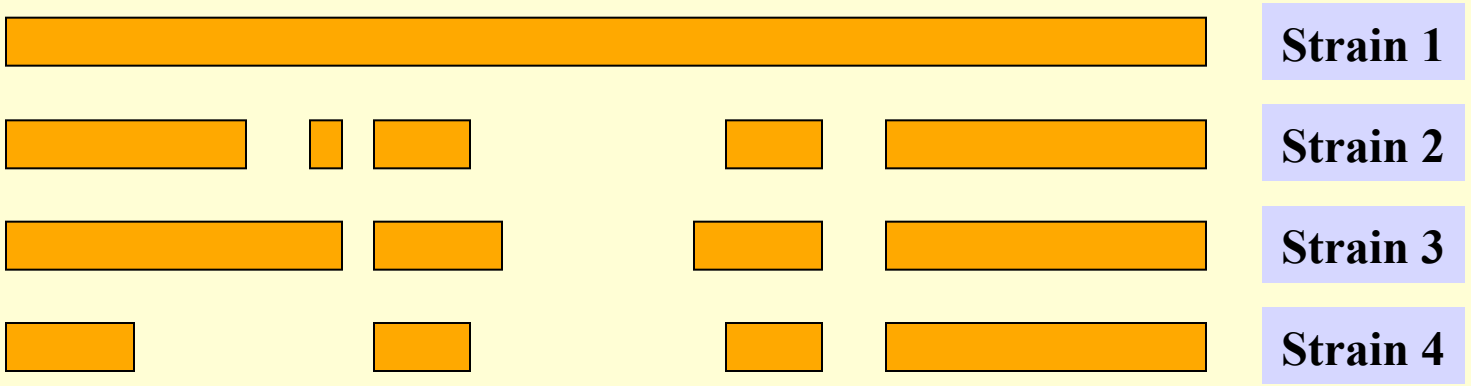
FIGURE 11.7 The initiation of a **BLAST** search. The search begins with query words of a given length (here, three amino acids) being compared against a scoring matrix to determine additional three-letter words "in the neighborhood" of the original query word. Any occurrences of these neighborhood words in sequences within the target database then are investigated. See text for details.

BLAST Strategy & Improvements

- Lipman et al.: speeded up finding "runs" of "hot spots".
- Eugene Myers '94: "Sublinear algorithm for approximate keyword matching".
- Karlin, Altschul, Dembo '90, '91: "Statistical Significance of Matches"

Why Gaps?

- Example: Aligning HIV sequences.



BLAST Variants

- **Nucleotide BLAST**
  - **Standard blastn**
  - **MEGABLAST** (Compare large sets, Near-exact searches)
  - **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering)
- **Protein BLAST**
  - **Standard blastp**
  - **PSI-BLAST** (Position Specific Iterated BLAST)
  - **PHI-BLAST** (Pattern Hit Initiated BLAST; reg expr. Or Motif search)
  - **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering, PAM-30)
- **Translating BLAST**
  - **Blastx**: Search nucleotide sequence in protein database (6 reading frames)
  - **Tblastn**: Search protein sequence in nucleotide dB
  - **Tblastx**: Search nucleotide seq (6 frames) in nucleotide DB (6 frames)

BLAST Cont'd

- **RPS BLAST**
  - Compare protein sequence against Conserved Domain DB; Helps in predicting rough structure and function
- **Pairwise BLAST**
  - blastp (2 Proteins), blastn (2 nucleotides), tblastn (protein-nucleotide w/ 6 frames), blastx (nucleotide-protein), tblastx (nucleotide w/6 frames-nucleotide w/ 6 frames)
- **Specialized BLAST**
  - Human & Other finished/unfinished genomes
  - *P. falciparum*: Search ESTs, STSs, GSSs, HTGs
  - VecScreen: screen for contamination while sequencing
  - IgBLAST: Immunoglobin sequence database

BLAST Credits

- Stephen Altschul
- Jonathan Epstein
- David Lipman
- Tom Madden
- Scott McGinnis
- Jim Ostell
- Alex Schaffer
- Sergei Shavirin
- Heidi Sofia
- Jinghui Zhang

Databases used by BLAST

- **Protein**
  - nr (everything), swissprot, pdb, alu, individual genomes
- **Nucleotide**
  - nr, dbest, dbsts, htgs (unfinished genomic sequences), gss, pdb, vector, mito, alu, epd
- **Misc**

BLAST Parameters and Output

- Type of sequence, nucleotide/protein
- Word size, w
- Gap penalties, p_1 and p_2
- Neighborhood Threshold Score, T
- Score Threshold, S
- E-value Cutoff, E
- Number of hits to display, H
- Database to search, D
- Scoring Matrix, M
- Score s and E-value e
  - E-value e is the expected number of sequences that would have an alignment score greater than the current score s.

Scoring Matrix to Use

- PAM 40 — Short alignments with high similarity (70-90%)
- PAM 160 — Members of a protein family (50-60%)
- PAM 250 — Longer alignments (divergent sequences) (~30%)

- BLOSUM90 — Short alignments with high similarity (70-90%)
- BLOSUM80 — Members of a protein family (50-60%)
- BLOSUM62 — Finding all potential hits (30-40%)
- BLOSUM30 — Longer alignments (divergent sequences) (<30%)

Rules of Thumb

- Most sequences with significant similarity over their entire lengths are homologous.
- Matches that are > 50% identical in a 20-40 aa region occur frequently by chance.
- Distantly related homologs may lack significant similarity. Homologous sequences may have few absolutely conserved residues.
- A homologous to B & B to C $\Rightarrow$ A homologous to C.
- Low complexity regions, transmembrane regions and coiled-coil regions frequently display significant similarity without homology.
- Greater evolutionary distance implies that length of a local alignment required to achieve a statistically significant score also increases.

Rules of Thumb

- Results of searches using different scoring systems may be compared directly using normalized scores.
- If S is the (raw) score for a local alignment, the **normalized** score S' (in bits) is given by

$$S' = \frac{\lambda - \ln(K)}{\ln(2)}$$

The parameters depend on the scoring system.

- **Statistically significant normalized score**,

$$S' > \log\left(\frac{N}{E}\right)$$

where E-value = E, and N = size of search space.