BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan

ECS 254A; Phone: x3748

giri@cs.fiu.edu

http://www.cs.fiu.edu/~giri/teach/BSC4934_Su11.html

July 2011

Overview of Course

- Sequence Alignment; Multiple Sequence Alignment
- Sequence Analysis
- Sequencing
- Phylogenetic Analysis
- Gene prediction techniques
- Pattern discovery techniques
- Protein structure alignment and analysis
- Genomics, Functional Genomics, Proteomics
- Gene Expression Data Analysis
- RNA Secondary structure
- RNA interference and small RNA
- Databases & Software Packages
- Statistics for Bioinformatics
- Computational Learning & Predictive Methods
- Emerging Biotechnologies

Software Packages

- Databases (GenBank, SwissPROT)
- Programming Environments (BioPerl)
- Sequence Alignment (BLAST, CLUSTALW)
- Phylogenetic Analysis (CLUSTALW)
- Learning Methods (HMMer)
- Pattern Discovery Techniques (GYM, TEIRESIAS)
- Molecular Structure Analysis (DALI, RASMOL, SPDBV)
- Microarray Analysis (CLUSTER, GeneCluster, TreeView)
- Statistical Software Packages (SAS, R)

Genomic Databases

- Entrez Portal at National Center for Biotechnology Information (NCBI) gives access to:
  - Nucleotide (GenBank, EMBL, DDBJ)
  - Protein (PIR, SwissPROT, PRF, and Protein Data Bank or PDB)
  - Genome
  - Structure
  - 3D Domains
  - Conserved Domains
  - Gene; UniGene; HomoloGene; SNP
  - GEO Profiles & Datasets
  - Cancer Chromosomes
  - PubMed Central; Journals; Books
  - OMIM
  - Database Neighbors and Interlinking

Evaluation

- Homework Assignments (35 %)
- Exam (35 %)
- Semester Project (25 %)
- Class Participation (5 %)

Course Homepage

http://www.cis.fiu.edu/~giri/teach/BSC4934_Su11.html

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

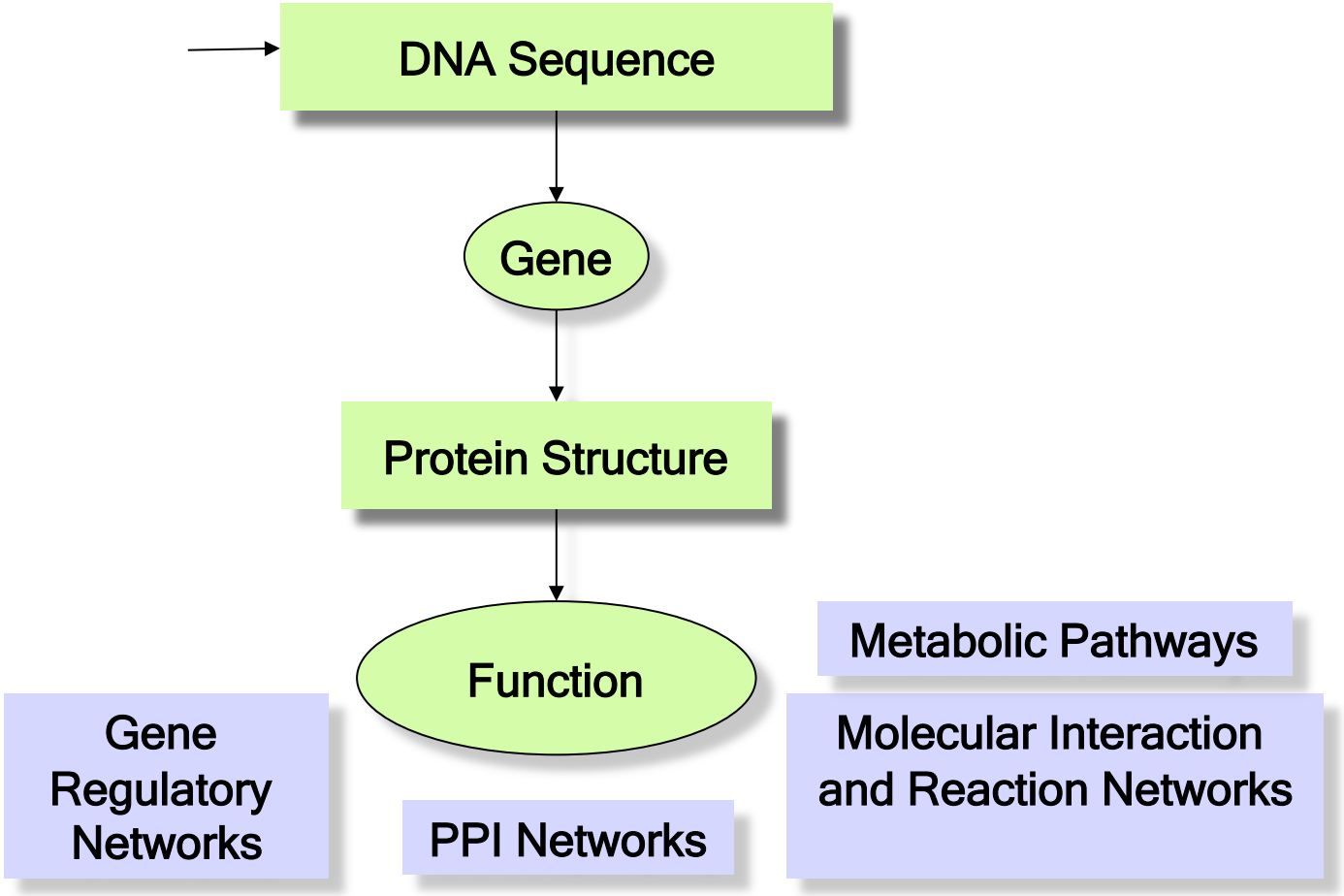
History

- What major discovery was made in 1953?
- What major world event took place on 26 June, 2000?
- 1975: Sanger Sequencing
- 1977: first bacteriophage sequenced
- 1990: HGP initiated

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



Perspective of Bioinformatics

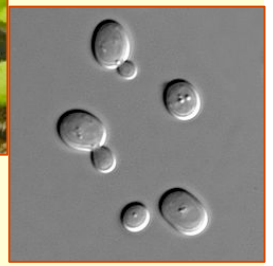
- Study of the cell: DNA, genes, proteins
- Study of the organism: genome, changes over time, over body regions, or over physiological or pathological states
- Study of all life: Tree of Life, Phylogeny, Variations, comparative genomics

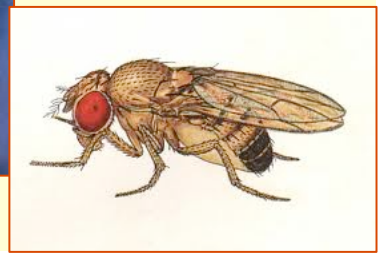
General Information

- GenBank Release 163/178/184 (Dec 2007/Aug 2009/Jun 2011) contains over 80/108/140 million sequence entries totaling over 83/106/129 Gb from over 2,500/12,000/15,000 organisms [http://www.ncbi.nlm.nih.gov] (Storage: ~150/476/506 GB uncompressed)
- Human Genome has ~3 billion bp with 32,000+ genes.
- 1673 complete bacterial (112 archaeal) genomes sequenced
- 3872 Viral genomes (300bp - 300Kb) (1st 1978: Simian virus; 5Kb).
- UniProtKB/Swiss-Prot Release 54.7/2010_06 (Jan'08/ May'10): 333K/517K entries; 120/182 million amino acids.

41 Complete Eukaryotic Genomes

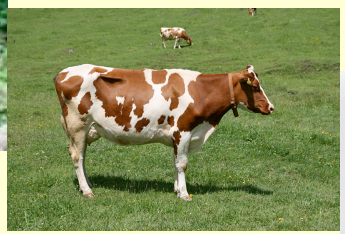
- *Caenorhabditis elegans*
- *Arabidopsis thaliana*
- *Saccharomyces cerevisiae*
- *Mus musculus*
- *Homo sapiens*
- *Oryza sativa*
- *Plasmodium falciparum*
- *Drosophila melanogaster*





434+731 Incomplete Eukaryotic Genomes

- *Anopheles gambiae*
- *Macaca mulatta*
- *Bos taurus*
- *Felis catus*
- *Gallus gallus*
- *Xenopus laevis*
- *Ovis aries*
- *Equus caballus*
- *Lama pacos*
- *Pan troglodytes*



Genome Sizes

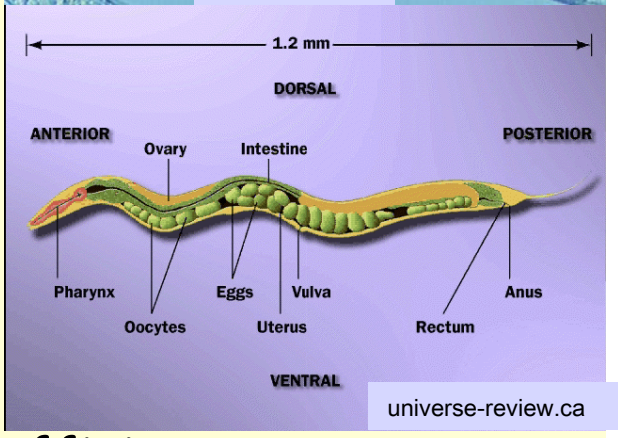
Organism	Size	Date	Est. # genes
HIV type 1	9.2 Kb	1997	9
H. influenzae	1.8 Mb	1995	1,740
M. genitalium	0.58 Mb	1998	525
E. coli	4.7 Mb	1997	4,000
S. cerevisiae	12.1 Mb	1996	6,034
C. elegans	97 Mb	1998	19,099
A. thaliana	100 Mb	2000	25,000
D. melanogaster	180 Mb	2000	13,061
M. musculus	3 Gb	2002	~30,000
H. sapiens	3 Gb	2001	32,000+

Short Homework 1

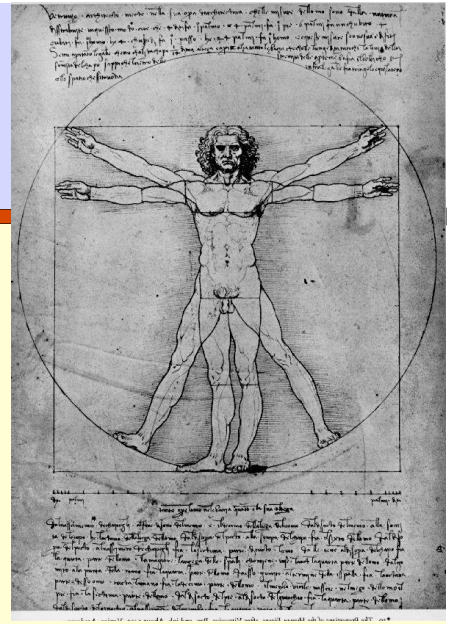
- List the most recent organism to be sequenced from each of the following categories. List the date they were sequenced and their genome sizes. When possible also list their common names.
  - Mammal
  - Land plant
  - Insect
  - Fish
  - Reptile
  - Bird
- Name the organism with the largest sequenced genome.
- Name the organism with the shortest sequenced genome.

Caenorhabditis Elegans

- Entire genome – 1998; 8 year effort
- 1st animal; 2nd eukaryote (after yeast)
- Nematode (phylum)
- Easy to experiment with; Easily observable
- 97 million bases; 20,000 genes;
- 12,000 with known function; 6 Chromosomes;
- GC content 36%
- 959 cells; 302-cell nervous system
- 36% of proteins common with human
- 15 Kb mitochondrial genome
- Results in ACeDB
- 25% of genes in operons
- Important for HGP: technology, software, scale/efficiency
- 182 genes with alternative splice variants



Homo sapiens



- Sequenced – 2001; 15 year effort
- 3 billion bases, 500 gaps
- Variable density of Genes, SNPs, CpG islands
- ~ 1.1% of genome codes for proteins; 99%?
- ~ 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)
- 223 genes common with bacteria that are missing from worm, fly or yeast.

Sequence Alignment – Why?

>gi|12643549|sp|O18381|PAX6_DROME Paired box protein Pax-6 (Eyeless protein)
MRNLPCLGTAGGSGLGGIAGKPSPTMEAVEASTASHRHSTSSYFATTYYHLTDDECHSGVNQLGGVFVGG
RPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKIS
QYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTSAGNSISAKVSV
SIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEP
ARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSWPPRHYSGSWYPTSLSEIPISSAPNIASVTAY
ASGPSLAHSLSPPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDD
QARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEK
LRNQRRTPNSTGASATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPTL
GAGIDSSESPTPIPHIRPSCTSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHAQGHALVPAISP
RLNFNSGSFGAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSPIPQQGDLTPSSLYPCHMTLRP
PPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGASCSGSGYEVLSAYALPPPPMASSSAADSSFSAASSAS
ANVTPHHTIAQESCPSPCSSASHFGVAHSSGFSSDPISPAVSSYAHMSYNYASSANTMTPSSASGTSAHV
APGKQQFFASCFYSPWV

>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRA
IGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGAD
GMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKL
QRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASN
TPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQ
TSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ

Drosophila Eyeless vs. Human Aniridia

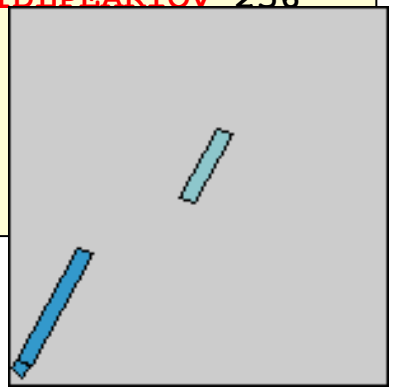
```
Query: 57   HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 116
            HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct: 5    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 64

Query: 117 SIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSIN 176
            SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E VCTNDNIPSVSSIN
Sbjct: 65   SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN 124

Query: 177 RVLRNLAAQKEQ 188
            RVLRNLA++K+Q
Sbjct: 125 RVLRNLASEKQQ 136
```

```
Query: 417 TEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQV 476
           +++ Q RL LKRKLQRNRTSFT +QI++LEKEFERTHYPDVFARERLA KI LPEARIQV
Sbjct: 197 SDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQV 256

Query: 477 WFSNRRAKWRREEKLRNQRR 496
           WFSNRRAKWRREEKLRNQRR
Sbjct: 257 WFSNRRAKWRREEKLRNQRR 276
```



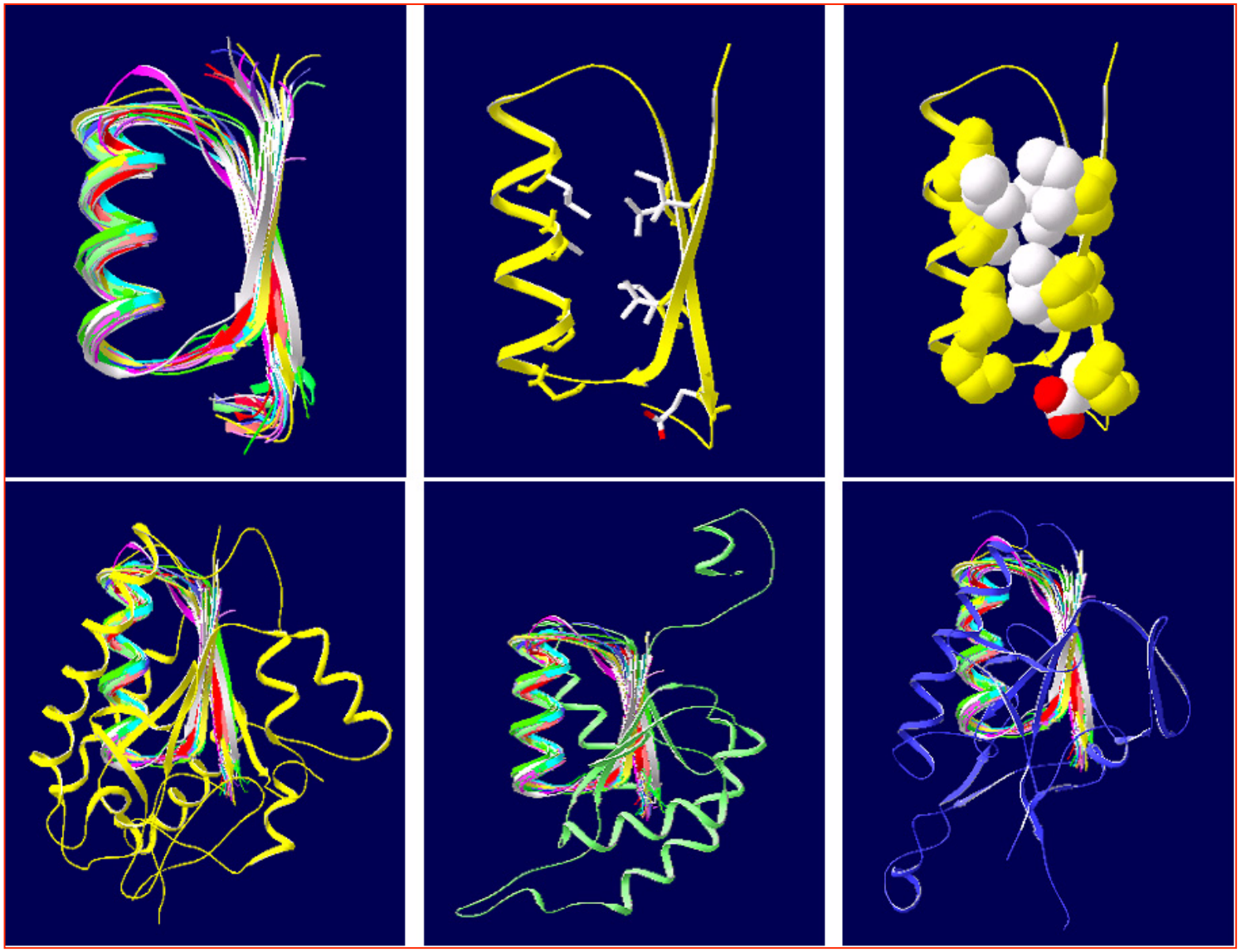
E-Value = 2e-31

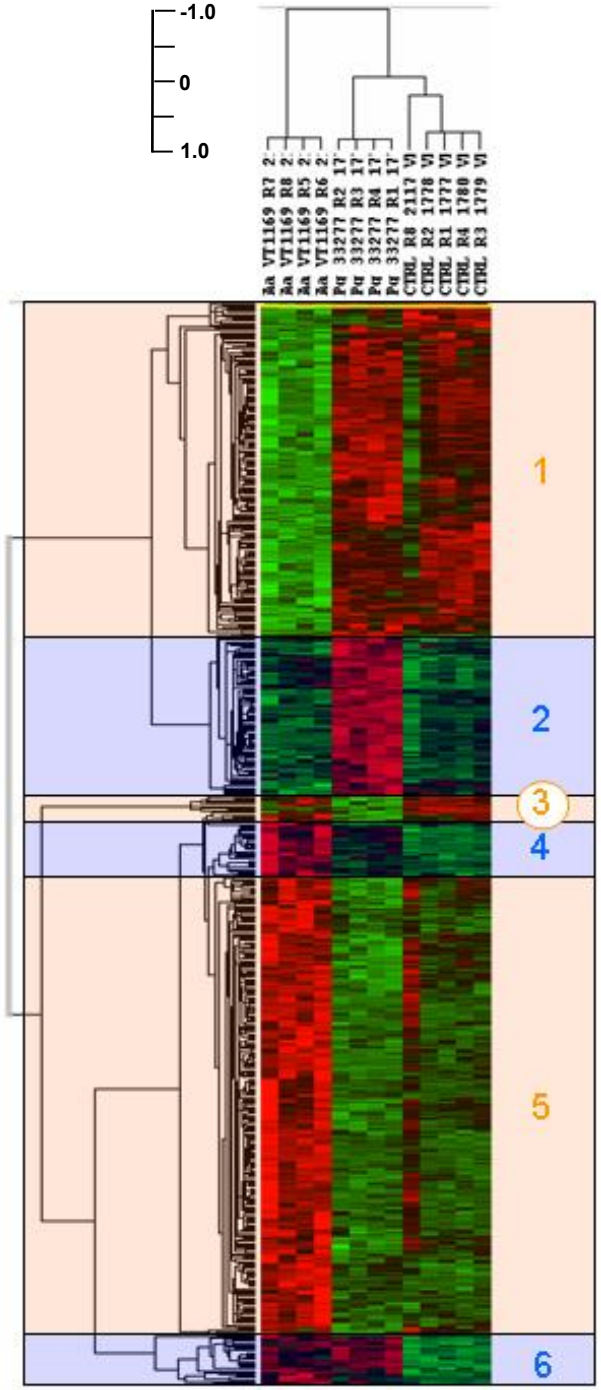
Motif Detection in Protein Sequences

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

Patterns in Protein Structures

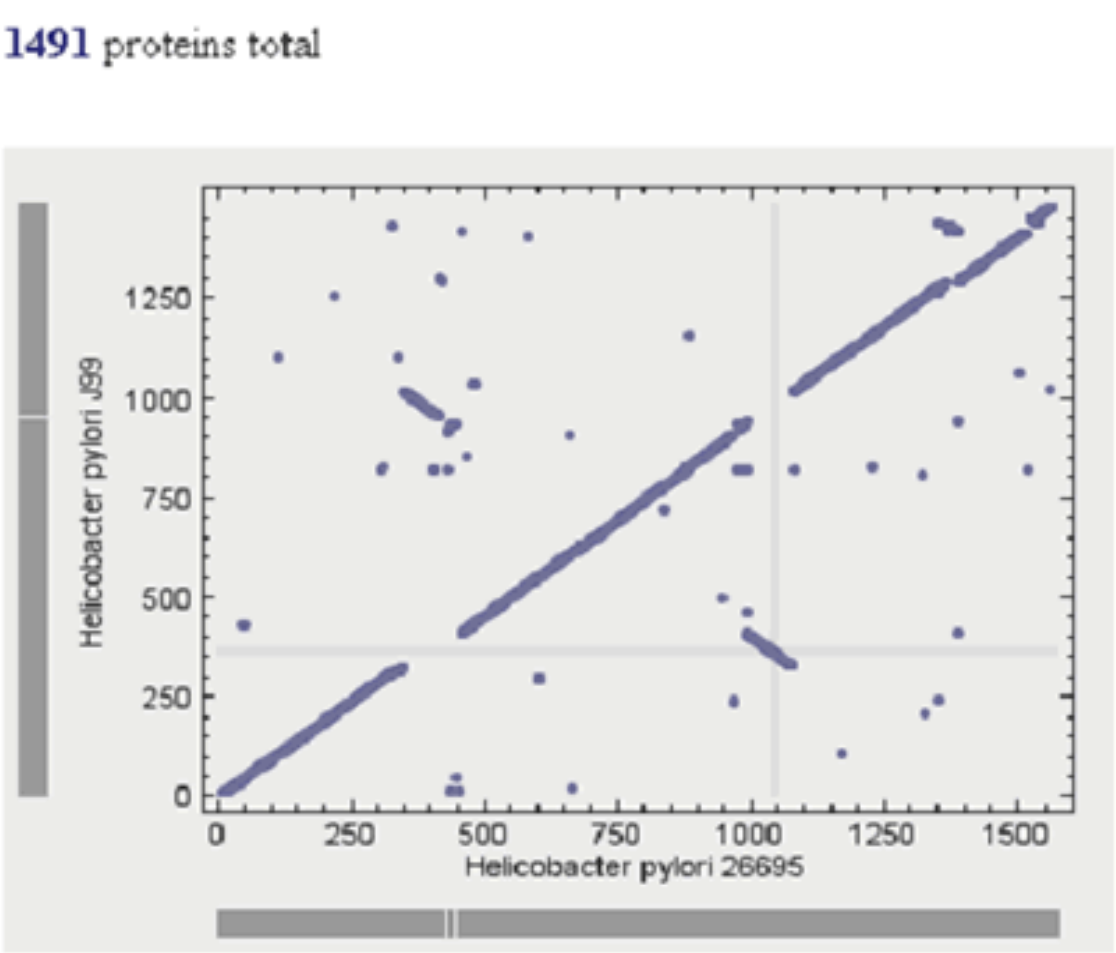




Microarray Analysis

Different patterns of gene expression of oral epithelial IHGK cells upon co-culture with A. actinomycetemcomitans or P. gingivalis.

Tools: GenePlot



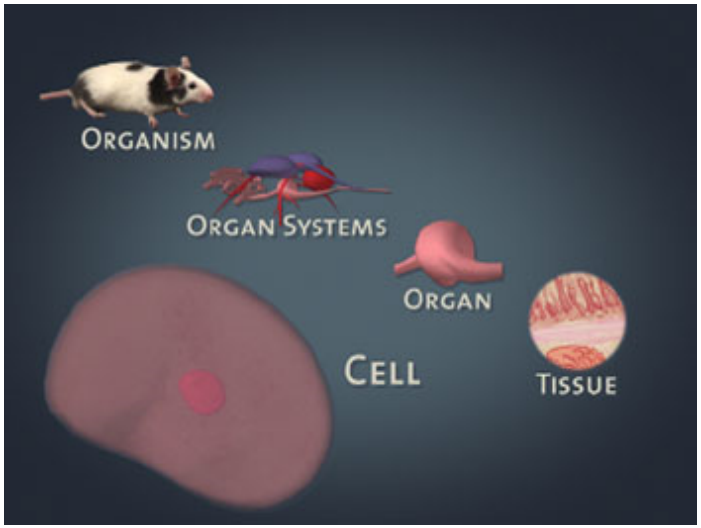
Comparison of proteins from two strains of Helicobacter Pylori, 26695 and J99. Each point represents a pair of proteins from the two organisms showing a symmetrical best BLAST score; the coordinates of each point correspond to the position of the protein genes in the 2 genomes. Note the juxtaposition and inversion of two segments of the genome between the two strains.

SIDS

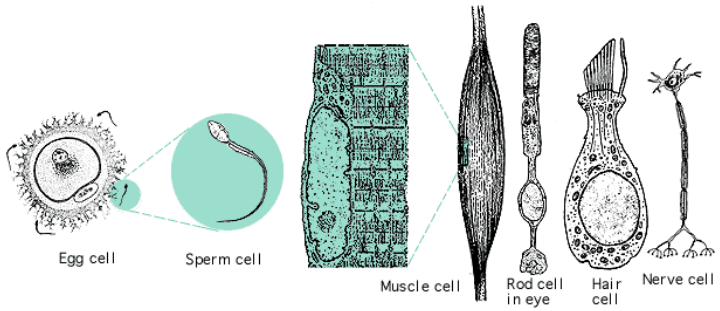
- 18000 Amish people in Pennsylvania
- Mostly intermarried due to religious doctrine
- rare recessive diseases occurred with high frequencies.
- SIDS: 3000 deaths/year (US); 21 deaths (Amish community)
- Many research centers failed to identify cause
- Collaboration between Affymetrix, TGEN & Clinic for special children solved the problem in 2 months
- Studied 10000 SNPs using microarray technology
- Their experiments showed that all the sick infants had two mutant copies of a specific gene, and their parents were carriers of the mutant gene.
- Conclusion: Disease caused by 2 abnormal copies of TSPYL gene
- Identified genes expressed in key organs (brainstem,testes)
- http://www.affymetrix.com/community/wayahead/modern_miracle.affx

Molecular Biology Background

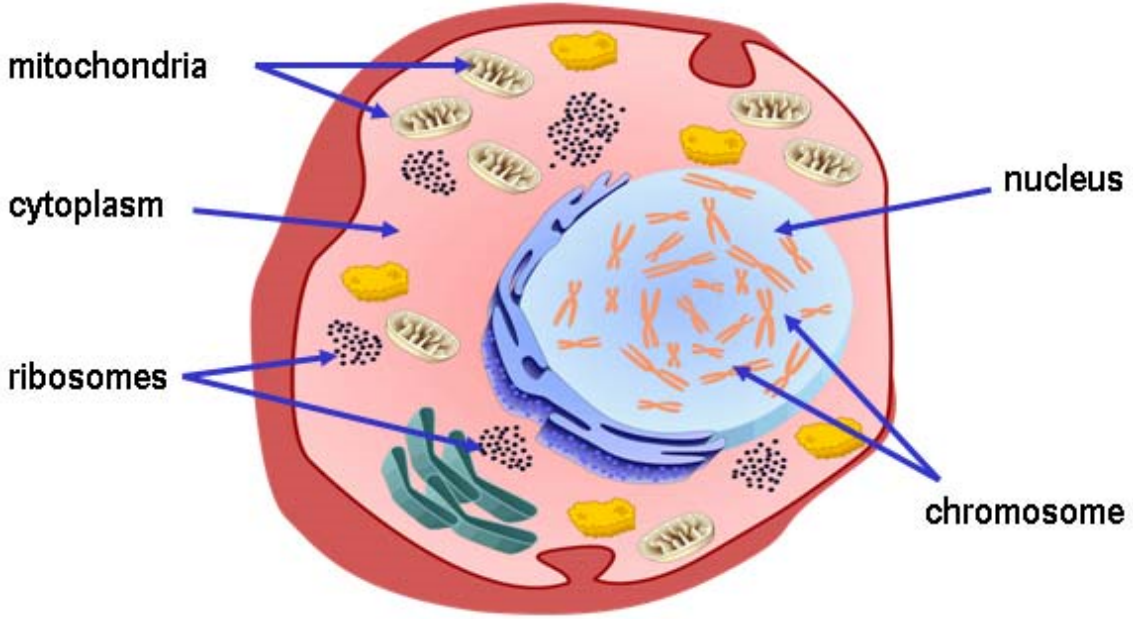
Cell



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

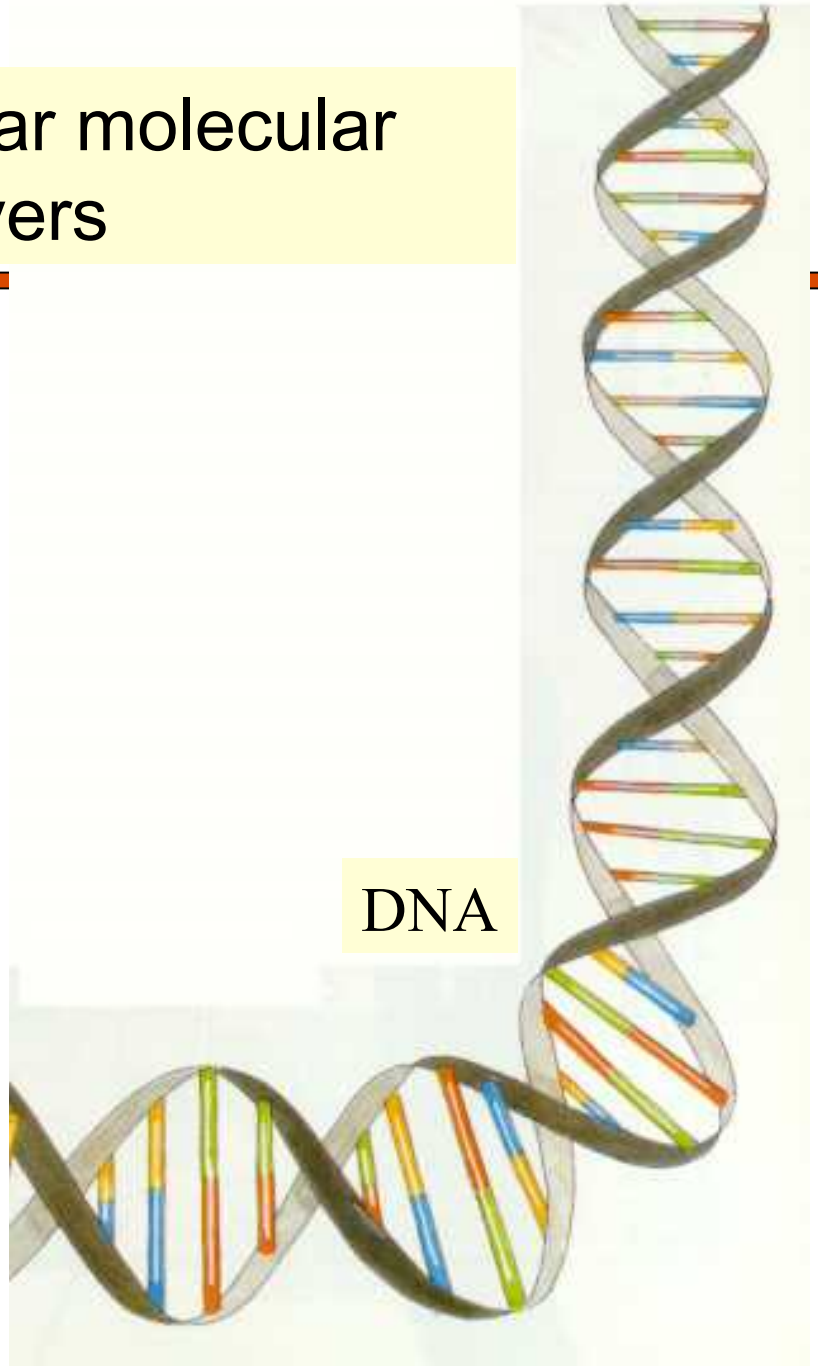


http://www.biology.eku.edu/RITCHISO/301notes1.htm

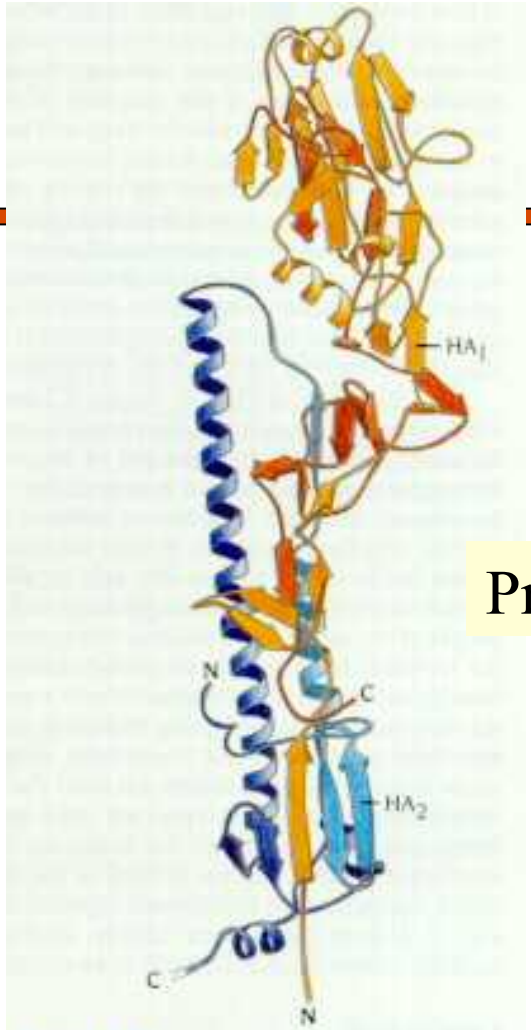


http://www.biotechnologyonline.gov.au/popups/img_cellwithlabels.cfm

2 star molecular 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 75Å long. The globular head is formed by residues only from HA_1. (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} **Bases**

Protein

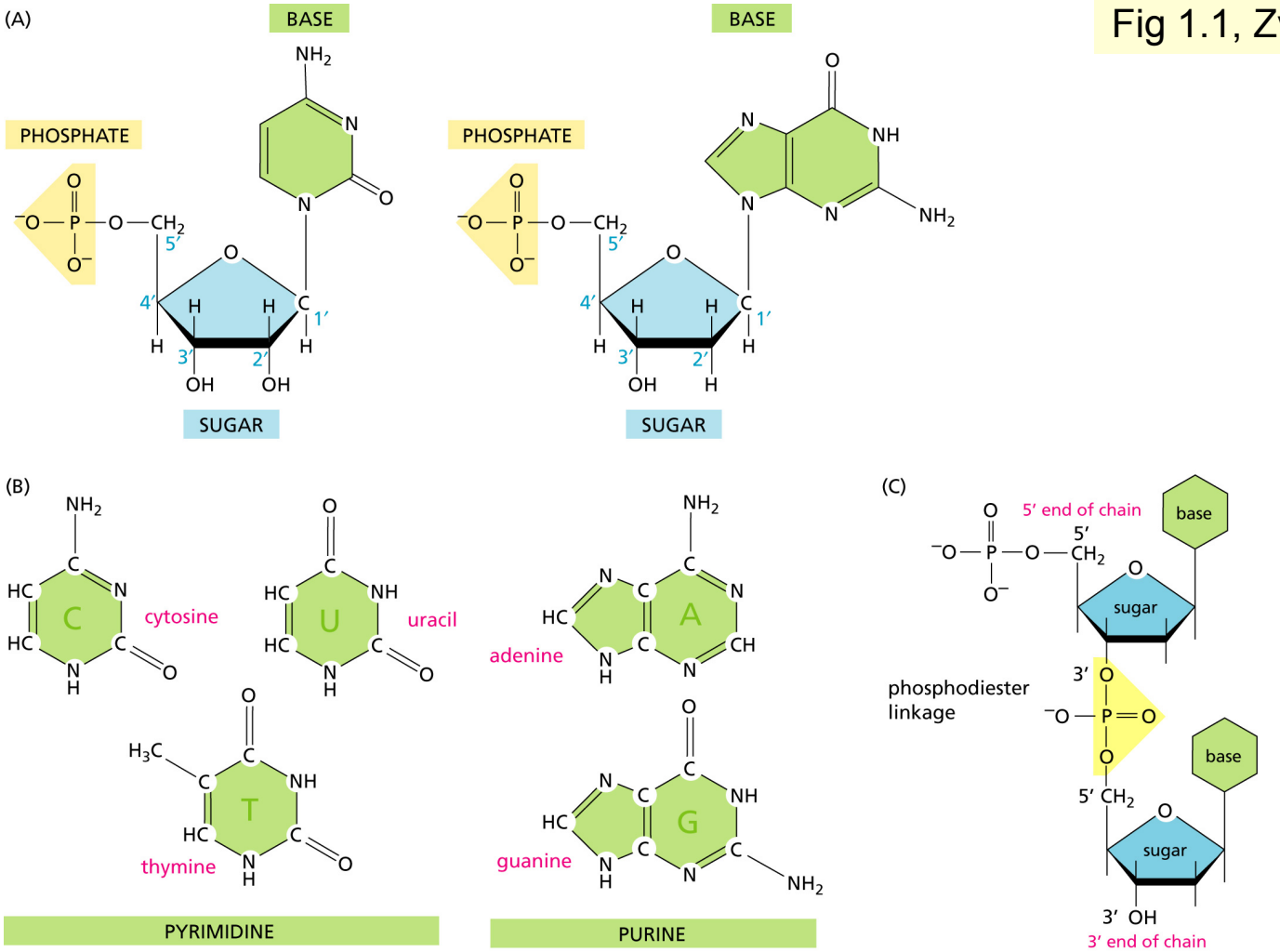
String with 20-letter alphabet **Amino acids/ Residues**

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 tgcccccta 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
```

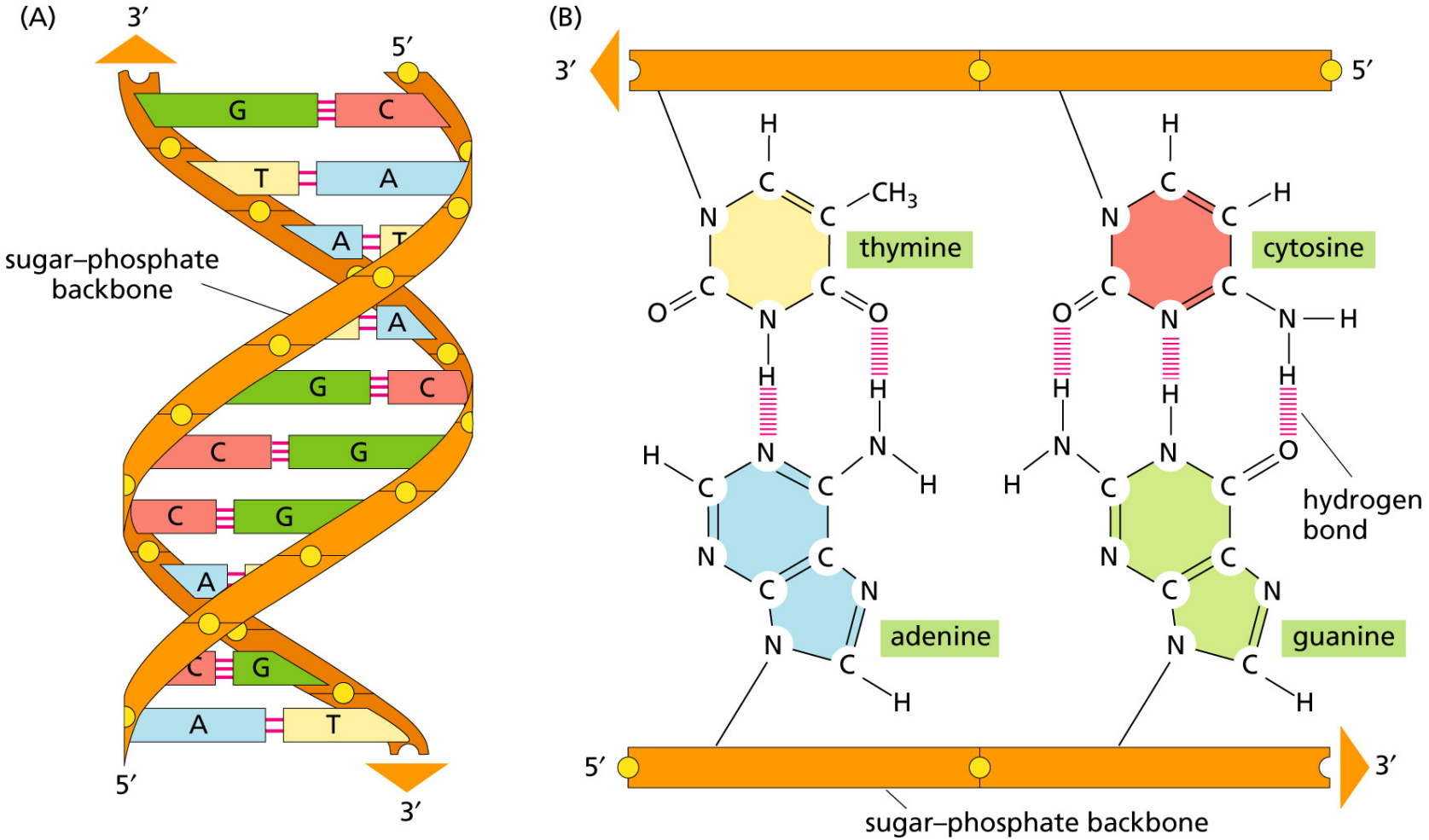
The building blocks of DNA & RNA

Fig 1.1, Zvelebil/Baum



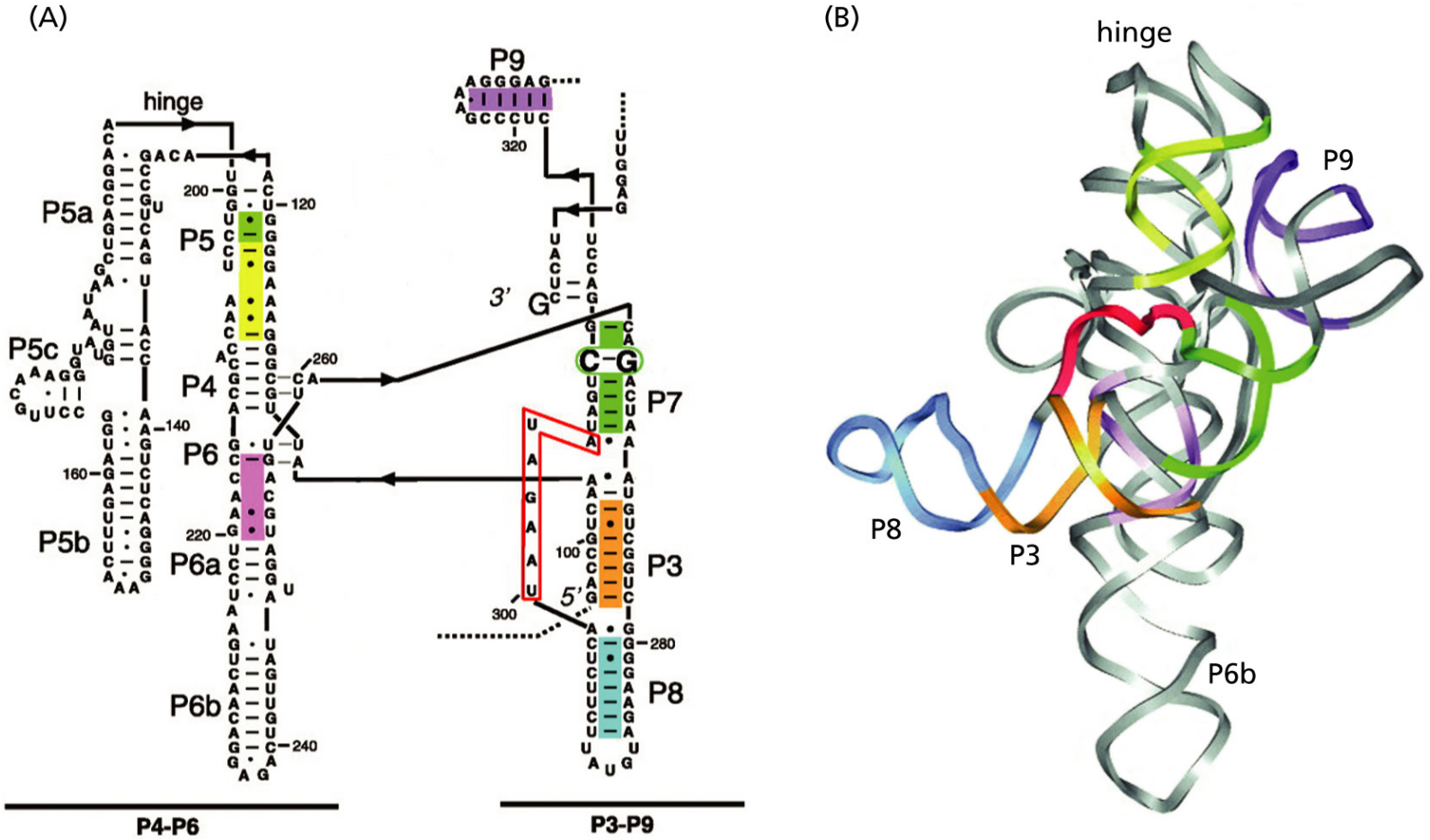
DNA double helix structure

Fig 1.3, Zvelebil/Baum



RNA molecule

Fig 1.5, Zvelebil/Baum



Typical protein sequence

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

Protein 3D Structure

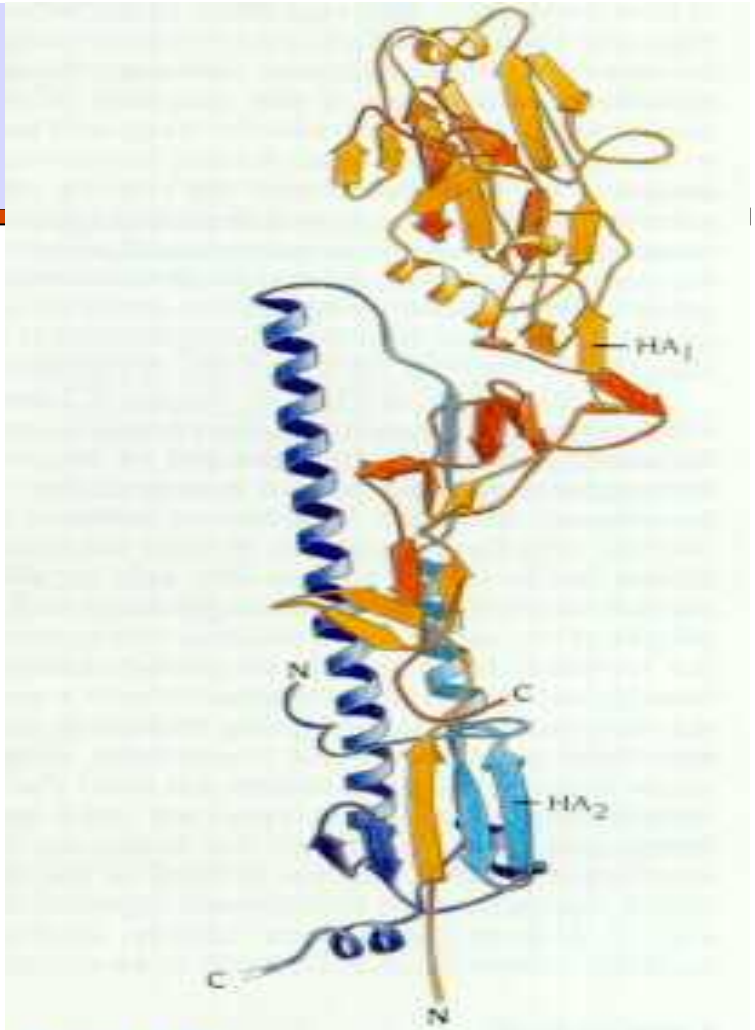
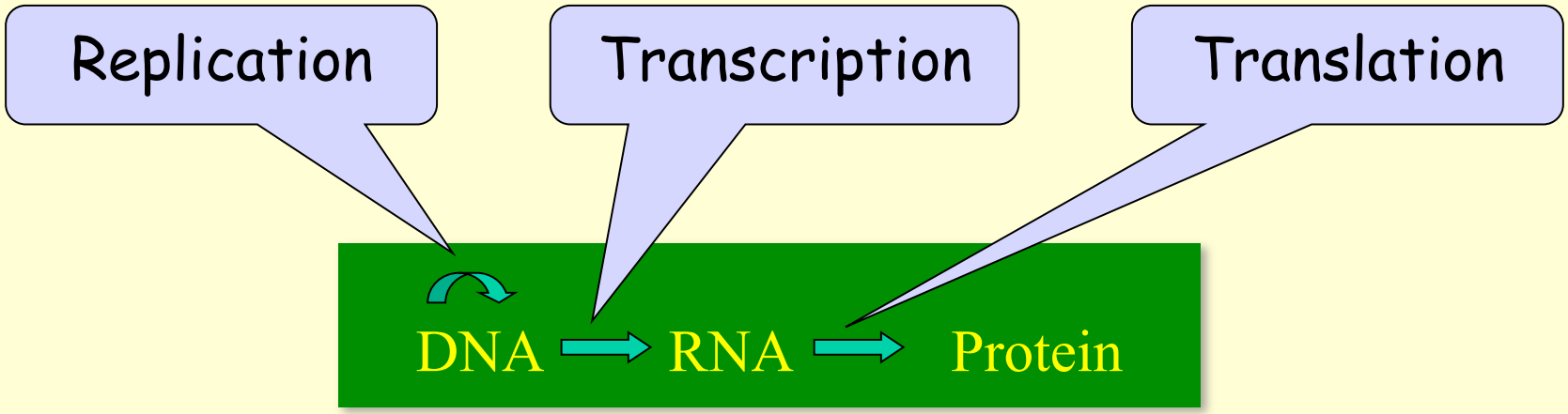


Figure 8.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75Å long. The globular head is formed by residues only from HA_1. (Courtesy of Don Wiley, Harvard University.)

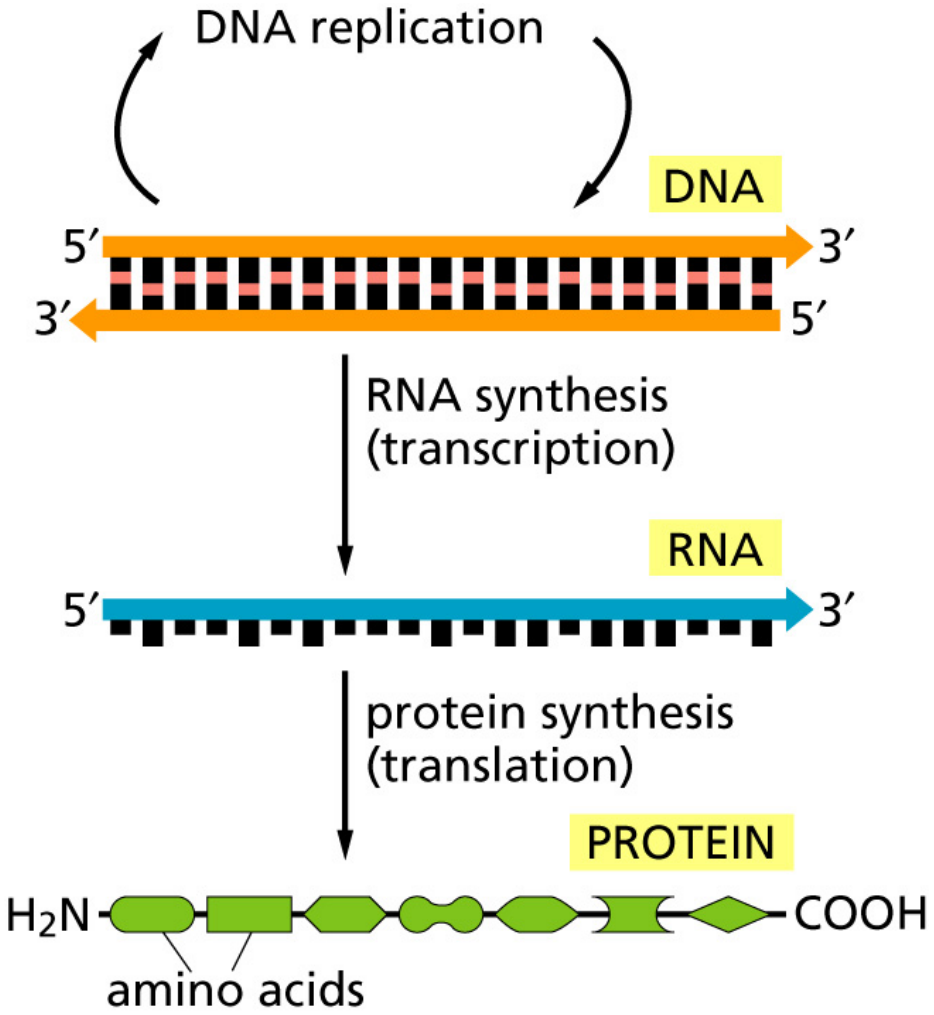
Central Dogma

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



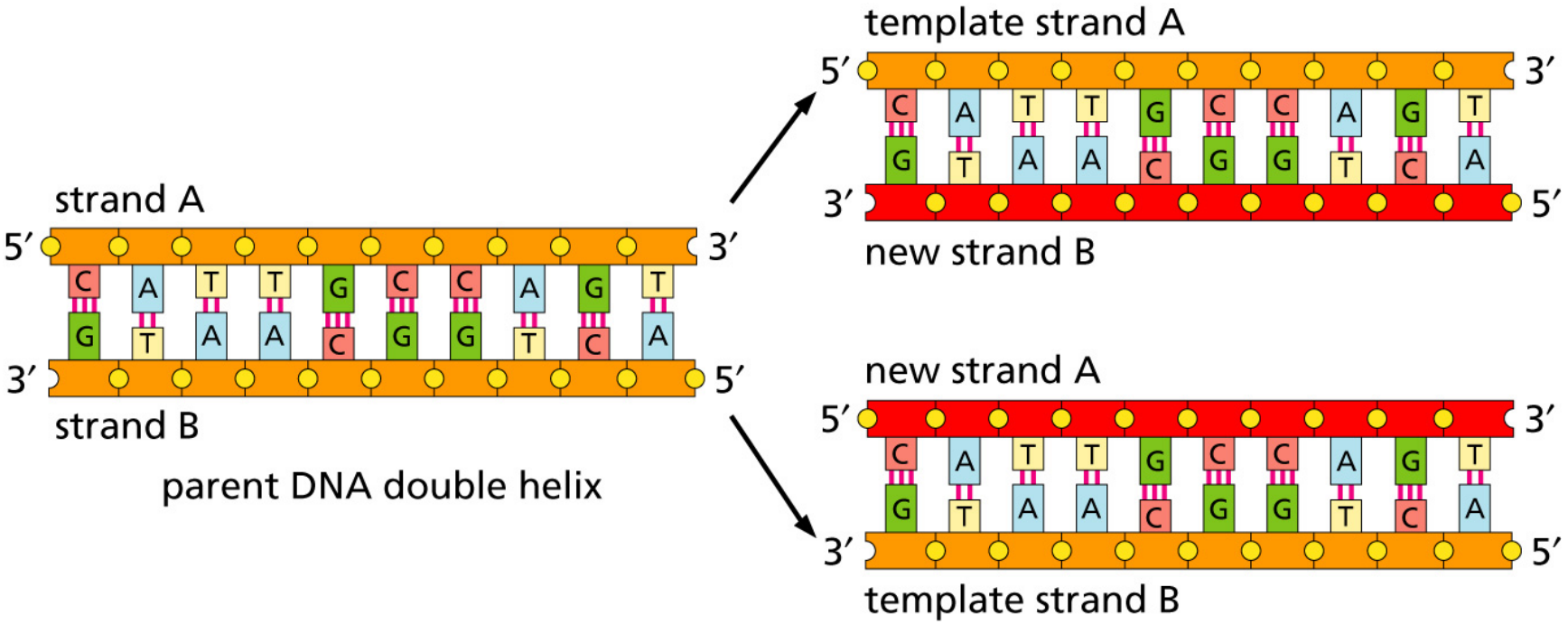
Central Dogma

Fig 1.6, Zvelebil/Baum

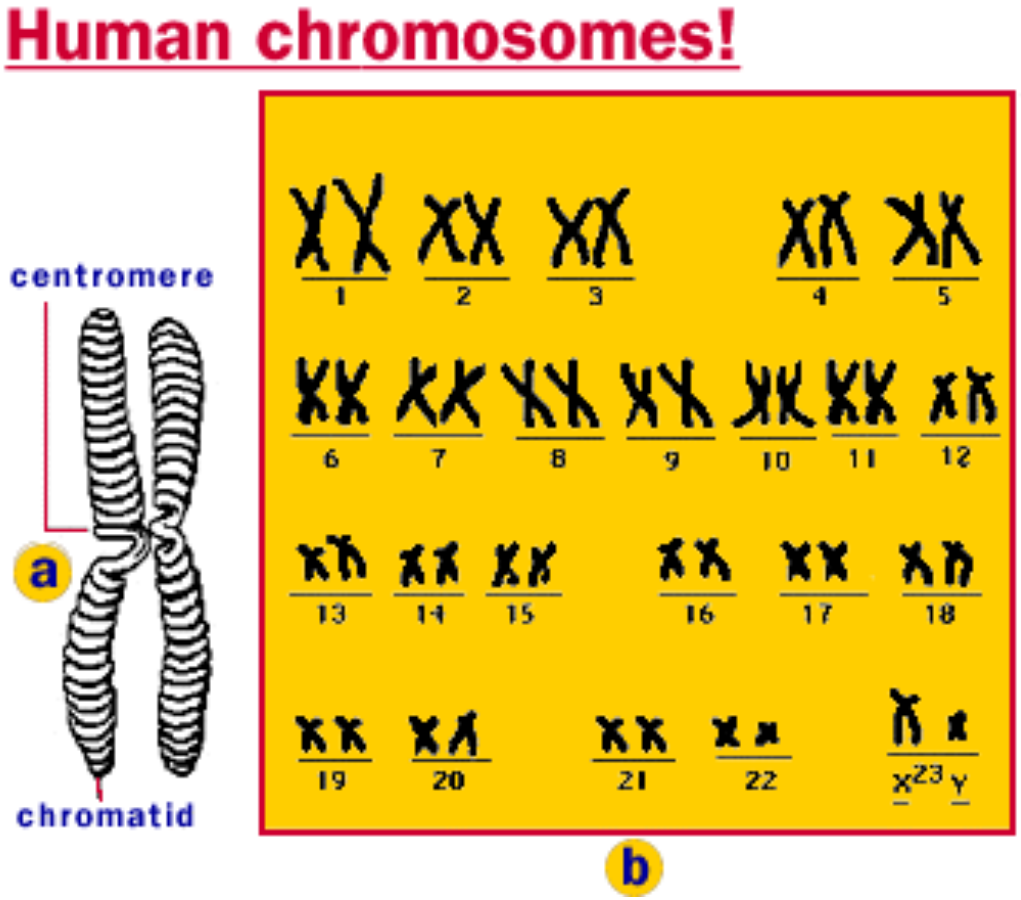


DNA Replication

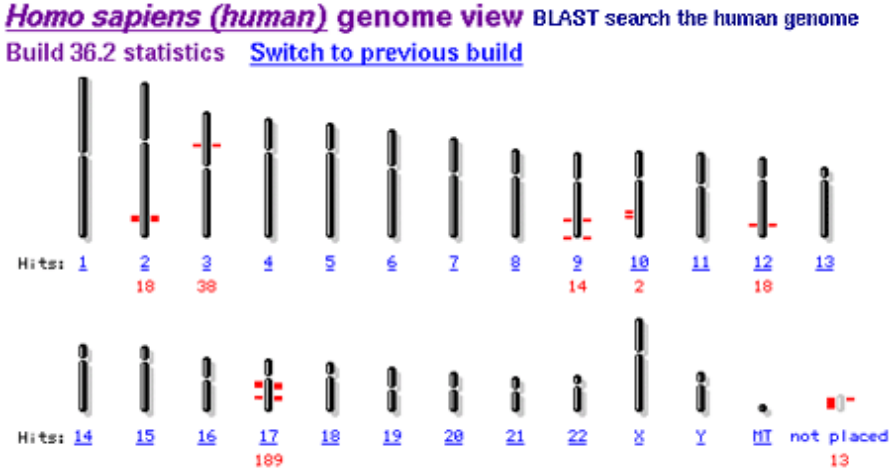
Fig 1.4, Zvelebil/Baum

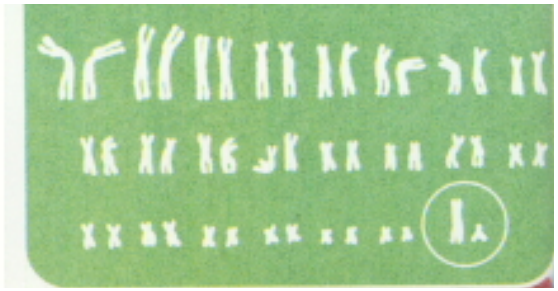


Chromosomes

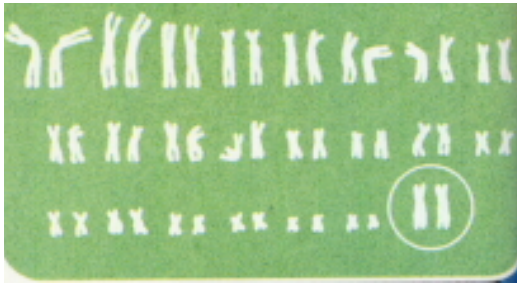


Chromosomes

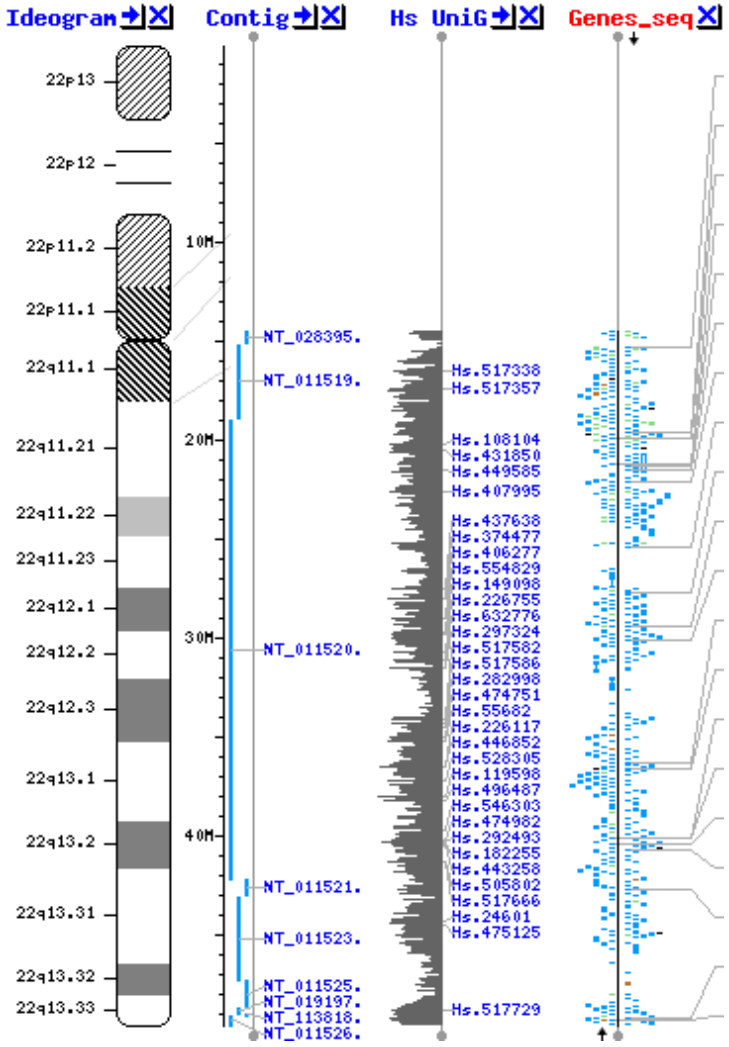




The chromosomal locations of several genes believed to be associated with the human BRCA1 gene implicated in breast cancer are highlighted.



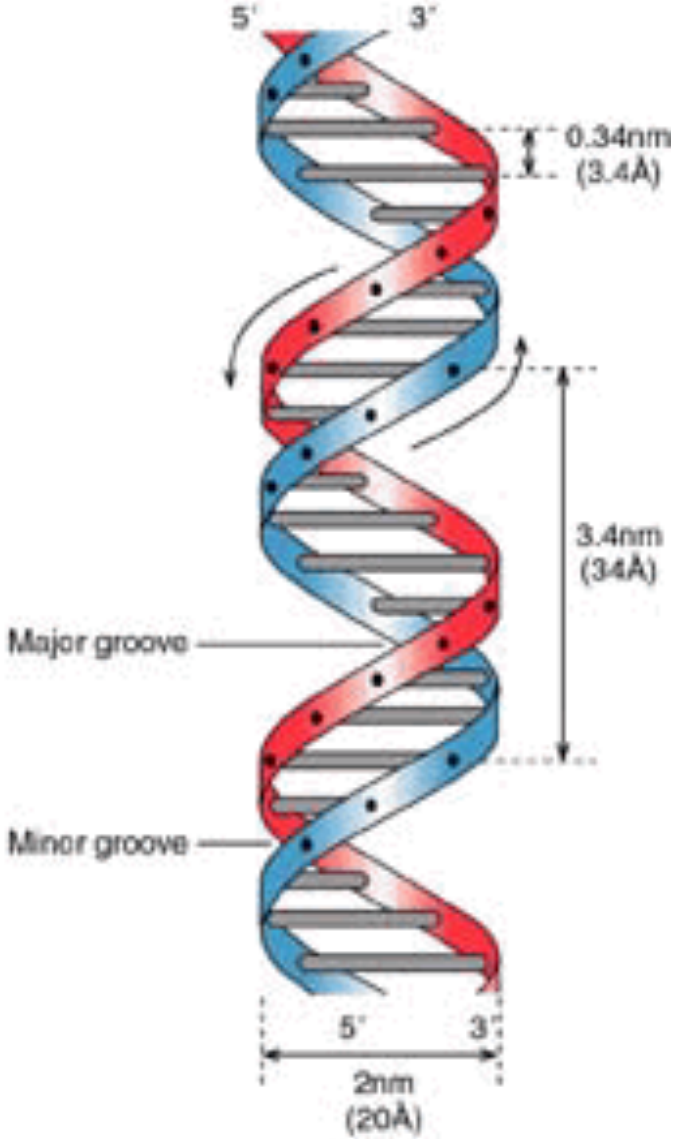
Human Chr 22



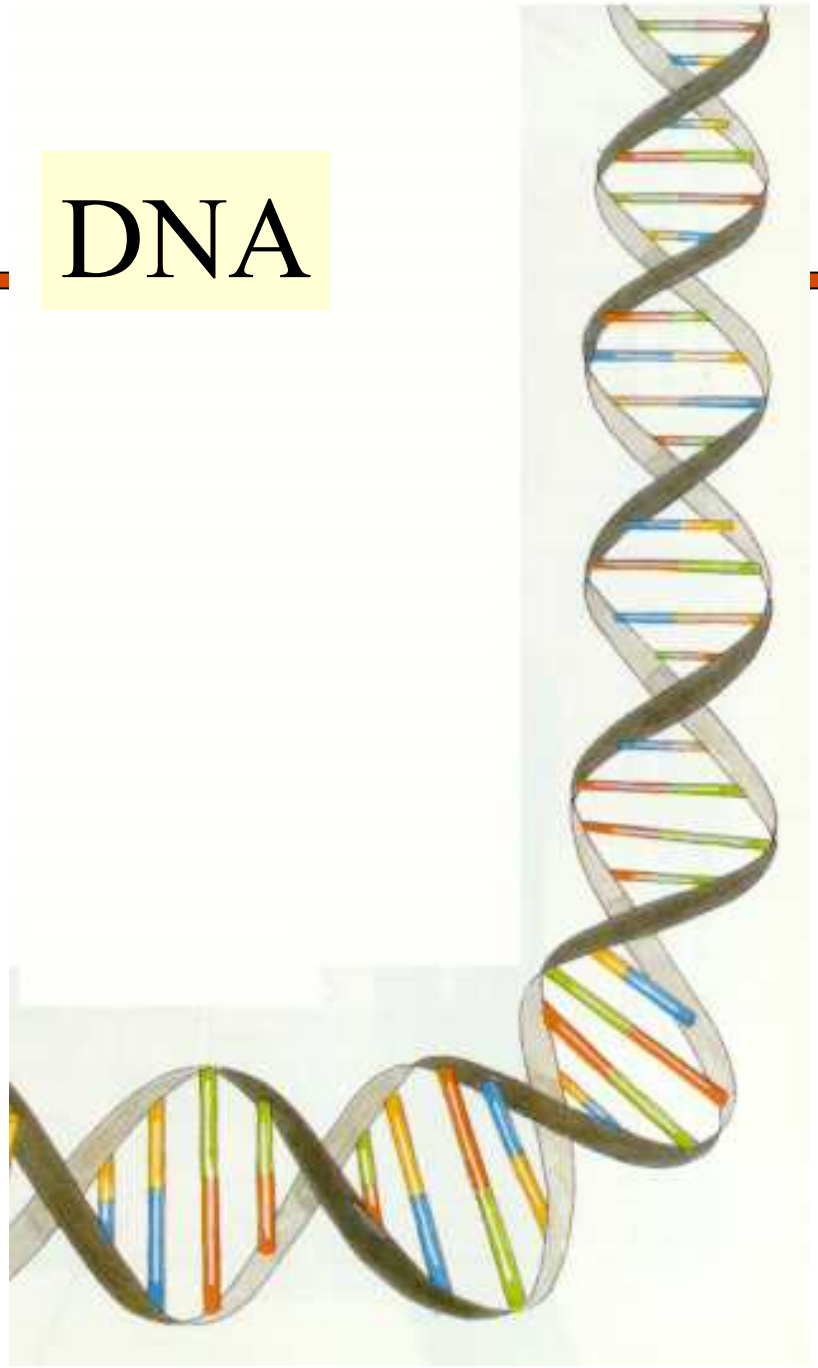
Symbol	Position	Description
ABCD1P4	22q11	ATP-binding cassette, sub-family D (ALD)
SNAP29	22q11.21	synaptosomal-associated protein

-
-
-

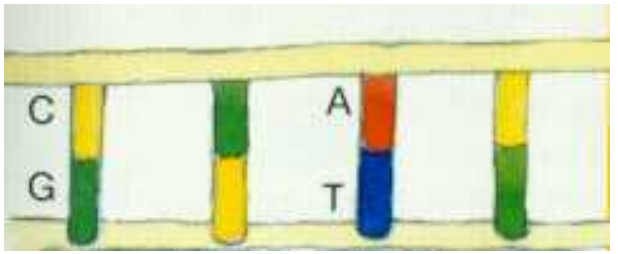
DNA Molecule



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

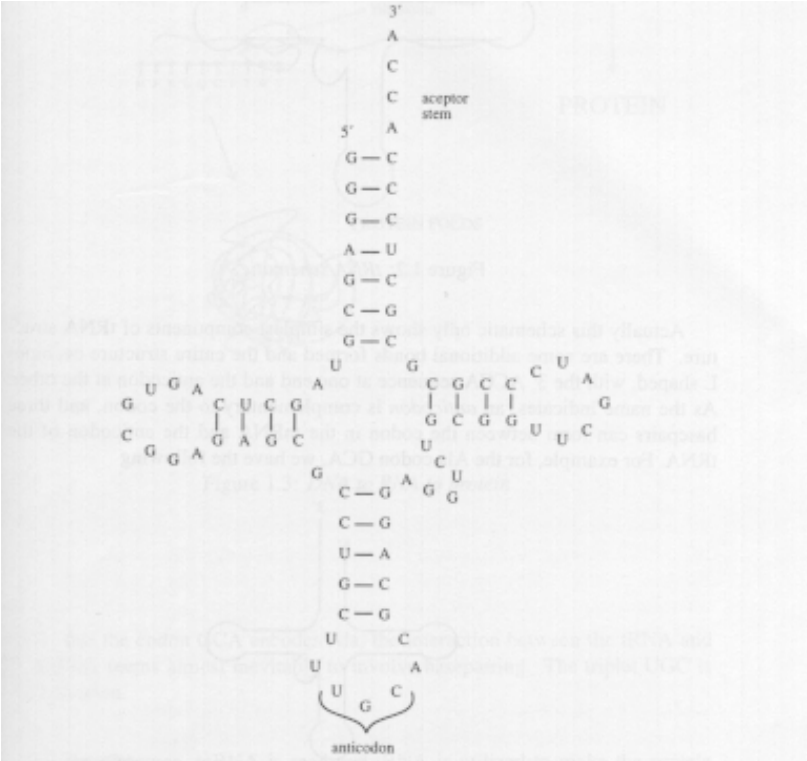


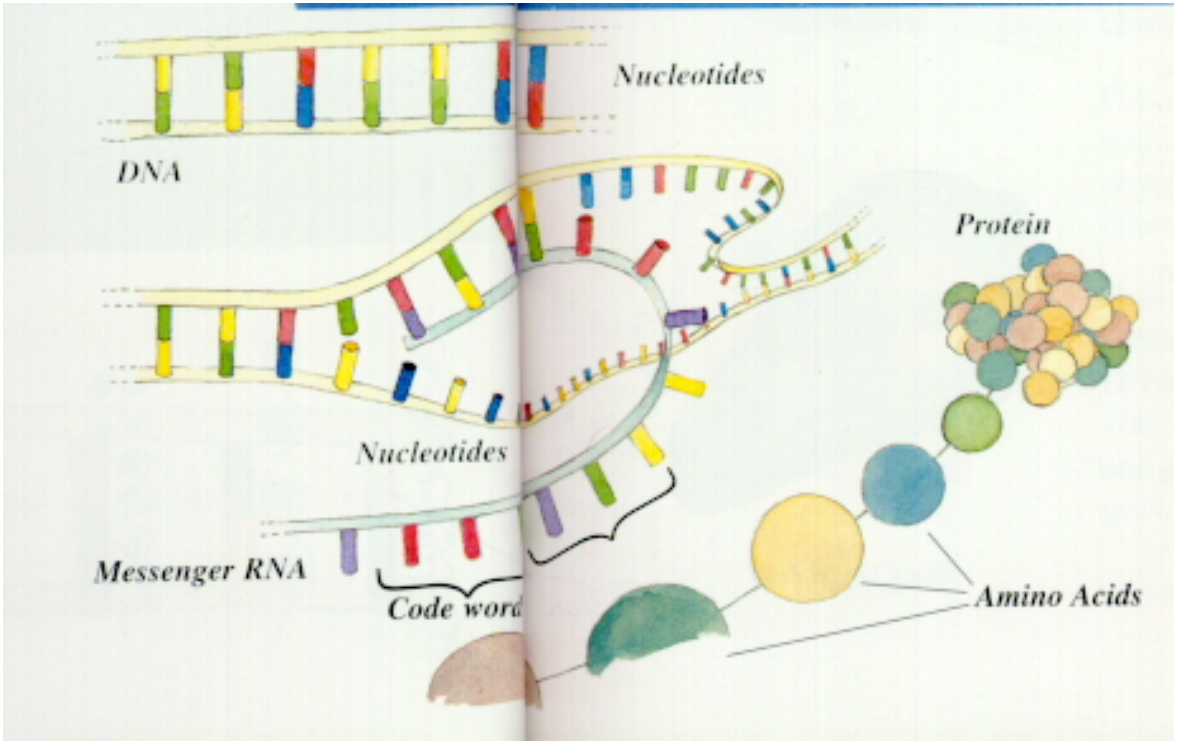
Figure 1.1: E. coli *Ala tRNA*

Genes

DNA

Gene A Gene B Gene C Gene D Gene E

DNA → RNA → Protein



Basic Genetic Processes

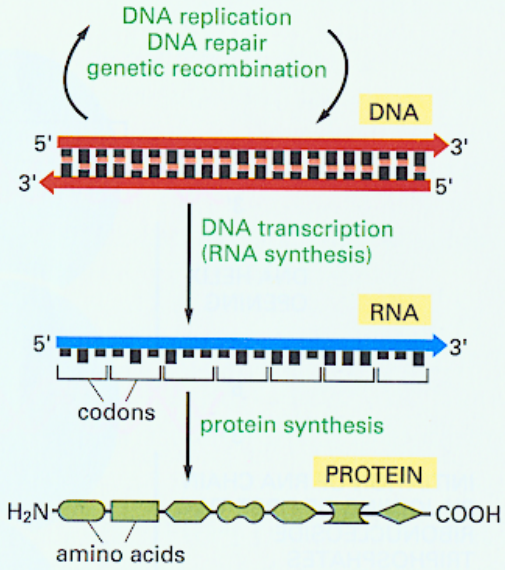
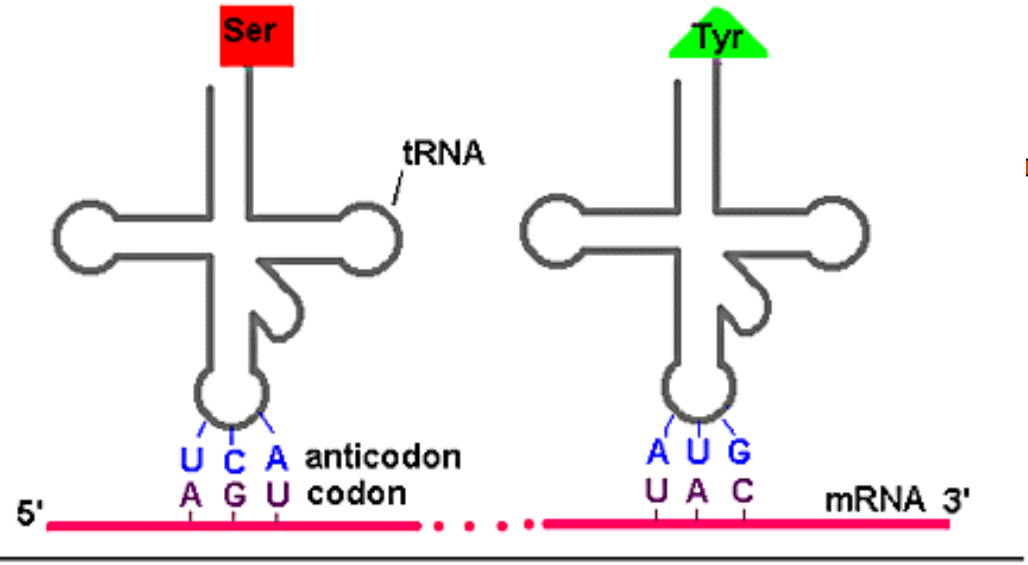


Figure 6–1 The basic genetic processes. The processes shown here are thought to occur in all present-day cells. Very early in the evolution of life, however, much simpler cells probably existed that lacked both DNA and proteins (see Figure 1–11). Note that a sequence of three nucleotides (a codon) in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

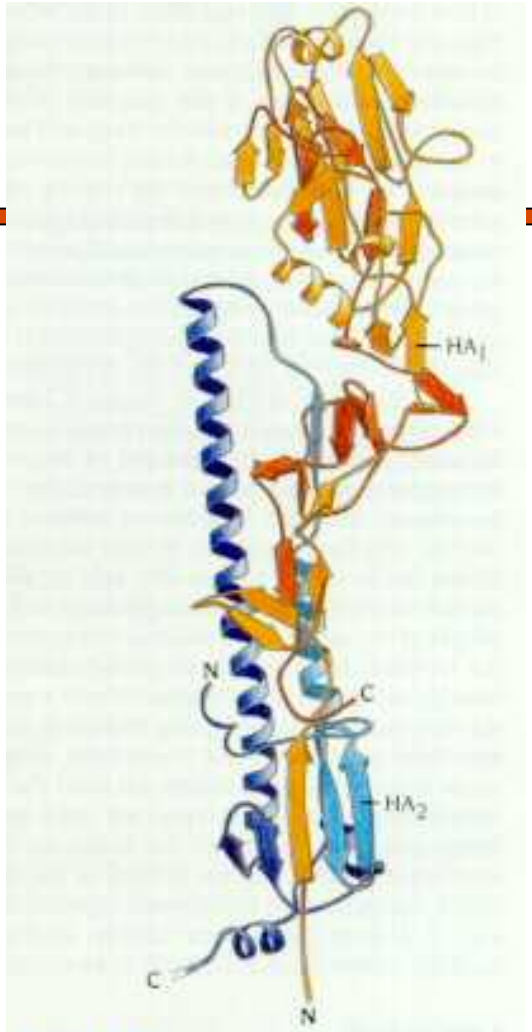
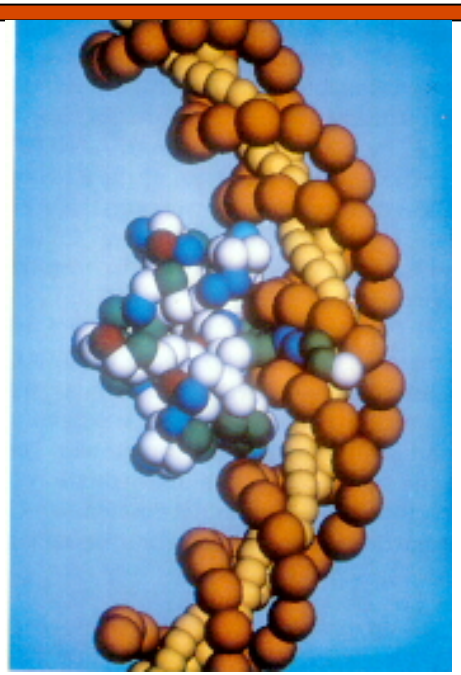
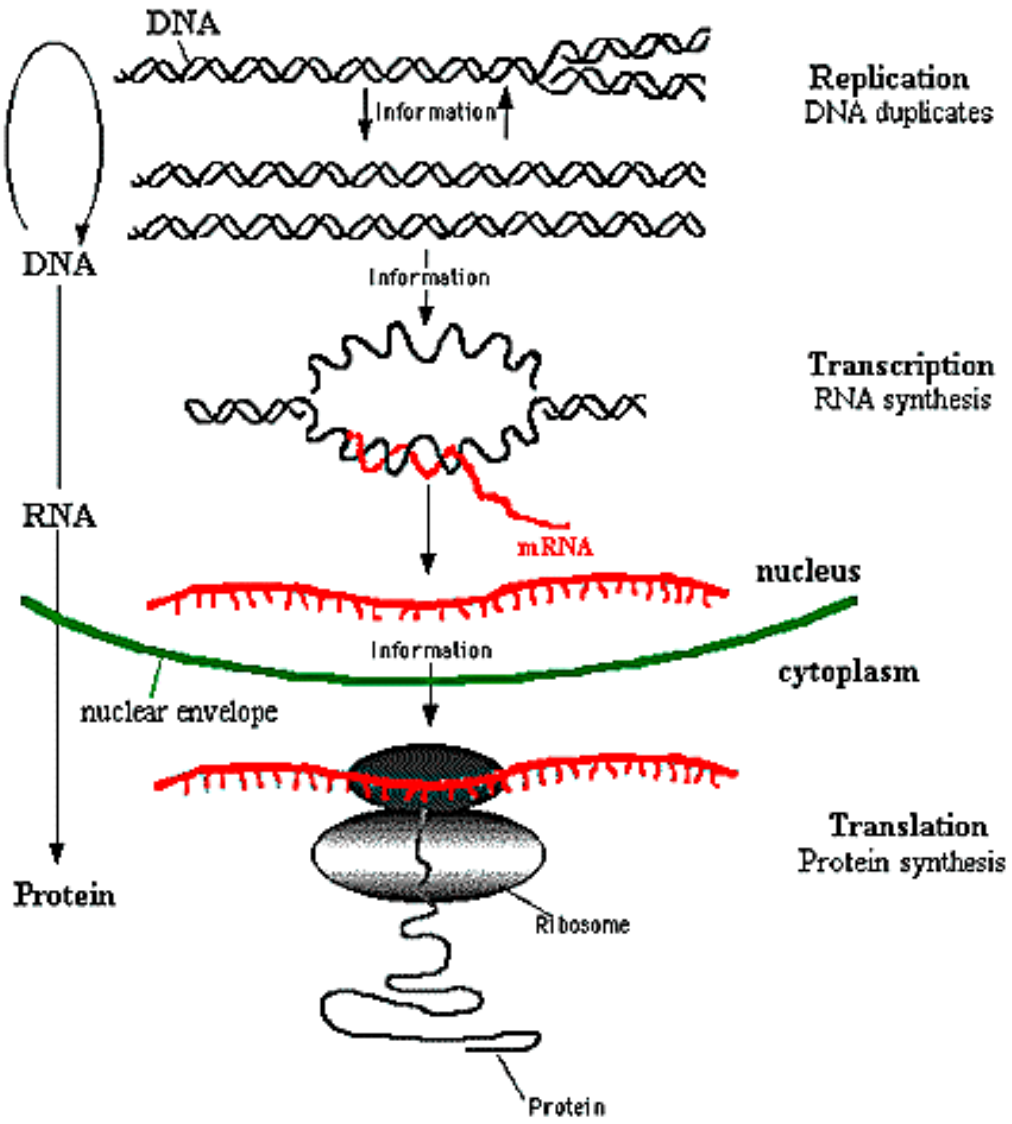


Figure 5.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike 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.)

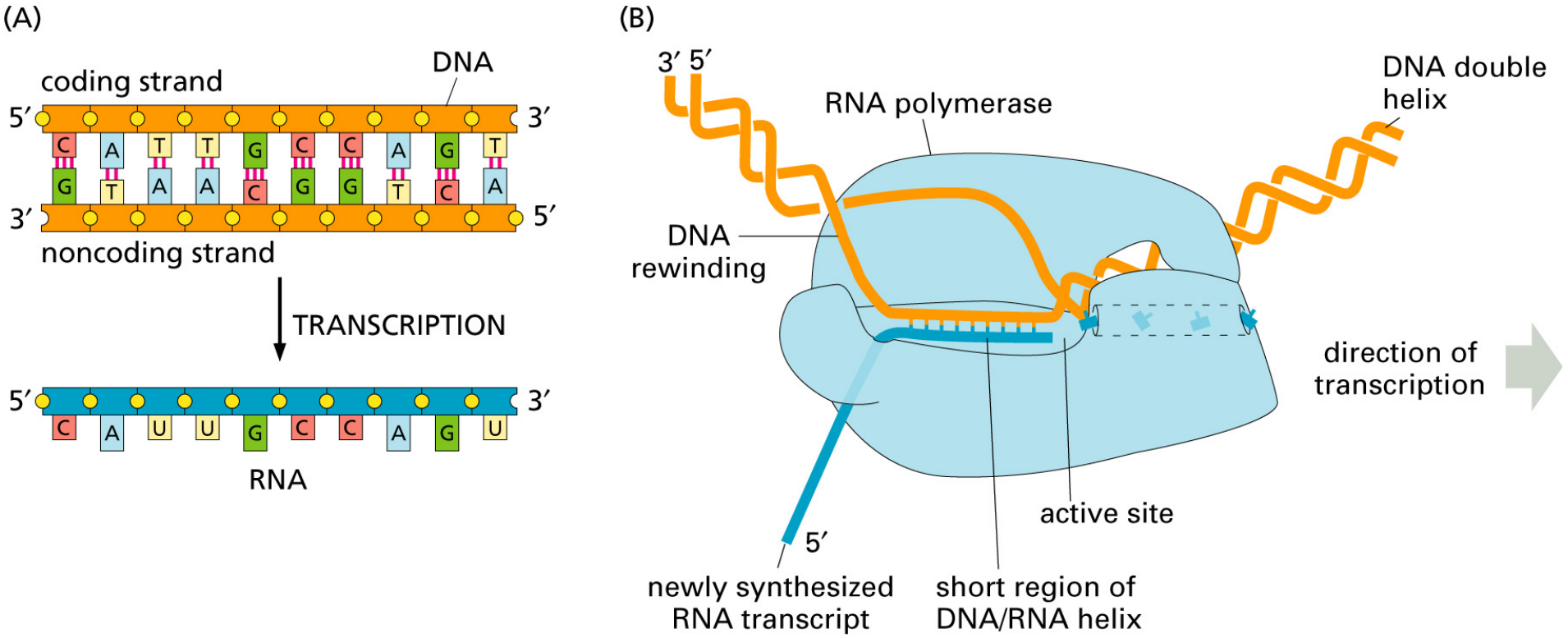


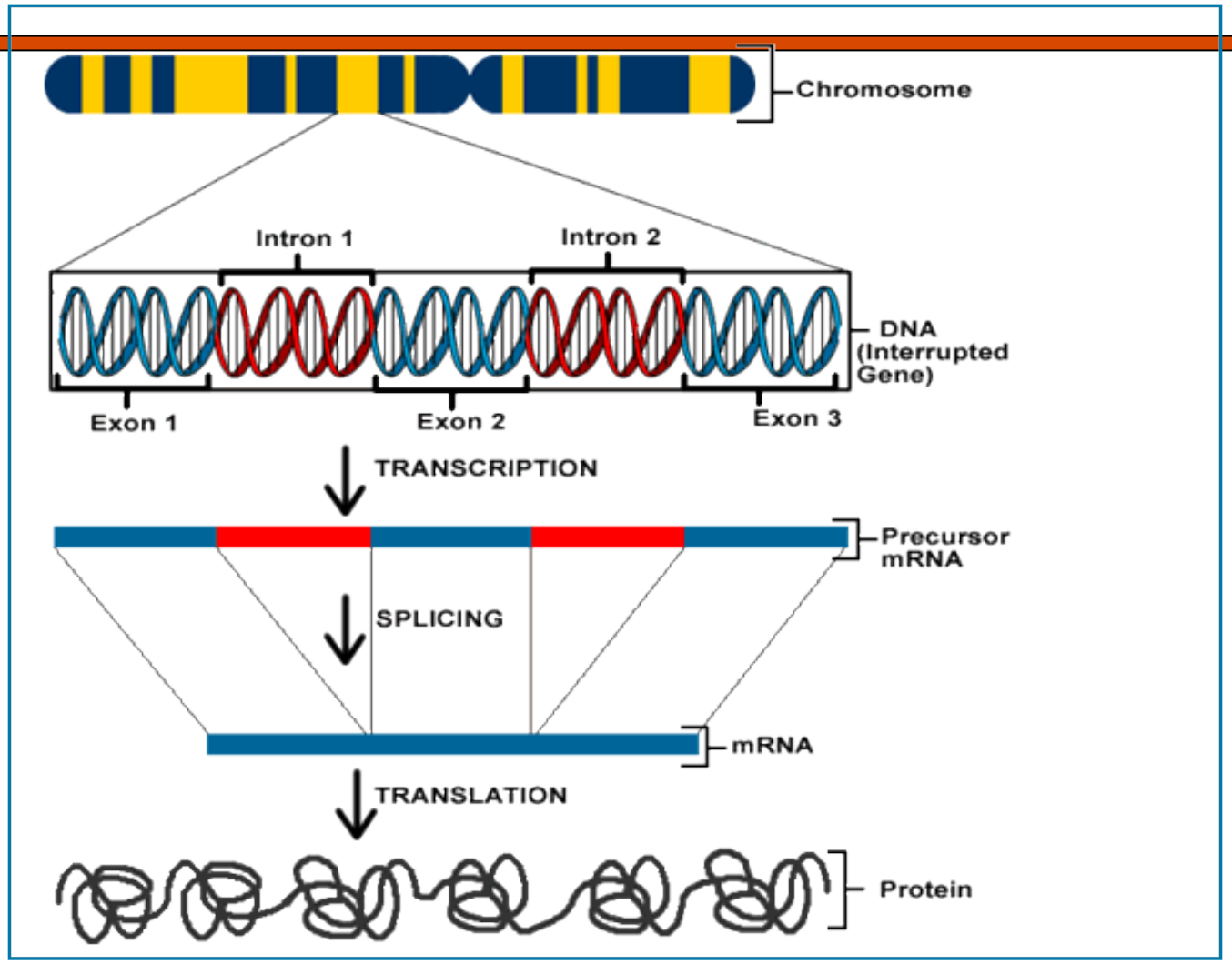


The Central Dogma of Molecular Biology

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