

# BSC 4934: Q'BIC Capstone Workshop

**Giri Narasimhan**

ECS 254A; Phone: x3748

[giri@cs.fiu.edu](mailto:giri@cs.fiu.edu)

[http://www.cs.fiu.edu/~giri/teach/BSC4934\\_Su10.html](http://www.cs.fiu.edu/~giri/teach/BSC4934_Su10.html)

July 2010

# Overview of Course

- ❑ Sequence Alignment; Multiple Sequence Alignment
- ❑ Sequence Analysis
- ❑ Sequencing and Mapping
- ❑ 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
- ❑ Ribozymes and Riboswitches
- ❑ Databases & Software Packages
- ❑ Statistics for Bioinformatics
- ❑ Computational Learning & Predictive Methods
- ❑ Biomedical Image Analysis
- ❑ Emerging Biotechnologies

# Software Packages

- ❑ Databases (*GenBank, SwissPROT*)
- ❑ Programming Environments (*BioPerl*)
- ❑ Sequence Alignment (*BLAST, CLUSTALW*)
- ❑ Phylogenetic Analysis (*CLUSTALW, Phylip, PAML*)
- ❑ Learning Methods (*HMMPPro, GeneCluster, ASOM*)
- ❑ Pattern Discovery Techniques (*GYM, TEIRESIAS, APRIORI*)
- ❑ 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

- |   |        |
|---|--------|
| <input type="checkbox"/> Homework Assignments | (35 %) |
| <input type="checkbox"/> Exam                 | (35 %) |
| <input type="checkbox"/> Semester Project     | (25 %) |
| <input type="checkbox"/> Class Participation  | (5 %)  |

# Course Homepage

[http://www.cis.fiu.edu/~giri/teach/BSC4934\\_Su10.html](http://www.cis.fiu.edu/~giri/teach/BSC4934_Su10.html)

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

# Introduction

## 1. What is Bioinformatics?

- Analysis of biological data with computing & statistical tools.

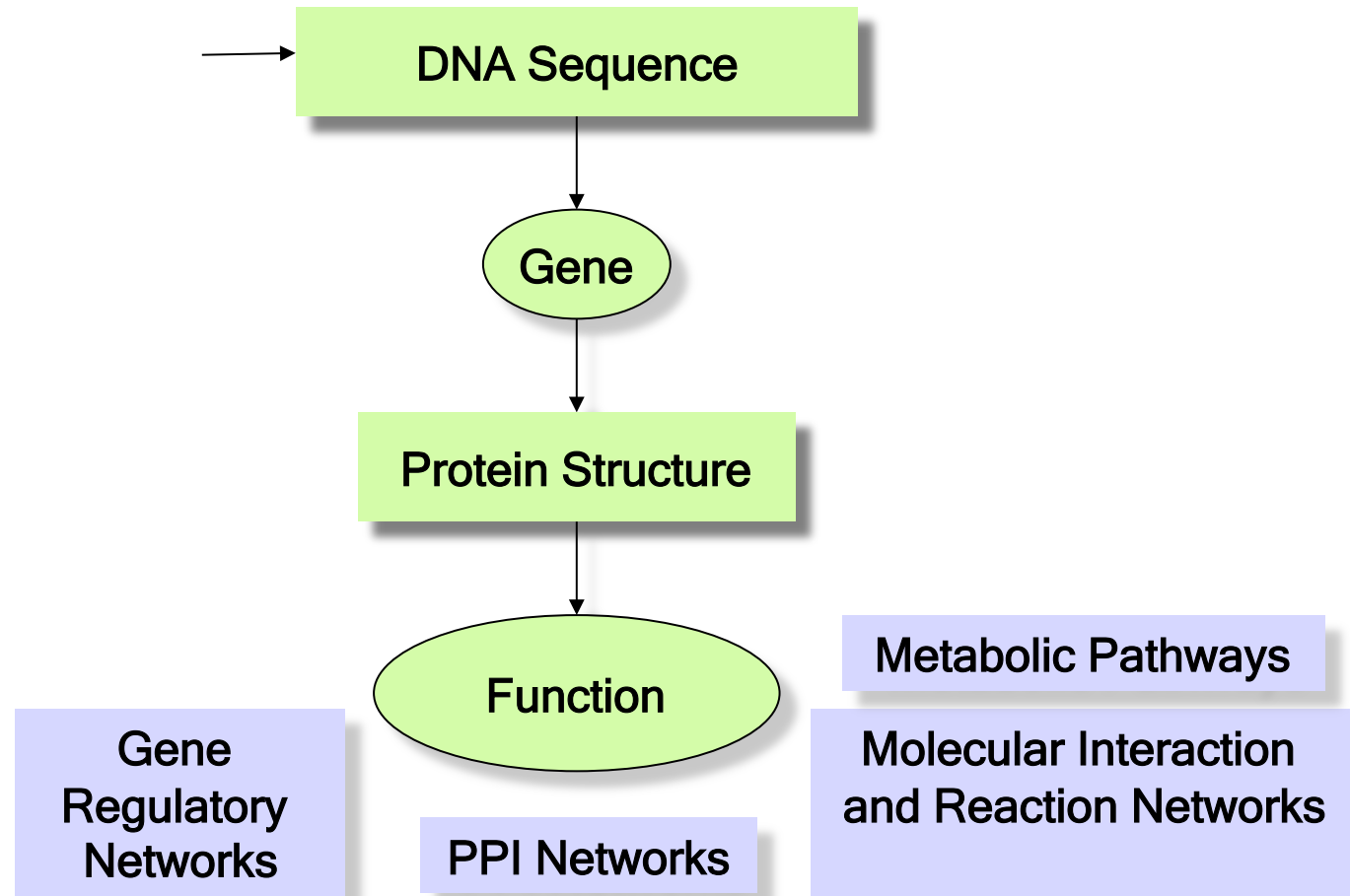
## 2. The different aspects of Informatics?

- 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

- ❑ **GenBank** Release 163/178 (Dec 2007/Aug 2009) contains over 80/108 million sequence entries totaling over 83/106 Gb from over 2,500/12,000 organisms [<http://www.ncbi.nlm.nih.gov>] (Storage: ~150/476 GB uncompressed)
- ❑ **Human Genome** has ~3 billion bp with 32,000+ genes.
- ❑ 1099 complete **bacterial** (86 **archaeal**) genomes sequenced
- ❑ 3041 **Viral** genomes (300bp - 300Kb) (1<sup>st</sup> 1978: Simian virus; 5Kb).
- ❑ 22/38 complete **eukaryotic** genomes sequenced:
  - Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa, Plasmodium falciparum, Drosophila melanogaster*
- ❑ 428 organisms have assemblies and chromosomal maps including:
  - Anopheles gambiae, Macaca mulatta, Bos taurus, Felis catus, Gallus gallus*
- ❑ **UniProtKB/Swiss-Prot** Release 54.7/2010\_06 (Jan'08/May'10): 333K/517K entries; 120/182 million amino acids.



# Short Homework

- What was the most recent large genome to be sequenced?
- List all mammals that have been sequenced. Also list their common names and their genome sizes.

# Genome Sizes

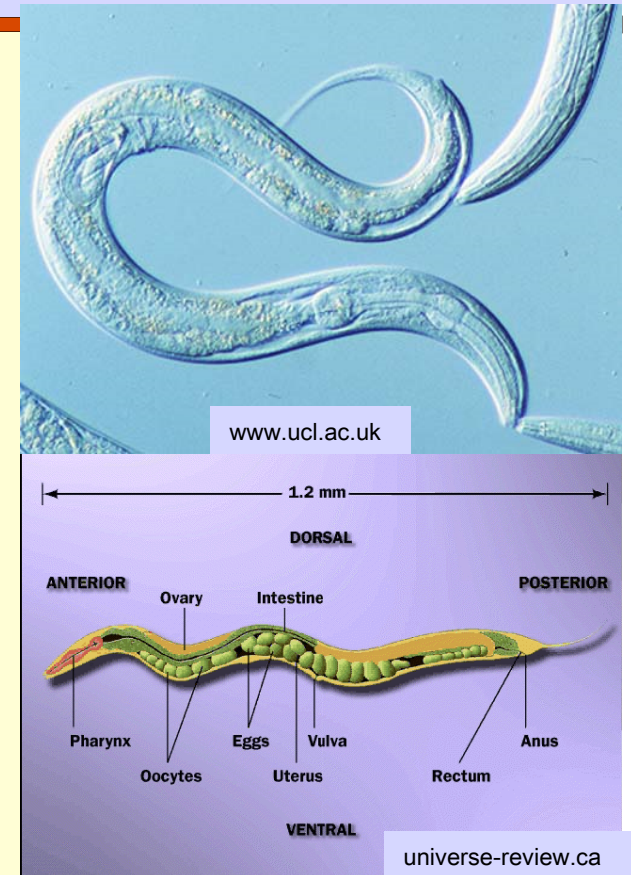
Organism	Size	Date	Est. # genes
<i>HIV type 1</i>	9.2 Kb	1997	9
<i>H. influenzae</i>	1.8 Mb	1995	1,740
<i>M. genitalium</i>	0.58 Mb	1998	525
<i>E. coli</i>	4.7 Mb	1997	4,000
<i>S. cerevisiae</i>	12.1 Mb	1996	6,034
<i>C. elegans</i>	97 Mb	1998	19,099
<i>A. thaliana</i>	100 Mb	2000	25,000
<i>D. melanogaster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

# Short Homework

- Find the organism with the largest genome known! How many chromosomes does it have?
- Do you think a larger genome implies a "more evolved" organism or a "less evolved" organism?

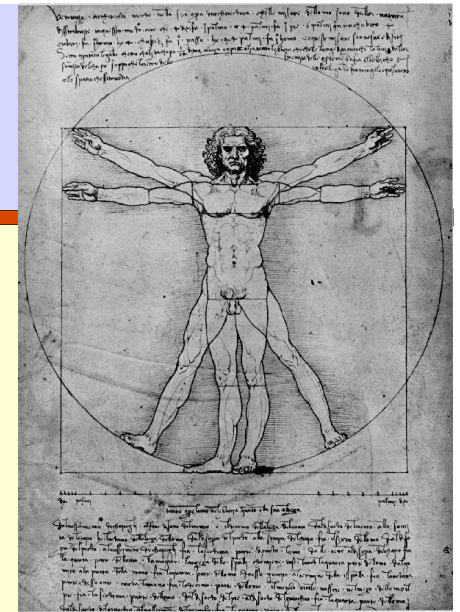
# Caenorhabditis Elegans

- ❑ Entire genome - 1998; 8 year effort
- ❑ 1<sup>st</sup> animal; 2<sup>nd</sup> 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)
MRNLPCLTAGGSGGLGGIAGKPSPTMEAVEASTASHRHSTSSYFATTYYHLTDDECHSGVNQLGGVVFVGG
RPLPDSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYETGSIRPRAIGGSKPRVATAEVSISKIS
QYKRECPSIFAWIIRDRLLEQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTSAGNSISAKVSV
SIGGNVSNVASGSRGTLSSSTDLMQTATPLNSESSEGGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEP
ARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSWPPRHYSWSWYPTSLSEIPISSAPNIASVTAY
ASGPSLAHSLSPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDD
QARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFAERERLAGKIGLPEARIQVWFSNRRAKWRREEK
LRNQRRTPNSTGASATSSSTSATASLTDSFNLSLACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPTL
GAGIDSESPTPIPHIRPCTSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHAQGHALVPAISP
RLNFNSGSGFGAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSPIPQQDLTPSSLYPCHMTLRP
PPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGASCSGSGYEVL SAYALPPPMASSAADSSSFAASSAS
ANVTPHHTIAQESCPSPCSSASHFGVAHSSGFSSDPI SPAVSSYAHMSYNYASSANTMTPSSASG TSAHV
APGKQOFFASC FYSPWV
```

```
>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
MQNSHSGVNQLGGVVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYETGSIRPRA
IGGSKPRVATPEVVS KIAQYKRECPSIFAWIIRDRLLEQENVCTNDNIPSVSSINRVLRNLASEKQQMGAD
GMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKL
QRNRTSFTQEIQIEALEKEFERTHYPDVFAERERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASN
TPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQ
TSSYSCLMPTSPSVNGRSYD TYTPPHMQTHMNSQPMGTS GTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ
```

# Drosophila Eyeless vs. Human Aniridia

```
Query: 57 HSGVNQLGGV FVGG RPLPDSTRQKIVELAHSGARPCDISRILQVSN GCVSKILGRYYETG 116
          HSGVNQLGGV FV GRPLPDSTRQKIVELAHSGARPCDISRILQVSN GCVSKILGRYYETG
Sbjct: 5  HSGVNQLGGV FVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSN GCVSKILGRYYETG 64

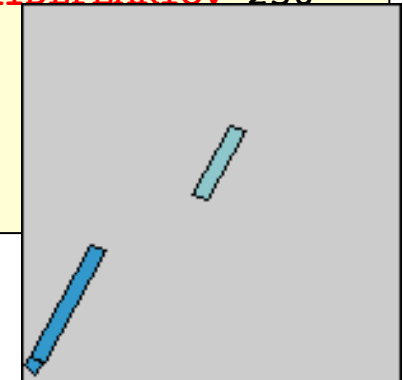
Query: 117 SIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAW EIRDRL LQENVCTNDNIPSVSSIN 176
          SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAW EIRDRL L E VCTNDNIPSVSSIN
Sbjct: 65 SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAW EIRDRL LSEG VCTNDNIPSVSSIN 124

Query: 177 RVLRLNLA AQKEQ 188
          RVLRLNLA ++K+Q
Sbjct: 125 RVLRLNLA SEKQQ 136
```

```
Query: 417 TEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFER THYPDVFARERLAGKIGLPEARIQV 476
          +++ Q RL LKRKLQRNRTSFT +QI++LEKEFER THYPDVFARERLA KI LPEARIQV
Sbjct: 197 SDEAQMRLQLKRKLQRNRTSFTQE QIEALEKEFER THYPDVFARERLAAKIDLPEARIOV 256

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

E-Value =  $2e^{-31}$



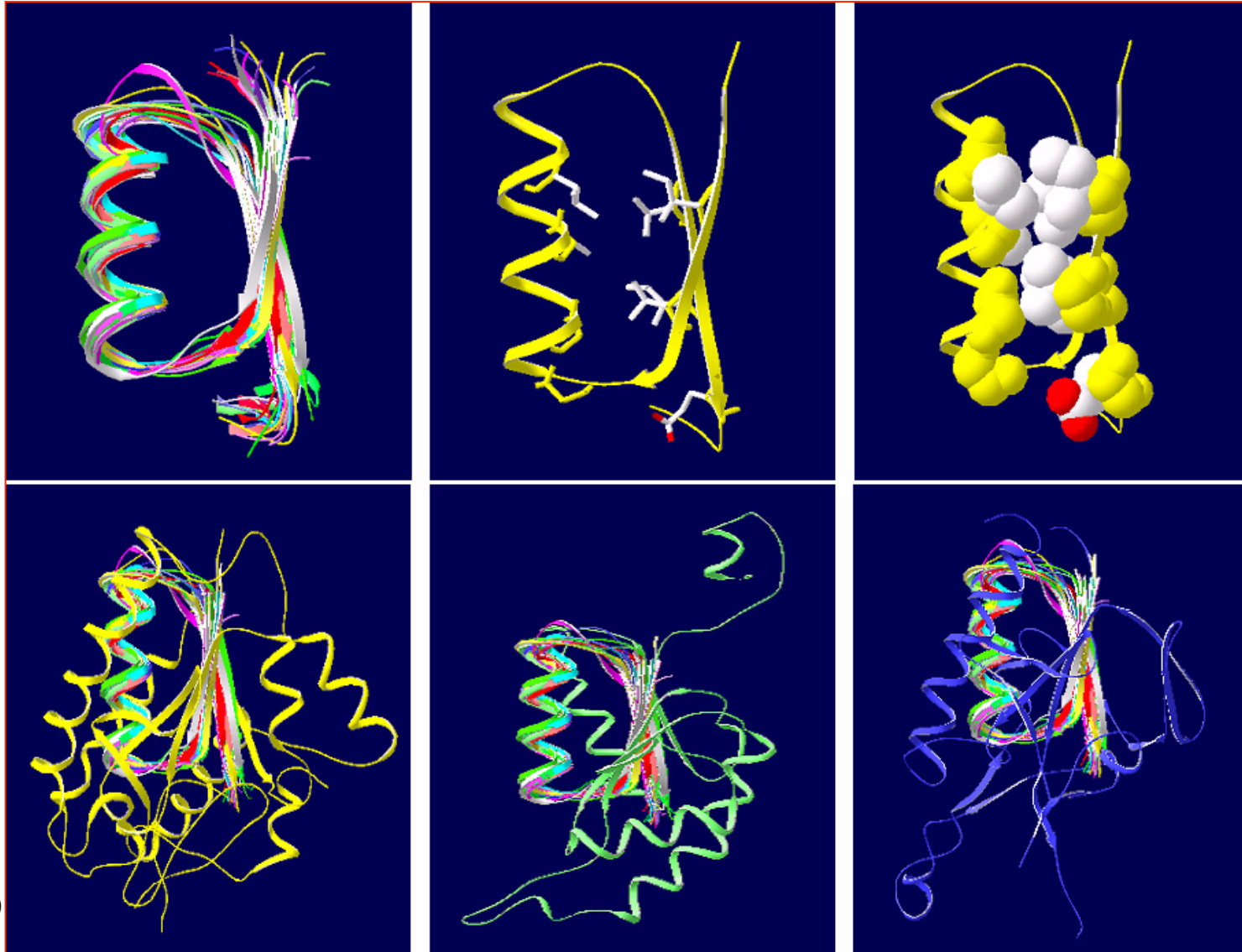
# Motif Detection in Protein Sequences

❑ MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLYHGDLEAA  
KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR  
LVSFVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFNLRKTKQRLGWFN  
QDEVEMVARELGVTSKDVREME SRMAAQDMTFDLSSDDSDSQPMAPVLY  
LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK  
STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

❑ MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLYHGDLEAA  
KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR  
LVSFVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFNLRKTKQRLGWFN  
QDEVEMVARELGVTSKDVREME SRMAAQDMTFDLSSDDSDSQPMAPVLY  
LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK  
STLQELADRYGVSAERVRQLEK NAMKKLRAAIEA

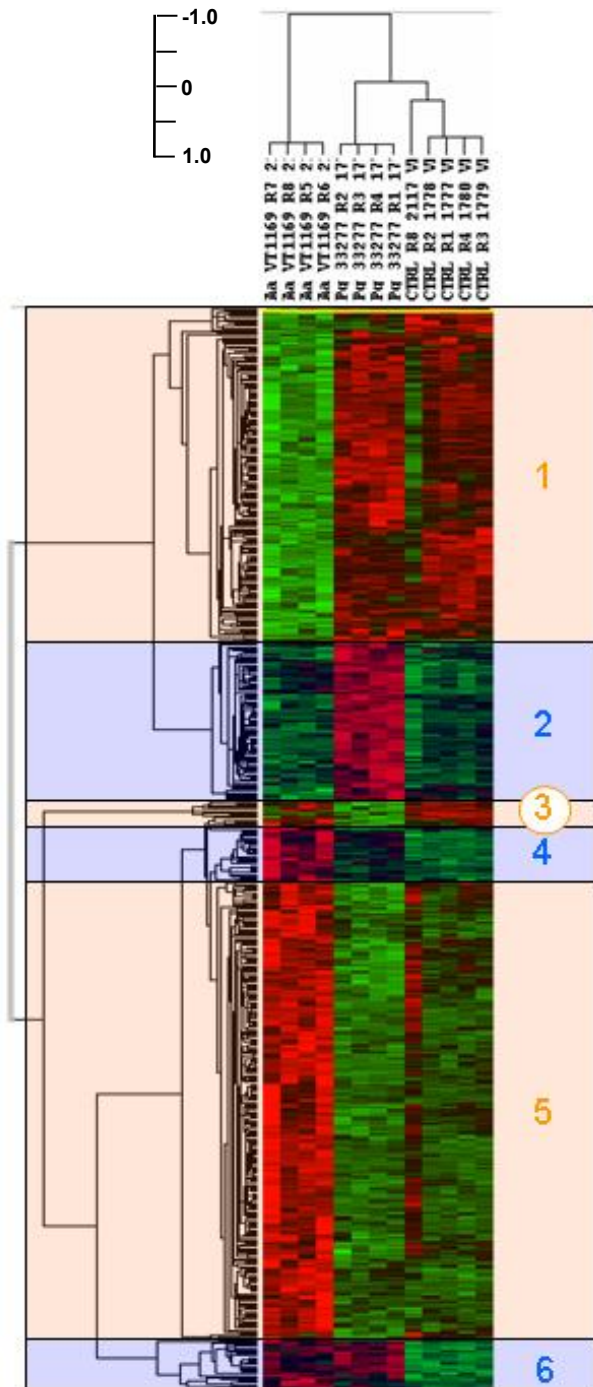


# Patterns in Protein Structures



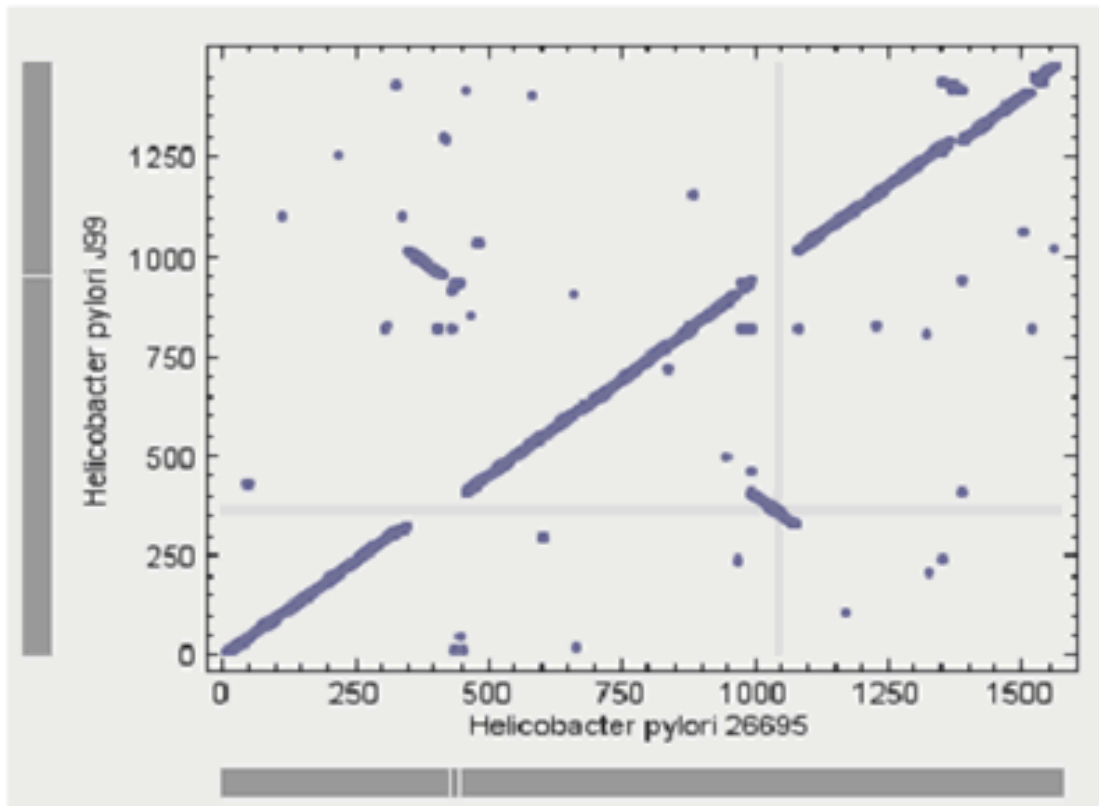
# Microarray Analysis

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



# Tools: GenePlot

1491 proteins total



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

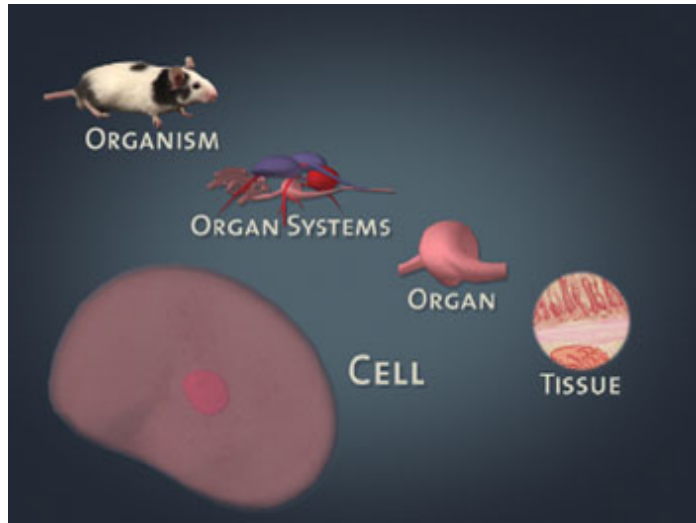


- ❑ 18000 Amish people in Pennsylvania
- ❑ Mostly intermarried due to religious doctrine
- ❑ rare recessive diseases occurred with high frequencies.
- ❑ SIDS: 3000 deaths/year (US); 21 deaths (Amish community)
- ❑ Many research centers failed to identify cause
- ❑ Collaboration between Affymetrix, TGEN & 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](http://www.affymetrix.com/community/wayahead/modern_miracle.affx)

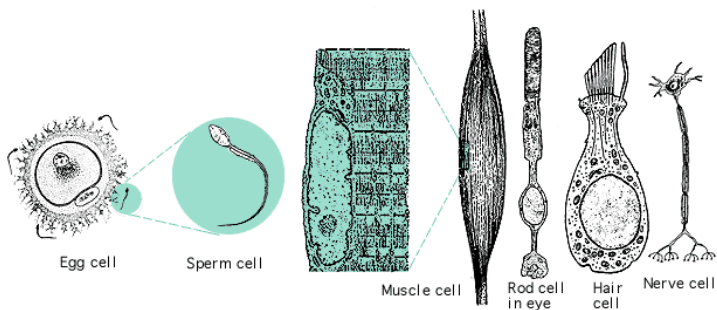
---

# Molecular Biology Background

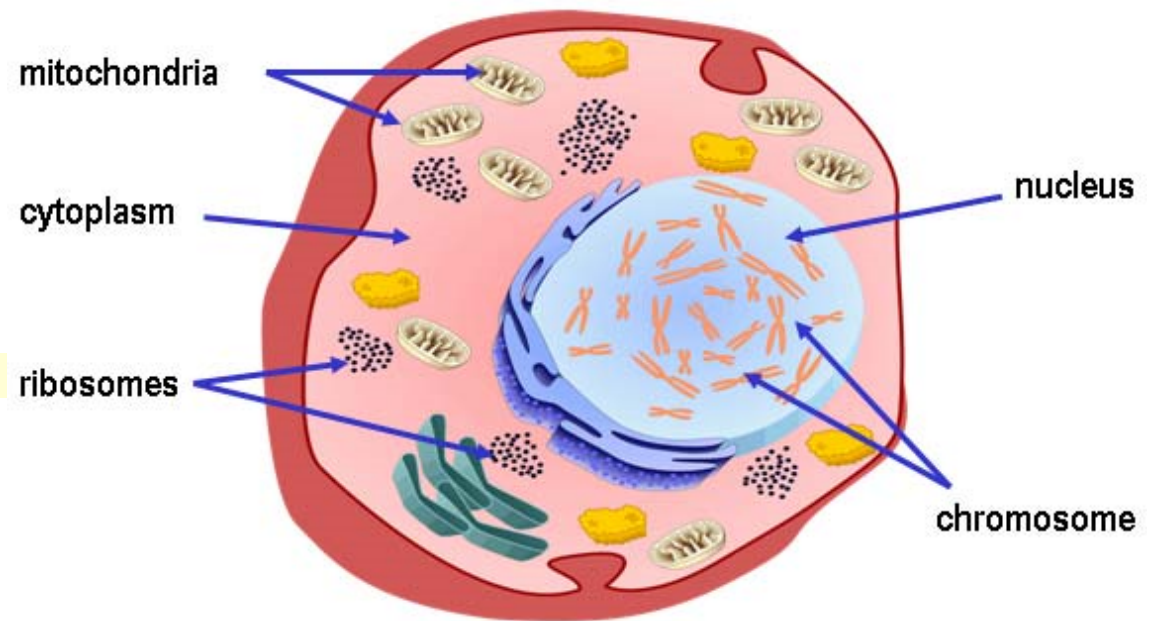
# Cell



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

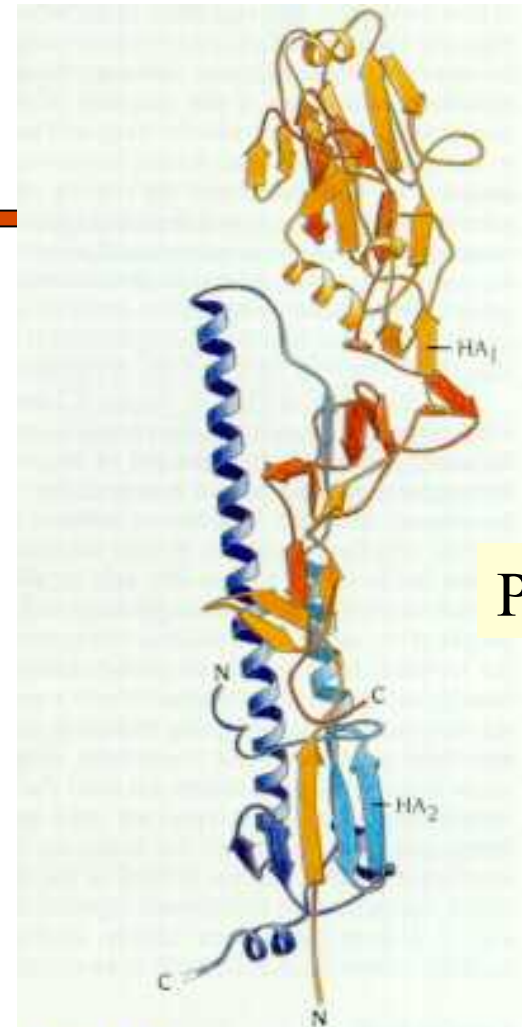
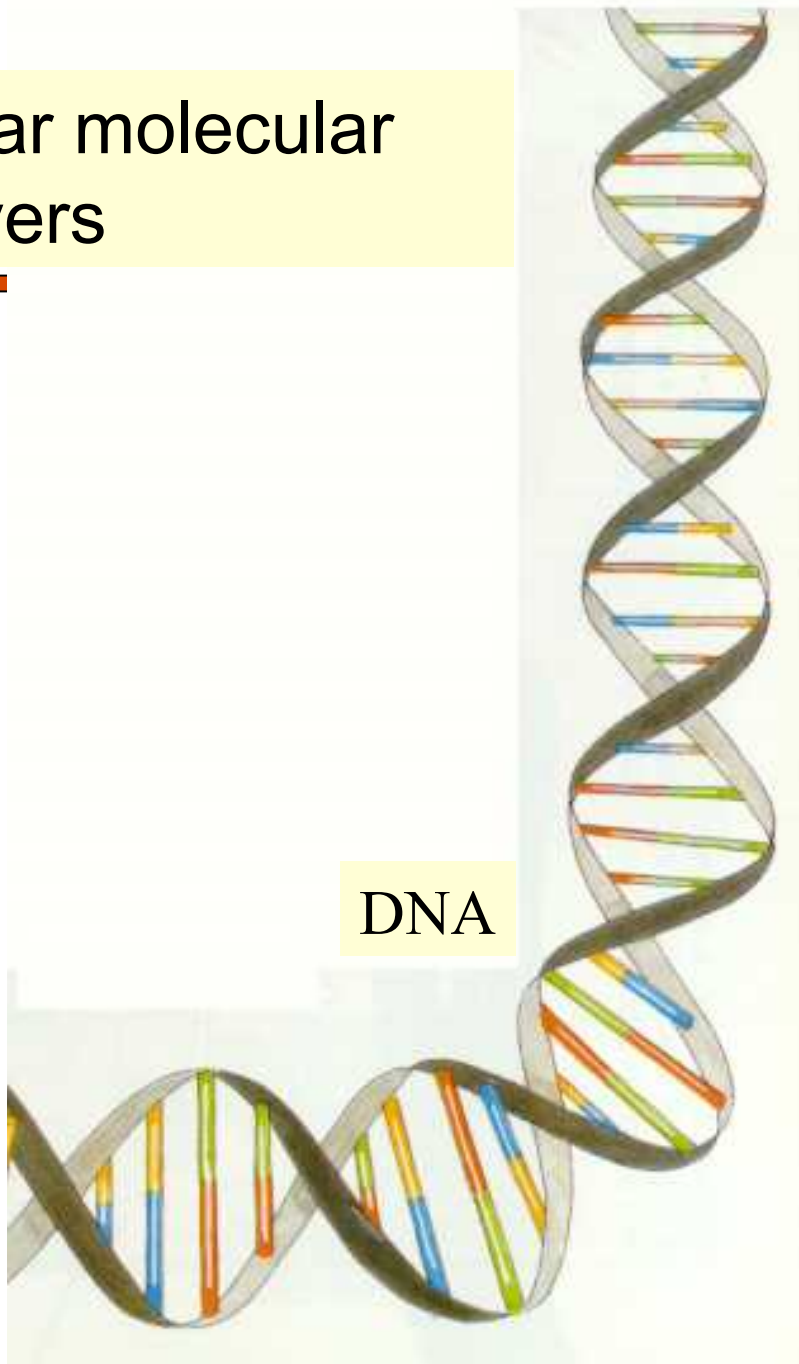


<http://www.biology.eku.edu/RITCHISO/301notes1.htm>



[http://www.biotechnologyonline.gov.au/popups/img\\_cellwithlabels.cfm](http://www.biotechnologyonline.gov.au/popups/img_cellwithlabels.cfm)

## 2 star molecular players



**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 HA<sub>1</sub> (red) and HA<sub>2</sub> (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  $\alpha$  helices known in a globular structure, about 75Å long. The globular head is formed by residues only from HA<sub>1</sub>. (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  cgggagaagg  aggaggaggc  gaagaaaagc  aacagaagcc  cagttgctgc
61  tccaggtccc  tcggacagag  ctttttccat  gtggagactc  tctcaatgga  cgtgccccct
121 agtgcttctt  agacggactg  cggctctccta  aaggctcgacc  atgggtggccg  ggaccctgctg
181 tcttctagtg  ttgctgcttc  cccaggctcct  cctgggcggc  gcggccggcc  tcattccaga
241 gctgggccgc  aagaagttcg  ccgcggcatc  cagccgacc  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

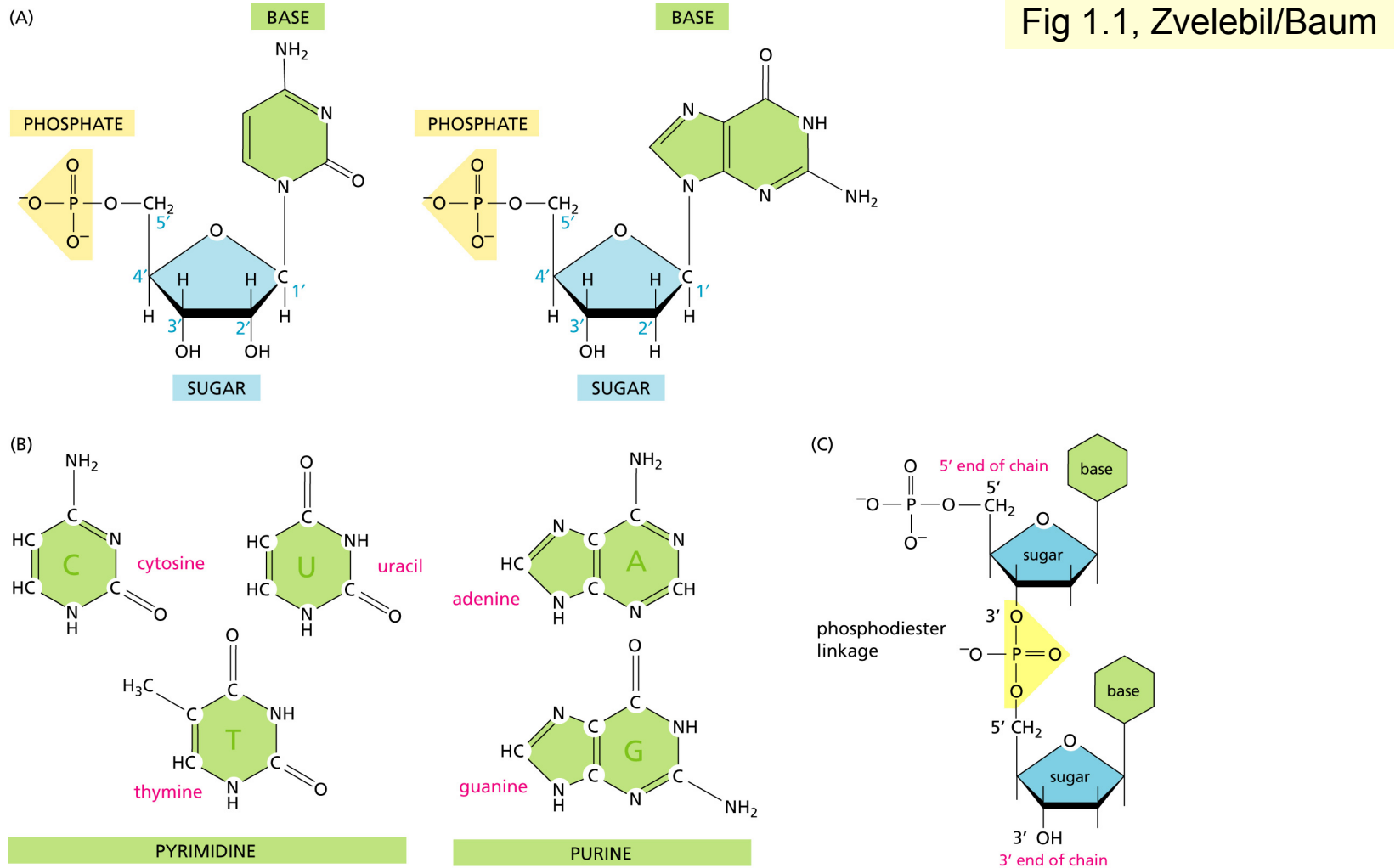
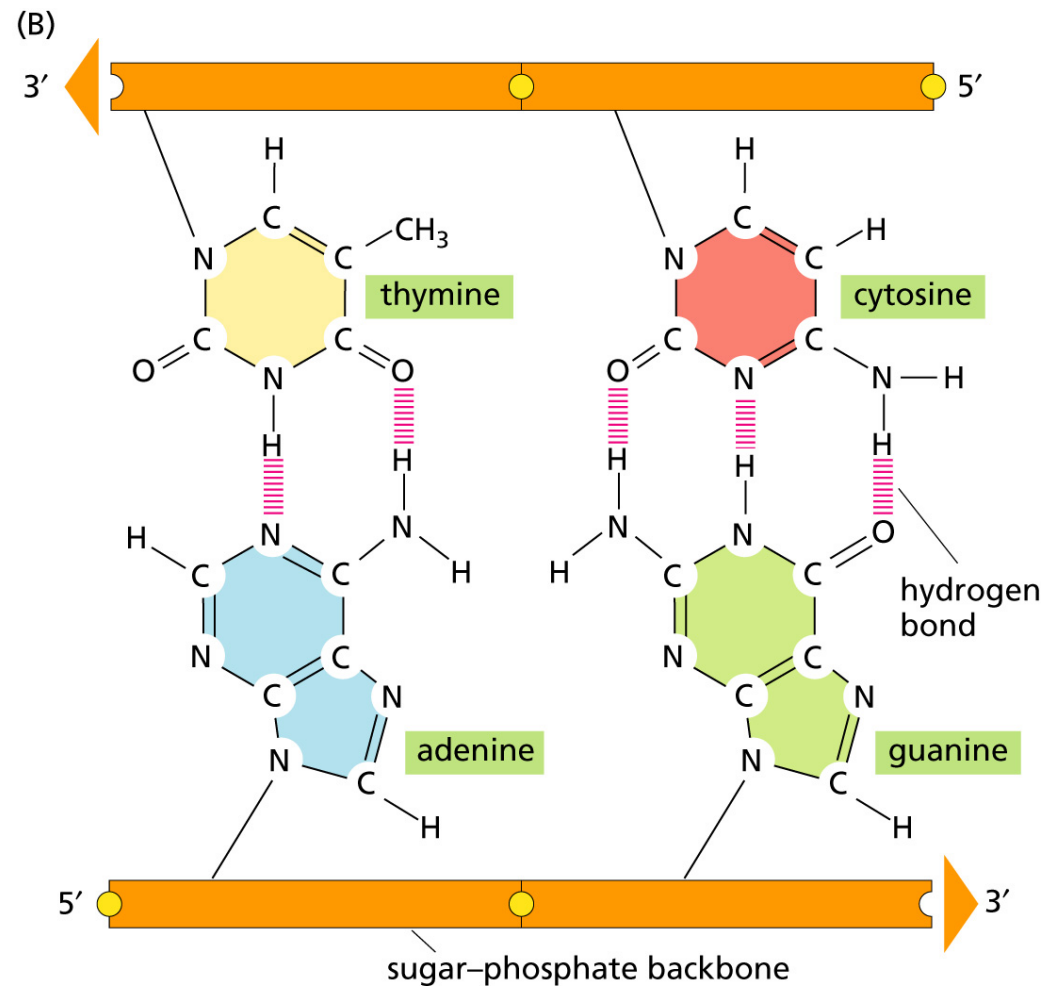
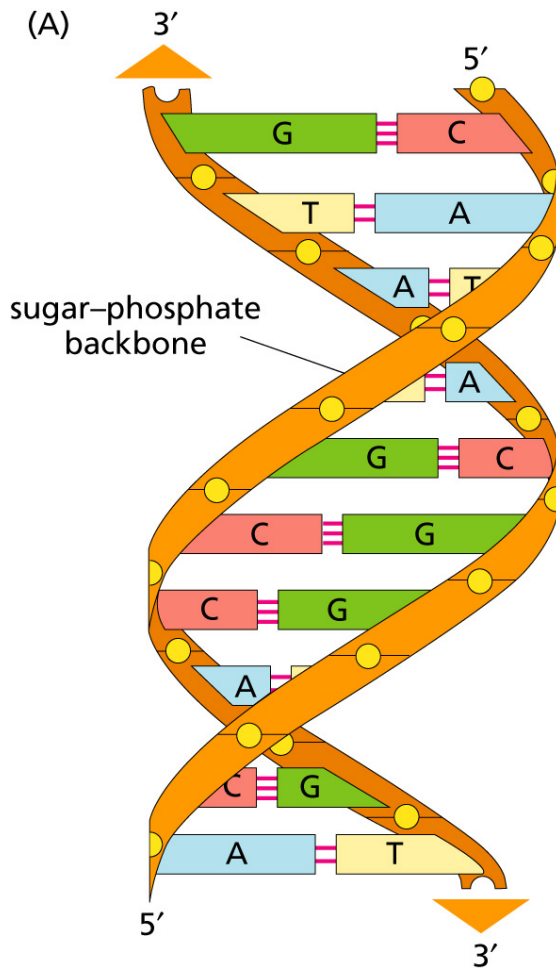


Fig 1.1, Zvelebil/Baum

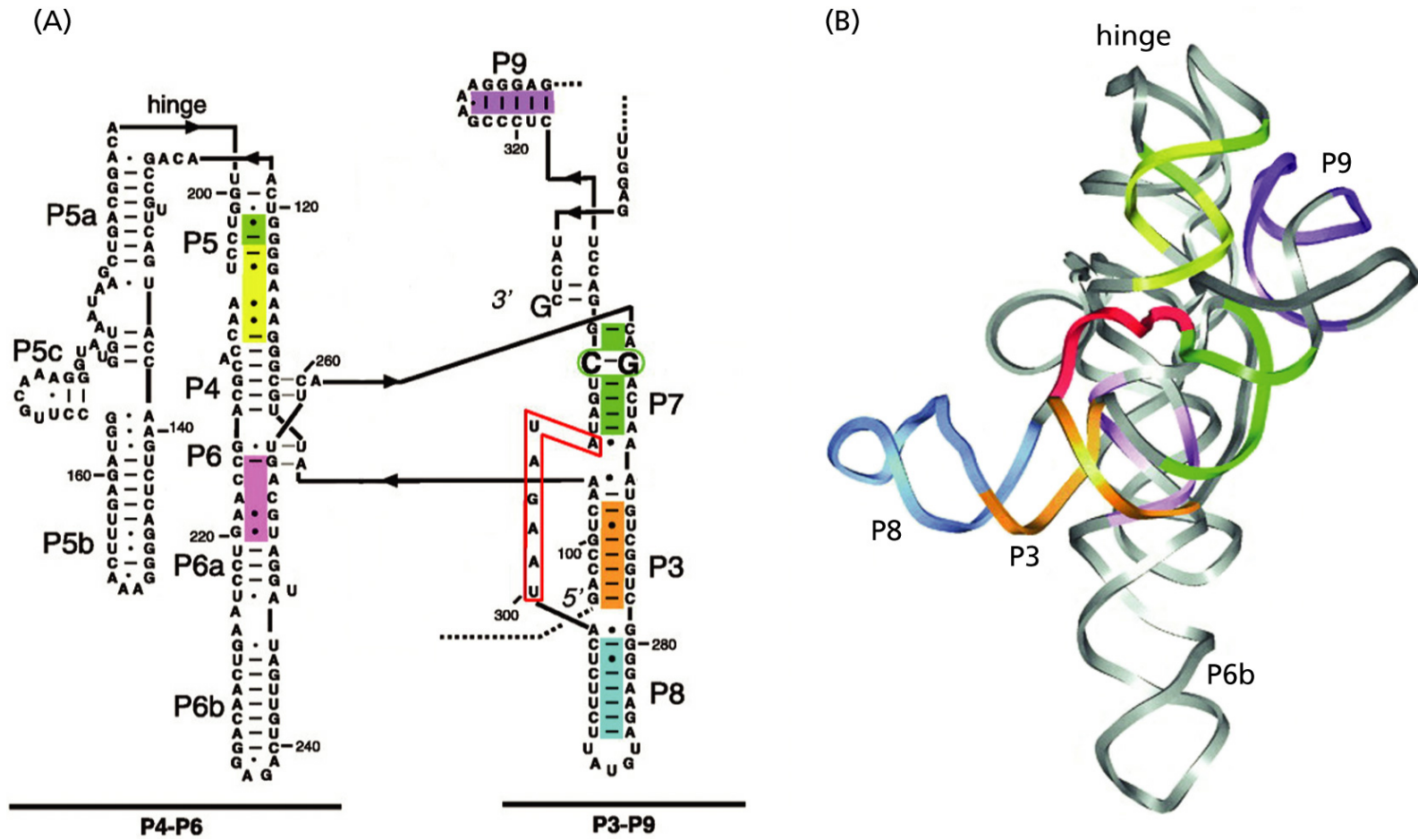
# DNA double helix structure

Fig 1.3, Zvelebil/Baum



# RNA molecule

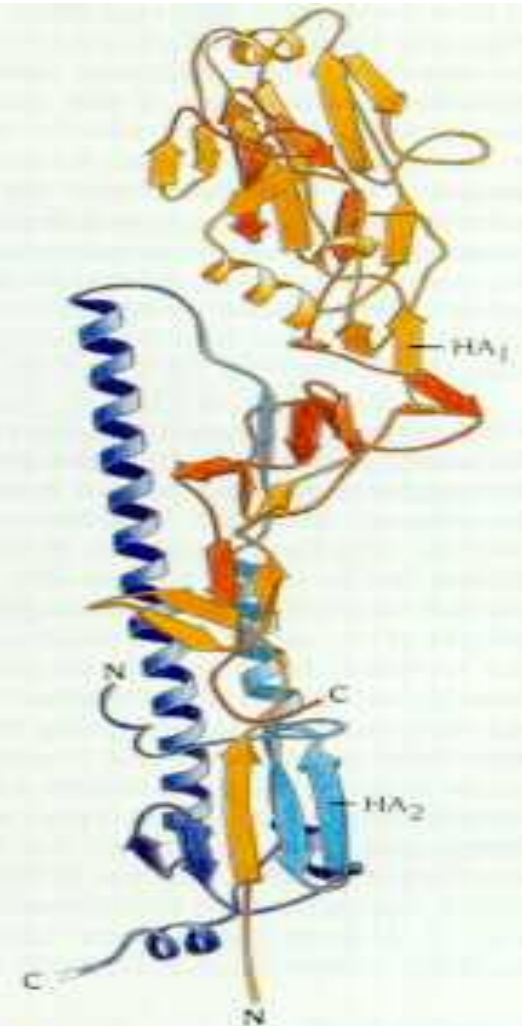
Fig 1.5, Zvelebil/Baum



# Typical protein sequence

```
/translation="MVAGTRCLLVLLLPQVLLGGAAGLIPELGRKKFAAASSRPLSRP  
SEDVLSEFELRLLSMFGLKQRPTPSKDVVVPPYMLDLYRRHSGQPGAPAPDHRLERAA  
SRANTVRSFHHEEAVEELPEMSGKTARRFFFNLSSVPSDEFLLTSAELQIFREQIQEAL  
GNSSFQHRINIYEI IKPAAANLKFVTRLLDTRLVNQNTSQWESFDVTPAVMRWTTQG  
HTNHGFVVEVAHLEENPGVSKRHVRI SRSLHQDEHSWSQIRPLLVTFGHDGKGHPLHK  
REKRQAKHKQRKRLKSSCKRHPLYVDFSDVGWNDWIVAPPGYHAFYCHGECPFPLADH  
LNSTNHAI VQTLVNSVNSKIPKACCVPTELSAISMLYLDENEKVVLKKNYQDMVVEGCG  
CR"
```

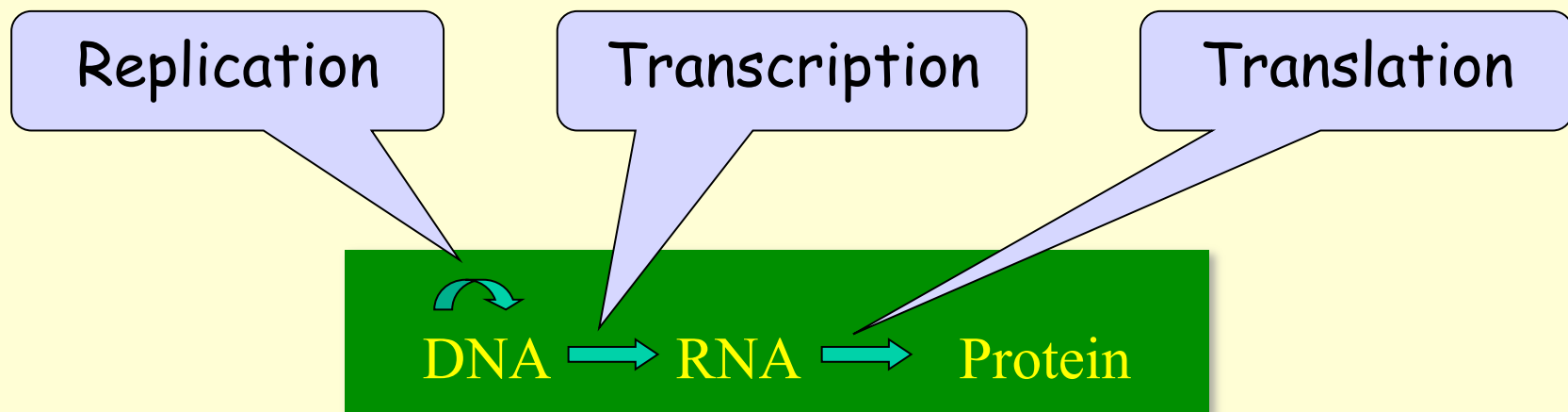
# Protein 3D Structure



**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 HA<sub>1</sub> (red) and HA<sub>2</sub> (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  $\alpha$  helices known in a globular structure, about 75 Å long. The globular head is formed by residues only from HA<sub>1</sub>. (Courtesy of Don Wiley, Harvard University.)

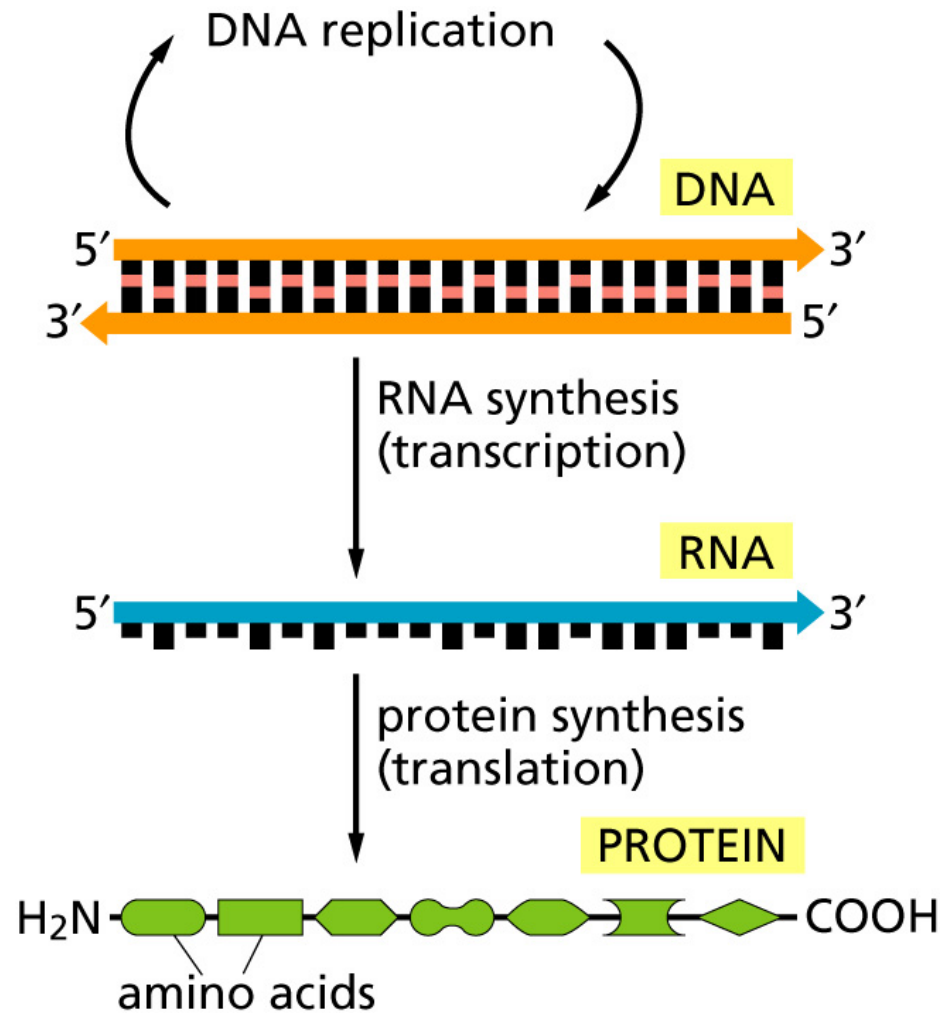
# Central Dogma

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



# Central Dogma

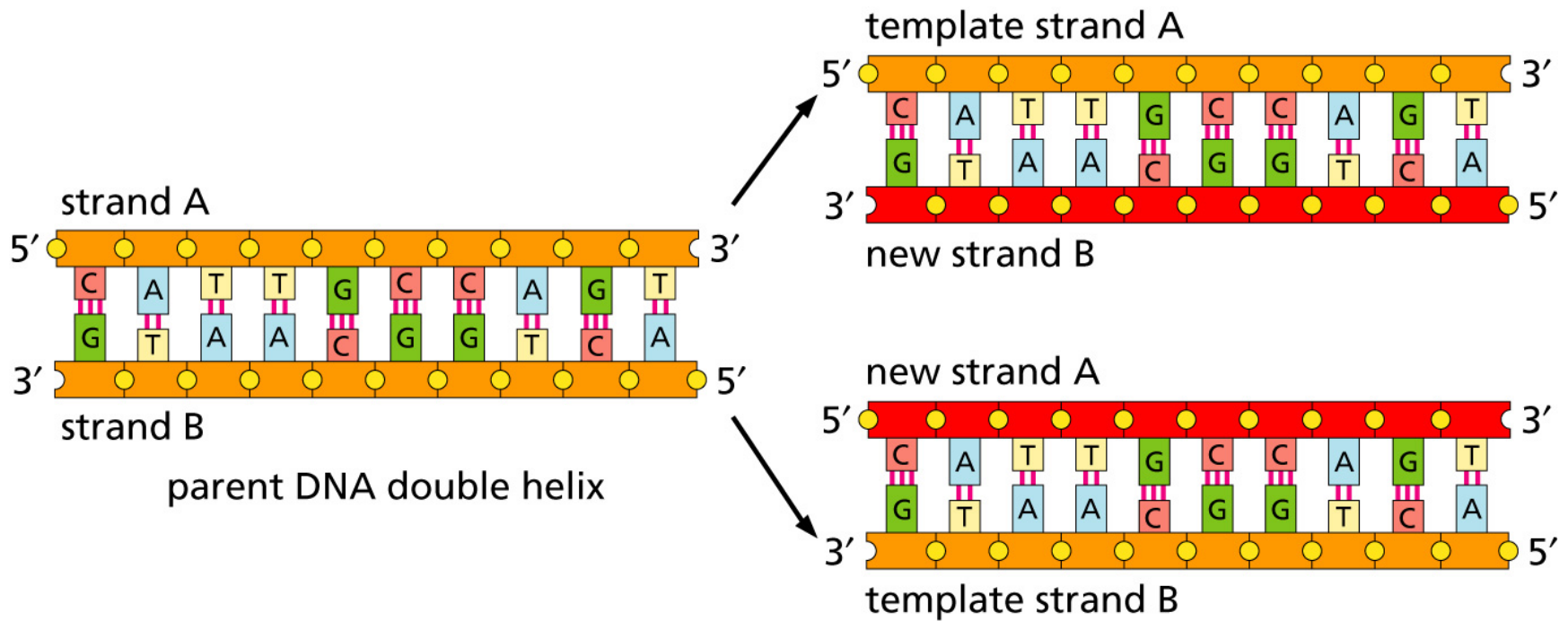
Fig 1.6, Zvelebil/Baum





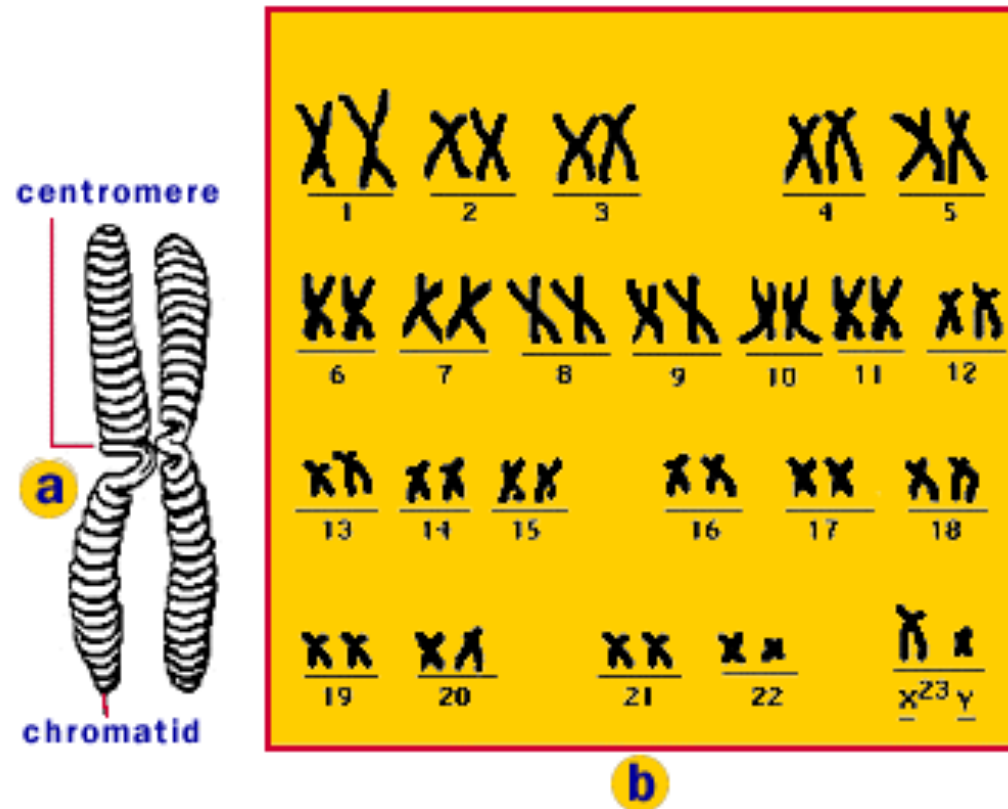
# DNA Replication

Fig 1.4, Zvelebil/Baum



# Chromosomes

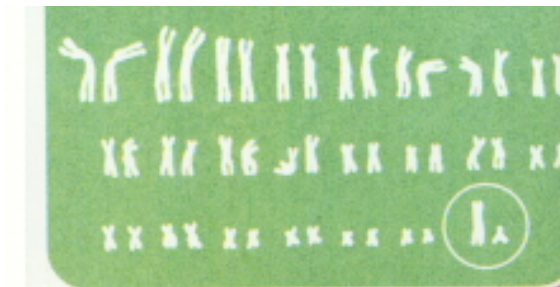
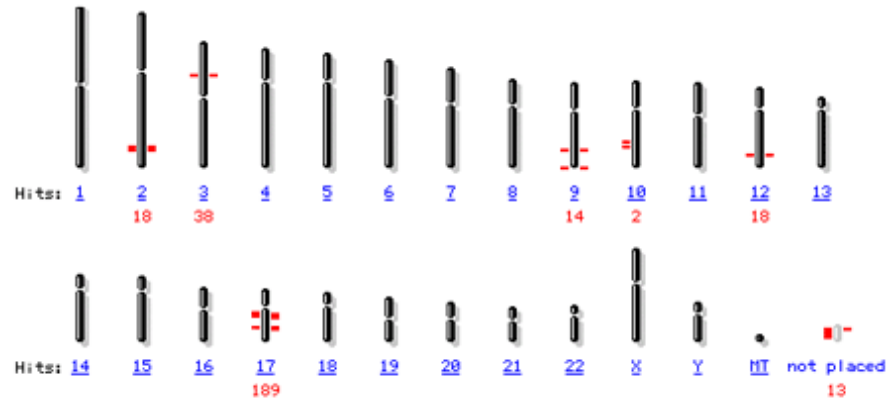
## Human chromosomes!



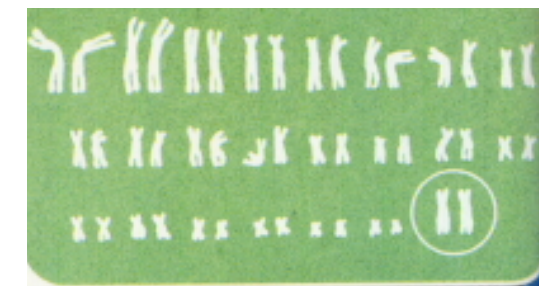
# Chromosomes

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

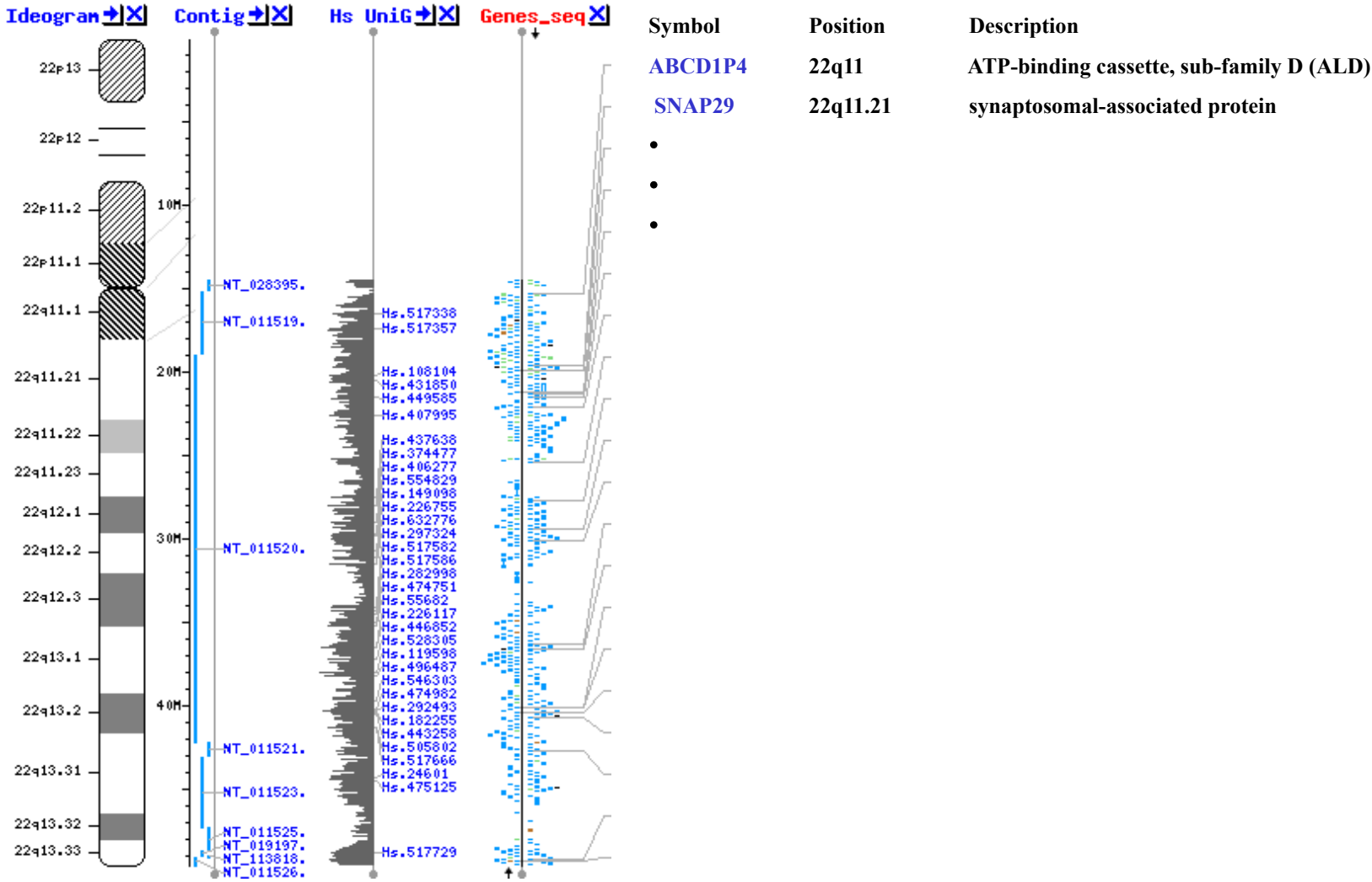
Build 36.2 statistics [Switch to previous build](#)



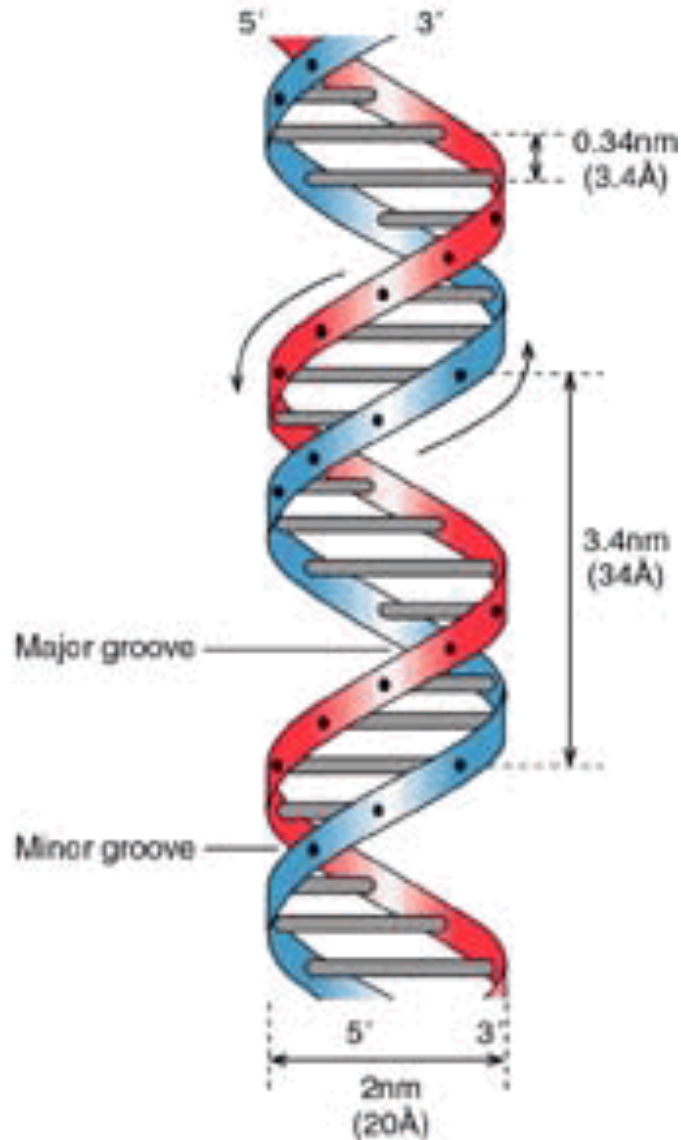
The chromosomal locations of several genes believed to be associated with the human BRCA1 gene implicated in breast cancer are highlighted.



# Human Chr 22



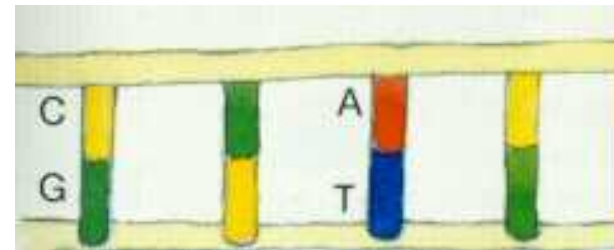
# DNA Molecule



# DNA



## Complementary Bases

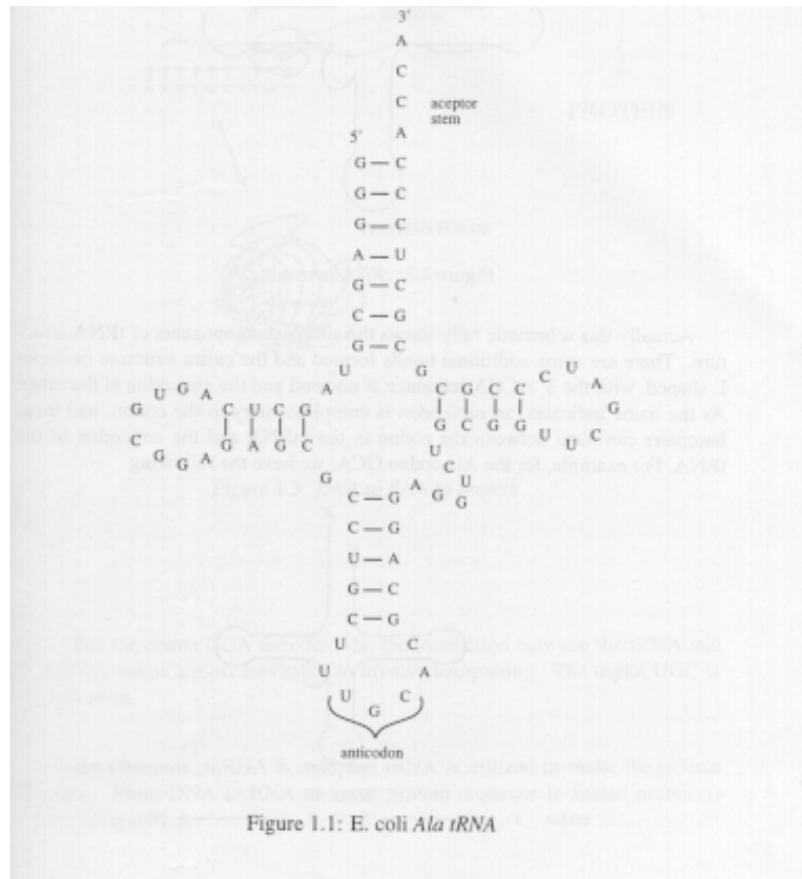


# Proteins – Amino acids

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

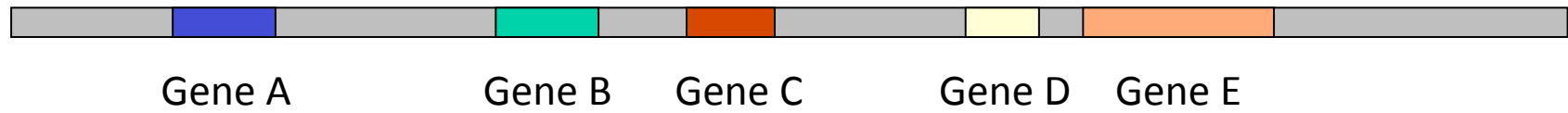
# RNA



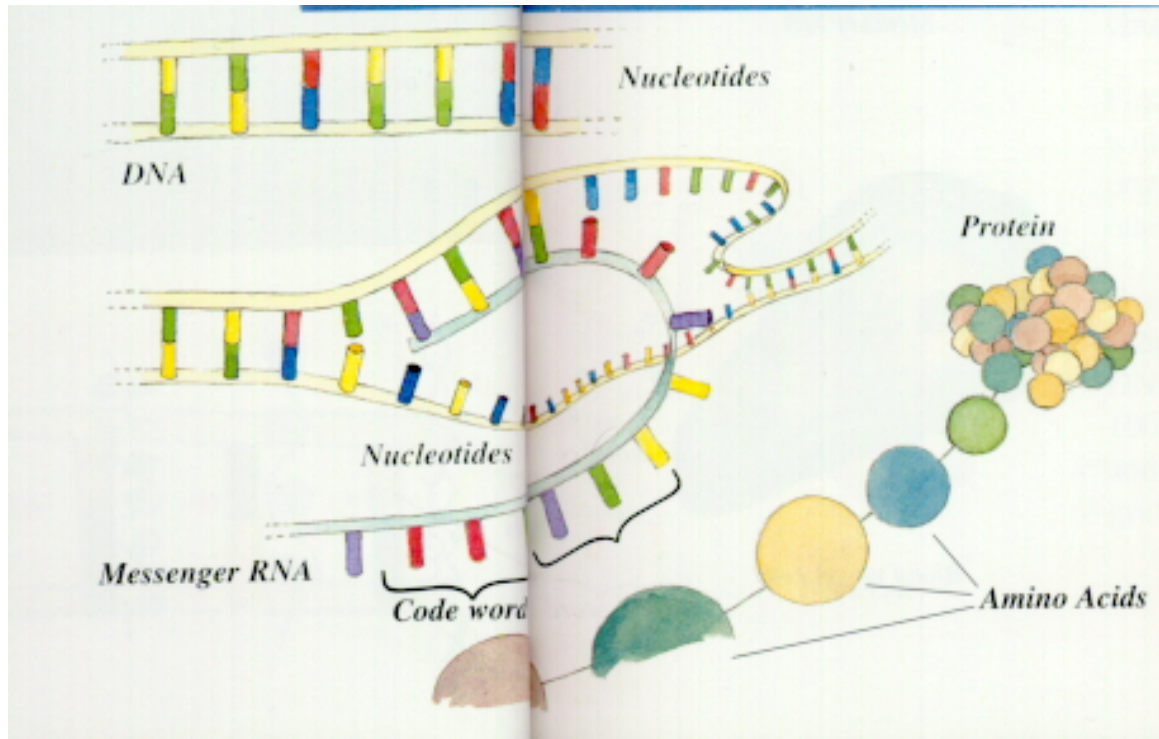


# Genes

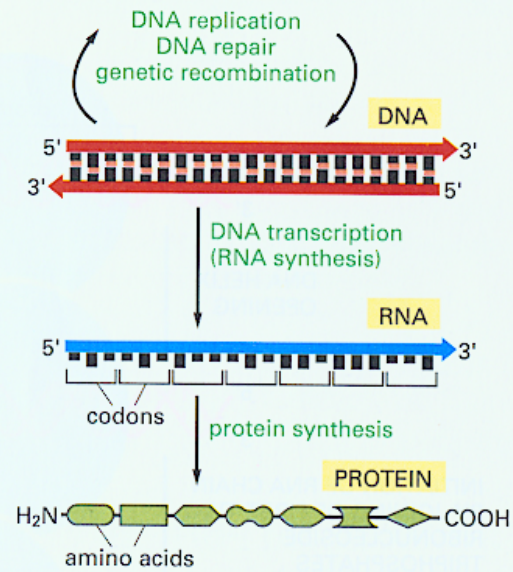
DNA



  
DNA → RNA → Protein

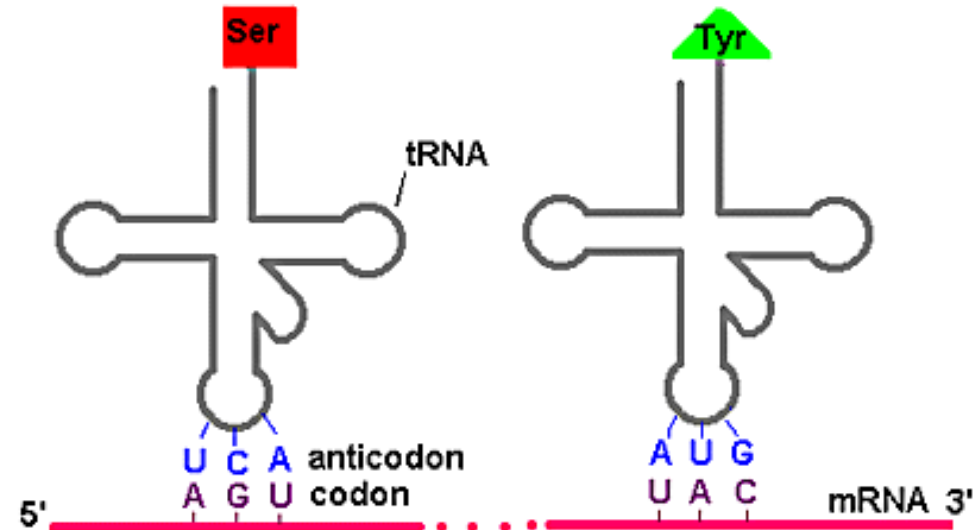


# Basic Genetic Processes



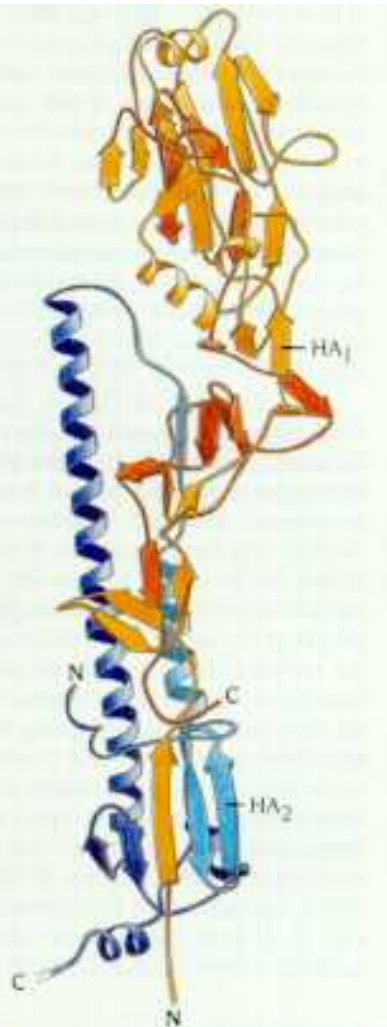
**Figure 6-1 The basic genetic processes.** The processes shown here are thought to occur in all present-day cells. Very early in the evolution of life, however, much simpler cells probably existed that lacked both DNA and proteins (see Figure 1-11). Note that a sequence of three nucleotides (a codon)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

# The Genetic Code

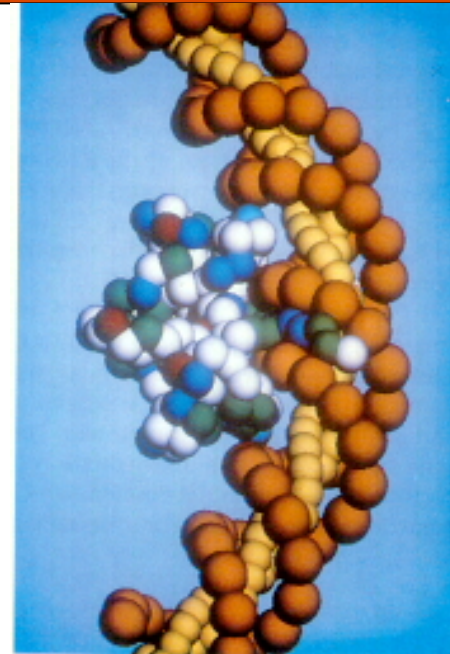


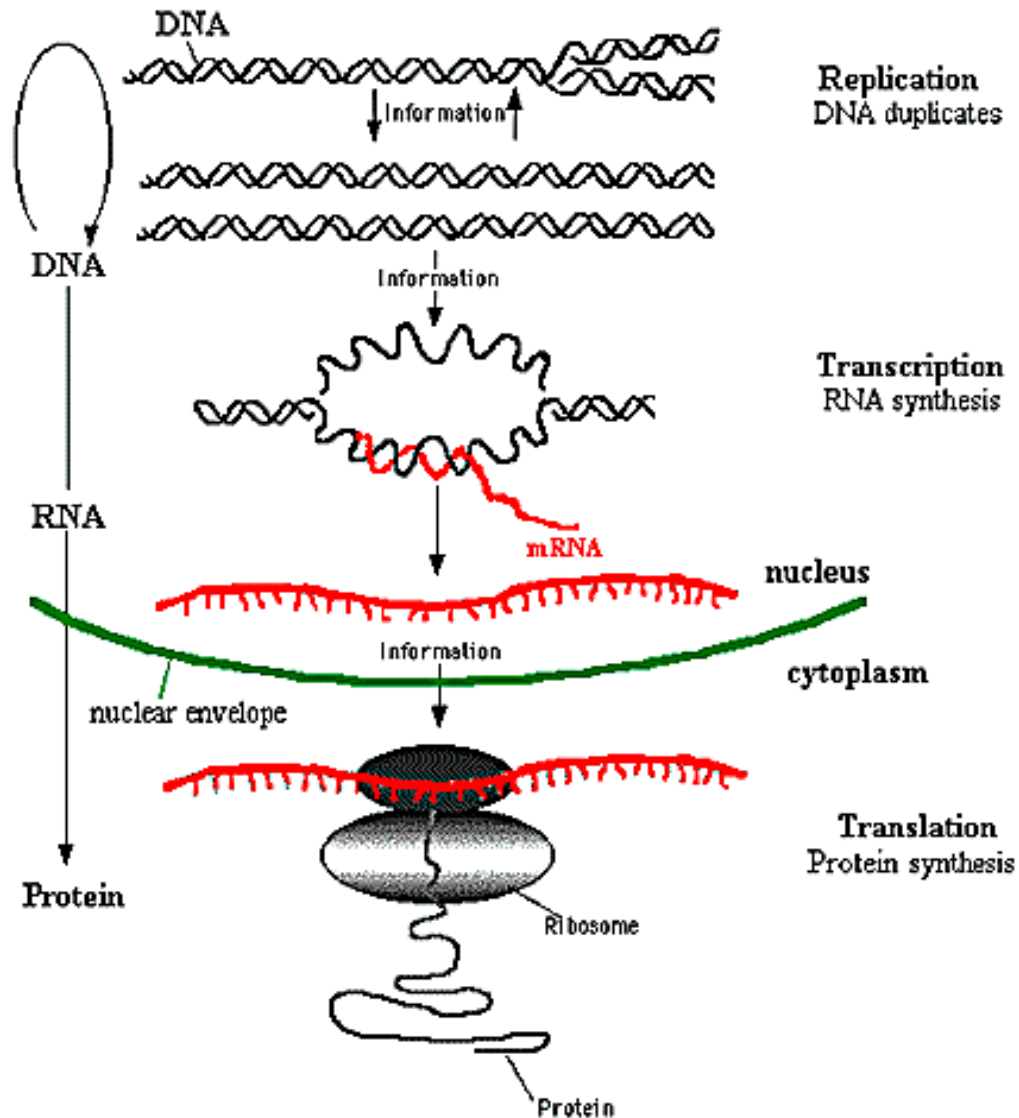
		2nd base in codon				
		U	C	A	G	
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
	G	Val Val Val Val	Ala 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 HA<sub>1</sub> (red) and HA<sub>2</sub> (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  $\alpha$  helices known in a globular structure, about 75Å long. The globular head is formed by residues only from HA<sub>1</sub>. (Courtesy of Don Wiley, Harvard University.)





### The Central Dogma of Molecular Biology

# Transcription

Fig 1.7, Zvelebil/Baum

