BSC 4934: Q’BIC Capstone Workshop

Giri Narasimhan
ECS 254A; Phone: x3748
giri@cs.fiu.edu
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

- DNA Sequence
  - Gene
  - Protein Structure
  - Function
  - Gene Regulatory Networks
  - PPI Networks
  - Metabolic Pathways
  - Molecular Interaction and Reaction Networks
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


- **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.
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

<table>
<thead>
<tr>
<th>Organism</th>
<th>Size</th>
<th>Date</th>
<th>Est. # genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV type 1</td>
<td>9.2 Kb</td>
<td>1997</td>
<td>9</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>1.8 Mb</td>
<td>1995</td>
<td>1,740</td>
</tr>
<tr>
<td>M. genitalium</td>
<td>0.58 Mb</td>
<td>1998</td>
<td>525</td>
</tr>
<tr>
<td>E. coli</td>
<td>4.7 Mb</td>
<td>1997</td>
<td>4,000</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>12.1 Mb</td>
<td>1996</td>
<td>6,034</td>
</tr>
<tr>
<td>C. elegans</td>
<td>97 Mb</td>
<td>1998</td>
<td>19,099</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>100 Mb</td>
<td>2000</td>
<td>25,000</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>180 Mb</td>
<td>2000</td>
<td>13,061</td>
</tr>
<tr>
<td>M. musculus</td>
<td>3 Gb</td>
<td>2002</td>
<td>~30,000</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>3 Gb</td>
<td>2001</td>
<td>32,000+</td>
</tr>
</tbody>
</table>
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)
MRNLPCLGTAGGSGLGGIAGKPSPTMEAVEASTASHRHTSSYFATTYYHLTDDECHSGLVQGLGFGVGG
RPLPDSTRQKIVELAHSGARPCDISRLQVSNGCVSKILGRYETGSIRPAIGGSKPRVATAEVSIS
QYKRECPSIFAWEIRDRLQLQENVCTNDNIPSVSSINRVLRNLAAQEQQSTSGSSSTSAGNSISAKSVSV
SIGGNVSNNVAGSRGLTSSSTDLMQTATPLNSSESAGASNSGEGSEQEIAYEKRLRNQVAHAAPGPLEP
ARAAPLVGQSPNHGTRSSHPQLVQHNGHQALQQHQQQSWPPRHYSGSWYPTSLSEIPISSAPNIASVTAY
ASGSLAHSLPNDIESLASIGHQRNPCVATEDIHLCKELDHQSDETGSGEGENSNGGASNIGNTEDD
QARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFAERLAKGKGLPEARIQVWFSNRRAKRREEK
LRNQRRTPNNGATSSDTSSATASLTDSNPSNLSACSSLLSGASAGGPSVSTINGLSSPSTLSTNVAPTL
GAGIDSESEPTIPPHIRPSCTSDNDNQSQEDCCRVCSCPCPLGVVGHQNTTHIQSNHQAQGHALVPAISP
RLNFNSGSFGAMSNMHTALSMSDSYGAVPITPSFNHSAVGPLAPPSPIPQGDLTPSSLPHCMTRLP
PPMAHAPPVGDGGRPAGVGLGQGSANLGCSCSGSYEYLSAYALPPPMASASAADDSSFASSAS
ANVTPHTIAQECSPCSASSHFGVAHSSGFSSDPIPASVYAHMSYVASSANTMTFSSASGTSASHV
APGKQFFASCYSPFWV

>gi|61748889|PAX6_HUMAN Paired box protein (Ocularhormbin) (Aniridia, type II protein)
MQNSHVSNQLGGVFGVNGRPLPDSTRQKIVELAHSGARPCDISRLQVSNGCVSKILGRYETGSIRPA
IGGSKPRVATPEVSVSKIAQYKRECPSIFAWEIRDLLSEQVCTNDNIPSVSSINRVLRNLAAQEQQSTG
QMQSDYHTSFTQOEIEALEKEFERTHYPDVFAERLAAKIDLPEARIQVWFSNRRAKRREEKLRNQRA
QNRTSTQOEIEALEKEFERTHYPDVFAERLAAKIDLPEARIQVWFSNRRAKRREEKLRNQRRQASN
TPSHIPSSSFSSTSYVQIPQPTTPVSSFTSGMLGRTDTALTNTYSAALPPMAPFTSMAANLPMQPVPSQ
TSSYSCMLPTSPVNGRISYDITYTPHMQTHMNSQPMGTSHTSTSTGGLSPGSVPQVPQGTPMDQVWPR
IQ

7/11/11
### Drosophila Eyeless vs. Human Aniridia

<table>
<thead>
<tr>
<th>Query</th>
<th>57</th>
<th>HSGVNQLGGVFVGRPLPDSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYETG</th>
<th>116</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>117</td>
<td>SIRPRAIGGSKPRVATAEVVSISQYKRECPSIFAEWEIRDRLQENVCTNDNIPSVSSIN</td>
<td>176</td>
</tr>
<tr>
<td></td>
<td>177</td>
<td>RVLRNLAAQKEQ</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td>417</td>
<td>TEEDQARLILKRLQQRNRTSFTNDQIDGELKFEKFTPFDVFARERLAGKIGLPEARIQV</td>
<td>476</td>
</tr>
<tr>
<td></td>
<td>477</td>
<td>WFSNRRRAKWRREEKLRNQRR</td>
<td>496</td>
</tr>
<tr>
<td>Sbjct</td>
<td>5</td>
<td>HSGVNQLGGVFVGRPLPDSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYETG</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>SIRPRAIGGSKPRVAT EVVSISQYKRECPSIFAEWEIRDRLQENVCTNDNIPSVSSIN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>RVLRNLASEKQO</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td>197</td>
<td>SDEAQMRQQLKRLQQRNRTSFTQEQIEALEKFEKFTPFDVFARERLAGKIDLPEARIQV</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>257</td>
<td>WFSNRRRAKWRREEKLRNQRR</td>
<td>276</td>
</tr>
</tbody>
</table>

**E-Value = 2e-31**
Motif Detection in Protein Sequences

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLLEAAKTLILSHLRFVVHIARYAGYGLPQADLIQEGINIQLMKAARRNFNPDEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTQAKRKLFFNLKRTKLQRLGWFNQDEVEMVARELGVTSDKVREMESRMMAQDMTDFLSSDDDSDSQPMAPVLYLQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDENKNSTLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLLEAAKTLILSHLRFVVHIARYAGYGLPQADLIQEGINIQLMKAARRNFNPDEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTQAKRKLFFNLKRTKLQRLGWFNQDEVEMVARELGVTSDKVREMESRMMAQDMTDFLSSDDDSDSQPMAPVLYLQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDENKNSTLQELADRYGVSAERVRQLEKNAMKKLRAAIEA
Patterns in Protein Structures
Different patterns of gene expression of oral epithelial IHGK cells upon co-culture with A. actinomycetemcomitans or P. gingivalis.
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.
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/RITCHIS0/301notes1.htm
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 HA₁ (red) and HA₂ (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stem-like region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75Å long. The globular head is formed by residues only from HA₁. (Courtesy of Don Wiley, Harvard University.)
## The Polymeric Players

<table>
<thead>
<tr>
<th>DNA</th>
<th>String with alphabet {A, C, G, T}</th>
<th>Nucleotides/Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA</td>
<td>String with alphabet {A, C, G, U}</td>
<td>Bases</td>
</tr>
<tr>
<td>Protein</td>
<td>String with 20-letter alphabet</td>
<td>Amino acids/Residues</td>
</tr>
</tbody>
</table>
Typical DNA Sequence

1  gggagaacac  cccgagaagg  agggagggcc  gaagaaagac  aacagaagcc  cagttgctgc
61 tccaggtccc  tccgacagag  ctttttccat  gtggagactc  tctcaatgga  cgtgcctccct
121 agtgcttctt  agacggactg  cggctctccta  aaggtcgacc  atggttggccg  ggacccgctg
181 tctttctagtg  ttgctgcttc  cccaggtcct  cctgggcggcg  gcggccggcc  tctattccaga
241 gctgggccgc  aagaagttcg  cccgcgcatc  cagccgacc  ttgtccccggc  ctctcgaaga
301 cgtcctcagc  gaaatttgagtt  cgaggctgct  cagcatgttt  ggcctgaagc  agagaccac
361 cccccagcaag  gacgctgctg  tgcctcccccta  tatgctagatg  ctgtacccgca  ggcaactcagg
421 ccagcccagg  gcgcctggccc  cagaccacggc  gctggagagg  gcagccagcc  ggcgaacacac
481 cgtgcgcagc  tttccatcagc  aagaagcct  ggaggaeactt  ccaagagatg  gttgggaaca
541 ggccccgcgc  ttccctcctg  attttaatccg  gtgcctcaggt  gcgaagtttc  aacatgacct
601 agaactcagc  atctttccggg  aacagataca  ggacagtttc  ggaacagta  gtttccagca
661 ccgaattaat  atttatgaaaa  ttataaagccc  tcacgagcag  aactgaaat  ttctgtgac
721 cagactattg  gacaccaggt  tagtgaatccg  gaacacaaagt  cagttggaga  gctttcgacg
781 caccctcagc  gctgatctggt  ggaccacaca  ggcagacagc  aacatgagtt  ttgcagtcgga
841 agtggcccat  ttagggaga  acccaggtgt  cttccagagag  catgtgagga  ttagcaggctc
901 ttgacccaa  gatgaacaca  gctgtgtcaca  gataagggcca  ttgctagttga  cttttggaca
961 tggatgaaaa  ggacatccgct  tccacaaacgc  agaaaaagcgt  caagcctaac  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="MVAGTRCLLLVLLLPQVLLGGAAGLIPELGRKFKAAAASSRPLSRPSEDVLSEFELRLLSMFGGLKQRPTPSKDVVPPYMLDLYRRHSGQPAPAPDHRLERAA
SRANTVRSHHEEAVEELEMPSEQKTARRFFFNLSVPSDEFUTSAELQIFREQIQEAL
GNSSFQHRINIEIIPAAANLKFVPTRLDDTRLVQNTSQWESFDVTAPMRWTTQG
HTNHGFVVEVAHLIEENPGVSRRHVIISRLHQDEHSWISQRPLLVTFGHDGKHPLH
REKRQAKHKQRKRLSSCKRHPLYVDFSDVGWDWIVAPPGYHAFYCHGECFPFPLADH
LNSTNHAIHVQTLVNSVNSKIPKACVPTELSAISMLYLDENEKVLKNYQDMVVEGC
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 HA₁ (red) and HA₂ (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α-helices known in a globular structure, about 75 Å long. The globular head is formed by residues only from HA₁. (Courtesy of Don Wiley, Harvard University.)
Central Dogma

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

Fig 1.6, Zvelebil/Baum
DNA Replication

Fig 1.4, Zvelebil/Baum

strand A

5’ C A T T G C C A G T 3’

3’ G T A A C G G T C 5’

strand B

5’ C A T T G C C A G T 3’

3’ G T A A C G G T C 5’

parent DNA double helix

template strand A

5’ C A T T G C C A G T 3’

3’ G T A A C G G T C 5’

template strand B

5’ C A T T G C C A G T 3’

3’ G T A A C G G T C 5’

new strand B

5’ C A T T G C C A G T 3’

3’ G T A A C G G T C 5’

new strand A

5’ C A T T G C C A G T 3’

3’ G T A A C G G T C 5’
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

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Position</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCD1P4</td>
<td>22q11</td>
<td>ATP-binding cassette, sub-family D (ALD)</td>
</tr>
<tr>
<td>SNAP29</td>
<td>22q11.21</td>
<td>synaptosomal-associated protein</td>
</tr>
</tbody>
</table>

•
•
•
DNA Molecule

[Diagram of DNA molecule with annotations for major groove, minor groove, 0.34nm (3.4Å), 3.4nm (34Å), and 2nm (20Å).]
Proteins – Amino acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>3 Letter Code</th>
<th>1 Letter Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>alanine</td>
<td>Ala</td>
<td>A</td>
</tr>
<tr>
<td>arginine</td>
<td>Arg</td>
<td>R</td>
</tr>
<tr>
<td>aspartic acid</td>
<td>Asp</td>
<td>D</td>
</tr>
<tr>
<td>asparagine</td>
<td>Asn</td>
<td>N</td>
</tr>
<tr>
<td>cysteine</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>glutamic acid</td>
<td>Glu</td>
<td>E</td>
</tr>
<tr>
<td>glutamine</td>
<td>Gln</td>
<td>Q</td>
</tr>
<tr>
<td>glycine</td>
<td>Gly</td>
<td>G</td>
</tr>
<tr>
<td>histidine</td>
<td>His</td>
<td>H</td>
</tr>
<tr>
<td>isoleucine</td>
<td>Ile</td>
<td>I</td>
</tr>
<tr>
<td>leucine</td>
<td>Leu</td>
<td>L</td>
</tr>
<tr>
<td>lysine</td>
<td>Lys</td>
<td>K</td>
</tr>
<tr>
<td>methionine</td>
<td>Met</td>
<td>M</td>
</tr>
<tr>
<td>phenylalanine</td>
<td>Phe</td>
<td>F</td>
</tr>
<tr>
<td>proline</td>
<td>Pro</td>
<td>P</td>
</tr>
<tr>
<td>serine</td>
<td>Ser</td>
<td>S</td>
</tr>
<tr>
<td>threonine</td>
<td>Thr</td>
<td>T</td>
</tr>
<tr>
<td>tryptophan</td>
<td>Trp</td>
<td>W</td>
</tr>
<tr>
<td>tyrosine</td>
<td>Tyr</td>
<td>Y</td>
</tr>
<tr>
<td>valine</td>
<td>Val</td>
<td>V</td>
</tr>
</tbody>
</table>

Table 1.1: Amino acid abbreviations
RNA
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

5' mRNA 3'

anticodon codon

Ser

tRNA

Tyr

The Genetic Code

2nd base in codon

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st base in codon</td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td>U</td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Ser</td>
<td>Tyr</td>
<td>STOP</td>
</tr>
<tr>
<td>A</td>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
</tr>
</tbody>
</table>

3rd base in codon

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>C</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>C</td>
<td>U</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>C</td>
<td>G</td>
</tr>
</tbody>
</table>
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 HA₁ (red) and HA₂ (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α-helices known in a globular structure, about 75 Å long. The globular head is formed by residues only from HA₁. (Courtesy of Don Wiley, Harvard University.)
The Central Dogma of Molecular Biology

- **Replication**: DNA duplicates
- **Transcription**: RNA synthesis
- **Translation**: Protein synthesis

Diagram showing the flow of information from DNA to RNA to protein in the cell.
Transcription

(A) **coding strand**

DNA

5' CAGTTGGCCAGT 3'

3' GTACGGGTTCAT 5'

**noncoding strand**

5' CAUUUGCCAGU 3'

3' CAGTGGCGGCA 5'

**TRANSCRIPTION**

(B) **RNA polymerase**

3' 5'

DNA rewinding

RNA polymerase

DNA double helix

direction of transcription

**newly synthesized RNA transcript**

short region of DNA/RNA helix

active site