CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan
ECS 389; Phone: x3748
giri@cis.fiu.edu
www.cis.fiu.edu/~giri/teach/BioinfS07.html
# 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><em>H. influenzae</em></td>
<td>1.8 Mb</td>
<td>1995</td>
<td>1,740</td>
</tr>
<tr>
<td><em>M. genitalium</em></td>
<td>0.58 Mb</td>
<td>1998</td>
<td>525</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>4.7 Mb</td>
<td>1997</td>
<td>4,000</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>12.1 Mb</td>
<td>1996</td>
<td>6,034</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>97 Mb</td>
<td>1998</td>
<td>19,099</td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>100 Mb</td>
<td>2000</td>
<td>25,000</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>180 Mb</td>
<td>2000</td>
<td>13,061</td>
</tr>
<tr>
<td><em>M. musculus</em></td>
<td>3 Gb</td>
<td>2002</td>
<td>~30,000</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>3 Gb</td>
<td>2001</td>
<td>32,000+</td>
</tr>
</tbody>
</table>
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 the 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.
- Completed in April 2003

http://www.ibiblio.org/wm/paint/auth/vinci/sketch/vitruvian.jpg
The Suffix Tree Data Structure

- *Borrelia burgdorferi*
  - 1 million bases
  - Shotgun Sequencing:
    - 4612 fragments
    - 2 million bases long totally
    - Using suffix trees - 15 min for Fragment Assembly
    - Using Dynamic Programming - 10 days
Sequence Alignment – Why?

>gi|12643549|sp|O18381|PAX6_DROME Paired box protein Pax-6 (Eyeless protein)
MRNLPCGLTAGGGSLGGGKAGKPSPTMEAVEASTASHRHSTSSYFATTYYHHTTDDECHGVSQQLGGGFVGG
RPLPDSTRQIVELAHSDGARPCDISRLQVNSGCVSKILGRYETGSIRPRAIGGKPRVATAEVSISIS
QYKRCEPSIFAEWEIRDLLQENVCTNDIPSVSSINRVLRLNAAKKEQQSTGSGSSSTSAGNSISAKVSV
SIGGNVSNSASRGTLSSSTDLMQTATPLNSESAGNASGSEGESEQAIEKLRLLLNTQHAAPGPLEEP
ARAAPLSVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSWPPHYSGSWYPTSLSEEIPISSAPNIAVTAY
ASGPSLAHSSLPPNDIESLASIHGQRNCPVATERDIHLLKELDHQSQDTEGSGEGENSNGGASNIGTEDD
QARILKRLQQRNRTSFTNDQIDSLEKEFTHERTHEYPDVFARERLAKIGKLPEARIQVWSNRRAKWRREEK
LRNQRTPNGSTGASSTASSTASATLTDPSNSLACSSTSSGSGAGGSPVSTINGLSSPSTLSTNVNAPTL
GAGIDSSESPTPIPHIRPSCTSNDNDNQRQSEDCRVCSCPCLVGGQGHQNTHISNGHAQGHALVPAISP
RLNFNSGSGAMASNMHTALSMDSYGAVTIPSFSNHSAVGPLAPPSSIPQQDLTPSSLYPCHMTRLNP
PPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGCSCSGSRYEVLSEAYLPPPPMASSAADSFFSAASSAS
ANVTPHHTIAQESCPSCSSASHFNGVAHASGFSDDPISPAPVSSYAHMSNYASSANTMTPSASGTSAHV
APGKQFFASCFSYFSPWV

>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
MQNSHSGVNGQQLGGGFVNGRPPLPDSLDRQIVELAHSGARPCDISRLQVNSGCVSKILGRYETGSIRPRA
IGGKPRVATPEVVSKIAYQKYRCEPSIFAEWEIRDLLQENVCTNDIPSVSSINRVLRLASEKQMQGAD
GMYDRLMLNQNTGSWGTPRGWYPGTSVPQPTGDCQECQEGGENTNSISSNGEDSDEAQMRMLQLKRKL
QRNRTSFTQEIEALEKEFTHERTHEYPDVFARERLAKIDLPEARIQVWSNRRAKWRREEKLRNQRRQQASP
TPSHIISSSSTSSTYQPIPQTTVPSSFSSTSGMGLRDTTALNMTYSLAPPPMSFMNMLPMQPVPSQ
TSSYSCLMLTPSVPNGRSYDYYTPHMQTHMNSQPMTGSGTTSTGLISPGVQVQVPGSEPDMQSYWPR
LQ
### Drosophila Eyeless vs. Human Aniridia

<table>
<thead>
<tr>
<th>Query:</th>
<th>57</th>
<th>HSGVNQLGGVFVGGRPPLPDSTROKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG</th>
<th>116</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct:</td>
<td>5</td>
<td>HSGVNQLGGVFVGRPLPDSTROKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG</td>
<td>64</td>
</tr>
<tr>
<td>Query:</td>
<td>117</td>
<td>SIRPRAIGGSKPRVATAEVSISKIQYKRECPSIFAEWIRDRLLQENVCTNDNIPSVSSIN</td>
<td>176</td>
</tr>
<tr>
<td>Sbjct:</td>
<td>65</td>
<td>SIRPRAIGGSKPRVATEVSVKI+QYKRECPSIFAEWIRDRLEVCTNDNIPSVSSIN</td>
<td>124</td>
</tr>
<tr>
<td>Query:</td>
<td>177</td>
<td>RVLRNLAAQKEQ</td>
<td>188</td>
</tr>
<tr>
<td>Sbjct:</td>
<td>125</td>
<td>RVLRNLASEKQQ</td>
<td>136</td>
</tr>
<tr>
<td>Query:</td>
<td>417</td>
<td>TEDDQARLILKRLKQRNRTSFTNDQIDSLEKEFERTHPDVFARERLAGKIGLPEARIQV</td>
<td>476</td>
</tr>
<tr>
<td>Sbjct:</td>
<td>197</td>
<td>SDEAQMLQLKRLQTNRSFTQIEAKEYFERTHPDVFARERLAKIDLPEARIQV</td>
<td>256</td>
</tr>
<tr>
<td>Query:</td>
<td>477</td>
<td>WFSNRRACKWRREEKLRNQRR</td>
<td>496</td>
</tr>
<tr>
<td>Sbjct:</td>
<td>257</td>
<td>WFSNRRACKWRREEKLRNQRR</td>
<td>276</td>
</tr>
</tbody>
</table>

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

MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLLEAA
KTLILSHLRFVVHIARNYAGYGLPQADLIQEGRINIGLMKAVRRFNPVEVGVR
LVSFAVHWIKAEIHEYVLRLRNWRIVKVATTKAQRKLFFNLKRTKQRLGWFN
QDEVMVARELGVTSDKVREMESRMAAQDMDTFDLSDDDSDSQPMPAVLY
LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDITRARWLDEDNKS
STLQELADRYGVAERVQLEKNAAMKKLRAAIEA

MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLLEAA
KTLILSHLRFVVHIARNYAGYGLPQADLIQEGRINIGLMKAVRRFNPVEVGVR
LVSFAVHWIKAEIHEYVLRLRNWRIVKVATTKAQRKLFFNLKRTKQRLGWFN
QDEVMVARELGVTSDKVREMESRMAAQDMDTFDLSDDDSDSQPMPAVLY
LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDITRARWLDEDNKS
STLQELADRYGVAERVQLEKNAAMKKLRAAIEA
Patterns in Protein Structures
Different patterns of gene expression of oral epithelial IHGK cells upon co-culture with A. actinomyces and P. gingivalis.

Microarray Analysis
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)
Molecular Biology Background
2 star molecular players
## The Players

<table>
<thead>
<tr>
<th>DNA</th>
<th>String with alphabet {A, C, G, T}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nucleotides/Bases</strong></td>
<td></td>
</tr>
</tbody>
</table>

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

1 gggagaacac ccggagaagg agggagggc gaagaaacgc aacagaagcc cagttgctgc
61 tccaggtccc tcggacagag cttttttccat gtggagactc tctcataagga cgtgcctccct
121 agtgtctttt agacggaactg cgggtctctca aaggtcgcacc atggtgcggcg ggacccgctcg
181 tctttctagtg ttgctgtcttc cccaggtcct ccggtggtggc gcggccggcc tcattccaga
241 gcttgggccgc aagaagttcg ccggtgcactc ccagggaccc ttgtccccggc ttctcggaga
301 cgctccctagc gaatttgagt tgagggtgct cagcatgttt ggctgaaggc agagaccac
361 cccagcaag gacgtcgttg gcgcccccta tatgctagat ctgtaccgca ggcactcagg
421 ccagccagga gcgcggcggcc cagaccaccg gcttgagagg gcagccagcc gcgcacacac
481 gctgcgcagc ttctcatcagc aagaagccgt ggaggaactt cccaggtcct cccttgtgaga
541 gcgcgggcccggc cttccttcttc ataataagttc tgcgcttcgat gcagatggttc cccatctgac
601 agaactccag atcttcgggg aacagataca ggaagctttg ggaaacagta ttctcacgca
661 cccgaattaat atttatgaaa ttataaagtc tccgacgcac aacttaagcc ttctctgtgac
721 cagactattg gacaccaggt tagtgaatca gaacahaagc gcttcgacgt cccagctgtg
781 caccctcgac ccggaggtcgtt gcggtggaga cgcagttgtt ggtgtgacacac ccttctgtgac
841 agtggcccat ttagaggaga acccaggtcgtttg tggaggttc ggttcaagc acacaccacac
901 tttggcccaa gatgaacaca gcttgacata gataaggcca tttgctagtga cttttggaca
961 tggtggaaaag ggcacagcccg tccacaaacc agaaaaagcgt caagccaaac acaaacagcg
Typical protein sequence

/translation="MVAGTRCLLLPVLLLPQVLLGGGALPGLPRLGRKFAAASSRPLSSRPSEDVLSEFELRLLSMFGGLKQRPTPSKDVVPPPYMLDLYYYRHSQPGAPAPDHRLERAA
SRANTVRSFHHEEAEVEELPMSGKTARRFFFNLSVPSDEFLTSAELQIFREQIQEAL
GNSSFQHRINIFYEIIKPAANLNKFPVTRLLDTRLVNLQNTSWSFVTPAVMRWTTQG
HTNHGFVVEAHLEENPGVSKRHVRISRLHQDEHSWSQIRPLLVTFGGDGKHPLHK
REKRQAKHKQRRKSLSSCKRHPLYVDSDVGVNWIVAPPYGHAHYCHGECFPFPLADH
LNSTNHAIVQTLVNSVNSKIPKACVPTELSAISMLYLDENEKVVLKNYQDMVVEGCGR"
Central Dogma

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

Human 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
## 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>histine</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
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

![Diagram of tRNA and mRNA with codons and anticodons]

<table>
<thead>
<tr>
<th>2nd base in codon</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<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>Pro</td>
<td>His</td>
<td>Arg</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

The Genetic Code
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 HA₂. (Courtesy of Don Wiley, Harvard University.)
The Central Dogma of Molecular Biology

Replication
DNA duplicates

Transcription
RNA synthesis

Translation
Protein synthesis

DNA

RNA

nucleus

cytoplasm

nuclear envelope

Protein

Ribosome

Protein
Transcription Regulation

- Exon 1
- Exon 2
- TATA Box
- CAT Box
- CCAAT
- TATAAT
- Gene-Specific TF Binding Sites
- Basal TF Binding Sites
- Transcription Start point
- Coding region
- Promoter region
- Enhancer
- Gene
DNA Transcription

RNA synthesis and processing
Transcription Initiation
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.
Transcription Steps

RNA polymerase needs many transcription factors (TFIIA, TFIIB, etc.)
(A) The promoter sequence (TATA box) is located 25 nucleotides away from transcription initiation site.
(B) The TATA box is recognized and bound by transcription factor TFIID, which then enables the adjacent binding of TFIIB. DNA is somewhat distorted in the process.
(D) The rest of the general transcription factors as well as the RNA polymerase itself assemble at the promoter. What order?
(E) TFIIH then uses ATP to phosphorylate RNA polymerase II, changing its conformation so that the polymerase is released from the complex and is able to start transcribing. As shown, the site of phosphorylation is a long polypeptide tail that extends from the polymerase molecule.
The general transcription factors have been highly conserved in evolution; some of those from human cells can be replaced in biochemical experiments by the corresponding factors from simple yeasts.
Protein Synthesis

1. Transcription

2. Translation

Protein synthesis
Protein Synthesis:

Incorporation of amino acid into protein