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/BioinfS08.html
Overview of Courses

- 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** *(HMMPro, 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

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semester Project</td>
<td>50</td>
</tr>
<tr>
<td>Homework Assignments</td>
<td>20</td>
</tr>
<tr>
<td>Exams</td>
<td>25</td>
</tr>
<tr>
<td>Class Participation</td>
<td>5</td>
</tr>
</tbody>
</table>

## Course Homepage

**www.cis.fiu.edu/~giri/teach/BioinfS08.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

DNA Sequence → Gene → Protein Structure → Function

- Gene Regulatory Networks
- PPI Networks
- Metabolic Pathways
- Molecular Interaction and Reaction Networks
General Information


- Human Genome has ~3 billion bp with 32,000+ genes.

- 435/624 complete microbial genomes sequenced (684/914 more in progress)

- 2540 Viral genomes (300bp - 300Kb) (1st 1978: Simian virus; 5Kb).

- 22 complete eukaryotic genomes sequenced (175 more in progress):
  - Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa, Plasmodium falciparum, Drosophila melanogaster

- 131 organisms have assemblies and chromosomal maps including:
  - Anopheles gambiae, Macaca mulatta, Bos taurus, Felis catus, Gallus gallus

- Swiss-Prot Release 51.3/54.7 (Dec'06/Jan'08): 250K/333K entries; 91/120 million amino acids.
<table>
<thead>
<tr>
<th>Organism</th>
<th>Size</th>
<th>Date</th>
<th>Est. # genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>HIV type 1</em></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 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)
MRNLPCGTVAGEGSGGLGIKPS2TPM8A40ESTASHHINTFATETYHILTDDECHGV1Q5LLGFGVG
RPLDSTRQKIVELAHSGARPCDISRLQVSNGCVSKILGRYETGSIRPRAIGGSKPRVATAEVDKSIS
YKRECPISIFAWEIRDRLQENVCTNDIPSVSINRVLRLNAAQKEemonsTGSGSSSTSAGNSISAKSVV
S1GNGVSNASRSRGTLSSSTDLMQRTATPLNSSESGAGSNSEGSGSIEAIYEKLRLLNTQHAAPGPPLP
ARAAPLVQQSPNHLGTRSSHPQLVHGNHQAQLQHHQQSSWPPRHYSGSWFYTPSLSLEIPISSAPNIASTVAY
ASGSPSLAHHLSSPPNDIESLASIGHQNRCPVATEDIHLKELDGHQSDETGSEGSGNNGASNGTTEDD
QARLILKRLQVRNRTSFTNDQISLEKEFERTHYPDVIFARERLAGKIGLPEARIPQVWSNRRRKRREEL
LRNQRTTPTNSTGATSSSTTNTASLTDSPNLSSACSSLSSGAGGSVSTINGLSSPSTLSTTVNAPTL
GAGISSESPTFIPHRPSCSTDNDNQRQSEDRCRRVCSCPLGVSQEHQNTHIQNSNGHAQGHALVPAISP
RLNFNSGSFGAMYNSMNHTALSMDSDYGTSNPSPSNHASVPLAPPSPIPQGDLTPSSLYPCHMTLRRP
PPMAPAHHRIVPDDGRPAGVGLSGQSANNLAGSACSGSVEYLSAYALPNNAMSSADASSFSASASSAS
ANVTIPHTIAQECSFCPSSASHFGVAHSSSGFSSDPIPAVSSYAHMSYASSANTMTPSSSATGSAHV
APGQKQFFSASCFSYPWV

>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
M QNSHSGVENQLGQVFVNGRPPLDSTRQKIVELAHSGRRCISCILQVSNGCVSKILGRYETGSIRPRA
IGGSKPRVATPEVSVKIAQKRECPISIFAWEIRDLLQENVCTNDIPSVSINRVLRLASEQKQMGAD
GMYDCLRMLNQQTGSWGRPGYWPQGTSGPQPTDGCQOPQEGGENTISISSNGEDSDAQMRVLRLKRL
QRNRTSFTQEOIEALEKEFERTHYPDVIFARERLAAKDLPEARIQIVWNSNRRRKRREELLRNQRRQASNN
TPSHPISISSFSSTSVQIPQPPTPVSSFTSGMLMRDSTALTNTYAALPPMPFMTANNLPMQPVPSQ
TSSYSCMLPTSPVNGRSDTYTPPMQTHMSQPMGTSGTTTSTGLISPGVSPVQVPQSEPDMGQYWPRLQ
Drosophila Eyeless vs. Human Aniridia

<table>
<thead>
<tr>
<th>Query:</th>
<th>Sbjct:</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPDCISRILQVSNGCVSKILGRYYETG 116</td>
</tr>
<tr>
<td>117</td>
<td>SIRPRAIGGSKPRVATAEVVSKISQYKRECPISIFAEIRDRLLLQENVCTNIDNIPSVSSIN 176</td>
</tr>
<tr>
<td>177</td>
<td>RVLRNLA+K+Q 188</td>
</tr>
<tr>
<td>417</td>
<td>TEDDQARLILKRKLRQRNRTSQTNDQIDSLLEKERFHTYPDFVARERLAKGIGLPERAIQV 476</td>
</tr>
<tr>
<td>477</td>
<td>WFSNRRAKWRREEKLRNQRR 496</td>
</tr>
</tbody>
</table>

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

- MTDKMQLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLLEAAKTLILSHLRVFVHIARNYAGYGLPQADLIQEGNIGLMAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIKVATTKAQRKLFNLRKTKQRLGWQFNDQDEVEVARELGVTSKDVREMERSMAAQQMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAAANRLTDAMQGLDERSQDIIRARWLEDDNKSSTLQELADRYGVSAERVQOLEKNAMKPLRAAIEA

- MTDKMQLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLLEAAKTLILSHLRVFVHIARNYAGYGLPQADLIQEGNIGLMAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIKVATTKAQRKLFNLRKTKQRLGWQFNDQDEVEVARELGVTSKDVREMERSMAAQQMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAAANRLTDAMQGLDERSQDIIRARWLEDDNKSSTLQELADRYGVSAERVQOLEKNAMKPLRAAIEA
Patterns in Protein Structures
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)