BSC 4934: Q’BIC Capstone Workshop

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

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

Course Homepage

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

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.
- 1099 complete bacterial (86 archaeal) genomes sequenced
- 3041 Viral genomes (300bp - 300Kb) (1st 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

<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

- 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
- 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)
MRNLPCCLGTAGGSGLGGIAGKPSPTMEAVEASTASHRTSTSYFATTYYHLTDDECCHSGVQLGGVFGG
RPPLPDRSQKIVALAHSGARPCDSRILQVSNCGVSKILGNYTETGSRPRAIGGSKPRVATAEVS Kis
QYKRECPSIFAWEIRDLLRQENVCTNIDPINPSVSINRVRNLNAACKEQQQSTSGSSSTSAGNSISAKV SV
SIGGNVSNVASGRGTLSSTDLQMTATPLNSESGGASNSGESTEQEAIYELKLRLLNTQHAAAGPGPLEP
ARAAPLTVGQSPNHLGTRSSHQPQLVHGHNQALQQHQQQSWPPRHYSGWSYPTSLSEIPISSAPNIA SVTAY
ASGSPAHLHLSPPNDIESLASHIGQRNCVPATIDHLKKELEDQGSDETGSGEQENSGGASPNI NIGN TEDD
QARLILRKLQRNRSTFNDQIDSLEKEFERTHPDVFARERLAGKIGLPEARIQVWSNRRAKWRREEK
LNNQRRTPPNSATNSSSTSHATLSATLSASCCSLSSGSAGGPSVSTINGLSSPSSTLSTVNAPTL
GAGIDSSESPTIPHPSPCSTSDNDNGRQSEDCCRVCSCPCLGVGHHQNTTHHIQSGNHAQGHALVPAISP
RLNFNSGSGGAMYSNMHHTALSMSDSYAVTPIPSNSHSAVGPLAPPSPPIQGGLTPSLYPCHMTRLR
PPMAHAPHHVPDGGRPGAVGGLGQQSANLASCSCSGYEVLSAYALPPPMFASSSAADDSSFSAA ASSAS
ANVTPHHTIAQESCPSRSSASHFVGAHSSGFSSDPISPAVSSYAHMSYASSANTMTFPSASSGTSAHV
APGKQQFFASCFSFPW

>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) ([Aniridia](#), type II protein)
MQNSHVSNQLGGGFGVGVRPLDPSTRQKIVALAHSGARPCDSRILQVSNCGVSKILGNYTETGSRPRA
IGGSKPRVATPVEVSKIAQYKRECPSIFAWEIRDLLRQENVCTNIDPINPSVSINRVRNLNAACKEQQQG AD
GMYDKLRMLNGQTGSGWTRPGWPHTVSVPQPTDQGCOQQCGGEGENTNSINSSNGEDSDAEQMRILQ LKRKL
QRNRTSFTQEIEALEKEFERTHPDVFARERLAGKIGLPEARIQVWSNRRAKWRREEKLNNQRRQAS N
TPSHIPISSSSFSSTSVYQIPQPTTPVSSFTSGLGRTDTALTNTYSALPPMPSFTMANNLPMQFPVPSQ
TSSYSCMLPTSVPNRSYDPTYTPPHMQTHMNQPMGTSSTTTSTGGLSPIVSGVVFQVPGSEPDMQSYWPR
LQ
Drosophila Eyeless vs. Human Aniridia

Query: 57  HSGVNQLGGVFGGVGRPLPDRQKIVELAHSGARCISRLQVSNGCVSKILGRYYETG  116
HSGVNQLGGVFGVRPLPDRQKIVELAHSGARCISRLQVSNGCVSKILGRYYETG
Sbjct:  5  HSGVNQLGGVFGVGRPLPDRQKIVELAHSGARCISRLQVSNGCVSKILGRYYETG  64

Query: 117  SIRPRAIGSGKPRVATAEVVSQSKISQYKRECPSIFAEWIDRULQENVCTNDNIPSVSSIN  176
SIRPRAIGSGKPRVATEVVSQSKISQYKRECPSIFAEWIDRULQENVCTNDNIPSVSSIN
Sbjct:  65  SIRPRAIGSGKPRVATPEVVSQSKIAQYKRECPSIFAEWIDRULSEGVCTNDNIPSVSSIN  124

Query: 177  RVLRNIAAQKEQ  188
RVLRNLA++K+Q
Sbjct:  125  RVLRNIAASEKQQ  136

Query: 417  TEDDQARLILKRKLRQRNRTSTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQV  476
+++ Q RL LKRKLRQRNRTST +QI++LEKEFERTHYPDVFARERLAGKIGLPEARIQV
Sbjct:  197  SDEAQMRLOKLKRKLRQRTSDTFQEQIEALEKEFERTHYPDVFARERLAGKIGLPEARIQV  256

Query: 477  WFSNRRAKWRREEKLRNQRR  496
WFSNRRAKWRREEKLRNQRR
Sbjct:  257  WFSNRRAKWRREEKLRNQRR  276

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

MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
KTLILSHLRVFVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPENVGVR
LVSFAVHWIKAEIHEYVLRNWRIVKVATTKARQKLFNFLRKQRLGWFN
QDEVEMVARELGVTSKDVREMESRMAAQDMDTFLSSDDDSDSQPMPAVLY
LQDKSSNFADGIEEDNWEEQAANRLTDAMQGLDERSQDIIRARWLDENK
STLQELADRYGVSAERVROLEKNAMKKLRAAIEA

MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
KTLILSHLRVFVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPENVGVR
LVSFAVHWIKAEIHEYVLRNWRIVKVATTKARQKLFNFLRKQRLGWFN
QDEVEMVARELGVTSKDVREMESRMAAQDMDTFLSSDDDSDSQPMPAVLY
LQDKSSNFADGIEEDNWEEQAANRLTDAMQGLDERSQDIIRARWLDENK
STLQELADRYGVSAERVROLEKNAMKKLRAAIEA
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/RITCHISO/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 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 Polymeric Players

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

```
1  gggagaacac  ccggagaagg  agggaggagg  gaagaaagac  aacagaagcc  cagttgctgc
61  tccaggtccc  tcggacagag  ctttttccat  gtggagactc  tctcaatgga  cgtgcctcccct
121  agtgctttct  agacggacctg  cggcttcctta  aaggtcgacc  atgggctggc  ggacccgctg
181  tctttctagtg  ttgctgcttc  cccaggtcct  cctggccggc  gggccggccc  tcattccaga
241  gctggggccgc  aagaagttcg  ccccgcaatc  cagccgaccc  ttgctccggc  cttcggaaaga
301  cgctcccagc  gaattttagt  tgggtgtgct  cagcatgttt  ggctgaagc  agacccacac
361  ccccgagcaag  gacgtcggtgg  tgcccccccta  tatgctagat  ctgtaccgca  ggcactcagg
421  ccagccagga  gcgcgggccc  cagaccacccg  gctggagagg  gcagccagcc  gcgccaacac
481  cgtgctgcagc  ttccactacgc  aagaagccgt  gggacagctt  ccagagatga  gttggaacac
541  ggcgcggcgc  ttctctctgc  atctaaagttc  tgggcttgcag  gacgagtttc  tcacatgctgc
601  agaactccag  atctcccggg  aacagataca  ggaagctttg  ggaaacagta  gtttccagca
661  cccgaattaat  atttatgaaa  ttataaagcc  tgccagagcc  aacccaggtt  tctccttgagc
721  cagactattg  gacaccaggt  tagtgaatca  gaacacaagt  cagttggga  gctttcgacgt
781  caccccacgt  ggtatgcggt  ggaccacaca  gggacacacc  aacatggtt  ttgtggtgga
841  agttgcccat  tttagaggaga  acccaggtgt  ctccaaagaga  catgtgagga  ttagcaggtc
901  ttgctcacaa  gatgaacaca  gctggctcac  gataaagccca  ttgcttgaat  ctttcggaca
961  ttagtgaaaaa  ggacatccgc  tccacaaccg  agaaagcgtc  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="MVAGTRCLLVLLLPQVLLGGAGLIPELGRKKFAAASSRPLSRP SEDVLSFELRLLSMFGLKQRPTPSKDVVPPYMLDLRYRHSGQPAPAPDHRLERAA SRANTVRSHHEEAVEELEPSMGKTARRFFFNLSSVPSDEFLTSQAELOIQREQQIEAQ EA GNSSFQHRINYEIIKPAANLKFVPTRLDDTRLVQNTSQWESFVTPAVMRWTQPQ HTNHGFVVEVAHLEENPGVSKRHRIRSLQDEHSQSIRPLLVTGHDGKHHPLHK REKFAQKHKQRKRLKSSCRKHPLYVDFSVDGWNAVPPGWRYHAFYCHGECFPFPLADH LNSTNHAIVQTLVNSVNSKIPKACVPTELSAISMLYLDNEKVLKNYQDMVVECGC 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

DNA replication

RNA synthesis (transcription)

protein synthesis (translation)

PROTEIN

amino acids

Fig 1.6, Zvelebil/Baum
DNA Replication

Fig 1.4, Zvelebil/Baum

DNA Replication diagram showing the replication process with strands A and B, template strands A and B, and new strands A and B.
Chromosomes

Human chromosomes!

[Diagram showing human chromosomes with labeled centromere and chromatid]
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.
<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
DNA

Complementary Bases

C
G
A
T
Proteins – Amino acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Three-letter Code</th>
<th>One-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
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
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 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

1. DNA replication: DNA duplicates
2. Transcription: RNA synthesis
3. Translation: Protein synthesis
Transcription

Fig 1.7, Zvelebil/Baum