CAP 5510/CGS 5166: Bioinformatics & Bioinformatic Tools

GIRI NARASIMHAN, SCIS, FIU
Course Preliminaries

  - Lecture Slides; Reading Material; Announcements; Homework
  - VISIT OFTEN!
- Class meets 5:00 – 6:15 PM, ECS 138, MW
- Office ECS 254B; Office Hours: By Appointment Only
- Phone: x-3748; Email: giri@cis.fiu.edu
- Final Exam: Monday, 12/3/2018, 5:00 – 7:00 PM, ECS 138
- Extra 1 credit for CGS 5166 students, if needed
Syllabus

- Fundamentals of Biology, Statistics, & Bioinformatics
- Databases; Data Integration; BioPerl & BioPython;
- Sequence Alignment, Multiple Sequence Alignment Sequencing; Next Generation Sequencing & Applications
- Pattern Discovery, Learning, Prediction & Inference; Machine Learning: NN, HMM, SOM, SVM, etc.
- Gene Regulation; Regulatory Elements; & networks Transcriptomics: Analysis of Gene Expression Data; Genomics, Proteomics, Transcriptomics; other Omics
- Gene Ontology and Pathways; Protein-protein interactions Comparative Genomics
- Phylogenetic Analysis
- Structural Bioinformatics: RNA and Proteins
- Genetics and Genome-Wide Association Schemes Single Nucleotide Polymorphisms Misc.: Omics; Alternative Splicing; Epigenetics;
- Cancer Bioinformatics; Microbiomes and Metagenomics;
- Software Engineering; Visualization;
Evaluation

- Semester Project (45 %)
- Homework Assignments (20 %)
- Exam (15 %)
- Quizzes (10 %)
- Summary Reports of Interest (5 %)
- Class Participation (5 %)

http://www.cs.fiu.edu/~giri/teach/BioinfF18.html
Some History ...

- What major world event took place on **26 June, 2000**?

- Other dates in Bioinformatics history:
  - 1758 – work of **Carl Linnaeus** – taxonomy
  - mid 1800s – **Gregor Mendel** – genetics
  - mid 1800s – **Charles Darwin** – evolution
  - 1953 – Watson, Crick, Franklin **Structure of DNA**

- More important dates
  - 1975 – Sanger Sequencing
  - 1977 – first bacteriophage sequenced
  - 1978 – Dayhoff’s Atlas of Protein Sequence and Structure
  - 1980s – EMBL, GenBank, SWISSProt, and DDBJ
  - 1990 – HGP initiated
Algorithms & Hardware

- Moore’s Law: Faster processors, larger and faster memory, larger external memories
- Optimization: “Linear Programming is tractable”
- Convex Programming: Interior Point Methods
- Energy Minimization: Soft Computing Methods (Simulated Annealing, Neural Networks, ...)
- Parallel/Grid/Cloud Computing: CHARMM ported to parallel environments
- GPU Computing: NVIDIA video cards do more than just graphics, and can be programmed (in C/C++) to deliver on high performance scientific computing
- Quantum Computing: Showed that some problems can be solved more efficiently on a quantum computer
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)
   - Data Analysis (Data Mining, Modeling, Statistics)
   - Development of Efficient Algorithms
   - Visualization and Interface Design (HCI, Graphics)

1. How to assist biological research?
   - Build databases for data
   - Build efficient tools for search, retrieval, analysis, & visualization
   - Propose models and efficient tools to verify the model using known data
   - use predicted information to narrow down search
   - propose new experiments based on model or analysis
   - Build smart, hyperlinked, integrated mining environments
Overall Goals

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

- **Molecular**: DNA, RNA, proteins, ligands, toxins, . . .
- **Cellular**: chromosome, nucleus, cell wall, chromatin, organelles, organization of a single cell
- **Tissue & Organ**: Collection of cells: gene expression
- **Organism or Systems Biology**: Genome, variations within organism, or over physiological or pathological states, epigenome
- **Community**: Metagenome, Microbiome, Ecology, ...
- **All life**: Tree of life, phylogeny, variations, comparative studies
Phenomenal Growth of Information

- **Human Genome** has 3 billion bp with 32,000+ genes.
- 435/624/3880/30,000 complete microbial genomes sequenced of which 4500 are virus genomes.
Phenomenal Growth of Information...2

Microbial Genome Growth (1995-2012)

Number of entries in UniProtKB/Swiss-Prot


Number of Genomes:
- 0 to 500
- 500 to 1000
- 1000 to 1500
- 1500 to 2000

Year:
- 1985
- 1988
- 1991
- 1994
- 1997
- 2000
- 2003
- 2006
- 2009
- 2012
- 2015
- 2018
- 2021
1800 Complete genomes

- Caenorhabditis elegans
- Arabidopsis thaliana
- Saccharomyces cerevisiae
- Mus musculus
- Homo sapiens
- Oryza sativa

- Plasmodium falciparum
- Drosophila melanogaster
- Anopheles gambiae
- Macaca mulatta
- Bos taurus
- Felis catus
- Gallus gallus
And Genome Sizes ...

<table>
<thead>
<tr>
<th>Organism</th>
<th>Size</th>
<th>Date</th>
<th>No. of Genes (est.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV Type I</td>
<td>9.2 Kb</td>
<td>1997</td>
<td>9</td>
</tr>
<tr>
<td>M. genitalium</td>
<td>0.58 Mb</td>
<td>1998</td>
<td>525</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>1.8 Mb</td>
<td>1995</td>
<td>1,740</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>
Caenorhabditis Elegans

www.ucl.ac.uk

universe-review.ca
C. elegans: The Model worm

- 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 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 Alignments: Why we need them?

>gi|12643549|sp|O18381|PAX6_DROME Paired box protein Pax-6 (Eyeless protein)
MRNLPCGLTGAGSGLGIGAGKPSPTMEAVEASTARHSTSFATTYHTTDCHGGNQNGGGVFGG
RPLPDSRKRKIVELAHSARGCDISRLQVSNCGVSKILGRYETGSIRPRAIGGSKRVTATAEVVSQ
YKRECSIFAEWIRDQLQENVCTNDNIPSVSNIRLRNLLAQKEQQSTGSRSSTSAGNSIASKVS
YGNSVNVASGSRTGSSTDLQQTATPLNSESQGASGRGEQSEAIEYLKLLLQHAAXPGLP
AREAPLVQASPNHLGTRSSHQVLVGHNHQAQQHQQQSWPPHRHYSGWSWPTSEIPISSAPNIAVTAY
ASGPALASHLSPPNDESLASIGHRQRCPTAVEIDHKLKEDGHQSDGETSGEGGNSNGGASNIGNTE
DDQARLILKRKLRNRTSFNTQIDSLEKEFERTHPDVFAERLAKTGLPEARIQVWSNRRRAKWRREE
KLRNQRTPNSGTASATSSATSATASLTDSPNLSSAASSLLSGSAGPSVSSTINGLSPSTLSTNVNAPTL
GAGDSSESPTIPHIPECTSNDNGQSESDDRCCPCLVGHHQNTTHIQRNSHGAQHALVPAISP
RLNFNSGSGFAMYSNMHHTALSMSDSYGVTPQIPSFHSAVGPLAPPSPQPPQDLTTPSLLPYCHMTLRP
PPMAPAHHIVPGDGRGAPAGVLGSGQAANLGRACSGGGYEVLSAYALPPPPMASSAADSFSASSAS
ANVTPHHTIAQESCPSCSSASHFVAGHSSFGFSSDIPSPAVSSYAHMSYNTASSANTMPSSASGTSAHV
APGQQFASCFYSPWV

>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
MQNSHSGNVQLGGVSVGVRPLDSTRKRKIVELAHSARGCISRLQVSNCGVSKILGRYETGSIRPRA
IGGSKRVTATPVSVKIAQYKRECSIFAEWIRDQLQENVCTNDNIPSVSNIRLRNLAASKQOMGAD
GYMDKLRLMLNGQTSGWGRPGVPGVTSPGQPTQDGQCQGEGGENTNISSSNGDSDEAQMRQLK
QRNRTSFTEQIEAALKEFERTHPDVFAERLAKTDLPEARIQVWSNRRRAKWRREEKLRNQRTQASN
TPSHIPISSFSSTSVQIPQPPTPVSSFTSSGMLSGRTDALTNTYSAALPPMSFTMANLPMOPPVPSQ
TSSYSCMLPTSPVSNGRSHTYTPPHMQTHMNSQMGMGSTTTSTGLISPVGVPQVPGSEPMDQWPR

Girl Narasimhan
**Drosophila Eyeless vs. Human Aniridia**

<table>
<thead>
<tr>
<th>Query</th>
<th>57</th>
<th>HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRLQVSNGCVSKILGRYYETG</th>
<th>116</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>5</td>
<td>HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRLQVSNGCVSKILGRYYETG</td>
<td>64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>117</th>
<th>SIRPRAIGGSKPRVATAEVS KISQYKRECPSIFAWEIRDRLLQENVCTNDNISVSSIN</th>
<th>176</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>65</td>
<td>SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E VCTNDNISVSSIN</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>177</th>
<th>RVLRLNIAAQEQ</th>
<th>188</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>125</td>
<td>RVLRLNASEKQQ</td>
<td>136</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>417</th>
<th>TEDDQARLILKRLQRNRTSFTNDQIDSLEKEFERTHPDVFARERLAGKIGLPEARIQV</th>
<th>476</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>197</td>
<td>SDEAQMRQLKRLQRNRTSFTQIEALEKEFERTHPDVFARERLAAKIDLPEARIQV</td>
<td>256</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>477</th>
<th>WFSNRRAKWRREEKLRNQRR</th>
<th>496</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>257</td>
<td>WFSNRRAKWRREEKLRNQRR</td>
<td>276</td>
</tr>
</tbody>
</table>

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

[MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDFEAA
KTLILSHLRFVVKHAIRNYAGYGLPQADLIIQEGNIGLMKAVRRFNPEVGVRLVSAVHWIKAIHEYVLRNWRIVKVATTKAQRKLFNLRTKKQLRQLGWFQDEVEMVARELGVTSDKVREMERSMAAQDMTFDLSDDDSDSQPMAPVLYLQDKSSNFADGIIEDDNWEEQAANRTLTDAMQGQDERSQDIIRARWLDDEDNKSTLQELADRYGVSASERVQQLENKAMKLRAAIEA]

[MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDFEAA
KTLILSHLRFVVKHAIRNYAGYGLPQADLIIQEGNIGLMKAVRRFNPEVGVRLVSAVHWIKAIHEYVLRNWRIVKVATTKAQRKLFNLRTKKQLRQLGWFQDEVEMVARELGVTSDKVREMERSMAAQDMTFDLSDDDSDSQPMAPVLYLQDKSSNFADGIIEDDNWEEQAANRTLTDAMQGQDERSQDIIRARWLDDEDNKSTLQELADRYGVSASERVQQLENKAMKLRAAIEA]

[G. Narasimhan, et al., “Mining Protein Sequences for Motifs,”
Patterns in Protein Structures
T. Milledge et al., “Sequence Structure Patterns: Discovery and Applications”, CBG 2005
Microarray Analysis

Handfield et al., Distinct Expression Profiles Characterize Oral Epithelium-Microbiota Interaction”, Cellular Microbiology, 2005
Comparative Genomics

Comparative Genomics

Microbiomes: The Ultimate “Social Network”

Fernandez, Riveros, Campos, Mathee, Narasimhan
Distinguishing SLE from MCTD
State of the Art
Distinguishing SLE from MCTD with ML

Distinguishing SLE from MCTD with ML


Table 4  Evaluating novel proposed classification rule for unclear SLE and MCTD patients

<table>
<thead>
<tr>
<th>Classification criteria sets</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>New proposed “Lu-vs-M” rule</td>
<td>96.30%</td>
<td>61.54%</td>
<td>50.00%</td>
</tr>
<tr>
<td>SLE</td>
<td>SLICC</td>
<td>62.96%</td>
<td>93.75%</td>
</tr>
<tr>
<td></td>
<td>ACR</td>
<td>55.55%</td>
<td>75.00%</td>
</tr>
<tr>
<td>MCTD</td>
<td>Alarcón-Segovia</td>
<td>22.22%</td>
<td>18.75%</td>
</tr>
<tr>
<td></td>
<td>Sharp</td>
<td>48.14%</td>
<td>62.50%</td>
</tr>
<tr>
<td></td>
<td>Kasukawa</td>
<td>50.00%</td>
<td>80.00%</td>
</tr>
<tr>
<td></td>
<td>Kahn</td>
<td>29.63%</td>
<td>6.25%</td>
</tr>
</tbody>
</table>

Analyses were performed in SPSS (version 18) and included unclear SLE ($n = 16$) and MCTD ($n = 11$) patients from the validation group. Lu-vs-M: SLE vs MCTD; SLICC: Systemic Lupus International Collaborating Clinics; ACR: American College of Rheumatology; SLE: systemic lupus erythematosus; MCTD: mixed connective tissue disease.
The SIDS Mystery

1. 18000 Amish people in Pennsylvania
2. Mostly intermarried due to religious doctrine
3. Rare recessive diseases occurred with high frequencies.
4. SIDS: 3000 deaths/year (US); 21 deaths (Amish community)
5. Many research centers failed to identify cause
6. Collaboration between Affymetrix, TGEN & Clinic for special children solved the problem in 2 months
7. Studied 10000 SNPs using microarray technology
8. 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.
9. Conclusion: Disease caused by 2 abnormal copies of TSPYL gene
10. Identified genes expressed in key organs (brainstem, testes)
11. [Link to Affymetrix website]
The Alzheimer’s Mystery

- Search for the “Alzheimer’s Laboratory”, an episode of 60 minutes that was aired by CBS in Nov 2016 and then again in Jan 2018.
- This is now in Homework 1. More later …