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/BioinfS06.html

Evaluation

- Semester Project (50 %)
- Homework Assignments (20 %)
- Exams (25 %)
- Class Participation (5 %)

Course Homepage

www.cis.fiu.edu/~giri/teach/BioinfS05.html

- Lecture notes, required reading material, homework, announcements, etc.

Course Schedules

- CAP 5510 (3 credits) and CGS 5166 (2 credits) will meet every Tue from 2 PM to 4:45 PM. The first half of the lecture will be meant for both classes. The second half will be more focused on the CS majors
- Different exams and evaluation.
- Please attend all classes regardless of registered course.

Genome Sizes

Organism	Size	Date	Est. # genes
HIV type 1	9.2 Kb	1997	9
H. influenzae	1.8 Mb	1995	1,740
M. genitalium	0.58 Mb	1998	525
E. coli	4.7 Mb	1997	4,000
S. cerevisiae	12.1 Mb	1996	6,034
C. elegans	97 Mb	1998	19,099
A. thaliana	100 Mb	2000	25,000
D. melanogaster	180 Mb	2000	13,061
M. musculus	3 Gb	2002	~30,000
H. sapiens	3 Gb	2001	32,000+

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

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)
- ~ 50 transposon-derived genes
- 223 genes common with bacteria that are missing from worm, fly or yeast.

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)
MRNLPCLGTAGGSGLGGIAGKPSPTMEAVEASTASHRHSTSSYFATTYYHLTDDECHSGVNQLGGVFVGG
RPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKIS
QYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTSAGNSISAKVSV
SIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEP
ARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSWPPHHYSGSWYPTSLSEIPISSAPNIASVTAY
ASGPSLAHSLSPPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDD
QARLILKRELQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEK
LRNQRRTPNSTGASATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPTL
GAGIDSSESPTPIPHIRPSCTSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHAQGHALVPAISP
RLNFNSGSPGAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSPIPQQGDLTPSSLYPCHMTLRP
PPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGASCSGSGYEVLSAYALPPPPMASSSAADSSFSAASSAS
ANVTPHHTIAQESCPSPCSSASHPGVAHSSGFSSDPISPAVSSYAHMSYNYASSANTMTPSSASGTSAHV
APGKQQPFASCFYSPWV
```

```
>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRA
IGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGAD
GMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKL
QRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREELRNQRRQASN
TPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQ
TSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ
```

Drosophila Eyeless vs. Human Aniridia

```
Query: 57   HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 116
            HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct: 5    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG

Query: 117  SIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSIN 176
            SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E VCTNDNIPSVSSIN
Sbjct: 65   SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN 124

Query: 177  RVLRNLAAQKEQ 188
            RVLRNLA++K+Q
Sbjct: 125  RVLRNLASEKQQ 136

Query: 417  TEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQV 476
            +++ Q RL LKRKLQRNRTSFT +QI++LEKEFERTHYPDVFARERLA KI LPEARIQV
Sbjct: 197  SDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQV 256

Query: 477  WFSNRRAKWRREEKLRNQRR 496
            WFSNRRAKWRREEKLRNQRR
Sbjct: 257  WFSNRRAKWRREEKLRNQRR 276
```

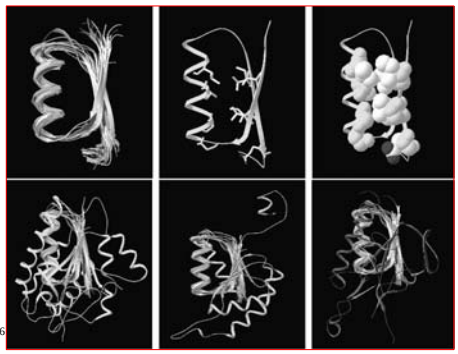
E-Value = 2e-31

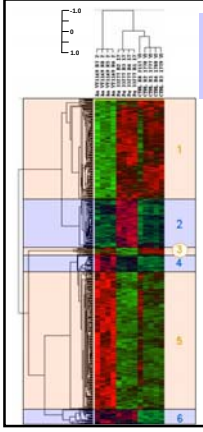
Motif Detection in Protein Sequences

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

Patterns in Protein Structures

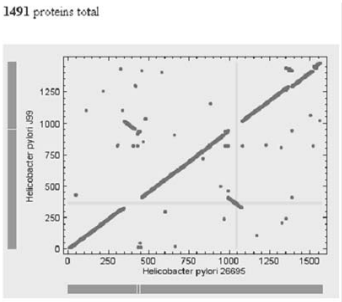




Microarray Analysis

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
- Identified genes expressed in key organs (brainstem,testes)
- Studied 10000 SNPs using microarray technology
- Conclusion: Disease caused by 2 abnormal copies of TSPYL gene