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/BioinfS11.html

Syllabus

- Fundamentals of Biology, Statistics, the Internet, and Bioinformatics
- Databases and Software Packages, BioPerl.
- Sequence Alignment, Multiple Sequence Alignment
- Sequencing; Next Generation Sequencing & Applications
- Predictive Methods: Nucleotide Sequences and Protein Sequences
- Pattern Discovery Techniques and applications
- Machine Learning: NN, HMM, SOM, SVM, etc.
- Gene Regulation; Predicting Regulatory Elements
- Analysis of Gene Expression Data
- Gene Ontology and Pathways; Protein-protein interactions
- Genomics, Proteomics, Comparative Genomics
- Phylogenetic Analysis
- Molecular Structural Analysis: RNA and Proteins
- Genetics and Genome-Wide Association Schemes
- Single Nucleotide Polymorphisms
- Advanced Topics: RNAi, Alternative Splicing, Epigenetics

Software Packages

- ❑ Databases: GenBank, SwissProt, BioPerl.
- ❑ Sequence Alignment: BLAST, CLUSTAL
- ❑ Sequencing Assembly: VELVET
- ❑ Pattern Discovery: PROSITE, Pfam, GYM, TEIRESIAS,
- ❑ Machine Learning Tools: HMMPro, GeneCluster, SVMLite.
- ❑ Useful Databases: RegulonDB, GO, KEGG
- ❑ Analysis of Gene Expression Data: MAS, GeneSpring
- ❑ Genomics, Proteomics, Comparative Genomics: GreenGenes
- ❑ Phylogenetic Analysis: PHYLIP, PAUP
- ❑ Molecular Structure Analysis DALI, RASMOL, SPDBV
- ❑ Statistical Software Packages SAS, R

Evaluation

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

Course Homepage

http://www.cis.fiu.edu/~giri/teach/BioinfS11.html

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

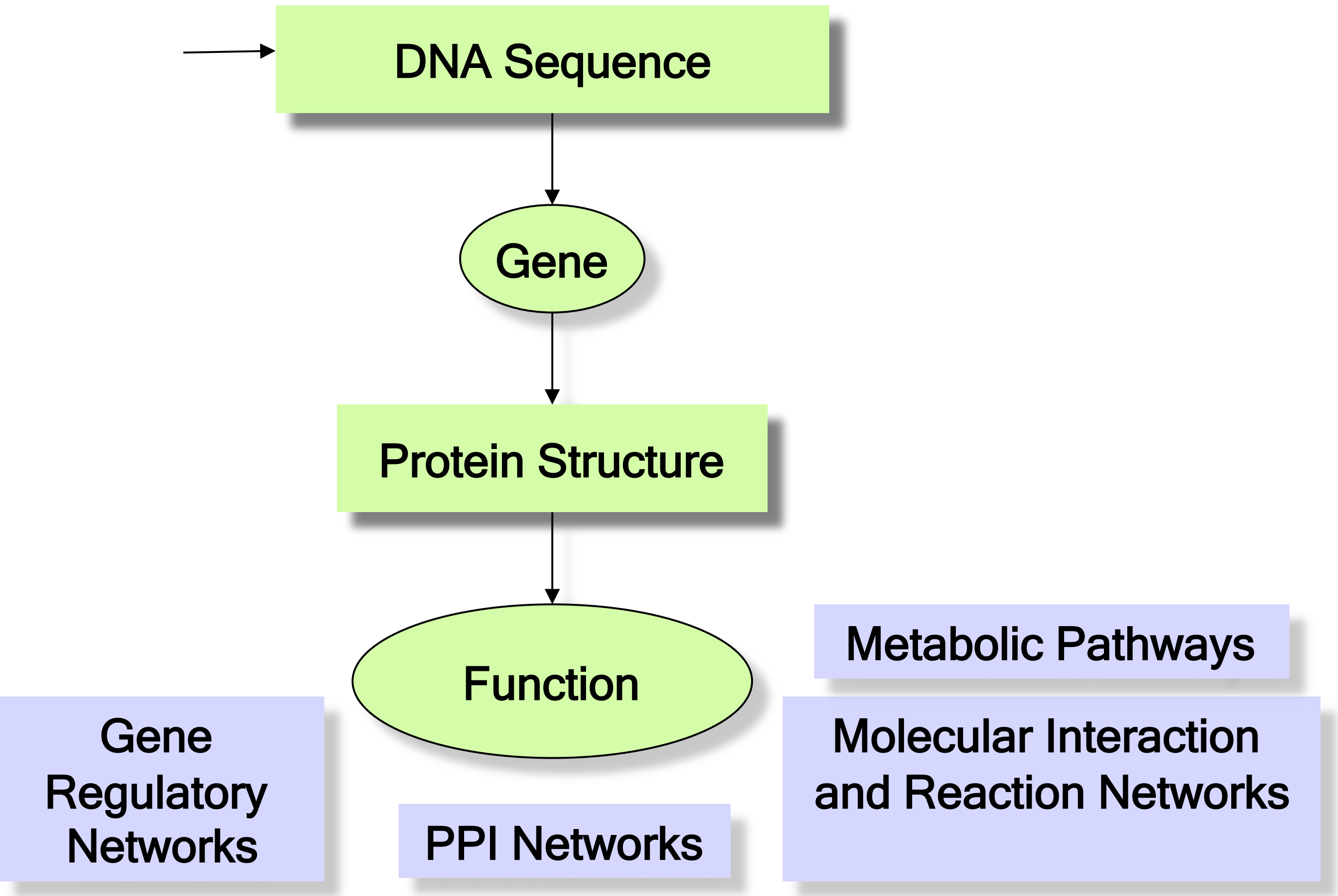
History

- ❑What major world event took place on 26 June, 2000?
- ❑What major discovery was made in 1953?
- ❑1975: Sanger Sequencing
- ❑1977: first bacteriophage sequenced
- ❑1990: HGP initiated

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



Perspective of Bioinformatics

- Study of the cell: DNA, genes, proteins
- Study of the organism: genome, changes over time, over body regions, or over physiological or pathological states
- Study of all life: Tree of Life, Phylogeny, Variations, comparative genomics

General Information

- GenBank Release 157/163 (Dec 2006/7) contains over 64/80 million sequence entries totaling over 83 Gb from over 2,500 organisms [http://www.ncbi.nlm.nih.gov] (Storage: ~150 GB uncompressed)
- 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.

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+

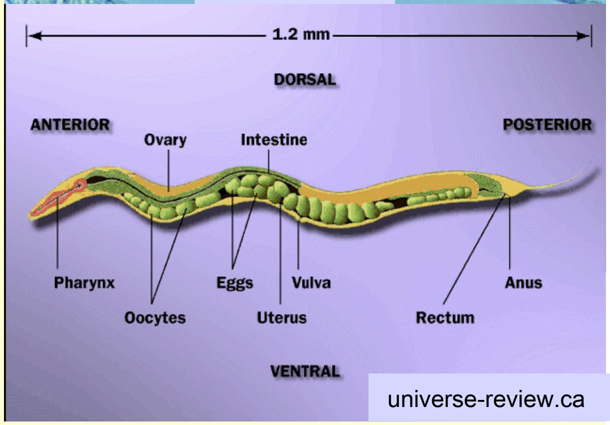
Short Homework

- Find the organism with the largest genome known! How many chromosomes does it have?
- Find the organism with the shortest genome known! How long is its genome?
- Something to think about: Do you think a larger genome implies a “more evolved” organism or a “less evolved” organism?

Caenorhabditis Elegans

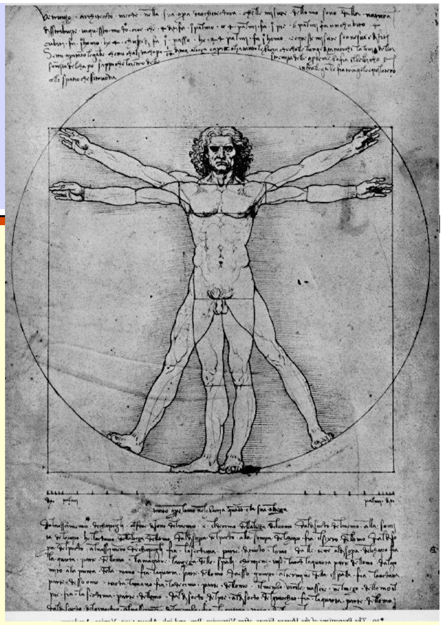
- ❑ Entire genome – 1998; 8 year effort
- ❑ 1st animal; 2nd eukaryote (after yeast)
- ❑ Nematode (phylum)
- ❑ Easy to experiment with; Easily observable
- ❑ 97 million bases; 20,000 genes;
- ❑ 12,000 with known function; 6 Chromosomes;
- ❑ GC content 36%
- ❑ 959 cells; 302-cell nervous system
- ❑ 36% 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

www.ucl.ac.uk



universe-review.ca

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)
MRNLPCLGTAGGSGLGGIAGKPSPTMEAVEASTASHRHSTSSYFATTYYHLTDDECHSGVNQLGGVFVGG
RPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKIS
QYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTSAGNSISAKVSV
SIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEP
ARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSWPPRHYSGSWYPTSLSEIPISSAPNIASVTAY
ASGPSLAHSLSPPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDD
QARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEK
LRNQRRTPNSTGASATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPTL
GAGIDSSESPTPIPHIRPSCTSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHAQGHALVPAISP
RLNFNSGSFGAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSPIPQQGDLTPSSLYPCHMTLRP
PPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGASCSGSGYEVLSAYALPPPPMASSSAADSSFSAASSAS
ANVTPHHTIAQESCPSPCSSASHFGVAHSSGFSSDPISPAVSSYAHMSYNYASSANTMTPSSASGTSAHV
APGKQQFFASCFYSPWV
```

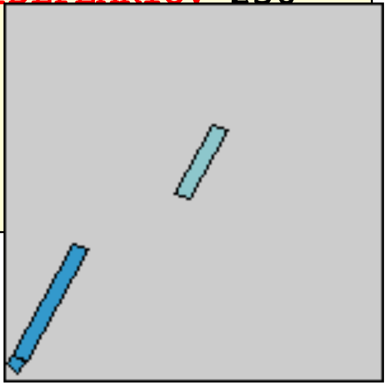
```
>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRA
IGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGAD
GMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKL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

Drosophila Eyeless vs. Human Aniridia

```
Query: 57  HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 116
           HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct: 5   HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 64

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

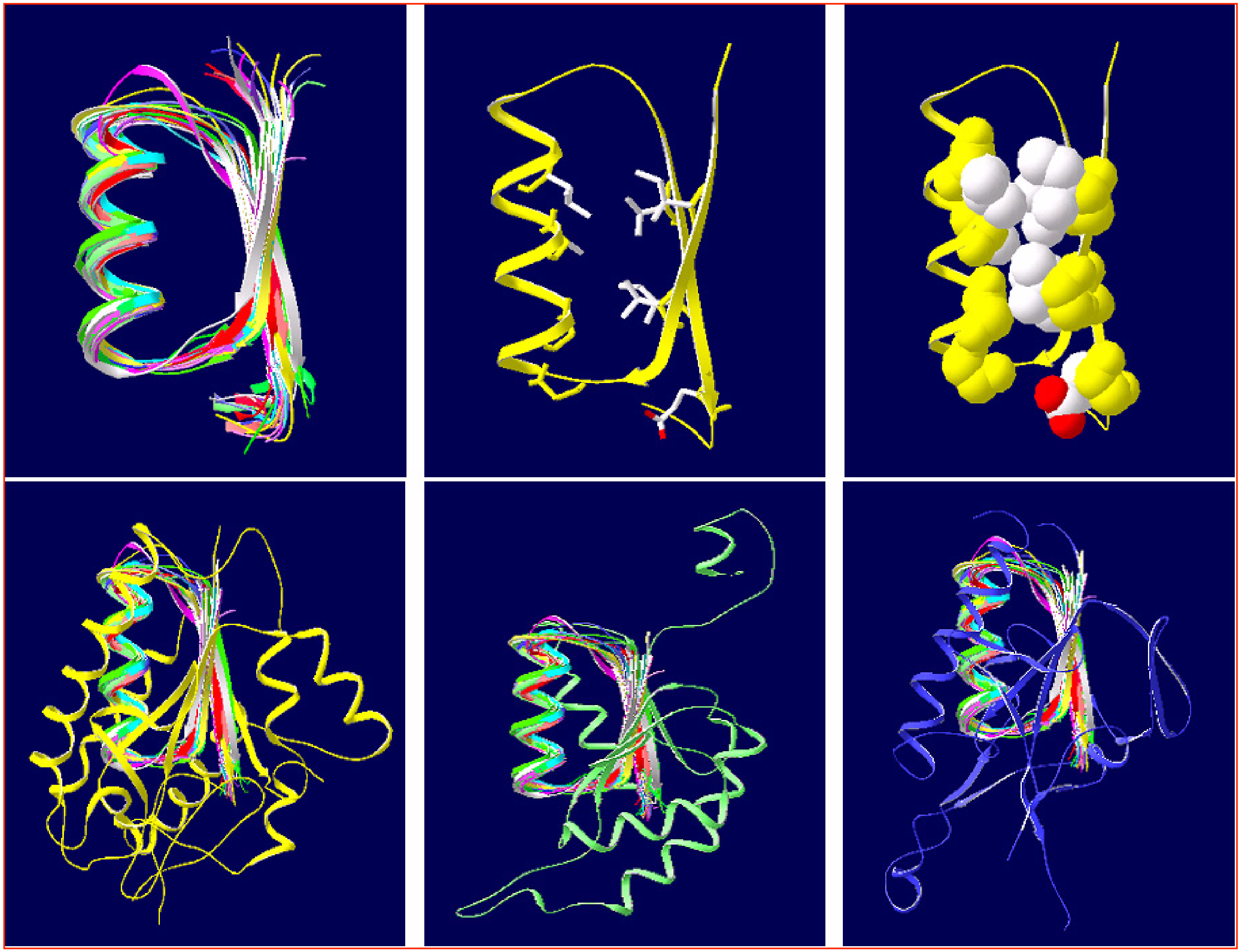


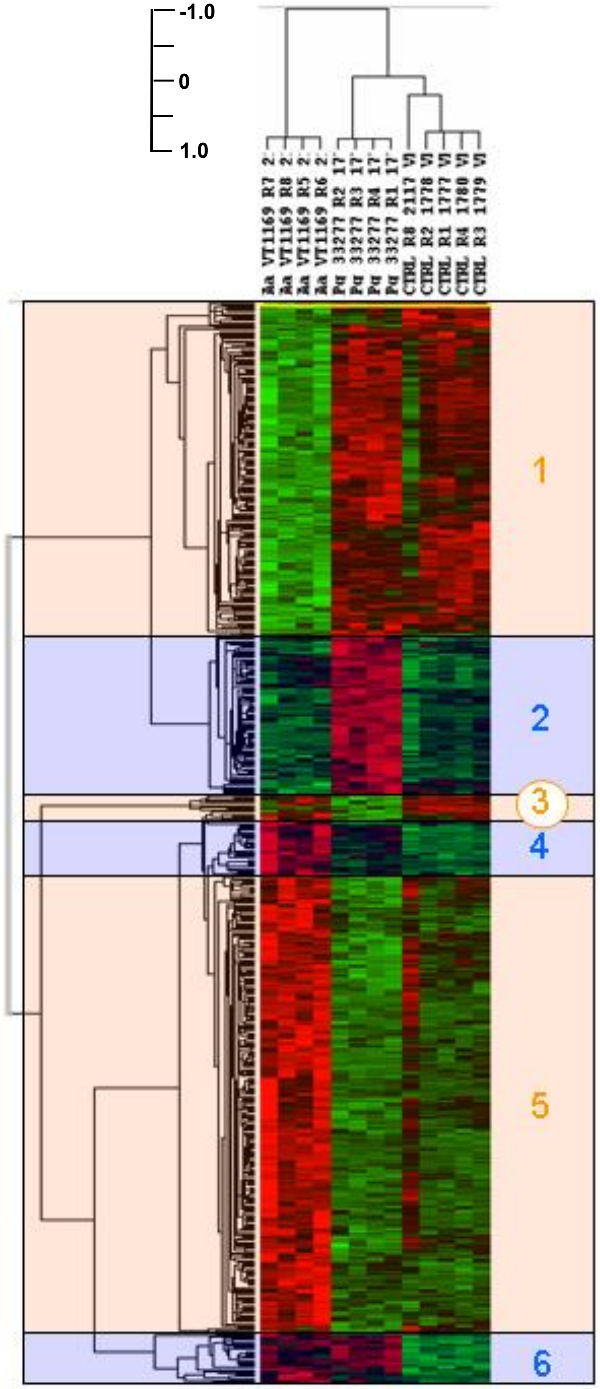
Motif Detection in Protein Sequences

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

Patterns in Protein Structures



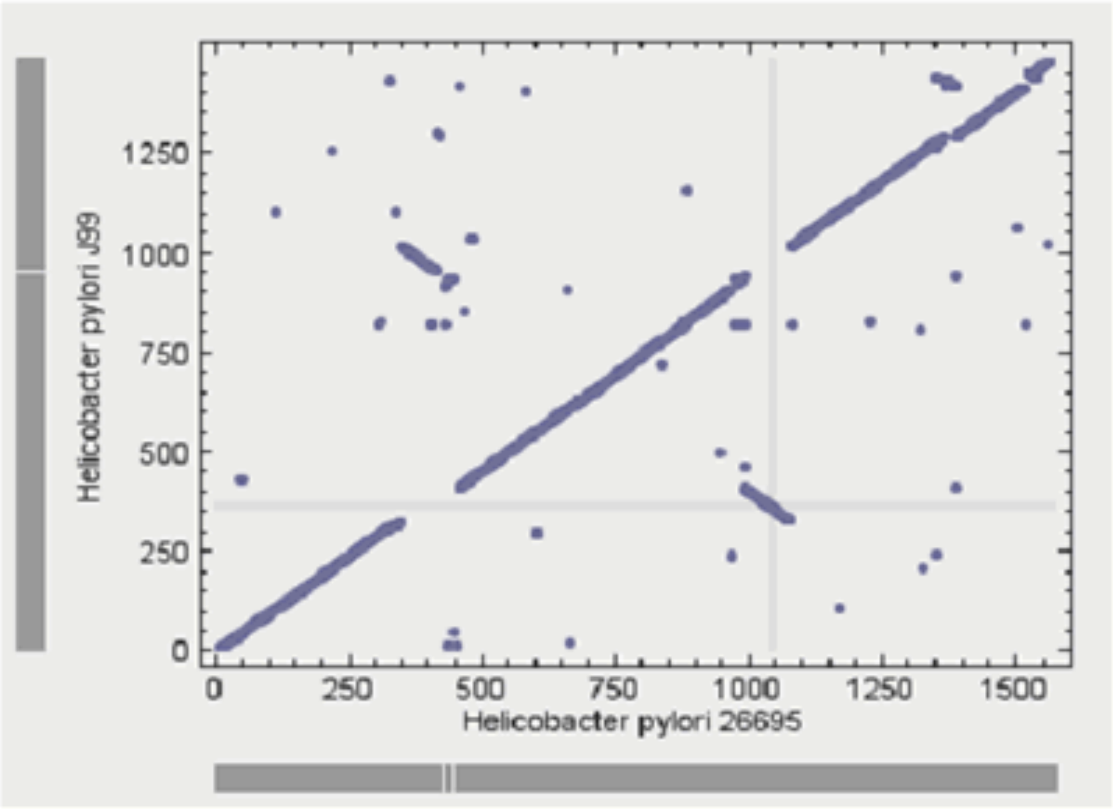


Microarray Analysis

Different patterns of gene expression of oral epithelial IHGK cells upon co-culture with A. actinomycetemcomitans or P. gingivalis.

Tools: GenePlot

1491 proteins total



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)
- http://www.affymetrix.com/community/wayahead/modern_miracle.affx