CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 389; Phone: x3748

giri@cs.fiu.edu

www.cs.fiu.edu/~giri/teach/5166S05.html

Course Schedules

- CAP 5510 (3 credit) will meet every Tue and Thu from 11 AM to 12:15 PM. For CS Majors.
- CGS 5166 (2 credit) will meet every Tue from 11AM to 12:15 PM. For non-CS majors.
- Different exams and evaluation.
- Please attend all classes regardless of registered course.

Overview of Course

- Sequence Alignment; Multiple Sequence Alignment
- Sequence Analysis
- Phylogenetic Analysis
- Gene recognition
- Pattern Discovery Techniques
- Sequencing and Mapping
- Structure Analysis
- Structure alignment
- Genomics, Functional Genomics, Proteomics
- Microarray Data Analysis
- Programming Environments
- 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, CLUSTALX)
- 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)

Evaluation

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

Course Homepage

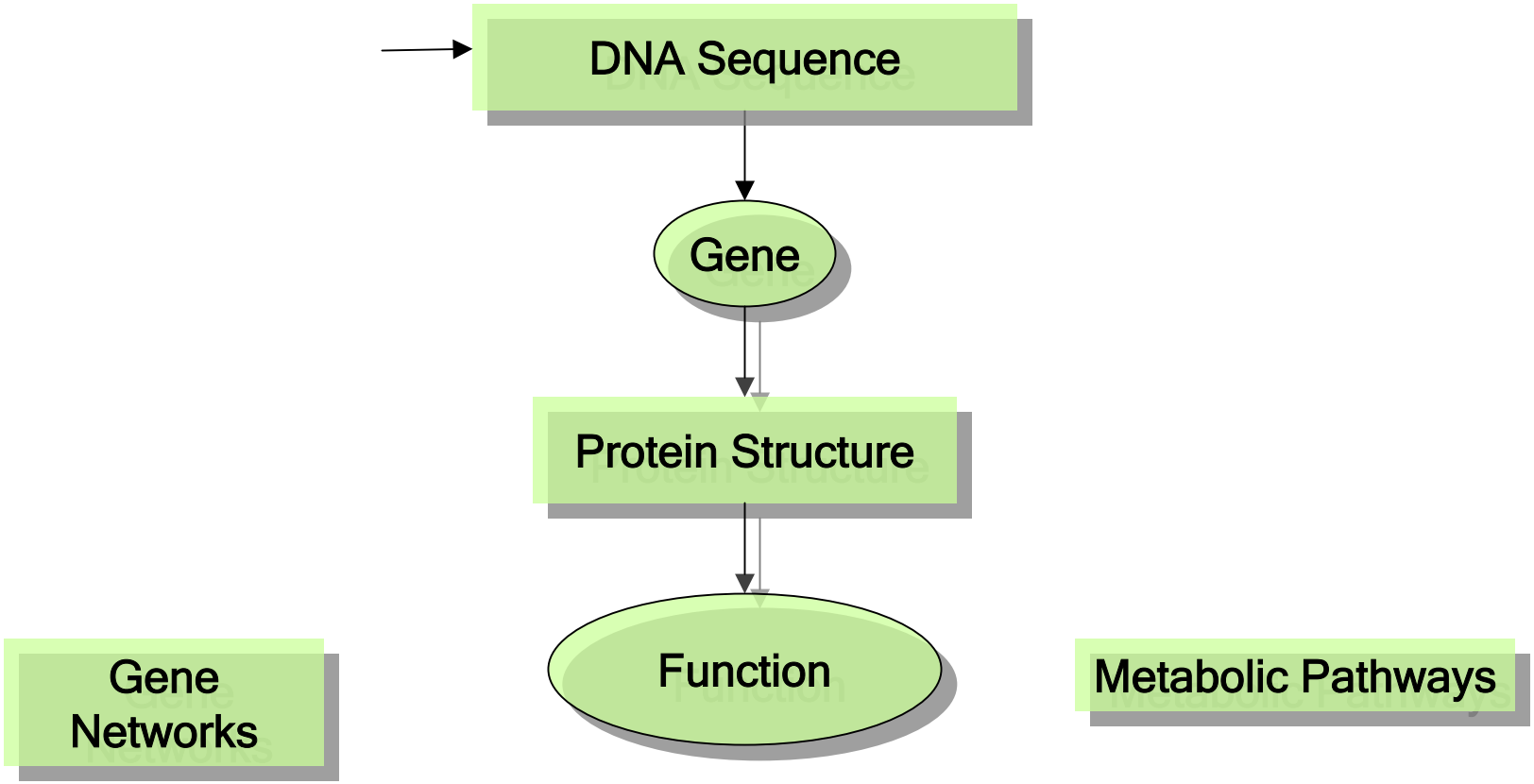
www.cs.fiu.edu/~giri/teach/5166S05.html

- 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



General Information

- GenBank Release 144 (Oct 2004) contains over 38 million sequence entries totaling over 43 Gb from 2,645 organisms [http://www.ncbi.nlm.nih.gov] (Storage: 147 GB uncompressed)
- Human Genome has ~3 billion bp with 32,000+ genes.
- 213 complete microbial genomes sequenced (274 more in progress)
- 1000 Viral genomes (300bp - 300Kb) (1st 1978: Simian virus; 5Kb).
- 14 complete eukaryotic genomes sequenced (46 more in progress):

  Caenorhabditis elegans, Arabidopsis thaliana, Drosophila melanogaster, Saccharomyces cerevisiae,
- Chromosomal maps for many organisms including:

  Mus musculus, Homo sapiens, Danio rerio, Zea mays, Oryza sativa
- Swiss-Prot Release 45.5 (Jan 2005): 167089 entries; 6 million amino acids.

Genome Sizes

Organism	Size	Date	Est. # genes
HIV type 1	10 Kb		
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+

Caenorhabditis Elegans

- Entire genome – 1998; 8 year effort;
- 97 million bases;
- 20,000 genes; 12,000 genes with known function
- 1st animal; Multicellular organism
- Nematode (phylum)
- Easy to experiment with; Easily observable
- 959 cells
- 302 nerve cells
- 36% of proteins common with human

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
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

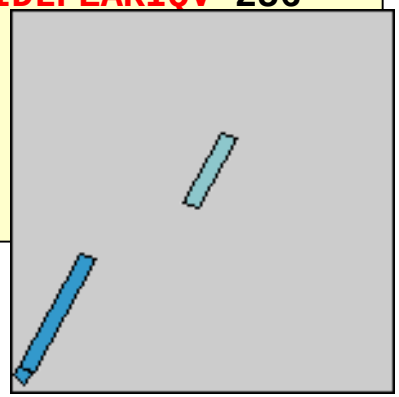
```
Query: 57   HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 116
            HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct: 5    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 64

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

Genomic Databases

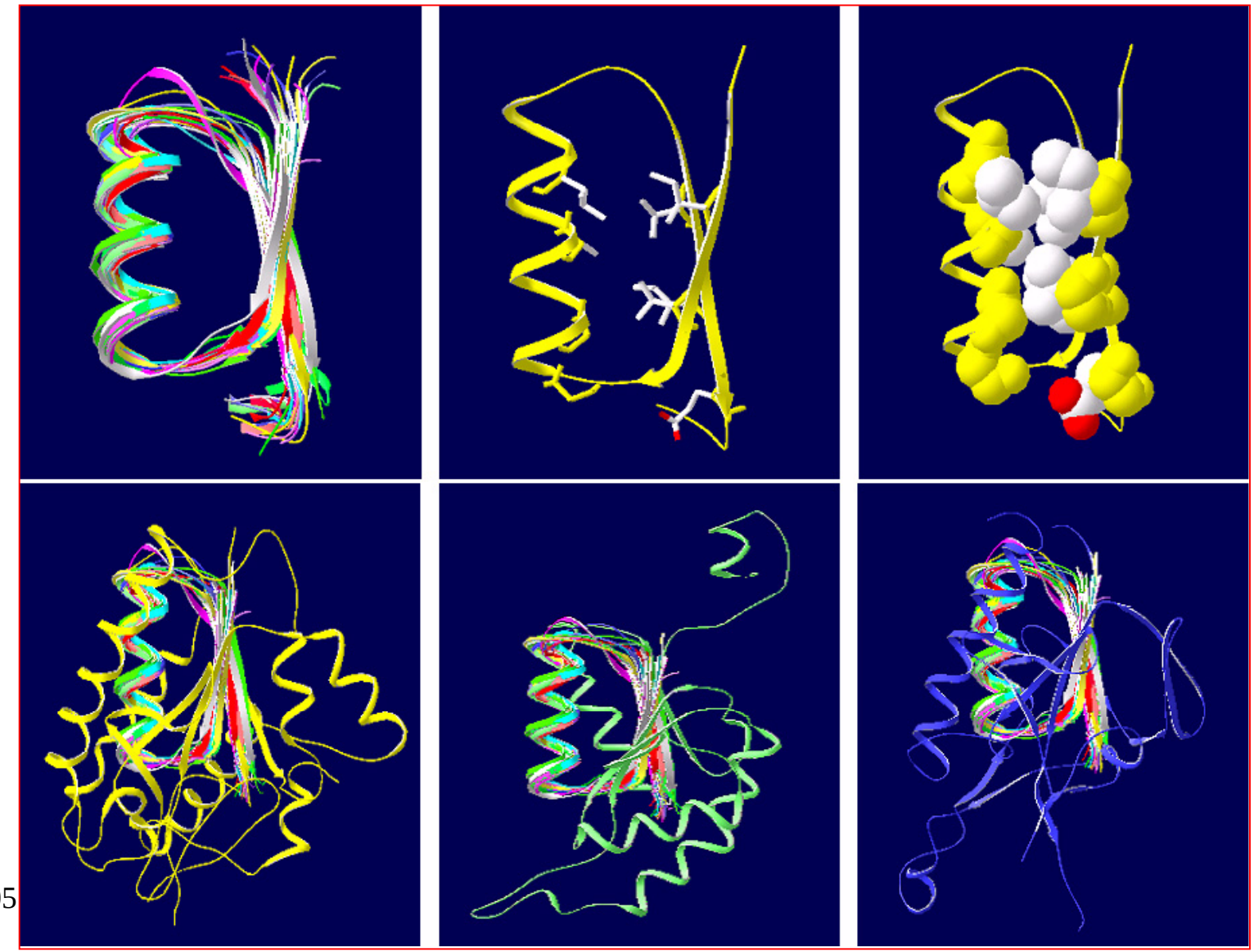
- Entrez Portal at National Center for Biotechnology Information (NCBI) gives access to:
  - Nucleotide (GenBank, EMBL, DDBJ)
  - Protein (PIR, SwissPROT, PRF, and Protein Data Bank or PDB)
  - Genome
  - Structure
  - 3D Domains
  - Conserved Domains
  - Gene; UniGene; HomoloGene; SNP
  - GEO Profiles & Datasets
  - Cancer Chromosomes
  - PubMed Central; Journals; Books
  - OMIM
  - Database Neighbors and Interlinking

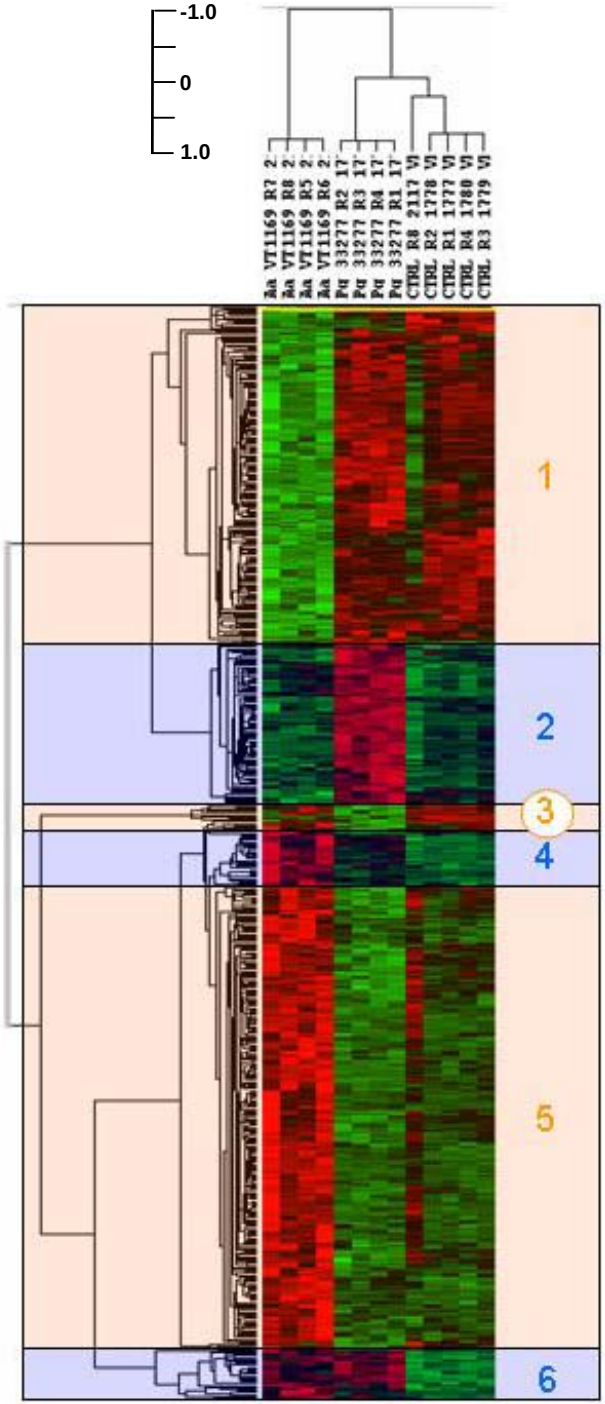
Motif Detection in Protein Sequences

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

Patterns in Protein Structures

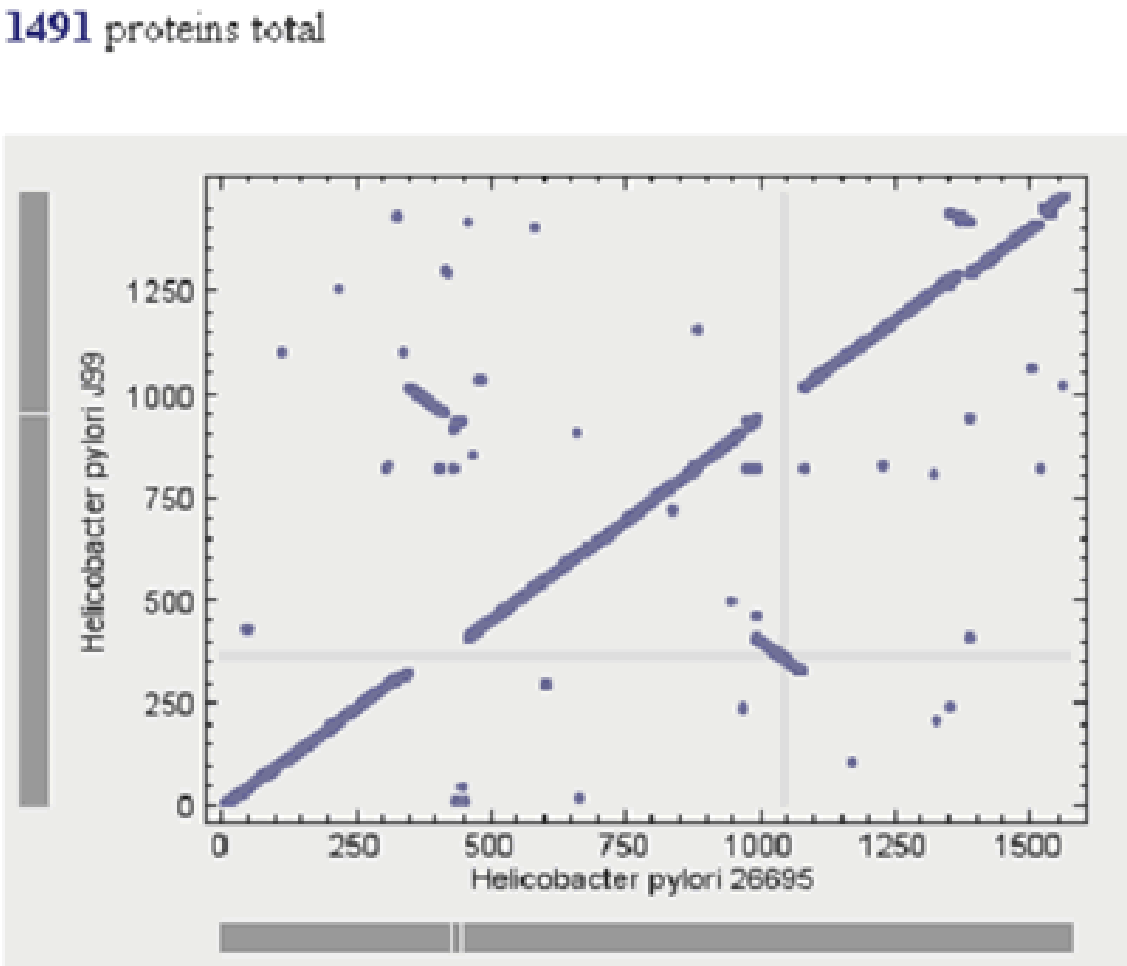




Microarray Analysis

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