CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254A / EC 2474; Phone x3748; Email: giri@cis.fiu.edu
My Homepage: http://www.cs.fiu.edu/~giri
http://www.cs.fiu.edu/~giri/teach/BioinfS15.html
Office ECS 254 (and EC 2474); Phone: x-3748
Office Hours: By Appointment Only

Jan 12, 2015

Presentation Outline

General Information

- Course Webpage: `http://www.cs.fiu.edu/~giri/teach/BioinfS15.html`
    - Lecture Slides
    - Announcements
    - Reading Material
    - Homework
    - VISIT OFTEN!
- Office Hours: By Appointment Only
- Phone: x-3748; Email: giri@cis.fiu.edu
- Office ECS 254 (and EC 2474);
- Extra 1 credit for CGS 5166 students

Core Syllabus

- Fundamentals of Biology, Statistics, & Bioinformatics
- Databases; Data Integration; BioPerl & BioPython;
- Sequence Alignment, Multiple Sequence Alignment
- Sequencing; Next Generation Sequencing & Applications
- Discovery, Learning, Prediction & Inference
- Machine Learning: NN, HMM, SOM, SVM, etc.
- Gene Regulation; Regulatory Elements; & networks
- Transcriptomics: 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
- Misc.: Omics; Alt. Splicing; Epigenetics; Visualization;

Evaluation

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

Presentation Outline

Some History . . .

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

Some History . . .

- What major world event took place on 26 June, 2000?
- Other important dates in Bioinformatics history:
    - 1758 – work of Carl Linnaeus –

Some History . . .

- What major world event took place on 26 June, 2000?
- Other important dates in Bioinformatics history:
  - 1758 – work of Carl Linnaeus – taxonomy
  - mid 1800s – work of Gregor Mendel –

Some History . . .

- What major world event took place on 26 June, 2000?
- Other important dates in Bioinformatics history:
    - 1758 – work of Carl Linnaeus – taxonomy
    - mid 1800s – work of Gregor Mendel – genetics
    - mid 1800s – work of Charles Darwin –

Some History . . .

- What major world event took place on 26 June, 2000?
- Other important dates in Bioinformatics history:
  - 1758 – work of Carl Linnaeus – taxonomy
  - mid 1800s – work of Gregor Mendel – genetics
  - mid 1800s – work of Charles Darwin – evolution
  - 1953 – a major discovery:

Some History . . .

- What major world event took place on 26 June, 2000?
- Other important dates in Bioinformatics history:
    - 1758 – work of Carl Linnaeus – taxonomy
    - mid 1800s – work of Gregor Mendel – genetics
    - mid 1800s – work of Charles Darwin – evolution
    - 1953 – a major discovery: Structure of DNA
    - 1975 –

Some History . . .

- What major world event took place on 26 June, 2000?
- Other important dates in Bioinformatics history:
    - 1758 – work of Carl Linnaeus – taxonomy
    - mid 1800s – work of Gregor Mendel – genetics
    - mid 1800s – work of Charles Darwin – evolution
    - 1953 – a major discovery: Structure of DNA
    - 1975 – Sanger Sequencing
    - 1977 –

Some History . . .

- What major world event took place on 26 June, 2000?
- Other important dates in Bioinformatics history:
  - 1758 – work of Carl Linnaeus – taxonomy
  - mid 1800s – work of Gregor Mendel – genetics
  - mid 1800s – work of Charles Darwin – evolution
  - 1953 – a major discovery: Structure of DNA
  - 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
  - Oct, 2013 – first Bioinformatics Nobel Prize (Chemistry) to Karplus, Warshel, and Levitt for computer models for xomplec chemical processes.

Algorithms and Hardware

Algorithms and Hardware

Moore's Law Faster processors, more and faster memory, larger external memories

Algorithms and Hardware

Moore's Law	Faster processors, more and faster memory, larger external memories
Optimization	“Linear Programming is tractable”

Algorithms and Hardware

Moore's Law	Faster processors, more and faster memory, larger external memories
Optimization	"Linear Programming is tractable"
Convex Programming	Interior Point Methods

Algorithms and Hardware

Moore's Law	Faster processors, more 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, ...)

Algorithms and Hardware

Moore's Law	Faster processors, more 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

Algorithms and Hardware

Moore's Law	Faster processors, more 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

Algorithms and Hardware

Moore's Law	Faster processors, more 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

What is Bioinformatics?

- Analysis of biological data with informatics tools

What is Bioinformatics?

- Analysis of biological data with informatics tools
- What is Informatics?
    - Data Management: Databases, internet

What is Bioinformatics?

- Analysis of biological data with informatics tools
- What is Informatics?
  - Data Management: Databases, internet
  - Data Analysis: Mining, Modeling, Statistics
  - Algorithm Design: Efficiency, Big Data
  - Visualization and Interface Design: HCI, Graphics
- Why?

What is Bioinformatics?

- Analysis of biological data with informatics tools
- What is Informatics?
    - Data Management: Databases, internet
    - Data Analysis: Mining, Modeling, Statistics
    - Algorithm Design: Efficiency, Big Data
    - Visualization and Interface Design: HCI, Graphics
- Why? – to aid biomedical research

What is Bioinformatics?

- Analysis of biological data with informatics tools
- What is Informatics?
  - Data Management: Databases, internet
  - Data Analysis: Mining, Modeling, Statistics
  - Algorithm Design: Efficiency, Big Data
  - Visualization and Interface Design: HCI, Graphics
- Why? – to aid biomedical research
  - Build databases and efficient tools for search, retrieval, analysis and visualization

What is Bioinformatics?

- Analysis of biological data with informatics tools
- What is Informatics?
  - Data Management: Databases, internet
  - Data Analysis: Mining, Modeling, Statistics
  - Algorithm Design: Efficiency, Big Data
  - Visualization and Interface Design: HCI, Graphics
- Why? – to aid biomedical research
  - Build databases and efficient tools for search, retrieval, analysis and visualization
  - Use predictions to narrow down search

What is Bioinformatics?

- Analysis of biological data with informatics tools
- What is Informatics?
  - Data Management: Databases, internet
  - Data Analysis: Mining, Modeling, Statistics
  - Algorithm Design: Efficiency, Big Data
  - Visualization and Interface Design: HCI, Graphics
- Why? – to aid biomedical research
  - Build databases and efficient tools for search, retrieval, analysis and visualization
  - Use predictions to narrow down search
  - Propose models and build tools to verify models efficiently

What is Bioinformatics?

- Analysis of biological data with informatics tools
- What is Informatics?
  - Data Management: Databases, internet
  - Data Analysis: Mining, Modeling, Statistics
  - Algorithm Design: Efficiency, Big Data
  - Visualization and Interface Design: HCI, Graphics
- Why? – to aid biomedical research
  - Build databases and efficient tools for search, retrieval, analysis and visualization
  - Use predictions to narrow down search
  - Propose models and build tools to verify models efficiently
  - Propose new experiments based on model/analysis

What is Bioinformatics?

- Analysis of biological data with informatics tools
- What is Informatics?
  - Data Management: Databases, internet
  - Data Analysis: Mining, Modeling, Statistics
  - Algorithm Design: Efficiency, Big Data
  - Visualization and Interface Design: HCI, Graphics
- Why? – to aid biomedical research
  - Build databases and efficient tools for search, retrieval, analysis and visualization
  - Use predictions to narrow down search
  - Propose models and build tools to verify models efficiently
  - Propose new experiments based on model/analysis
  - Build smart, hyperlinked, integrated analytical envoronments

What is Bioinformatics?

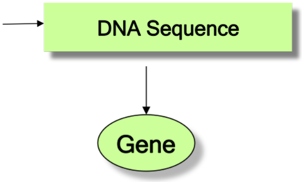
- Analysis of biological data with informatics tools
- What is Informatics?
  - Data Management: Databases, internet
  - Data Analysis: Mining, Modeling, Statistics
  - Algorithm Design: Efficiency, Big Data
  - Visualization and Interface Design: HCI, Graphics
- Why? – to aid biomedical research
  - Build databases and efficient tools for search, retrieval, analysis and visualization
  - Use predictions to narrow down search
  - Propose models and build tools to verify models efficiently
  - Propose new experiments based on model/analysis
  - Build smart, hyperlinked, integrated analytical envoronments

Presentation Outline

Broad Goals of Bioinformatics

DNA Sequence

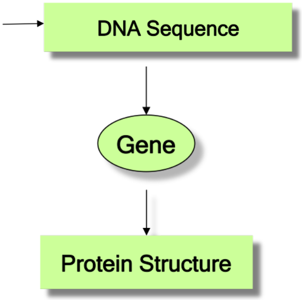
Broad Goals of Bioinformatics



DNA Sequence

Gene

Broad Goals of Bioinformatics

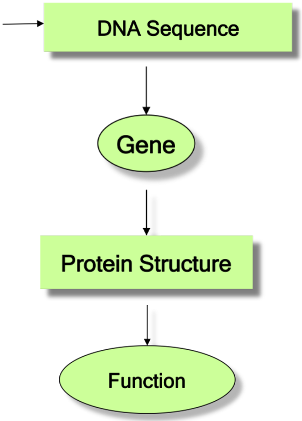


DNA Sequence

Gene

Protein Structure

Broad Goals of Bioinformatics



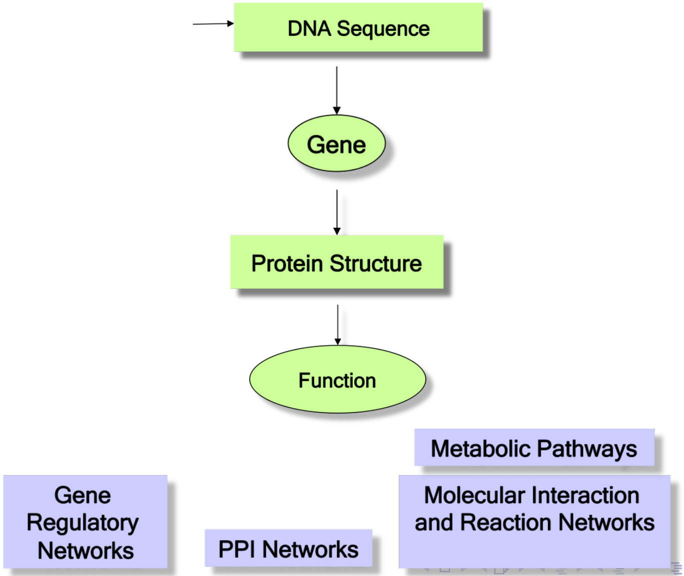
DNA Sequence

Gene

Protein Structure

Function

Broad Goals of Bioinformatics


DNA Sequence

Gene

Protein Structure

Function

Metabolic Pathways

Gene Regulatory Networks

PPI Networks

Molecular Interaction and Reaction Networks

Perspectives in Bioinformatics

- **Molecular:** DNA, RNA, proteins, ligands, toxins, . . .

Perspectives in Bioinformatics

- **Molecular:** DNA, RNA, proteins, ligands, toxins, . . .
- **Cellular** chromosome, nucleus, cell wall, chromatin, organelles, organization of a single cell

Perspectives in 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

Perspectives in 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

Perspectives in 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

Perspectives in 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
- **All life:** Tree of life, phylogeny, variations, comparative studies

Presentation Outline

Growth of Genomic Databases . . .?

- Life on Earth is diverse & complex. About 8.5M eukaryotic species [Mora, C., et al., (2011). PLoS Biol, 9(8)];

Growth of Genomic Databases . . .?

- Life on Earth is diverse & complex. About 8.5M eukaryotic species [Mora, C., et al., (2011). PLoS Biol, 9(8)];
- Human Genome has 3 billion bp with 32,000+ genes.

Growth of Genomic Databases . . . ?

- Life on Earth is diverse & complex. About 8.5M eukaryotic species [Mora, C., et al., (2011). PLoS Biol, 9(8)];
- Human Genome has 3 billion bp with 32,000+ genes.
- GenBank Release 157/175/193/**205** (Dec 2006/09/12/**14**) contains over 64/112/161/**179** million sequence entries totaling over 69/110/126/**184** Gb from over 2,500/?/9000/**11000** organisms (Storage: 600 GBytes uncompressed); More at `http://www.ncbi.nlm.nih.gov/genbank/statistics`

Growth of Genomic Databases . . . ?

- Life on Earth is diverse & complex. About 8.5M eukaryotic species [Mora, C., et al., (2011). PLoS Biol, 9(8)];
- Human Genome has 3 billion bp with 32,000+ genes.
- GenBank Release 157/175/193/**205** (Dec 2006/09/12/**14**) contains over 64/112/161/**179** million sequence entries totaling over 69/110/126/**184** Gb from over 2,500/?/9000/**11000** organisms (Storage: 600 GBytes uncompressed); More at `http://www.ncbi.nlm.nih.gov/genbank/statistics`
- 435/624/3880/**30,000** complete microbial genomes sequenced of which **4500** are virus genomes.

Growth of Genomic Databases . . .?

- Life on Earth is diverse & complex. About 8.5M eukaryotic species [Mora, C., et al., (2011). PLoS Biol, 9(8)];
- Human Genome has 3 billion bp with 32,000+ genes.
- GenBank Release 157/175/193/**205** (Dec 2006/09/12/**14**) contains over 64/112/161/**179** million sequence entries totaling over 69/110/126/**184** Gb from over 2,500/?/9000/**11000** organisms (Storage: 600 GBytes uncompressed); More at `http://www.ncbi.nlm.nih.gov/genbank/statistics`
- 435/624/3880/**30,000** complete microbial genomes sequenced of which **4500** are virus genomes.
- UniProtKB/Swiss-Prot Release 54.7/2012_11/**2015_01** (Jan08/Nov'12/**Jan'15**): 333K/530K/**550K** entries; 120/191/**194** million amino acids.

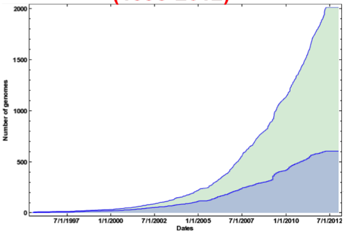
Growth of Genomic Databases . . .?

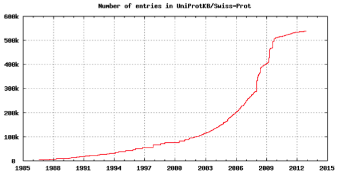
- For GenBank growth see: `http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/` & `http://www.ncbi.nlm.nih.gov/genbank/statistics`

Growth of Genomic Databases . . .?

- For GenBank growth see: `http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/` & `http://www.ncbi.nlm.nih.gov/genbank/statistics`
- Microbial and SwissProt growth:

Microbial Genome Growth (1995-2012)





SwissProt Growth (1985-2012)

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals:

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa, Plasmodium falciparum,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa, Plasmodium falciparum, Drosophila melanogaster,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa, Plasmodium falciparum, Drosophila melanogaster, Anopheles gambiae,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa, Plasmodium falciparum, Drosophila melanogaster, Anopheles gambiae, Macaca mulatta,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa, Plasmodium falciparum, Drosophila melanogaster, Anopheles gambiae, Macaca mulatta, Bos taurus,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa, Plasmodium falciparum, Drosophila melanogaster, Anopheles gambiae, Macaca mulatta, Bos taurus, Felis catus,

And the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: 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 the Diversity . . .

- Over 1800 complete eukaryotic genomes sequenced with 159 mammals, 203 plants, 380 other animals: 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
- Read more at: `http://www.the-scientist.com/?articles.list/tagNo/2416/tags/whole-genome-sequencing/`

And the Genome Sizes . . .

Organism	Size	Date	No. of Genes (est.)
HIV Type I	9.2 Kb	1997	9
M. genitalium	0.58 Mb	1998	525
H. influenzae	1.8 Mb	1995	1,740
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

Presentation Outline

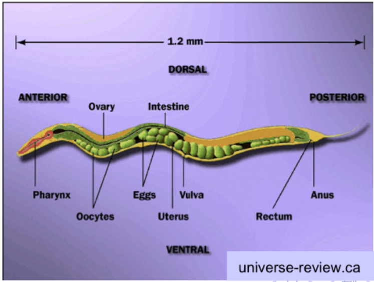
Short Homework

- What is the C-value of an organism? What is its relationship to the genome size? Read aboutthe *C-value enigma* in:
  - Gregory, TR (2004). “Macroevolution, hierarchy theory, and the C-value enigma”. *Paleobiology* **30**:179-202. or
  - Gregory TR (2001). ”Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma”. *Biological Reviews* **76**(1): 65101
- Name at least one organims whose genome is considerably larger than that of humans. How big is that genome?
- What is the *Ferrari of the virus world* and why is it called so?
- What is the *1000 Genomes* project? Look at: `http://www.nature.com/nature/focus/1000genomes/`
- What is *antibiotic resistance* and why is it a public health crisis? What is the name of a new antibiotic reported in *Nature* in 2015 that kills pathogens without detectable

Presentation Outline

Caenorhabditis Elgeans: the model worm

www.ucl.ac.uk

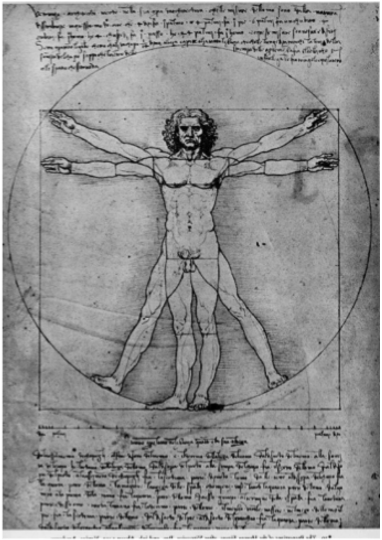


universe-review.ca

Caenorhabditis Elgeans: 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

H. sapiens

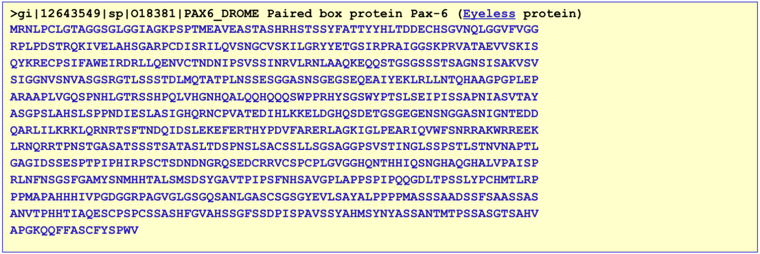


H. sapiens

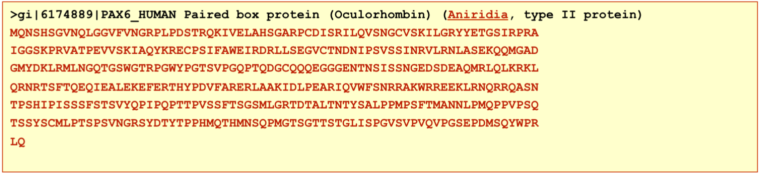
- Sequenced 2001; 15 year effort
- 3 billion bases, 500 gaps
- Variable density of Genes, SNPs, CpG islands
- 1.1% of genome codes for proteins;
- 40-48% of the genome consists of repeat sequences
- 10 % of the genome consists of repeats called ALUs
- 5 % of the genome consists of long repeats (¿1 Kb)
- 223 genes common with bacteria that are missing from worm, fly or yeast.

Presentation Outline

Sequence Alignment

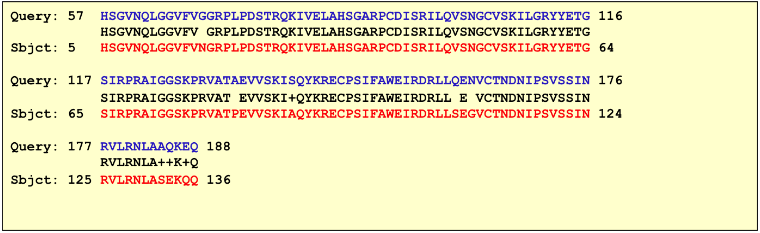


```
>gi|12643549|sp|O18381|PAX6_DROME Paired box protein Pax-6 (Eyeless protein)
MRNLPCLGTAGGSGLGGIAGKPSPTMEAVEASTASHRHSTSSYFATTYYHLTDDECHSGVNQLGGVFVGG
RPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKIS
QYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTSAGNSISAKVSV
SIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEP
ARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSWPPRHYSGSWYPTSLSEIPISSAPNIASVTAY
ASGPSLAHSLSPPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDD
QARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEK
LRNQRRTPNSTGASATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPTL
GAGIDSSESPTPIPHIRPSCTSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHAQGHALVPAISP
RLNFNSGSFGAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSPIPQQGDLTPSSLYPCHMTLRP
PPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGASCSGSGYEVLSAYALPPPPMASSSAADSSFSAASSAS
ANVTPHHTIAQESCPSPCSSASHFGVAHSSGFSSDPISPAVSSYAHMSYNYASSANTMTPSSASGTSAHV
APGKQQFFASCFYSPWV
```

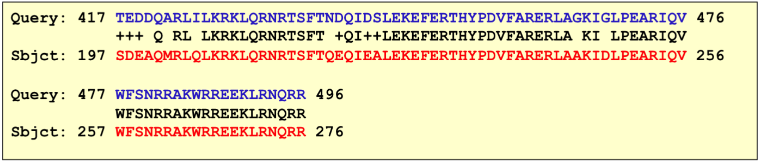


```
>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRA
IGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGAD
GMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKL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

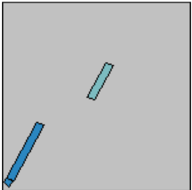
Drosophila Eyeless vs. Human Aniridia



```
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
```



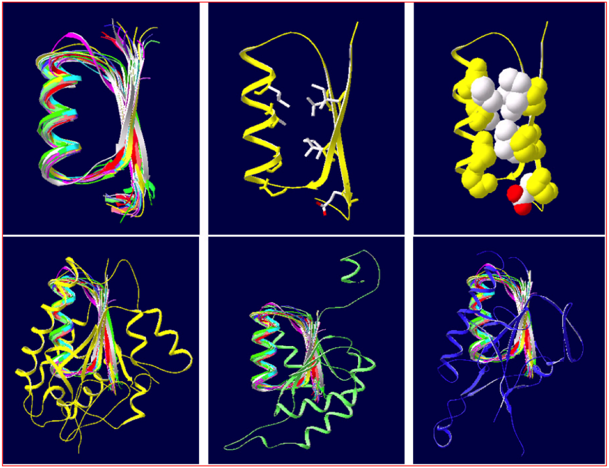
Motif Detection in Protein Sequences

☐ MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

☐ MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

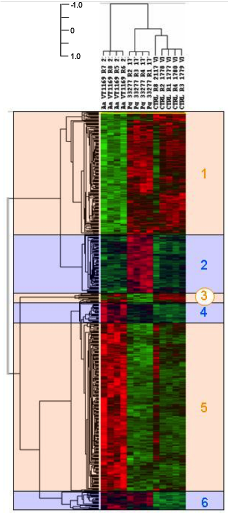
[G. Narasimhan, et al., “Mining Protein Sequences for Motifs,” *J. of Comput Biol*, **9**(5):707-720, 2002.]

Patterns in Protein Structures



[T. Milledge, et al., "Sequence Structure Patterns: Discovery and Applications," *Proc 6th Symposium on Comput Biol & Genome Informatics (CBG)*, 1282-1285, 2005.]

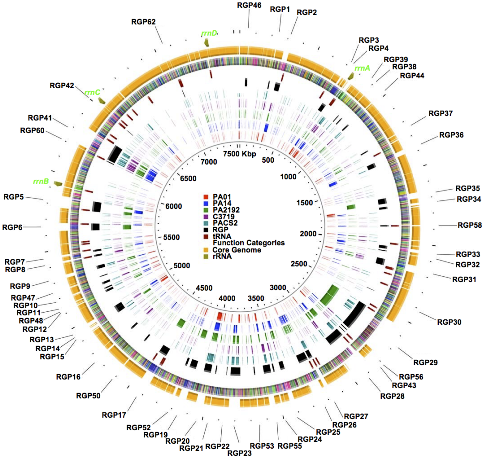
Microarray Analysis



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

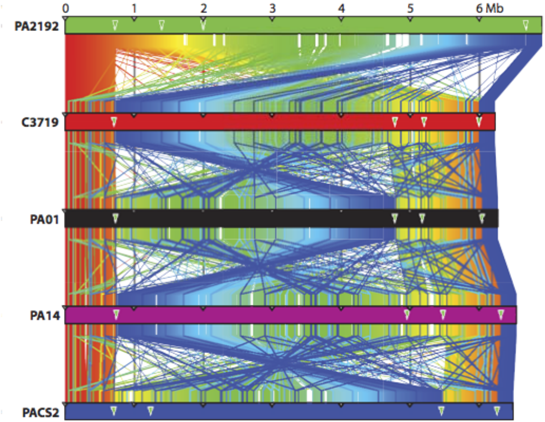
[M. Handfield, et al., "Distinct Expression Profiles Characterize Oral Epithelium-Microbiota Interactions," *Cellular Microbiol*, **7**(6):811-823, 2005.]

Tools: CGView for Comparative Genomics



[K. Mathee, et al., "Dynamics of Pseudomonas aeruginosa genome evolution," *Proc Natl Acad of Sciences (PNAS)*,

Tools: Murasaki for Comparative Genomics



[K. Mathee, et al., "Dynamics of Pseudomonas aeruginosa genome evolution," *Proc Natl Acad of Sciences (PNAS)*,

SIDS

SIDS

- 18000 Amish people in Pennsylvania
- Mostly intermarried due to religious doctrine
- rare recessive diseases occurred with high frequencies

SIDS

- 18000 Amish people in Pennsylvania
- Mostly intermarried due to religious doctrine
- rare recessive diseases occurred with high frequencies
- SIDS: 3000 deaths/yr (US); 21 deaths/yr (Amish)

SIDS

- 18000 Amish people in Pennsylvania
- Mostly intermarried due to religious doctrine
- rare recessive diseases occurred with high frequencies
- SIDS: 3000 deaths/yr (US); 21 deaths/yr (Amish)
- Many research centers failed to identify cause
- Collaboration between Affymetrix, TGEN & Clinic for special children solved the problem in 2 months

SIDS

- 18000 Amish people in Pennsylvania
- Mostly intermarried due to religious doctrine
- rare recessive diseases occurred with high frequencies
- SIDS: 3000 deaths/yr (US); 21 deaths/yr (Amish)
- 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
- Experiments showed that all sick infants had two mutant copies of a specific gene; parents were carriers of mutation.

SIDS

- 18000 Amish people in Pennsylvania
- Mostly intermarried due to religious doctrine
- rare recessive diseases occurred with high frequencies
- SIDS: 3000 deaths/yr (US); 21 deaths/yr (Amish)
- 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
- Experiments showed that all sick infants had two mutant copies of a specific gene; parents were carriers of mutation.
- Conclusion: Disease caused by 2 abnormal copies of TSPYL gene & is expressed in key organs (brainstem,testes) `http://www.affymetrix.com/community/wayahead/modern_miracle.affx`