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

1. Course Preliminaries
2. Introduction
3. Goals & Perspectives
4. Phenomenal Growth
5. Short Homework
6. Model Organisms
7. Motivating Bioinformatics
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

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semester Project</td>
<td>45%</td>
</tr>
<tr>
<td>Homework Assignments</td>
<td>20%</td>
</tr>
<tr>
<td>Exam</td>
<td>15%</td>
</tr>
<tr>
<td>Quizzes</td>
<td>10%</td>
</tr>
<tr>
<td>Summary Reports of Interest</td>
<td>5%</td>
</tr>
<tr>
<td>Class Participation</td>
<td>5%</td>
</tr>
</tbody>
</table>
Presentation Outline

1. Course Preliminaries
2. Introduction
3. Goals & Perspectives
4. Phenomenal Growth
5. Short Homework
6. Model Organisms
7. Motivating Bioinformatics
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: Structure of DNA
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 complex chemical processes*. 
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
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
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 environments
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 environments
Broad Goals of Bioinformatics

- DNA Sequence

Course Preliminaries
Introduction
Goals & Perspectives
Phenomenal Growth
Short Homework
Model Organisms
Motivating Bioinformatics
Broad Goals of Bioinformatics

DNA Sequence

Gene
Broad Goals of Bioinformatics

- DNA Sequence
- Gene
- Protein Structure
Broad Goals of Bioinformatics
Broad Goals of Bioinformatics

1. DNA Sequence → Gene
2. Gene → Protein Structure
3. Protein Structure → Function
4. Function → Metabolic Pathways
5. Function → Molecular Interaction and Reaction Networks
6. Gene Regulatory Networks
7. PPI 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

1. Course Preliminaries
2. Introduction
3. Goals & Perspectives
4. Phenomenal Growth
5. Short Homework
6. Model Organisms
7. Motivating Bioinformatics
Growth of Genomic Databases . . .?

- **Human Genome** has 3 billion bp with 32,000+ genes.

Human Genome has 3 billion bp with 32,000+ genes.


- **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.
Growth of Genomic Databases . . .?

- 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.
Growth of Genomic Databases . . .?

Growth of Genomic Databases . . .?


- Microbial and SwissProt growth:

  ![Microbial Genome Growth (1995-2012)](image1)

  ![SwissProt Growth (1985-2012)](image2)
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,
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 . . .

And the Diversity . . .

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

### And the 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><em>M. genitalium</em></td>
<td>0.58 Mb</td>
<td>1998</td>
<td>525</td>
</tr>
<tr>
<td><em>H. influenzae</em></td>
<td>1.8 Mb</td>
<td>1995</td>
<td>1,740</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>4.7 Mb</td>
<td>1997</td>
<td>4,000</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>12.1 Mb</td>
<td>1996</td>
<td>6,034</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>97 Mb</td>
<td>1998</td>
<td>19,099</td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>100 Mb</td>
<td>2000</td>
<td>25,000</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>180 Mb</td>
<td>2000</td>
<td>13,061</td>
</tr>
<tr>
<td><em>M. musculus</em></td>
<td>3 Gb</td>
<td>2002</td>
<td>30,000</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>3 Gb</td>
<td>2001</td>
<td>32,000</td>
</tr>
</tbody>
</table>
Presentation Outline

1. Course Preliminaries
2. Introduction
3. Goals & Perspectives
4. Phenomenal Growth
5. Short Homework
6. Model Organisms
7. Motivating Bioinformatics
Short Homework

- What is the C-value of an organism? What is its relationship to the genome size? Read about the C-value enigma in:
  - 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 organism 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/](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

1. Course Preliminaries
2. Introduction
3. Goals & Perspectives
4. Phenomenal Growth
5. Short Homework
6. Model Organisms
7. Motivating Bioinformatics
Caenorhabditis Elgeans: the model worm
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

1. Course Preliminaries
2. Introduction
3. Goals & Perspectives
4. Phenomenal Growth
5. Short Homework
6. Model Organisms
7. Motivating Bioinformatics
Sequence Alignment

>gi|12643549|sp|Q18381|PAX6_DROME Paired box protein Pax-6 (Eyeless protein)
MRNLPCGLTAAGSLGGGIAAGKSPTMEAVEASTSSHRTSTSFATYYHHTDCHSGVNLGTVFGG
RPLPDDTRQKIPELASHGPPCPDISLRLQVNGSCVSISLIGYETGRPRAIGSKPRVATAEVSISKIS
QYKREDPSIFAIWEIRDDLQENVCTNNDIPSSINRVLRLNAQKEQQSTGSSSTSAAGSIAKVSV
SIGGNVSNVASGSRGTSSLSDLMQTATPLNSESGGASNSGEGSEQEAIYKLRLLNTQHAAGPGPEP
ARAAPLVQGPSNHGTRSSHPOQVGNHQAQLQHQQQSWPPHRYSGWSYPTLSLEIPISSAPNISAVTAY
ASGSPSAILSISPNDIESLASICGHQRGCPVATEDILHKELDGHQSDETSGEGENSNNGASNITDEDD
QRALILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVDFARERLAKIKGLPEARIOQVFWSNNRAKWRREEK
LRNQRPTNPTSTGASATSSSTASATSLTSDPSNLSSACSSLSSGAGPSVSTINGLSSPSTLSTNVNAPTL
GAGIDSSESPTPIPHIRSPCTSDNDRGQSEDCCRCVSCPCPLGQVGHQNTTHIQSNGHQAQGLVPAISP
RLNFNSGSFGAMYSNMMHTALSMSDSYGAIVPSSIPNHSAVGFLAPPSSIPQGDLTPSSLYPCHMTLRP
PPMAMPHHIVPDCGRVQPAGVGLGQSANLGAACSNGSYEVLASYALPPPMASAADSSFSAASSAS
ANVTPTHTIAQESCPSFCSSASHFVHASSFSGFSDFIPAVSSYAHMSYASSANTMPSSAAGSFAHV
APGKQQFFASCFSYPWV

>gi|61748889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
MQNSHSGVNLGTVFGYVGRPLPDDTRQKIPELASHGPPCPDISLRLQVNGSCVSISLIGYETGRPRA
IGSKPRVATPEVSKIAQYKREDPSIFAIWEIRDDLQENVCTNNDIPSSINRVLRLNAQKEQQSTGSSSTSAAGSIAKVSV
GMYDKLRMLNQGTSGWTRGRQPTGVPSGVQPQDGQQEGGEINSSNSNGEDSDEAQMRPLQRLKLKRKLR
QRNRTSFTQIEALEKEFERTHYPDVDFARERLAAKIDLPEARIOQVFWSNNRAKWRREEKLRNQRQASN
TPSHIIPSSFSFTSVQPIQPFTPPSVSFTSGMLRGTDLANTYLSALPPMPSTMTANNLMPFHPSPQSVS
TSSYSCMLPTSPSVNRGSDYTPTPHMQTHMSQPMGTSSTGTTSGTISPGVSVPVQVPSPEDMSQYWPR
LQ
**Drosophila Eyeless vs. Human Aniridia**

**Query:** 57  
HSGVNQLGGVFVGGRPLDPDRQKIVELAHSGARPCDISRLQVSNVCVSILGRYETG 116  
HSGVNQLGGVFVGGRPLDPDRQKIVELAHSGARPCDISRLQVSNVCVSILGRYETG 64

**Sbjct:** 5  
HSGVNQLGGVFVGGRPLDPDRQKIVELAHSGARPCDISRLQVSNVCVSILGRYETG 116  
HSGVNQLGGVFVGGRPLDPDRQKIVELAHSGARPCDISRLQVSNVCVSILGRYETG 64

**Query:** 117  
SIRPRAIGGSKPRVATAEVVSKISPYKRECPSIFAWEIRDRLLQENVCTNPDIPSUSSIN 176  
SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E VCTNPDIPSUSSIN

**Sbjct:** 65  
SIRPRAIGGSKPRVATPEVSVKIAQYKRECPSIFAWEIRDRLLSEGVCNTNPDIPSUSSIN 124

**Query:** 177  
RVRNLAQQEIQ 188  
RVRNLA+++K+Q  
RVRNLAQQEIQ 136

**Sbjct:** 125  
RVRNLAQQEIQ 136

**Query:** 417  
TEDDOQRALILKRLQRNRSTNDQIDSLEKEFERTYPDVFARERLAGKIGLPEARIQV 476  
+++ Q RL KLRQNRNRTSF +QI++LEKEFERTYPDVFARERLA KI LPEARIQV

**Sbjct:** 197  
SDEAQMLQLKRLQRNRSTQEQIEALEKEFERTYPDVFARERLAAKIDLPEARIQV 256

**Query:** 477  
WFSNRRAKWRREEKLRNQR 496  
WFSNRRAKWRREEKLRNQR  
WFSNRRAKWRREEKLRNQR

**Sbjct:** 257  
WFSNRRAKWRREEKLRNQR 276
Motif Detection in Protein Sequences

- MTDKMQSLALAPVGNLDSYIRAAANAWPLSADDEERALAEKLHYHGDLLEAAKTLILSHLRFVVHIARNYAGYGLPQADLIGEIGNIGLMIKAVRRFNPEVGPVRLVSFAVHWIKAIEIHEYVLRNWRIVKVATTKAQRLFFNLRKTKQRLGWFRQDEVEMVARELGVTSDKVREMERSMAAQDMDTLSDPPDSDSQPMPAVLYLQDKSSNFDAGIEDDNWEEQAAANRTLDMQGLDERSQDIIRARWLDENKSTLQELADRYGVSAERVQLEKNAMKKLRAAIEA

- MTDKMQSLALAPVGNLDSYIRAAANAWPLSADDEERALAEKLHYHGDLLEAAKTLILSHLRFVVHIARNYAGYGLPQADLIGEIGNIGLMIKAVRRFNPEVGPVRLVSFAVHWIKAIEIHEYVLRNWRIVKVATTKAQRLFFNLRKTKQRLGWFRQDEVEMVARELGVTSDKVREMERSMAAQDMDTLSDPPDSDSQPMPAVLYLQDKSSNFDAGIEDDNWEEQAAANRTLDMQGLDERSQDIIRARWLDENKSTLQELADRYGVSAERVQLEKNAMKKLRAAIEA

[G. Narasimhan, et al., “Mining Protein Sequences for Motifs,” 
Patterns in Protein Structures

Microarray Analysis

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

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

18,000 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)

SIDS

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