

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/BioinfS07.html

Overview of Courses

- Sequence Alignment; Multiple Sequence Alignment
- Sequence Analysis
- Sequencing and Mapping
- Phylogenetic Analysis
- Gene prediction techniques
- Pattern discovery techniques
- Protein structure alignment and analysis
- Genomics, Functional Genomics, Proteomics
- Gene Expression Data Analysis
- RNA Secondary structure
- RNA interference and small RNA
- Ribozymes and Riboswitches
- 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*)
- ❑ Phylogenetic Analysis (*CLUSTALW, Phylip, PAML*)
- ❑ Learning Methods (*HMMPPro, GeneCluster, ASOM*)
- ❑ Pattern Discovery Techniques (*GYM, TEIRESIAS, APRIORI*)
- ❑ Molecular Structure Analysis (*DALI, RASMOL, SPDBV*)
- ❑ Microarray Analysis (*CLUSTER, GeneCluster, TreeView*)
- ❑ Statistical Software Packages (*SAS, R*)

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

Evaluation

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

Course Homepage

www.cis.fiu.edu/~giri/teach/BioinfS07.html

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

Introduction

1. What is Bioinformatics?

- Analysis of biological data with computing & statistical tools.

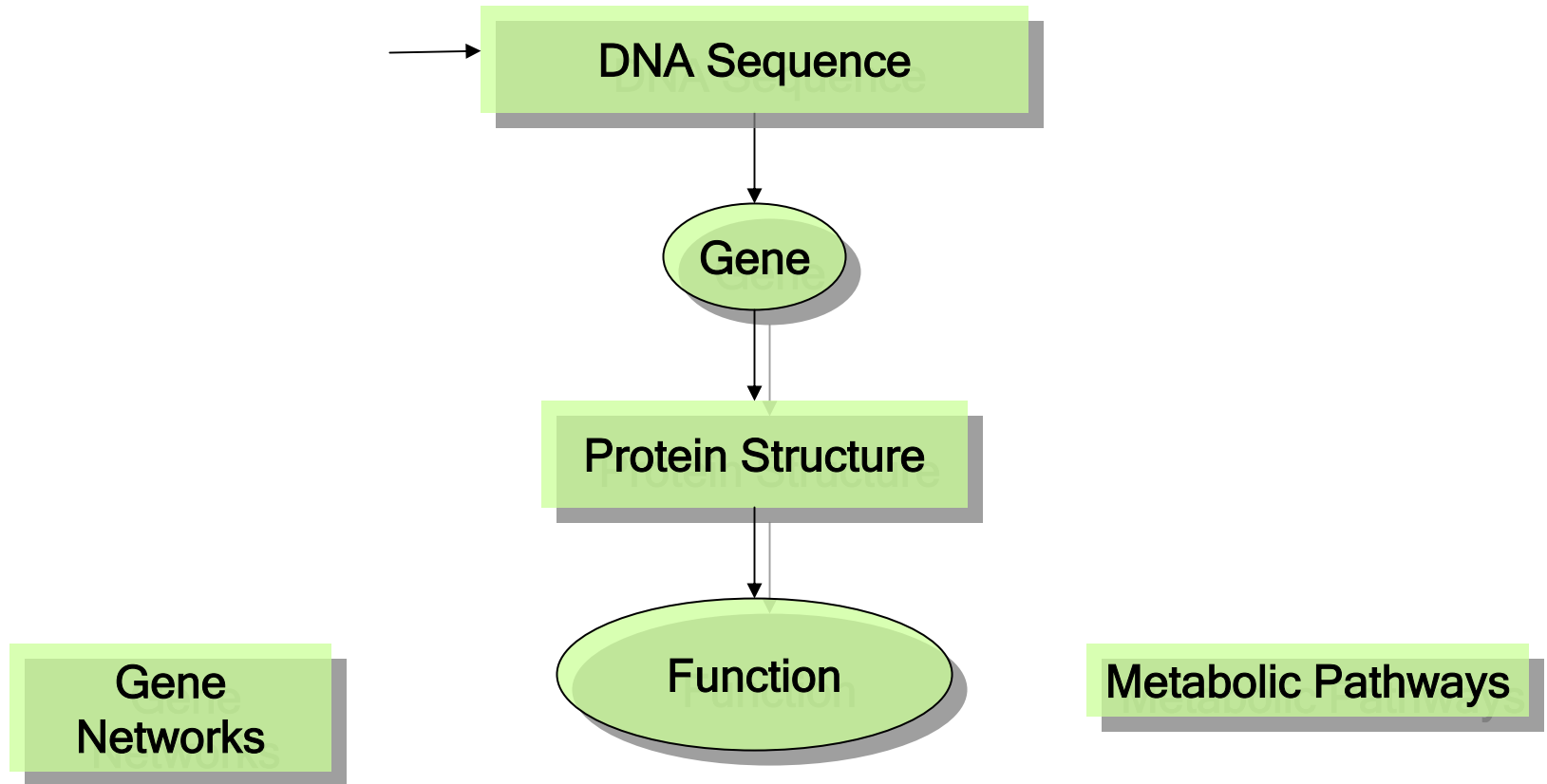
2. The different aspects of Informatics?

- 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 157 (Dec 2006) contains over 78 million sequence entries totaling over 135 Gb from over 2,645 organisms [<http://www.ncbi.nlm.nih.gov>] (Storage: 147 GB uncompressed)
- ❑ **Human Genome** has ~3 billion bp with 32,000+ genes.
- ❑ 435 complete **microbial** genomes sequenced (684 more in progress)
- ❑ 2540 **Viral** genomes (300bp - 300Kb) (1st 1978: Simian virus; 5Kb).
- ❑ 23 complete **eukaryotic** genomes sequenced (175 more in progress):
Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus, Homo sapiens, Oryza sativa,
- ❑ 124 organisms have assemblies and chromosomal maps including:
Drosophila melanogaster, Anopheles gambiae, Macaca mulatta, Bos taurus, Felis catus
- ❑ **Swiss-Prot** Release 51.3 (Dec 2006): 250296 entries; 91 million amino acids.

Genome Sizes

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

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)
- ❑ 223 genes common with bacteria that are missing from worm, fly or yeast.