COT 6936 (3)
Introduction to Bioinformatics

CAP 6990 (2)
Bioinformatics Tools

Giri Narasimhan

Course Schedules

- COT 6936 (3 credit) will meet every Tue and Thu at 9:30 AM
- CAP 6990 (2 credit) will meet every Tue and Thu at 9:30 AM except for the Tuesdays between Sep 17 and Nov 19, 2002. This course is not for CS students.
- Different exams and evaluation.

COT 6936: Topics in Algorithms (Introduction to Bioinformatics)

Overview of Course

- Preliminaries
- Sequence Alignment
- Multiple Sequence Alignment
- Phylogenetic Analysis
- Molecular Structure Analysis
- Gene Recognition
- Genomics, Functional Genomics
- Proteomics
- Databases and Software Packages
- Statistical Analysis of Sequences
- Sequencing and Mapping
- Computational Learning Methods
  HMM, NN, SOM, SVM, GA
- Computational Predictive Methods
- Pattern Discovery Techniques
- Emerging Biotechnologies

COT 6936/6990

Software Packages

- Databases and Software Packages (GenBank, SWISS-PROT, GCG, Sequencher, DNAStar, Vector Nti)
- Sequence Alignment & Multiple Sequence Alignment (BLAST, Pileup, CLUSTAL, DBCLUSTAL)
- Phylogenetic Analysis (CLUSTALW, Phylip, LAMARC)
- Sequencing and Mapping (Genotyper, MapMaker)
- Learning Methods (HMMPro, GeneCluster)
- Pattern Discovery Techniques (GYM, TEIRESIAS)
- Molecular Structure Analysis (DALI, RASMOL)
- Microarray Analysis (CLUSTER, SAM, GeneCluster, TreeView)

Evaluation

- Semester Project (50 %)
- Homework Assignments (20 %)
- Exams (20 %)
- Class Participation (10 %)

Introduction

• 1. What is Bioinformatics?

- Investigation of biological questions with computer technology.

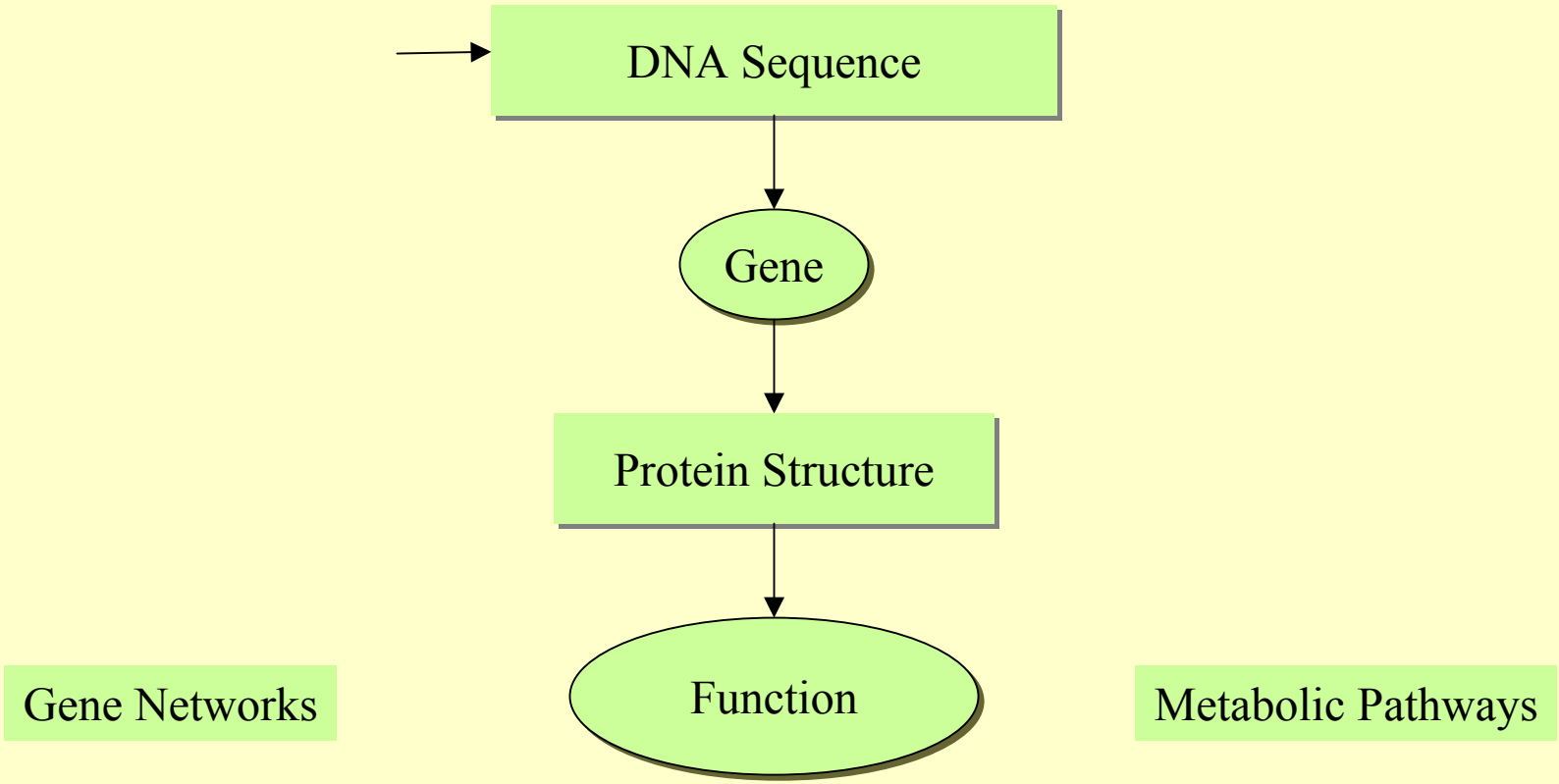
• 2. The different aspects of Informatics:

- 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?

- use predicted information to narrow down search
- verify a proposed model
- propose experiments based on model or information

Overall Goals



General Information

- **Number of bases in the NCBI database is over 20 billion.**
- **Human Genome has ~3 billion bp.**
- **Human Genome has 26,383 + 12,000 genes.**
- **86 complete microbial genomes sequenced.**
- **Number of whole genomes sequenced – 91, including:**
  Caenorhabditis Elegans, Arabidopsis Thaliana, Drosophila Melanogaster, Saccharomyces Cerevisiae,
- **Chromosomal maps for many organisms including:**
  Mus musculus, Homo Sapiens, Danio rerio, Zea mays, Oryza Sativa

Genome Sizes

Organism	Size	Date	Est. # genes
H. influenzae	1.8 Mb	1995	1,740
E. Coli	4.7 Mb	1997	
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		??
H. Sapiens	3 Gb		30,000+

Caenorhabditis Elegans

- Entire genome - 1998
- 1st animal; 26th organism
- 8 year effort
- 97 million bases
- 19,099 genes
- 402 gene clusters
- 12,000 genes with known function
- thousands of mutants
- 7000 families of repeats
- Multicellular organism
- Nematode (phylum)
- Easy to experiment with
- Easily observable
- 959 cells
- 302 nerve cells
- 36% of proteins common w/ human

Homo Sapiens

- 15 year effort, 3 billion bases, 100,000 gaps
- Variable density of:

  Genes, SNPs, CpG islands, recombination rates
- ~1.1% of the 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)
- ~ 50 transposon-derived genes
- 223 genes common with bacteria that are missing from worm, fly or yeast.

(Approximate) String Matching

Input: Text **T** , Pattern **P**

Question(s):

- Does **P** occur in **T**?
- Find one occurrence of **P** in **T**.
- Find all occurrences of **P** in **T**.
- Count # of occurrences of **P** in **T**.
- Find longest substring of **P** in **T**.
- Find closest substring of **P** in **T**.
- Locate direct repeats of **P** in **T**.
- *Many More variants*

Applications:

- Is **P** already in the database **T**?
- Locate **P** in **T**.
- Can **P** be used as a primer for **T**?
- Is **P** homologous to anything in **T**?
- Has **P** been contaminated by **T**?
- Is *prefix*(**P**) = *suffix*(**T**)?
- Locate tandem repeats of **P** in **T**.

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**

Repeats in DNA Sequences

Genomic Imprinting: Some genes are expressed only when inherited from one specific parent.

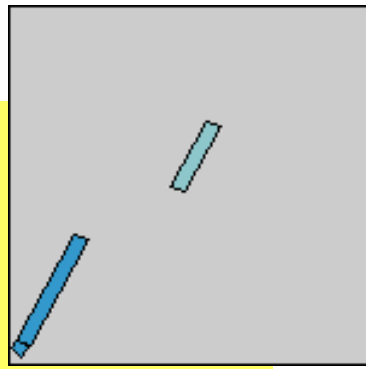
16 such genes are known; 5 inherited from mother; rest from father.

These 16 genes have a lot of **repeats**.

Repeats are of size 25 to 120 bp and of total length 1500.
The repeats are unique to these imprinted regions.
They have no obvious homolgy to each other
or to other highly repetitive mammalian sequences.

Repeats are also known to be responsible for several genetic diseases:
Fragile X, Huntington's disease, Kennedy's disease,
myotonic dystrophy, ataxia.

Drosophila Eyeless vs. Human Aniridia



```
24 IERLPSLEDMAHKGHSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSN 83
   I R P+   M +  HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSN
17 IPRPPARASMQNS-HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSN 75

84 GCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQEN 143
   GCVSKILGRYYETGSIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E
76 GCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEG 135

144 VCTNDNIPSVSSINRVLRNLAAQKEQ 169
    VCTNDNIPSVSSINRVLRNLA++K+Q
136 VCTNDNIPSVSSINRVLRNLASEKQQ 161

398 TEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQV 457
    +++ Q RL LKRKLQRNRTSFT +QI++LEKEFERTHYPDVFARERLA KI LPEARIQV
222 SDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQV 281

458 WFSNRRAKWRREEKLRNQRR 477
    WFSNRRAKWRREEKLRNQRR
282 WFSNRRAKWRREEKLRNQRR 301
```

Sequence Alignment

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++           +L+ L+++H+ K
LGB2_LUPLU  NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
            GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
  F11G11.2  GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

HBA_HUMAN: Human Alpha Globin
HBB_HUMAN: Human Beta Globin
F11G11.2 : Leghaemoglobin from yellow lupin

- **Needleman-Wunsch**
- **Smith-Waterman**

Sequence Alignment

Input: Sequence **A** , Sequence **B**, Database **D**

Question(s):

Align **A** and **B.**

Determine *similarity* (**A** , **B**).

Align **A** and **D.**

Find sequence in **D** with max similarity.

Many More variants

Reading List and Schedule

Watch that Course Home Page!

http://www.cs.fiu.edu/~giri/teach/6936F02.html