BCB 590

Practical Bioinformatics

1

Introductions

- Name
- Lab
- Research Project
- What do you hope to learn in the next two weeks about bioinformatics?

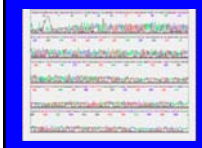
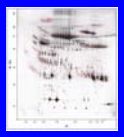
2

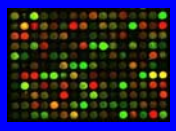
Course Format

- Short introductory lecture
- Lab exercise
- Answer summary questions about the exercise and email answers to the instructor
- Final exam – lab practical
- Show up, do the labs, get an A

3

What is *Bioinformatics? Computational Systems Biology?*

Thanks to Drena Dobbs (ISU), Mark Gerstein (Yale) & Eric Green (NIH) for borrowed & modified PPTs

4

What is Bioinformatics? (& Computational Biology?)

***Wikipedia*:**

- Bioinformatics & computational biology involve the use of techniques from **mathematics**, **informatics**, **statistics** & **computer science** (& engineering) to solve ***biological problems***

5

What is Bioinformatics? (& Computational Biology?)

***Gerstein*:**

- (Molecular) Bioinformatics is conceptualizing **biology in terms of molecules** & applying "**informatics" techniques** - derived from disciplines such as mathematics, computer science, and statistics - to **organize and understand information associated** with these molecules, **on a large scale**

Modified from Mark Gerstein

6

What is Systems Biology?

Wikipedia:

- **Systems biology -** is a relatively new ***biological study field*** that focuses on the systematic study of complex interactions in biological systems, thus using a new perspective (integration instead of reduction) to study them

7

What is the Information?
Biological Sequences, Structures, Processes

Central Dogma of Molecular Biology

- DNA sequence
  - -> RNA
    - -> Protein
      - -> Phenotype
- Molecules
  - Sequence, Structure, Function
- Processes
  - Mechanism, Specificity, Regulation

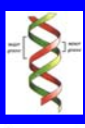
Central Paradigm for Bioinformatics

- Genomic (DNA) Sequence
  - -> mRNAs & other RNA sequences
    - -> Protein sequences
      - -> RNA & Protein Structures
        - -> RNA & Protein Functions
          - -> Phenotype
- Large Amts of Information
  - Standardized
  - Statistical

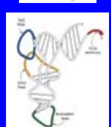
Modified from Mark Gerstein idea from D Brutlag, Stanford, graphics from S Strobel)

8

Explosion of "Omes" & "Omics!"
Genome, Transcriptome, Proteome



- **Genome** - the complete collection of DNA (genes and "non-genes") of an organism



- **Transcriptome** - the complete collection of RNAs (mRNAs & others) expressed in an organism

- **Proteome** - the complete collection of proteins expressed in an organism

9

Even MORE *Omes:*

Glycome

Metabolome

Cytome

Diseasome!!!

10

Genome = Constant *(MOSTLY)*
Transcriptome & Proteome = Variable

- **Genome** - the complete collection of DNA (genes and "non-genes") of an organism
- **Transcriptome** - the complete collection of RNAs (mRNAs & others) expressed in an organism*
- **Proteome** - the complete collection of proteins expressed in an organism*

* Note: Although the DNA is "identical" in all cells of an organism, the sets of RNAs or proteins expressed in different cells & tissues of a single organism vary greatly -- and depend on variables such as environmental conditions, age, developmental stage disease state, etc.

11

Molecular Biology Information:
DNA & RNA Sequences

Functions:

- Genetic material
- Information transfer (mRNA)
- Protein synthesis (tRNA/mRNA)
- Catalytic & regulatory activities (some very new!)

DNA sequence:

**atggcaattaaaattggtatcaatggttttggtcgtat
gcacaacaccgtgatgacattgaagttgtaggtattaa
atggcttatatgttgaaatatgattcaactcacggtcg
aaagatggtaacttagtggttaatggtaaaactatccg
Gcaaacttaaactggggtgcaatcggtgttgatatcgctttaactgatgaaa
ctgctcgtaaacatatcactgcaggcgcaaaaaaagtt**

RNA sequence has "U" instead of "T"

Information:

- 4 letter alphabet (DNA nucleotides: AGCT)
- ~ 1,000 base pairs in a small gene
- ~ 3 X 10^9 bp in a genome (human)

- **Where are the genes?**
- **Which DNA sequences encode mRNA?**
- **Which DNA sequences are "junk"?**
- **Which RNA sequences encode protein?**

Modified from Mark Gerstein

12

Molecular Biology Information: Protein Sequences

Functions: *Most cellular functions are performed or facilitated by proteins*

- Biocatalysis
- Cofactor transport/storage
- Mechanical motion/support
- Immune protection
- Regulation of growth and differentiation

Protein sequences:

d1dhfa_
LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTT
d8dfr__
LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTS

Information:

- 20 letter alphabet (amino acids)
  - ACDEFGHIKLMNPQRSTVWY (but not BJOUXZ)
- ~ 300 aa in an average protein (in bacteria)
- ~ 3 X 10^6 known protein sequences

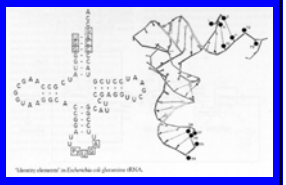
· What *is* this protein?
· Which amino acids are most important -- for folding, activity, interaction with other proteins?
· Which sequence variations are harmful (or beneficial)?

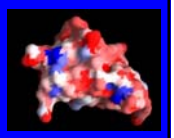
Modified from Mark Gerstein 13

Molecular Biology Information: Macromolecular Structures

DNA/RNA/Protein Structures

- How does a protein (or RNA) sequence fold into an active 3-D structure?
- Can we predict structure from sequence?
- Can we predict function from structure (or perhaps, from sequence alone?)



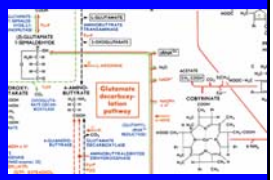


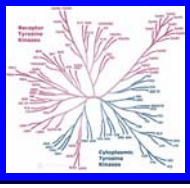
Modified from Mark Gerstein 14

Molecular Biology Information: Biological Processes & Systems

Functional Genomics

- How do patterns of gene expression determine *phenotype*?
- Which genes and proteins are required for differentiation during during normal development? Are involved in disease?
- How do proteins, signaling molecules and metabolites interact in biological networks?
- Which genes and pathways have been most highly conserved during evolution? & *Why*?





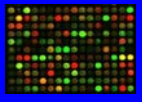
15

Gene Expression Data: *Transcriptome (& Proteome)*

MicroArray Data

Yeast Expression Data:

- Levels for all 6,000 genes!
- Experiments to investigate how genes respond to changes in environment or how patterns of expression change in normal vs cancerous tissue



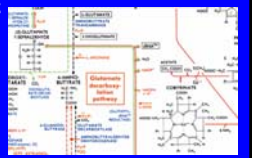
ISU's Biotechnology Facilities include state-of-the-art Microarray & Proteomics instrumentation

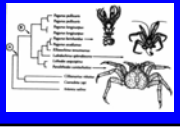
Modified from Mark Gerstein (courtesy of J Hager) 16

Molecular Biology Information: Integrating Data

Understanding the function of genomes requires integration of many diverse and complex types of information:

- Metabolic pathways
- Regulatory networks
- Whole organism physiology
- Evolution, phylogeny
- Environment, ecology
- Literature (MEDLINE)





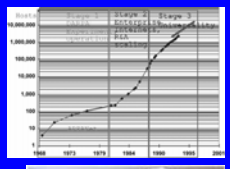
Modified from Mark Gerstein 17

Storing & Analyzing Large-scale Information: Exponential Growth of Data Matched by Development of Computer Technology

CPU vs Disk & Net

- Both the increase in computer speed and the ability to store large amounts of information on computers have been crucial
- Improved computing resources have been a driving force in Bioinformatics

ISU's supercomputer "CyBlue" is (WAS…) among 100 most powerful in the world



Srinivas Aluru with CyBlue and its 2,048 processors. At left is the data storage and equipment to make the supercomputer operational. *Photo by Bob Elbert.*

Modified from Mark Gerstein (Internet picture adapted from D Bruhn, Stanford) 18

Resources for Bioinformatics & Computational Biology

- *Wikipedia: Bioinformatics*
- *Wikipedia: Systems Biology*

- *NCBI - National Center for Biotechnology Information*
- *ISCB - International Society for Computational Biology*
- *JCB - Jena Center for Bioinformatics*
- *UBC - Bioinformatics Links Directory*

19

(a few) ISU Resources & Experts

ISU Research Centers & Graduate Training Programs:

BCB - Bioinformatics & Computational Biology
Baker Center - Bioinformatics & Biological Statistics
CIAG - Center for Integrated Animal Genomics
CILD - Computational Intelligence, Learning & Discovery

ISU Facilities:

Biotech - Instrumentation Facilities
CIAG - Center for Integrated Animal Genomics
PSI - Plant Sciences Institute
PSI Centers

20

Things we won't cover

- Comparative and functional genomics
- Gene prediction
- Microarray analysis
- Proteomics
- Phylogenetics
- Multiple sequence alignment
- Networks
- And more …

21

Topics for our course

- Biological databases
- Making sense of DNA and protein sequences
- BLAST
- RNA structure prediction
- Protein structure and function
- Protein structure prediction
- Protein structure visualization

22

Our focus

- Effective use of some of the most commonly used bioinformatics tools and databases
- How to interpret your results
- How to find the information you need
- How to learn to use new tools

23

- http://dobbslab.gdcb.iastate.edu/BCB590

24