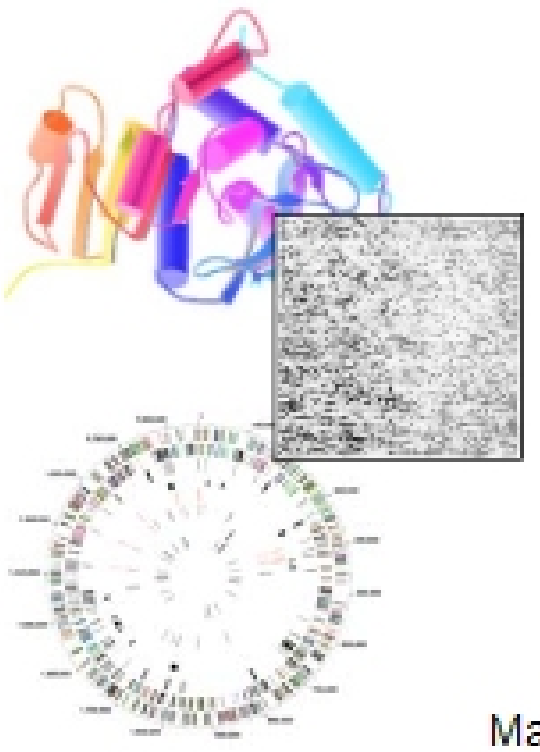
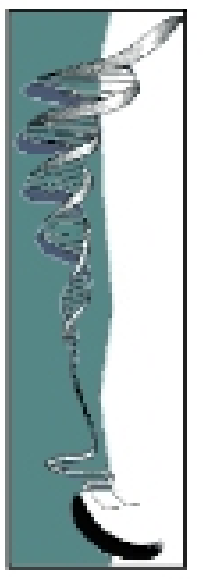
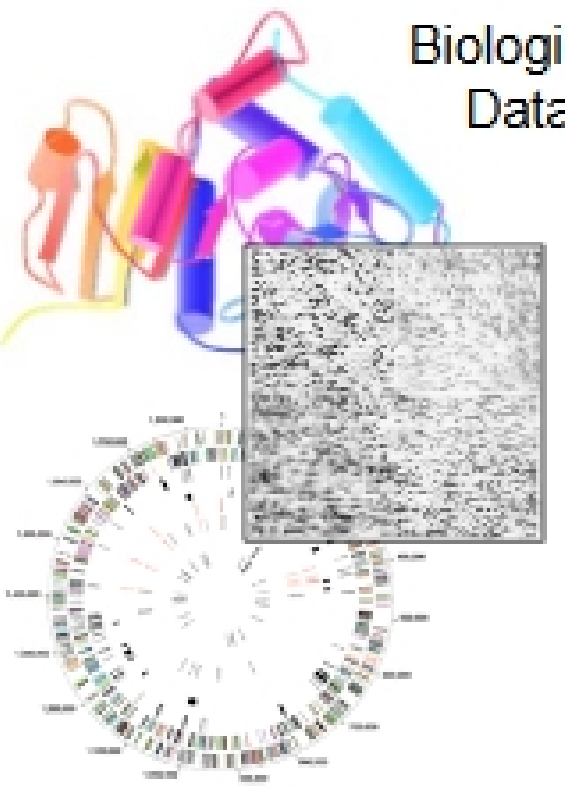
BIOINFORMATICS
Introduction

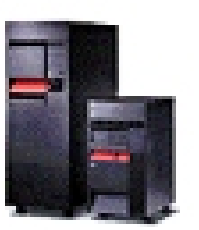
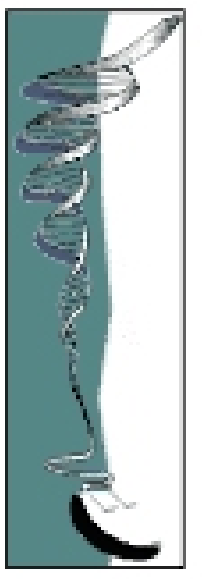




Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mbb452a
(last edit in fall 2002)

Bioinformatics



Biological Data + Computer Calculations


What is Bioinformatics?

Core

- *(Molecular)* **Bio** - `informatics`
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale.**
  - Bioinformatics is "MIS" for Molecular Biology Information. It is a practical discipline with many **applications**.

What is the Information?
Molecular Biology as an Information Science

- Central Dogma of Molecular Biology

```
DNA
 -> RNA
  -> Protein
   -> Phenotype
    -> DNA
```

- Molecules
  - Sequence, Structure, Function
- Processes
  - Mechanism, Specificity, Regulation

- Central Paradigm for Bioinformatics

```
Genomic Sequence Information
 -> mRNA (level)
  -> Protein Sequence
   -> Protein Structure
    -> Protein Function
     -> Phenotype
```

- Large Amounts of Information
  - Standardized
  - Statistical

•Genetic material
- •Information transfer (mRNA)
- •Protein synthesis (tRNA/mRNA)
- •Some catalytic activity

- •Most cellular functions are performed or facilitated by proteins.
- •Primary biocatalyst
- •Cellular transport/storage
- •Mechanical motion/support
- •Immune protection
- •Control of growth/differentiation

[Idea from D Brutlag, Stanford, graphics from S Strobel]

Molecular Biology Information - DNA

- Raw DNA Sequence
  - ◊ Coding or Not?
  - ◊ Parse into genes?
  - ◊ 4 bases: AGCT
  - ◊ ~1 K in a gene, ~2 M in genome

Molecular Biology Information: Protein Sequence

- 20 letter alphabet
  - ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- ~200 K known protein sequences