BIOINFORMATICS
Introduction

Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mbb452a
Bioinformatics

Biological Data + Computer Calculations
What is Bioinformatics?

- **(Molecular) Bio-informatics**
  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 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

(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

atggcaat...
Molecular Biology Information:
Protein Sequence

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

```plaintext
d1dhfa_ LNCIVAVSQNMIGKNGDLPWPRNHEFYFQRMTTTSSVEGKQ-NLVMGKKTWFSI
d8dfr_ LNSIVAVCNMGKGDGNLPPRLNEKYKYFQRMTSTSHVEGKQ-NAVMGKKTWFSI
d4dfr_ ISLIAALAVDRVIGMENAMPW-LPADLAWFKNRLD----------KPVIMGRHTWESI
d3dfr_ TAFLWAQDRDGLIGKDHLPW-LPDDLHYFAQTV----------GKIMVGVRTYESF

d1dhfa_ LNCIVAVSQNMIGKNGDLPWPRNHEFYFQRMTTTSSVEGKQ-NLVMGKKTWFSI
d8dfr_ LNSIVAVCNMGKGDGNLPPRLNEKYKYFQRMTSTSHVEGKQ-NAVMGKKTWFSI
d4dfr_ ISLIAALAVDRVIGMENAMPW-LPADLAWFKNRLD----------KPVIMGRHTWESI
d3dfr_ TAFLWAQDRGNLIGKDHLPW-LPDDLHYFAQTV----------GKIMVGVRTYESF

d1dhfa_ VPEKNRPLKRINLVSLRELEKPPQAGHFLRSLSLDALKLTEQPELANKVMVWIVGGSYYKEAMNH

d8dfr_ VPEKNRPLKRINLVSLRELEKPPQAGHFLRSLSLDALKLTEQPELANKVMVWIVGGSYYKEAMNH

d4dfr_ ---G-RPLGRKNIILS-SSQGTDVR-TWVKSVDIAACGDPV--------EIMVGGGRVYEQFLPKA

d3dfr_ ---PKRP-LPERTNVLTHQEDYQAQFGA-VVHDVAAVFAYAKQHLDQ---ELVIAAGQAIFTAFKDDV

d1dhfa_ -PEKNRPLKRINLVSLRELEKPPQAGHFLRSLSLDALKLTEQPELANKVMVWIVGGSYYKEAMNH

d8dfr_ -PEKNRPLKRINLVSLRELEKPPQAGHFLRSLSLDALKLTEQPELANKVMVWIVGGSYYKEAMNH

d4dfr_ -G----RP-LGRKNIIILS-SSQGTDVR-TWVKSVDIAACGDPVE-------IMVGGGRVYEQFLPKA

d3dfr_ -P---KRPLPERTNVLTHQEDYQAQFGA-VVHDVAAVFAYAKQHLDQ---ELVIAAGQAIFTAFKDDV
```
Molecular Biology Information: Macromolecular Structure

- DNA/RNA/Protein
  - Almost all protein

(RNA Adapted From D. Soll Web Page, Right Hand Top: Protein from M. Levitt web page)
Molecular Biology Information: Whole Genomes

- The Revolution Driving Everything


(Picture adapted from TIGR website, http://www.tigr.org)

- Integrative Data
  1995, HI (bacteria): 1.6 Mb & 1600 genes done
  1997, yeast: 13 Mb & ~6000 genes for yeast
  1998, worm: ~100Mb with 19 K genes
  1999: >30 completed genomes!
  2003, human: 3 Gb & 100 K genes...

Genome sequence now accumulate so quickly that, in less than a week, a single laboratory can produce more bits of data than Shakespeare managed in a lifetime, although the latter make better reading.

---

Gene Expression Datasets: the Transcriptome

Young/Lander, Chips, Abs. Exp.

Brown, Botstein $\mu$array, Rel. Exp. over Timecourse

Also: SAGE; Samson and Church, Chips; Aebersold, Protein Expression

Snyder, Transposons, Protein Exp.
Array Data

Yeast Expression Data in Academia:
levels for all 6000 genes!

Can only sequence genome once but can do an infinite variety of these array experiments

at 10 time points,
6000 x 10 = 60K floats

telling signal from background

(courtesy of J Hager)
microarrays

• Affymetrix
  • Oligos

• Glass slides
  ◊ Pat brown
Systematic Knockouts


Other Whole-Genome Experiments

2 hybrids, linkage maps


For yeast:
6000 x 6000 / 2
~ 18M interactions
Molecular Biology Information: Other Integrative Data

• Information to understand genomes
  ◦ Metabolic Pathways (glycolysis), traditional biochemistry
  ◦ Regulatory Networks
  ◦ Whole Organisms Phylogeny, traditional zoology
  ◦ Environments, Habitats, ecology
  ◦ The Literature (MEDLINE)

• The Future....

(Pathway drawing from P Karp’s EcoCyc, Phylogeny from S J Gould, Dinosaur in a Haystack)
Large-scale Information: GenBank Growth

<table>
<thead>
<tr>
<th>Year</th>
<th>Base Pairs</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>680338</td>
<td>606</td>
</tr>
<tr>
<td>1983</td>
<td>2274029</td>
<td>2427</td>
</tr>
<tr>
<td>1984</td>
<td>3368765</td>
<td>4175</td>
</tr>
<tr>
<td>1985</td>
<td>5204420</td>
<td>5700</td>
</tr>
<tr>
<td>1986</td>
<td>9615371</td>
<td>9978</td>
</tr>
<tr>
<td>1987</td>
<td>15514776</td>
<td>14584</td>
</tr>
<tr>
<td>1988</td>
<td>23800000</td>
<td>20579</td>
</tr>
<tr>
<td>1989</td>
<td>34762585</td>
<td>28791</td>
</tr>
<tr>
<td>1990</td>
<td>49179285</td>
<td>39533</td>
</tr>
<tr>
<td>1991</td>
<td>71947426</td>
<td>55627</td>
</tr>
<tr>
<td>1992</td>
<td>101008486</td>
<td>78608</td>
</tr>
<tr>
<td>1993</td>
<td>157152442</td>
<td>143492</td>
</tr>
<tr>
<td>1994</td>
<td>217102462</td>
<td>215273</td>
</tr>
<tr>
<td>1995</td>
<td>384939485</td>
<td>555694</td>
</tr>
<tr>
<td>1996</td>
<td>651972984</td>
<td>1021211</td>
</tr>
<tr>
<td>1997</td>
<td>1160300687</td>
<td>1765847</td>
</tr>
<tr>
<td>1998</td>
<td>2008761784</td>
<td>2837897</td>
</tr>
<tr>
<td>1999</td>
<td>3841163011</td>
<td>4864570</td>
</tr>
<tr>
<td>2000</td>
<td>8604221980</td>
<td>7077491</td>
</tr>
</tbody>
</table>
Large-scale Information: Exponential Growth of Data Matched by Development of Computer Technology

- **CPU vs Disk & Net**
  - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial.

- **Driving Force in Bioinformatics**

(Internet picture adapted from D Brutlag, Stanford)
The Next Step after the sequence:

Comprehensive Understanding of Gene Function on a Genomic Scale

- Proteomics
- Expression Analysis
- Structural Genomics, Protein Interactions

Step 1: The genome sequence and genes

Evolutionary Implications of Intergenic Regions as Gene Graveyard

Pseudogenes, Regulatory Regions, Repeats

BEHAVIOR OF THE GENES
INTERGENIC REGIONS
The next step:

What is a Conserved Domain?

Domains can be thought of as functional and/or structural units of a protein. These two classifications coexist rather often, and what is found as an independent folding unit of a polypeptide chain also carries out a specific function. Typically domains are identified as recurring (sequence or structure) units, which may exist in various contexts. The image below illustrates a domain, identified as structural units in the PDB entry 1BEP, chain A. (Click on the figure to launch this view in Cn3D.)

For this query sequence, the CD-Search service would identify the conserved domains indicated below (click on the image below to launch the actual search). Good correspondence exists between structural units, identified by purely geometric criteria, and units asserted to be evolutionarily conserved. The region annotated as “Furc-like” was split in two by the PDB domain parser.

For instance, the domain at 90-91 was split into two homologous domains at 90-91 and 92-94 (for which the program used the description “Furc-like”).

Molecular evolution readily utilizes such domains as building blocks which may be recombined in different arrangements to modulate protein function. We define conserved domains as recurring units in molecular evolution whose extents can be determined by sequence and structure analysis.

Conserved domains contain conserved sequence patterns or motifs, which allow for their detection in polypeptide sequences. The distinction between domains and motifs is not sharp, however, especially in the case of short repetitive units. Functional motifs are also present outside the scope of conserved or extended domains. The CD database does not attempt to systematically collect them.
Organizing Molecular Biology Information: Redundancy and Multiplicity

- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genes are grouped into Pathways
- Genomic Sequence Redundancy due to the Genetic Code
- How do we find the similarities?.....

Integrative Genomics -
genomes ↔ structures ↔
functions ↔ pathways ↔
expression levels ↔
regulatory systems ↔ ....
Integrative Genomic Surveys of Many Proteins from Many Perspectives vs “Prediction” Bioinformatics (focused on individual genes and structures)
Major Application I: Designing Drugs

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).
Bioinformatics Subtopics

- Docking & Drug Design
- Protein Geometry
- Protein Flexibility
- Structure Classification
- Homology Modeling
- Sequence Alignment
- Gene Prediction
- Function Classification
- Database Design
- Genome Annotation
- Expression Clustering
- E-literature
- Large-Scale Genomic Surveys

Fold Recognition
Secondary Structure Prediction

Fold Recognition
Secondary Structure Prediction
What is Bioinformatics?

• (Molecular) **Bio-informatics**
  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 a practical discipline with many **applications**.