Introduction to Bioinformatics

Shifra Ben-Dor

Irit Orr

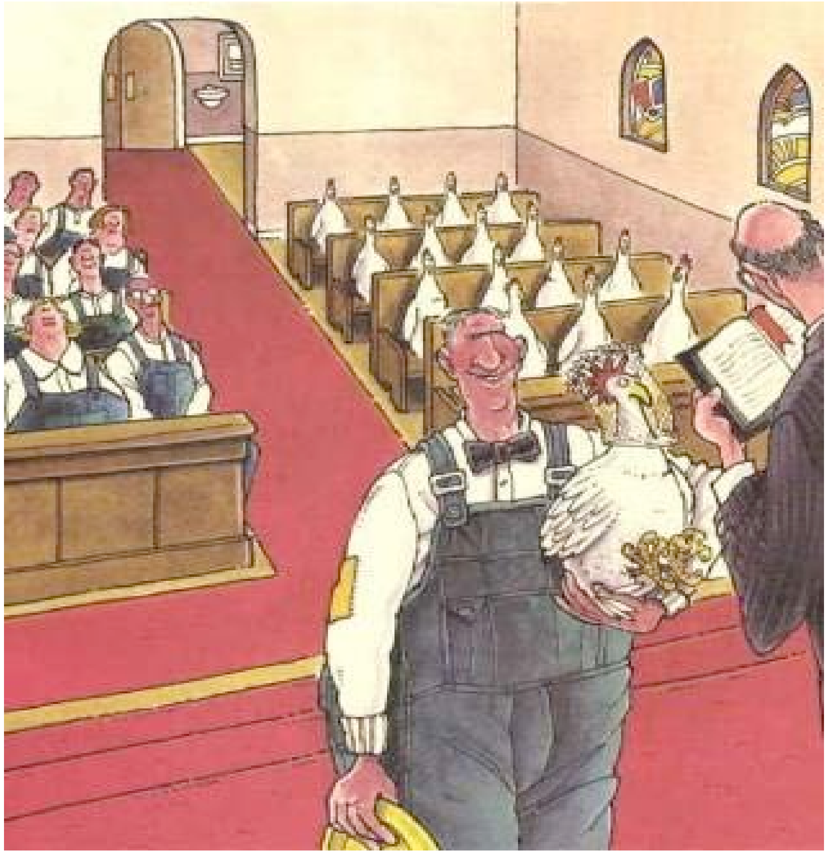
WEIZMANN INSTITUTE OF SCIENCE מכון ויצמן למדע

BBCU

Israeli National Node

What is bioinformatics?

A marriage between Biology and Computers!



What is bioinformatics?

- Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline.

- Bioinformatics is the science of managing and analyzing biological data using advanced computing techniques.

- Bioinformatics ultimate goal, (as is described by an expert), is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

Computers & Bioinformatics

- **Bioinformatics** is the computer-assisted data management discipline that helps us:

  Gather, store, analyze, integrate biological and genetic information (data), and represent this information efficiently.

- Bioinformatics experts claim that

  "Bioinformatics is the electronic infrastructure of molecular biology".

What is done in bioinformatics?

- Analysis and interpretation of various types of biological data including: nucleotide and amino acid sequences, protein domains, and protein structures.

- Development of new algorithms and statistics with which to assess biological information, such as relationships among members of large data sets.

What is done in bioinformatics?

- Development and implementation of tools that enable efficient access and management of different types of information, such as various databases, integrated mapping information.

Exponential Growth of Data Matched by Development of Computer Technology

- CPU vs Diskspace & 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 and needs special attention as well.

What “units of information” do we deal with in bioinformatics?

- DNA
- RNA
- Protein

- Sequence
- Structure
- Evolution

- Pathways
- Interactions
- Mutations

Examples of biological data used in bioinformatics

- DNA (Genome)
- RNA (Transciptome)
- Protein (Proteome)

DNA

- Simple Sequence Analysis
  - Database searching
  - Pairwise analysis…
- Regulatory Regions
- Gene Finding
- Whole Genome Annotations
- Comparative Genomics (Analyses between Species and Strains)

DNA

Raw DNA Sequence

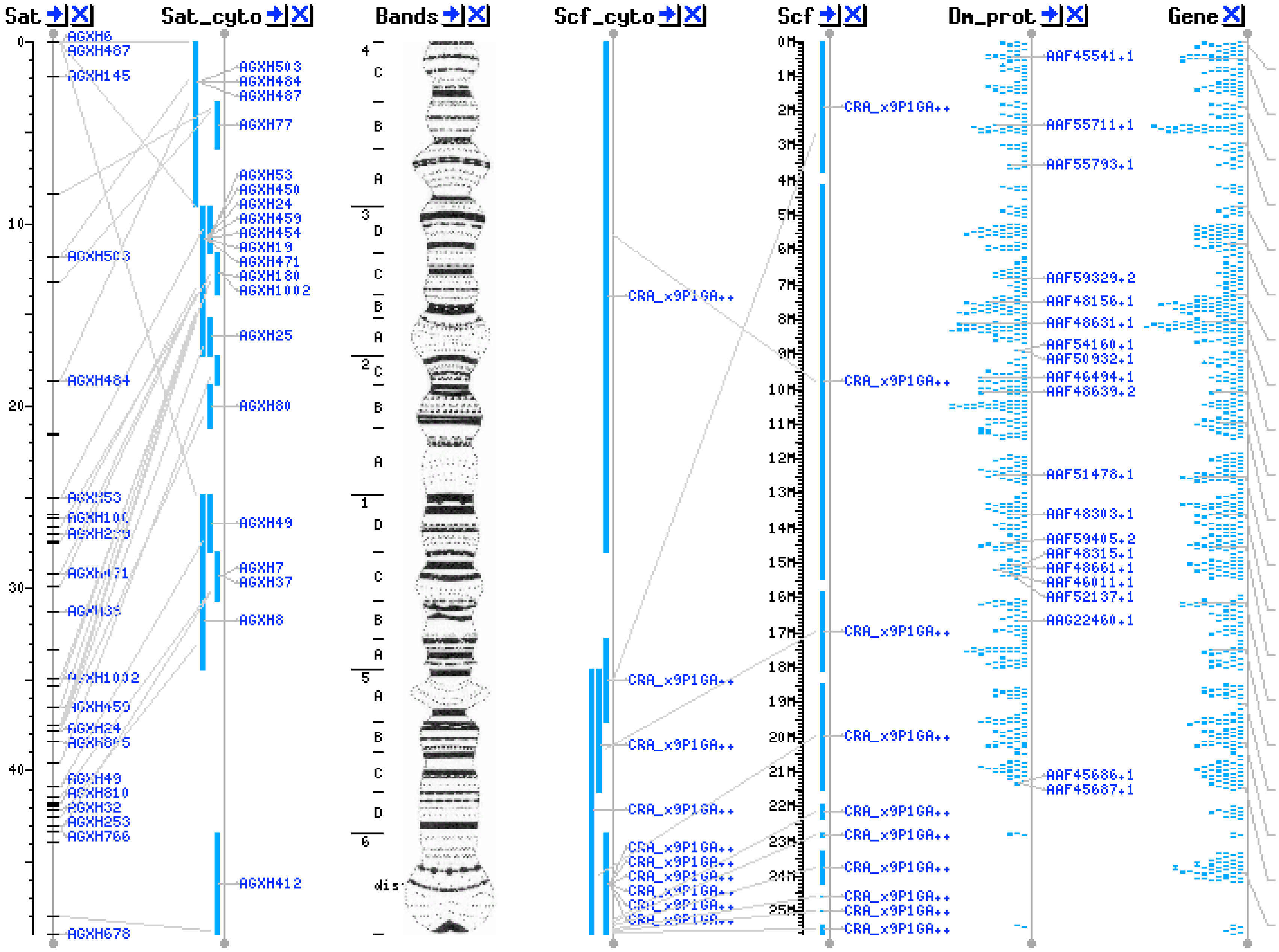
- Coding or Not coding?
- Parse into genes?
- 4 bases: AGCT

```
atggcaattaaaattggtatcaatggt
tttggtcgtatcggccgtatcgtattc
cgtgcagcacaacaccgtgatgacatt
gaagttgtaggtattaacgacttaatc
gacgttgaatacatggcttatatgttg
aaatatgattcaactcacggtcgtttc
gacggcactgttgaagtgaaagatggt
aacttagtggttaatggtaaaactatc
cgtgtaactgcagaacgtgatcca
```

DNA

- Complete genome annotation

- Comparative Genomics (Analyses between Species and Strains)


Sat
Sat_cyto
Bands
Scf_cyto
Scf
Dm_prot
Gene
0
10
20
30
40
AGXH6
AGXH487
AGXH145
AGXH503
AGXH484
AGXH53
AGXH100
AGXH239
AGXH471
AGXH35
AGXH1002
AGXH459
AGXH24
AGXH805
AGXH49
AGXH810
AGXH32
AGXH253
AGXH766
AGXH678
AGXH503
AGXH484
AGXH487
AGXH77
AGXH53
AGXH450
AGXH24
AGXH459
AGXH454
AGXH19
AGXH471
AGXH180
AGXH1002
AGXH25
AGXH80
AGXH49
AGXH7
AGXH37
AGXH8
AGXH412
4
C
B
A
3
D
C
B
A
2
C
B
A
1
D
C
B
A
5
A
B
C
D
6
dis
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
0M
1M
2M
3M
4M
5M
6M
7M
8M
9M
10M
11M
12M
13M
14M
15M
16M
17M
18M
19M
20M
21M
22M
23M
24M
25M
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
CRA_x9P1GA++
AAF45541.1
AAF55711.1
AAF55793.1
AAF59329.2
AAF48156.1
AAF48631.1
AAF54160.1
AAF50932.1
AAF46494.1
AAF48639.2
AAF51478.1
AAF48303.1
AAF59405.2
AAF48315.1
AAF48661.1
AAF46011.1
AAF52137.1
AAG22460.1
AAF45686.1
AAF45687.1

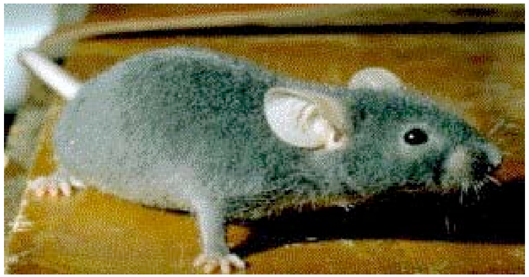
Whole Genomes

Drosophila

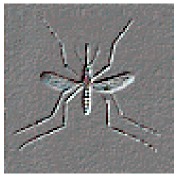
C. elegans

Rat

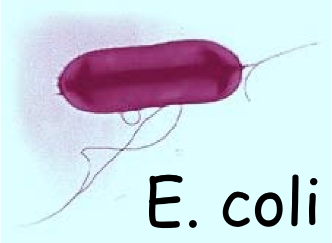
Human

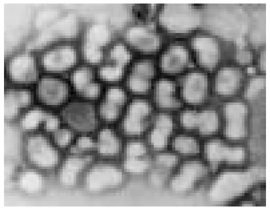

Mouse

Rice


Mosquito

Yeast


E. coli


H. influenza

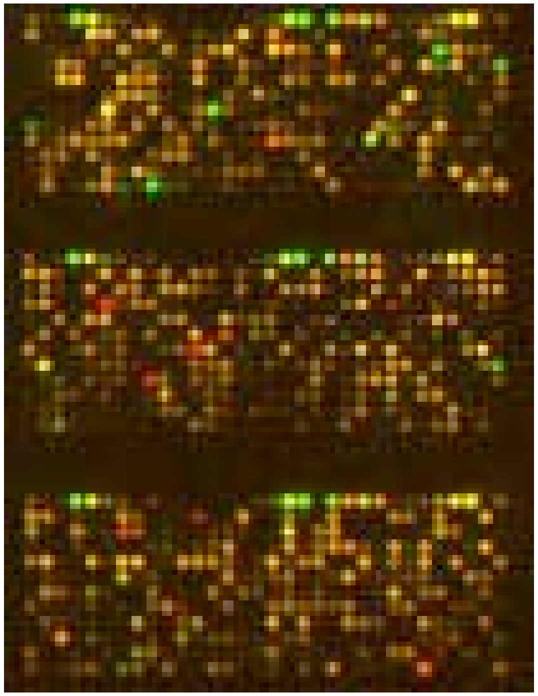
Arabidopsis

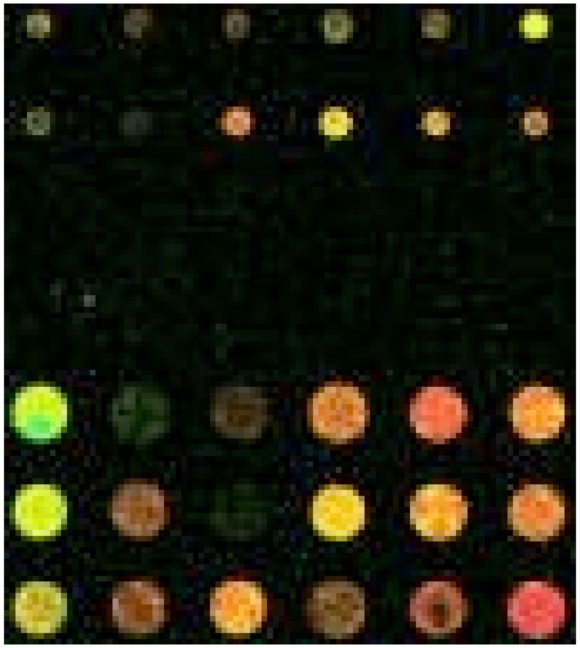
RNA

- Splice Variants
- Tissue specific expression
- Structure
- Single gene analysis (various cloning techniques…)
- Experimental data involving thousands of genes simultaneously
- DNA Chips, MicroArray, and Expression Array Analyses

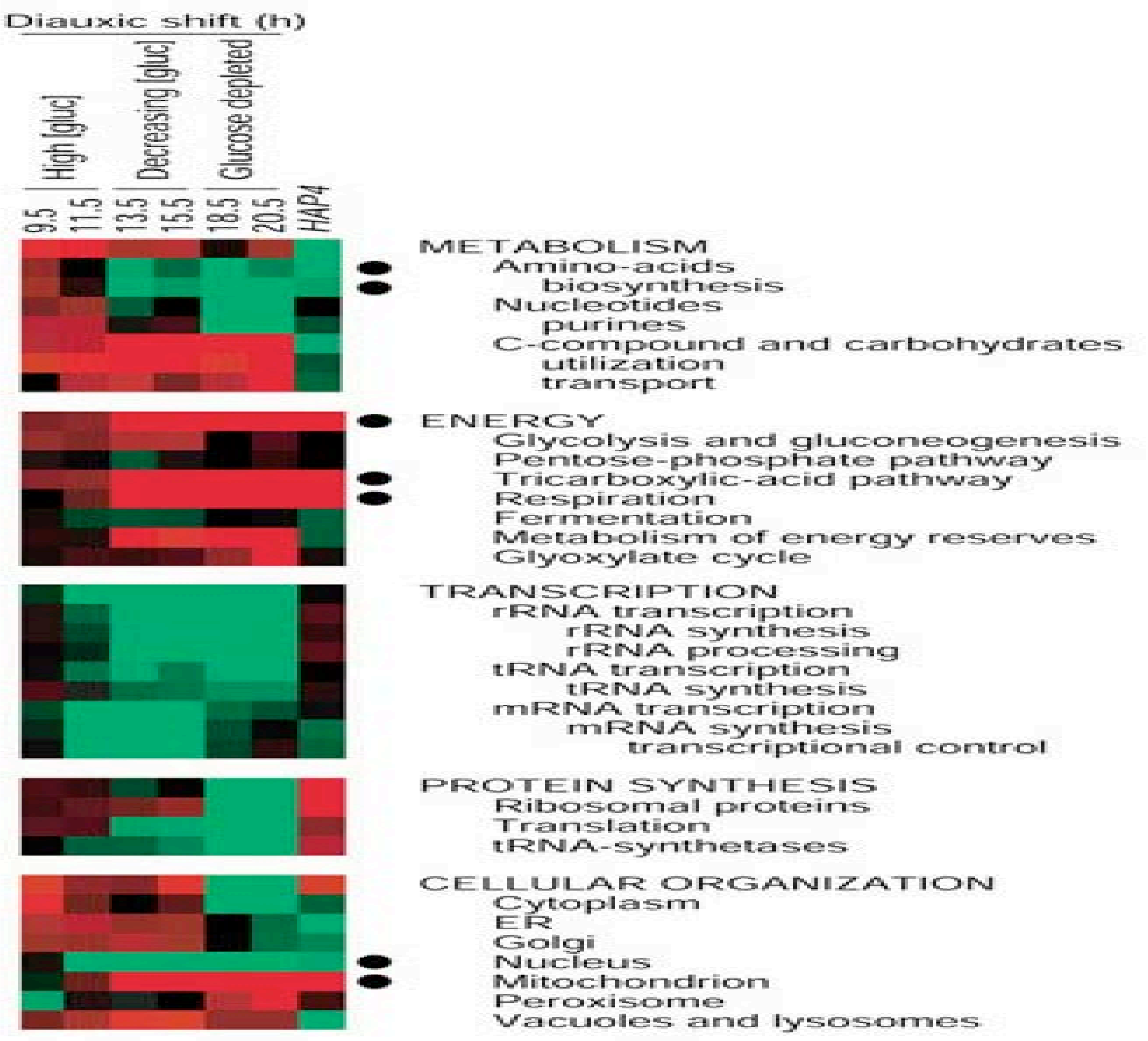
Examples of biological data for bioinformatics

- DNA Chips, MicroArray, and Gene Expression Data to be Analyzed.





Expression data

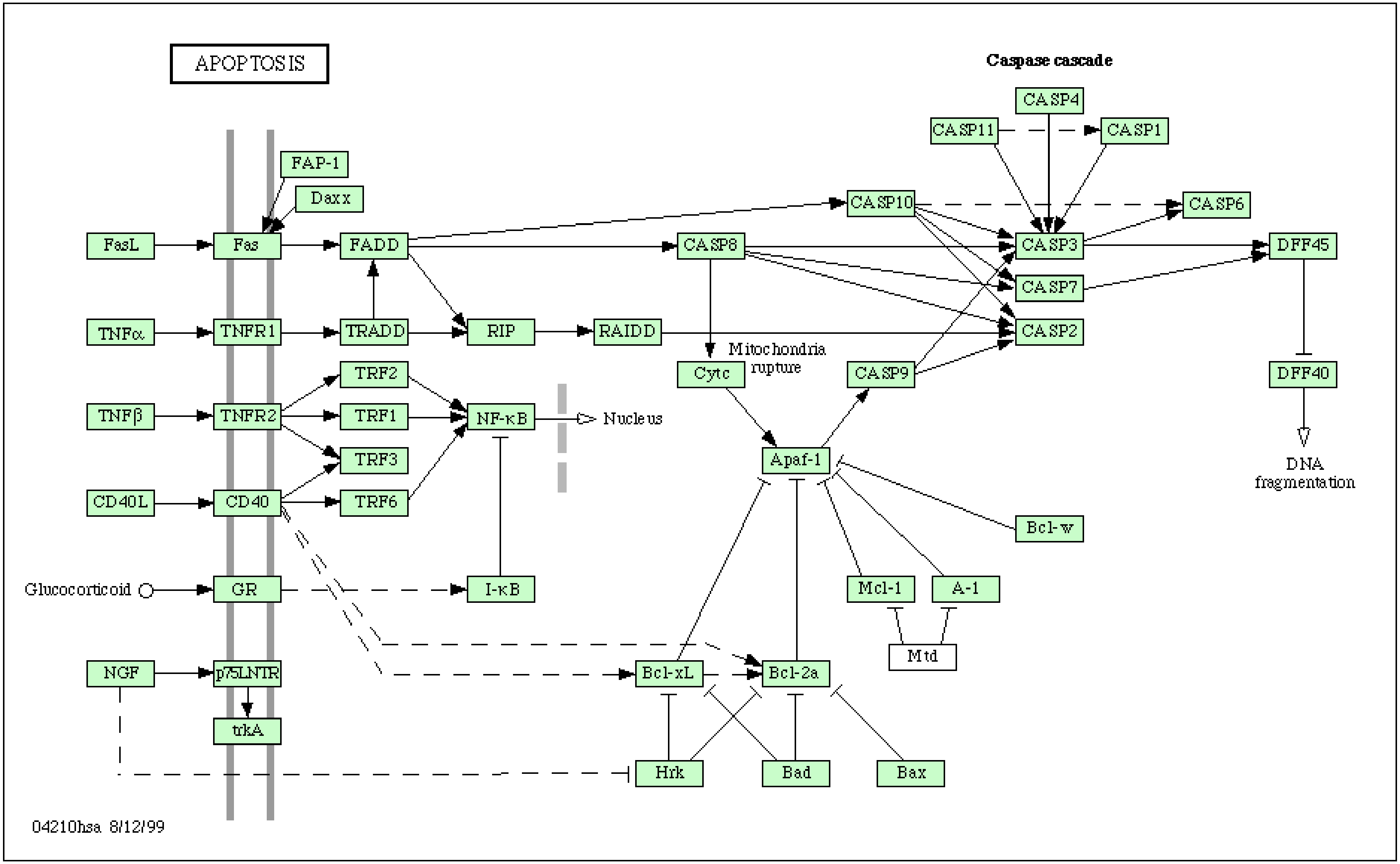


Where Bioinformatics is being used

Information to understand systems biology:

- Metabolic Pathways
- Regulatory Networks

Pathway of Apoptosis in Homo sapiens



Protein

- Proteome of an Organism
- 2D gels
- Mass Spec
- 2D Structure
- 3D Structure

Protein

- 20 letter alphabet

  ACDEFGHIKLMNPQRSTVWY

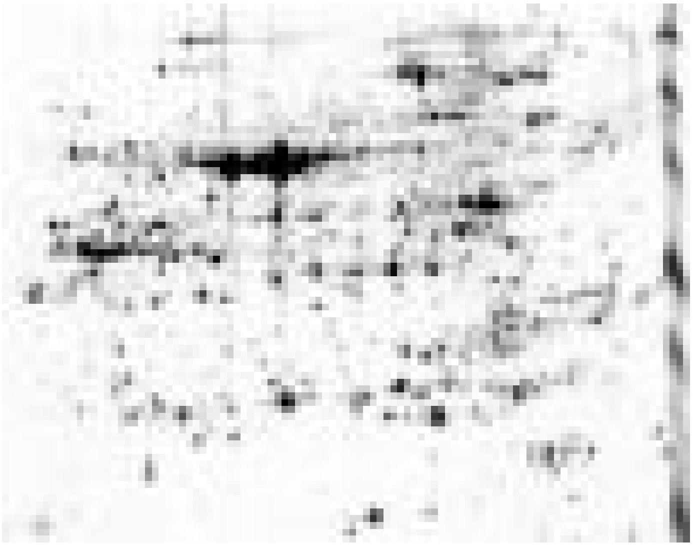
  But not BJOUXZ

- Strings of ~300 aa in an average protein

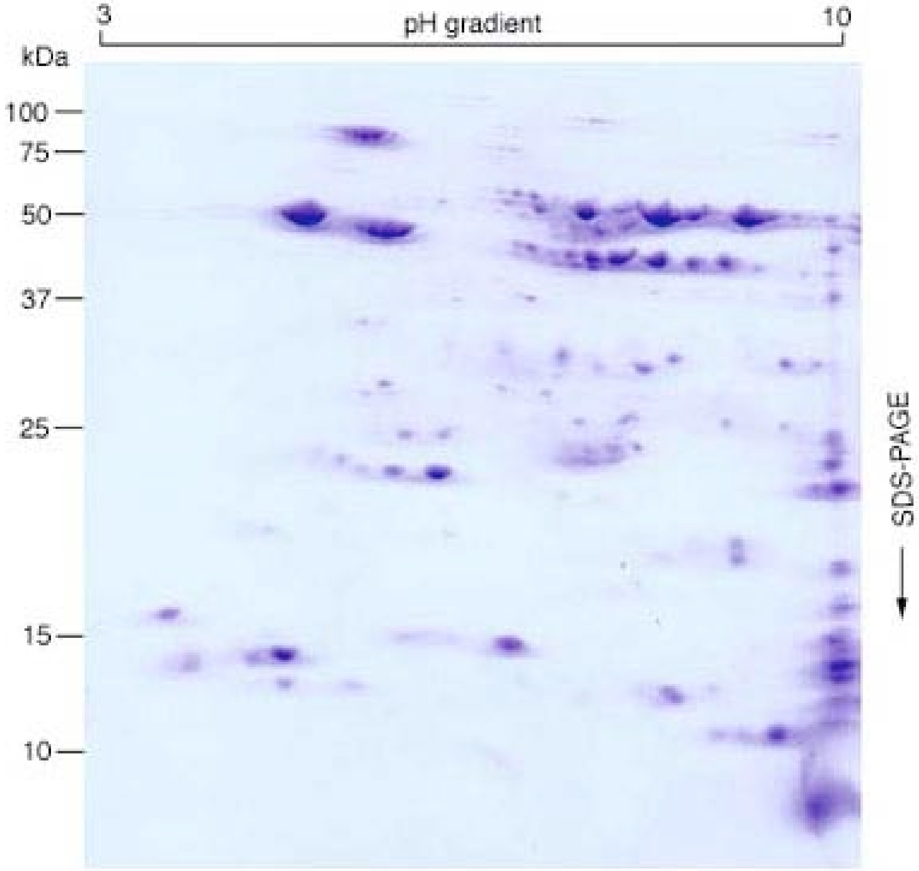
  (e.g. bacteria),

- protein are divided into domains

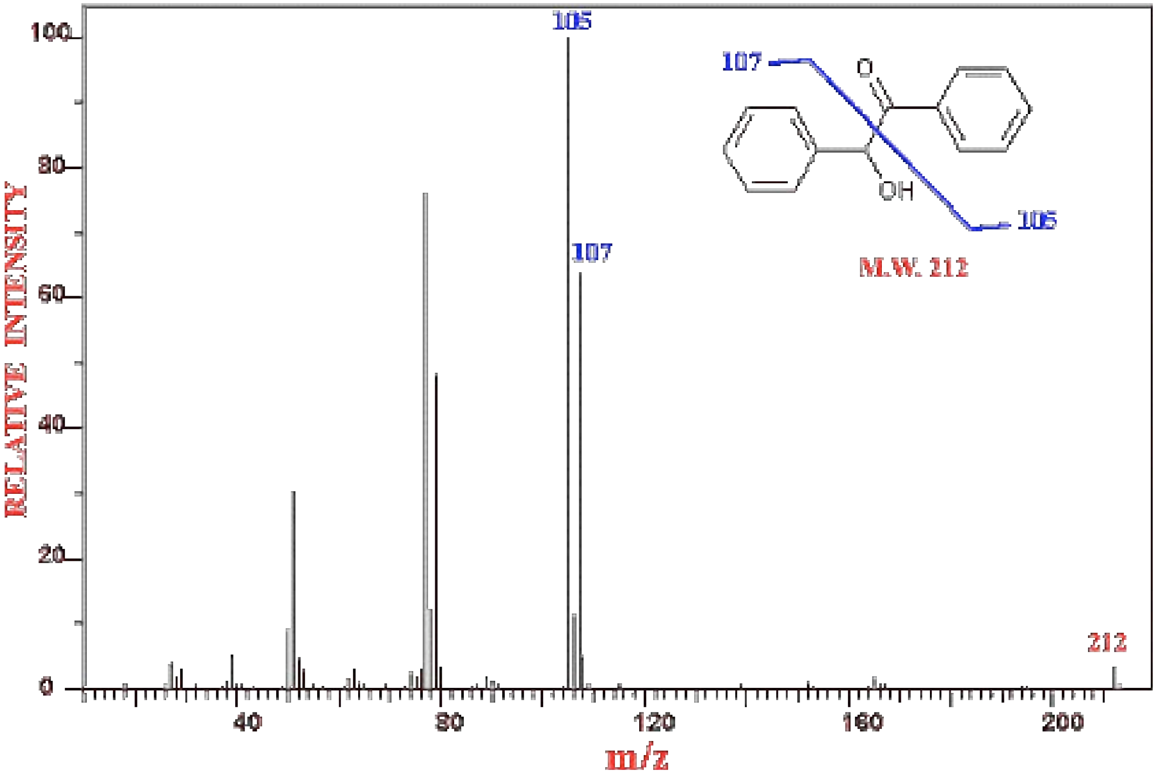
```
LNCIVAVSQNMGIGKNGDLPW
PPLRNEFRYFQRMTTTSSVEG
KQNLVIMGKKTWFSILNSIVA
VCQNMGIGKDGNLPWPPLRNE
YKYFQRMTSTSHVEGKQNAVI
MGKKTWFSIISLIAALAVDRV
IGMENAMPWNLPADLAWFKRN
TLDKPVIMGRHTWESITAFLW
AQDRNGLIGKDGHLPWHLPDD
LHYFRAQTVGKIMVVGRRTYE
SF
```



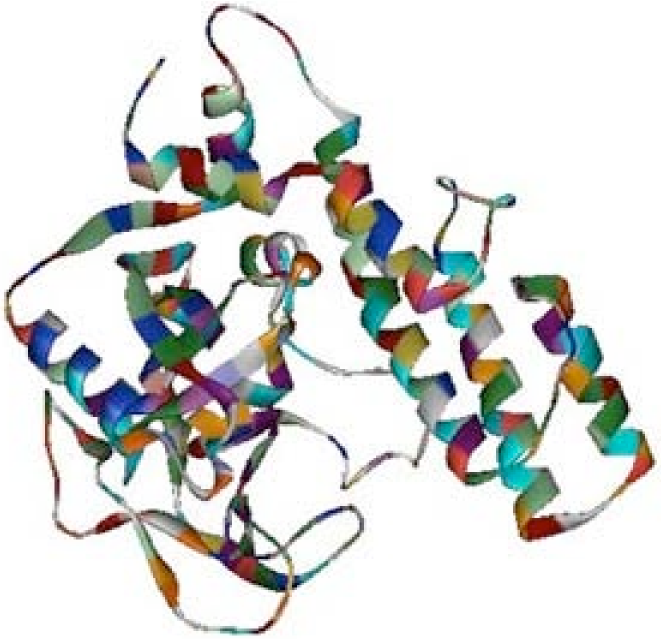
2D Gel Electrophoresis



Mass Spec data



PROTEIN 3D Structure



Molecular Biology Information

Other Integrative Data

¥Metabolic Pathways

traditional biochemistry

¥Regulatory Networks

¥Whole Organisms Phylogeny

¥Environments, Habitats, ecology

¥The Literature (MEDLINE)

Molecular Biology Information
Redundancy and Multiplicity

- Different sequences have the same structure.
- One organism has many similar genes
- A single gene may have multiple functions
- Redundancy due to the genetic code

All this information is applied to…

- Medical applications:
- Understand life processes in healthy and disease states.
- Genetic Disease (SNPs)

- Pharmaceutical and Biotech Industry
- To find (develop) new and better drugs.
- Gene-based or Structure-based Drug Design

- Agricultural applications
- Disease, Drought Resistant Plants
- Higher Yield Crops

Why use bioinformatics?

- ✓ The explosive growth in the amount of biological information necessitates the use of computers for cataloging and retrieval of this information.

- ✓ A more global perspective in experimental design. As we move from the one scientist-one gene/protein/disease paradigm of the past, to a consideration of whole organisms, we gain opportunities for new, more general insights into health and disease.

Why use bioinformatics?

- Data-mining - the process by which testable hypotheses are generated regarding the function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms.
  - For example, new insight into the molecular basis of a disease may come from investigating the function of homolouge of the disease gene in other organisms.
- Equally exciting is the potential for uncovering phylogenetic relationships and evolutionary patterns.

Biological problems that computers can help with:

- I cloned a gene - is it a known gene?
- Does the sequence match? Is the sequence any good?
- Does it look like anything else in the database?
- Which family does it belong to?
- How can I find more family members?
- I have an orphan receptor, how can I find its ligand?
- The gene I'm interested in was found in another organism, but not mine. How can I look for it?
- I have linkage to a specific region on chromosome x, how do I find genes in that region?

- My advisor wants me to construct a chimeric gene - how do I plan primers? How do I check to know that I have the right sequence?
- I have an RNA sequence with poor expression and I'd like to know its structure.
- I have a protein sequence, how can I find out what it's structure and/or function is?
- How can I cluster protein sequences into families of related sequences and develop protein models?
- I'd like to align similar proteins (or DNA) and generate phylogenetic trees.
- How can I find out which other proteins my sequence interacts with?

What will we cover in this course?

- Introduction to databases
- Working with sequences
  - Issues in plasmid and primer design
  - Sequencing DNA
  - Translation to protein
- Pairwise comparison
- Database similarity searching
- Multiple alignment

What will we cover in this course?

- Protein 2D structure, topology
- Introduction to Phylogenetic Analysis

- Genomics
  - How were genomes sequenced?
  - What benefits can we get from the sequence?

- Introduction to high throughput analysis

What won't we cover in this course?

- Detailed structural analysis of proteins
- Algorithm Development
- In-depth chip analysis methods
- In-depth phylogenetics or evolutionary biology
- In-depth systems biology
- Promoter Analysis
- Graphics programs, word processing, endnote, electronic journals