BioInformatics and Computational Molecular Biology

Course Website

http://BioInformatics.uchc.edu

What is Bioinformatics

Bioinformatics upgrades the information content of biological measurements.

“Discovery”

WHAT KIND OF MEASUREMENTS?

- SEQUENCE
- PHYSIOLOGIC
- EXPRESSION
- STRUCTURE

Where do these data come from?

Computational Molecular Biology

- Strings
  - nucleotides
  - amino acids
  - Sequence reads
- Gene expression
  - Large data arrays

Kinds of problems we will study: Homology/Alignment

How similar is the *catfish* Somatostasin Precursor

mpstriqcalallavalsvcsvsgapsdaklrqflqrsilapsvkqeltryttlaellaelaqaenevld
sdevsraaesegarlemeraagpmlaprerkagcknffwktftsc

to the *human* Somatostatin Precursor

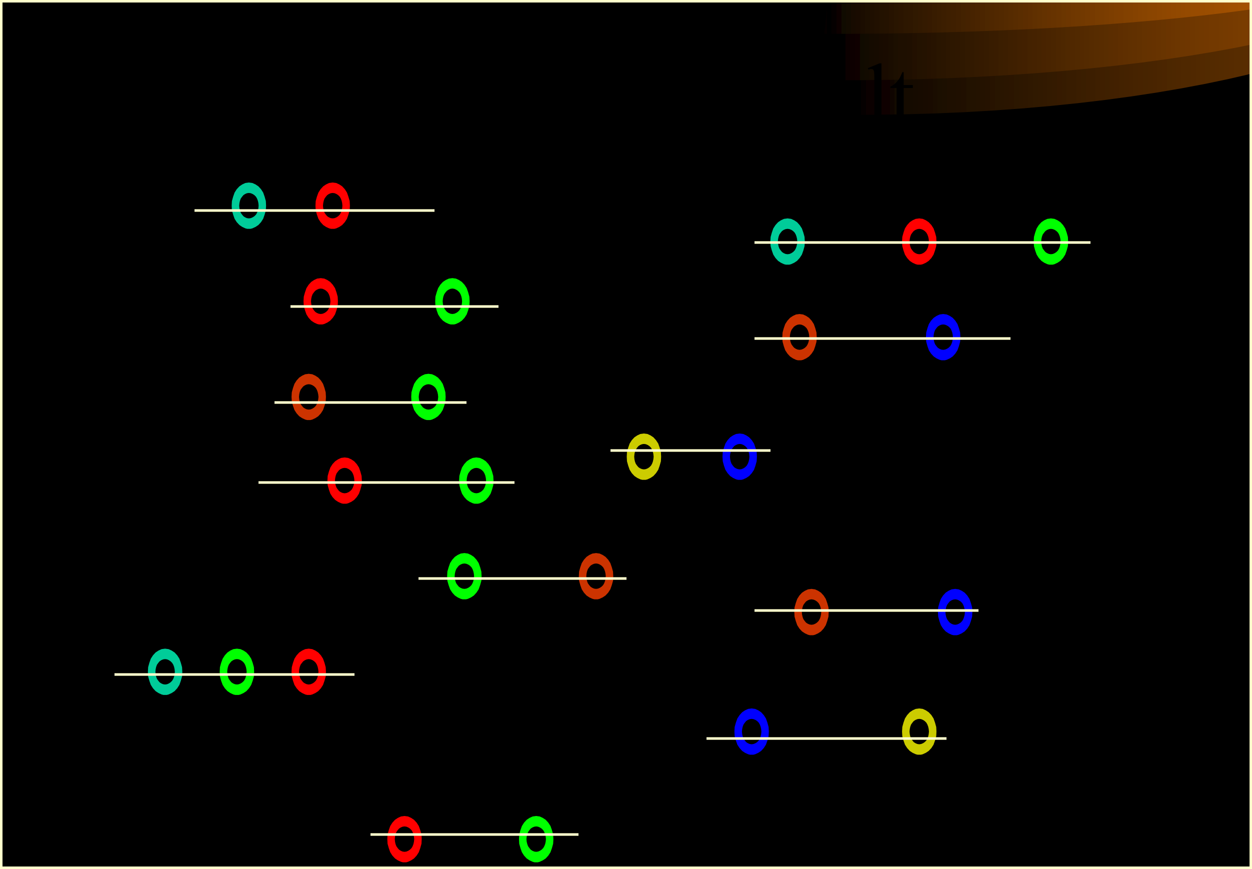
mlscrlqcalaalcivlalggvtgapsdprlrqflqkslaatgkqelakyflaellsepnqtendalepedlpqaaeqde
mrlelqrsansnpamaprerkagcknffwktftsc

Kinds of problems we will study: Phylogeny

Given corresponding regions of DNA in the gene causing cell wall permeability, can we determine which of the bacteria is most likely the ancestor and what the lineage is?

Bacterium A	ACGTAC
Bacterium B	ACTTTG
Bacterium C	ACTATG
Bacterium D	AGGTAG

Kinds of problems we will study: Constructing a genome *de novo* Mapping/Fragment Assembly



Kinds of problems we will study: Gene-finding

- Given this sequence of bases, can we find a substring which is a gene?

CTG TTA GTA GCC CAT TTC CAA ATA ATA ATA ATA ATA ATA ATA ATA

ATG CCT GAC ACC ATG CTG CCC GCC TGC TTC CCT GGC CTA CTG GCC
TTC TCC TCC GCG TGC TAC TTC CAG AAC TGC CCG AGG GGC GGC
AAG AGG GCC ATG TCC GAC CTG GAG CTG AGA CAG TGC CTC CCC
TGC GGC CCC GGG GGC AAA GGC CGC TGC TTC GGG CCC AGC ATC
TGC TGC GCG GAC GAG CTG GGC TGC TTC GTG GGC ACG GCT GAG
GCG CTG CGC TGC CAG GAG GAG AAC TAC CTG CCG TCG CCC TGC
CAG TCC GGC CAG AAG GCG TGC GGG AGC GGG GGC CGC TGC GCC
GCC TTC GGC GTT TGC TGC AAC GAC GAG AGC TGC GTG ACC GAG
CCC GAG TGC CGC GAG GGC TTT CAC CGC CGC GCC CGC GCC AGC
GAC CGG AGC AAC GCC ACG CAG CTG GAC GGG CCG GCC GGG GCC
TTG CTG CTG CGG CTG GTG CAG CTG GCC GGG GCG CCC GAG CCC
TTC GAG CCC GCC CAG CCC GAC GCC TAC TGA

TAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA

Kinds of problems we will study: Gene-expression in large data arrays

An example from the M.I.T. Lab of
G. Stephanopoulos

7000 gene expression levels provided among 24 different tissues-some normal, some diseased. No labels for the genes.

The task: Discover ways that the gene expression levels (numbers) can be viewed so as to give insight into their function or tissue of origin.

STRATEGY

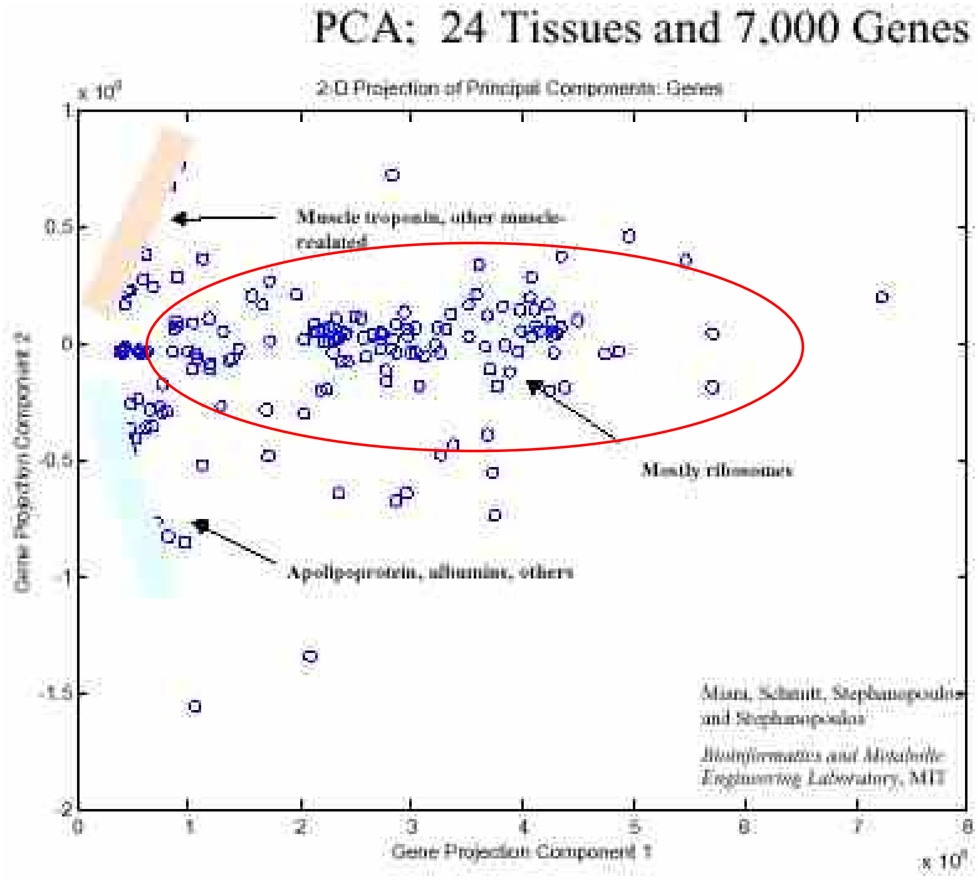
1. Arrange/transform the data into ‘natural’ clusters

2. Reduce the *dimension* of the data by identifying and reducing *correlated data*

2. Reveal the labels to see if the clustering made sense

Principal Component Analysis

By Function



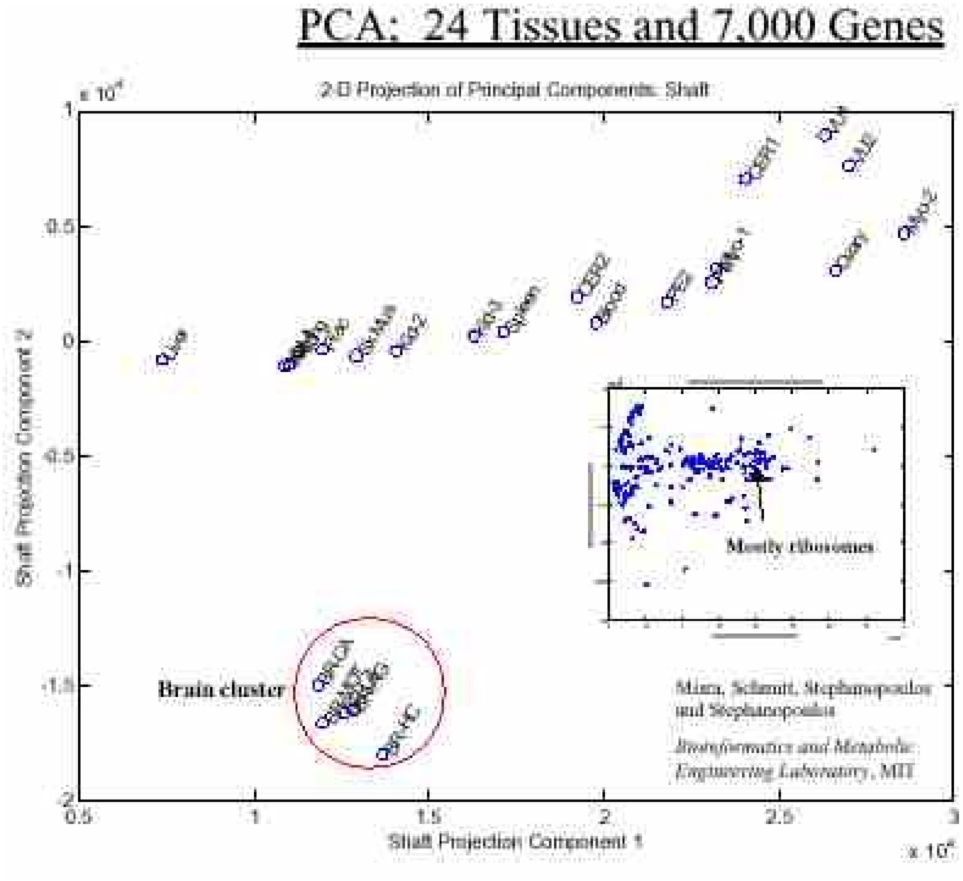
PCA: 24 Tissues and 7,000 Genes

Misra, Schmitt, Stephanopoulos and Stephanopoulos

Bioinformatics and Metabolic Engineering Laboratory, MIT

Principal Component Analysis

BY Tissue



PCA: 24 Tissues and 7,000 Genes

Misra, Schmitt, Stephanopoulos and Stephanopoulos

Bioinformatics and Metabolic Engineering Laboratory, MIT

ANOTHER EXAMPLE

Golub,TR, Slonim, DK, Tamayo, P *et al.*

SCIENCE 286:531-37,1999

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

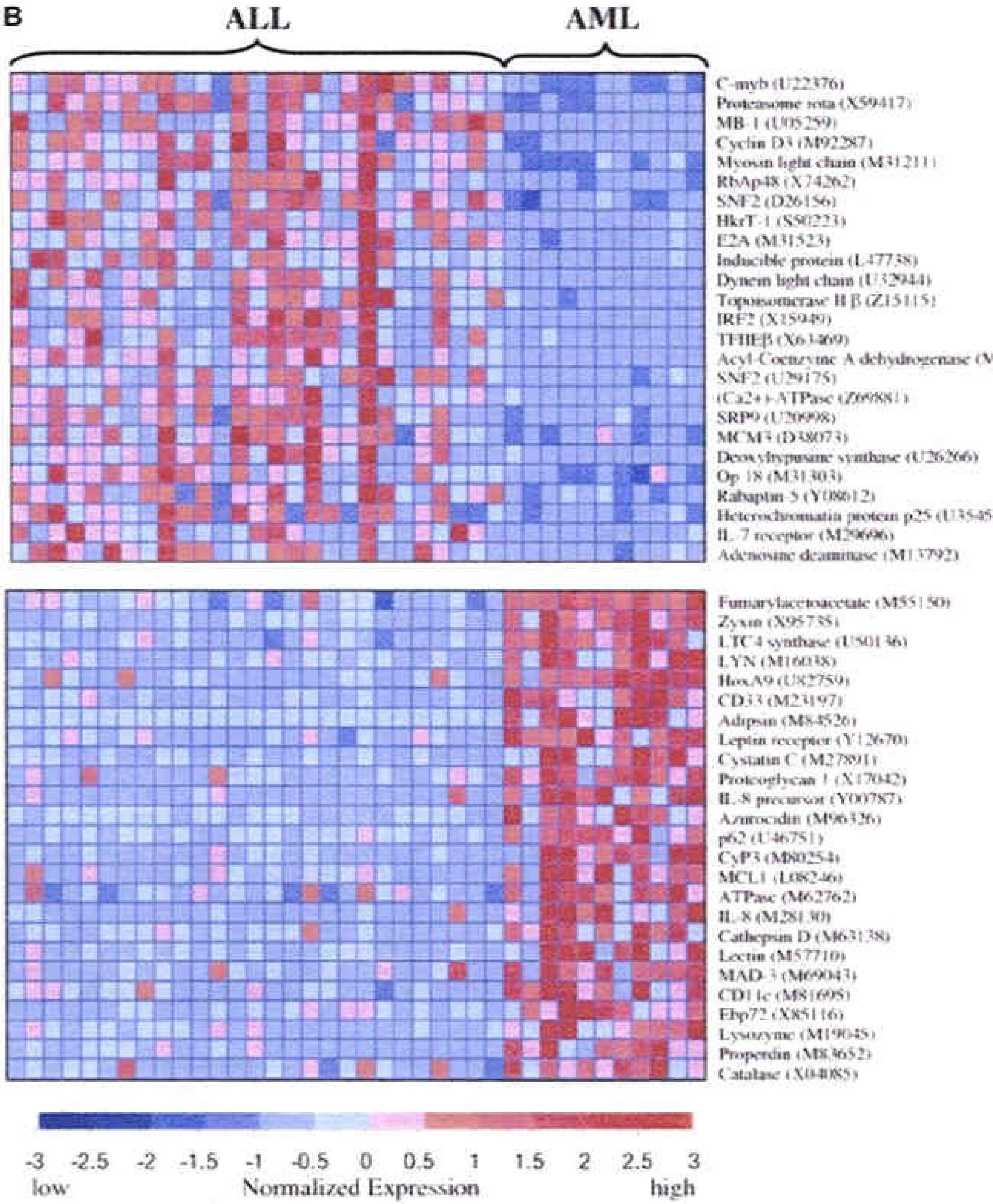
Golub,TR, Slonim, DK, Tamayo, P *et al.* SCIENCE 286:531-37,1999

- 38 Marrow Samples
  - 11 Acute Myelocytic Leukemia
  - 27 Acute Lymphocytic Leukemia
- Probes (cDNA’s) for 6817 different Human Genes were arranged in a Gene Chip (microarray). The marrow sample caused each cDNA to react with a measurable level (expression level)
- Gene expression levels for each gene were measured in each tissue sample

Golub,TR, Slonim, DK, Tamayo, P *et al.* SCIENCE 286:531-37,1999

- The preliminary technique, neighborhood analysis, implicated 1100 genes as related to the two tissue types, AML and ALL
- 50 ‘Best’ were used
- Given a 2 class model, the 50 genes classified 36/38 samples accurately. 2 were ‘uncertain’

Golub,TR, Slonim, DK, Tamayo, P *et al.* SCIENCE 286:531-37,1999



Golub,TR, Slonim, DK, Tamayo, P *et al.* SCIENCE 286:531-37,1999

- Self-organizing map applied to the data
  - 2 classes found -slightly less accurate than Neighborhood Analysis
- SOM applied using 4 class model. SOM found 3 classes accurately:
  - AML
  - B-cell ALL
  - T-cell ALL

So, what exactly is Bioinformatics?

For sure, it is the application of analytical tools to analyze data and set the foundation for discovery

- Pattern Recognition
- AI
- ANN
- Probabilistic Models
- Machine learning
- Optimization
- Dynamic Programming

Using computational methods, we are going to study…

- The Genome
  - How to assemble fragments
  - How to identify genes
- Genes
  - Relation to other genes in other species
  - How are they transcribed
    - RNA structure prediction
  - How do they respond to the environment
    - Gene expression

Using computational methods, we are going to study…

- Protein Structure
  - Identification
  - Folding
  - Comparison
    - Primary structure homology
    - Secondary structure: Motifs
    - Tertiary structure homology