Analysis & Visualization of large-scale genomic data

Olga Troyanskaya, Ph.D.

About the course

Instructor information

- Olga Troyanskaya
- Best way to contact is by e-mail: ogt@cs.princeton.edu, please put course number (597F) in subject line
- Office: 204 in 35 Olden Street

The course

- In **bioinformatics** – a field that brings together computer science and biology to study the flow of information in biological systems and in biological research
- This course will focus on **analysis of large-scale functional data**: gene expression, proteomics, data integration, data visualization

What this course is and is not

- A course on analysis of gene expression, proteomic, and other high-throughput functional biological data
- A course in applied computer science (with some statistics in the mix)
- Not an overview of bioinformatics – this is a depth-first course, although a brief intro to bioinformatics and biology will be provided (very soon)

Who should take this course

- Graduate or advanced undergraduate students from any department
- Interested in genomics, bioinformatics, or applied computer science
- Have some computational background
- Are interested in learning about genomics

Prerequisites

- SEAS students: ability to program a computer at CS 217 (intro to programming) level in a language of your choice
- Biology students: GENERAL understanding of computation and mathematical concepts on the level of SVD
- If in doubt, talk to me or e mail me- most likely there isn't a problem

Course format

- Lectures to introduce topics
- Student presentations of literature papers
- Discussion of presented papers in seminar format following the presentation
- Students will complete a team project during the duration of the course and write a paper on it

Grading

- Project – ~45%
- Presentations – ~35%
- Discussion of assigned reading (& attendance) – ~20%

Presentations

- Two 30-min presentations per class, plus 20 minutes discussion
- Each presentation is of 1 paper
  - Describe major points of the paper, including methods details and evaluation
  - Outline what you think are strong/weak points of the paper
  - Suggest what would improve the paper and what the future steps could be

Presentation (cont.)

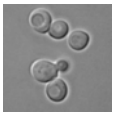
- Do:
  - Make you presentation accessible to everyone in the class by explaining methods (both computational and relevant experimental techniques)
  - Skip minor points, but do not just gloss over important method details or evaluation
- Do Not:
  - Go over time – 25 mins is good, 31 mins is bad
  - Be afraid to point out important points you are confused about even after you looked into them
- Presentations judged mainly on content, but delivery does matter

The project

- A team or individual project (up to 3 people/team)
- Involves designing, implementing and evaluating a novel bioinformatics method
  - Can be a known computational or statistical technique not yet applied to bioinformatics
  - Can be a novel visualization tool
  - I would be happy to provide ideas
- Project can be applicable to your research
- Biology students who cannot program can instead do a longer in depth review paper of methods in one area of informatics we covered (e.g. microarray image analysis), including ideas for novel methods and their necessary characteristics
- At the end of fall – submission of project/review writeups or project papers

Molecular biology 101
or
“why bother?”

Cells are fundamental working units of all organisms

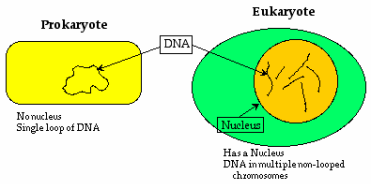


Yeast are unicellular organisms

Humans are multi- cellular organisms

Understanding **how a cell works** is critical to understanding how the organism functions

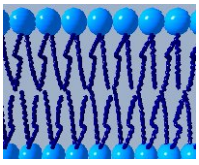
Prokaryotes vs. Eukaryotes



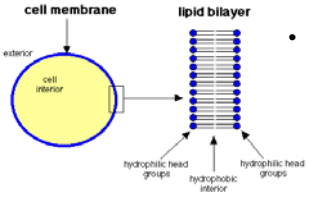
Yeast is a eukaryote just like humans. Fundamental biological processes are very similar.

Key biological macromolecules

- Lipids:
  - mostly structural function
  - Construct compartments that separate inside from outside
- DNA
  - Encodes hereditary information
- Proteins
  - Do most of the work in the cell
  - Form 3D structure and complexes critical for function

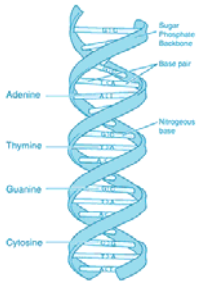


Lipids



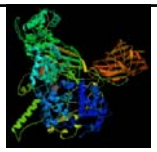
- Each lipid consists of a hydrophilic (water loving) and hydrophobic fragment
- Spontaneously form lipid bilayers => membranes

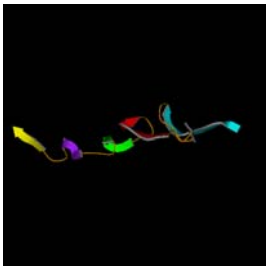
DNA



- Uses alphabet of 4 letters {ATCG}, called bases
- Encodes genetic information in triplet code
- Structure: a double helix

Proteins

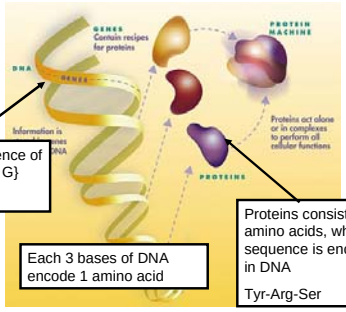




Courtesy of the Zhou Laboratory, The State University of New York at Buffalo

- A sequence of amino acids (alphabet of 20)
- Each amino acid encoded by 3 DNA bases
- Perform most of the actual work in the cell
- Fold into complex 3D structure

How does a cell function?



DNA is a sequence of bases {A, T, C, G}

TAT-CGT-AGT

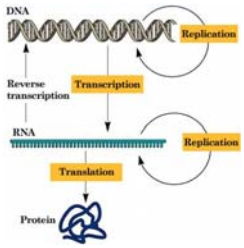
Each 3 bases of DNA encode 1 amino acid

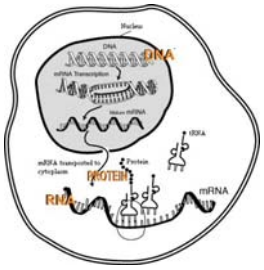
Proteins consist of amino acids, whose sequence is encoded in DNA

Tyr-Arg-Ser

Courtesy U.S. Department of Energy Genomes to Life program

The Central Dogma of biology





The "omes"

- Genome – organism's complete set of DNA
  - Relatively stable through an organism's lifetime
  - Size: from 600,000 to several billion bases
  - Gene is a basic unit of heredity (only 2% of the human genome)
- Proteome – organism's complete set of proteins
  - Dynamic – changes minute to minute
  - Proteins actually perform most cellular functions, they are encoded by genes (not a 1-to-1 relationship)
  - Protein function and structure form molecular basis for disease

Beyond the "omes" – systems biology

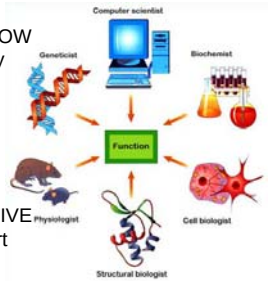
- Understanding the function and regulation of cellular machinery, as well as cell-to-cell communication on the molecular level
- Why? Because most important biological problems are fundamentally systems-level problems
  - Systems-level understanding of disease (e.g. cancer)
  - Molecular medicine
  - Gene therapy

Systems-level challenges

- **Gene function annotation – what does a gene do**
  - ~30,000 genes in the human genome => systems-level approaches necessary
  - A modern human microarray experiment produces ~500,000 data points => computational analysis & visualization necessary
  - Many high-throughput functional technologies => computational methods necessary to integrate the data
- **Biological networks – how do proteins interact**
  - Large amounts of high-throughput data => computation necessary to store and analyze it
  - Data has variable specificity => computational approaches necessary to separate reliable conclusions from random coincidences
- **Comparative genomics – comparing data between organisms**
  - Need to map concepts across organisms on a large scale => practically impossible to do by hand
  - High amount of variable quality data => computational methods needed for integration, visualization, and analysis
  - Data often distributed in databases across the globe, with variable schemas etc => data storage and consolidation methods needed

Function

- To study WHAT proteins DO, HOW they INTERACT, and HOW they are REGULATED, need data beyond genomic sequence
- Genomics/Bioinformatics is fundamentally a COLLABORATIVE and MULTIDISCIPLINARY effort

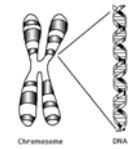


Gene expression – one type of high-throughput functional data

Why microarray analysis: the questions

- Large-scale study of biological processes
- What is going on in the cell at a certain point in time?
- On the large-scale genetic level, what accounts for differences between phenotypes?
- Sequence important, but genes have effect through expression

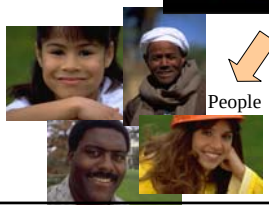
Why study gene expression



DNA

Gene Expression

Proteins



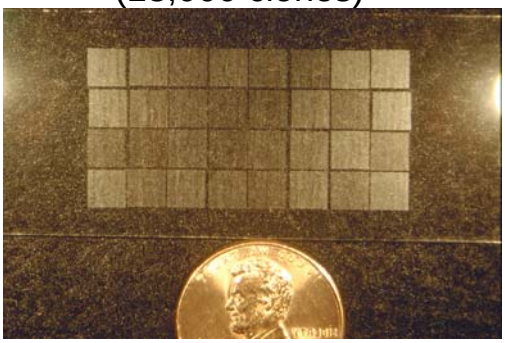
People

Microarray technology

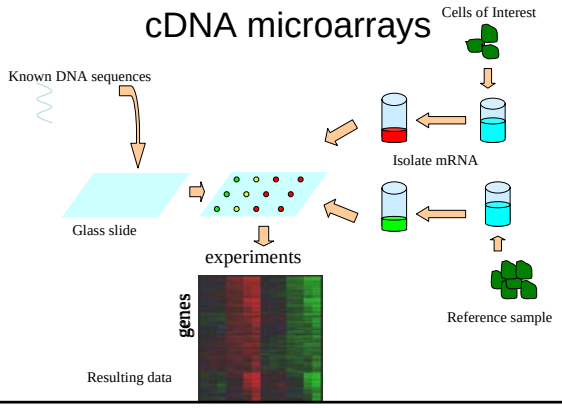
Microarray technologies

- Spotted cDNA arrays
  - Developed by Pat Brown (Stanford U)
  - Robotic microspotting
  - PCR products of full-length genes (>100nts)
- Affymetrix GeneChips
  - Photolithography (from computer industry)
  - Each gene represented by many n-mers
- Bubble jet / Ink jet arrays
  - Oligos (25-60 nts) built directly on arrays (in situ synthesis)
  - Highly uniform spots, very expensive

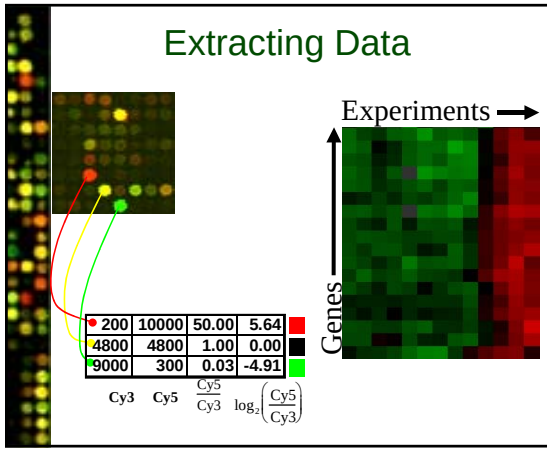
Early cDNA microarray (18,000 clones)



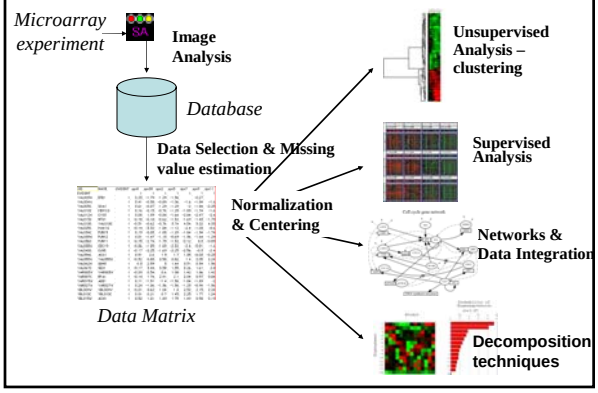
cDNA microarrays



Extracting Data



Microarray Data Flow



Experimental design of microarrays

What can microarrays tell us?

- What genes are involved in specific biological processes (e.g. stress response)
- Assumption = guilt by association (similar expression pattern => same pathway)
- Tumor classification for treatment guidance & outcome prediction

Types of experiments

- Time series vs.
- Comparison of groups of samples

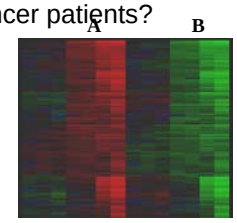
- Common reference vs.
- Using reference to compare

Time series

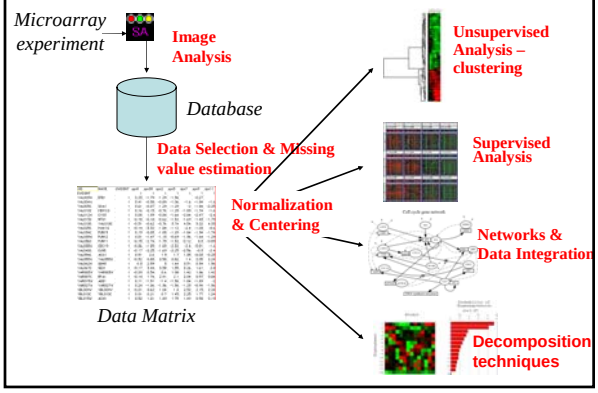
- Measurements taken throughout the time course
- Each array (column of the expression matrix) corresponds to a specific time point
- Can use common reference, or zero-time-point reference

Comparing groups of samples

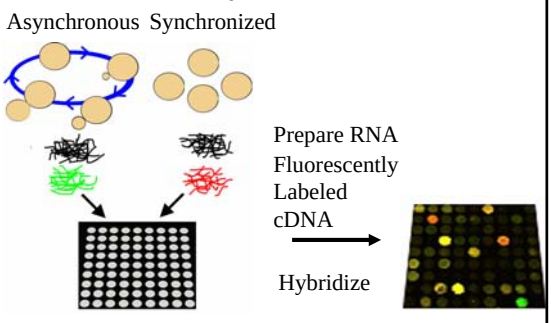
- Often in clinical studies – can we find similarities or differences within a group of lung cancer patients?



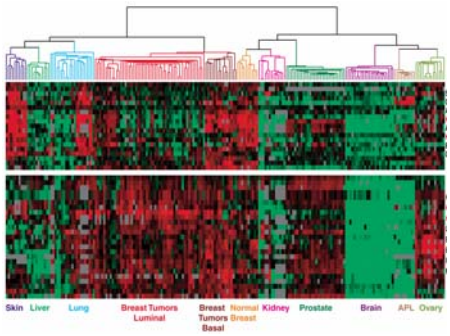
Issues in microarray analysis



Using reference for comparison



Common reference



Common reference problem

- Comparison of array experiments from different technologies (even labs) is difficult
- For spotted arrays, data is ratios of sample fluorescence (red) to reference fluorescence (green)
- To compare between experiments, need consistent reference
- “common reference” – a pool of reference mRNA from over 22 cell lines