Quiz Today

Covers:

- Concepts from lab 5
- Material from lecture 5
- Reading from:
  - Model Chap 7

Time:
Programming for Biologists and Biochemists

Lecture 6: High Throughput Data, Statistics, and Accessing Online Data

Brian Kidd

May 3, 2010
Overview

1. High-throughput Data
2. Statistics
3. Accessing Online Information
Overview

1. High-throughput Data
2. Statistics
3. Accessing Online Information
We are drowning in information and starving for knowledge.

–Rutherford D. Roger
Entering an Age of Exploration

21\textsuperscript{st} century science has shifted from being data-limited to hypothesis, and analysis limited.

“Tools are needed to support the research cycle—from data capture and data curation to data analysis and data visualization.”*

High-Throughput Data

Micro-arrays

DNA sequencers

SNP chips

ChIP-Seq
Microarrays: Overview

**Multiplexed** assay that **quantifies** the amounts of thousands of biomolecules present in a **sample**

- measures multiple molecules at same time
- functional view of biological activity within sample
- cells, tissues (blood, skin, muscle), organs

Techniques developed for examining:
DNA, RNA, proteins, lipids
Microarrays: Steps*

Obtain microarray chip containing probe molecules

Extract mRNA from cells in sample

Convert mRNA to cDNA**

Apply cDNA mixture to chip and allow to hybridize

Scan chip with multicolor laser

Quantify fluorescent intensity of colors

*http://www.bio.davidson.edu/Courses/genomics/chip/chip.html

**Fluorescently labeled targets
Introduction to DNA Microarrays
Motivation

Suppose you are studying a strain of heat resistant bacteria, and want to find the genes that are responsible for the heat resistance.

You want to find the genes the bacteria express at higher levels when exposed to heat.

You have two samples of bacteria, one exposed to heat and one not.

You use a green (no heat) and red (heat) spotted array to measure the expression levels of the genes in the bacteria’s genome.
Hybridizing Microarrays

Note that one can do one-color arrays as well.
One Chip = 1 Array of Data

Heat-exposed vs Non-heat-exposed bacteria

<table>
<thead>
<tr>
<th>Gene</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>-0.01</td>
</tr>
<tr>
<td>Gene 2</td>
<td>+0.30</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Gene 118</td>
<td>+4.00</td>
</tr>
<tr>
<td>Gene 194</td>
<td>-3.00</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Hybridizing Microarrays

Heat exposed vs Non-heat exposed

Multiplying vs Stationary

<table>
<thead>
<tr>
<th>Gene</th>
<th>Heat</th>
<th>Mult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>Gene 2</td>
<td>+0.30</td>
<td>+0.90</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 118</td>
<td>+4.00</td>
<td>+0.01</td>
</tr>
<tr>
<td>Gene 194</td>
<td>-3.00</td>
<td>-2.00</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DNA Microarrays

different experiments

gene is “on”
gene is “off”
Current Generation of μArrays

~30,000 human genes from ~48,000 probes
Microarray File

The data from a microarray run will be stored in a tab-delimited file – in Python this is ‘\t’

<table>
<thead>
<tr>
<th>Gene</th>
<th>13164</th>
<th>HMEC X NB2</th>
<th>This experiment was an out</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-0.24300</td>
<td>-0.02100</td>
<td>-0.20500</td>
</tr>
<tr>
<td>10000</td>
<td>-1.17800</td>
<td>-0.79000</td>
<td>0.06300</td>
</tr>
<tr>
<td>10016</td>
<td>-0.63500</td>
<td>-0.79200</td>
<td>-0.54500</td>
</tr>
<tr>
<td>10019</td>
<td>2.08300</td>
<td>-0.11400</td>
<td>-0.72600</td>
</tr>
<tr>
<td>10023</td>
<td>-0.02700</td>
<td>0.63800</td>
<td>-0.61000</td>
</tr>
<tr>
<td>10024</td>
<td>-0.73200</td>
<td>-0.85800</td>
<td>-0.83500</td>
</tr>
<tr>
<td>10026</td>
<td>-0.45400</td>
<td>-0.66900</td>
<td>0.60100</td>
</tr>
<tr>
<td>10036</td>
<td>-0.76000</td>
<td>0.48400</td>
<td>-0.89800</td>
</tr>
<tr>
<td>1004</td>
<td>-0.36000</td>
<td>-0.70400</td>
<td>0.14900</td>
</tr>
<tr>
<td>10046</td>
<td>0.12800</td>
<td>-0.53800</td>
<td>0.13400</td>
</tr>
</tbody>
</table>

~
# define input filename
fileName = 'arrayDataV1.txt'

# create empty list, represents rows
inputArray = []

# open file and go through each line, i.e. rows
fh = open(fileName)
for line in fh:
    # split line into columns and append to array
    arrayCols = line.strip().split('	')
    inputArray.append(arrayCols)
Loading Data Challenge

Load data that has:
- a header line
- accounts for row labels
- make sure data is float
454 and Polony Sequencing

(1) Sheering to small fragments

(2a) Linker attachment

(3a) Amplification on beads

(4a) Beads applied to picotiter plate

(5a) Pyrosequencing by primer extension

(2b) Circularization and linker attachment

(3b) Amplification on ePCR beads

(4b) Beads immobilized in a monolayer in an acrylamide matrix

(5b) Ligation with oligo pools added in cycles and four-color imaging

Sequencing Data

>seq1
TCACATCTCTACGTACTGAATTTAAAGGCTTTTTTGTCCTTTCTCGTTTTCTTTTCTGTTT
AATGATGTTTCAAGCGTAACCTCGGAAAAATGTGTGATAAATCGCCATATT
TACCCTTTTCTTAGCCAAATATTCCATGACACAAACCTAGCTGTAGGCCTTTGCTGCT
TTTAGCCAAAAACTTTGCCTATATTTTTTTATGCAAAAAATCGAGAAATGATGGTAAGACGTT
CGCGATTATATCTCTAATTGGTTTGCCGGTTAGTTGTTTACCCGGTTGCTTTCTTGGCTGTC

>seq2
TCAGGAGAATGCAGATGACAGCAGTAGCGCACCAAGTAACCCCTTTTCTAACGTCTTACG
AAGTTATGGCTCGTTACCACATTAGCTATACGACGCTCTCGGCAGAATAAAAGATGGA
GCTTGCCCGCAACCTCGTATCAACCAGAATACACGAAAACAAGCTGTGCGACATTGAAGACT
TGGAGGAGTATGAGGAATAGGAAGAGATACGCTGTACTTTTTCTTGGTGGACCTT
GAGCTAAGCAGCTTTTGAACGCGCCGCTTGTGCGCAACAAATATTAGTTATATATGGATACACCAACTT
AGGCTAAGGATAGCAGCAAGGATTTTTTTATGATTGGATGGATGAGTGGACTCTCT
CAAGAAACGCAACAGGTATTACAAATGCGTCGATAAAAA

>seq3
AAAAAAGGTTTACCCAAAAAGGCAGTCAAGGCGAGGGGGCGGTATTCTATTTGCTATAATAT
GCTGCAACAGCAACCTGGCTTACGAGAAGACAATCTGAGAATGCACGTATTACCATAGAC
TGGTCGGGTATTAAAACCCTGATCAAAAACTTTTGAGTGGGTGATAACGGATATGAGTAGT
ATTCTTTCAATGGAAGTGGGGACATCAGTATTTCTTACGGGCATTACGAGTGAAGAAATCT
CAAAAAACTCCCTAAATATGAAAAATATCATTTGAGAGGCTTTTCAAAATGAAATTGGTTGGAAG
CTAATAGTTTTTTCTAACCTAAAAATATCGGTAAATAGAGTCCGAGT
Overview

1. High-throughput Data
2. Statistics
3. Accessing Online Information
Statistics

What are they?

Summary stats

Hypothesis testing

Probability

Log-likelihood and risk ratios

*This is a large field of study, see the instructor if you have any interest in material that goes beyond what is presented in class.
Statistics Outline

Loose Definition:
Set of mathematical operations/calculations used to describe and interpret data

Where/How Used:
Designing of experiments and surveys
Modeling or predicting trends/outcomes using data

Types:
Descriptive – summarize data
Inferential – model or predict based on data
Describing Central Tendency

Mean – arithmetic

\[ \mu = \frac{1}{n} \sum_{i=0}^{n-1} x_i \]

arithmetic average of a set of values: “average”

Mean – geometric

\[ \bar{x} = \left( \prod_{i=0}^{n-1} x_i \right)^{1/n} \]

product average

Median

middle value of a data set
Describing Central Tendency

Mean – arithmetic

arithmetic average of a set of values: “average”

\[ \mu = \frac{1}{n} \sum_{i=0}^{n-1} x_i \]

How would you implement this in Python?

```python
X = [1, 2, 3, 4, 5]
# option A
total = 0
for i in x:
    total += i
mean = total / len(x)
# option B
total = 0
mean = sum(x) / len(x)
```
A number of mathematical formulas are summations – repetitive operations

Repetitive operations are perfect for loops

for loop for summation

\[ \sum_{i=0}^{n-1} X_i \]

\[ X = [1, 2, 3, 4, 5] \]

for loop for multiplication

\[ \prod_{i=0}^{n-1} X_i \]

\[ X = [1, 2, 3, 4, 5] \]
Median

The value of the middle number

To identify the median, you need to have a sorted list of numbers

Use the list method `sort` in Python

```python
x = [5, 4, 3, 2, 1]
x.sort()
mid = len(x)/2
median = x[mid]
```

Does this solve the problem completely?
Median

The value of the middle number

```python
x = [5, 4, 3, 2, 1]
x.sort()
mid = len(x)/2
if len(x) % 2 == 0:
    median = (x[mid] + x[mid-1]) / 2.
else:
    median = x[mid]
```

How would you turn this into a function?

*Depends on whether we have an odd or even number of elements.*
Statistics Challenge

\[ X = [6, 4, 8, 10, 2] \]

mean = 
median =

\[ Y = [1, 1, 9, 2, 2] \]

mean =
median =
Describing Spread of Data

Variance

$$\sigma^2 = \frac{1}{n-1} \sum_{i=0}^{n-1} (X_i - \mu)^2$$

Expected square of deviation from the mean

Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

Square root of the variance
T-statistic

Mean

\[ \mu = \frac{1}{n} \sum_{i=0}^{n-1} x_i \]

Variance

\[ \sigma^2 = \frac{1}{n-1} \sum_{i=0}^{n-1} (X_i - \mu)^2 \]

Standard Deviation

\[ \sigma = \sqrt{\sigma^2} \]

T-statistic (one-sample)

\[ t = \frac{\mu - \mu_0}{\sigma / \sqrt{n}} \]
Test for Mean Difference

T-statistic (one-sample)

\[ t = \frac{\mu - \mu_0}{\sigma / \sqrt{n}} \]

Test whether our measured mean (\(\mu\)) differs from the expected mean (\(\mu_0\)).

1. Calculate \(t\)
2. Check if \(|t| > \text{threshold}\)
Test for Mean Difference

T-statistic (one-sample)

\[ t = \frac{\mu - \mu_0}{\sigma / \sqrt{n}} \]

Example **Gene 1**:  
\( \mu = 0.5, \sigma = 1, \mu_0 = 0, n = 4 \)  
\[ t = \text{Yes/No} \]

Example **Gene 2**:  
\( \mu = 2, \sigma = 1, \mu_0 = 0, n = 4 \)  
\[ t = \text{Yes/No} \]
Test Between Success/Failure

Die throw or coin toss

\[ Pr = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \]

\( k \) successes in \( n \) trials with probability \( p \)

Example **Coin Toss**

\( p = 0.5, \ k = 3, \ n = 3 \quad Pr = 0.1250 \)

Example **Coin Toss 2**

\( p = 0.5, \ k = 10, \ n = 12 \quad Pr = 0.0193 \)
Statistics Module in Python

Scipy - scientific python
Numpy - numeric python

```
import scipy

scipy.mean()
s scipy.median()
s scipy.std()
s scipy.ttest_ind()
...
```
Overview

1. High-throughput Data
2. Statistics
3. Accessing Online Information
Accessing Online Information

- HTML pages
- Public Databases
  - microArrays
  - PDB
- PubMed
- Queries and downloads
The Rise of the Database

Information from the scientific literature, experimental results and computational analyses are now online in databases.

Cheaper storage space and faster computers have allowed databases to flourish (increase in number and size).

Requires computational tools to search through and extract meaningful information from these repositories.
Why are DBs Important?

Warehouses of information

Collect, curate, and catalogue the information

Provide single point of entry for accessing information

Provide “public” access to information
Rise in Indexed Articles

#Citations in Medline

*http://www.nlm.nih.gov/bsd/medline_lang_distr.html*
Rise in Biological Structures

#Structures Published in the PDB

Rise in Genetic Information

[Bar chart showing the increase in GenBank Entries over time, with a significant rise in the last decade.]

Online Genetic Databases

**NCBI** – National Center for Biotechnology Information*

GEO** – functional genomics data repository
GenBank – genetic sequence data bank
BLAST† – find similarity between sequences
OMIM‡ – human genes & genetic phenotypes
Pubmed – 19 million citations for biomedical articles

**EMBL** – European Molecular Biology Lab

**DDBJ** – DNA Data Bank of Japan **Gene Expression Omnibus
†Basic Local Alignment Search Tool
‡Online Medelian Inheritance in Man
*Division of the National Library of Medicine at the NIH
Online Protein Databases

**PDB*** – experimentally determined structures of biomolecules

**UnitProt**** – curated information about protein function, classification and cross-reference

**ExPASY**† – domains, families and functional sites

*Protein Data Bank
**Universal Protein Resource
†Expert Protein Analysis System
Strategy to Process Information

Obtain information from database

Extract information from file(s)

Parse extracted information

Process parsed information

Output processed information
Getting Info Online

Go direct from python

# tools for directly accessing urls
import urllib

# download a page and save it in a file. this will download
# the homepage of the new york times online.
urllib.urlretrieve("http://nytimes.com", "nytHomepage.html")

Web interface

From a colleague
Organize Info from Database

Single or multiple files?

```python
fh = open( "something.txt" )
for line in fh:
    do something

fh = open( "file_list.txt" )
```

Formats similar or different

- fasta, pdb, etc.
Process Parsed Information

Perform statistics or other mathematical operations

Important step for answering questions about information stored in database
Homework

Reading:

Scipy documentation on stats.py


Quiz:

Quiz 6 next Tues (5/10)
Covers readings, lectures, and lab

Practice problems (include with Lab 6 submission):
http://www.soe.ucsc.edu/classes/bme060/Spring11/examples/

   practiceProblemsLab06.py
   arrayDataInput.txt
   arrayDataHeaderRowlables.txt