Why do you need probability and statistics to analyze modern biological data?

**Reason 2:**
Life is random and messy
Life is messy, random, and noisy

Yet it is beautifully complex and has many parts
(see statistics)
Why life is so complex?

Primer on complex biomolecular networks
Intra-cellular Networks operate on multiple levels

Slides by Amitabh Sharma, PhD
Northeastern University & Dana Farber Cancer Institute
Sea urchin embryonic development (from endomesoderm up to 30 hours) by Davidson’s lab

May 29, 2007

Ubiq = ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; γ = γ-catenin source;
nβ-TCF = nuclearized β-κatenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Copyright © 2001–2007 Hamid Belouri and Eric Davidson
Protein-Protein binding
IntAct Database (Dec 2015)
Interactions: 577,297  Proteins: 89,716

Baker’s yeast S. cerevisiae (only nuclear proteins shown)
From S. Maslov, K. Sneppen, Science 2002

Worm C. elegans
From S. Lee et al, Science 2004
Metabolic pathway chart by ExPASy: 5702 reactions as of December 2015
Brain and nerves of a worm

- Worm (C. elegans) has 302 neurons
- Our brain has 100 billion \((10^{11})\) neurons
Credit: XKCD comics
Foundations of Probability

Random experiments

Sample spaces

Venn diagrams of random events
Random Experiments

• An **experiment** is an operation or procedure, carried out under controlled conditions
  – Example: measure the metabolic flux through a reaction catalyzed by the enzyme A

• An experiment that can result in **different outcomes**, even if repeated in the same manner every time, is called a **random experiment**
  – Cell-to-cell variability due to history/genome variants
  – Noise in external parameters such as temperature, nutrients, pH, etc.

• **Evolution** offers ready-made random experiments
  – Genomes of different species
  – Genomes of different individuals within a species
  – Individual cancer cells
Variability/Noise Produce Output Variation

- **Input**: What I want to change in the experiment, e.g., expression level of a gene A
- **Biological system: cell/organism/population**
- **Output**: What I measure in the experiment, e.g., metabolic flux catalyzed by the enzyme encoded by the gene A
- **Controlled variables**: e.g., Temperature, Nutrients, pH
- **Noise variables**: Internal state of individual cells, Signals from neighbors
Sample Spaces

• Random experiments have unique outcomes.
• The set of all possible outcomes of a random experiment is called the sample space, $S$.
• $S$ is discrete if it consists of a finite or countable infinite set of outcomes.
• $S$ is continuous if it contains an interval (either a finite or infinite width) of real numbers.
Examples of a Sample Space

• Experiment measuring the abundance of mRNA expressed from a single gene
  \[ S = \{ x \mid x > 0 \} \]: continuous.

• Bin it into four groups
  \[ S = \{ \text{below 10, 10-30, 30-100, above 100} \} \]: discrete.

• Is gene “on” (mRNA above 30)?
  \[ S = \{ \text{true, false} \} \]: logical/Boolean/discrete.
Event

An event \((E)\) is a **subset of the sample space** of a random experiment, i.e., **one or more** outcomes of the sample space.

- The **union** of two events is the event that consists of all outcomes that are contained in either of the two events. We denote the union as \(E_1 \cup E_2\).

- The **intersection** of two events is the event that consists of all outcomes that are contained in both of the two events. We denote the intersection as \(E_1 \cap E_2\).

- The **complement** of an event in a sample space is the set of outcomes in the sample space that are not in the event. We denote the complement of the event \(E\) as \(E'\) (sometimes \(E^c\) or \(\overline{E}\) ).
Examples

Discrete
1. Assume you toss a coin once. The sample space is $S = \{H, T\}$, where $H =$ head and $T =$ tail and the event of a head is $\{H\}$.

2. Assume you toss a coin twice. The sample space is $S = \{(H, H), (H, T), (T, H), (T, T)\}$, and the event of obtaining exactly one head is $\{(H, T), (T, H)\}$.

Continuous

Sample space for the expression level of a gene: $S = \{x \mid x \geq 0\}$

Two events:
- $E_1 = \{x \mid 10 < x < 100\}$
- $E_2 = \{x \mid 30 < x < 300\}$

- $E_1 \cap E_2 = \{x \mid 30 < x < 100\}$
- $E_1 \cup E_2 = \{x \mid 10 < x < 300\}$
- $E_1' = \{x \mid x \leq 10 \text{ or } x \geq 100\}$
Venn diagrams

Find 5 differences in beard and hairstyle

John Venn (1843-1923)  
British logician

John Venn (1990- )  
Brooklyn hipster
Venn diagrams

Which formula describes the blue region?

A. \( A \cup B \)
B. \( A \cap B \)
C. \( A' \)
D. \( B' \)

Get your i-clickers
Venn diagrams

Which formula describes the blue region?
A. $A \cup B$
B. $A \cap B$
C. $A'$
D. $B'$
Venn diagrams

Which formula describes the blue region?
A. \((A \cup B) \cap C\)
B. \((A \cap B) \cap C\)
C. \((A \cup B) \cup C\)
D. \((A \cap B) \cup C\)

Get your i-clickers
Which formula describes the blue region?

A. \((A \cup B) \cap C\)

B. \((A \cap B) \cap C\)

C. \((A \cup B) \cup C\)

D. \((A \cap B) \cup C\)
Which formula describes the blue region?

A. $A \cap C$
B. $A' \cup C'$
C. $(A \cap B \cap C)'$
D. $(A \cap B \cap C)$

Get your i-clickers
Venn diagrams

Which formula describes the blue region?

A. $A \cap C$

B. $A' \cup C'$

C. $(A \cap B \cap C)'$

D. $(A \cap B) \cap C$
Definitions of Probability
Two definitions of probability

• (1) **STATISTICAL PROBABILITY**: the relative frequency with which an event occurs in the long run

• (2) **INDUCTIVE PROBABILITY**: the degree of belief which it is reasonable to place in a proposition on given evidence

Bulmer, M. G.. Principles of Statistics (Dover Books on Mathematics)
Statistical Probability

A statistical probability of an event is the limiting value of the relative frequency with it occurs in a very large number of independent trials

Empirical
Statistical Probability of a Coin Toss

Excess of heads among 2,000 coin tosses (Kerrich 1946)

N(Heads out of T tosses) - N(Tails out of T tosses)

John Edmund Kerrich (1903–1985)
British/South African mathematician
Statistical Probability of a Coin Toss

Probability(Heads) = \frac{N(\text{Heads out of } T \text{ tosses})}{T}

limit for large $T$

John Edmund Kerrich
(1903–1985)
British/South African mathematician

Proportion of heads among 10,000 coin tosses (Kerrich 1946)
Let’s reproduce and improve Kerrich results without going to jail

Show of hands:
Who has a laptop with **Matlab installed** now?
Matlab is easy to learn

- Matlab is the lingua franca of all of engineering
- Use online tutorials e.g.: https://www.youtube.com/watch?v=82TGGQApFlQ
- Matlab is designed to work with Matrices → symbols * and / are understood as matrix multiplication and division
- Use .* and ./ for regular (non-matrix) multiplication
- Add ; in the end of the line to avoid displaying the output on the screen
- Loops: for i=1:100; f(i)=floor(2.*rand); end;
- Conditional statements: if rand>0.5; count=count+1; end;
- Plotting: plot(x,y,’ko’); or semilogx(x,y,’ko’); or loglog(x,y,’ko’); . To keep adding plots onto the same axes use: hold on; To create a new axes use figure;
- Generating matrices: rand(100) – generates square matrix 100x100. **Confusing!** Use rand(100,1) or zeros(30,20), or randn(1,40) (Gaussian);
- If Matlab complains multiplying matrices check sizes using whos and if needed use transpose operation: x=x’;
A Matlab Cheat-sheet (MIT 18.06, Fall 2007)

Basics:

- **save** 'file.mat': save variables to file.mat
- **load** 'file.mat': load variables from file.mat
- **diary on**: record input/output to file diary
- **diary off**: stop recording
- **whos**: list all variables currently defined
- **clear**: delete/undefine all variables
- **help command**: quick help on a given command
- **doc command**: extensive help on a given command

Constructing a few simple matrices:

- **rand(12,4)**: a 12×4 matrix with uniform random numbers in [0,1)
- **randn(12,4)**: a 12×4 matrix with Gaussian random (center 0, variance 1)
- **zeros(12,4)**: a 12×4 matrix of zeros
- **ones(12,4)**: a 12×4 matrix of ones
- **eye(5)**: a 5×5 identity matrix I("eye")
- **eye(12,4)**: a 12×4 matrix whose first 4 rows are the 4×4 identity
- **linspace(1.2,4.7,100)**

Defining/changing variables:

- **x = 3**: define variable x to be 3
- **x = [1 2 3]**: set x to the 1×3 row-vector (1,2,3)
- **x(2) = 7**: change x from (1,2,3) to (1,7,3)

Portions of matrices and vectors:

- **x(2:12)**: the 2nd to the last elements of x
- **x(1:3:end)**: every third element of x, from 1st to the last
- **A(5,:)**: the row vector of every element in the 5th row of A
- **A(5,1:3)**: the column vector of every element in the 2nd column of A
- **diag(A)**: diagonal of the matrix A

Arithmetic and functions of numbers:

- **exp(12)**: compute e for A matrix and b column vector, the solution x to Ax=b
- **log(3)**, **log10(100)**: compute the natural log (ln) and base-10 log (log10)
- **sqrt(-5)**: compute the square root of −5
- **sqrtm(A)**: the matrix whose square is A

Arithmetic and functions of vectors and matrices:

- **x * y**
- **x.* y**: element-wise product of vectors x and y
- **x ^ y**: multiply, add, subtract, and divide numbers
- **x + y**: element-wise addition of two vectors x and y
- **x + 2**: add 2 to every element of x
- **x * A**: product of a matrix A and a vector y
- **A * y**: product of two matrices A and B
- **A * B**: product of two matrices A and B
- **A'**: the complex-conjugate of the transpose of A
- **x'**: the transpose of x

Solving linear equations:

- **A \\ b**: for A a matrix and b a column vector, the solution x to Ax=b
- **inv(A)**: the inverse matrix A⁻¹
- **[L,U,P] = lu(A)**: the LU factorization PA=LU
- **A \ b**: the columns of A are the eigenvectors of A, and the diagonals diag(D) are the eigenvalues of A
- **eig(A)**: the eigenvalues of A
- **[V,D] = eig(A)**: the columns of V are the eigenvectors of A, and the diagonals diag(D) are the eigenvalues of A

Transposes and dot products:

- **x.'**: the transposes of x and A
- **A.'**: the complex-conjugate of the transposes of x and A
- **x' * y**: the dot (inner) product of two column vectors x and y
- **x * y**: the outer product of two column vectors x and y

http://web.mit.edu/18.06/www/Spring09/matlab-cheatsheet.pdf