Randomness and information in biological data
BIO-369

Prof. Anne-Florence Bitbol
TAs: Cecilia Fruet, Margaret Lane, Benjamin Martin, Shuhao Zhang

EPFL

First class

Outline of the course

I Randomness in biological processes and biological data

- 1 Randomness and random variables
  - 1.1 Coins and dice: discrete random variables
  - 1.2 Medical testing and conditional probabilities
  - 1.3 Luria-Delbrück experiment: Poisson distribution vs. jackpot distribution
- 2 Importance of thermal fluctuations at the cellular scale
  - 2.1 Thermal fluctuations and associated energy scale
  - 2.2 Strength of various chemical bonds
  - 2.3 Flexibility of biopolymers and biomembranes
- 3 Random walks
  - 3.1 Population genetics
  - 3.2 Protein abundances in single cells
  - 3.3 Importance of random walks in biological systems

Outline of the course

II Extracting information from biological data

- 1 Quantifying randomness and information in data: entropy
  - 1.1 Notion of entropy
  - 1.2 Interpretation of entropy
  - 1.3 Entropy in neuroscience data: response of a neuron to a sensory input
- 2 Quantifying statistical dependence
  - 2.1 Covariance and correlation
  - 2.2 Mutual information
  - 2.3 Identifying coevolving sites in interacting proteins using sequence data
- 3 Inferring probability distributions from data
  - 3.1 Model selection and parameter estimation: maximum likelihood
  - 3.2 Introduction to maximum entropy inference
  - 3.3 Predicting protein structure from sequence data
- 4 Finding relevant dimensions in data: dimension reduction
  - 4.1 Principal component analysis
  - 4.2 Beyond principal component analysis
- 5 Introduction to Bayesian inference

Organization

- **Problem classes:** 3:15pm-5pm on Mondays, room CE1106
- **Lectures:** 10:15am-12noon on Wednesdays, room BS170
- Problem class about *previous* lecture

- **First two weeks:**

- Monday, Feb. 19, 3:15pm-5pm: lecture 1; Wednesday, Feb. 21, 10:15am-12noon: lecture 2;
- Monday, Feb. 26, 3:15pm-5pm: problem sets 1&2; Wednesday, Feb. 28, 10:15am-12noon: lecture 3.

- All questions are welcome:
  - During lectures and problem classes
  - Outside of lecture hours, using **Ed Discussion**

- All class material is available on **Moodle**
- Computer-based questions in problems + project: **Jupyter Notebooks with Python3**
- Some quiz questions during lectures, using **TurningPoint**

- **Numerical mini-project** during the semester (40% of the final grade)
  Weeks 10 & 11 (TBC), with two problem sessions on week 10 and one on week 11 devoted to it
- **Written exam** during the exam session (60% of the final grade)
  Mainly classic problems + some coding questions

A few words about TurningPoint

- Need a device connected to the Internet to answer the question

- No need for a TurningPoint account / license – all answers will be **anonymous** (as “guests”)

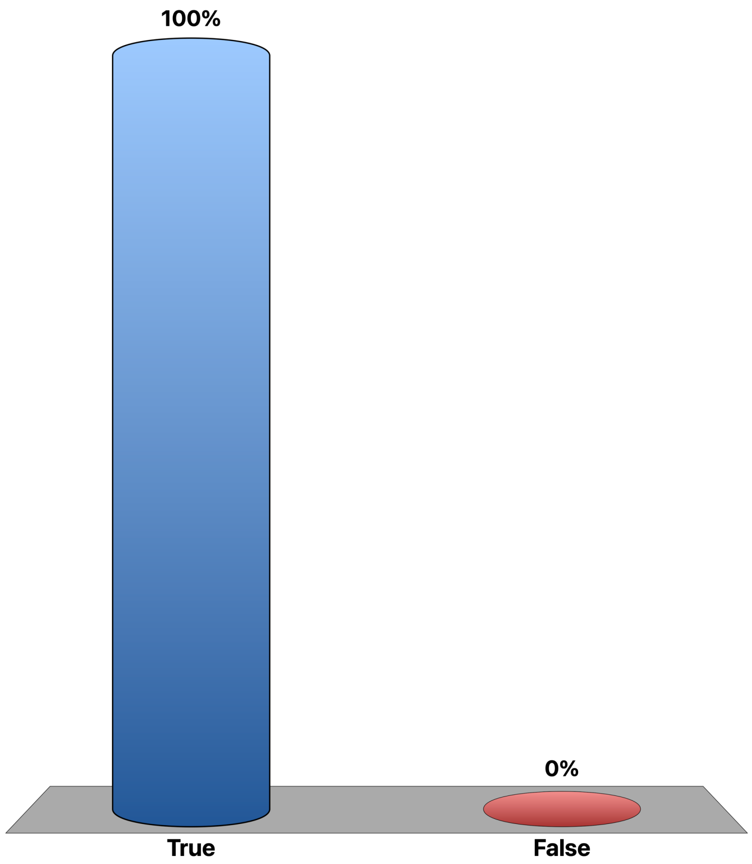
- **Goals:**
  - Think about a new notion
  - Recall a previously seen notion
  - Encourage active participation
  - Get feedback

- **Not an evaluation**

- Legal information: Data is processed outside Switzerland, which may include the USA or EU countries
  Contract with EPFL = Turning Technologies will not reuse the data collected for any other purpose, provided that you use the “guest mode”

Have you used Python before?

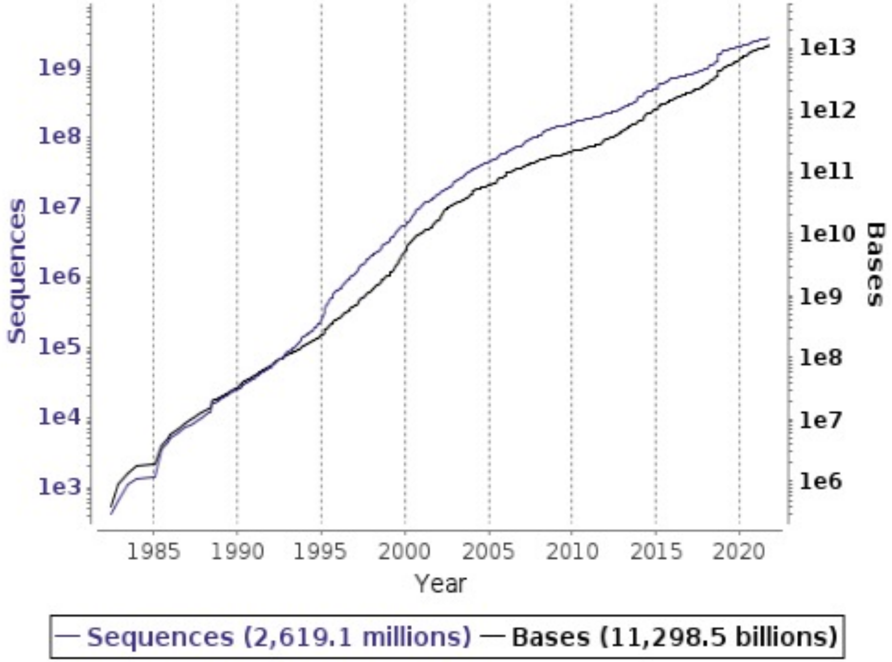
A. True
B. False



To answer, please:
- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

Motivation

- **Biology is becoming more and more a data science**
  - **Example: sequencing**



    - Sequence data

```
ISHDLKTPITAIILLDLMLPGIDG
VSHELKTPLTSIVILDLNLPKQDG
VSHELRTPLTSILVLDLMLPEIGG
ASHELRTPISVIVLLDIMLPGLSG
ISHDLKTPITAIILLDLMLPGIDG
ASHELRTPISVIVLLDIMLPGLSG
VSHELRTPLTSILVLDLMLPEIGG
```

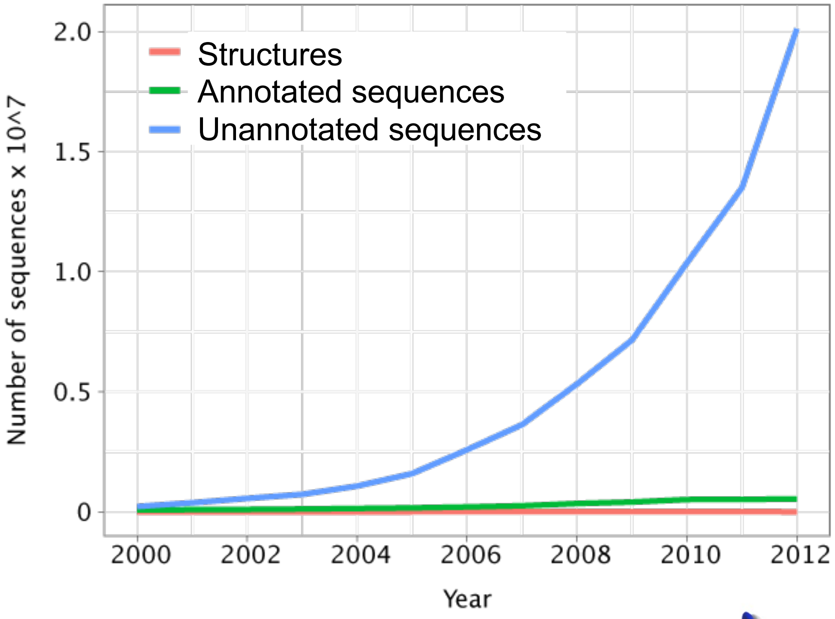
  - Accumulating unannotated sequence data (currently > 10^9 sequences)

→ **Great opportunity to learn about proteins employing inference, machine learning, statistical physics, information theory**

Motivation

- **Biology is becoming more and more a data science**
  - **Example: sequencing**
    - Sequence data

```
ISHDLKTPITAIILLDLMLPGIDG
VSHELKTPLTSIVILDLNLPKQDG
VSHELRTPLTSILVLDLMLPEIGG
ASHELRTPISVIVLLDIMLPGLSG
ISHDLKTPITAIILLDLMLPGIDG
ASHELRTPISVIVLLDIMLPGLSG
VSHELRTPLTSILVLDLMLPEIGG
```



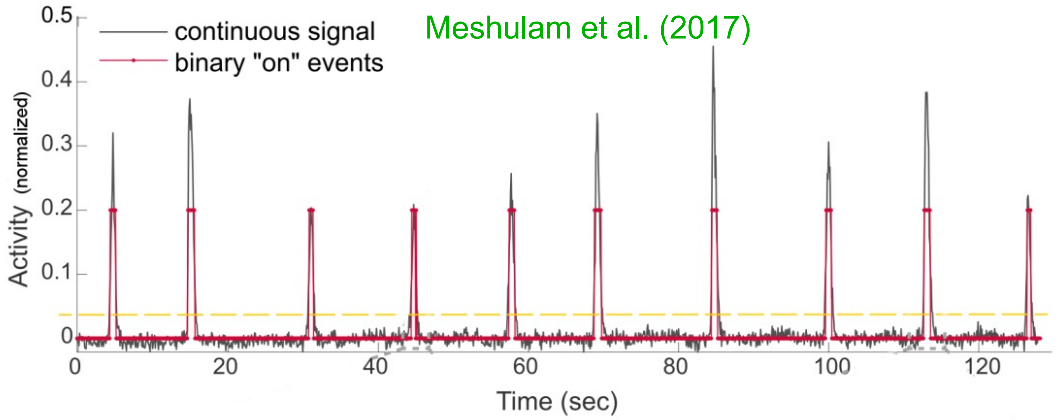
- Accumulating unannotated sequence data (currently > 10^9 sequences)

→ **Great opportunity to learn about proteins employing inference, machine learning, statistical physics, information theory**

Motivation

- **Biological data can be viewed as sampled from distributions of random variables**

- **Neuroscience data:**

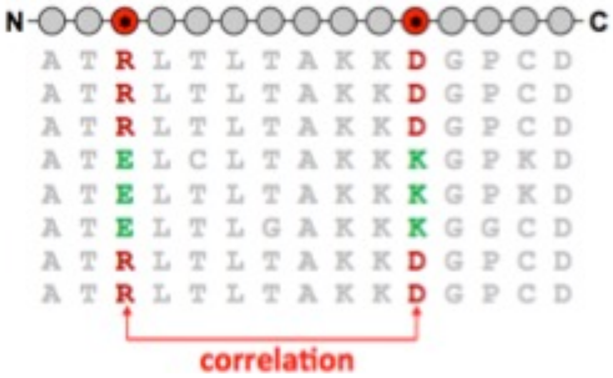


Meshulam et al. (2017)

$$P(\{\sigma_i\}) = \frac{1}{Z}\exp[-E(\{\sigma_i\})].$$

$$E(\{\sigma_i\}) = -\sum_{i=1}^{N} h_i\sigma_i - \frac{1}{2}\sum_{i,j=1}^{N} J_{ij}\sigma_i\sigma_j$$

- **Protein sequence data:**



$$P(\alpha_1, \ldots, \alpha_L) = \frac{1}{Z}\exp\left\{-\left[\sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i, \alpha_j)\right]\right\}$$

Weigt, White et al. (2009)
Morcos et al. (2011)
Marks, Colwell et al. (2011)

Laboratory of Computational Biology and Theoretical Biophysics

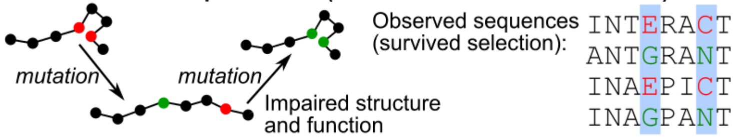
molecular scale

Optimization & historical contingency

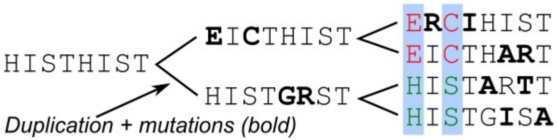
population scale

- **Aim 1:**
  Understanding how optimization & phylogeny shape protein sequences

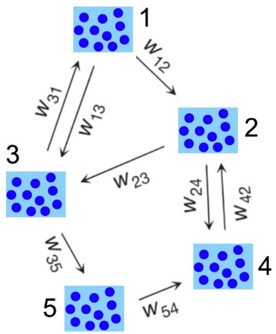
Correlations from optimization (maintain structure and function):

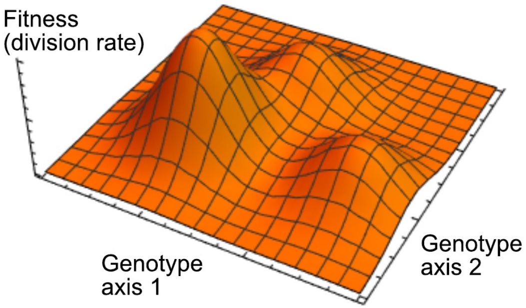


Correlations from historical contingency (phlogeny):



- **Aim 2:**
  Assessing the predictability of evolution in subdivided populations





Outline of the course

What is the name of the distribution that describes the outcome of a coin flip, including the case where the coin is not fair?

11%	A.	Uniform distribution
43%	B.	Bernoulli distribution
46%	C.	Binomial distribution
0%	D.	Poisson distribution

To answer, please:

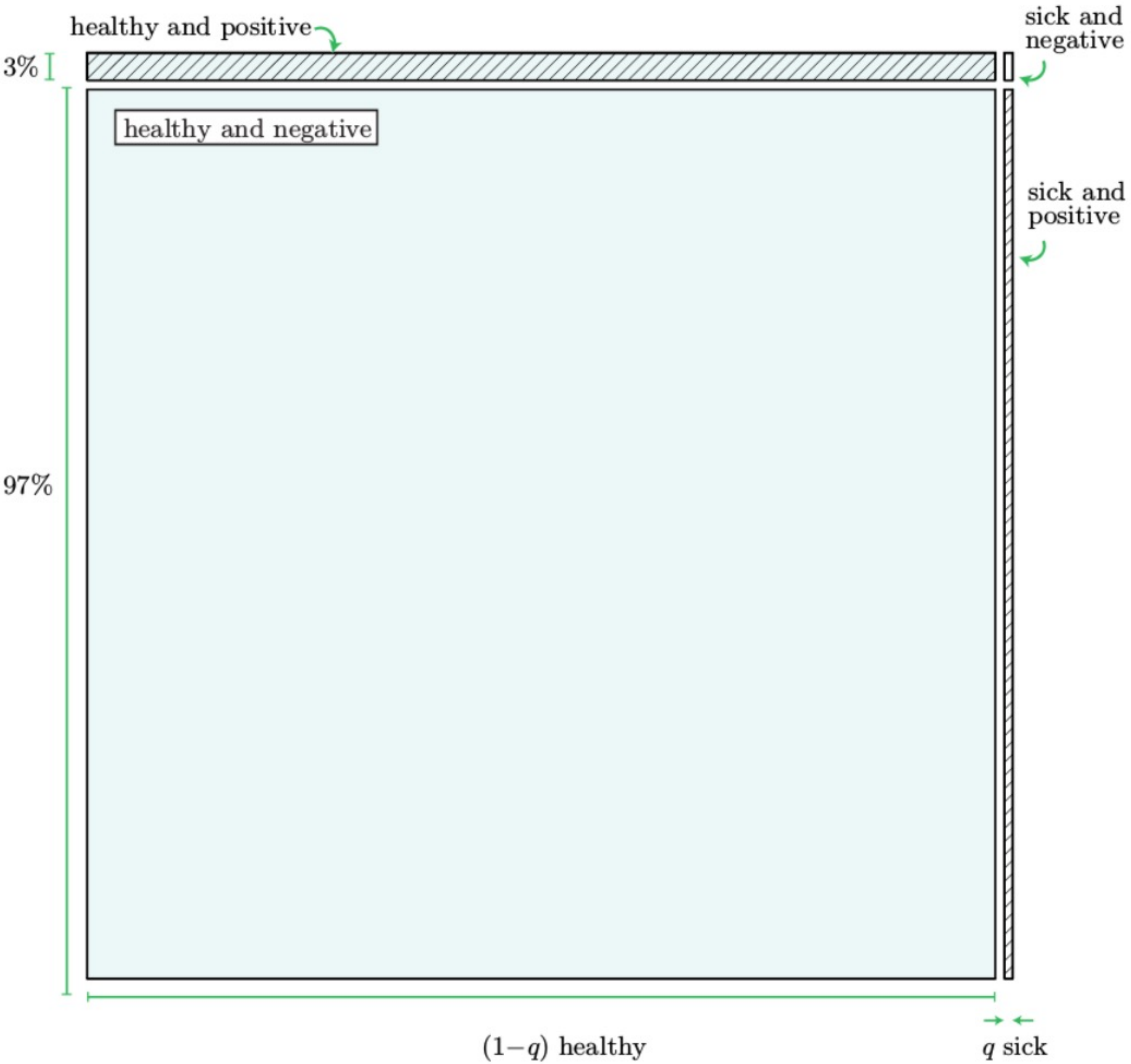
- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

Consider two random variables X and Y.
When can we write, for all x and y, $P(X=x, Y=y) = P(X=x)\,P(Y=y)$?

0%	A.	Always
0%	B.	Never
0%	C.	When X and Y have the same distribution
78%	✓ D.	When X and Y are independent
22%	E.	When X and Y have the same distribution and are independent

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer


healthy and positive
sick and negative
3%
healthy and negative
sick and positive
97%
(1−q) healthy
q sick

Imagine that you are a doctor. Your patient has taken a test, and tested positive (p). Before taking action, what would you like to know?

2%	A.	P(s)
2%	B.	P(p)
0%	C.	P(s,p)
83%	✔ D.	P(s|p)
12%	E.	P(p|s)

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

All patients are either sick (s) or healthy (h), and all patients are tested (p or n). What is P(p,s)+P(p,h) equal to?

0% A. P(p)

0% B. P(s)

0% C. 1

0% D. P(p|s)+P(p|h)

0% E. None of the above

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

All healthy (h) patients are tested, and all tests are either positive (p) or negative (n). What is P(p|h)+P(n|h) equal to?

39%	A.	P(h)
0%	B.	P(n,h)
55%	✓ C.	1
3%	D.	P(p,h)+P(n,h)
3%	E.	None of the above

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer