Randomness and information in biological data
BIO-369

Prof. Anne-Florence Bitbol

EPFL

Lecture 10

Organization

Evaluation

- **Numerical mini-project** (**40%** of the final grade)
  - Three problem sessions devoted to working on the mini-project (weeks 10-11):
    - **Monday April 28 at 3:15pm** in room CE1106
    - **Monday May 5 at 10:15am** in room BS170
    - **Wednesday May 7 at 3:15pm** in room CE1106 (lecture slot)
  - Deadline to hand in the mini-project: **Friday May 9 (11:59pm)**
  - You will have to hand in a **Jupyter Notebook in Python 3 via Moodle ("Project assignment")**
  - Communicating is allowed, asking questions to TAs during problem sessions, or on EdDiscussion, is allowed – but we won't answer the questions of the project
  - Personal thought will be valued; detected plagiarism will result in a reduction of your grade
  - Coding style only evaluated as a bonus

- **Written exam** during the session (**60%** of the final grade) – **Monday June 30 from 9:15 to 12:15**
  - One "formula sheet" (formulaire) allowed:
    - Hand-written (can be hand-written on a tablet and printed, but not typed)
    - Maximum size: one two-sided standard A4 sheet
  - No other documents allowed
  - Calculator (standard)

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

What do you think about the ratio P(model)/P(model')?

57%	A.	We can calculate it if we know what the models are
0%	B.	It worries me, I don't think we can calculate it
43%	C.	It will not matter in the end

To answer, please:

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

What is a prior?

https://www.youtube.com/watch?v=A4QcyW-qTUg

To find the maximum likelihood estimate of the bias of the coin, you need to:

11%	A.	Maximize P(m) with respect to N, m and p
33%	B.	Maximize P(m) with respect to N and m
44%	C.	Maximize P(m) with respect to p
11%	D.	Maximize P(m) with respect to p and m
0%	E.	Maximize P(m) with respect to m

To answer, please:

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

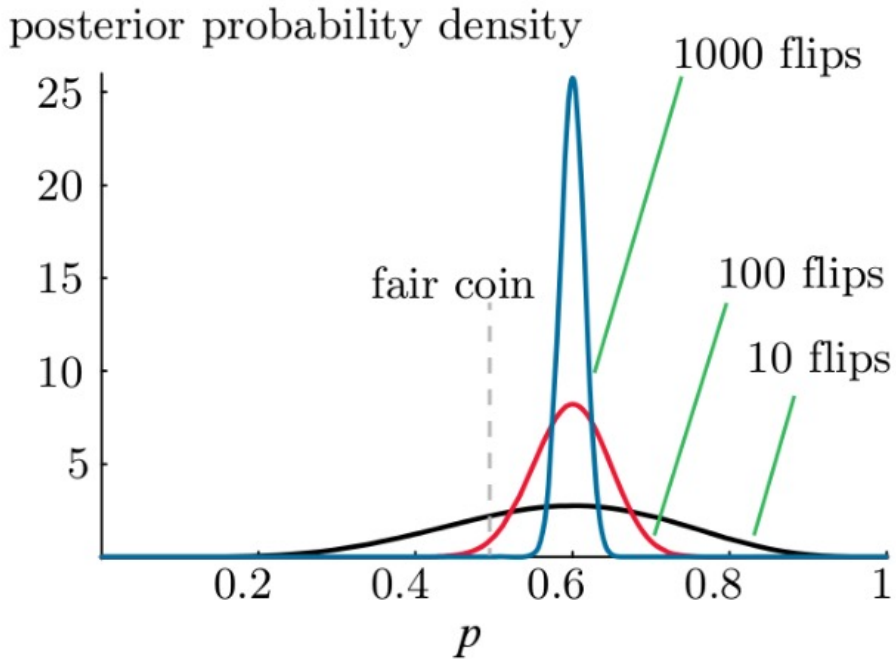
What do you expect for the maximum-likelihood estimate of p?

0%	A.	p=1/2
0%	B.	p=0
0%	C.	p=1
0%	D.	p=m/N

To answer, please:

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

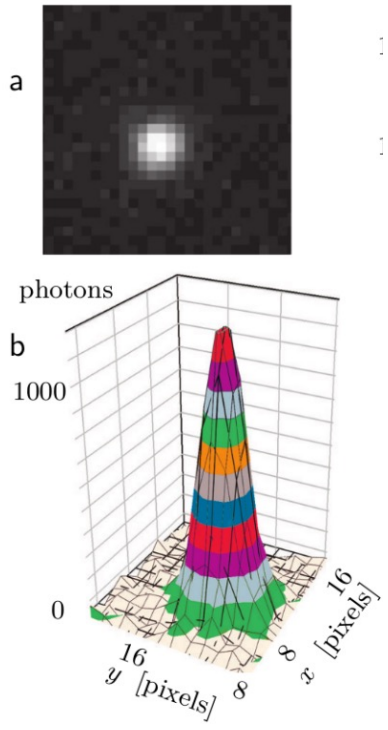
Likelihood analysis of the bias of a coin



Posterior probability distributions P (model, p | data) for the probability p of getting “heads” upon a flip

Black is 10 flips, of which 6 were heads; red is 100 flips, of which 60 were heads; blue is 1000 flips, of which 600 were heads

Superresolution microscopy: FIONA

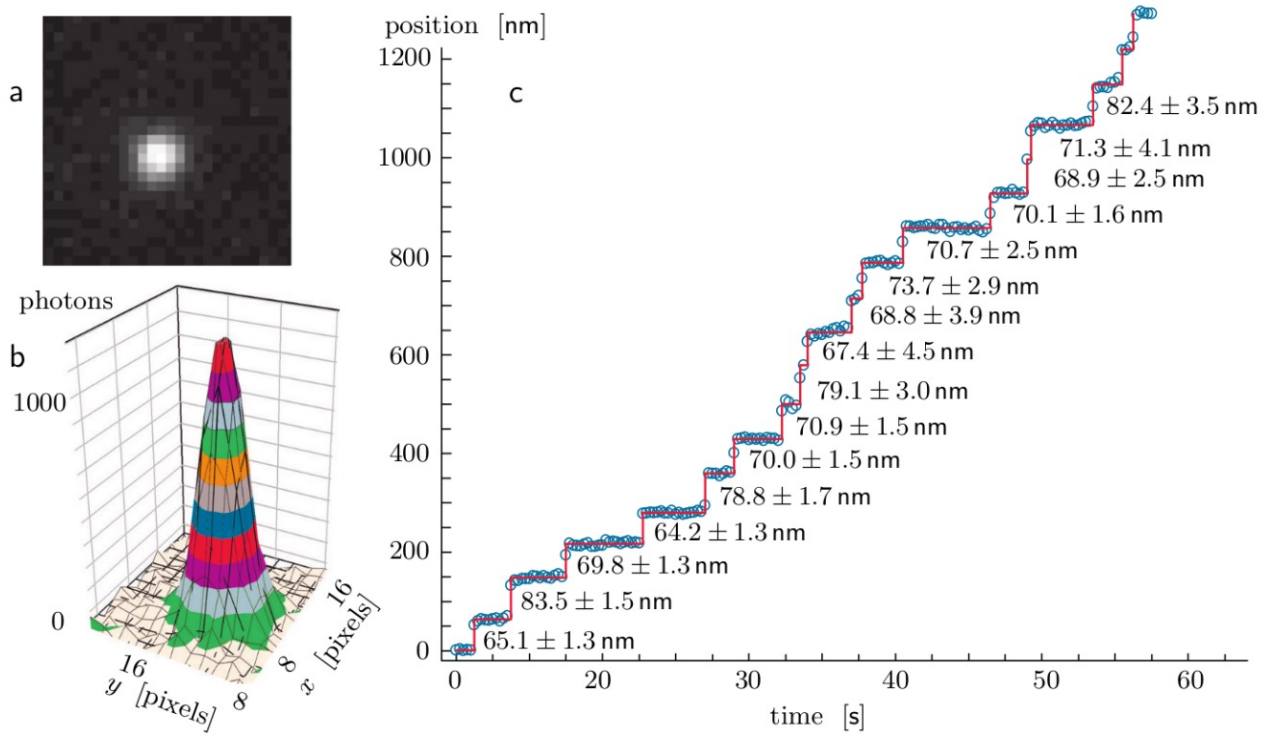


a: Single image of a single fluorophore attached to the molecular motor protein myosin-V. Each camera pixel represents 86 nm in the system

b: Number of photons collected in each pixel for the image in (a)

→ movie “showing” myosin stepping on actin

Superresolution microscopy: FIONA



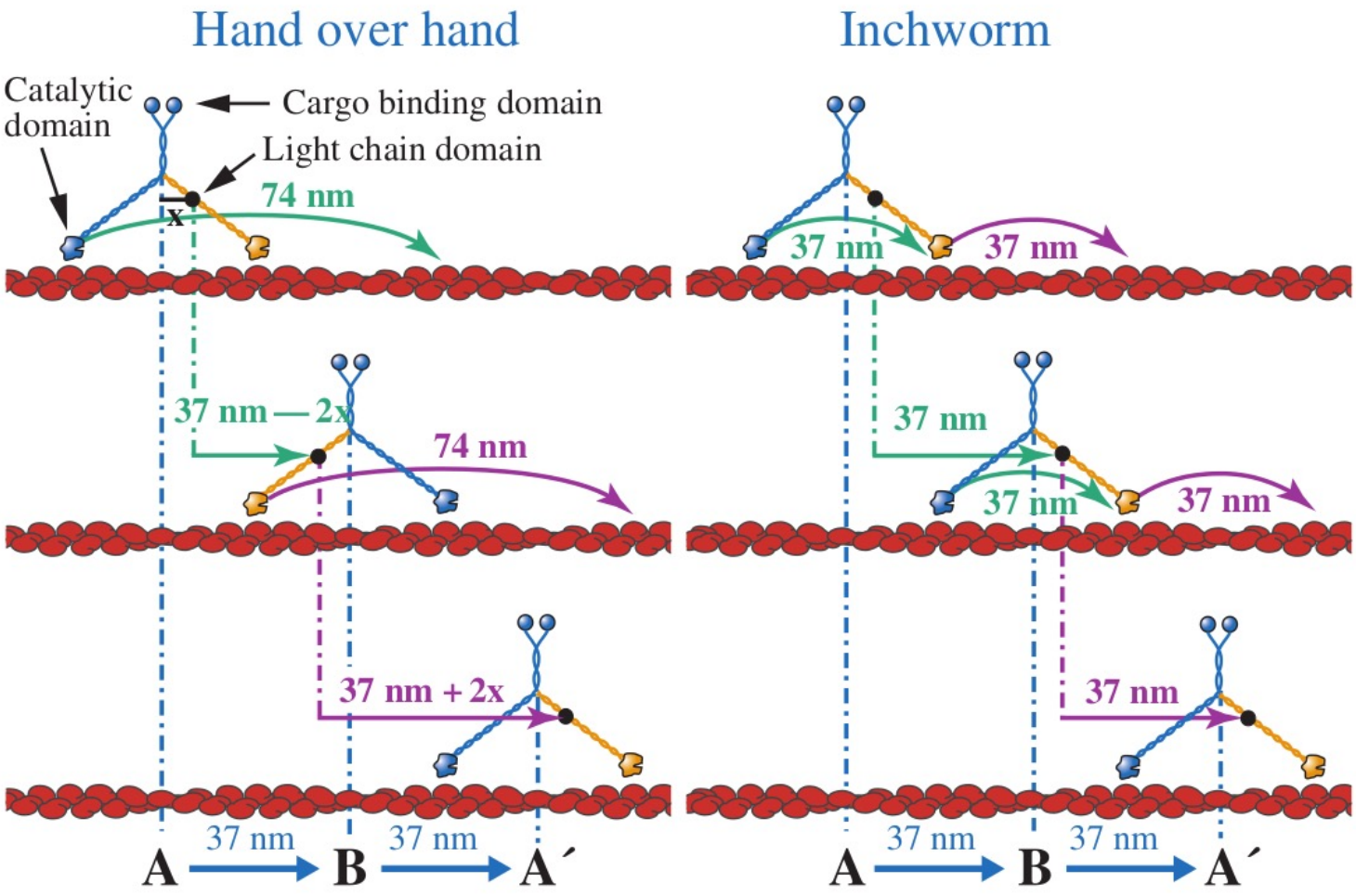
Yildiz et al, 2003

a: Single image of a single fluorophore attached to the molecular motor protein myosin-V. Each camera pixel represents 86 nm in the system

b: Number of photons collected in each pixel for the image in (a)

c: Maximum likelihood estimates of the position of the fluorophore versus time, revealing a sequence of ~74 nm steps

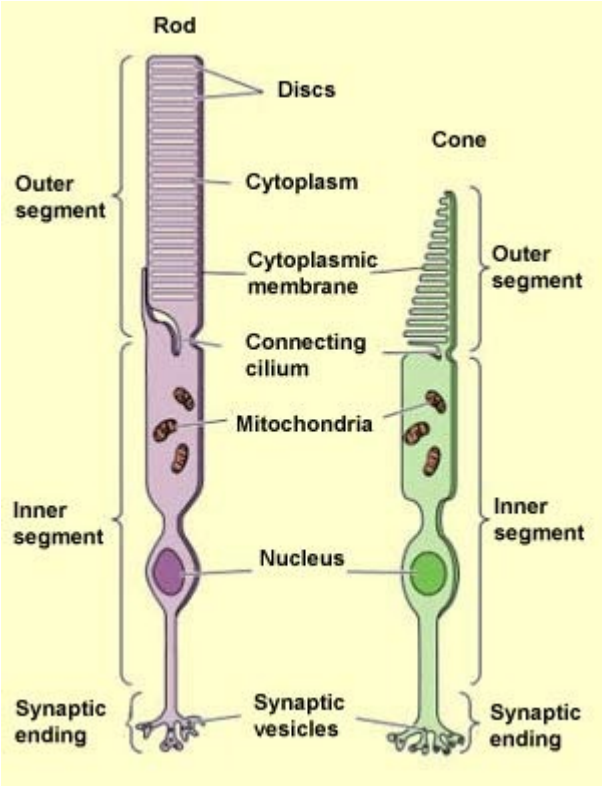
Superresolution microscopy: FIONA



Yildiz et al, 2003

Allowed to settle how myosin “walks” on actin filaments

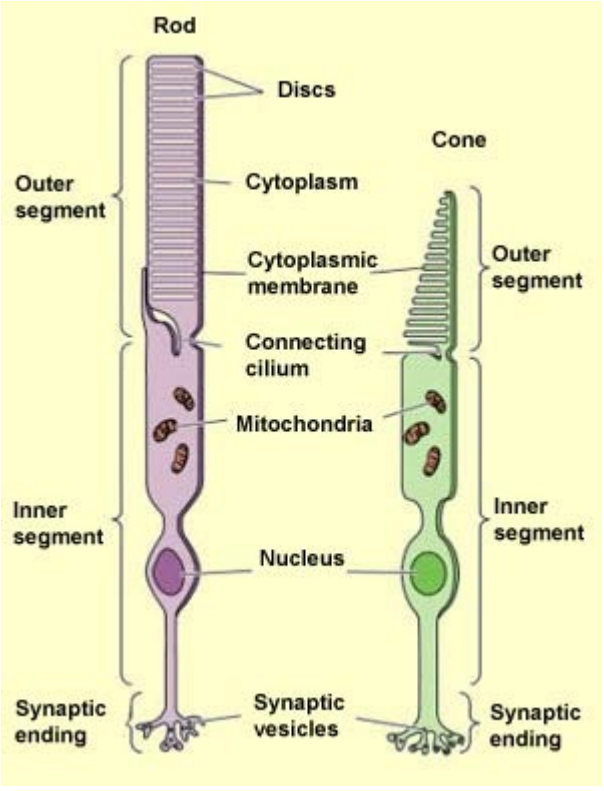
Photoreceptor cells in the retina



The outer segment consists of a stack of discs embedded in the cell membrane
Light-sensitive pigments are located on these discs

Rod cells can function in lower light better than cone cells, but have little role in color vision

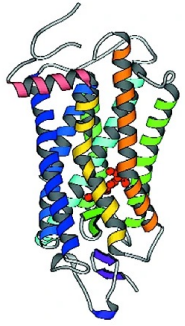
Photoreceptor cells in the retina



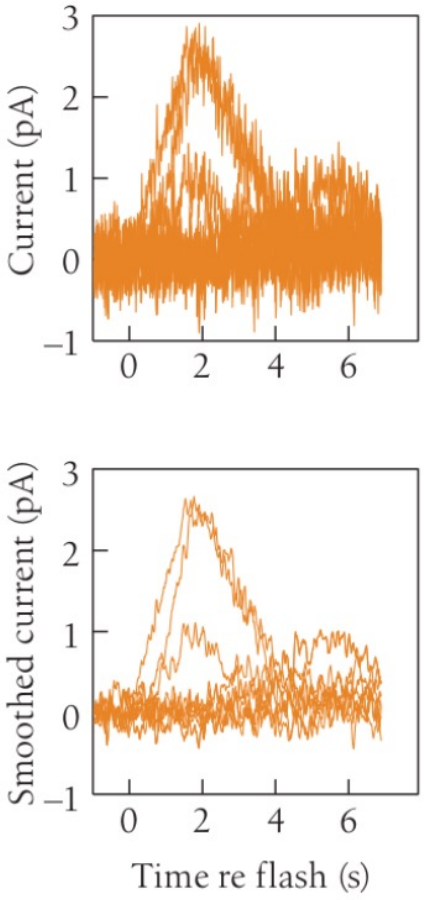
The outer segment consists of a stack of discs embedded in the cell membrane
Light-sensitive pigments are located on these discs

Rod cells can function in lower light better than cone cells, but have little role in color vision

Rod cells contain rhodopsin, a light-sensitive transmembrane protein (and a GPCR)
Light → structural change that increases its affinity for another protein and triggers a signaling pathway
→ closing of ion channels and hyperpolarization
→ change in the current across the membrane of the rod cell

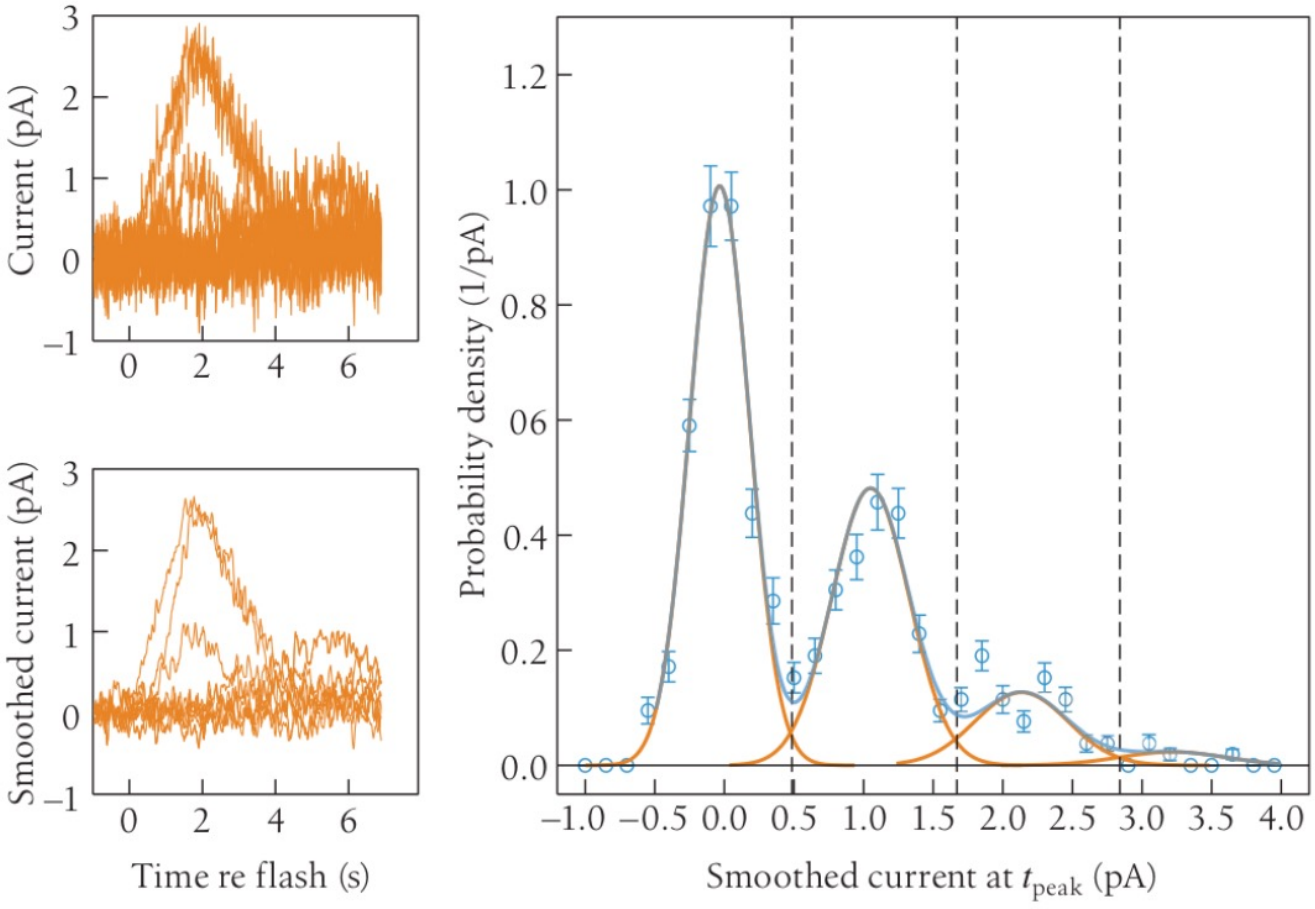


Current in a rod cell exposed to a dim flash of light



Left panels (top: raw data; bottom: data smoothed by moving average on a 100 ms window): 5 instances in which the rod cell is exposed to a dim flash at t = 0

Current in a rod cell exposed to a dim flash of light



Left panels (top: raw data; bottom: data smoothed by moving average on a 100 ms window): 5 instances in which the rod cell is exposed to a dim flash at t = 0

Right panel: distribution of smoothed currents at t_{peak}, mean and standard error from 350 flashes in one cell
Blue line: fit to distribution, composed of contributions from N = 0, 1, etc. (orange)

The probability density p of observing a given intensity i for a dim flash can be expressed as the sum over the number N of photons received by the rod cell of:

0% A. p(N)

0% B. p(N,i)

0% C. p(N|i)

To answer, please:

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

Assume that the number of photons is 0 or 1. If we choose a threshold theta to decide this, then the probability of making an error on our conclusion on the number of photons is:

0% A. P(conclude that N=0 | N=1)

0% B. P(conclude that N=0, N=1)

0% C. P(conclude that N=0 | N=1)+P(conclude that N=1 | N=0)

0% D. P(conclude that N=0, N=1)+P(conclude that N=1, N=0)

To answer, please:

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

What do you expect the optimal threshold theta to satisfy?

0% A. P(i=theta | N=0) = P(i=theta | N=1)

0% B. P(i=theta, N=0) = P(i=theta, N=1)

0% C. P(N=0 | i=theta) = P(N=1 | i=theta)

0% D. P(i=theta) = 1/2

To answer, please:

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

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

Now that we have found the form of P(x), what should we do?

0% A. We are done, this probability distribution works for any lambda

0% B. We should choose the value of lambda such that the distribution is normalized

0% C. There is only one value of lambda that works, and it depends on the data

To answer, please:

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