Randomness and information in biological data
BIO-369

Prof. Anne-Florence Bitbol

EPFL

Lecture 14

Announcements

- **Documents and grades:**
  - Full typed notes + mock exam now available on Moodle
  - Project solution and grades now available on Moodle

- Extra problem session during exam preparation: **Friday June 20 at 10:15am**, room TBD

- Please fill in the **in-depth evaluation of the class** (closes on June 8 at midnight)
  → Moodle homepage/dashboard, « in-depth evaluation » tile, or EPFL Campus App (cf. email)

Thank you very much!

About the exam

- **Written exam** (60% of the final grade)

- 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 allowed, no other electronic device allowed

- **Date and place:** June 30, 9:15am to 12:15pm, rooms CE 1 104-5-6 (see IS-Academia)

- **Format:**
  - Classic written exam, problems with questions related to lectures and problem classes
  - A few « coding questions » that can be answered in Python or in pseudocode, but this will remain a small proportion of the questions

Outline of the course

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

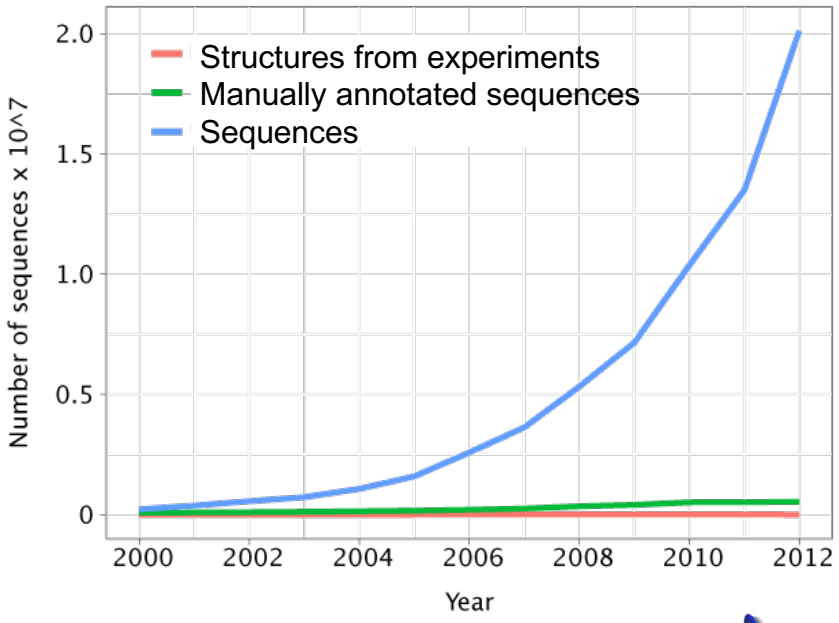
3 Inferring probability distributions from data
- 3.1 Model selection and parameter estimation: maximum likelihood
- 3.2 Introduction to maximum entropy inference
- 3.3 Predicting protein structure from sequence data

4 Finding relevant dimensions in data: dimension reduction
- 4.1 Principal component analysis
- 4.2 Beyond principal component analysis

Motivation

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

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

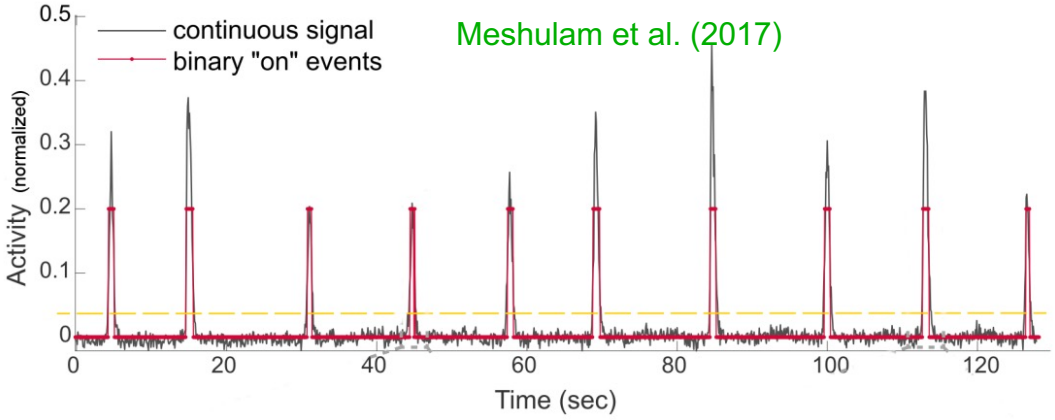


- Accumulating unannotated sequence data (currently > 10^9 sequences)

→ **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:**

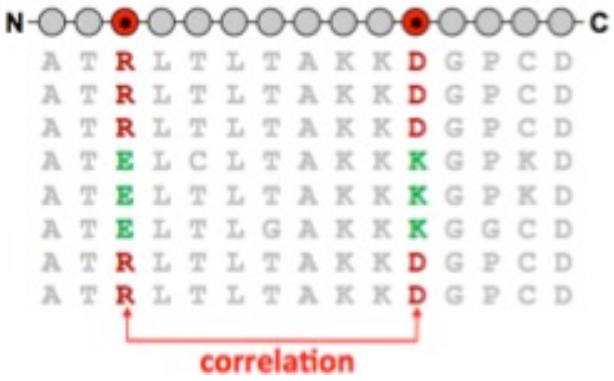


Meshulam et al. (2017)

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

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

  - **Protein sequence data:**



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

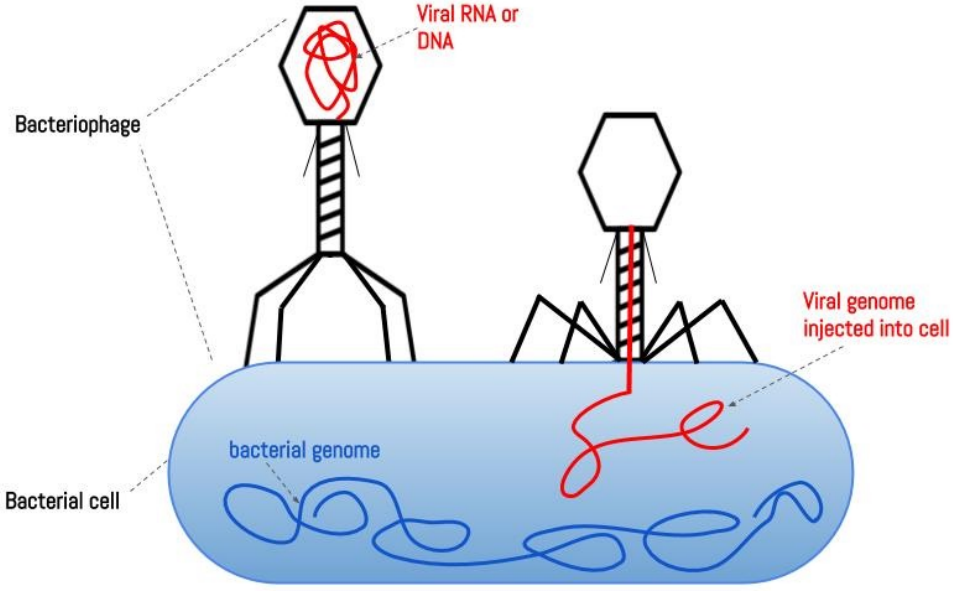
Weigt, White et al. (2009)
Morcos et al. (2011)
Marks, Colwell et al. (2011)

Luria-Delbrück experiment

- **Use of probabilities to test hypotheses about evolution**

Phage and bacteria (phage T1, obligately lytic virus of *E. coli*)

Bacteria can develop resistance to phage infection



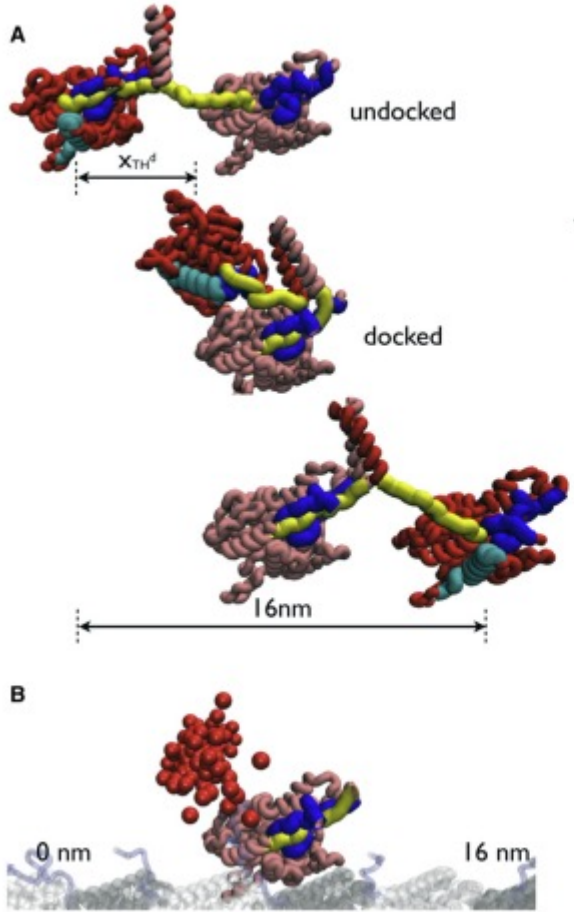
- Is resistance to phage in bacteria a trait that:
  - appears randomly and is then selected upon exposure of bacteria to phage, or
  - appears upon exposure of bacteria to phage (in response to it)?

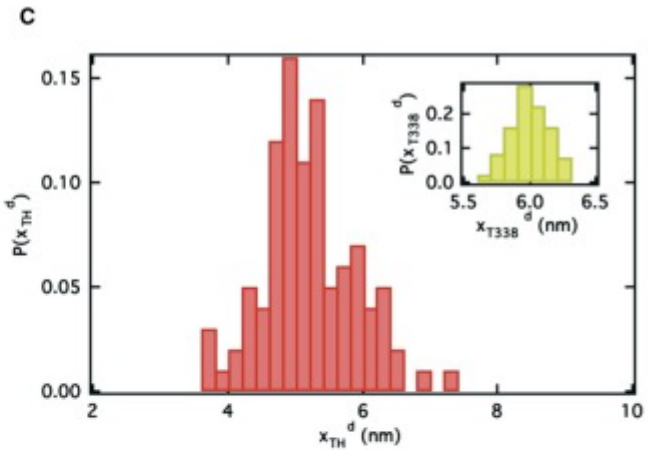
- Each hypothesis yields a different **probability distribution** for the number of phage-resistant bacteria

→ Quantitative test

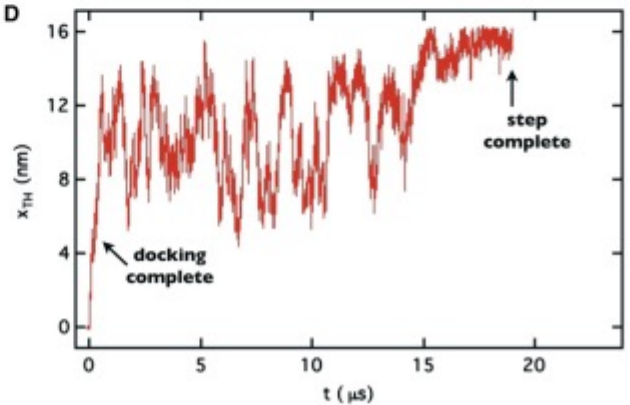
Importance of thermal fluctuations at the cellular scale

- **Kinesin walking on a microtubule**
  - Brownian dynamics simulation (Zhang & Thirumalai 2012)





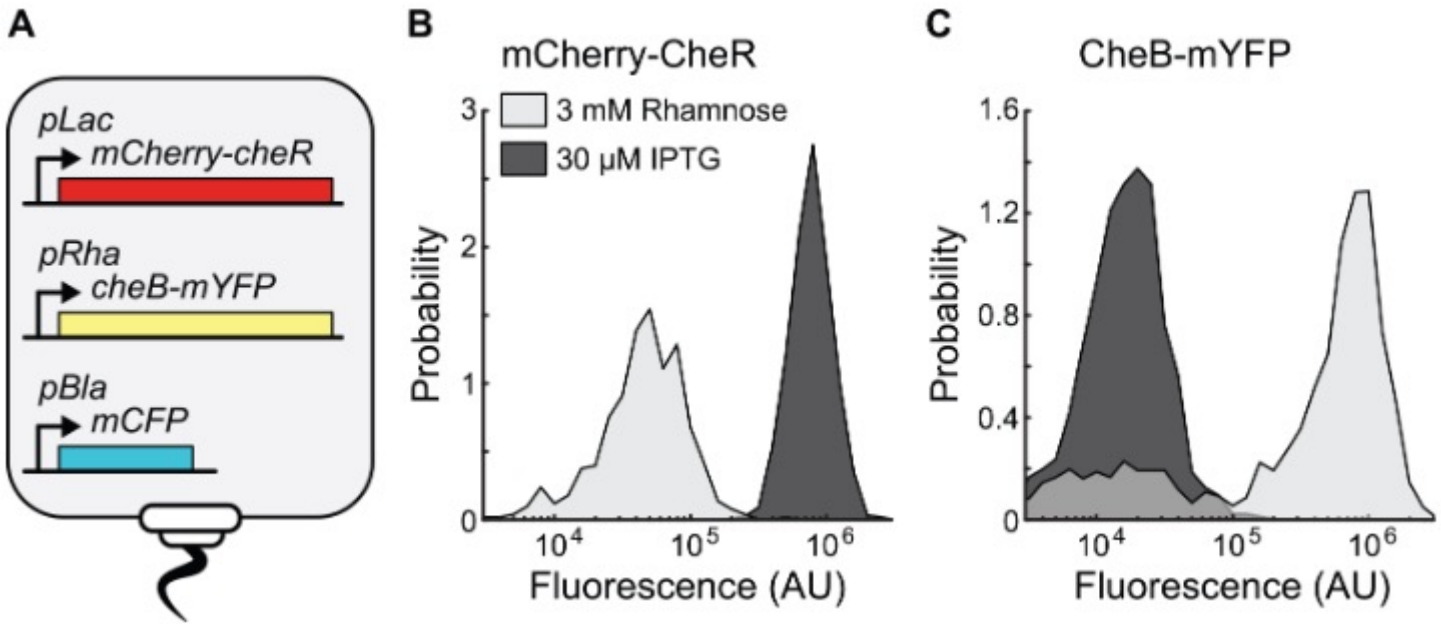
Histograms, based on 99 trajectories, of the TH (trailing head) movement



Time-dependent changes in the center of mass of the TH as a function of t for a sample trajectory

Protein abundances

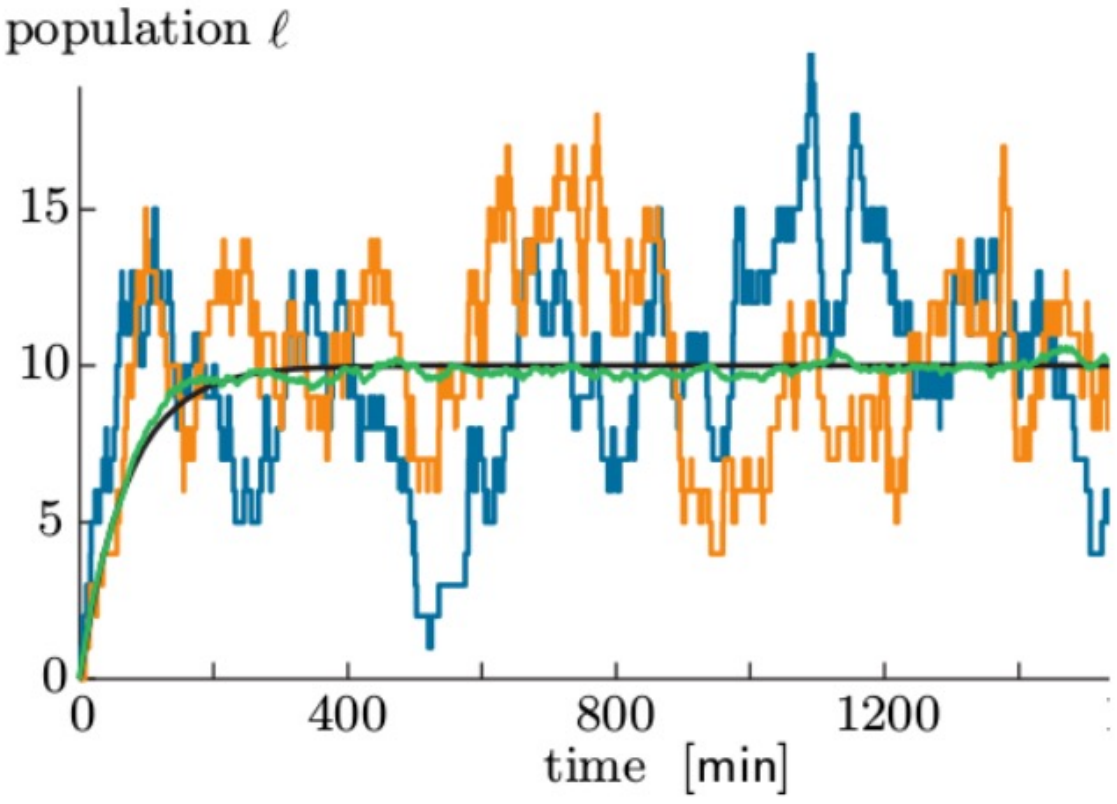
- **Protein abundances are heterogeneous across cells**
  - Expression of fluorescently labeled chemotaxis proteins with inducible promoters under two conditions



Dufour et al 2016

Random walks

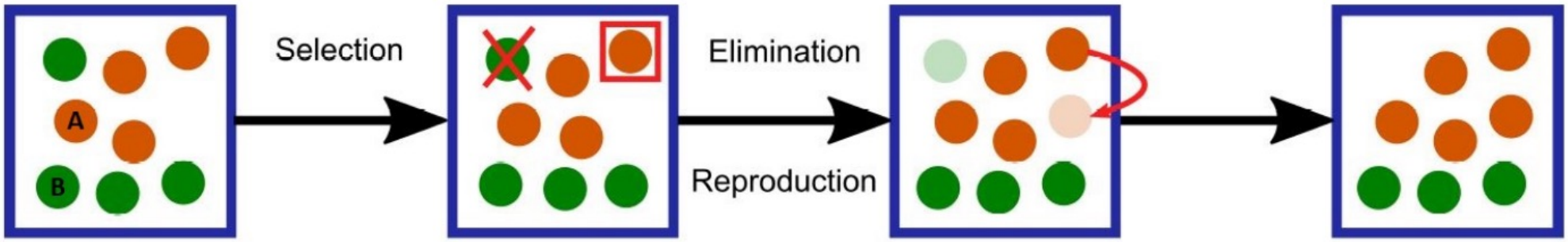
- **Protein abundance as a random walk**



- Simulation of a simple model for the abundance of a protein expressed in small copy numbers
- Orange / blue: number of proteins versus time in individual realizations ℓ starting from 0 protein
- Green: mean over 200 such replicates; Black: deterministic approximation

Random walks

- **The Moran model in population genetics**

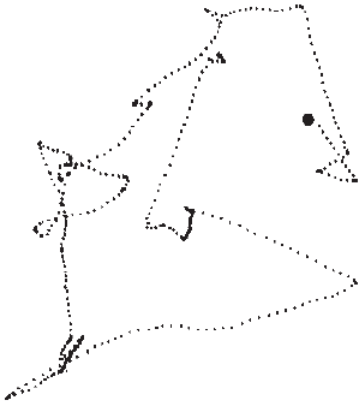


- Schematic of one step of the Moran process
- The population comprises two types of individuals, type A (orange) and type B (green)

Random walks

Bacterial chemotaxis

- Motion of swimming, flagellated bacteria: 3D random walk with “runs” and “tumbles”

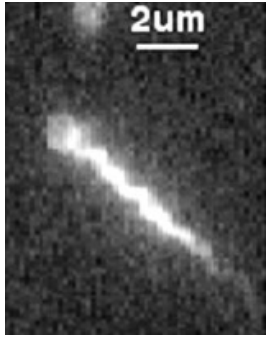


Wild type
29.5s
26 runs
Mean speed 21.2 μm/s

50μm

Berg & Brown (1972)

- Runs → CCW rotation of the flagella
- Tumbles → CW rotation of the flagella



Turner et al. (2000)

- Gradients of chemicals can induce a change of tumbling frequency & CW bias

→ Chemotaxis: bacteria are attracted or repelled by various chemicals

Quantifying randomness and information in data: entropy

Questions:

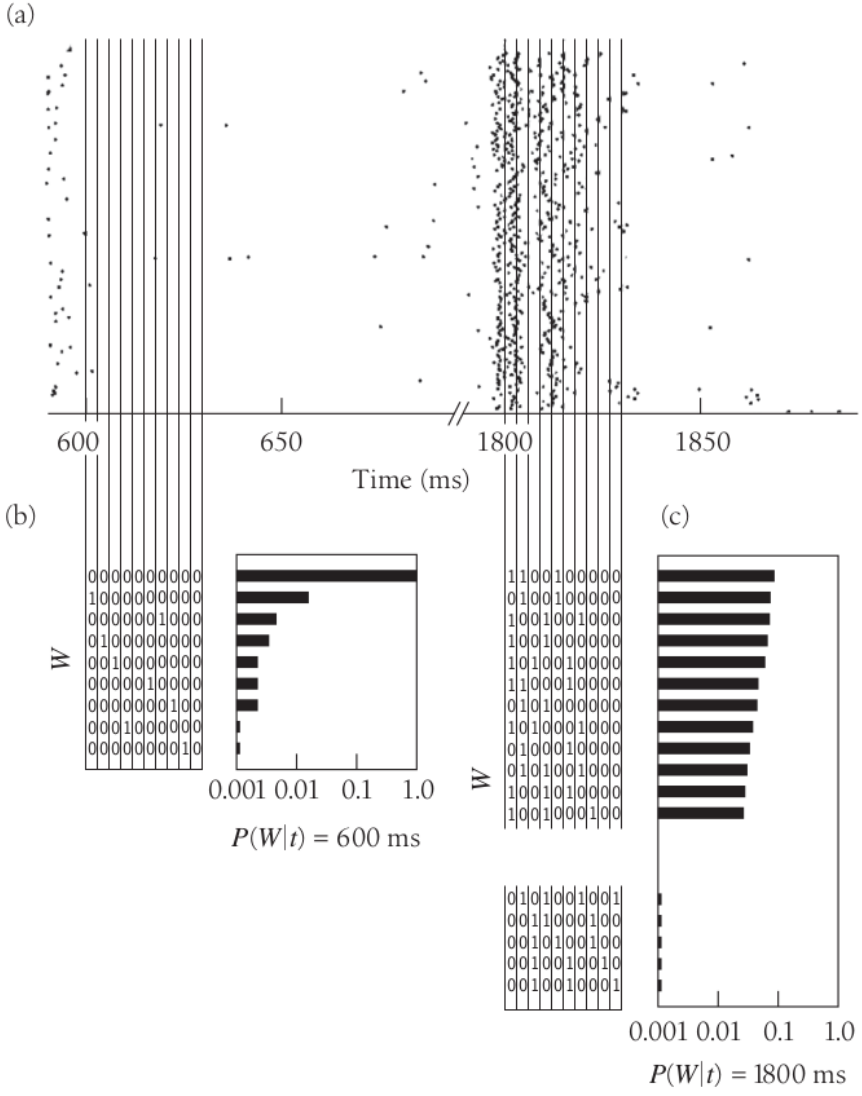
- "How random" is a random variable?
- How much information are we missing when we don't know the outcome of a random variable (but know its distribution)?
- How much information do we gain when learning about the outcome of a random variable (if we know its distribution)?

→ Can we *quantify* randomness and information?

And also...

- How different are two probability distributions?
- Can we quantify statistical dependence between two random variables?

Information in neuroscience data



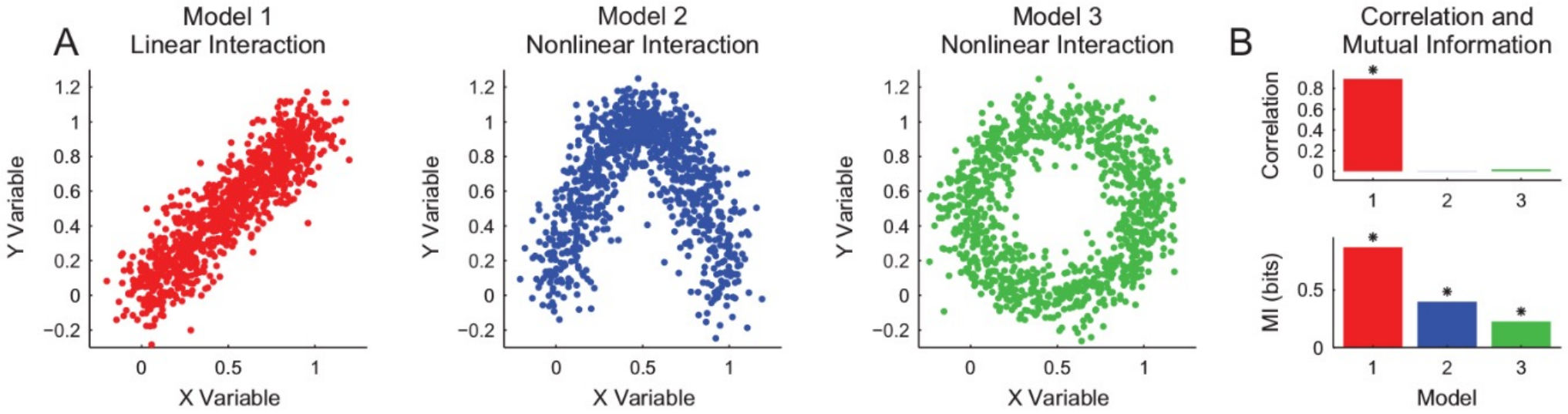
How do sequences of spikes represent the sensory world?

Study of the motion-sensitive neuron H1 in the fly's visual system while a fly is shown the same movie several times (a pattern of random bars that moves across the visual field at variable velocity)

"Words" of 10 characters correspond to τ = 30 ms (~ behavior reaction time), with each binary character corresponding to $\Delta\tau$ = 3 ms

The distribution of words that occur at a particular moment in the movie, P(W|t), is shown for t = 600 ms and t = 1800 ms de Ruyter et al, 1997

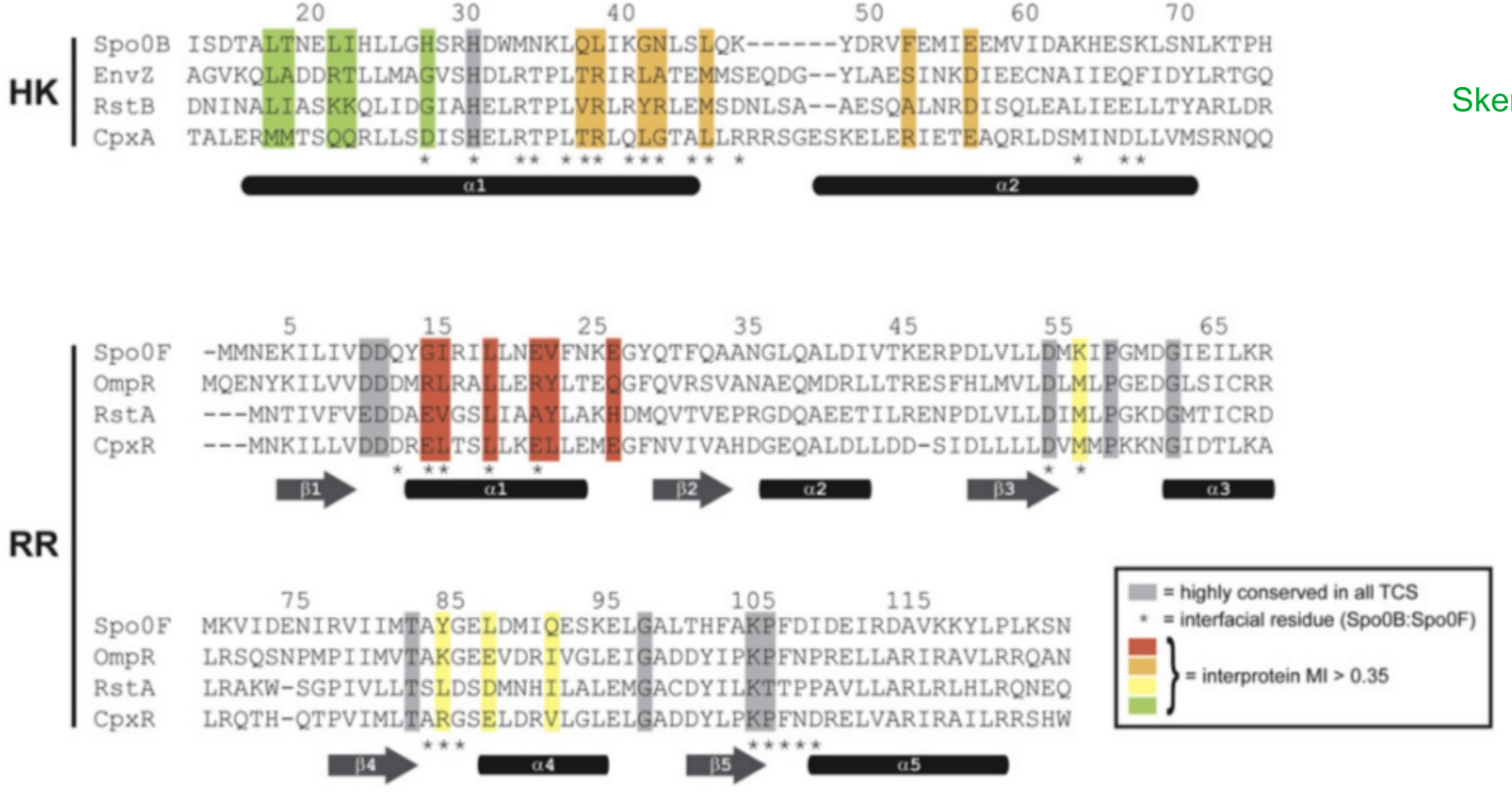
Quantifying statistical dependence



Correlation and mutual information between random variables X and Y
Different draws are performed, yielding values x and y, and the correlation and mutual information are estimated

Mutual information is able to detect some nonlinear forms of statistical dependence that are missed by correlation

Identifying coevolving sites in interacting proteins



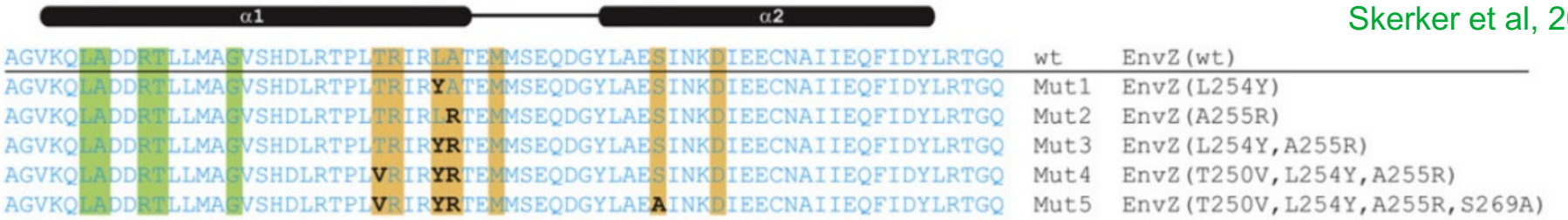
Skerker et al, 2008

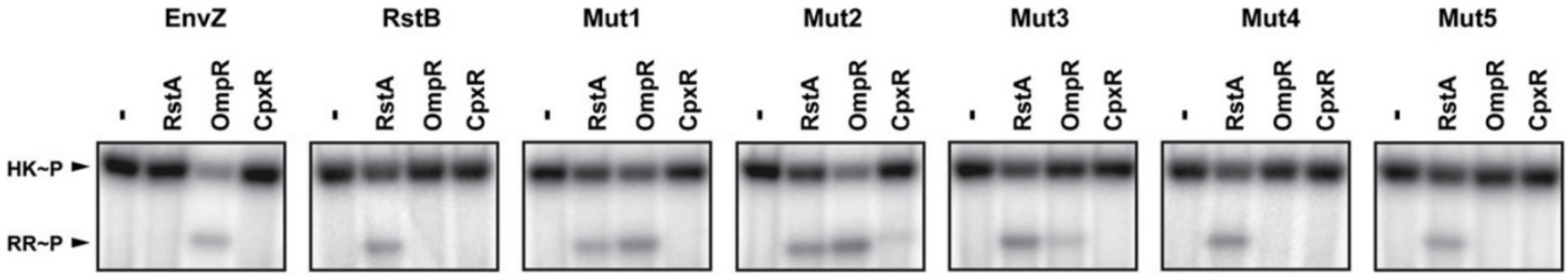
Top: Sequence alignment of the histidine kinases EnvZ, RstB, and CpxA with the histidine phosphotransferase Spo0B

Bottom: Sequence alignment of their cognate response regulators Spo0F, OmpR, RstA, CpxR

Rewiring two-component systems

Skerker et al, 2008

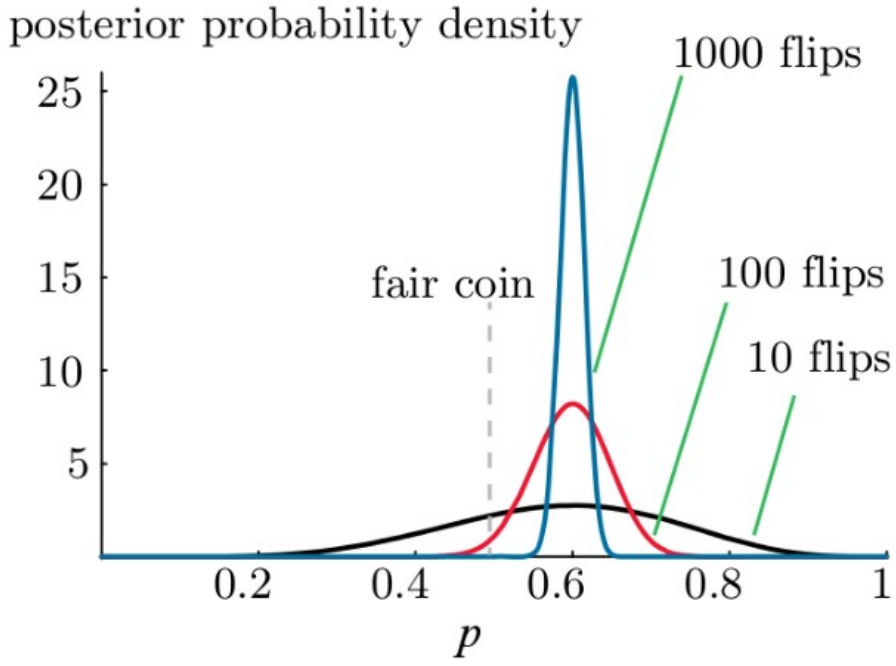




Experimental assay of phosphotransfer specificity of EnvZ, RstB, and mutants Mut1-Mut5
In each case, the kinase was autophosphorylated and then incubated alone or examined for phosphotransfer to RstA, OmpR, and CpxR after 10 s incubations
Black or gray bands = RR-P or HK-P is present (^{32}P, radioactive; electrophoresis + radiography)

Maximum likelihood

- **Likelihood analysis of the bias of a coin**



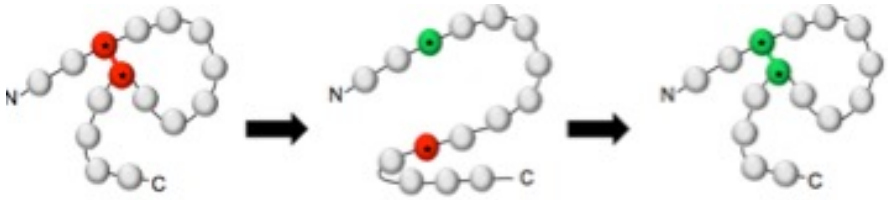
Posterior probability distributions P (model, p | data) for the probability p of getting “heads” upon a flip

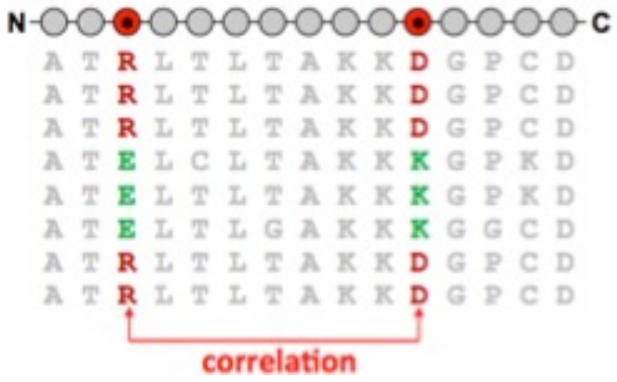
Black is 10 flips, of which 6 were heads; red is 100 flips, of which 60 were heads; blue is 1000 flips, of which 600 were heads

Protein sequence data

- **Inferring structure and function from sequences**

  - Recent data-driven approaches





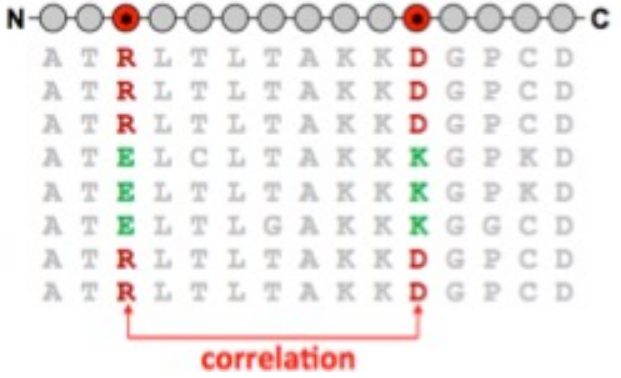
homologs -
a protein family

Evolutionary coupling between interacting residues
→ correlations in multiple sequence alignments inform us about structure and function
BUT... observed correlations can be indirect A ↔ B ↔ C

Maximum entropy model of protein sequence data

- **Goal: joint probability distribution**

  $P(\alpha_1, \alpha_2, \ldots, \alpha_L)$ probability of a sequence in the protein family



- **Observations retained: one- and two-body frequencies**

```
...ISHEL...
...VSHDI...
...VSHEL...
```

$$\rightarrow \begin{cases} f_i(\alpha) & i \in \{1, .., L\} \\ f_{ij}(\alpha, \beta) & \alpha \in \{A_1, .., A_{20}, A_{21} = -\} \end{cases}$$

- **Maximum entropy model consistent with these observations**

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

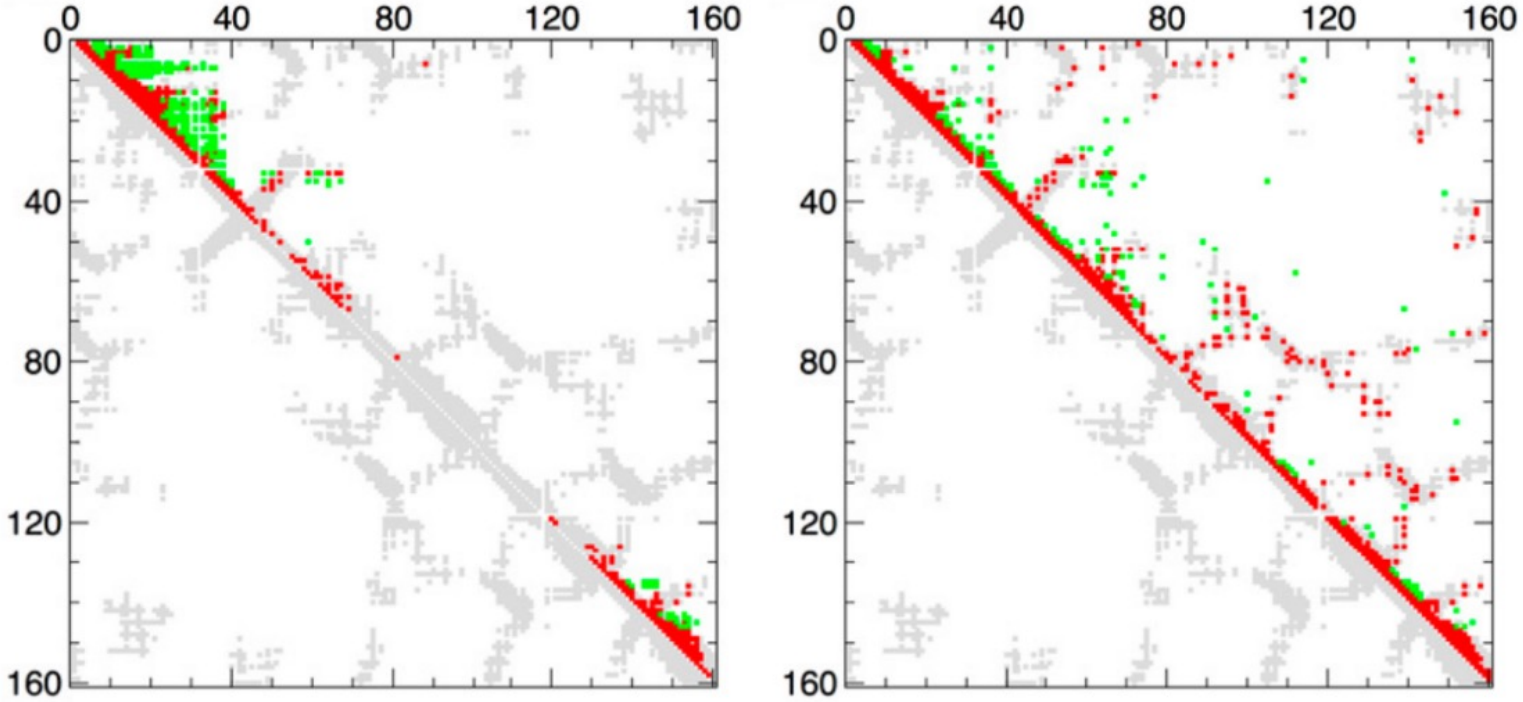
one-body terms - fields

two-body terms - (direct) couplings

Structure prediction

$e_{ij}(\alpha, \beta)$ much better predictor of 3D contact than | $C_{ij}(\alpha, \beta)$ Mutual Information

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



Contact map prediction for the eukaryotic signaling protein Ras

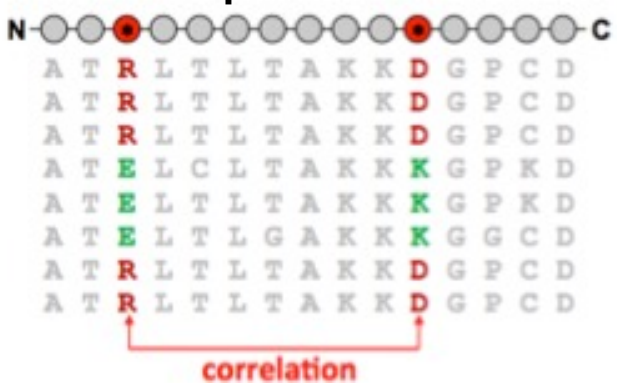
Mutual information (left)
Direct couplings (right)

Morcos, Pagnani et al (2011)

Gray: experimental contacts (cutoff: 8 A)
Red: correct predictions; green: incorrect ones

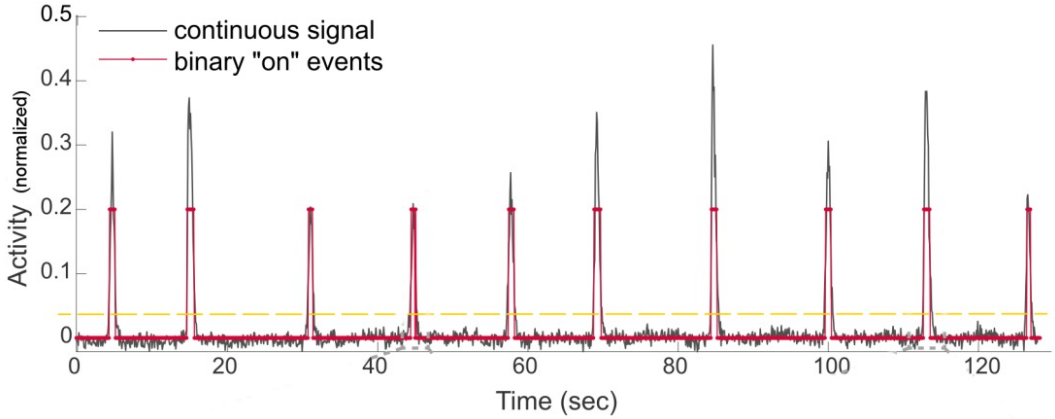
Some applications of maximum entropy modeling

- **Protein sequence data:**



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

- **Neuroscience data:**

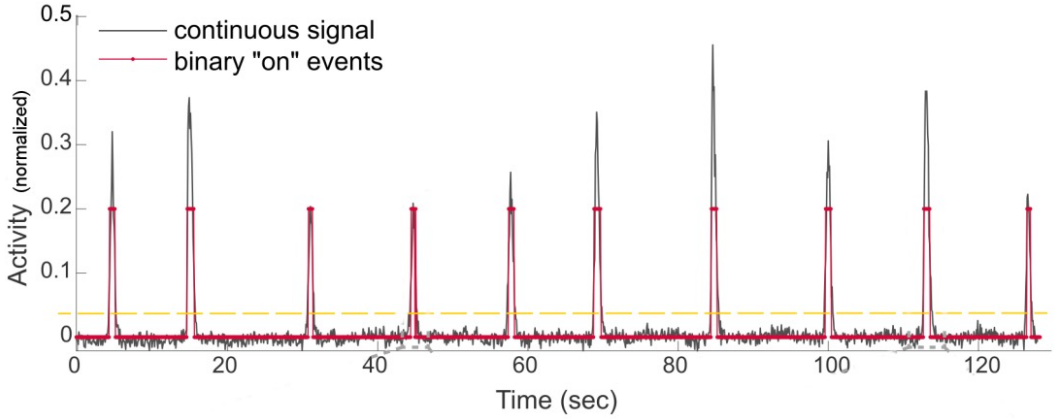


$$P(\{\sigma_i\}) = \frac{1}{Z} \exp[- E(\{\sigma_i\})].$$

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

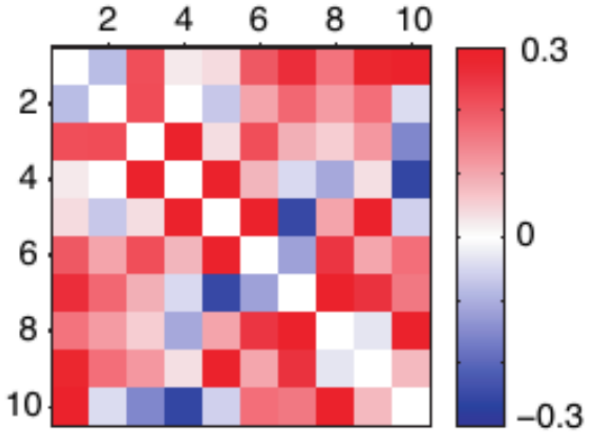
Some applications of maximum entropy modeling

- **Neuroscience data:**



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

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

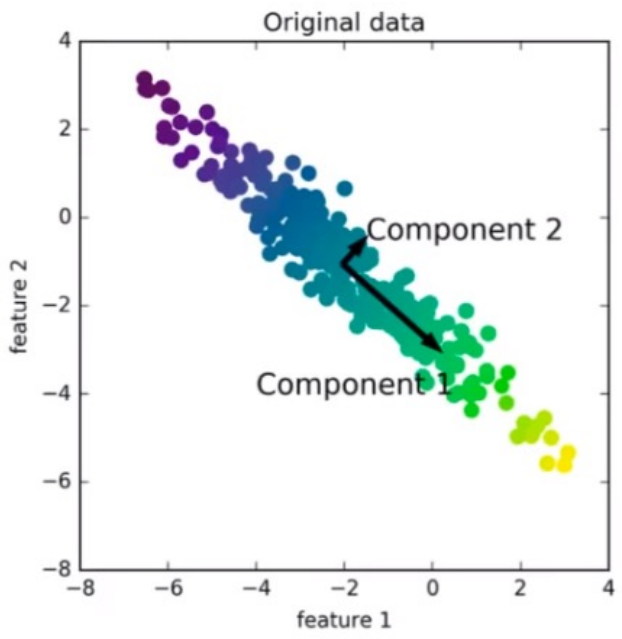


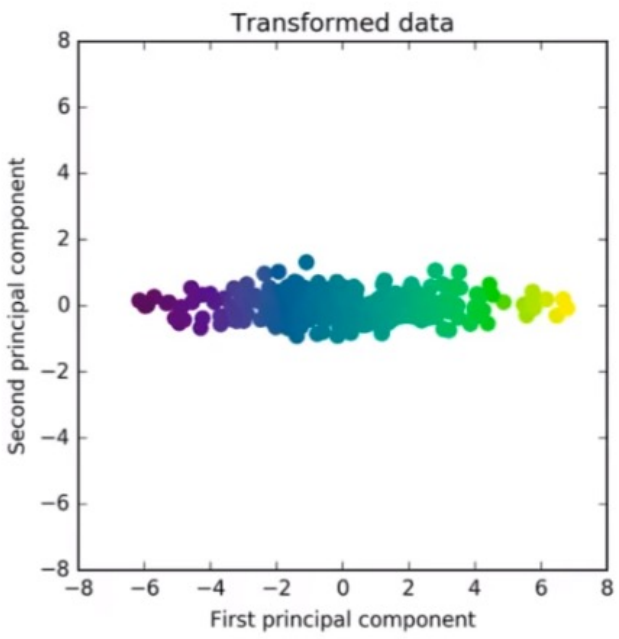
Inferred J_{ij} values for a set of 10 neurons

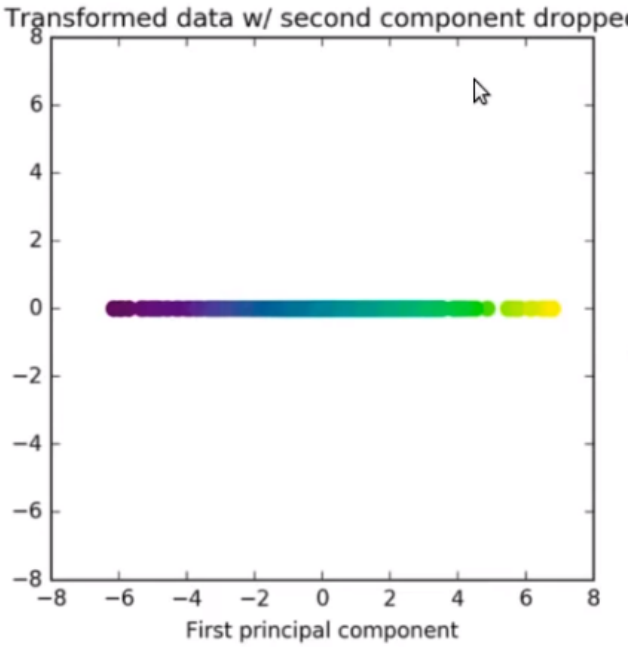
Schneidman et al, 2006

Dimension reduction: principal component analysis

- **Two-dimensional data**

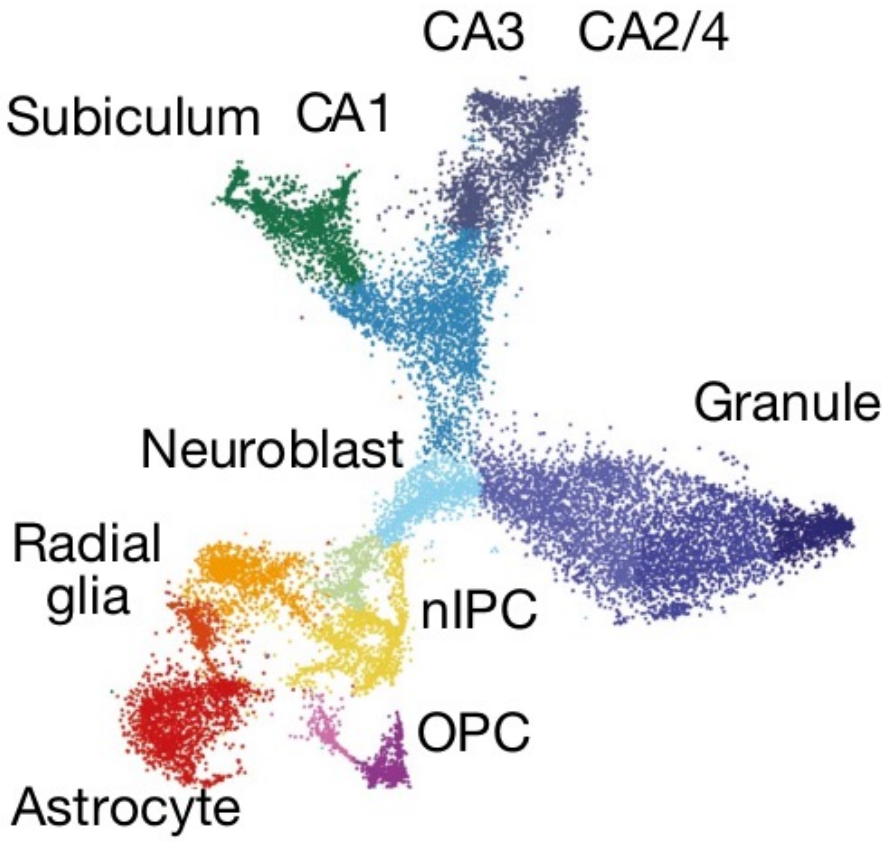






Dimension reduction: nonlinear methods

- **Single-cell RNA sequencing (sc-RNA) and cell types** La Manno et al 2018

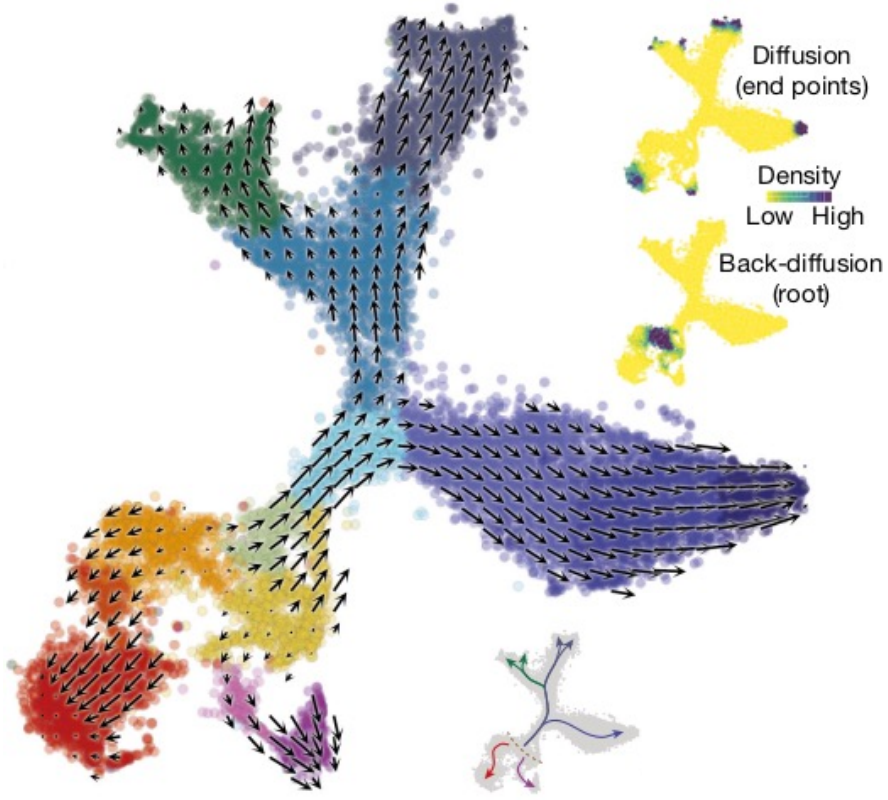


t-SNE plot of developing mouse hippocampus cells (18,213 cells), showing major transient and mature subpopulations

Dimension reduction: nonlinear methods

- **RNA velocity and cell differentiation trajectories**

La Manno et al 2018



Unspliced vs. spliced RNA → time evolution of gene expression: RNA velocity

Velocity field (arrows) projected onto the t-SNE plot
Top inset, differentiation endpoints (mature cell types) and root (progenitor cells)
Bottom inset, summary schematic of the RNA velocity field