Randomness and information in biological data
BIO-369

Prof. Anne-Florence Bitbol

EPFL

Lecture 8

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

Reminder: Motivation

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?

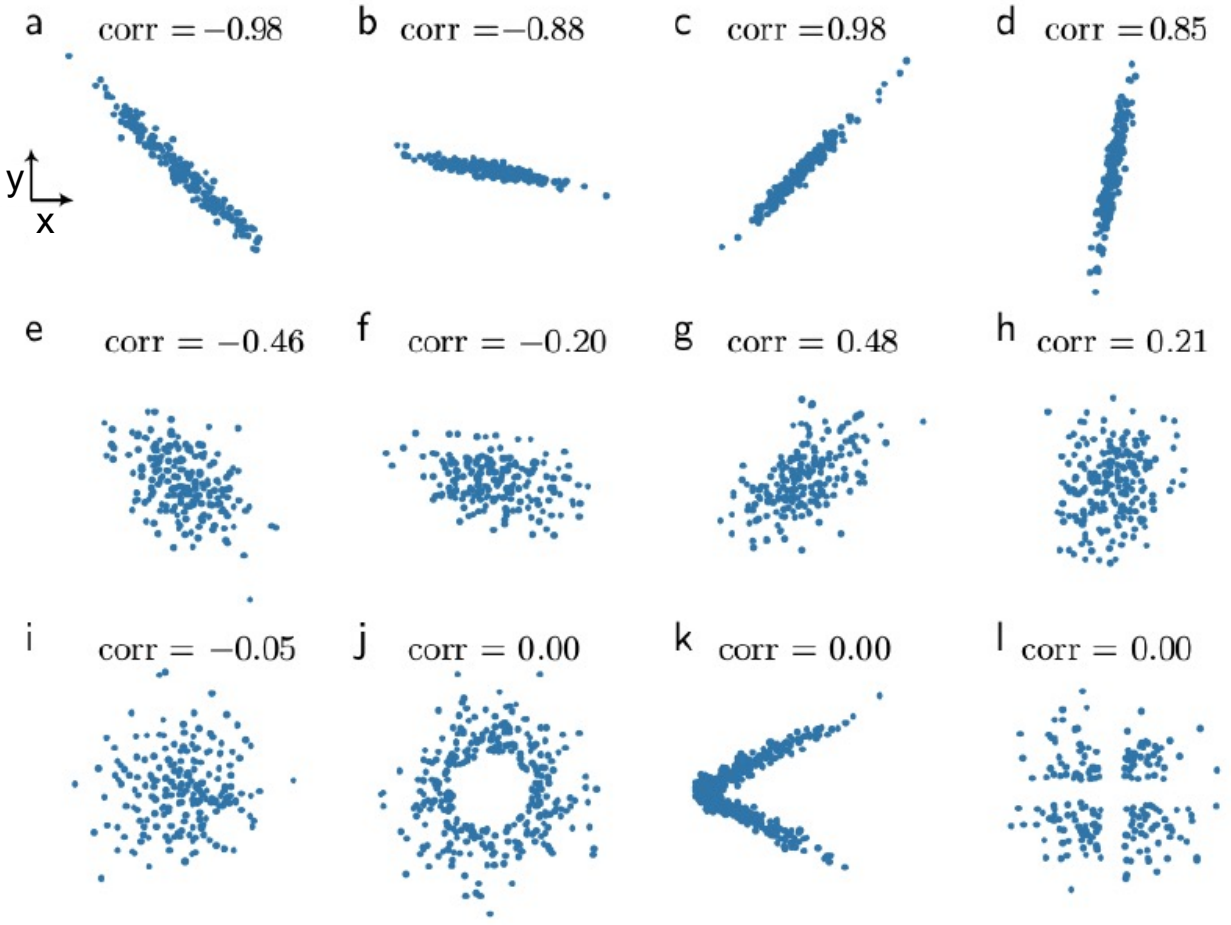
Consider two random variables X and Y with zero covariance. Are they are statistically independent?

A. Yes
B. No
C. It depends, they could be statistically independent or not

To answer, please:

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

Correlation coefficients



Correlation coefficients between some example random variables X and Y

Different draws are performed, yielding values x and y, and the correlation is estimated from this data

In each case, the correlation coefficient was estimated from 5000 draws of both random variables, but only the first 200 are shown.

What is the sum over all y of P(x,y) equal to (for a given x)?

0% A. P(x)

0% B. P(y)

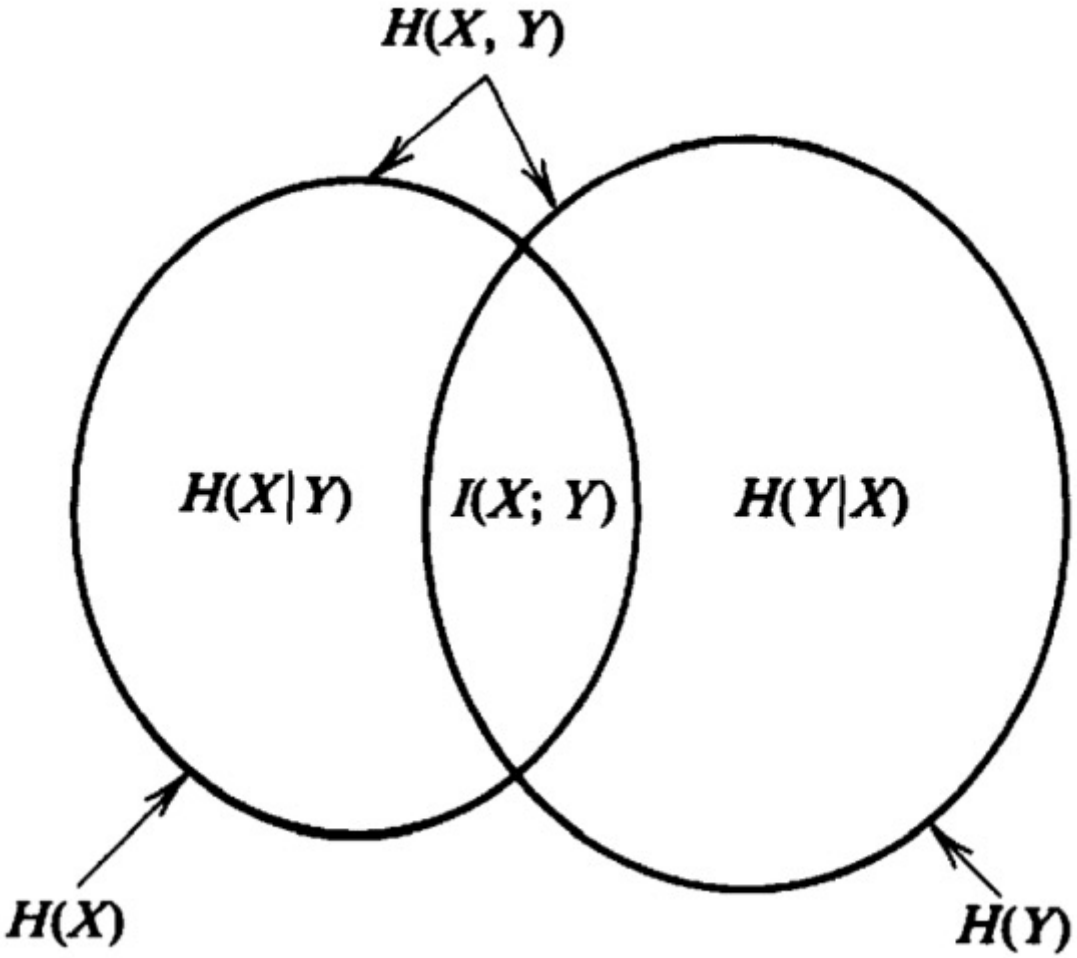
0% C. P(x|y)

0% D. 1

To answer, please:

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

Mutual information



Relationship between entropy, joint entropy, conditional entropy and mutual information

The relationship is illustrated schematically by ensemble unions and intersections, meaning that
$H(X) = H(X|Y) + I(X, Y)$

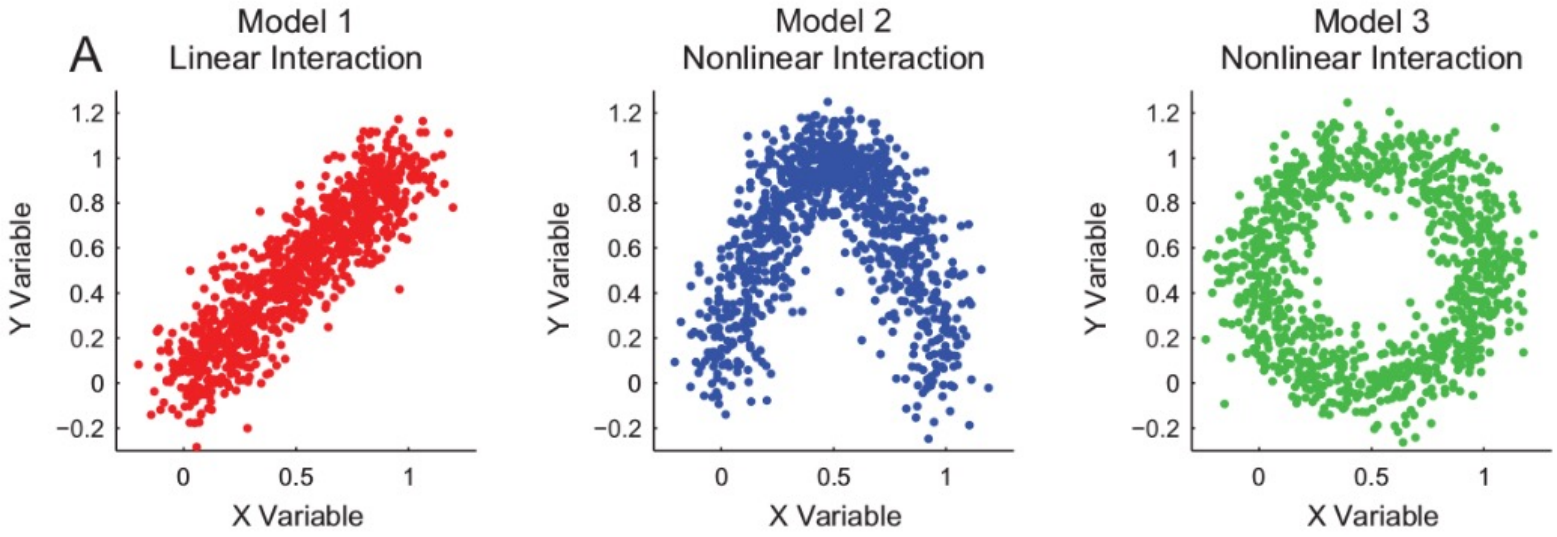
What can we say about the mutual information between X and Y? Is it the amount of uncertainty (i.e. missing information)…

A. Remaining about y after x is known
B. Remaining about x after y is known
C. About x that is removed after learning y
D. About y that is removed after learning x

To answer, please:

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

Mutual information



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

What do you expect for the mutual information between X and Y for models 1, 2 and 3?

0% A. Substantially larger than 0 in model 1 but close to 0 in models 2 and 3

0% B. Substantially larger than 0 in models 1 and 2 but close to 0 in model 3

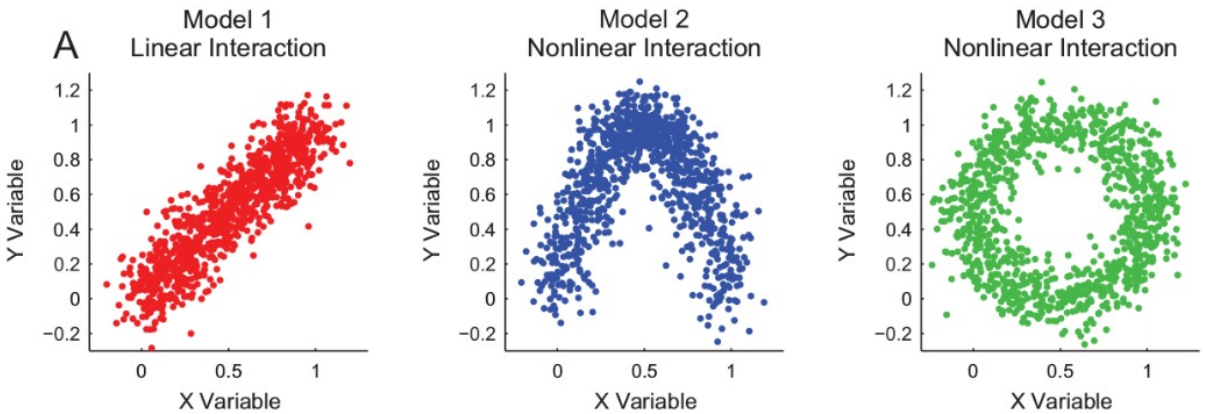
0% C. Substantially larger than 0 in models 1 and 3 but close to 0 in model 2

0% D. Substantially larger than 0 in models 1, 2 and 3

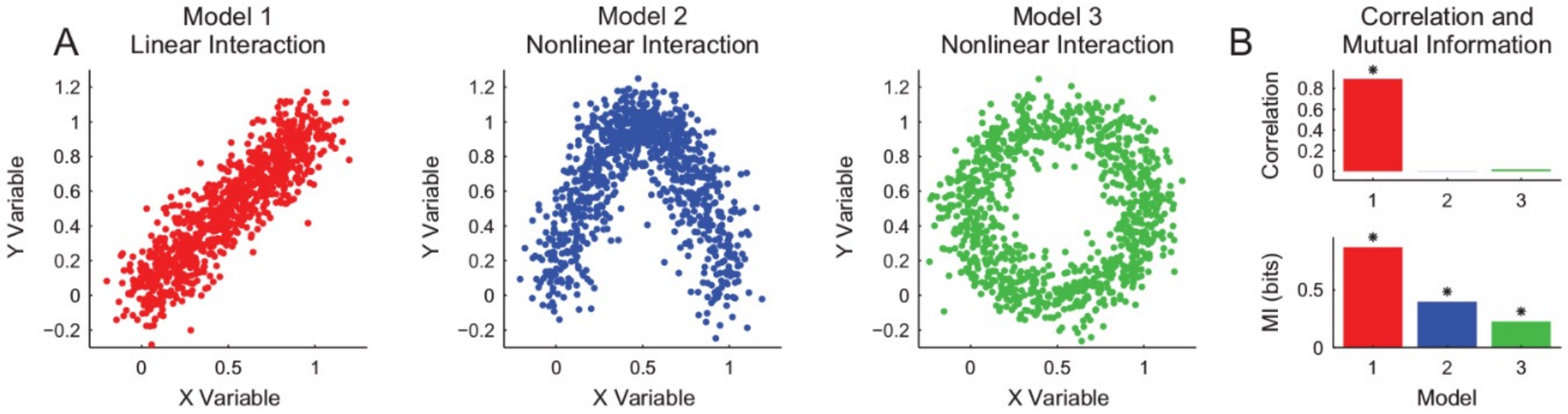
0% E. Close to 0 in models 1, 2 and 3

To answer, please:

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



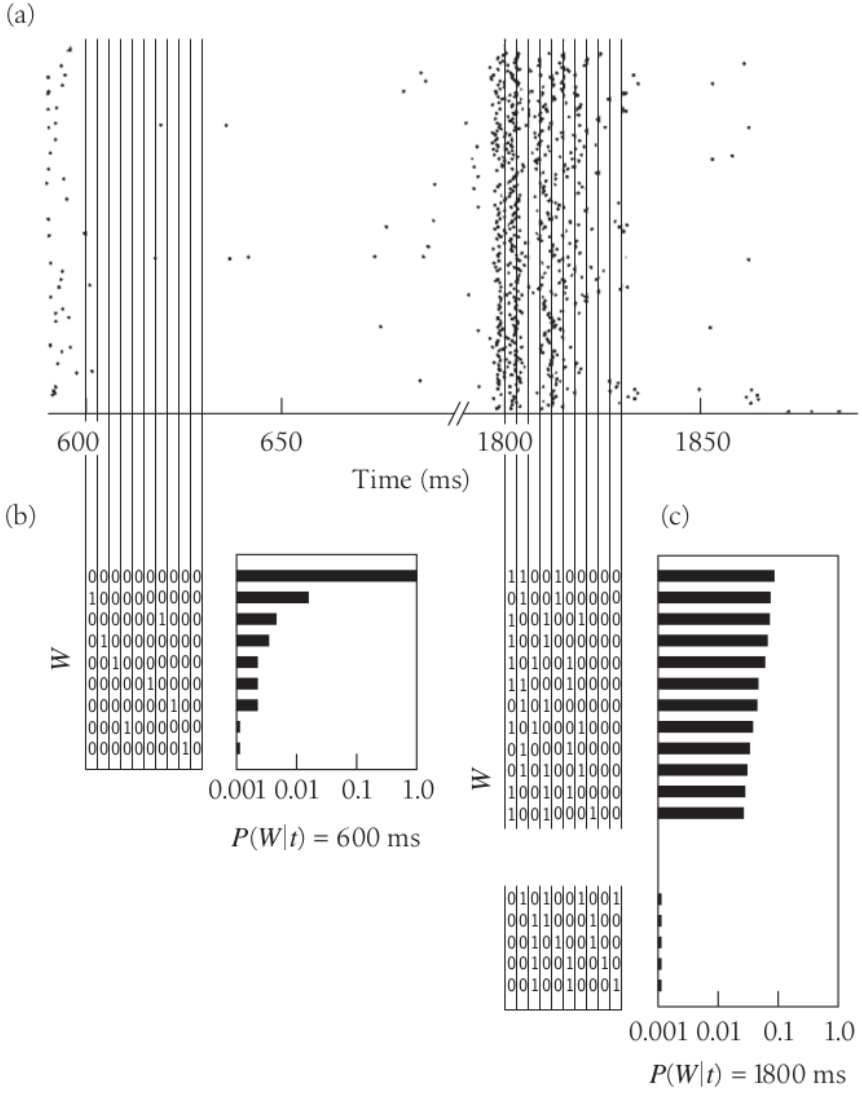
Mutual information



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

Reminder: 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)

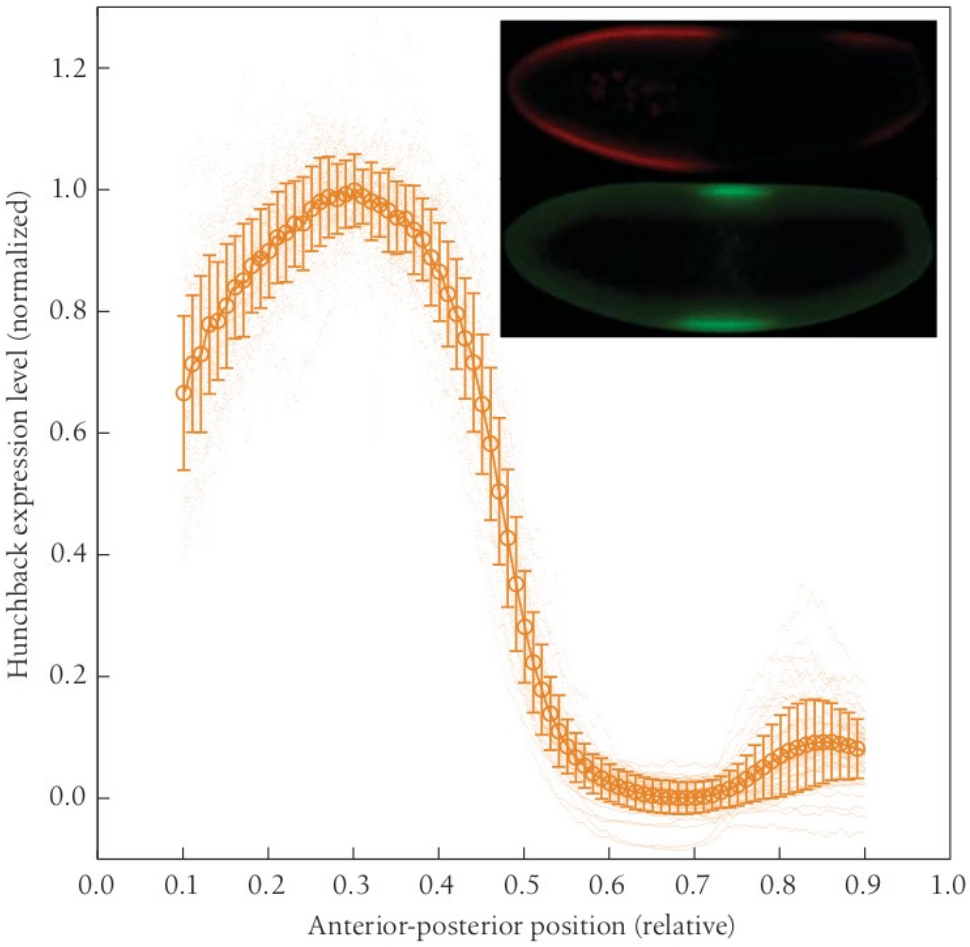
The distribution of words that occur at a particular moment in the movie, P(W|t), is shown for t = 600 ms and t = 1800 ms

$I(W;T) = H(W) - H(W|T)$

$I(W;T) = H(W) - \langle H(W|t) \rangle_t$

de Ruyter et al, 1997

Positional information in development



Inset: early *Drosophila* embryo
Red: fluorescent antibody staining for Hunchback; green: same for Krüppel (inhibited by high levels of Hunchback)

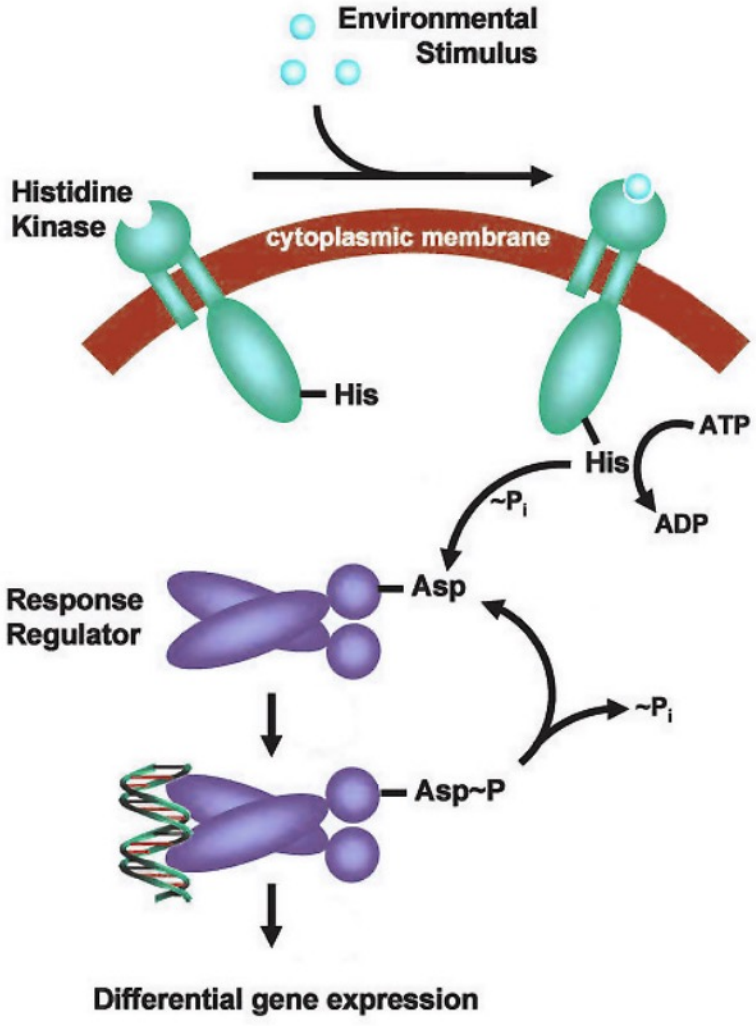
How much information do Hunchback levels carry about position?

Main panel: spatial profiles of Hunchback expression
Mean and standard deviation across 51 embryos
→ P(g|x) (Gaussian approximation)
→ $I(G;X) = H(G)-H(G|X) = H(G)-\langle H(G|x)\rangle_x$

I(G;X) = 2.3 bits (lower bound)
→ Gap genes do more than specifying on/off boundaries

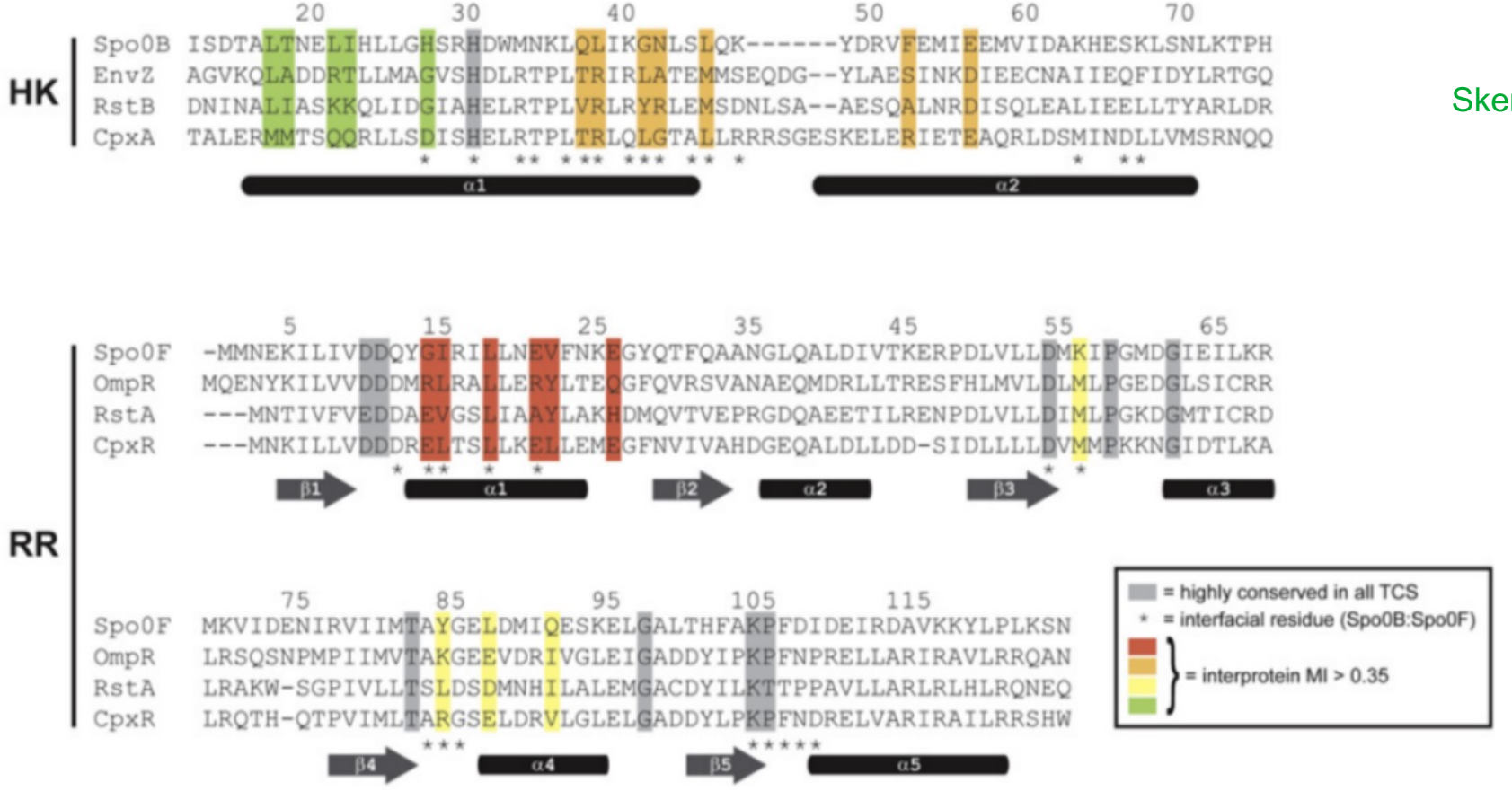
Dubuis et al, 2012

Two-component signaling systems



- Schematic of signal transduction in bacterial two-component systems
- The response regulator is assumed to be a transcription factor
- Many (~dozens) of such system in each bacterial genome, each responding to a different stimulus
- High specificity

High interprotein mutual information sites in two-component systems

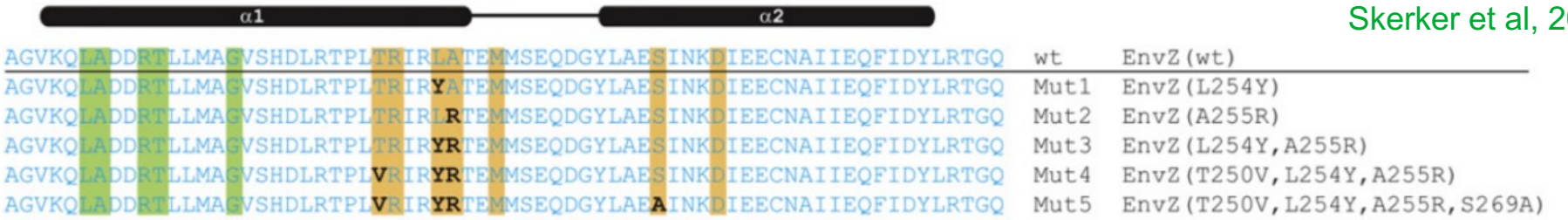


Skerker et al, 2008

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

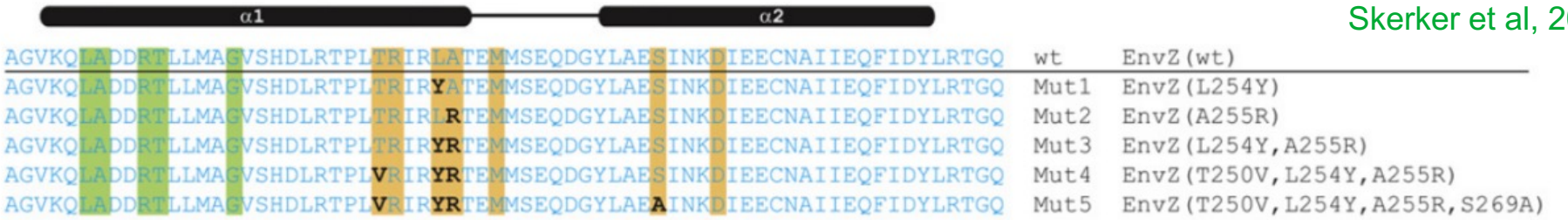
Sequences of point mutants of HKs constructed in the study (mutations in black)

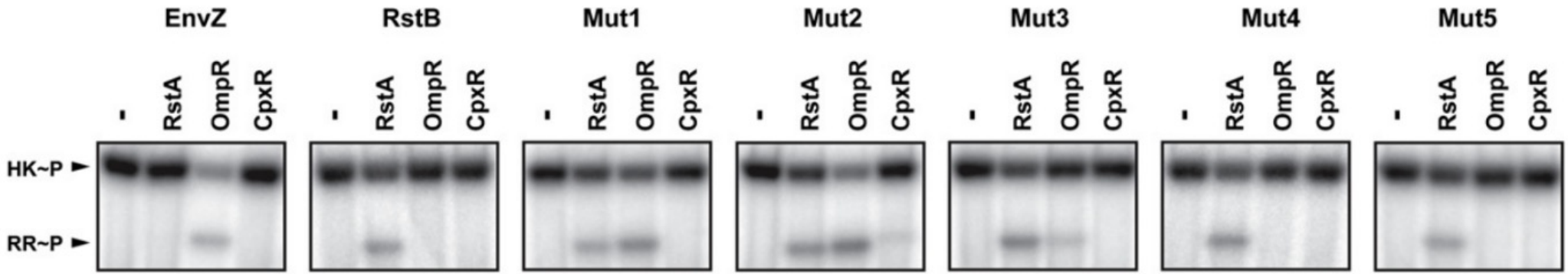
Amino acids with high interprotein mutual information scores were mutated in the HK EnvZ to match the corresponding amino acid in the other HK RstB

→ Do these mutations change the specificity of the interaction and of the phosphotransfer?

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)