Theoretical physics in biology

Curtis G. Callan, Jr.
Physics Department,
Princeton University

Mike Cornwall's career is an example of the power of the theoretical physics way of thinking to illuminate problems across the physical sciences and on into technology. One area that I believe he has left untouched is biology. I would argue that the study of living matter needs the attention of people who think like Mike. I will make some general remarks about why this might be so and then walk you through a problem where mathematical principles lead to interesting insights into an important area of biology.

Broad Subject of This Talk

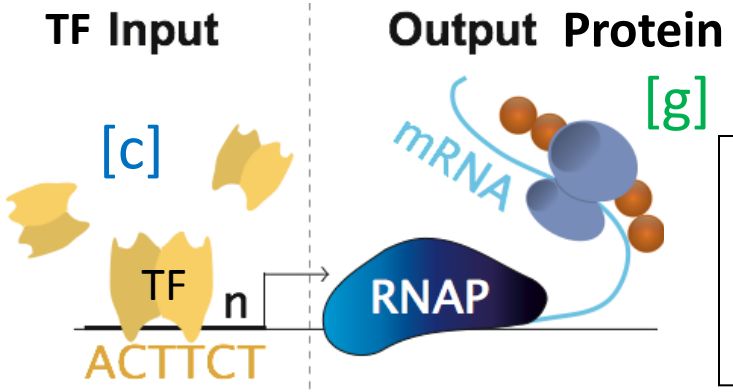
- How biology is constrained by physical principles has always been a legitimate subject of study for theoretical physics. The recent explosion of quantitative data, due to whole genome sequencing, expression profiling, etc. has placed this question in a qualitatively new context -- one in which the making of quantitative models, and even theories, will be essential for understanding biological data. This provides an opportunity for theoretical physics (and physicists) to contribute to the advance of biology.
- Gene regulation presents an instructive particular case. Each cell in our body contains the same genetic information, coded in a few DNA molecules. Each cell controls how much of each protein is made via the binding of special transcription factor proteins (TFs) to short segments of DNA lying upstream of protein-coding regions (genes). Despite the relentless advance of molecular biology over the last fifty years, deep physical questions about the working of this mechanism (in particular regarding specificity, kinetics and noise) remain imperfectly understood.
- This area is a rich source of problems for theoretical physicists who see living matter as a physics challenge. We'll talk about a couple of them.

Some Motivation

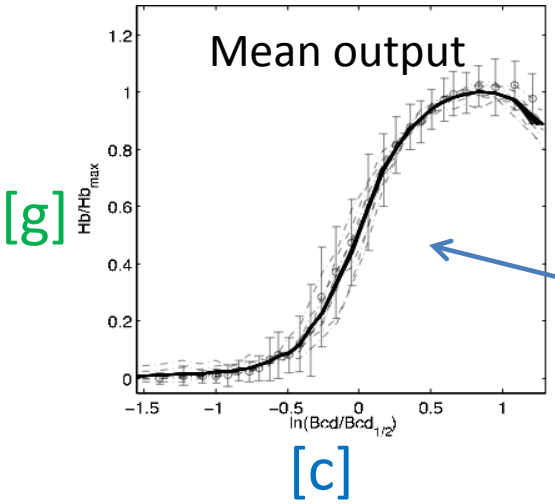
- How biology is constrained by physical principles has always been a legitimate subject of study for theoretical physics (think of Helmholtz).
- One can argue that biology is reaching the point where theory, as we know it in physics, will become essential for understanding.
- We can identify a number of physics issues which cut across the usual biological divisions of kingdom and function (from bacteria to brains)
- Using such insights, we can attempt to develop general principles which have quantitative predictive power for real biological systems.
- In this talk, I will take you through a worked example of what I mean: using information theory to understand the development of the fly embryo.
- It used to be said that there is no theory in biology … the day may be coming when there is no biology without theory.

Overview of Gene Regulation

Transcription factor concentration



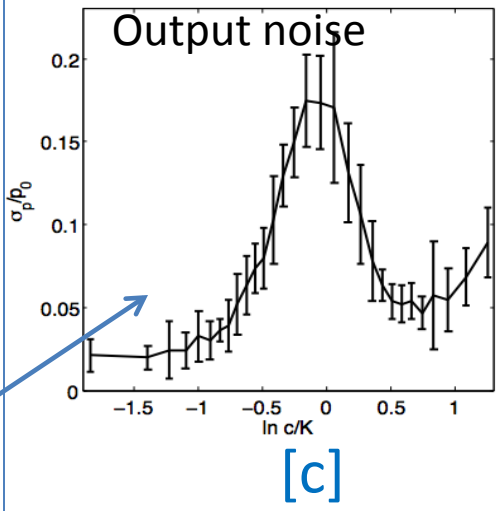
Regulated gene expression



Special DNA site occupied by TF with thermodynamic oc-cupancy *n* driven by TF conc'n [c] and DNA binding energy ε

Occupancy *n* enhances RNAP binding, drives transcription. Typical on/off mean response.

But, diffusion and shot noise cause output variance



Noise in Gene Regulation (and signaling)

- If sensors are of size a, and molecules are at concentration c, the sensor can try to "count" $N \approx ca^3$ molecules (it measures the local concentration)
- Because of statistics, relative errors $\approx 1/N^{1/2}$ are inevitable
- Averaging for a time τ, we can make $K \approx \tau / T$ measurements, where $T \approx a^2 / D$ is time to "clear" sensor volume, we reduce the fractional error by $1/K^{1/2}$
- Result is a hard limit on the precision of concentration sensing:

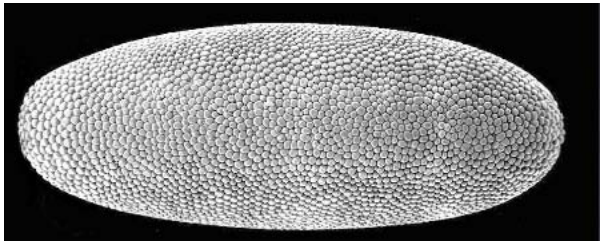
$$\frac{\delta c}{c} \sim \frac{1}{(\pi c D a \tau)^{1/2}}$$

This old argument of Berg+Purcell can be made rigorous in new contexts (Bialek+Setayeshgar)

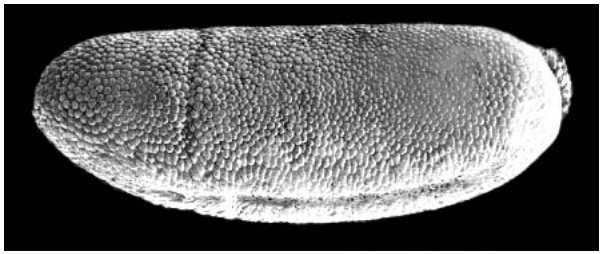
- Keep this in mind as we talk now about TF molecules (sometimes just a few per cell) turn genes on and off to make cell fate decisions in an embryo.

How does a fruit fly get its parts?

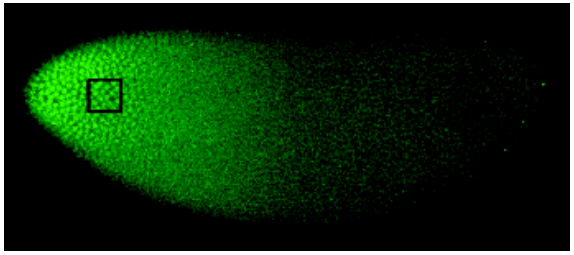
Nuclei divide over and over again within the egg shell. At some point, they undergo choreographed motion to make body parts. Each nucleus knows its place …. how?



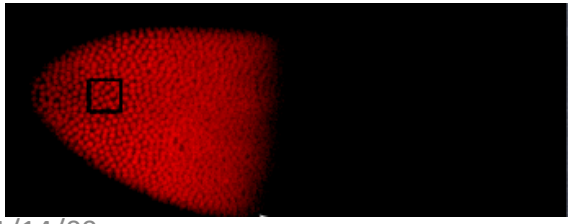
Up to a point, cells divide in place. Then they move to form differentiated organs.

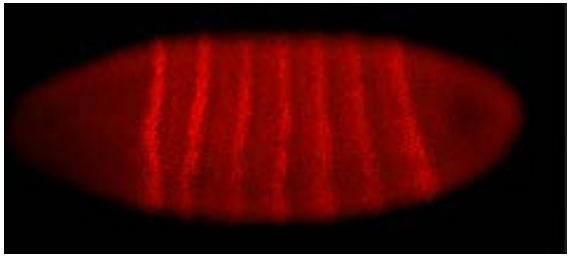


Transcription Factors turn genes on and off. Development is a cascade of TFs turning on other TFs. Staining for particular TFs, see patterns with sharp location dependence.

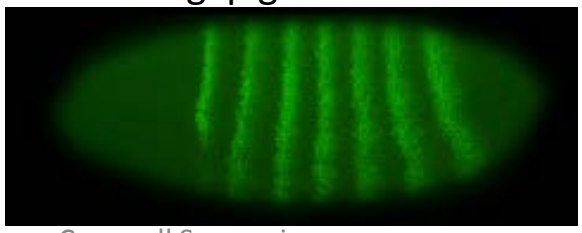


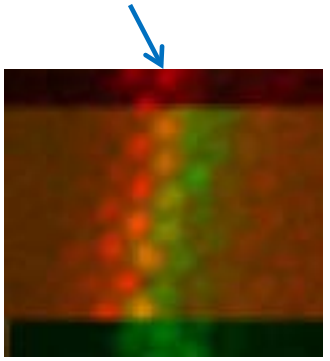
Early: Bicoid → Hunchback





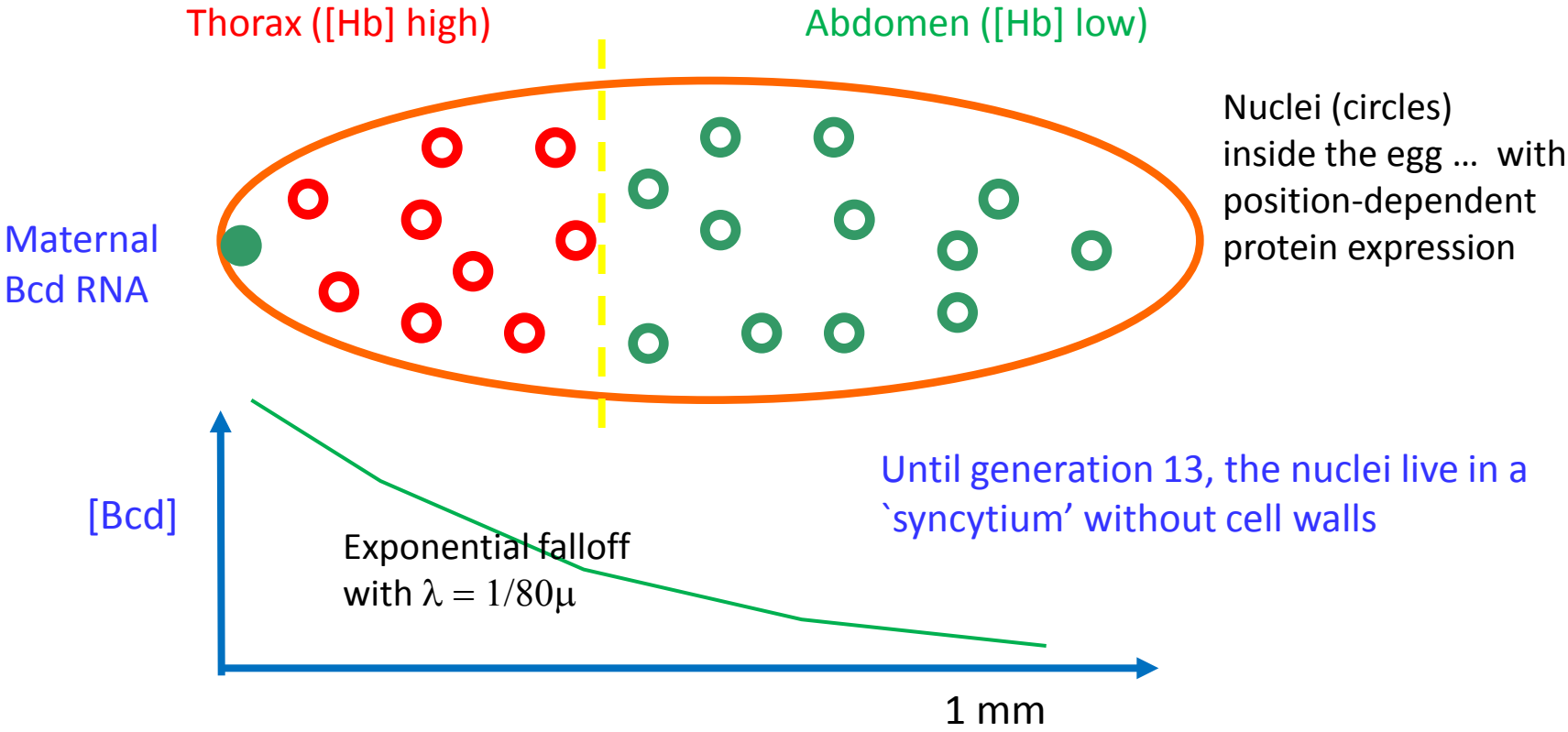
Later: "gap genes" interact





Neighboring cells have different fates … how? Focus on first step …

Cell Fate in Fruit Fly Embryogenesis



Nuclei need to know `where they are' in order to make cell fate choices (thorax v. abdomen). The concentration of Bicoid protein diffusing from the head of the embryo is the signal: if [Bcd] is big enough, a nucleus expresses Hb (red), if [Bcd] too small, Hb is not expressed (green). This pattern repeats in a TF cascade (Bcd → Hb → Kr, Gt, Kn, ..) leading to the expression "stripes" needed to make a multi-segment body. Etc.

Accuracy of Pattern Formation?

- Maternal Bcd a has longitudinal gradient: Hb transcription reads out [Bcd]. Roughly, it turns on when threshold concentration is exceeded.
- Hb expression threshold is sharp, separating neighboring nuclei ($\delta x \approx 6\mu m$ out of 1mm egg length). ***Readout accuracy of [Bcd] must be ≈ 10%.***
- But readout by transcription factor (TF) binding to small DNA sites has an accuracy limited by $N^{1/2}$ fluctuations, TF Brownian motion (Berg/Purcell)
  - ***For a≈3nm, D ≈ 5μm^2/s, [c] ≈ 2nM, T ≈ 60s limit is $\delta c/c > 50\%$!***
- Noise is reduced by time averaging. But, to achieve 10% accuracy, promoter would have to average for 1 hour (several cell division times)
  - Perhaps [Bcd] is sensed as a *spatial* average? If so [Hb] fluctuations in nearby nuclei must be *correlated*. They are … but mechanism for this is not known.`
- Clearly, *Drosophila* is operating near statistical mechanical limits. To get more insight, we should measure TF input/output relations for the nuclei.

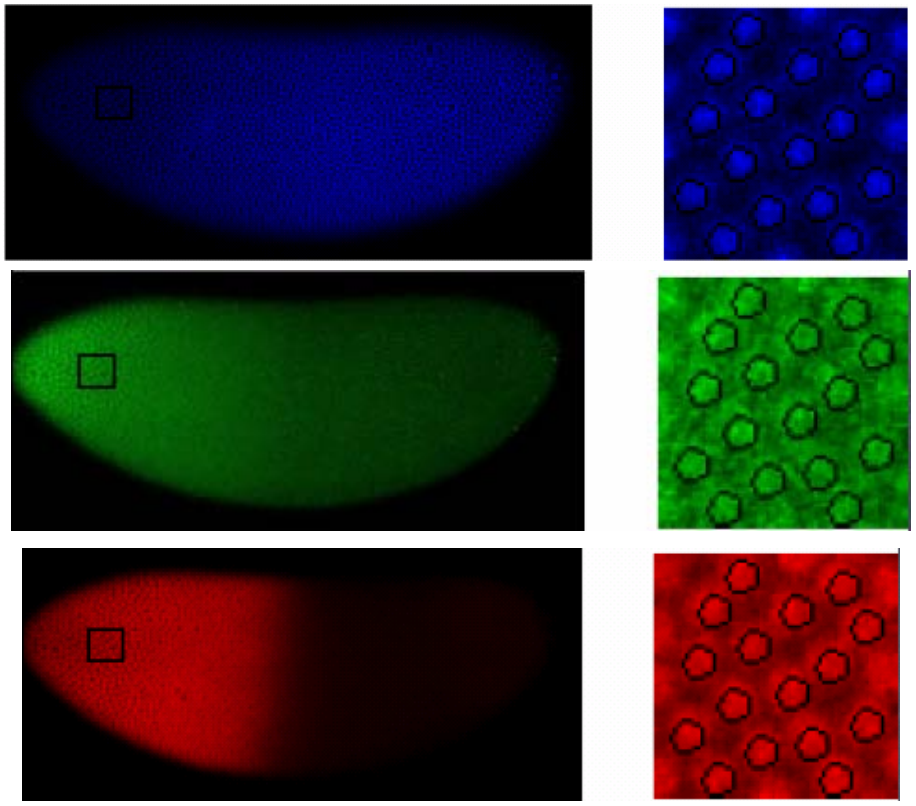
Observations on Triply Immuno-stained Drosophila Embryos (Gregor et al):

Measure [Bcd] and [Hb] in each and every nucleus of cycle 14 embryo with ~2300 nuclei:

DNA

Bcd

Hb



Positions of nuclei in embryo identified by DNA signal (blue)

Observe diffusion gradient of primary morphogen [Bcd]: above some threshold, it turns on [Hb]

Infer [Hb] and [Bcd] within each nucleus from R/G signals

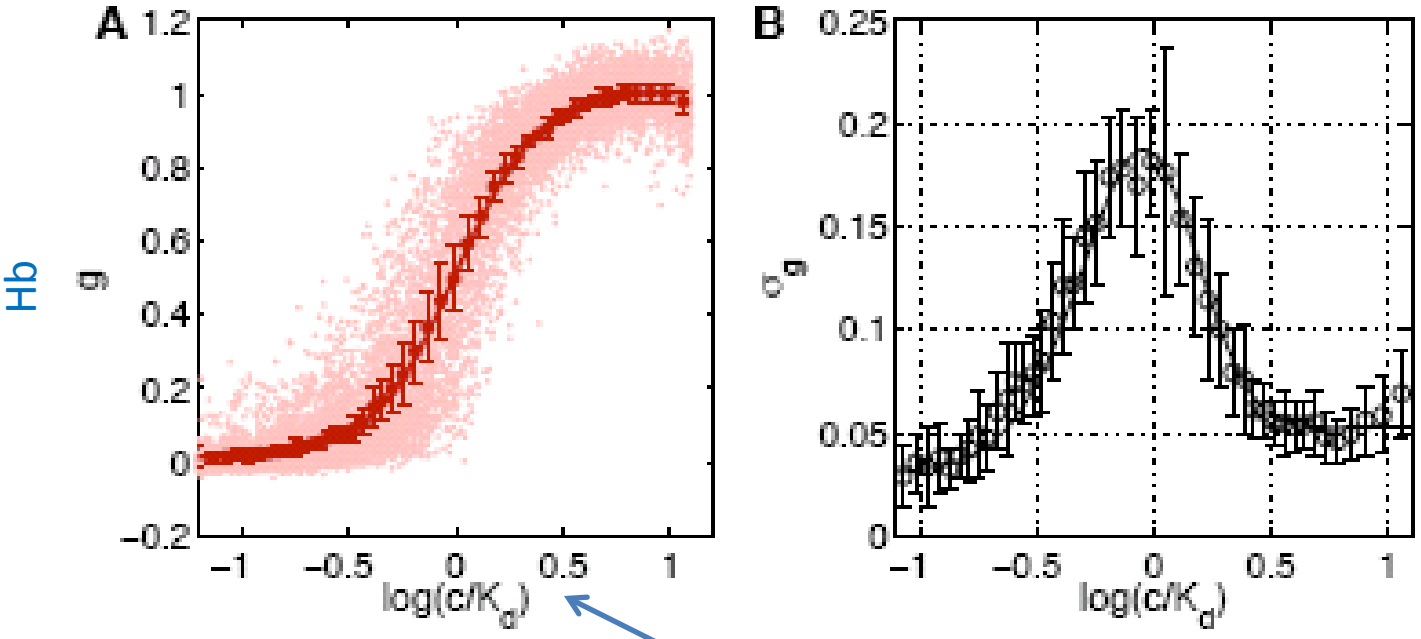
10^4s of nuclear data points give the input/output relation for expression control system:

Powerful tool for studying issues of noise and fluct-uations in gene expression

Expression PDF from Measurements

Input vs output

Noise vs input



Normalized noise in readout of [g=Hb] from input [c=Bcd] . Note approx 15% accuracy in decision region. It would be hard to do better than this because of small-number fluctuations in the [Bcd] molecules arriving at the genomic site where expression of Hb is controlled.

Mutual information: a primer

Can a noisy switch do more than just be "on" or "off"? How many bits can it set? Mutual information between output [g] and input [c] is the best measure:

Shannon entropy of random variable g with probdist $p(g)$:

$$S[p(g)] = -\int dg\, p(g) \log_2 p(g)$$

Mutual information of variables g,c with joint probdist $p(c;g)$:

$$I(c;g) = \int dc\, dg\, p(c;g) \log \frac{p(c;g)}{p(c)p(g)} = S[p(g)] - < S[p(g|c)] >_c$$

Model-independent and unbiased way to quantify how two "related" variables inform abt each other … central quantity in comms & neuroscience. In cellular biology, cells in environment [c] respond with output [g] … in the embryo, "where" is [c] and "what" is [g] … MI constrains how finely gene regulation can "divide up" the embryonic real estate (stripes in fly embryo)

Mutual Information in the Genetic Switch

Rewrite MI in an alternate form that organizes things in a convenient:

$$I(c;g) = \int dc\, P_{TF}(c) \int dg\, P(g|c) \log_2 \left[\frac{P(g|c)}{P_{\exp}(g)} \right]$$

Positive, measured in "bits". 0 if (g,c) uncorrelated; 1 if two levels can be sensed, ...

Input/output function *P(g|c)* reflects the physics of TF binding and transcription. Take it as a Gaussian prob distn, defined by its measured mean and variance (analytic convenience)

$$P(g|c) = \frac{1}{\sqrt{2\pi\sigma_g^2(c)}} \exp\left\{ - \frac{[g - \bar{g}(c)]^2}{2\sigma_g^2(c)} \right\}$$

$P_{TF}(c)$ = "environmental" TF dist'n

$$P_{\exp}(g) = \int dc\, P(g|c) P_{TF}(c)$$

The quantities needed to evaluate *P(g|c)* have all been measured. No need to know *why* it has the form it does (for now) …,

The role of *p(c)* -- analogy with neuroscience

Sensory coding: tune the 'filter' *p(g|c)* so that the information transmission is maximized under the 'natural' ensemble of stimuli

$$p(\text{response}, \text{stimulus}) = p(\text{response}|\text{stimulus})p(\text{stimulus})$$

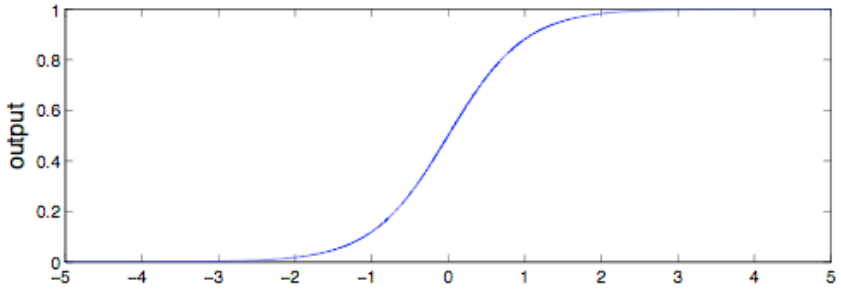
joint response *organism* *environment*

Transcriptional control: tune the 'input distribution' so that information transmission is maximized given the 'input-output relation'

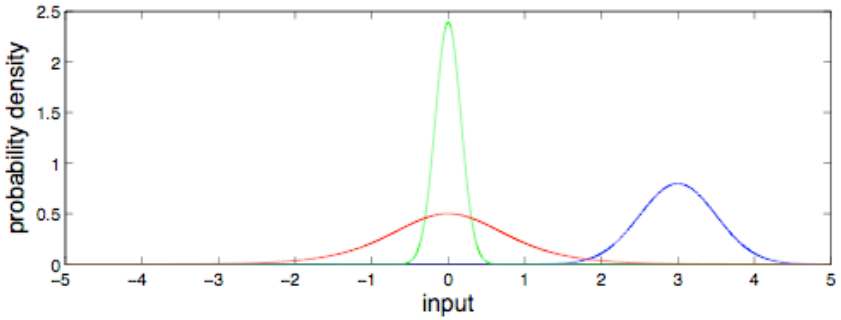
$$p(g, c) = p(g|c)\, p(c)$$

In neural "adaptation", the organism can control the mutual information by varying the response function, using learned info about the stimulus. There is real evidence for information maximization. The cell is clearly different .. but we can explore the question of info maximization … varying *p(c)* rather than *p(g|c)* .

Intuitive view of input optimization:



Is this typical input/output relation "good" for MI?



It depends on the context. Look at different input distributions…

If inputs are chosen from the blue distributions, then the output is always fully on, and so isn't related to the input.
If the inputs are chosen from the green distribution, then the output is stuck in its mid-range, and doesn't get modulated much.
The red distribution seems well matched: typical variations in the input are enough to modulate the output through its full dynamic range

Information maximization for real

The input and output distributions $P_{TF}(c)$ and $P_{exp}(g)$ are related by fixed $P(g|c)$. Take $I(c,g)$ to be a functional of $P_{exg}(c)$ and maximize it. Gives max mutual info for the regulated gene *and* identifies the distributions of input c and output g that achieve this optimum.

Bialek, CGC, Tkacik (08)

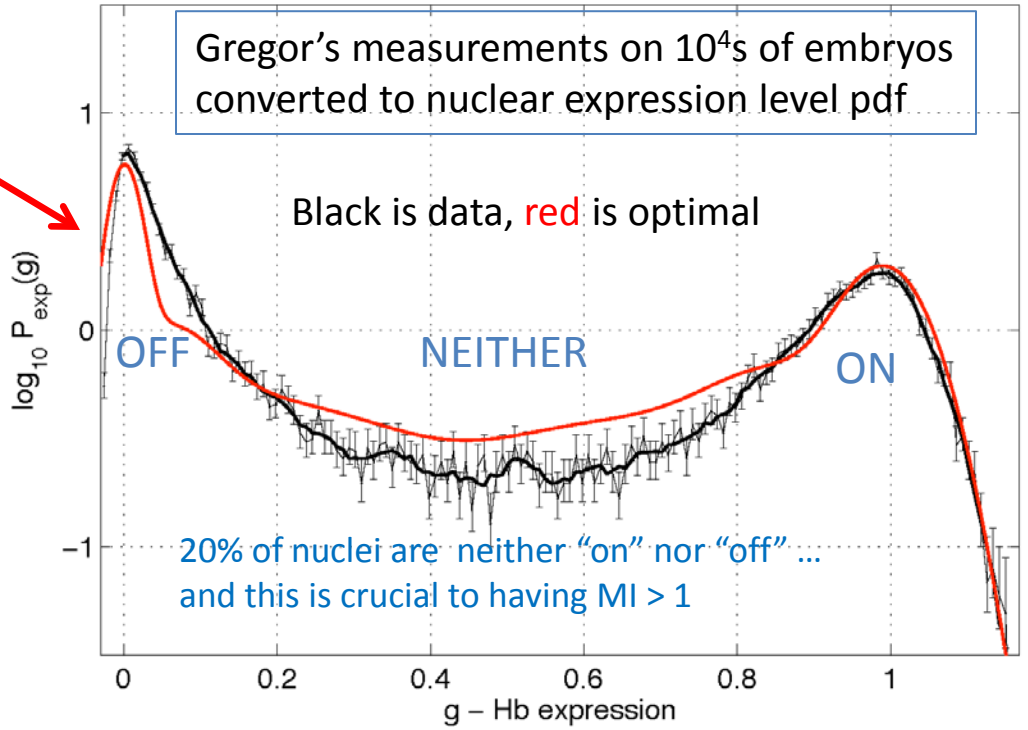
The variational solution is simple:

$$\hat{P}^*_{\rm exp}(\bar{g}) = \frac{1}{Z}\cdot\frac{1}{\sigma_g(\bar{g})}$$

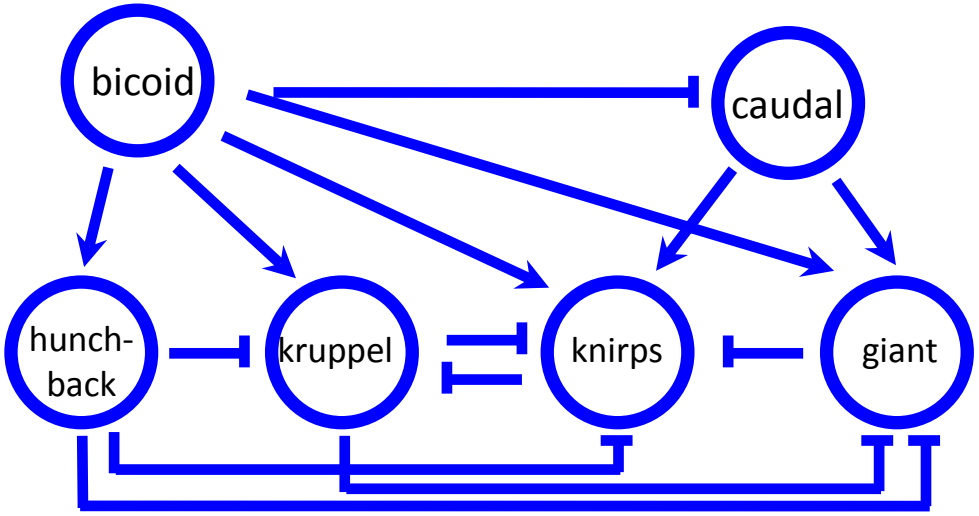
Results for MI are impressive:
I_{opt}=1.7 bits, I_{data}=1.5 bits

Optimal distribution matches observed distribution of [Bcd]

Doing better than 1 bit is not easy: the cell has to be fine-tuned. Why?



Theoretical challenge: role of information in real gap gene network



Gap gene network in D. melanogaster

Expression levels of {hb,kr,kn,gt} could be a code for "where" on the ant-post axis.

Nuclei need to "locate" about 100 rows … which requires 6.6 bits. At 1.5 bits (or so) per gene, we need 4 readout genes … exactly the number of gap genes.

Suggests that information (and constraints on it) plays a central role in development

Theory problem: Derive the networks - with all the numbers on the arrows! - maximize information transmission subject to physical constraints (number of molecules).

Theory/experiment collaboration: Measure information transmission; measure the structure of real networks.

Infomax and gene network structure?

- This encourages us to think about using MI maximization as a principle to understand gene regulatory networks more generally. Instead of $c \to g$, we might have $c \to \{g_1, \ldots, g_n\}$ (or worse). MI between *in* and *out* variables still meaningful.
- More excitingly, physics allows us to parametrize $P(g|c)$ .. not take it as an input!

$$\bar{g}_i(c) = \frac{c^n}{c^n + K_i^n}$$

Hill function … phenomenological

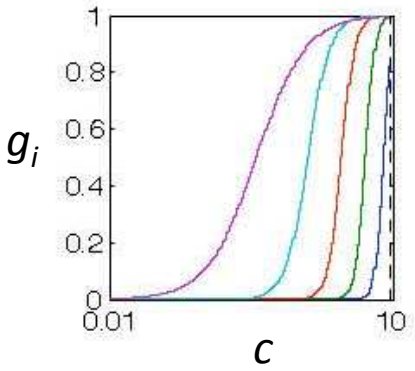
$$\sigma_g^2(c) = \frac{c}{D a \tau}\left[\frac{d\bar{g}(c)}{dc}\right]^2 + \frac{1}{N_g}\bar{g}(c)$$

Input noise from diffusion

Output noise from Poisson statistics

- Thus we can express the *MI* between c and $\{g_1, \ldots, g_n\}$ as a function of the "environmental" distribution $P_{TF}(c)$ *and* the parameters which determine $P(g|c)$. Now optimize over parameters (K_i, n) *as well as* the environment:
  - Single input c independently drives genes $g_i(c)$
  - Maximize MI with a constraint on the total number of molecules available to the system.
  - Non-obvious, biologically reasonable answer ….
  - What happens when we allow feedback loops?

Tkacik, Walczak, Bialek (08)



Remarkable that infomax imposes fine structure on a simple feed-fwd network …..

Many Opportunities for Theory in Biology

- Small number fluctuations impose limits on precision of biochemical signaling:
  - Bacteria make decisions on basis of small changes in signal molecule concentration.
  - What is the "information capacity" of gene regulatory elements (on/off or better?)
  - How do we understand accurate pattern formation in embryogenesis (fruit fly studies)
- Statistical inference (in the sense of particle physics and cosmology) is needed to exploit the avalanche of "high-throughput" experimental data
  - Microchip expression array data and "deep sequencing" provide masses of noisy data
  - Theory is needed to extract precise info, viz. about thermodynamics of gene regulation
  - Unlike Big Bang cosmology, the "noise" is very ill-understood & new methods needed
- Biological systems are characterized by correlated probability distributions on high-dimn'l data: how can we infer "true" pdf from finite data? how do we learn?
  - Multi-neuron firing patterns in the retina when exposed to "scenes"
  - The repertoire of antibodies in an individual immune system can be sequenced
  - The distribution on amino acids in a particular protein type across species
  - Physicist-friendly models, such as maximum entropy, work remarkably well.
- Looking for general principles/approaches that apply across systems and kingdoms ("from bacteria to brains"), yet make contact with biological data begins to look realistic.
  - A new level of cooperation between theorist and experiment will be needed.

Acknowledgments

- Bill Bialek, my Princeton colleague from whom I have learned most of what I know about theoretical physics in biology
- Justin Kinney (L: Cold Spring Harbor Lab) and Gasper Tkacik (R: U Penn Physics) … two fabulous Princeton physics graduate students who came to do string theory and left doing biology … proving once again that physicists are good for a lot more than physics ….
- Something which Mike Cornwall has demonstrated over and over in his career …
- HAPPY 75th MIKE!