From Particle Physics to the Physics of Biology: A tribute to Pierre Fayet

Curtis Callan
Princeton University

Over the last century, physics has constructed a theoretical framework for describing the world on all scales, up to and including the universe as a whole. The next hundred years are likely to see equally dramatic advances in biology. This will open up a new agenda for theoretical physics, if only because the flood of data from biological experiment will create an urgent need for mathematical frameworks to make the data intelligible. I will sketch some specific examples of how theoretical physicists are taking up this challenge and speculate on the future of this marriage of theoretical physics and biology.

Personal Notes and Reminiscences

- Les Houches '75: Pierre was a young postdoc, I was a youngish lecturer. We hit it off, and I entrusted my life to him as my alpine guide. For the ensuing unforgettable weekends of high mountaineering, thanks!
- Phenomenology: I visited LPTENS pretty much annually in the following years, and hearing from Pierre was a highlight. He embodied the best French tradition of bringing rigorous theory to bear on experimental fact.
- Rencontres de Moriond: For many years, Pierre was the particle theory organizer for Moriond. He often called on me for a "theory summary" or equivalent … indulging my love of high altitude physics. Thank you!
- LPTENS Spirit: Pierre was a major contributor to the uniquely collegial, serious yet friendly, atmosphere of LPTENS over the years. His personality and friendship were major reasons why I kept coming back year after year.
- Setting off into uncharted territory: I eventually wandered away from my first love in physics, particles and their phenomenology, while Pierre remained faithful to his roots. I first went into string theory and then (worse yet) into biology. I owe him some explanations and, in this offering on the occasion of Pierre's "fête", I will try to provide some ….

The primary task of theoretical physics

- To discover the fundamental, mathematically expressed, Laws of Nature … as well as the "stuff" that obeys those laws .. in our universe (?).
- These laws are valid within broad domains of phenomena and are in some sense "simple", fitting handily on a postcard in shorthand form.
- They were discovered in steps over recent centuries:
  - Newtonian mechanics & the gravitational force law (1670s)
  - Maxwell's equations of E&M (1860s) (and special relativity)
  - Einstein's general relativity (gravity as dynamical geometry) (1910s)
  - Quantum mechanics of electrons, atoms and molecules (1920s)
  - Quantum field theory and the Standard Model of "everything" (1970s)
  - Big Bang cosmology and a theory of the origin of our universe (2000s)
- The culmination of this development, the outcome of an explosion of discovery over the last 50 years, is a Standard Model of our universe.
  - It is a precise mathematical theory whose scope is … everything.
  - Some modesty is of course in order, but a victory lap or two is justified.

The "other agenda" of theoretical physics

- Discovering the fundamental laws is the historic core mission of theoretical physics .. but that's not all theoretical physicists do!
- In addition, we want to explain phenomena that are not directly baked into the fundamental laws … we call them emergent phenomena.
  - E.g. show that superconductivity (a macroscopic quantum effect) follows from the Schrödinger equation for many electrons moving in a host atomic lattice.
  - Or prove "confinement", namely that the quarks of QCD can never be seen outside the hadrons (neutrons and protons) of which they are constituents.
- Better yet, we want to *predict* unknown emergent phenomena (i.e. derive them from fundamental law) *before* their experimental discovery:
  - It doesn't happen often, but it is not impossible: cosmologists *did* anticipate Big Bang phenomena like CMB. Topological insulators were also anticipated.
- Life is the "emperor of all emergent phenomena". Maybe it is time to bring living matter into the realm of predictive, mathematical, science.
  - But looking for "beautiful equations" is not the way to go. Modern physics in fact uses a methodology that I would argue is more applicable to biology.

The methodology of contemporary physics

In cosmology/particle physics we do not directly measure what really interests us. We want to know about certain "hidden variables" and we need a "theory" to learn about them .. and convert data into useful knowledge.

We use our beautiful equations to construct a model (with parameters) for how hidden things generate the noisy data sets we *do* measure. Then use statistical inference to "see through the noise" to discover the underlying physical entities and the parameters describing those entities.

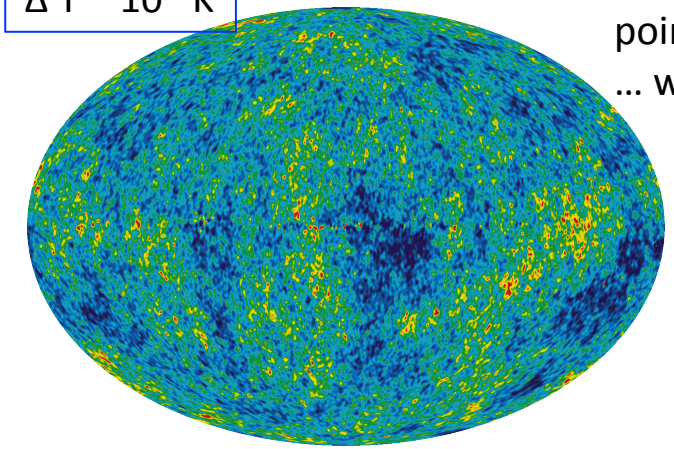
This is a universal method in modern physical science. My point is that biology will need it too: vast amounts of data are being collected … that mean nothing without a mathematical framework to hang them on (and make predictions).

This is how discoveries have been made recently in cosmology (how did the universe we see come into being?) and particle physics (what are the material constituents and forces in our universe?).

A quick look at cosmology will give a concrete and specific idea of what I mean.

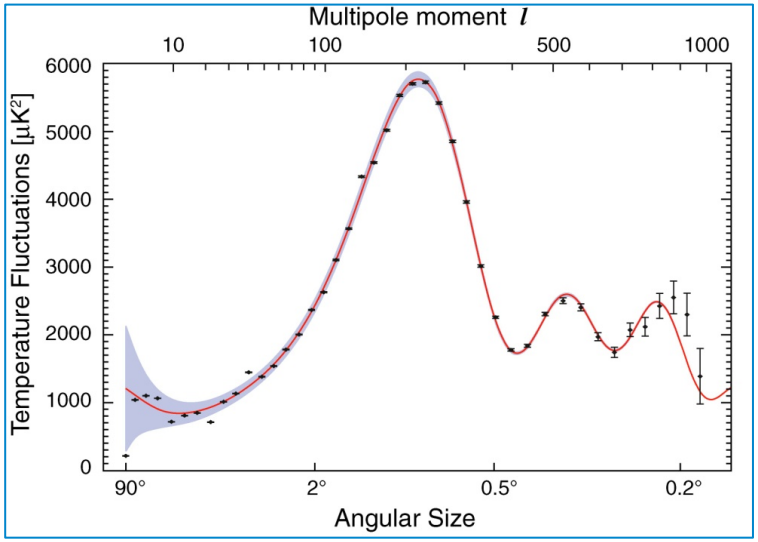
How we know the age of the universe (and more)

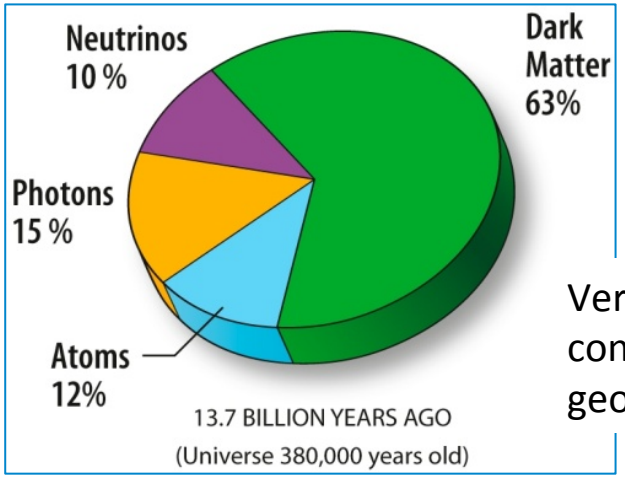
$T_{sky} = 2.7$ K
$\Delta T \sim 10^{-5}$ K



Millions of data points. Signifying … what?

Standard correlation analysis tells you that something interesting is going on! But what?





Very specific matter content and universe geometry come out.

We can only tell by positing a cosmological model with generic geometry, energy content, initial fluctuations .. to see what works!

Planck Collaboration Cosmological parameters[12]

	Description	Symbol	Value
Independent parameters	Physical baryon density parameter[a]	$\Omega_b h^2$	0.022 30 ±0.000 14
	Physical dark matter density parameter[a]	$\Omega_c h^2$	0.1188 ±0.0010
	Age of the universe	t_0	13.799 ±0.021 × 10^9 years
	Scalar spectral index	n_s	0.9667 ±0.0040
	Curvature fluctuation amplitude, $k_0 = 0.002$ Mpc^{-1}	Δ^2_R	$2.441\ ^{+0.088}_{-0.092} \times 10^{-9}$[15]
	Reionization optical depth	τ	0.066 ±0.012
Fixed parameters	Total density parameter[b]	Ω_{tot}	1
	Equation of state of dark energy	w	−1
	Sum of three neutrino masses	$\Sigma\, m_\nu$	0.06 eV/c^2[c][11]:40
	Effective number of relativistic degrees of freedom	N_{eff}	3.046[d][11]:47
	Tensor/scalar ratio	r	0
	Running of spectral index	dn_s / d ln k	0
Calculated values	Hubble constant	H_0	67.74 ±0.46 km s^{-1} Mpc^{-1}
	Baryon density parameter[b]	Ω_b	0.0486 ±0.0010[e]
	Dark matter density parameter[b]	Ω_c	0.2589 ±0.0057[f]
	Matter density parameter[b]	Ω_m	0.3089 ±0.0062
	Dark energy density parameter[b]	Ω_Λ	0.6911 ±0.0062
	Critical density	ϱ_{crit}	(8.62 ±0.12) × 10^{-27} kg/m^3[g]
	Fluctuation amplitude at $8h^{-1}$ Mpc	σ_8	0.8159 ±0.0086
	Redshift at decoupling	z_*	1 089.90 ±0.23
	Age at decoupling	t_*	377 700 ±3200 years[15]
	Redshift of reionization	z_{re}	$8.8\ ^{+1.2}_{-1.1}$

Modern biology has a similar method problem!

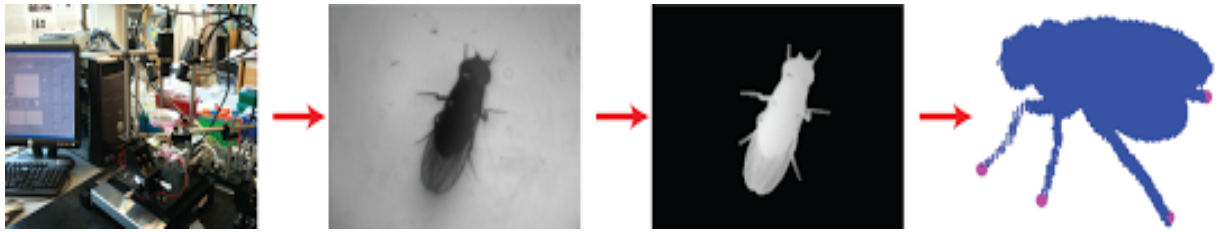
My point is going to be that much of today's biology has the same issue as modern physics: the data is too copious and indirect to mean anything without a mathematical framework to hang it on. This is true in many areas:

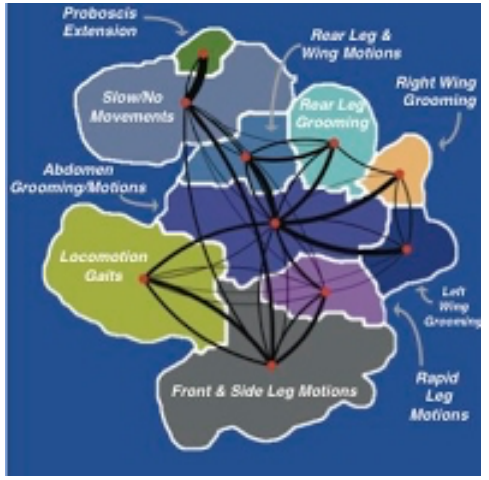
- Next-gen DNA and RNA sequencing. Its not just genomes any more!
- Recording from 10^3s of neurons in a behaving animal (retina, brain).
- Automated quantitative behavior tracking (worms, flies, birds).
- Experimental high-throughput ecology (counting foxes is old hat).
- Single-cell sequencing and expression level tracking .. for 10^4s of cells.
- Synthetic biology lets us make 10^6s of variations on an organism.

I will walk you through a couple of these, to give you a quantitative sense of the problem posed by this ever-more generalized "data explosion". But keep in mind .. it is also an exploding "understanding" problem.

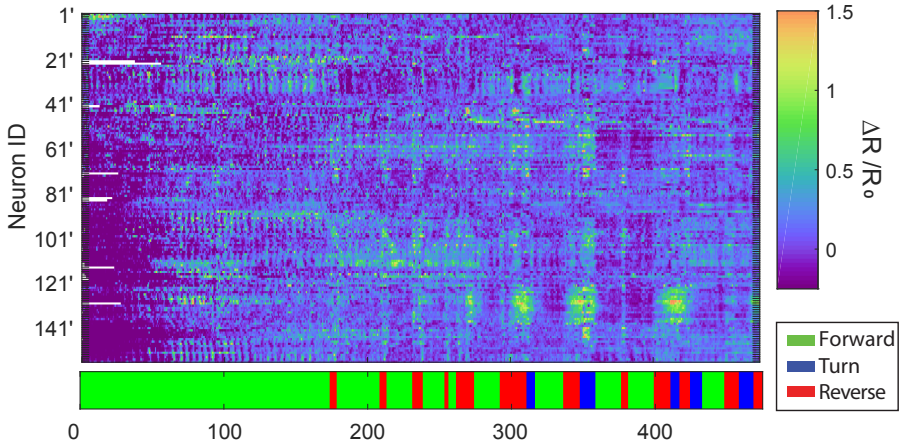
Biological big data in pictures

Fly walking: image, quantify, model, classify





Activity of all brain neurons in a worm



DNA sequences of all T cells in a mouse



The ever-growing suite of ‘omes’: genomes, exomes, proteomes, interactomes give vast amounts of information about the internal states of cells. It is even entering medicine: cancer centers are now routinely sequencing tumors.

The theoretical physics of Life: notable past attempts

Historical instances of using theoretical principles to illuminate and make predictions about a broad class of biological phenomena include:

- Schroedinger's "What is Life?": genes are carried by a polymer molecule
  - Basic quantum mechanics and the known rate of induction of mutations by x-rays (Morgan, 1910) led him to the conclusion that genes had to be carried by a molecule. Many genes -> information -> linear polymer molecule of heredity
- Hopfield and Ninio's "kinetic proofreading" for DNA replication accuracy
  - Boltzmann equilibrium statistics and the binding energy differences between base pairs would give a copying error rate of 1 per 10^4 bases per generation. The actual rate is more like 1 per 10^8, and this led them to propose an energy consuming enzyme that checks fidelity and corrects mistakes. Found!
- Berg and Purcell's explanation of how bacteria locate their next meal
  - Bacteria must move up an odor gradient to find a food source. B&P showed that a bacterium's size, and the physics of diffusion, mean that it can't measure the food gradient across its body. How does it know which way to go? They do a funny "run and tumble" random walk: as long as density is increasing, go straight … else pick a new direction at random. True!

Theoretical physics of Life: contemporary themes

Theoretical physicists, once they get hooked on biology, try to study issues that cut across kingdoms (" from bacteria to brains"), are suitable for mathematical analysis, and are illuminated by modern biological "big data":

We'll look at two examples that exemplify the theoretical physics method of "understanding" complex phenomena through inference of a much simpler underlying "hidden" (mathematically expressed) mechanism.

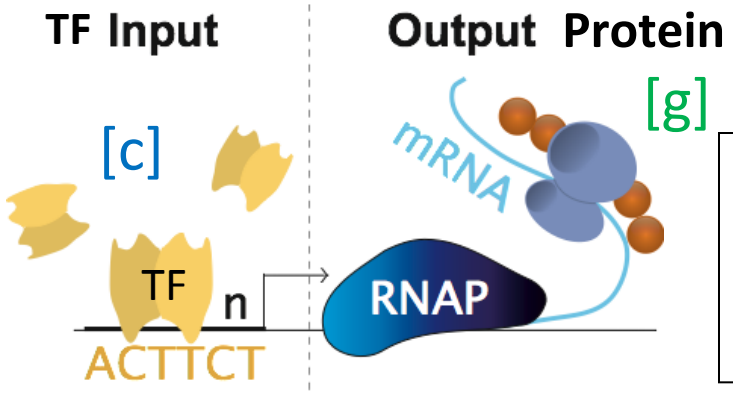
Theme I: Information transmission problems are ubiquitous in biology, and data adequate for quantifying how well it is transmitted can now be collected in many systems.

Theme II: Correlated probability distributions on high-dimensional data, and the problems of learning them from sparse data are similarly ubiquitous in biology.

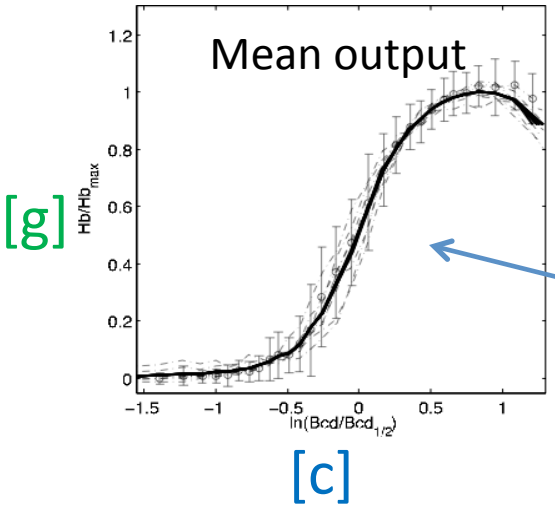
I will tell you about two problems that fall under these headings. They are problems I know well … so please forgive me if I give you more detail than you really want. I will try to make it quick, if not painless.

Quick study: how genes are turned on and off

Transcription factor concentration



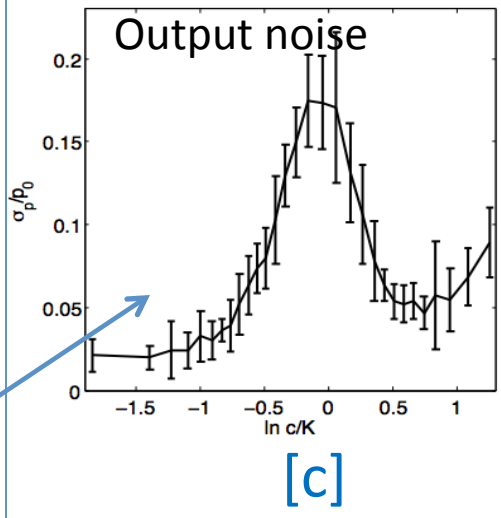
Regulated gene expression



Special DNA site (promoter) is occupied by TF with occupancy τ driven by TF concentration [c] and DNA binding energy ε.

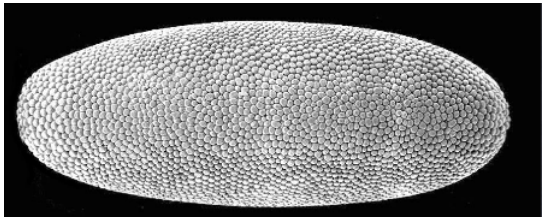
Occupancy τ enhances RNAP binding, drives transcription. Typical on/off mean response.

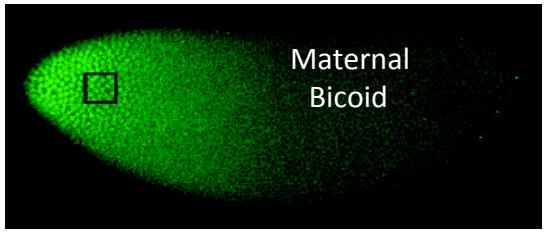
But diffusion and small numbers make the output noisy.

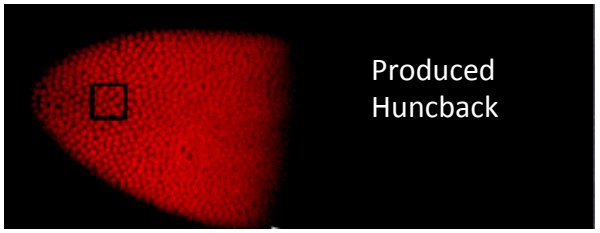


(I) Building a fly embryo bit by bit

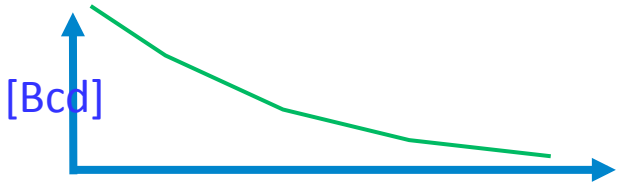
Embryonic development is driven by transcription factor (TF) proteins. TF Hunchback comes first. It separates the thorax from the abdomen in response to the level of maternally supplied TF Bicoid. This pattern repeats in a TF cascade (Bcd → Hb → Kr, Gt, Kn, ..) leading to the expression "stripes" needed to make a multi-segment body.

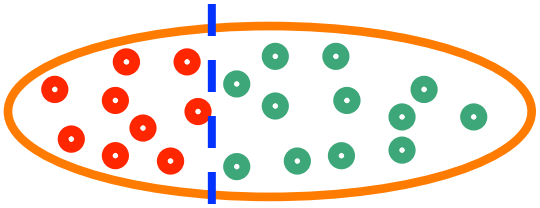






Nuclei must know `where they are' to make cell fate choice (thorax v. abdomen). [Bcd] is the signal: if [Bcd] big enough, Hb is expressed

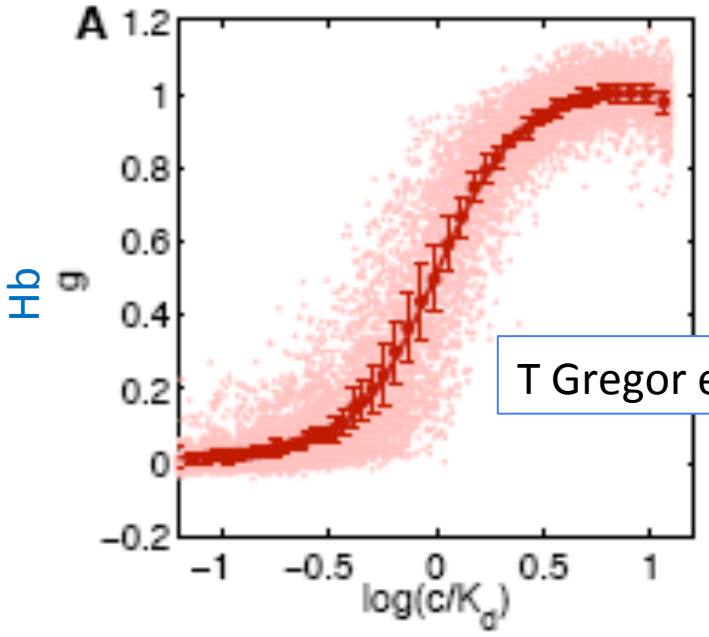




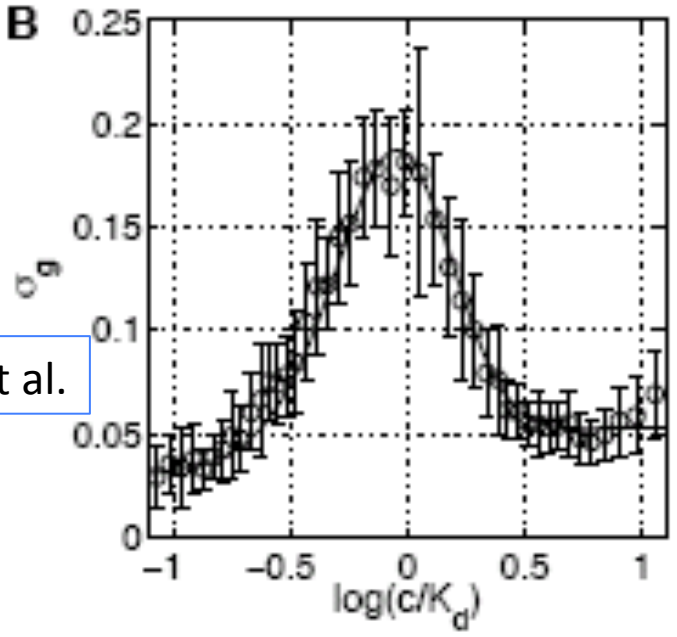
[Hb] and [Bcd] levels in 10^4s of nuclei can be measured by fluorescent immunostaining (ditto for other TFs). We can measure the distribution of 10^6s of input/output pairs P([Hb] , [Bcd]) over the nuclei in 10^2s of eggs. What we see is a noisy "switch". Can it reliably switch when it should … and can it do more than signal just "on" vs "off"?

Measurement of 10^4s of Nuclei (Gregor etal) can be Summarized in a Probabilistic "Genetic Switch" I/O Relation

Input vs output



Noise vs input



T Gregor et al.

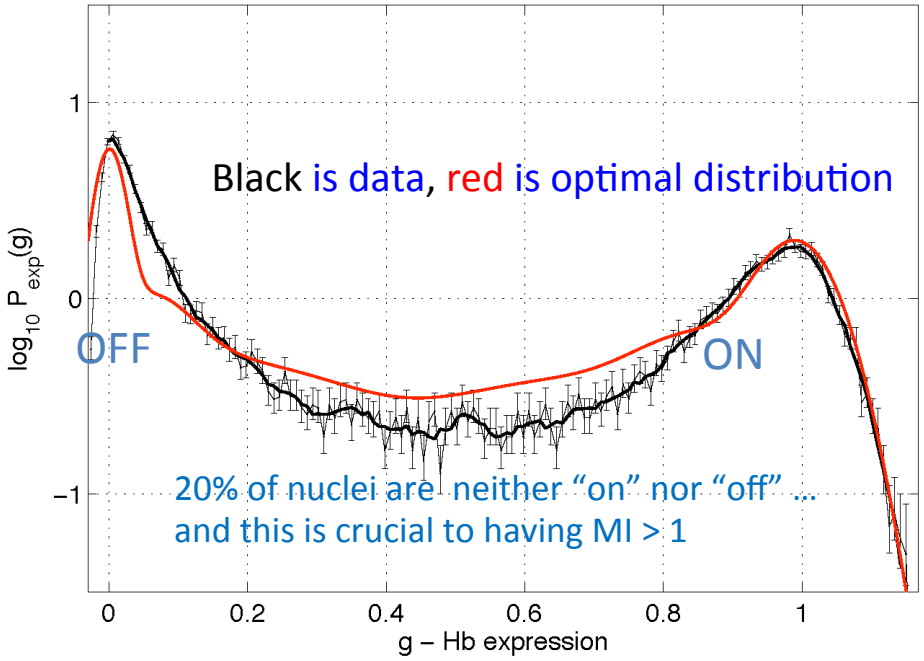
The Bcd/Hb relation is a noisy switch. Can it do more than signal "on" vs. "off", i.e. distinguish multiple levels of the input? mutual information between output [g] and input [c] is the best measure of this.

Normalized noise in readout of [g=Hb] from input [c=Bcd] (~15% accuracy in decision region. Can't do better because of small-number fluctuations in the [Bcd] molecules arriving at the Hb promoter.

(I) Building a fly embryo bit by bit

Given two variables *g* and *c* with joint pdf P([g],[c]) their correlation is best quantified by "mutual information" (MI). MI is positive and measured in "bits". One bit means that knowing [g] will tell you just that [c] was "high" or "low". If more bits --> more discrimination is possible. The joint pdf P([Hb],[Bcd]) has been measured, so we can evaluate MI([Hb],[Bcd]) from the data. We find MI_{data} = *1.5* bits. So, the fruit fly nucleus is more than a simple on/off switch! … Why?

Here is where theoretical physics comes in! We can ask how big MI([Hb],[Bcd]) could be, given the known I/O rule of the nucleus. This is a variational problem, with simple solution. Find MI_{max} = 1.7 bits (true MI is very close to optimal). Plus, we get the specific distribution of TF concentrations that achieves this optimum … it matches the data!



Mutual Information and the Genetic Switch

If we have two variables *g* and *c* with joint pdf *P(g,c)* their correlation is best quantified by mutual information (MI). Positive, measured in bits, basically counts the number of distinguishable levels. We write MI in a convenient alternate form:

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

$P_{TF}(c)$ = "environmental" TF dist'n
(Count how many nuclei in each [c] bin)

$$P_{\exp}(g) = \int dc\, P(g \mid c) P_{TF}(c)$$
(Count how many nuclei in each [g] bin)

All the quantities needed to evaluate *I(g;c)* are measured. No need to know *why* they have the forms they do (for now). Hence, we can evaluate *I*(g,c) from data. We find $I_{data} = 1.5$ bits …. the fruit fly nucleus is not a simple on/off switch!

How big *could* *I*(g;c) have been, given the observed noise? Regarding *P(g|c)* as a fixed quantity, we can take *I(c;g)* to be a functional of $P_{exp}(g)$ and maximize over it. The variational solution gives the max mutual info for the regulated gene *and* identifies the distributions of input *c* and output *g* that achieve this optimum.

A Biological Variational Principle …

For convenience, we use a gaussian approx to *P(g|c)* in the *I* functional. The solution to the variational problem is simple:

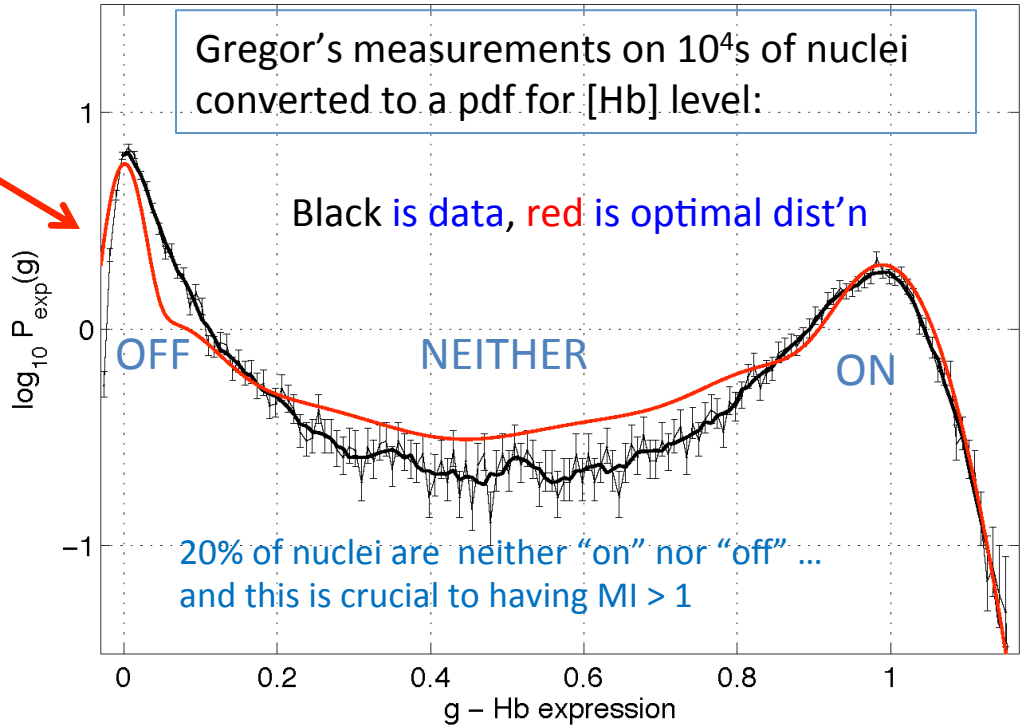
$$P(g|c) = \frac{1}{\sqrt{2\pi\sigma_g^2(c)}} \exp\left\{ - \frac{[g - \bar{g}(c)]^2}{2\sigma_g^2(c)} \right\}.$$

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

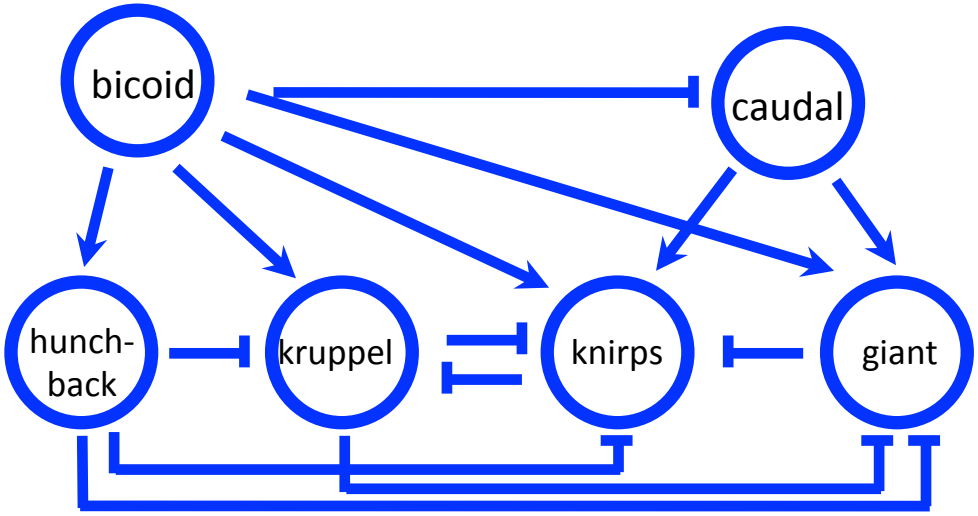
Substitute back into the expression for MI to get the optimal value:
I_{opt}=1.7 bits vs I_{data}=1.5 bits

The true MI is close to optimal, and the actual [Hb] distribution also matches the observed distribution. Doing better than 1 bit is not easy: the cell has to be fine-tuned. Why?

Bialek, CGC, Tkacik (08)

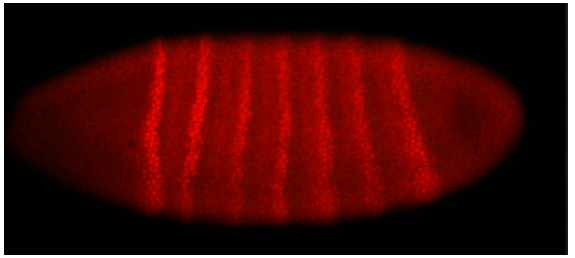


(I) The larger theoretical challenge: role of information in the gap gene network

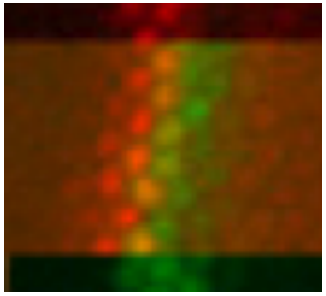


Gap gene network in D. melanogaster

Gap genes express in stripes



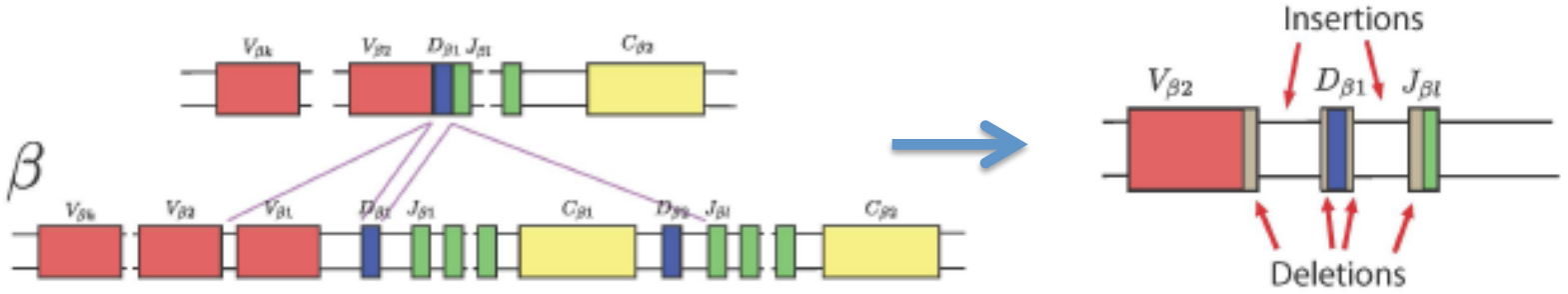
Joint expression patterns of {hb,kr,kn,gt} could be a code for "where you are" on the ant-post axis. NB: neighboring nuclei have different states!



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. Information optimization can be used to "derive" the regulatory network.

(II) Probability distributions & immune diversity

- T&B cells implement adaptive immunity. These cells have surface receptors to recognize pathogens floating loose or (for T cells) infecting our cells.
- A T/B-cell that recognizes a pathogen proliferates to create a big clone to clear the infection. A "memory" clone is left to protect against re-infection.
- New T/B-cells are made by stem cells. In each creation event, the germline DNA for the receptor is "edited" to create a new, unique, receptor gene.



- You have ~ 10^7 unique, randomly-generated immune cell types in your body. Can we quantify this diversity and understand how it is generated?
- By harvesting T cells, extracting/amplifying DNA, and then sequencing, we obtain > 10^5 distinct examples of receptor genes from one blood sample.
- There is virtually no overlap between repertoires of different individuals. We need to understand their statistics in order to learn anything useful.

(II) Probability distributions & immune diversity

The genome editing event is carried out by a few DNA repair enzymes. There are only a few "moves" they can apply to the germline DNA. Any new receptor gene sequence σ is the result of a "scenario" *E*, a set of values for these actions.

$$E_{CDR3}: \begin{cases} \text{V, D, J} \\ \text{del}V,\ \text{del}J,\ \text{del}D5,\ \text{del}D3 \\ \text{pal}V,\ \text{pal}J,\ \text{pal}D5,\ \text{pal}D3 \\ \text{ins}VD,\ \text{ins}DJ \\ (x_1,\ldots,x_{insVD}),\ (y_1,\ldots,y_{insDJ}) \end{cases}$$

Choose possible V,D,J gene segments.

Chew away some number of bases from each

Don't ask! A bit exotic ….

Insert some number of bases between genes

Account for different base insertion probabilities

For each type of move there is an unknown probability distribution: each V gene has its own likelihood of choice, ditto for each number of VD or DJ insertions, etc. We want to infer a pdf for the generative scenarios. To do so, we assume a plausible structure:

$$\mathrm{P}^{\mathrm{recomb}}(\text{scenario}) = P(V)P(D,J)P(\text{deletions}V|V)P(\text{insertions}DJ)\ldots$$

$$P_{gen}(\vec{\sigma}) = \sum_{\substack{\text{scenarios:} \\ V,D,J,\ldots\to\vec{\sigma}}} P^{\mathrm{recomb}}(\text{scenario})$$

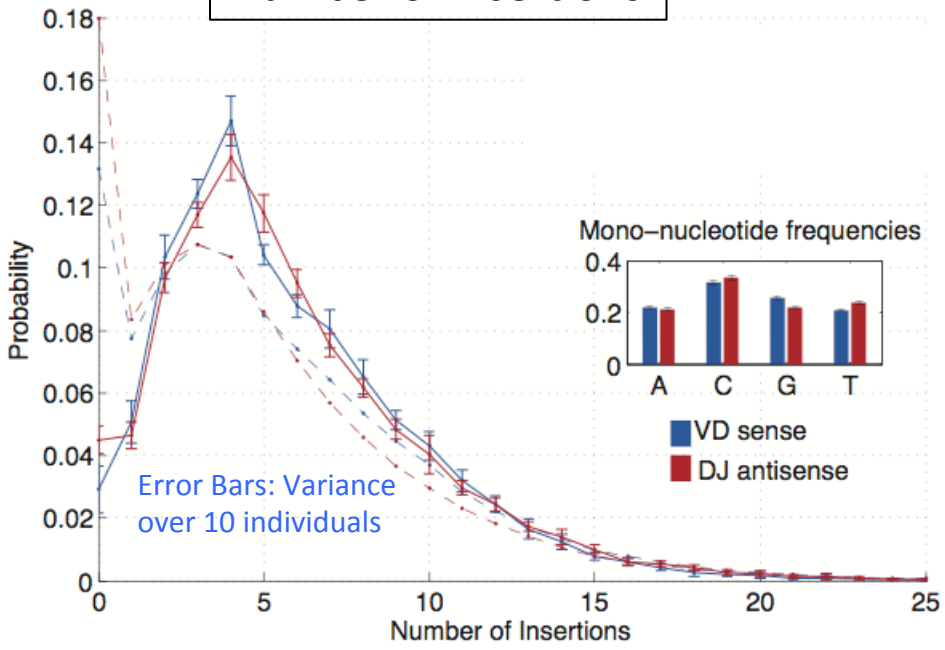
N.B. Many "scenarios" can yield the same sequence read σ.

P_{gen}(σ) is the net probability that the result σ is produced in one stem cell event. Once we know the right component pdf's we will be able to evaluate this hidden variable.

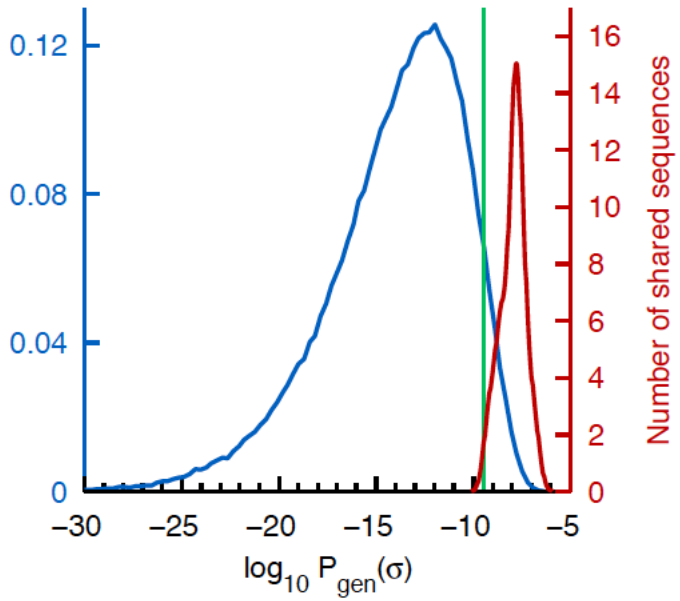
(II) Probability distributions & immune diversity

Key theoretical idea: The true component probability distributions are those that maximize the likelihood (product of individual probabilities over all data sequences) of the sequences you observe. “Easy” search problem .. Typical results are as follows:

Number of insertions



P_{gen} distribution & Shared Seqs



The total possible diversity is enormous: something like 10^{13} sequences. The unique clones we observe come nowhere near sampling sequence diversity. We can get sharp results because we assume that the sequence diversity has a simple hidden source.

Why did I tell you these particular stories?

First: Each is a specific instance of a broader class of conceptually similar problems spanning the tree of life .. and for which big data is available

Theme I: Information transmission problems are ubiquitous, and data adequate for quantifying how well it is transmitted can now be collected in many systems.

- Higher organisms convert sensory inputs to spikes on neurons. Reconstructing a sensory input from the spike train is an information problem.
- Eukaryotic signaling converts external chemical signal into phosphorylation state of internal messenger proteins. Quantify with mass-spec.

Theme II: Correlated probability distributions on high-dimensional data, and the problems of learning them from sparse data are everywhere.

- A given visual scene produces a noisy train of "spikes" on the optic nerves. P(spikes|scene) is the mother of all high-dim'l pdf's
- The same enzymatic function is provided in different species by proteins with varied sequence. Is there a P(aa seq|protein type) to be learned?

Second: Each exemplifies in its own way the theoretical physics method of "understanding" complex phenomena through inference of a simpler underlying "hidden" (mathematically expressed) mechanism.

The theoretical physics of Life: the future

Bigger theoretical questions are waiting in the wings. They are way beyond our grasp today but not (I hope) forever! Examples I like are:

- What is it about non-equilibrium statistical mechanics that makes it possible for populations of distinct, stably reproducing entities to arise? We hardly know where to begin in trying to answer this question.
- Are there equations describing evolution of organisms with 10^3s of genes; can they be solved and can anything be said about their global behavior?
- Can we capture the dynamics of a cell? Can equations accurately describe its behavior, given that a cell has thousands of interacting parts? Can we see tissue types as basins of attraction of a large dynamical system?
- Brains carry out tasks that have an abstract representation. Can dynamical network models capture the processing power of a human brain (or eye)?
- These are some of the “big questions” that a future theoretical physics of biology will want to tackle. Today we are doing “warm-up exercises” to prepare ourselves for the big task that lies ahead.

Summing Up

- It seems blindingly obvious that biology will need some kind of general mathematical framework to organize the data flood that is coming.
- I am very skeptical that generic "data mining" approaches (à la Google) will do the job. Finding and using simpler underlying structures will be key.
- The deep problem is figuring out what complexities can be ignored in creating an "adequate" theory or model of a cell (or brain). Do the "coarse graining" and "universality" that work in physics have a role in biology?
- We need to rise above specific models for each specific biological system. I don't claim that physicists are uniquely equipped for this task .. but they are naturally inclined to look for the generality that we hope for.
- The concrete examples I described in the talk are a pretty pale imitation of the sort of analysis we will eventually need. They are faltering baby steps.
- The problem is a major intellectual challenge, just as hard and deep as the fundamental physics challenges we have solved in the past. Accepting it will ensure the future vitality of theoretical physics .. and biology.