Ambitions for theory in the physics of life

William Bialek

Joseph Henry Laboratories of Physics and Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544 USA

wbialek@princeton.edu

Part of the 2023-07: Theoretical Biological Physics 2023 collection
Session 121 of the Les Houches School, July 2023
published in the Les Houches Summer School Lecture Notes series

Abstract

Theoretical physicists have been fascinated by the phenomena of life for more than a century. As we engage with more realistic descriptions of living systems, however, things get complicated. After reviewing different reactions to this complexity, I explore the optimization of information flow as a potentially general theoretical principle. The primary example is a genetic network guiding development of the fly embryo, but each idea also is illustrated by examples from neural systems. In each case, optimization makes detailed, largely parameter–free predictions that connect quantitatively with experiment.

Copyright W. Bialek.
This work is licensed under the Creative Commons Attribution 4.0 International License.
Published by the SciPost Foundation.

Received 11-04-2024
Accepted 25-06-2024
Published 15-08-2024
doi:10.21468/SciPostPhysLectNotes.84

Contents

1 Introduction

The history of physics teaches us that qualitatively striking phenomena have correspondingly deep theoretical explanations. In some cases the relevant phenomena are quite mundane, and it takes time to appreciate just how surprised we should be. The (literally) everyday observation that the sky gets dark at night turns out to be one of these familiar but profound facts [1,2], as does the rigidity of solid objects [3]. Today, however, the search for fundamentally new physics is concentrated in places very far from our immediate experience: looking back to the earliest times in our universe's history; at the shortest distances and highest energies; at the lowest temperatures; and in materials that do not occur in nature. One might be tempted to conclude that everyday phenomena are understood, at least in outline.

For theoretical physicists, declaring something to be "understood" requires meeting a high standard. We expect a wide range of phenomena to be explained using a small set of general principles; we expect these principles to be summarized in compact mathematical form; and we expect this framework to be tested in quantitative experiments, often with little room for adjusting parameters as we try to reach detailed numerical agreement between theory and experiment. By these standards, the everyday phenomena of life are *not* understood.

In living systems matter organizes itself with an intricacy that is unmatched in the inanimate world. This organized state maintains and even reproduces itself, with extraordinary fidelity. Once organized, living systems behave in ways that are reasonably described as functional or even purposeful and intelligent. While it seems silly to say that water is trying to flow downhill, it would seem equally silly not to admit that a predator is trying to catch its prey.

We should be careful not to anthropomorphize, and certainly we no longer believe in a vital force, but surely there is something different about life.

A physicist's understanding would make the difference between animate and inanimate matter precise. This should yield a classification of complexity in cellular and animal behavior, and predict quantitative connections between this macroscopic complexity and the richness of underlying microscopic mechanisms. These are ambitious goals, but theoretical physics is not a modest enterprise.[1]

1.1 An example, and a problem

Crucial facts about living systems provide ingredients for sharpening our questions. A famous example is the discovery in the 1930s that genes are the size of molecules, or more precisely that the targets for radiation to produce mutations are of molecular dimensions [4, 5], which was the foundation for the questions and conjectures in *What is Life?* [6]. The explosive growth of molecular biology has given us a veritable encyclopedia of facts about life's microscopic mechanisms [7, 8], and it is natural to try and summarize these facts in mathematical terms, perhaps leading to something that we would recognize as a theory of the phenomena that first attracted our attention. But such efforts lead to a forest of arbitrary parameters.

Let us start with an unambiguously striking phenomenon, the development of a single cell into a complete multicellular organism. Building on a century of foundational work by biologists, we will focus on the case of a fruit fly, *Drosophila melanogaster,* where it takes just twenty–four hours to go from one cell to a larva (maggot) that emerges from the egg shell and walks away, ready to navigate the world. The maggot has a segmented body, and it is remarkable that if you know which molecules to look at then you can measure striped patterns in the concentration of these molecules (Fig 1); these stripes provide a preview of the segmented structure. It is perhaps even more striking that these stripes develop just *three hours* after the egg is laid, a time when almost all the $\sim 2^{14}$ cells are geometrically equivalent, arrayed in a featureless lattice covering the embryo's surface. Thus (in this case) Nature has divided the problem of development into laying out a blueprint—transmitting to each cell information about its position in the embryo, and hence its ultimate fate in the final structure—and then actually building the structure by changing the shape of the embryo.

How does the fly embryo make stripes? By the turn of this century, there was a clear outline of how this works. To start, all the relevant molecules were identified, and this was one of the greatest triumphs of using genetic methods to dissect a complex phenomenon in living systems [9, 11, 12]. The molecules with striped patterns of concentration are a group of eight proteins that are encoded by the "pair–rule" genes. Whether these proteins are synthesized[2] is controlled largely by the concentrations of another set of four proteins that are encoded in the "gap genes," so named because a mutation in one of these genes causes a large gap in the body plan. The gap genes not only regulate the pair–rule genes, they also regulate one another, forming a network. Finally, there are three inputs to the gap gene network that are provided by the mother when she makes the egg. This flow of information through three layers of a molecular network is schematized in Fig 2.

Schematics with nodes connected by arrows are common in descriptions of living systems. How do we turn these schematics into equations? There is no unique mapping but there are

[1] I am hoping that these notes capture some of the fun and informality of the original lectures. One advantage of the written version is that I can give references. I may have been over–enthusiastic about this, but perhaps the long bibliography will provide guidance to a literature that sprawls across physics and multiple subfields of biology.

[2] When the information encoded in a particular gene is read out to make the corresponding protein we say that the gene is "expressed." This occurs in two steps, the synthesis of mRNA from the DNA template (transcription) and then the synthesis of protein from the mRNA template (translation). Both steps are regulated, although our emphasis will be on the regulation of transcription. When someone talks about the "expression level of a gene" it can be ambiguous whether this means the number of mRNA molecules or the number of protein molecules.

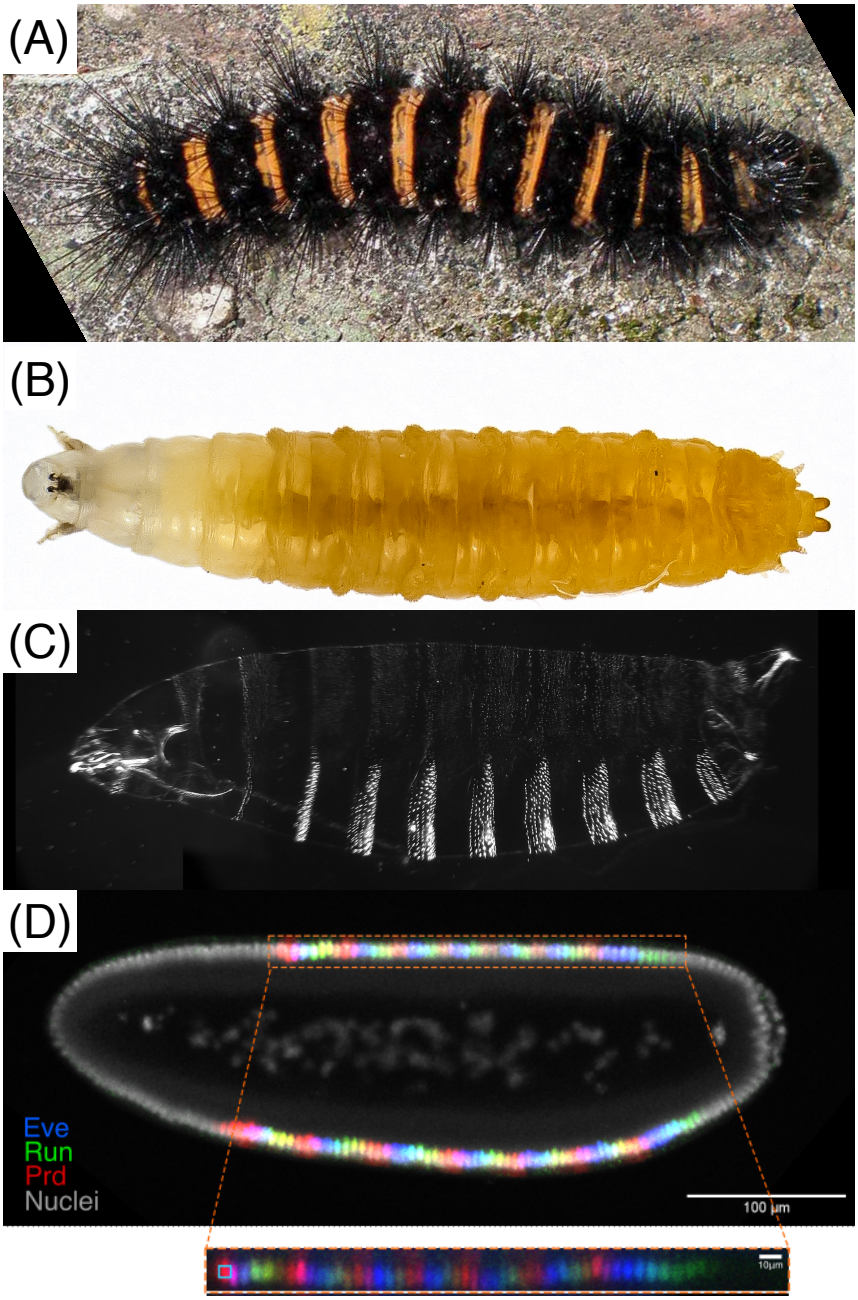

Figure 1: Segmented body plans of larval insects, and the underlying molecular blueprint. (A) Caterpillar of the agreeable tiger moth, in which the segments are especially visible. Image by Cyndy Syms Parr, from Wikipedia under the CC–SA 2.0 license. (B) Larva (maggot) of the fruit fly *Drosophila melanogaster*. Image by Salvatore Vitanza, with permission. (C) The "cuticle preparation" of the *Drosophila* maggot shortly after hatching, highlighting the segmented structure. Thanks to Eric Wieschaus for the image, from experiments described in Ref [9]. (D) An optical section through an embryo stained for three of the "pair–rule" proteins, showing striped patterns that align with the body segments; data from Ref [10], with thanks to M Nikolić for the image.

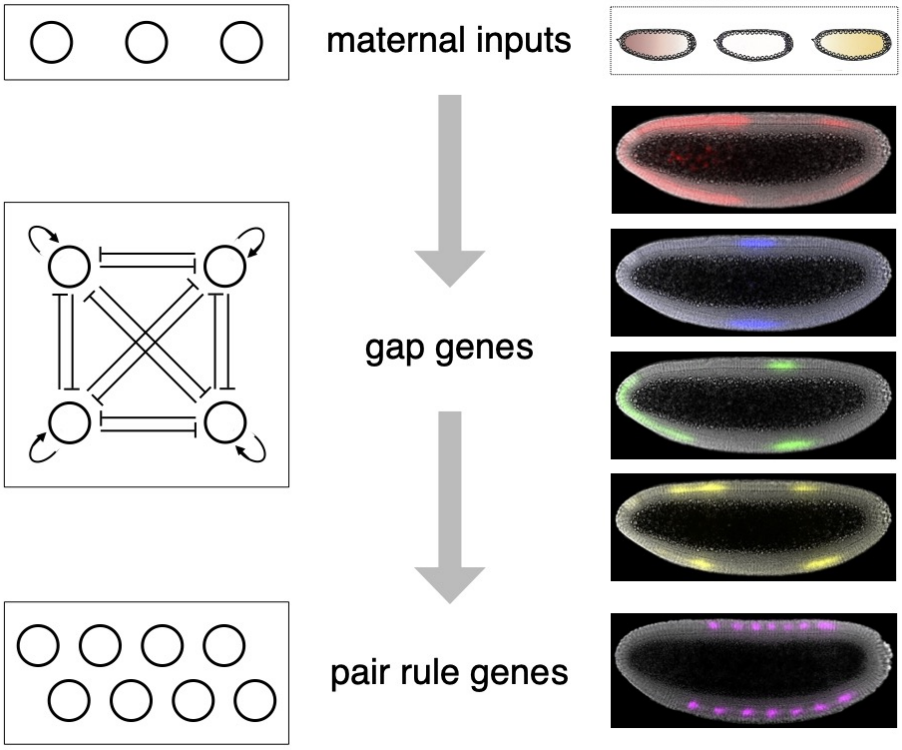

Figure 2: A schematic of information flow in the early fly embryo. The maternal inputs have simple spatial profiles: one with high concentration at the anterior end (left), one with high concentration at the posterior end (right), and one that is symmetrically high at the both ends (middle). These molecules activate the expression of four gap genes, which also regulate one another. Finally the gap genes modulate the expression of eight pair–rule genes, whose concentration profiles consist of striped patterns; one example is shown. Not illustrated are paths by which the maternal inputs can reach around the gap genes to regulate the pair–rule genes directly. Images reproduced from Ref [10], with permission.

some common themes. In a genetic network such as the one relevant for the fly embryo, drawing an arrow $A \to B$ means that the rate of synthesis of B molecules depends on the concentration of A molecules (Fig 3), so at the very least we must have something like

$$\frac{dB}{dt} = r_{\rm max} f(A) - \frac{1}{\tau} B \,, \tag{1}$$

where $r_{\rm max}$ is the maximum synthesis rate, $1 > f(A) > 0$ is a normalized regulatory function, and τ is the lifetime of B molecules; throughout these lectures I'll usually use the same symbol for the name of the molecular species and for its concentration. We can use τ to set the units of time, and the combination $r_{\rm max}\tau$ to set the units of B, but we still need to describe the regulatory function. Plausibly it is monotonic, increasing if A activates the expression of B and decreasing if A represses the expression of B.[3] A smooth function running monotonically between 0 and 1 has at least two parameters, roughly the concentration A at which $f(A) = 1/2$ and the slope or sensitivity at this point. We can imagine that the slope is controlled by the number of A molecules that bind cooperatively to the relevant sites along the DNA and regulate the gene encoding B, although we should not take this too literally.

[3] In networks of neurons we speak conventionally of excitation and inhibition, rather than activation and repression as in genetic networks. I am not sure how these differences in jargon arose.

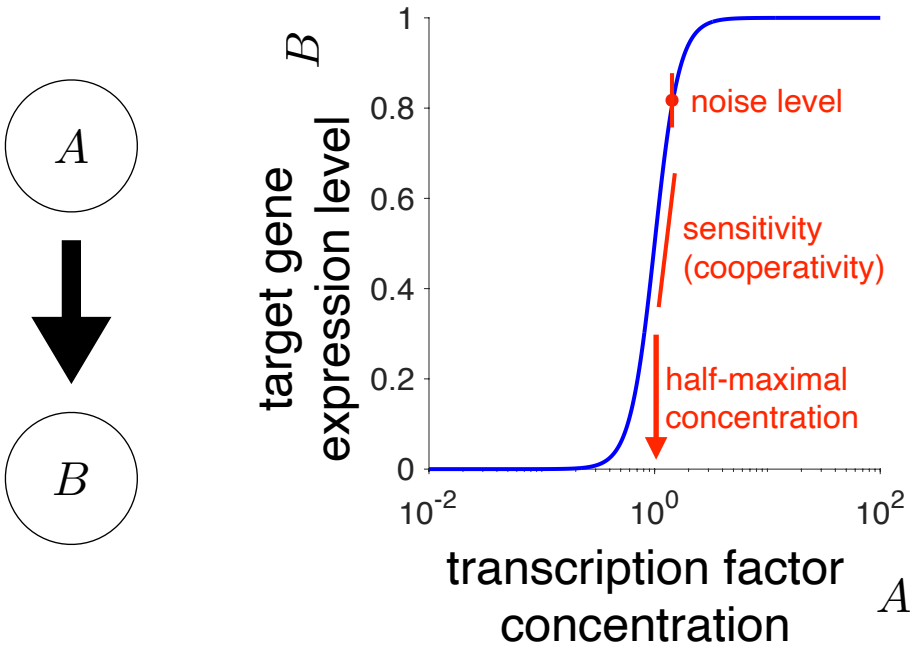

Figure 3: Parameters hiding under the arrow. A is a transcription factor activating the expression of B, as in Eq (1). If this system comes to steady state then the expression level $B = r_{\max}\tau f(A)$, and we are free to choose units such that $r_{\max}\tau = 1$, so that $B = f(A)$ as shown here. The regulatory function $f(A)$ has at least two parameters, as shown. The system also is noisy, which we neglect in Eq (1), but this will be important below.

So, to give a quantitative description we need to attach at least two numbers to every arrow in our schematic, and even this leaves things out:

- Something complicated could happen at places where multiple arrows converge.
- We have associated one molecular species with each gene, but there could be at least two—the protein and the corresponding messenger RNA.
- We have assumed the dependence of synthesis rates on regulatory inputs is instantaneous, but switching among regulatory states could introduce relevant time scales.
- In the embryo there are many cells,[4] so there is a separate copy of these dynamics at each of $\sim 2^{14}$ sites. Exchange of molecules between sites can be described by an effective diffusion constant, but the dynamics could be more complicated.
- During the relevant time period there are multiple rounds of nuclear division, and the regulation of gene expression could interact with this cycle.

Even with these simplifications, a network of four gap genes with three maternal inputs could have $(4 \times 4) + (3 \times 4) = 28$ arrows, and in fact there is evidence that most of these exist; certainly there is no reason to exclude any of them *a priori* from a theory of this network. But then there are at least 56 parameters needed to describe the first few hours of development

[4] In the fly, there are actually no membranes separating the cells until well into the fourteenth cycle. It would be more precise to say "many nuclei."

along one axis of the embryo of one particular organism. It seems fair to say that this kind of complex model is uncomfortable for most theoretical physicists.

In the traditional core of theoretical physics our models of the world have parameters [13], but somehow fitting these parameters does not seem to be the central focus. We assume that if there are too many parameters there must be something that unifies them, and if we need to set these parameters precisely to non–generic values we hope there is some extra dynamics that can make this happen more naturally. More strongly, their are spectacular successes which are almost parameter independent, such as the BCS theory of superconductivity [14], the renormalization group theory of critical phenomena [15–18], the theory of the fractional quantum Hall effect [19], and more. These theories make detailed quantitative predictions about the properties of real materials, despite the complexity of these materials on an atomic scale. In fact the periodic table of the elements, and hence most of the rules of chemical bonding, can be derived from quantum mechanics by knowing only one parameter ($\alpha \sim 1/137$), and even this isn't so important in the first rows of the table. Perhaps ironically, the standard model of elementary particles has many more parameters [13], but non–trivial predictions that are the foundation for our confidence in this model are nearly parameter–free, as with the connection of deep inelastic scattering experiments to the asymptotic freedom of QCD [20–23].

Not all our efforts at theorizing reach the lofty heights of BCS or QCD. But generations of theoretical physicists have developed a distaste for highly parameterized models, and by and large this bias has served the community well.[5] If we need 50+ parameters to describe one genetic network in one organism, and there are no principles that cut through the arbitrariness of these parameters, then we will be led to a different model for each of the many different genetic networks relevant in the life of complex organisms. The same concern applies to other classes of processes. The resulting collection of independent models for each of many different but related phenomena is almost the opposite of the physicist's search for unification.

1.2 Reacting to complexity

In many ways, theoretical physicists' engagement with the phenomena of life can be classified by their reaction to this proliferation of parameters. At the risk of being cartoonish, let me list some of these broad classes:[6]

0. Give up, life really is that complicated. A positive way of saying this is that living systems really must be described by models with many parameters, as in the example of the fly embryo, and so the interesting theoretical problems concern how we infer these parameters from data, or how prediction may be possible even when parameters are underdetermined. This once seemed pessimistic, but it has received renewed attention in response to the dramatic successes of deep neural networks [24] and large language models [25,26]. These models grew, over a long period, out of efforts in the physics community to build theories of brain function [27–31], and many physicists now are interested in the question of why these networks 'work' as well as they do [32–36]. It is not unreasonable to think that a theory of deep networks will circle back to influence how we think about the physics of life.

1. We should study theories that remind us of the real thing, and not try for quantitative comparison of theory with experiments on real living systems. There was a period in which this style of work came under the heading of "biologically inspired physics" [37]. As an example, much of the work on soft and active matter grew out of efforts to create simpler, better controlled examples of phenomena that we first encounter in the living world, from

[5] It will be interesting to see whether this view survives the current revolution in artificial intelligence.

[6] References are meant to be illustrative rather than exhaustive, and are a mix of original papers and reviews.

fluid membranes [38] to flocks and swarms [39]. In the same spirit, neural networks are a source of statistical physics problems that now are quite independent of efforts to understand how real brains process information and learn from their experience. In the background of this approach is the worry that experiments on living systems are irreducibly messy, and so we will never have the kind of theory/experiment comparison in biological physics that is characteristic of physics more broadly. This concern is addressed explicitly below.

2. The only real theory is of how things are related to one another. One sometimes hears the claim that biology is different from physics because biology is historical and physics is not. This is a complicated claim,[7] but it suggests that even if we can't have a theory of life as we see it today, we could have a quantitative theory of the relationships among different life forms, over time—evolution. Indeed, circa 2000, a number of physicists realized that population genetics and evolutionary dynamics could be seen as statistical physics problems, and this has been extraordinarily productive: even the simplest models of evolutionary change have subtle properties, the progress of a population of organisms over a fitness landscape is dominated by individuals in the tail of the distribution [40, 41], and more realistic contexts lead to fascinating interacting many–body problems [42–44]. This work has involved both sophisticated theory and quantitative connections to experiment, both in the laboratory [45–47] and in the populations of viruses that infect humans all over the world [48–50].[8]

3. We are interested in (relatively) macroscopic behaviors, and these could be more universal than their microscopic mechanisms, in the spirit of the renormalization group. Biologists often complain that physicists oversimplify when we think about living systems, but we also simplify when we think about inanimate matter. These simplifications work not just because we are lucky. The renormalization group teaches us that if we start with a detailed microscopic description of a system and coarse–grain to arrive at a model for behavior on long distances and long time scales, then in this process many of the microscopic details will be lost. In technical language, there usually are only a small number of relevant operators [15–18]. Thus, quantitative descriptions of macroscopic phenomena can be simpler and more universal that the underlying microscopic mechanisms. This inspires us to think that essentially macroscopic functional phenomena in living systems could be similarly independent of microscopic details. Possibly related ideas of simplification arise from thinking about the broad spectrum of sensitivities to different parameters [51, 52], and the universal behavior of dynamical systems near transition or bifurcation points [53].

4. The fact that living systems function (often quite well!) can be promoted to a principle that selects parameters or behaviors, circumventing details. On a dark night, our visual system can count single photons [54]; in bright daylight, insect eyes reach a resolution close to the diffraction limit [55]; bacteria navigate chemical gradients so reliably that they must be counting every molecule that arrives at the their surface [56]. These and other examples suggest that organisms can reach levels of functional performance close to the limits of what is allowed by the laws of physics [57]. We can turn this around, promoting evidence of near optimal performance to a principle from which we can derive aspects of the underlying mechanisms: rather than fitting highly parameterized models to data, we can hope that parameters

[7] In our modern understanding, the particular locations of stars or galaxies are accidents of history, but the distribution out of which these positions are drawn is not. Thus one can see cosmology as historical. But this absolutely does not preclude having theories in the same way as in other areas of physics—based on fundamental principles and tested by detailed comparison to (very!) quantitative experiments.

[8] As a result there have been important practical consequences of this work, both in designing next season's flu vaccines and in the global response to covid–19. See, for example, https://nextstrain.org.

have been driven to values that optimize performance. This is in keeping with the formulation of many ideas in the core of physics as variational principles, such as least action or the minimization of free energy.

Exploring each of these reactions to complexity would make for a long review article, or perhaps a whole semester's worth of lectures. With four lectures, it seems best to choose one approach and explore it more deeply, so these lectures will be about approach #4. We will keep coming back to the example of the fruit fly embryo, but I also will try in each lecture to connect what we have been doing with the embryo to work on other examples, encouraging you to think about the generality of the principles involved. You will have to decide how far we have come, but I hope to communicate my ambitions. But first...

1.3 A few words about experiments

Theoretical physics aims at a compact and compelling mathematical description of the world. Because our theories are mathematical, our predictions are numerical. Testing our theories thus involves measuring numbers, and this is such a central feature of physics that we can take it for granted. But can we do this in living systems, with all their functional complexity?

Not so long ago, experiments on living systems seemed hopelessly noisy and messy. Few things were measured quantitatively. Some suggested that this absence of quantification was an essential difference between biology and physics, that crucial features of living systems were not reproducible as we expect in experiments on the inanimate world. This view—which was surprisingly popular—always involved ignoring particular fields where experiments had reached physics–level precision, e.g. in studying the ability of the visual system to count single photons or the connection of neural dynamics to the properties of single ion channel molecules. If correct, the view of biological systems as intrinsically messy would mean that there simply is no path to build an understanding of life that parallels the theoretical physicists' understanding of the inanimate world. It is not just that we would need new principles, which would be welcome, but that we would have to retreat from what we mean by "understand." This all has changed dramatically. There has been an explosion of opportunities for physics experiments on the functional behavior of living systems, across all scales from molecules to ecosystems. I hope we can banish forever the prejudice that life is a mess.[9] As data improve, we should ask more from our theories.

As a theorist I can be an unapologetic fan of what my experimental colleagues are doing. Since the early fly embryo will be our prime example, let's focus on how measurements are made in this system, and then circle back to give a quick survey of how things work in other contexts. In the embryo we have a network of interacting genes, so we'd like to monitor the dynamical variables at each node of the network. How do we do this? In Les Houches I only drew schematics on the blackboard, but here I take the liberty of showing real data, including some of the raw microscope images from which these data are extracted. I find these very beautiful, and hope you will too.

The relevant variables in a genetic network are the concentrations of proteins and mRNAs. There are ways of measuring both, and this can be done in live embryos and in fixed embryos that give us snapshots of the underlying dynamics. Each method has pros and cons, and methods evolve with time. I am writing this just after hearing about (admittedly preliminary) measurements that were not possible a few months before when I gave these lectures in Les Houches. As a theorist you will need to keep up with what can be done experimentally.

The oldest methods for exploring the genetic networks of the fly embryo involve measurements of protein concentrations in fixed samples. One gently cooks the egg, stopping all the action, and then does some chemistry to make the embryo permeable and cross–link

[9] Certain aspects of life, of course, will remain messy, and delightfully so.

the proteins so that they don't move around. At this point we have plenty of time to make measurements. Cells make many thousands of different proteins, and we want to know the concentrations of just a few of these, marked with the labels on the network nodes. To do this we exploit the specificity of life's own mechanisms.

We can purify the protein we are interested in and inject it into an animal which will then mount an immune response. To a good approximation, we can extract antibodies that bind to the protein of interest and nothing else. If we can tag these antibodies with a fluorescent molecule, then when they diffuse into the embryo they will stick to the target protein and the local fluorescence intensity will be proportional to the protein concentration; this is immunofluorescent staining.[10] With care in the choice of fluorophores, one can now measure, simultaneously, the concentrations of four proteins, sufficient to probe all the nodes of the gap gene network, for example.

Figure 4 shows an example of these measurements in the embryo, for one of the gap genes. The embryo is intact, but the plane of focus is midway through its depth, so cells are arrayed along the rim of the optical slice. One can resolve individual nuclei, but here we just plot the fluorescence intensity averaged over a window with diameter equal to that of the nuclei, sliding along the rim; we measure the position x of the window by projecting onto the midline, so that $x = 0$ is the anterior end (future head) and $x = L$ is the posterior end (future tail). It is hard to convert the raw fluorescence intensity into absolute protein concentrations, so we choose units where the maximum mean concentration across embryos is $\max_x \langle g(x/L)\rangle = 1$, and we subtract a background so that $\min_x \langle g(x/L)\rangle = 0$, where $\langle g(x/L)\rangle$ is the average across the ensemble of embryos at fixed x/L.[11] To be clear, this single normalization is applied to all the embryos in the sample. Plausible alternatives, such as normalizing each image by the maximum intensity in that image, are unphysical and distort our estimates of noise, which will be important below. The small fluctuations from embryo to embryo that we see here provide a first glimpse of how the picture of biology as noisy can be conquered by careful experiments.

It would be nice to take a shortcut and have the proteins themselves be fluorescent. Most of the fluorescence, and indeed most of the color that we see in living systems is generated by medium–sized organic pigment molecules that are synthesized through pathways that engage several enzymes (proteins that catalyze specific chemical reactions). A major advance was the discovery that the fluorescence we see in some species of jellyfish and other sea creatures arises directly from a single protein molecule, with no accessory pigments, soon named the "green fluorescent protein" or GFP [59, 60]. It would take thirty years until it was possible to clone and sequence the gene that encodes this protein, and then insert this gene into other organisms [61, 62]. Importantly one can attach the gene for GFP to the gene for a protein you are interested in, with a short linker, so that the protein is synthesized with an intrinsic fluorescent tag; these are called GFP fusions. Considerable effort has gone into engineering the fluorescent proteins so that they have a range of emission and absorption spectra [63].

All of the proteins we are interested in here are present at very low concentrations, so the proportionality of GFP fluorescence to its concentration is guaranteed. With immunofluorescent staining it is a bit less obvious, because nonlinearities might creep in through the two steps of antibody binding. By making a GFP fusion with a particular protein and then

[10] This technique is a little more complicated. Rather than tagging the (precious) antibodies against the protein of interest, one makes a large batch of general purpose anti–antibody antibodies, and tags these. This depends on the fact that antibody molecules have two parts, one specific to their target and one common to all targets but varying from animal to animal. There thus are two steps: exposing the embryo first to antibodies against the protein whose concentration we want to measure, and then to the fluorescently tagged secondary antibodies.

[11] These experiments are done in inbred laboratory stocks of flies, minimizing genetic variation. Nonetheless there are small variations in the embryo length L, even in eggs laid by the same mother in succession. Our (very conventional) choice to plot always vs x/L suggests that the underlying pattern formation dynamics have some mechanism that compensates for these variations, achieving a kind of scale–invariance. This is a subtle problem that I stayed away from in my lectures, but we have come back to it since then [58].

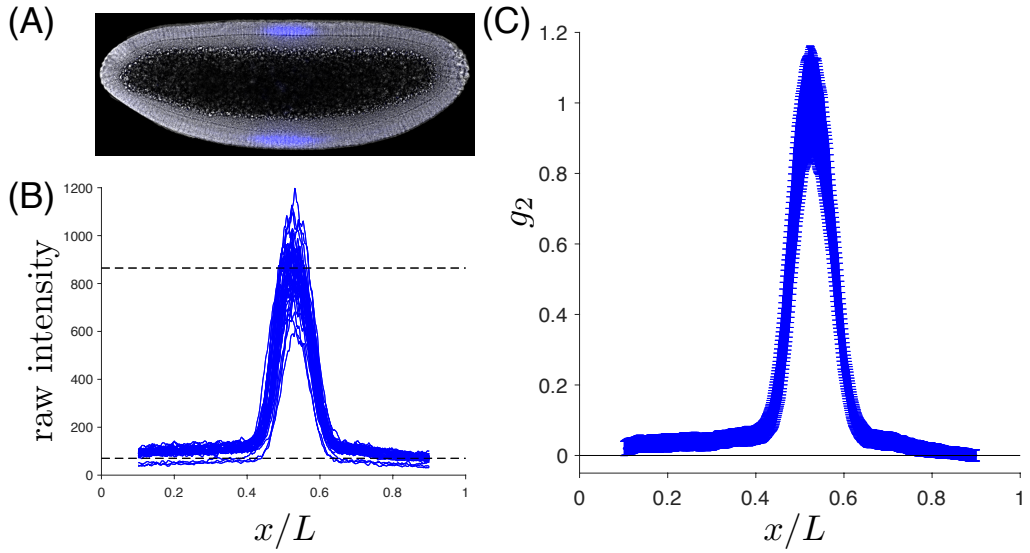

Figure 4: Measuring protein concentrations in the *Drosophila* embryo. (A) Image of fluorescence from labeled antibodies against one of the proteins encoded by the gap genes, corresponding to g_2 in Fig 11 below; the gene is named *krüppel*. Embryo is $\sim 500\,\mu$m long. Reproduced from Ref [10], with permission. (B) Raw fluorescence intensity from images of many such embryos, measured along the (straigher) upper edge. Embryos are chosen to be a in narrow time window after the egg is laid, and in a narrow range of orientations. We define a background (lower dashed line) and a scale (upper dashed line) so that the mean concentration is in the range $0 < \langle g_2(x)\rangle < 1$. (C) Mean and standard deviation across this large ensemble of embryos. Note that fluctuations in the peak concentration are $\sim 10\%$, and the flanks around the peak are defined very sharply. Data from Ref [10], with thanks to T Gregor, MD Petkova, G Tkačik, and EF Wieschaus.

directing immunofluorescent stains against both the protein and GFP itself, one can verify that immunofluorescence intensity really is proportional to protein concentration, and that nothing funny happens to break the 1–to–1 link between the protein and the GFP tag [64].

The construction of GFP fusion proteins has the obvious advantage that one can measure protein concentrations in live cells or embryos, making the dynamics of these signals (literally) visible in real time. The disadvantage is that GFP does not immediately fold into the structure that supports maximal fluorescence, but rather takes time to "mature." There is a continuing stream of work to engineer GFP variants that have shorter maturation times, but we are not quite where we would like to be. Thus, if we want to measure the concentration of a maternal morphogen during nuclear cycle fourteen, we are looking at proteins that have been synthesized more than two hours ago and everything is fine.[12] On the other hand, if we make GFP fusions with the pair–rule genes we can see the stripes emerging during cycle fourteen but the dynamics we see probably are lagging the true dynamics.

The obvious disadvantage of immunofluorescent staining is that one has only a snapshot of a dynamic process. We can do better by fixing and staining hundreds of embryos at once so that we get many snapshots, and unless we make a big effort to synchronize things these snapshots will be at different times after the eggs are laid. Just by counting we can see whether we have stopped the action in nuclear cycle 12, 13, or 14, but we can do better. As noted above, the first cycles of nuclear division are completed without pause to make membranes that define

[12] Dual immunofluorescence experiments (above) can detect small corrections due to the maturation time.

separate cells. During nuclear cycle fourteen these membranes are constructed by infolding of the membrane that surrounds the whole embryo—the membrane of the initial fertilized egg cell. If you watch this process in many live embryos, you see that the distance the membrane has progressed is such a reproducible function of time that you can take it as a clock accurate to one minute [65]. If you are not careful about this and mix together embryos fixed at different times, you vastly over estimate the noise in the expression levels.[13] As an example, in Fig 4 we look only at embryos in the window $40 < t < 44$ min into nuclear cycle fourteen.

Rather than measuring protein concentrations one can observe the mRNA molecules. Again we rely on the specificity of interactions among biological molecules to point accurately to the mRNAs that are transcribed from particular genes. Short segments of DNA can be synthesized that are complementary to different pieces of an mRNA sequence, each labelled with a fluorescent molecule. As with the antibodies directed at proteins, we can diffuse these molecules into a fixed and permeabilized embryo, in effect "lighting up" each individual mRNA molecules with dozens of fluorophores, making it bright enough that we can count molecules one by one, as shown in Fig 5 [66]. This can be done in multiple colors, counting the mRNAs from multiple genes, and again this is sufficient to monitor all of the gap genes simultaneously. These methods can be extended using combinations of fluorophores and cycles of washing and relabelling, so that eventually one can count hundreds of different mRNA species, each with single molecule resolution [67, 68].

In Figure 5 you see not only the individual mRNA molecules in the cytoplasm, but also one or two exceedingly bright spots inside the nuclei. These are the locations along the two chromosomes where the mRNA for this gene is being transcribed. The gene is long enough that many copies of the transcriptional apparatus can operate simultaneously, with the result that many mRNA strands are "in progress" and still tethered to the DNA. With care one can calibrate against the fluorescence intensity of the cytoplasmic spots and effectively count these nascent transcripts. By tagging probes directed against the early and late parts of the sequence one gets a distribution of both colors and intensities across spots, and this can be used to test models of the underlying dynamics [69]. By interleaving the different colors one can check the precision of the measurement.

As with proteins, there is a strategy for live measurements of transcriptional activity. Instead of attaching the DNA sequence encoding GFP to the gene of interest, one can add a sequence drawn from a virus that carries its genome as RNA rather than DNA. To package the genome, proteins that form the coat of the virus bind to these specific RNA sequences, which fold into three–dimensional structures called "stem loops." If the fly also has been genetically engineered to produce the coat protein tagged with GFP, then as the gene we are interested in gets transcribed these molecules, initially distributed throughout the nucleus, will bind to the nascent transcript [70–72]. If there are many stem loops there will be many GFPs, enough to light up the mRNAs much as in the nuclear spots of Fig 5. After a decade of development the noise levels in these measurements now are at the point where one can almost count transcripts one by one [73].

This strategy for visualizing transcriptional activity again exploits the specificity of interactions among biological molecules (here the coat protein and the stem loop), it makes use of sophisticated genetic engineering, and it depends on pushing the state of the art in optical microscopy. These experiments get directly at the dynamics, showing for example that genes switch between near zero (silent) and near maximal (active) transcription rates, with the probability of being active responding to the input transcription factors. These switching dynamics are essentially universal across all four gap genes. Relatedly, the maximum number of mRNA molecules that one finds in cell–sized volumes surrounding a nucleus also is the same for all

[13] More subtly, the embryo is not cylindrically symmetric, so you also have to be careful not to mix measurements from different orientations. For a full discussion of all these concerns see Ref [65].

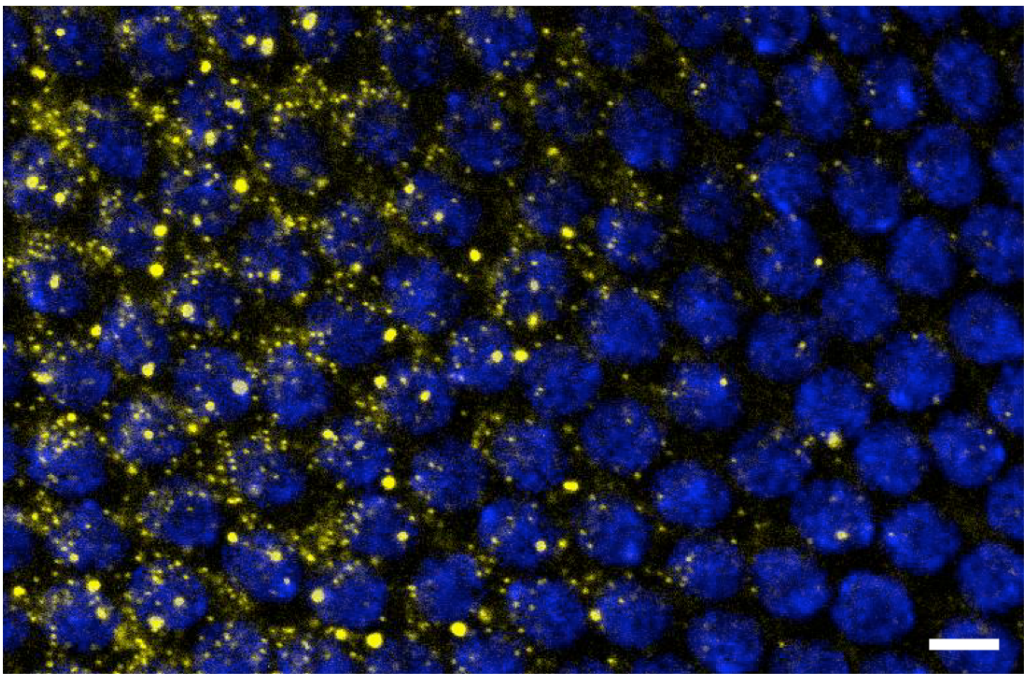

Figure 5: Single molecule mRNA detection in a small region of the *Drosophila* embryo. Yellow probes are complementary to the mRNA of the *hb* gene (protein concentration is shown as g_1 in Fig 11) and all nuclei are labeled in blue. Detailed analysis shows that all the spots outside the nuclei have brightness drawn from a narrow distribution, and a size consistent with the point spread function of the microscope; these and other observations indicate that they are single molecules. The one or two brighter spots inside the nuclei are points at which mRNAs are being transcribed, as explained in the text. Expression is larger toward the left, corresponding to the anterior of the embryo. Scale bar 5 μm. Image courtesy of T Gregor, from the experiments of Ref [66].

the gap genes. These kinds of absolute statements would have been impossible not so long ago, and this is just a start.

Before leaving the fly, let me note that one can adapt these live fluorescence experiments for other purposes, probing the nanoscale molecular events that underlie the control of concentrations and flow of information through gene networks. As an example one can label a point along the DNA close to the start of transcription for one gene (the promoter), introduce stem loops into the sequence of that gene, and also label a point along the DNA where regulatory proteins bind (an enhancer). With these three labels you can measure the distance between the promoter and enhancer, and find that these must be in close proximity in order for transcription to start [74]. Higher resolution versions of these experiments show that proximity is not contact, so that even when transcription is active the promoter and enhancer are separated by ~ 150 nm [75]. We don't know how this apparent action at a distance is achieved.

It is important that the ability to do physics experiments in living systems extends far beyond measuring protein concentrations and counting mRNA molecules. Indeed, studying systems in which signals are carried by changing concentrations of particular molecules is challenging in part because monitoring each different species of molecule could require a different probe, especially if we want to make real time measurements on live cells. In contrast, cells in the brain communicate by generating voltage differences across their membranes, and of course if we can record voltage in one cell we can in principle record from all cells with the same methods. The currents that support transmembrane voltage changes are large enough

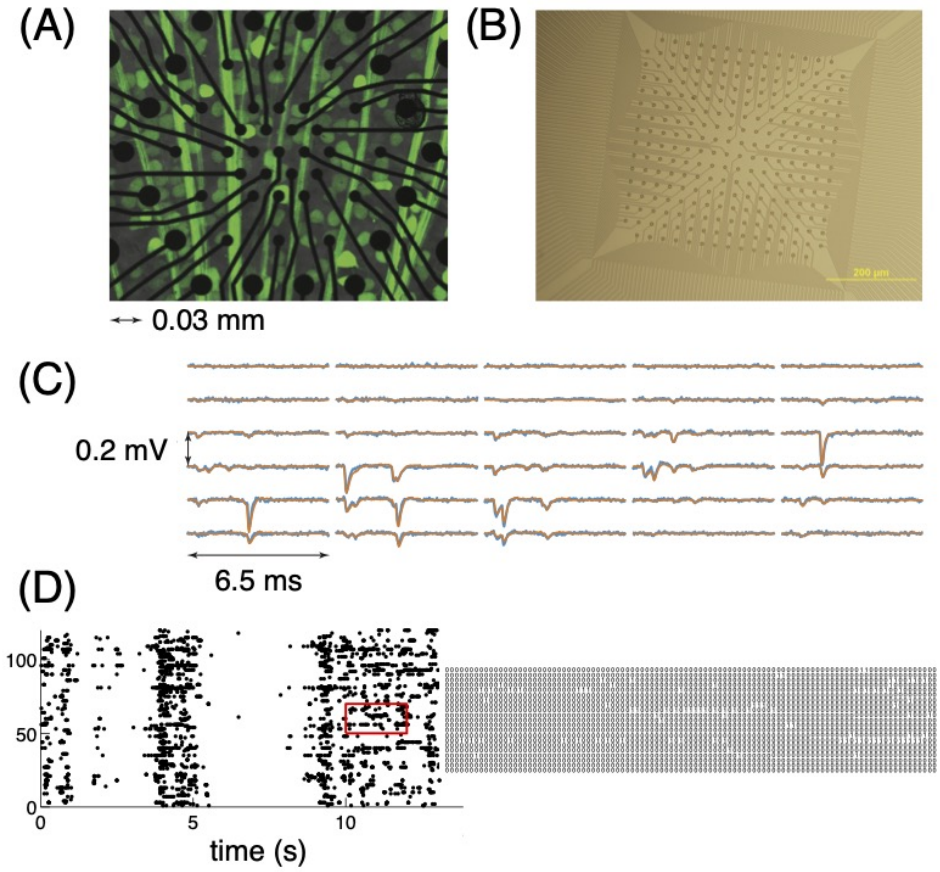

Figure 6: Recording action potentials from retinal ganglion cells. (A) Salamander retina on an array of electrodes. The electrodes, and the leads that carry signals away from the electrodes, are black features on a transparent slide; the ganglion cells of the retina have been filled with a green dye. Round green objects are cell bodies, and long lines are bundles of axons that eventually converge to form the optic nerve. Note that the number of electrodes is comparable to the number of cells (B) The next generation of electrode arrays. (C) Voltage traces from a selection of these electrodes during an experiment on the salamander retina. Blue traces are the actual voltages, and orange traces are a reconstruction of the voltages as a superposition of stereotyped waveforms—action potentials (spikes) from individual neurons—learned from a different part of the data. (D) At left, a raster plot of the action potentials found in (C) for 100+ cells, a dot showing the time of a spike from each neuron. At right, expanded version of the red box, with responses expressed as binary (spike/silence) words in 20 ms bins. (A–C) reproduced from Ref [57], with permission; data from experiments by MJ Berry II and colleagues, including D Amodei, O Marre, JL Puchalla, and R Segev, with thanks [80, 82].

that they result in voltage differences that are measurable in the salty water outside the cell; in particular many cells generate discrete pulses termed action potentials or spikes, and these are easier to identify. Action potentials from single neurons were first measured nearly one hundred years ago [76]. The exploration of electrical activity in single neurons has been a source of productive interactions between physics and biology, exploiting very sensitive electronics and shaping theoretical ideas about the microscopic mechanisms of this activity [77, 78] and about the abstract structure of spike sequences as a code for sensory inputs and motor outputs [79].

There has been dramatic progress in our ability to monitor the activity of many single neurons simultaneously, where "many" began with ~ 10 and now is $\sim 10^5$ and soon 10^6. An early strategy was to focus on relatively flat pieces of brain tissue, such as the retina, which can be dissected out and placed onto an array of electrodes (Fig 6). This led to experiments monitoring essentially all of the hundreds of cells that provide the brain with information about a small patch of the visual world as the retina is driven by complex visual inputs, including fully naturalistic movies [80–82]. There are three–dimensional arrays of electrodes that can be inserted into thicker tissue [83, 84], and most recently there are flexible polymer based electrodes [85]. An alternative to direct electrical measurements is to genetically engineer organisms so that neurons make proteins whose fluorescence is sensitive to electrical activity. The ideal would be to have proteins that insert into the cell membrane and report directly on trans–membrane voltage, but these have developed slowly [86–89]. Much better established are fluorescent proteins that respond to changes in calcium concentration inside the cell, which are (slower) corollaries of electrical activity [90–92].

Fluorescent calcium indicators turn the problem of recording from large numbers of neurons into a problem of imaging (Fig 7). To reach cellular resolution requires sophisticated microscopy methods, often built around scanning two–photon microscopy [93, 94], with recent developments involving more advanced optics to allow better access to depth [95–97]. To obtain high resolution images it is easiest if the brain is not moving, but many aspects of brain function are tightly coupled to behavior, and one solution is to construct virtual reality for experimental animals [98]. The combination of genetic engineering, state of the art microscopy, and virtual reality in these experiments is impressive. In particular, note in Fig 7C the very quiet baseline in individual cells, which demonstrates the generally low noise levels that can be achieved with these methods. These tools are being used in a wide variety of organisms, from worms to mammals, and in many different regions of the brain devoted to different tasks.

For smaller animals, such as the worm *C elegans* or the larval zebrafish, we are getting close to recording from every single neuron in the brain at reasonable time resolution [99, 100]. The same techniques of genetic engineering that have led to calcium–sensitive fluorescent proteins have also led to proteins that act as light–sensitive ion channels [101–103]. Combining these tools means that one can both inject currents into neurons and record their responses, all using light and working at single cell resolution. In *C elegans* this has led to a nearly complete "pump–probe" experiment probing signal transfer among all $\sim 10^4$ pairs of neurons [104].

It had long been possible to do physics experiments on isolated parts of living systems, culminating in the observation and manipulation of single molecules [107–109], but what has emerged more recently is the ability to tame complexity and do physics experiments on an ever wider range of intact, functioning systems. The genetic and neural networks discussed here are good examples, but one can also look to measurements on flocks of birds and swarms of insects [110, 111], populations of bacteria [112, 113], pattern formation in groups of stem cells [114, 115], and much more.[14] While many systems remain to be tamed, the enormous variety of systems where it is possible to do physics experiments has removed a major obstacle to theorizing. Importantly, as data improve we should expect more from our theories.

1.4 Agenda

This explosion of experimental developments obviously creates a need for new methods of data analysis. Indeed, as these approaches to high dimensional data collection have penetrated the mainstream of biology, the biology community itself has emphasized the urgency of this need for mathematical analysis. But for physicists theory is more than data analysis. The most powerful analyses are grounded in theories, and there is a strong case that *all* analysis

[14] Again, references are illustrative rather than exhaustive. Broader coverage, including historical context, can be found in the first Decadal Survey of our field [116].

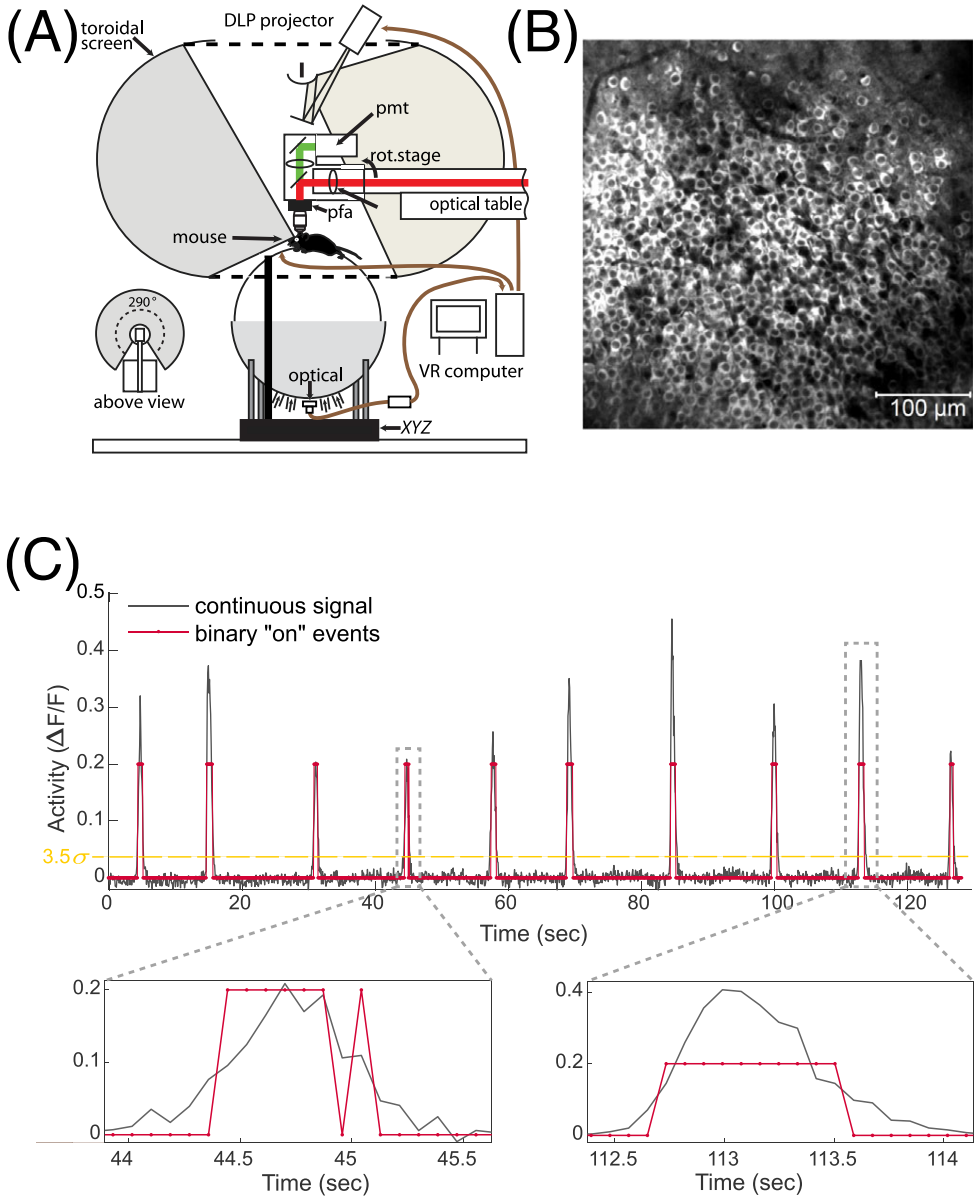

Figure 7: Monitoring electrical activity in the brain by imaging of calcium sensitive fluorescent proteins as a mouse moves in virtual reality. (A) Schematic of the experiment, with mouse (in black) running on a styrofoam ball. Motion of the ball advances the position of a virtual world projected on a surrounding toroidal screen, and a scanning two–photon microscope is focused on a single layer of cells in the hippocampus. (B) Fluorescence image of these neurons expressing calcium sensitive fluorescent protein. (C) Integrated fluorescence from a single cell vs time. Note the very quiet baseline interrupted by discrete events. These can be binarized, and (as shown in the insets) we understand enough about the dynamics of the indicator molecule to assign unusually slow transient decays as flickering off and on. (A) and (B) reproduced from Ref [105]; (C) reproduced from Ref [106], with permission. With thanks to L Mehsulam, JL Gauthier, CD Brody, and DW Tank.

methods embody some theoretical prejudice.[15] Something we do well in physics is to make these theoretical prejudices explicit.[16] Returning to the fly embryo, as we discussed at the outset any attempt to give a realistic quantitative description immediately leads into a dense forest of parameters. Do we have clues about what principle(s) might cut through this complexity?

The system is extraordinarily precise. It has been thought for decades that just ~ 3 hrs after the egg is laid, every cell along the anterior–posterior axis knows its fate [119]. Apparently the genetic network schematized in Fig 2 carries enough information to do this. Since cell fates are tied to positions, and there are fewer than 100 rows of cells along the anterior–posterior axis, this means that the concentrations of just a handful of molecules specify position with $\sim 1\%$ accuracy. This can be made explicit, e.g. by measuring the reproducibility of the pair–rule stripe positions as in Fig 8 [120, 121].

The concentrations of relevant molecules are low. Almost all the molecules in the network of Fig 2 are transcription factors—proteins that bind to specific sites along DNA and regulate the expression of other genes, in this case other genes in the network. Again it has been known for decades that transcription factors function at concentrations measured in nanoMolar [122–124], and there is no reason to doubt that this is true of the relevant molecules in the embryo [125]. Cell nuclei have dimensions measured in microns, and $1\,\mathrm{nM} = 0.6\,\mathrm{molecules}/\mu\mathrm{m}^3$, so even the absolute numbers of molecules can't be very large.

These two facts (here somewhat stylized) might be in conflict—at low concentrations things are noisy (because of physics not biology), and it is hard for molecules to convey much information. On the other hand there might be a principle that reconciles the conflict: parameters of the relevant networks have been selected to convey as much information as possible from these physically limited signals. This principle will be the focus of what follows.

Before getting started let me acknowledge that optimization principles engender strange reactions. For some, optimization is obvious because evolution has had billions of years to get things right. For others, optimization is nonsense because evolution is not about being best, but about being better than the competition. Things get worse when we are optimizing abstract quantities such as information—why should the organism care about bits? These discussions can devolve into debates about beliefs rather than evidence. Nobody knows how to do a calculation that weighs the benefits of optimizing performance (e.g., counting single photons in vision) against the costs of the underlying mechanisms (energy dissipation in the biochemical amplification of single molecular events), and we certainly don't know enough to calculate how long it would take evolution to find the optimal tradeoff.

Optimization comes along with an aesthetic that you might or might not find appealing. But what matters is that optimization principles make quantitative predictions that can be tested in modern experiments.[17] In many cases, as we will see, these predictions are essentially parameter free and accomplish the goal of circumventing highly parameterized models. Importantly, the claim that information flow is being optimized can be tested by measuring the information flow itself, in bits or as an effective noise level against which independently measured signals must be compared in order for the system to function. If we go back to the picture of information flow from maternal inputs through the gap genes to the pair–rule stripes, we can identify three distinct opportunities for optimization.

[15] In particle physics, for example, the signals from thousands of detector elements are reduced to a plot of the rate for some class of events vs some energy variable. An enormous amount of theoretical understanding is contained in the idea that this is the right thing to plot, even before we ask theory to predict what the plot shows.

[16] At the risk of being pedantic, consider the simple idea that high dimensional data—expression levels of hundreds of genes, electrical activity of thousands of neurons, and more—live in low dimensional spaces. This seems "theory free," testable by standard methods for linear [117] or nonlinear [118] dimensionality reduction. But low dimensionality is a theoretical claim: what is the principle that limits the dimensionality of the dynamics? More subtly, to measure how well a low dimensional description works, we need a metric, and this metric is a theoretical claim about which variations are most relevant, or perhaps which variations are measured most reliably.

[17] See also Chapter 3 in Ref [57].

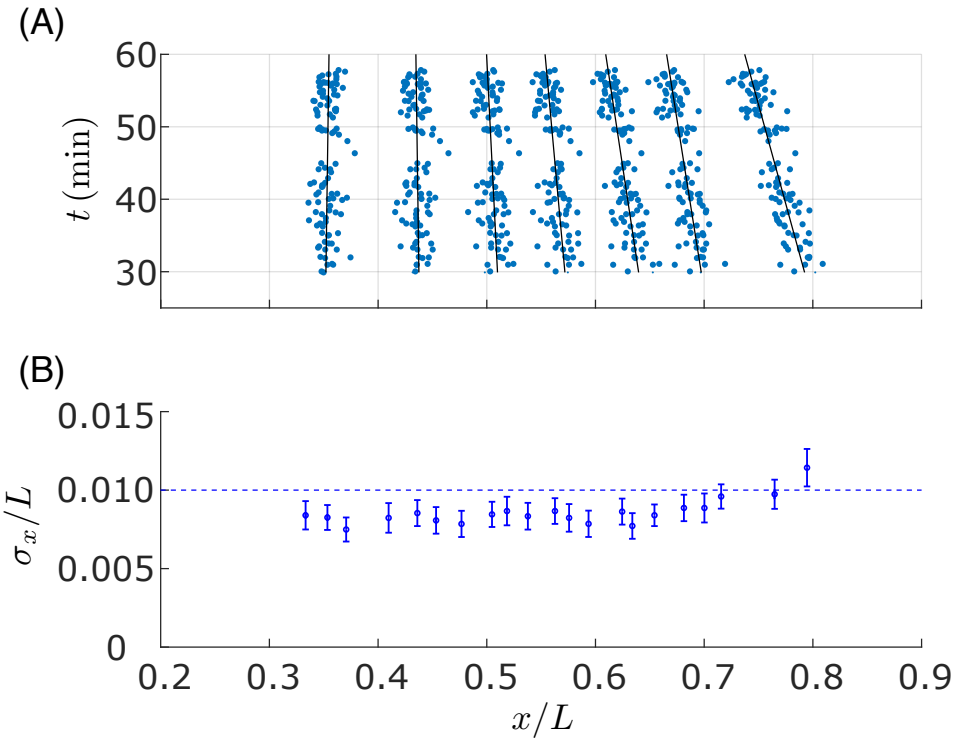

Figure 8: Pair–rule stripes have positional noise $\sigma_x/L \sim 0.01$. (A) Positions of the seven pair–rule stripes for the gene shown at bottom in Fig 2 drift with time during the fourteenth cycle of nuclear divisions, illustrated here for 100+ embryos. (B) Once corrected for drift, the variance of stripe positions is very small. Results for seven stripes from three different pair–rule genes (Fig 1, bottom); error bars are standard errors of the mean. Dashed line is a positional noise equal to one percent of the embryo length. Redrawn from Refs [120,121], with thanks to H Casademunt, JO Dubuis, T Gregor, L McGough, M Nikolić, MD Petkova, G Tkačik, and EF Wieschaus.

Optimal decoding. Information about position is encoded in the concentrations of the four gap gene products. Given the (measured) noise levels in these concentrations, there is an optimal strategy for decoding this information, extracting an estimate of position that is as accurate as possible. Does this optimal precision match the observed $\sim$ 1% precision of the pair–rule stripes? Can we test whether the embryo really implements something functionally equivalent to the optimal decoding algorithm?

Matching distributions. The information that a system's output provides about its inputs depends not only on the internal dynamics and noise level in the system, but also on the distribution of its inputs. In transmitting positional information, the embryo can't choose the distribution of positions, but it can adjust the representation of position by the maternal morphogens, which provide the direct input to the gap gene network. How should these inputs be adjusted to optimize information transmission? Can we see signatures of this optimization?

Network architecture. Finally, we can go back to the 50+ parameters of the gap gene network and ask if these parameters have been set to optimize information transmission. This is a hard problem, and we have tried to make progress by breaking off small pieces. Recently there has been a major step forward, optimizing the whole gap gene network over a class of models that is (almost) rich enough to include the real network. We are coming close to deriving the properties of this network from a general physical principle, with no free parameters.

We will take these problems in sequence, and in each case we'll see how the same principles are relevant to very different systems. The first efforts to use these principles came in the context of neural information processing, and the idea that neural and genetic networks might be shaped by common principles is appealing in itself. I take the liberty of including a few subsections that I didn't have time for in the lectures, and hopefully these provide more context.

2 Optimal decoding

If we think that the genetic regulatory networks in the embryo have been selected to transmit as much positional information as possible while using a limited number of molecules, it would be very odd if the cells then used this information inefficiently. So it is reasonable to ask how the embryo could use the concentrations of the gap proteins to draw the most reliable inferences about position, and whether we can find signatures of this optimization. We'll need to build up some tools to answer these questions.

2.1 A warmup exercise

The essential problem of inference is that we can only measure things that are related to what we care about, we can't get direct access. This is familiar in the physics lab, where (for example) we measure currents in response to applied voltages to estimate the resistance of a material. What we measure in this case is proportional to what we want to know, but also noisy. So the simplest version is that we want to know x but we measure y, and these are connected by

$$y = x + \eta \,. \tag{2}$$

If, as often is the case, the noise is Gaussian with zero mean, then

$$P(y|x) = \frac{1}{\sqrt{2\pi\langle\eta^2\rangle}} \exp\left[-\frac{(y-x)^2}{2\langle\eta^2\rangle}\right] . \tag{3}$$

This predicts the values of y that we will observe if we control x. But the problem we face is that x varies in ways outside our control, and we would like to infer these variations based on measurements of y. Everything that we can say about this inference problem is contained in the conditional probability distribution $P(x|y)$.[18]

We recall that what we need can be constructed from Bayes' rule:

$$P(x,y) = P(y|x)P_X(x) = P(x|y)P_Y(y) \tag{4}$$

$$\Rightarrow P(x|y) = \frac{1}{P_Y(y)} P(y|x)P_X(x) \,. \tag{5}$$

Just to be clear, $P(x,y)$ is the probability (density) for observing particular values of x and y together, while $P_X(x)$ and $P_Y(y)$ are the probabilities for observing each value independent of the value for the other.[19] Bayes' rule tells us that to make inferences we combine the data, which sits in $P(y|x)$, with prior knowledge or expectations, encoded in $P_X(x)$. Let's try assuming that x is drawn from a Gaussian distribution with zero mean, so that

$$P(x|y) = \frac{1}{P_Y(y)} \frac{1}{\sqrt{2\pi\langle\eta^2\rangle}} \exp\left[-\frac{(y-x)^2}{2\langle\eta^2\rangle}\right] \frac{1}{\sqrt{2\pi\langle x^2\rangle}} \exp\left[-\frac{x^2}{2\langle x^2\rangle}\right] . \tag{6}$$

[18] I am a little embarrassed that what I give here as a warmup exercise also appeared in my Les Houches lectures 20+ years ago [126]. It's still a good exercise, but I'll go more quickly, as I did in the 2023 lectures.

[19] I am being a bit pedantic to emphasize that the two marginals are different functions. I think most of us would write $P(x)$ and $P(y)$ when calculating in private, and trust that we can keep things straight.

We could work a little harder on the algebra, but you can see that this is a Gaussian function of x. The mean is the same as the most likely value, which we can find from

$$0 = \frac{\partial}{\partial x}\left[\frac{(y-x)^2}{2\langle\eta^2\rangle} + \frac{x^2}{2\langle x^2\rangle}\right]_{x=x_*} \tag{7}$$

$$\Rightarrow x_* = \frac{\langle x^2\rangle}{\langle x^2\rangle + \langle\eta^2\rangle} y\,. \tag{8}$$

This most likely value of x given our observation of y is one definition of the "best estimate." Another definition is to find the estimator $x_{\rm est}(y)$ which makes the smallest mean–square error

$$\chi^2 = \int dx \int dy\, P(x,y)[x_{\rm est}(y) - x]^2\,. \tag{9}$$

The best estimate in this sense, that is the solution to $\delta\chi^2/\delta x_{\rm est}(y) = 0$, can be found for arbitrary distributions, and is equal to the conditional mean,

$$x_{\rm est}^{\rm opt}(y) = \int dx\, x P(x|y)\,. \tag{10}$$

In the case where both the signal x and the noise η are Gaussian, as above, then these two different definitions of the optimal estimate agree. More generally if different but plausible definitions of "best" lead to significantly different estimators, it probably is a sign that $P(x|y)$ has a complicated structure, such as multiple peaks, so that inference is not just noisy but genuinely ambiguous.

Notice that with Gaussian signals x and Gaussian noise η, and a linear input/output relation for the $x \to y$ transformation, the optimal estimate of x from y also is linear. This doesn't generalize. Suppose that

$$P_X(x) = \frac{a}{2} e^{-a|x|}\,. \tag{11}$$

Then

$$P(y|x) \propto \exp\left[-a|x| - \frac{(y-x)^2}{2\langle\eta^2\rangle}\right], \tag{12}$$

and one can see that the most likely x is a thresholded function of y,

$$x_*(y) = 0\,, \qquad |y| < a\langle\eta^2\rangle\,, \tag{13}$$

$$x_*(y) = y - a\langle\eta^2\rangle \mathrm{sgn}(y)\,, \qquad |y| > a\langle\eta^2\rangle\,. \tag{14}$$

If we compute the conditional mean then the threshold is softened but the optimal estimator still is nonlinear. This emphasizes that the statistical structure of the inputs can shape the *qualitative* structure of the estimator.

One other point is that the optimal estimator is not a perfect estimator. There is noise, which leads to random errors, but also there are systematic errors. If you work through, you can see that (for example) Eq (8) describes an estimator that systematically underestimates the magnitude of x, and this also is obvious in Eqs (13, 14). There is a tradeoff between systematic and random errors, and there is a best tradeoff, but no way to escape from both.

With these remarks in mind, let's do a real problem.

2.2 Counting photons and estimating motion

The ability of visual systems to count single photons remains, for me, one of the most striking facts about the physics of life. There is an obvious sense in which this provides an example of near optimal performance, close to the limits of what the laws of physics allow. To reach this level of performance one has to think about physics at many scales [57]:

- the dynamics of the rhodopsin molecules, where photon–driven transitions are so fast that they compete with the loss of quantum coherence, and spontaneous transitions are so slow that individual molecules are stable for a thousand years;
- the biochemical mechanisms of amplification, which allow the photoreceptor cell to "smell" one rhodopsin molecule out of one billion that has changed structure in response to photon absorption, and generate a macroscopic response;
- the circuitry of the retina, which preserves and processes the single photon responses of individual receptor cells amid a sea of noise from other cells;
- and neural computations that combine single photon signals with prior expectations, for example to compensate for long delays in the retinal response.

There is much to discuss here, much of it now classical but still some questions remain open. I want to focus on one part of the retinal circuitry problem.

Let's consider the limit where each receptor cell i receives either zero or one photon. Each photon triggers a rather reproducible pulse of current, and we can choose units in which this pulse has unit amplitude. Then the signal from each cell becomes

$$y_i = n_i + \eta_i \,, \tag{15}$$

where $n_i = 0$ or 1 is the number of photons and η_i is a Gaussian noise source with variance σ^2. While $\sigma \ll 1$, there is a problem in combining signals from many cells. As an example, the initial experiments showing that the statistics of human responses to dim light flashes are consistent with photon counting involved flashes that delivered ~ 5 photons distributed over ~ 500 receptor cells. In some species we know that integration over such a large area happens as receptor cell signals cross the first synapse, converging on the bipolar cell. If the bipolar cell just adds up the signals, then the fluctuations in the sum are $> 20\sigma$, and now single photon signals will be lost.

Although y_i is on average equal to n_i the best estimate of photon count is not the receptor output itself. Following the argument above, we have

$$P(n_i|y_i) = \frac{1}{P(y_i)} P(y_i|n_i) P(n_i) \,, \tag{16}$$

$$P(y_i) = \sum_{n_i} P(y_i|n_i) P(n_i) \,. \tag{17}$$

From Eq (15) we have

$$P(y_i|n_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - n_i)^2/2\sigma^2} \,. \tag{18}$$

Further, since the lights are low the only nonzero values for $P(n_i)$ are $P(0)$ and $P(1)$, so that

$$P(n = 1|y) = \frac{P(1)\exp\left[-(y-1)^2/2\sigma^2\right]}{P(1)\exp\left[-(y-1)^2/2\sigma^2\right] + P(0)\exp\left[-(y)^2/2\sigma^2\right]} \tag{19}$$

$$= \frac{1}{1 + e^{-\beta(y-\theta)}} \,, \tag{20}$$

which is a sigmoidal function. In this regime the conditional mean of n_i, which we recall is the best estimate in the least squares sense, is just $P(n_i = 1|y_i)$. So the best estimate of the photon count is the sigmoidal function in Eq (20), with "threshold"

$$\theta = \sigma^2 \ln\left[\frac{P(0)}{P(1)}\right] + \frac{1}{2} \,, \tag{21}$$

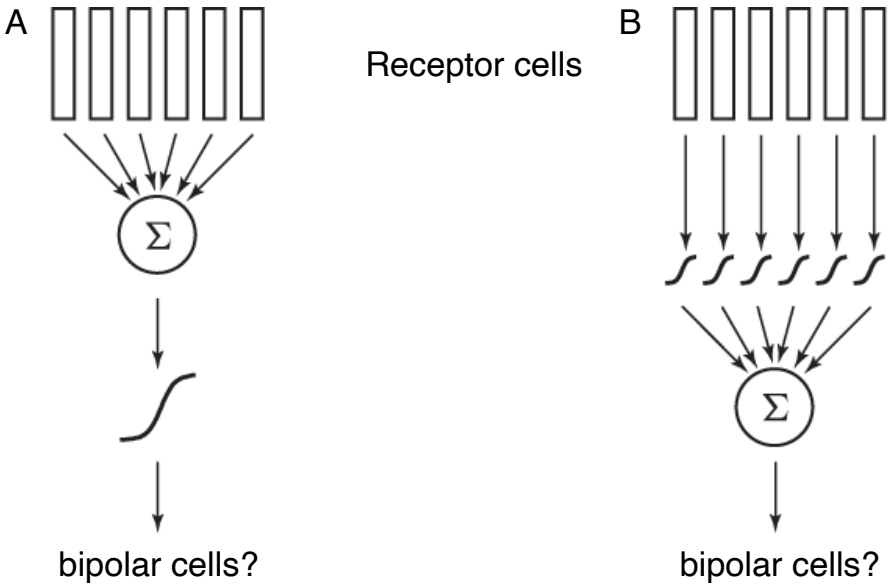

Figure 9: Two models of integrating single photon signals at the first synapse in the retina [57]. (A) Receptor cell signals are summed, and the target (bipolar) cell applies a nonlinearity, as in conventional neural network models. (B) Each receptor cell signal is passed through a nonlinearity, as suggested by Eq (20), and the resulting estimates of photon count are summed.

and sensitivity

$$\beta = \frac{1}{\sigma^2}\,. \tag{22}$$

After passing through this near threshold nonlinearity, we can sum the signals from many cells with much less sensitivity to the background noise.

We can think of these arguments as contrasting two models, shown in Fig 9. In (A) we see something like the conventional neural network model: inputs to a single neuron sum linearly, and a sigmoidal nonlinearity is applied after the summation [24, 27]. In (B) we have the opposite, where nonlinearities are applied to each input signal as it crosses the synapse, and these processed signals are summed. Despite expectations, optimal estimation of photon counts requires something more like (B). Happily, these synaptic nonlinearities have been detected directly in the mouse retina, in at least semi–quantitative agreement with theory [127]; importantly these nonlinearities do not appear at synapses to bipolar cells which are not involved in processing single photon signals. This discussion has emphasized instantaneous nonlinearities, but temporal filtering can also help to separate single photon signals from noise at this first synapse [128].

In addition to detecting light and estimating its intensity, the brain of course draws many more sophisticated inferences from its visual inputs. An interesting example, especially appealing for physicists, is the inference of movement. The fact that we have the *appearance* of motion from discrete flashing lights is also among the foundational observations of gestalt psychology. We can study visual motion perception in humans, and the behavioral responses of insects to visual motion, and there are striking similarities across this enormous evolutionary distance. In particular, humans and insects make similar systematic errors in estimating movement velocities, especially in scenes with low contrast.

One might think that estimating movement velocity is not so hard. Let's work in one dimension (which admittedly hides some important issues), so the image that falls on our

retina is $I(\phi, t)$ where ϕ is the azimuthal angle. If a small patch of the visual world is moving relative to us at velocity v, then we should have

$$I(\phi, t) = I_0(\phi - vt) . \tag{23}$$

This suggests that

$$v_{\rm est} = v_{\rm deriv} = -\frac{\partial I/\partial t}{\partial I/\partial \phi} , \tag{24}$$

provides a direct estimate of velocity as the ratio of temporal and spatial derivatives. This is overly optimistic, because we have neglected noise; more subtly we have neglected any dynamics in the image that cannot be ascribed directly to movement. In the presence of noise it is dangerous to differentiate, because noise typically extends to higher frequencies than the signal, and it is dangerous to divide, because the denominator might fluctuate to zero. In Eq (24) we commit both these sins.

Equation (24) says that, in the presence of motion, the spatial and temporal derivatives should be proportional to one another. A gentler statement is that these derivatives are correlated, and the strength of this correlation should be related to the movement velocity. This suggests that we might be able to estimate

$$v_{\rm est} = v_{\rm corr} \propto \frac{\partial I}{\partial t} \times \frac{\partial I}{\partial \phi} . \tag{25}$$

Notice that if we double the variations in light intensity, then this estimate will increase by a factor of four, unless we do some sort of normalization. But this confounding of contrast and velocity is one of the systematic errors made by humans and insects alike, at least at low contrast. The idea that brains estimate motion based on spatiotemporal correlation goes back to classical experiments on insect behavior [129, 130] and reappears decades later as a model of "motion energy" in human and non–human primate vision [131, 132].

It is especially attractive to study visual motion estimation in flies because there is a very accessible and beautifully laid out population of neurons that encode these estimates and ultimately drive behaviors such as flight control [133–135]. In this system one can also give a detailed characterization of signals and noise in the photoreceptors and second order neurons, including evidence that photon shot noise is dominant at counting rates up to $\sim 10^6\,{\rm s}^{-1}$ [136, 137]. By the early 1990s, a theory/experiment collaboration with Rob de Ruyter van Steveninck had shown that sequences of action potentials from the motion sensitive neurons in flies encoded motion estimates with a reliability close to the physical limits set by noise in the photoreceptors and diffraction blur in the compound eye [138, 139].

Motivated by the observation of near–optimal performance, we developed a theory of optimal motion estimation following the lines sketched above [140]. We were excited that, as a function of signal strength or noise level the optimal estimator interpolated between something like the ratio of derivatives in Eq (24) and the correlator in Eq (25). This suggested that some of the systematic errors of real visual motion estimation might actually be features of the optimal estimator. The problem is that the detailed form of the optimal estimator depends, as expected from §2.1, on the statistical structure of the visual inputs, and in the early work we just had to guess at this. In particular, depending on this structure, the crossover between correlator and ratio estimators might or might not occur in a regime that is relevant for real brains under natural conditions.

To make progress let's focus on motion estimates derived from very small patches of the visual world; in the fly this could just mean a handful of neighboring receptors on the lattice of the compound eye. Rob and his colleagues built a special purpose camera that samples the visual world at high frame rates and with optics that matches those of the compound eye [141]. From a long walk in the woods, one can derive many samples of the local image derivatives

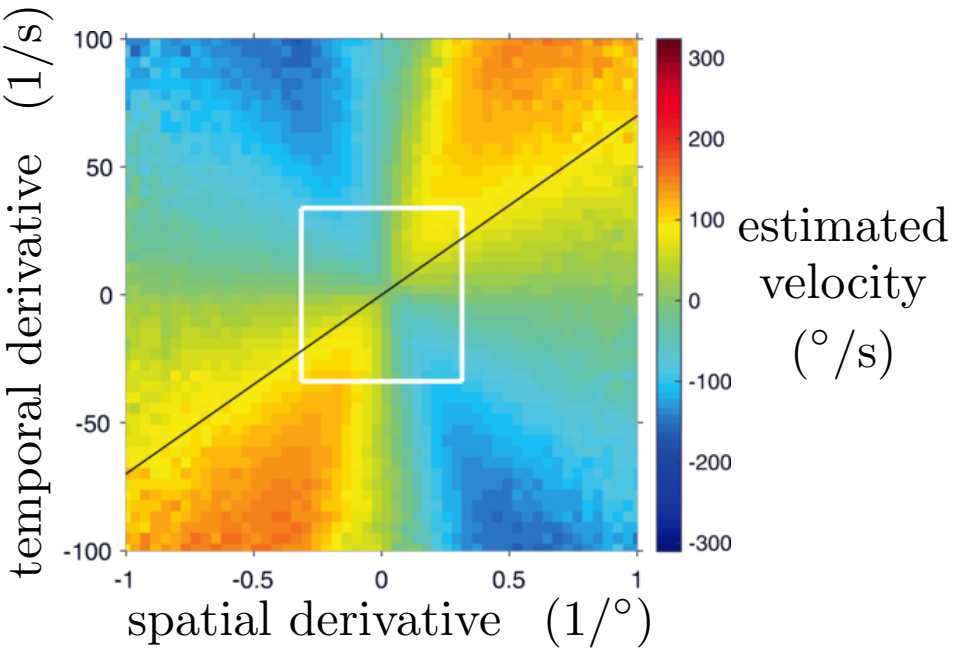

Figure 10: Estimation of visual motion from local derivatives. The optimal estimator in Eq (26) is computed from data collected by a special purpose camera during a walk in the woods. Black line indicates the predictions of Eq (24) for $v_{\rm deriv} = 70\,^\circ/\mathrm{s}$. White square encloses 90% of the data. Redrawn from Ref [141], with thanks to S Sinha and RR de Ruyter van Steveninck.

and the angular velocity of the camera at the same moment in time. To remove the (largely irrelevant) absolute light level we can take the log intensity as the raw data, so we can think of this experiment as providing an estimate of the joint distribution $P(\partial_t \ln I, \partial_\phi \ln I, v)$. Then we can form the conditional distribution $P(v|\partial_t \ln I, \partial_\phi \ln I)$, and finally the optimal estimator

$$v_{\rm est}(\partial_t \ln I, \partial_\phi \ln I) = \int dv\, v P(v|\partial_t \ln I, \partial_\phi \ln I)\,. \tag{26}$$

Results are shown in Fig 10.

If the optimal estimator were the ratio of derivatives, then contours of constant $v_{\rm est}$ would be straight lines in Fig 10, as with the black line at $70\,^\circ/\mathrm{s}$. We see that this is a decent description of the estimator at large spatial and temporal derivatives. On the other hand, the correlator model predicts that contours of constant $v_{\rm est}$ are hyperbolic, and this curvature is what we see at small values of the derivatives; we can make this clearer by taking different slices through the data [141]. If we test the system with a rigidly moving spatial pattern then in the small derivative regime the optimal estimator will be systematically wrong, with these errors arising as a by product of insulation against random errors. What we have called "small" and "large" can be read from Fig 10, but importantly real derivatives are mostly small: the white box shows a range of spatial and temporal derivatives that contains 90% of the samples collected on an hour long walk through the woods. These results provide direct evidence that motion estimation in a naturalistic context really is in the regime where correlation is optimal.

There is much more to be done. We have not yet added back the effects of photon shot noise[20] and other sources of noise in the receptor cells; these will widen the dynamic range

[20] The camera is built to have a much larger collecting area than the fly's receptor cells, so the signals analyzed here are essentially noiseless.

over which the correlator model is optimal. Asymmetries in the underlying distributions should lead to asymmetries in the optimal estimator, which are barely visible in Fig 10 and should be connected to the separate processing of on and off signals [142, 143]. If the walk takes us through regions of very different statistical structure we may be able to divide the data accordingly and predict adaptation of the optimal estimator to the input statistics, perhaps connecting to adaptation seen in the responses of motion–sensitive neurons. It also will be interesting to understand the rules for optimal combination of these local estimators into wide field motion signals.

After many decades we have gotten used to the idea that the visual system can count single photons, and perhaps we forget that this provides evidence for optimal performance—functional behavior near the limits of what is allowed by the laws of physics. It is tempting to think that such physical limits are irrelevant to vision on a bright sunny day, but Fig 10 shows us that this is not true. From data collected literally at noon, we see that the physical structure of visual input is such that to make maximally precise estimates the brain must do unexpected things, including making systematic errors of the same form made by humans and insects. These errors are driven by physics, not by biological limitations.

2.3 Concentration measurements, revisited

How do these ideas play out in the fly embryo? Roughly three hours after the egg is laid, individual cells have access to the concentrations (expression levels) of the four proteins encoded by the gap genes. In order to do the right thing, cells need to know where they are in the embryo. So it is natural to ask how a cell could infer its position from the gap gene expression levels. This idea that cells extract positional information from the concentration of specific morphogen molecules is very old [144]. The fact that in the fly we can identify *all* the relevant molecules gives us a chance to turn these words into mathematics.

Figure 11 shows the spatial profiles of the gap gene expression levels along the long (anterior–posterior) axis of the embryo.[21] These data are extracted from the fluorescent antibody staining experiments discussed in §1.3. We will refer to the concentration of each molecule at position x as $g_i(x)$, with the index $i = 1, 2, 3, 4$. Solid lines show the mean concentrations $\langle g_i(x)\rangle$, with cyan shading to indicate the standard deviation of fluctuations around this mean.[22] An important qualitative observation is that these fluctuations in fact are quite small. These data sets are large enough that we can estimate, reliably, the 4×4 covariance matrix of fluctuations at each point,

$$\left[\hat{C}(x)\right]_{ij} = C_{ij}(x) = \langle \delta g_i(x) \delta g_j(x)\rangle \,, \tag{27}$$

where as usual

$$\delta g_i(x) = g_i(x) - \langle g_i(x)\rangle \,. \tag{28}$$

If we can make the approximation that fluctuations δg are Gaussian, then armed with these measurements we can write the probability distribution

$$P(\{g_i\}|x) = \frac{1}{Z(x)} \exp\left[-\frac{1}{2}\chi^2(\{g_i\};x)\right], \tag{29}$$

$$\chi^2(\{g_i\};x) = \sum_{ij} \delta g_i(x) \left[\hat{C}^{-1}(x)\right]_{ij} \delta g_j(x) \,, \tag{30}$$

$$Z(x) = \sqrt{(2\pi)^4 \det \hat{C}(x)} \,, \tag{31}$$

[21] I tried to give this course without mentioning the names of these (and other) molecules, because I don't think it matters. But, to connect with the literature, the names are (1) Hunchback, (2) Krüppel, (3) Giant, and (4) Knirps.

[22] Recall from Fig 4 that we choose units such that $\langle g_i(x/L)\rangle$ runs between 0 and 1 for each gene.

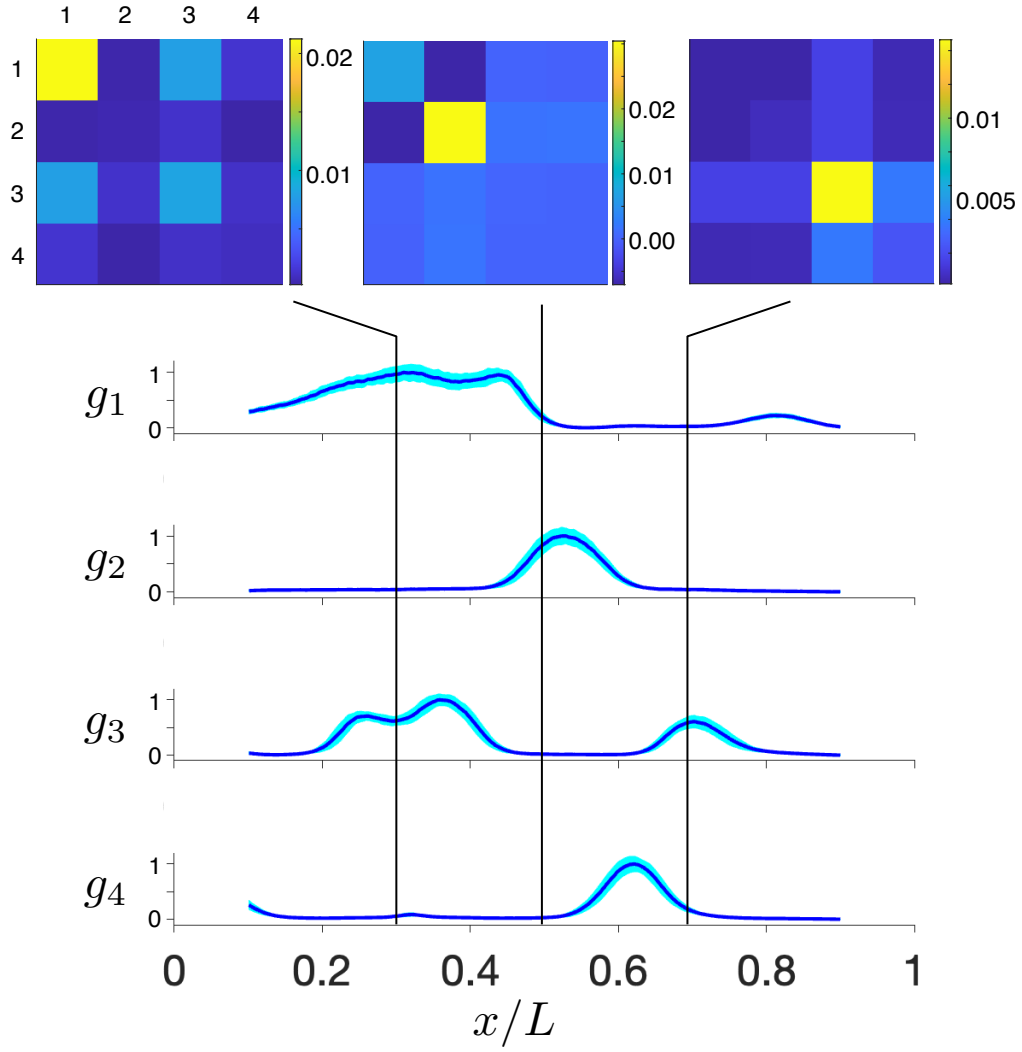

Figure 11: Expression levels of the four gap genes as a function of position along the long (anterior–posterior) axis of the embryo. Bottom panels show the means $\langle g_i(x)\rangle$ (blue lines) and standard deviations $\sqrt{\langle[\delta g_i(x)]^2\rangle}$ (cyan shading). Top panels show examples of the covariance matrix $C_{ij}(x) = \langle \delta g_i(x) \delta g_j(x)\rangle$ at three positions, $x/L = 0.3, 0.5$, and 0.7. Data from Ref [10], with thanks to T Gregor, MD Petkova, G Tkačik, and EF Wieschaus.

where $\hat{C}^{-1}$ is the inverse of the matrix $\hat{C}$ and $\det \hat{C}$ is its determinant.

As is in the examples above, the problem facing the embryo is inverse to the problem we face in characterizing the patterns of gene expression. A cell (again, more precisely, a nucleus) has access to the concentrations $\{g_i\}$ and must infer its position x. Everything that can be known about x by observing $\{g_i\}$ is contained in the conditional probability distribution

$$P(x|\{g_i\}) = \frac{P(\{g_i\}|x)\, P_X(x)}{P(\{g_i\})}. \tag{32}$$

In the embryo all positions are equally likely a priori, so $P_X(x) = 1/L$, and

$$P(\{g_i\}) = \int dx\, P(\{g_i\}|x)\, P_X(x). \tag{33}$$

From the distribution $P(x|\{g_i\})$ we can compute many things.

In particular it is tempting to think about constructing a single estimator $x_{\rm est}(\{g_i\})$. As above, this could be the optimal estimator in the least–square sense, the conditional mean

$$x^{(1)}_{\rm est}(\{g_i\}) = \int dx\, x P(x|\{g_i\})\,, \tag{34}$$

or it could be the maximum likelihood estimator

$$x^{(2)}_{\rm est}(\{g_i\}) = \arg\max_x P(x|\{g_i\})\,. \tag{35}$$

If the distribution $P(x|\{g_i\})$ has a single sharp peak in the neighborhood of $x^{(2)}_{\rm est}(\{g_i\})$, then these two estimators will be very close to one another, and to any other reasonable estimator, e.g. the one that minimizes the mean absolute error (the L_1 estimator). On the other hand, if there is genuine ambiguity, so that $P(x|\{g_i\})$ has more than one peak, or if the estimation is very uncertain, so that $P(x|\{g_i\})$ is extremely broad, then no single estimator really captures what a cell "knows" based on the expression levels $\{g_i\}$. At the start, it is not obvious that cells won't be in one of these ambiguous or uncertain situations, so we would like to keep all the available information. This requires us to visualize $P(x|\{g_i\})$ more directly.

Consider a cell at (actual) position x along the anterior–posterior axis. In one particular embryo α, this cell has expression levels $\{g_i^\alpha(x)\}$ at this position. If we ask what this cell knows about its possible or estimated position x^*, it is chosen from

$$P^\alpha(x^*|x) = P(x^*|\{g_i\})\Big|_{\{g_i = g_i^\alpha(x)\}}\,. \tag{36}$$

For simplicity it is useful to look at the average of these "decoding maps" across all $N_{\rm em}$ embryos in an experimental ensemble,

$$P(x^*|x) = \frac{1}{N_{\rm em}}\sum_{\alpha=1}^{N_{\rm em}} P(x^*|\{g_i\})\Big|_{\{g_i = g_i^\alpha(x)\}} \to \int \left(\prod_i dg_i\right) P(x^*|\{g_i\})\, P(\{g_i\}|x)\,. \tag{37}$$

To understand how this works, let's start not with all four gap genes but with one, as shown in Fig 12, which focuses on the information contained in g_2.

The concentration g_2 peaks roughly in the middle of the embryo, and falls to be very low in both the front quarter and the back quarter. Thus cells in these regions would be very uncertain about their position if they had access to only this one gene. In contrast, cells in the middle of the embryo experience near maximal concentrations and this "points" to a relatively narrow region along the anterior–posterior axis. This peak rises to an amplitude $P(x^*/L|x/L) \sim 25$ which means that the width of the distribution must be $\sigma_x/L \sim 1/25 \sim 4\%$. Because the mean profile $\langle g_2(x)\rangle$ is non–monotonic, rising and falling almost symmetrically around the peak, cells that are on these flanks have a two–fold ambiguity in the inference of x^* from g_2. This combination of uncertainty at the ends, precision in the middle, and ambiguity on the flanks gives rise to the X–like pattern that we see when representing $P(x^*|x)$ in gray levels in Fig 12C. If g_2 were the only signal available, embryos literally would not know their head from their tail, and no single cell could reach the precision of $\sigma_x/L \sim 1\%$ that is seen all along the anterior–posterior axis (§3.3).

Fortunately cells have access to multiple gap genes. We see in Fig 13 that as we use more of these genes to infer position, ambiguities are removed and uncertainty is reduced. Finally when we use all four genes the distribution $P(x^*|x)$ narrows essentially to a single band around the diagonal $x^* = x$. The peak height is ~ 100, which means that the width is ~ 0.01. Thus,

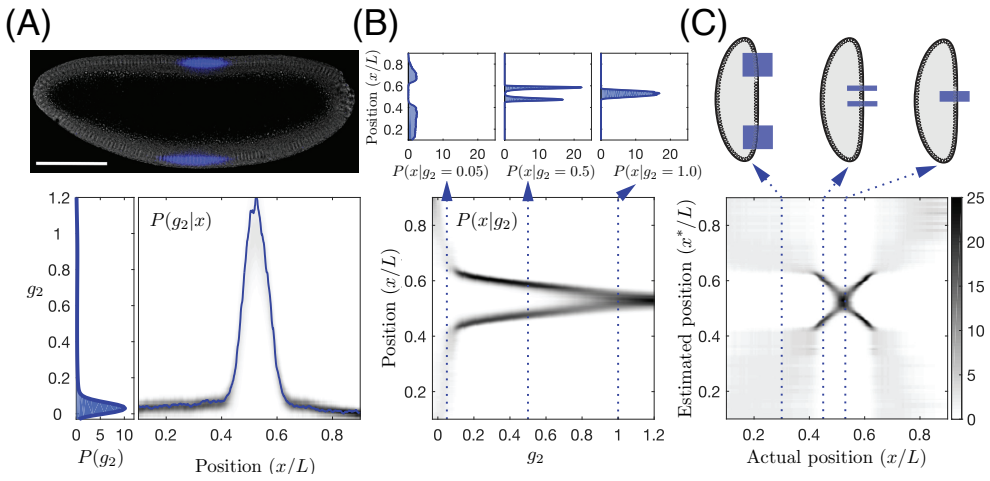

Figure 12: Decoding position from a single gene. (A) Top panel is an optical section through the midsagittal plane of a *Drosophila* embryo with immunofluorescence labeling for the protein encoded by a single gap gene, corresponding to g_2 in Fig 11 (scale bar 100 μm). Bottom panel shows the normalized $g_2(x)$ for this particular embryo (blue) and a gray–level visualization of $P(g_2|x)$; $P(g_2)$ is at left. (B) The distribution $P(x|g_2)$, shown in gray levels (bottom) and slices at fixed g_2 (top). (C) The average decoding map $P(x^*|x)$ from Eq (37). Top panel schematizes the combination of ambiguity and uncertainty. Reproduced from Ref [10], with permission.

the gap genes provide enough information to specify position with 1% accuracy, but only if cells read this information optimally.

We recall that the pair–rule stripes are positioned with $\sim$ 1% accuracy, and similarly the location of the cephalic furrow is reproducible from embryo to embryo with $\sim$ 1% precision. A possible conclusion is that embryos "read out" the information carried by local gap gene concentrations and use this to guide subsequent events. This read out is optimal, and sets the precision of the body plan. Certainly what we see is *consistent* with this conclusion, but the evidence is not unambiguous.

As an example, we could imagine that cells make dramatically sub–optimal use of the local concentration information, but compensate at the next stage through interactions among many cells. More subtly, driven by the nature of experiments we have emphasized reading the signals from gap genes at a single moment in time, while any realistic mechanism will involve some integration over time, perhaps providing another opportunity to reduce noise. There are reasons to think that these options are more limited than they seem: noise in the expression levels of the gap genes is correlated over long distances [121,145], and momentary expression

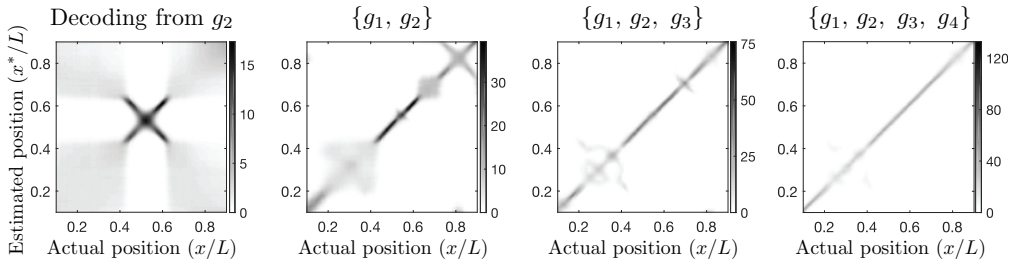

Figure 13: Decoding position from an increasing number of gap genes reduces uncertainty and eliminates ambiguity. Redrawn from Ref [10], with permission.

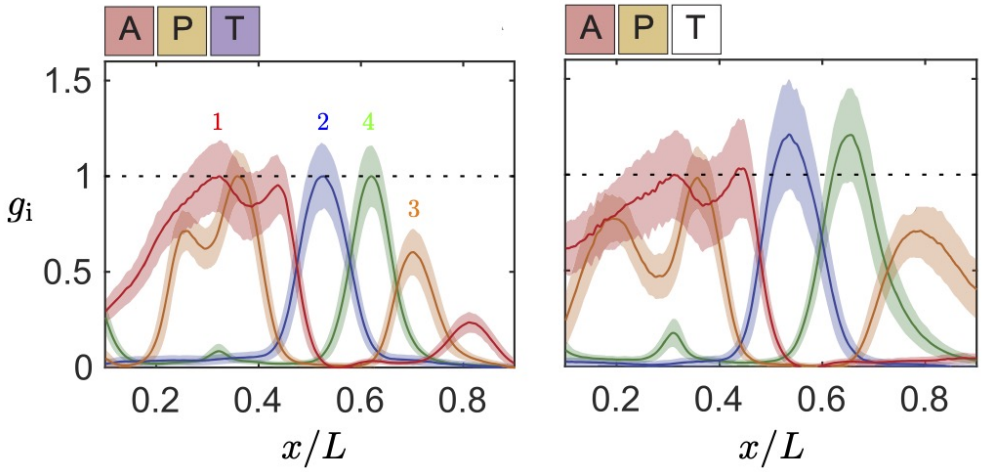

Figure 14: Gap gene concentration profiles change in mutant flies missing one of the maternal inputs. At left are the profiles in wild type flies, as in Fig 11 but with the four genes represented by different colors. At right are the results of the same measurements done in flies that are missing the terminal inputs T. Reproduced from Ref [10], with permission. It is essential that although concentrations are normalized (maximum mean concentration is one in the wild type), the normalization is the same in both panels so that concentrations can be compared meaningfully.

levels are the output of a network whose dynamics imposes temporal correlations on these noise levels; both these effects limit the possibilities for noise reduction by averaging. Still, one would like a more positive and convincing test of the idea that cells are performing the optimal readout of the positional information encoded in the gap genes.

If the embryo performs an optimal readout, then if the spatial patterns of gap gene expression are perturbed cells will get the wrong answer for their estimates of position, and this effect should be both systematic and predictable. We can perturb the gap gene patterns by knocking out one or two of the three maternal inputs to the gap gene network. We recall that one of the maternal inputs has high concentration at the anterior (A) of the embryo, one has high concentration at the posterior (P), and one has high concentration at both ends (the "terminal" inputs T). Figure 14 shows what happens in mutant flies that are missing the T inputs. As expected there is very little change to the gap gene expression levels in the middle of the embryo, with larger perturbations at both the anterior and posterior extremes, but in general it seems fair to say that these perturbations are fairly gentle.

Figure 15 shows what happens when we try to decode the patterns of gap gene expression found in the mutants of Fig 14. The idea is that we take the measured $\{g_i\}$ in the mutant and pass it through the function $P(x^*|\{g_i\})$, which was constructed from data taken in the normal (wild–type) embryos, and then we average as before to get a new map $P_{\rm mut}(x^*|x)$. By analogy with Eq (37) we can write

$$P_{\rm mut}(x^*|x) = \int \left(\prod_i dg_i\right) P\left(x^*|\{g_i\}\right) P_{\rm mut}\left(\{g_i\}|x\right) . \tag{38}$$

As expected, the map is only slightly perturbed in the central region of the embryo. Signals at small x/L are noisy and ambiguous, while at large x/L the ridge of maximum probability is systematically at $x^* < x$, along a gently curving trajectory.

We can test the predicted $P_{\rm mut}(x^*|x)$ with a simple idea. If the only information that cells have about position is their estimate x^*, then pair–rule stripes should be at locations $x^* = x_s$

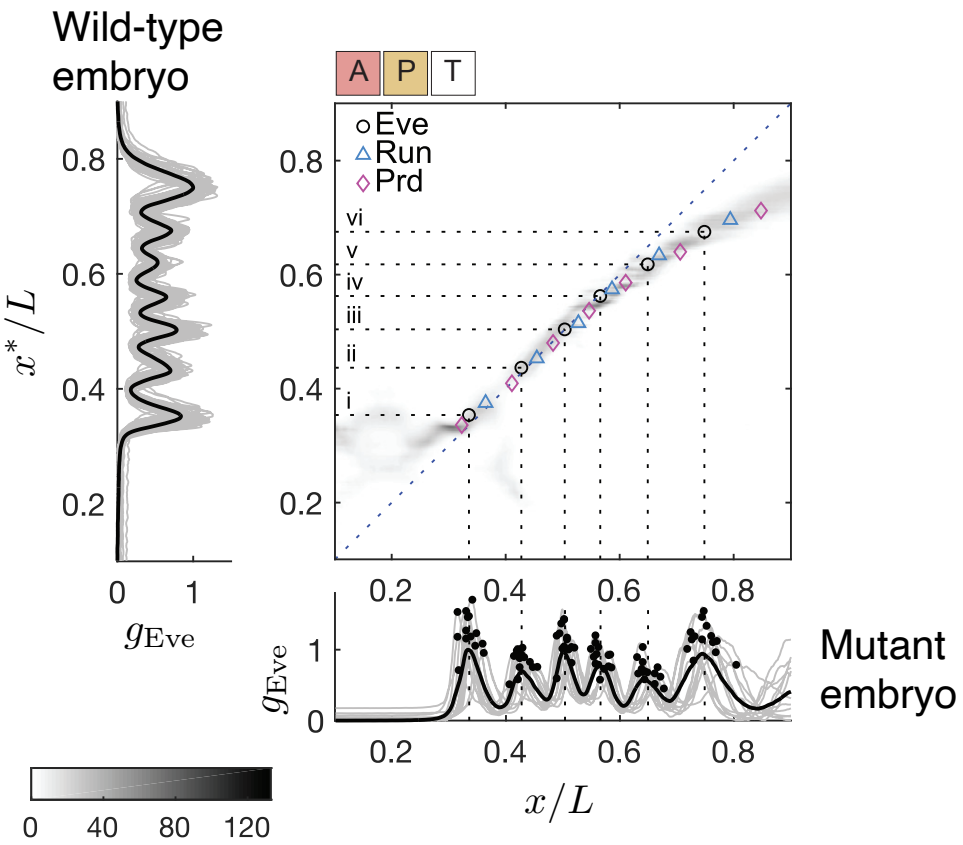

Figure 15: Optimal decoding predicts pair–rule stipe shifts in mutant embryos. We use the distribution $P(x^*|\{g_i\})$ constructed from data on wild type flies to interpret the gap gene expression signals in flies missing terminal inputs T (Fig 14). Results are shown as the average decoding map $P_{\mathrm{mut}}(x^*|x)$ from Eq (38); density in grey scale (bottom left). At left is the expression of Eve, one of the pair–rule genes; individual embryos in grey, mean in black, illustrating the seven stripes at positions x_i. When $P_{\mathrm{mut}}(x^* = x_i|x)$ is large we expect a stripe at (real) position x in the mutant, as confirmed at the bottom. Markers summarize these results, and those for two other pair–rule genes. Reproduced from Ref [10], with permission.

not $x = x_s$ as in the wild–type embryo, for each stripe s; this is the construction shown by dashed lines in Fig 15 for the gene *eve*.[23] We can test all of the stripes, for each of several pair–rule genes, and we see that the results track along the ridge of $P_{\mathrm{mut}}(x^*|x)$, with high precision.

We can redo the analysis of Fig 15 six times, deleting the three maternal inputs singly and then in pairs. A sanity check is that if we delete all three inputs there is no positional information, so any nonzero g_i is uniform in x. Detailed descriptions of the results can be found in the original paper [10], so let me draw attention here to a few points:

- In most cases when we delete maternal inputs the density in $P(x^*|x)$ fails to intersect $x^* = x_s$ for many values of the stripe index s. This predicts that certain stripes should be missing, and these (many) predictions are correct.

[23] This is one place where I'll use names. Eve is short for "even–skipped," which gives a sense for what goes wrong in the structure of the mutant embryo. The convention is that genes are named in lower case italics, while the protein products are described with capitalized Roman text. Also, beware that biochemists name molecules after what they do, while geneticists name molecules after what goes wrong when they are mutated.

- We can analyze maps $P^{\alpha}_{\rm mut}(x^*|x)$ constructed from data on individual mutant embryos (before averaging), and in some cases the density is sufficiently variable at $x^* = x_s$ that we predict stripes to be present in some but not all of the embryos. This variability never happens in wild–type embryos but it happens in the mutants, where we predict it.[24]

- In mutants where the anterior input is deleted, the most likely x^* is a non–monotonic function of x. Most of the *eve* stripes are predicted to be missing, but the seventh stripe is predicted to be duplicated, and this happens at the predicted location. Details of the underlying molecular biology show that this really is a duplicate of stripe seven and not a shifted version of a more anterior stripe.

- In mutants where both the anterior and posterior inputs are deleted, we predict that there should be only two *eve* stripes located symmetrically along the anterior–posterior axis, and this is confirmed.

In total we have 70 of these examples, and almost all predictions are confirmed within experimental error. More subtly, the predicted noise in stripe position, from the width of $P_{\rm mut}(x^*|x)$, agrees with the measured variability. The few errors are in places where the map $P_{\rm mut}(x^*|x)$ has discontinuities, so that a little bit of spatial averaging (which we neglect) would have large effects, or where expression levels in the mutant are very deep in the tails of $P(\{g_i\})$.

We have constructed the optimal decoder $P(x^*|\{g_i\})$ from measurements in a small window of time during the fourteenth nuclear cycle. This window is chosen to surround the point at which positional information is maximized, and is as narrow as possible while still leaving a reasonable number of samples. But gap gene expression profiles vary slowly throughout cycle fourteen. If the embryo implements the optimal decoding, tuned to the time of maximal positional information, then pair–rule stripes are predicted to evolve with time as well. This effect has been known for a long time, and subject to multiple interpretations. We were surprised to find that these details—shifts of $\Delta x(t)/L$ corresponding to just a few percent over half an hour—are predicted correctly as well.

Optimization principles provide a compact formulation for much of physics. As applied to living systems we typically use such principles to select the behavior of the particular systems that we see in nature out of the universe of possibilities available. In the past, this sort of argument has led to a single number, or a scaling relation between different numbers. What is new, I think, in this analysis of positional information in the fly embryo is that a single optimization principle leads to a rich set of subtle parameter free predictions, essentially all of which agree quantitatively with experiment. It is the same physical principle that leads to predictions about visual motion estimation in Fig 10.

3 Matching distributions

In the fly embryo information flows from maternal inputs to the gap gene network to the striped patterns of pair–rule gene expression. The previous lecture was about optimization at the output of the gap gene network. Can we also optimize the inputs? This is part of a more general question: given some signal processing system with fixed signal and noise characteristics, how can we choose the inputs to optimize information flow?

To be more explicit, a random cell along the length of the embryo encounters concentrations of maternal input morphogens that are drawn out of a distribution. We would like to

[24] A limitation of the experiments is that they measure all the gap genes or the pair–rule genes, but not in the same embryo. Thus we know that pair–rule stripes are variable where we predict them to be variable based on measurements of the gap genes in many mutant embryos, but we don't know if the presence or absence of stripes is connected deterministically to the gap gene profiles in single embryos, as we predict.

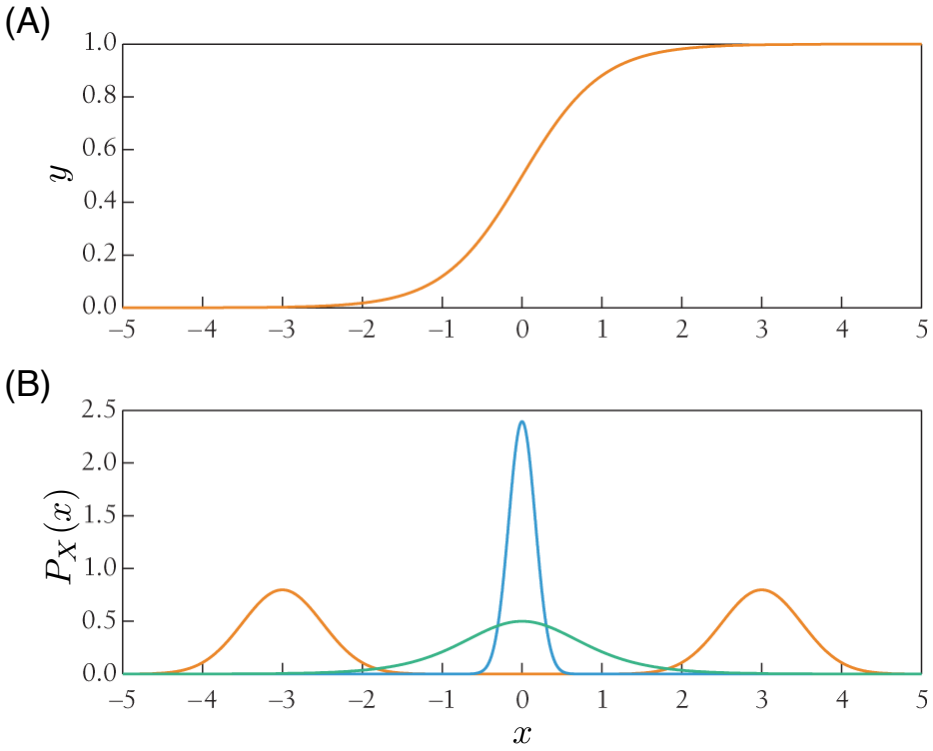

Figure 16: Matching distributions to the input/output relation [57]. (A) Example of an input/output relation. (B) Different possible probability distributions for the inputs. With either of the orange distributions, the output is either fully on or fully off and provides no information about the input signal. With the blue distribution the modulations of the output are very small, and likely to be obscured by noise (which is missing from the figure). The green distribution seems to be a better match, moving the output through its full dynamic range.

formulate an optimization principle for this distribution. Some intuition for this is given in Fig 16, where we compare the distribution of inputs to the structure of a hypothetical input/output relation. In this example, if the inputs x are concentrated at high or low values, then the output y is "stuck" in a fully on or fully off state, and not sensitive to any variations in the input. If the inputs are concentrated near the point of maximum sensitivity (here, zero input) then the responses will be larger, but if the distribution is too narrow than the variations in the output still will be small and can easily be masked by noise. It seems sensible that the best we can do is to have inputs that are centered on the point of maximum sensitivity and have a distribution wide enough to drive the outputs through their full dynamic range.

The intuitions of Fig 16 can be formalized in the language of information theory. Information theory is a beautiful subject that has deep connections to statistical mechanics, but physicists vary in their level of comfort and familiarity with the ideas. I gave the lecture in Les Houches as if people knew the basics, but there was a strong desire for a more pedagogical introduction. So we followed the regular lecture with a tutorial [146], which we plan to have as an Appendix to the Summer School proceedings.[25] For completeness let me also note that (in contrast to many other subjects) one really can learn much of what you need to know about information theory by reading Shannon's original paper [147]. The standard textbook is by Cover and Thomas [148], a version aimed at the physics community is by Mézard and

[25] My sincere thanks to Tarek Tohme, who will co-author this Appendix, having recorded the tutorial and helped turn it into coherent prose. He even captured many of the excellent questions from the students.

Montanari [149], and I have tried to give a fuller account of these ideas in the context of biophysics [57]. Each of these texts provide more than you need to make sense out of these lectures, but perhaps you will want to explore more deeply.

3.1 One input, one output

The simplest version of information transmission is a system in which one input x drives one output y. The mutual information $I(x;y)$ between these variables can be thought of either as the information that the output provides about the input or vice versa, as a measure of control power. This measure is unique, and can be written as

$$I(x;y) = \int dx \int dy\, P(x,y) \ln\left[\frac{P(x,y)}{P_X(x)P_Y(y)}\right], \tag{39}$$

where, as usual, $P(x,y)$ is the joint distribution while $P_X(x)$ and $P_Y(y)$ are the two marginals.

Mutual information is a measure of correlation between x and y. If the variables were independent, then the entropy of the joint distribution would be equal to the sum of the entropies of the two marginal distributions, but in fact it is smaller. The mutual information is exactly this decrease in entropy:

$$I(x;y) = S[P_X(x)] + S[P_Y(y)] - S[P(x,y)], \tag{40}$$

where $S[P]$ is the entropy of the distribution P,

$$S[P(z)] = -\int dz\, P(z) \log P(z). \tag{41}$$

We also can think of the mutual information as a functional of the two distributions,

$$I(x;y) = I[P_X(x), P(y|x)] = \int dx\, P_X(x) \int dy\, P(y|x) \log\left[\frac{P(y|x)}{P_Y(y)}\right], \tag{42}$$

$$P_Y(y) = \int dx\, P(y|x) P_X(x). \tag{43}$$

The convexity of the entropy implies that $I(x;y)$ has a maximum with respect to variations in $P_X(x)$ and a minimum with respect to variations in $P(y|x)$. Thus if we fix the way in which y responds to x, as encoded in $P(y|x)$, we can maximize information transmission by adjusting the distribution of inputs.

To be concrete, let's assume that

$$P(y|x) = \frac{1}{\sqrt{2\pi\sigma_y^2(x)}} \exp\left[-\frac{(y-\bar{y}(x))^2}{2\sigma_y^2(x)}\right]. \tag{44}$$

Notice that

$$\begin{aligned} I(x;y) &= \int dx\, P_X(x) \int dy\, P(y|x) \log\left[\frac{P(y|x)}{P_Y(y)}\right] \\ &= -\int dy\, P_Y(y) \log P_Y(y) - \int dx\, P_X(x) \left[\int dy\, P(y|x) \log P(y|x)\right] \qquad (45) \\ &= S[P_Y(y)] - \langle S[P(y|x)] \rangle^{(x)}, \qquad (46) \end{aligned}$$

where $\langle\cdots\rangle^{(x)}$ denotes an average over the distribution $P_X(x)$. We can compute the conditional entropy from Eq (44), now using natural logs:

$$S[P(y|x)] \equiv -\int dy\, P(y|x)\ln P(y|x) \tag{47}$$

$$= -\int dy\, \frac{1}{\sqrt{2\pi\sigma_y^2(x)}} \exp\left[-\frac{(y-\bar{y}(x))^2}{2\sigma_y^2(x)}\right]\left[-\frac{1}{2}\ln\left[2\pi\sigma_y^2(x)\right] - \frac{(y-\bar{y}(x))^2}{2\sigma_y^2(x)}\right] \tag{48}$$

$$= \frac{1}{2}\ln\left[2\pi\sigma_y^2(x)\right] + \frac{1}{2} \tag{49}$$

$$= \frac{1}{2}\ln\left[2\pi e\sigma_y^2(x)\right]. \tag{50}$$

This result for the entropy of Gaussian distributions is very useful.

Substituting into Eq(46), we have

$$I(x;y) = S[P_Y(y)] - \langle S[P(y|x)]\rangle^{(x)}$$

$$= -\int dy\, P_Y(y)\ln P_Y(y) - \int dx\, P_X(x)\frac{1}{2}\ln\left[2\pi e\sigma_y^2(x)\right]. \tag{51}$$

Now if the function $\bar{y}(x)$ is monotonic, and the noise is small, we can approximate

$$P_Y(y)dy \approx P_X(x)dx\,, \tag{52}$$

so that

$$I(x;y) \approx -\int dx\, P_X(x)\ln\left[P_X(x)\left|\frac{dy}{dx}\right|^{-1}\right] - \int dx\, P_X(x)\frac{1}{2}\ln\left[2\pi e\sigma_y^2(x)\right] \tag{53}$$

$$= -\int dx\, P_X(x)\ln\left[P_X(x)\sqrt{2\pi e}\sigma_x^{\rm eff}(x)\right], \tag{54}$$

where we identify

$$\frac{1}{\sigma_x^{\rm eff}(x)} = \frac{1}{\sigma_y(x)}\frac{d\bar{y}(x)}{dx}. \tag{55}$$

We can understand this as the propagation of the noise σ_y back into an estimate of x with effective noise level $\sigma_x^{\rm eff}(x)$, illustrated in Fig 17. This approximation is self–consistent if $\sigma_x^{\rm eff}(x)$ is small on the scale over which $\bar{y}(x)$ and $P_X(x)$ vary.

Starting from Eq (54), we can vary the input distribution $P_X(x)$ to maximize the information $I(x;y)$, introducing a Lagrange multiplier to enforce normalization:

$$0 = \frac{\delta}{\delta P_X(x)}\left[I(x;y) - \lambda\left(\int dx\, P_X(x) - 1\right)\right] \tag{56}$$

$$= \frac{\delta}{\delta P_X(x)}\left[-\int dx\, P_X(x)\ln\left[P_X(x)\sqrt{2\pi e}\sigma_x^{\rm eff}(x)\right] - \lambda\left(\int dx\, P_X(x) - 1\right)\right] \tag{57}$$

$$= -\ln\left[P_X(x)\sqrt{2\pi e}\sigma_x^{\rm eff}(x)\right] - 1 - \lambda \tag{58}$$

$$\Rightarrow P_X(x) = \frac{1}{Z}\frac{1}{\sigma_x^{\rm eff}(x)}, \tag{59}$$

where we collect all the normalization constants into

$$Z = \int \frac{dx}{\sigma_x^{\rm eff}(x)}. \tag{60}$$

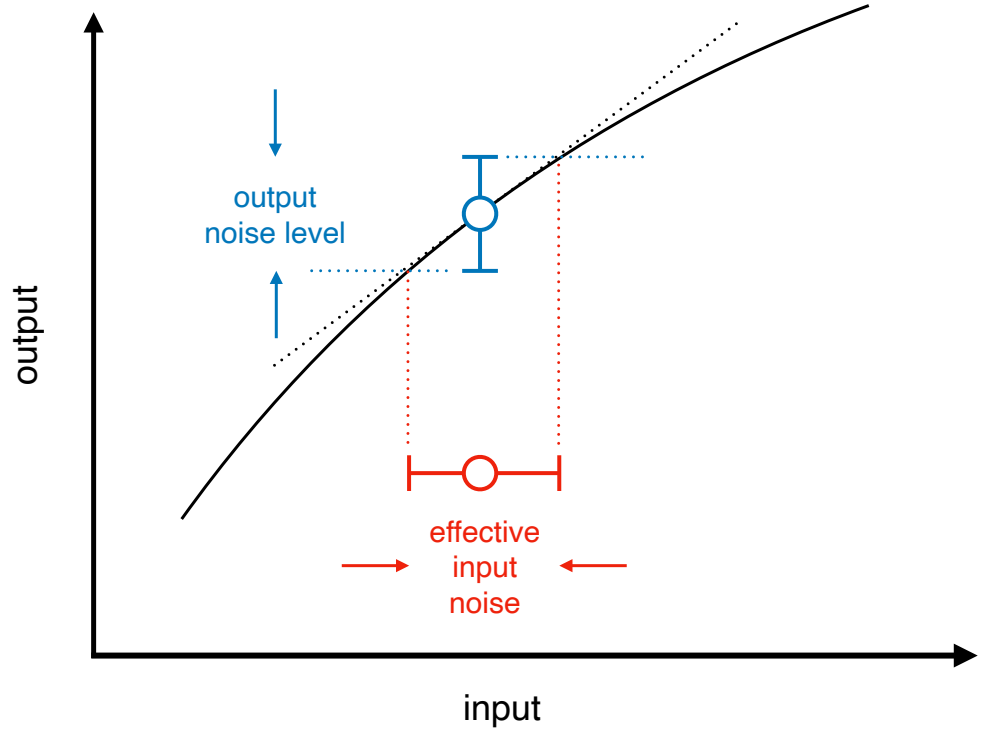

Figure 17: Error propagation, from Eq (55). The noise level at the output—which often can be estimated directly from experiment—is translated to an effective noise level at the input, using the local slope of the input/output relation.

Notice that with this result $\ln\left[P_X(x)\sqrt{2\pi e}\sigma_x^{\text{eff}}(x)\right] = \ln[Z/\sqrt{2\pi e}]$, so that

$$I_{\text{max}}(x;y) = \ln\left(\frac{Z}{\sqrt{2\pi e}}\right). \tag{61}$$

The crucial result here is Eq (59): to transmit the maximum information (at low noise levels) we should use inputs in inverse proportion to their effective noise levels. In this sense the optimal input distribution matches the input/output relation and the associated noise levels.

The result that inputs should be used in inverse proportion to their noise level is a precise version of familiar ideas. When writing we avoid words that we don't know how to spell, and when speaking a foreign language we avoid constructions which we suspect we might get wrong. Different species of frogs call to one another in different frequency bands, and these match the bands where they hear best. In the low noise limit we can think of the matching condition as applying equally well to the input or the output,

$$P_X(x) \propto \frac{1}{\sigma_x^{\text{eff}}(x)} \quad \Rightarrow \quad P_Y(y) \propto \left.\frac{1}{\sigma_y(x)}\right|_{x=\tilde{x}(y)}, \tag{62}$$

where $\tilde{x}(y)$ is the inverse of the function $\bar{y}(x)$.

3.2 Neural input/output relations

An important special case of these arguments is where the noise at the output is constant, $\sigma_y(x) = \sigma_y$. Then Eq (62) tells us that optimal information transmission corresponds to $P_Y(y)$ also being constant. Outputs always have limited dynamic range, so this means that the system transmits the maximum information by using this dynamic range uniformly, maximizing the

entropy of the outputs.[26] A familiar example is where the outputs are quantized, as with a digital image. Then the dominant source of (effective) noise often is the discretization itself, and thus is constant—we distinguish 1 vs 2 as reliably as we distinguish 254 vs 255. Then optimal information transmission occurs when all the output values are used equally often, and this is called "histogram equalization" or adaptive binning.

We can go one step further and choose units such that $0 < y < 1$, so the optimized uniform distribution of outputs is $P_Y(y) = 1$. In the low noise limit we have Eq (52), so that

$$P_Y(y)dy = P_X(x)dx$$
$$dy = P_X(x)dx \tag{63}$$
$$\frac{dy}{dx} = P_X(x)\,. \tag{64}$$

Again because this is the low noise limit the y which appears here can be taken as $\bar{y}(x)$, and so we have

$$\frac{d\bar{y}(x)}{dx} = P_X(x)\,. \tag{65}$$

Integrating, the optimal input/output relation becomes the cumulative distribution of inputs.

Two interesting things just happened. First, we started by optimizing the distribution of inputs and ended up by expressing the result as an optimal input/output relation. Second, we have a prediction that the input/output relation should match the distribution of inputs, quantitatively and with no free parameters. If the distribution $P_X(x)$ has a single peak, $\bar{y}(x)$ should be roughly sigmoidal, as one finds for the input/output behaviors of many biological systems.

In an inspiring paper, now 40+ years ago, Laughlin took these theoretical ideas seriously and applied them to the responses of neurons in the fly retina, the "large monopolar" cells (LMCs) that take inputs directly from the photoreceptors [150].[27] These cells produce a graded voltage in response to changes in light intensity around a background. Laughlin built a photodetector to match the optics of a single receptor cell in the fly's eye and measured the distribution of light intensities found by scanning natural scenes. He then measured the input/output relations of the LMCs, with the comparison shown in Fig 18. Note that there are no free parameters.

To be fair, the input/output relation is not a very complicated function, so saying that we predict its form with no free parameters may be an overstatement. What we predict is that, in the normalized units of Fig 18, the maximum slope should be at a location near $x = 0$ and the width of the response should span $x \in [-0.5, 0.5]$; these predictions come from the shape of the distribution of light intensities.

The light intensity that we see is the product of a source intensity and the reflectivity of the surface we are looking at. As the overall brightness changes, e.g. from dawn through noon to dusk, reflectivities are constant. This suggests that the distribution of log intensity should keep roughly the same shape, just shifting along the intensity axis. The matching condition Eq (65) then predicts that the neural responses to log intensity should keep the same shape, just shifting their midpoints, and this is a good zeroth order theory of what happens during light and dark adaptation.

There are many over–simplifications here, but the idea is powerful. When we ask why the input/output relation of a neuron looks the way it does, the standard answer is to start explaining all the molecular, cellular, and synaptic mechanisms that lead to the final phenomenology.

[26] Maximizing information transmission generally is *not* the same as maximizing the entropy of the outputs. The difference arises both because the noise can have structure $[\sigma_y(x) \neq \sigma_y]$ and because we can depart from the low noise limit where Eqs (59, 62) were derived. This will be important below.

[27] These cells are in the same position in the fly's retina as the bipolar cells in the vertebrate retina (§2.2).

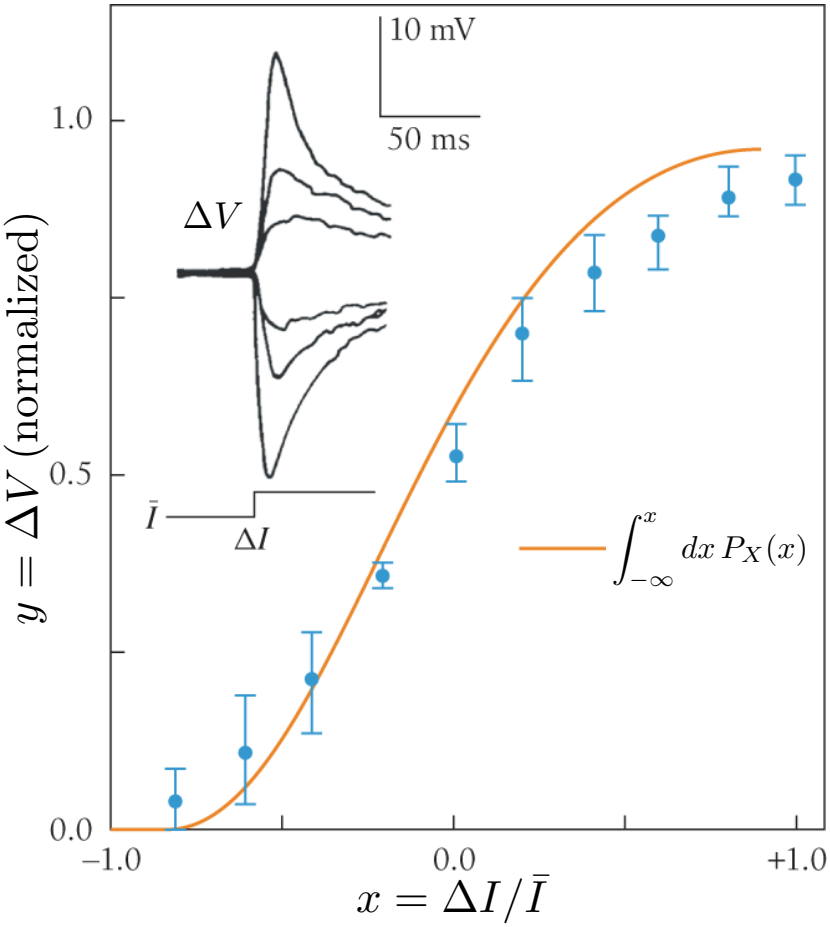

Figure 18: Input/output relations match the distribution of inputs. Brief changes in light intensity ΔI above or below background $\bar{I}$ produce transient voltage changes in the large monopolar cells (inset), and the peaks of these responses are taken as the cell's output ΔV . Normalized responses are compared to the cumulative probability distribution of light intensities, testing the predictions of Eq (65). Data redrawn from Ref [150], under the Creative Commons Attribution NC-ND 3.0 License.

Such explanations, as emphasized in the introductory lecture, lead us into a forest of parameters. Worse yet, these parameters are adjustable, for example if the cell in question had expressed different ion channels, different neurotransmitter receptors, etc. The idea here is different: input/output relations have the form that they do because they are matched to their inputs, optimizing information transmission. The fact that the predictions from such a potentially general, parameter free theory are even approximately right is very encouraging.

If we think that input/output relations are matched to the distribution of inputs, it is natural to ask on what time scale this matching occurs. More critically, one might worry that "the" distribution is such a dynamic object that it is not well defined. But maybe this is a good thing, since we know that input/output relations in neurons are themselves dynamic objects.

The fact that input/output relations of neurons change in response to background conditions is called adaptation.[28] In the original descriptions of adaptation in sensory neurons, the focus was on the fading response to constant stimuli, in effect subtracting a constant from the output, but this is far from the whole story. The time scales of response often change with the

[28] Surely "adaptation" is one of the most over used words in the description of biological systems.

background, e.g. as visual responses become slower in the dark. More subtly there are changes in the gain of the response, e.g. as one extra photon produces a smaller response when added to a brighter background of light. All of these effects, however, can be seen as driven by the mean background signal. Maximizing information transmission predicts that neurons should match their input/output relations to the whole *distribution* of input signals.

Maybe the simplest question is whether neurons adapt to the variance of inputs as well as to the mean. As an example, if we look at the activity of the neurons that carry the output of the retina to the brain (retinal ganglion cells), then stimulate the retina with time varying light intensity, there is a clear and rather immediate response to changing the mean intensity and that gradually relaxes as the retina adapts. The same thing happens if we suddenly change the variance of the light intensity, or its spatial correlations [151], which certainly is consistent with the idea that input/output relations are changing to match the input distribution.

We would like to map the input/output relations in the steady states that are reached after the system has adapted to different distributions of inputs. How to do this is a subject in itself. Briefly, we are interested here in neurons that respond to inputs with a sequence of discrete, identical electrical pulses called action potentials or spikes. We will assume for simplicity that the input is a single function of time $s(t)$. One characterization of the input/output relation is then to give the probability per unit time that the neuron will generate a spike near time t given the history of inputs up to this moment, $s(t-\tau)$ for $\tau > 0$. If the inputs were weak one could try a linear approximation,

$$P_{\rm spike}(t) = \bar{r}\left[1 + \int_0^\infty d\tau\, g(\tau)s(t-\tau) + \cdots\right], \tag{66}$$

where $\bar{r}$ is the mean probability per unit time or rate of generating spikes, and $g(\tau)$ is the linear response function. This is too restrictive, and easily runs up against the constraint that $P_{\rm spike} \geq 0$. A natural extension is to say that only the filtered inputs are important, but these could be processed nonlinearly,

$$P_{\rm spike}(t) = \bar{r}F\left[\int_0^\infty d\tau\, g(\tau)s(t-\tau)\right], \tag{67}$$

where F is arbitrary but a natural choice might be something sigmoidal as in Fig 18.

The "linear–nonlinear" model in Eq (67) is quite popular [79, 152]. It is interesting in part because it is tractable. If the input signals $s(t)$ are Gaussian with zero mean,[29] then we can separate the filter from the nonlinearity by computing a correlation function between the spike sequence and the input, or equivalently an average of the input triggered on spike times:

$$\langle s(t_{\rm spike} - t)\rangle \propto \int d\tau g(\tau)\langle s(t)s(\tau)\rangle\,. \tag{68}$$

In the simplest case where the inputs are both Gaussian and white, so that $\langle s(t)s(\tau)\rangle \sim \delta(t-\tau)$, we have just $\langle s(t_{\rm spike} - t)\rangle \propto g(t)$. This strategy for describing neural responses came from work on the neurons that first encode sound in the inner ear, as people tried to separate the mechanical filtering of sound from the nonlinearity of spiking [153].

Another way of expressing Eq (67) is to say that the input is a high dimensional vector—the function $s(t < t_{\rm spike})$, perhaps sampled at discrete times—and that the neural response depends on only one projection of this vector. Then there is a natural generalization

$$P_{\rm spike}(t) = \bar{r}F\left[s_1(t), s_2(t), \cdots, s_K(t)\right], \tag{69}$$

$$s_\mu(t) = \int_0^\infty d\tau\, g_\mu(\tau)s(t-\tau)\,. \tag{70}$$

[29] I think in the computer science and applied mathematics literature one would say "Gaussian stochastic process." Physicists use "Gaussian" more vaguely to describe anything from a single random variable to a free field theory.

This description emerged from analysis of the distribution of inputs conditional on the occurrence of a spike [154] and would be used as a model of responses only later [155–157]. Again the key is that triggered averages are connected to the underlying structure. If we compute the covariance of inputs in the neighborhood of a spike, and compare with the total (not triggered) covariance, then Eq (69) implies that

$$\Delta C(t,t') = \left[\langle s(t_{\rm spike}-t)s(t_{\rm spike}-t')\rangle - \langle s(t_{\rm spike}-t)\rangle\langle s(t_{\rm spike}-t')\rangle\right] - \langle s(t)s(t')\rangle\,, \qquad (71)$$

is an operator of rank K. Further, the eigenfunctions of this operator span the same space as the filters $\{g_\mu(\tau)\}$, blurred by the input correlation function. If the rank K is reasonably small this gives us a path to identify the relevant projections of the input and map the input/output relation $F[\{s_\mu\}]$. We can then change the distribution of inputs $P[s]$ and ask if this relation changes in ways that we expect for the optimization of information transmission.

To be fair, we don't really know how to solve the relevant optimization problem for spiking neurons. Obviously the assumption of constant noise at the output, as above, doesn't really make sense when the output is the probability of generating a spike. But if the noise levels are small, then in any reasonable scenario the scales of the optimal input/output relation $F[\{s_\mu\}]$ will be set by the scale of the probability distribution $P[s(t)]$ itself. Concretely, if we rescale the input signals we expect a compensatory rescaling of the response function,

$$s \to \lambda s \quad \Rightarrow \quad F[\{s_\mu\}] \to F[\{s_\mu/\lambda\}]\,. \qquad (72)$$

Maybe a clearer way to say this is that the input/output relations should be different in response to different distributions of input, but these should collapse if we plot the response not vs the projected signals s_μ but rather vs the normalized projections $s_\mu/s_\mu^{\rm rms}$.

These predictions were tested in experiments on motion–sensitive neurons in the fly visual system [155]. The fly watches a movie of a random spatial pattern that moves with velocity $s(t)$ chosen from a Gaussian distribution with a short (2 ms) correlation time, and the standard deviation of this distribution could be changed by rescaling as above. Mean rates of spiking were $\bar r \sim 70$ spikes/s, and the spike–triggered covariance $\Delta C(t,t)$ had only two significantly nonzero eigenvalues. The filter $f_1(\tau)$ smooths the velocity over a ~ 50 ms window, while $f_2(\tau)$ is almost exactly the derivative of $f_1(\tau)$. It then is natural to express s_1 in the same physical units as the input angular velocity (°/s), and s_2 as an angular acceleration (°/s^2). Figure 19 shows the projections of $F(s_1,s_2)$ onto the two axes as measured under four different conditions where the standard deviation of the input velocity varies by an order of magnitude.

Figures 19A and C show the input/output relation in physical units. We see that when the same velocity is chosen from a different distribution, the response of the neuron can differ by orders of magnitude. But if we rescale the inputs by their standard deviations, these enormous variations collapse onto a single function, as in Figs 19B and D. As an aside, we note that this approach maps the input/output relations over a dynamic range of $\sim 10^3$, and the rescaling is essentially perfect across this full range. Although harder to visualize, we can see the same rescaling in the (s_1,s_2) plane. This is exactly what we expect from Eq (72).

One might worry that the changing input/output relations are an artifact of sampling a fixed function of many variables in different parts of the space as we change distributions. We are reassured by the fact that we see the same behavior along two dimensions, separately and together, and that ΔC is very accurately of low rank. We can also check that the information that single spikes convey about the pair (s_1,s_2) is equal to the information that is conveyed about the stimulus as a whole, within experimental error. The same effects can be recapitulated in response to inputs that have orders of magnitude slower time scales.

These results show that the system can implement not just a single input/output relation $F(s_1,s_2)$ but rather a whole family of relations $F(bs_1,bs_2)$, where the scale factor is inversely

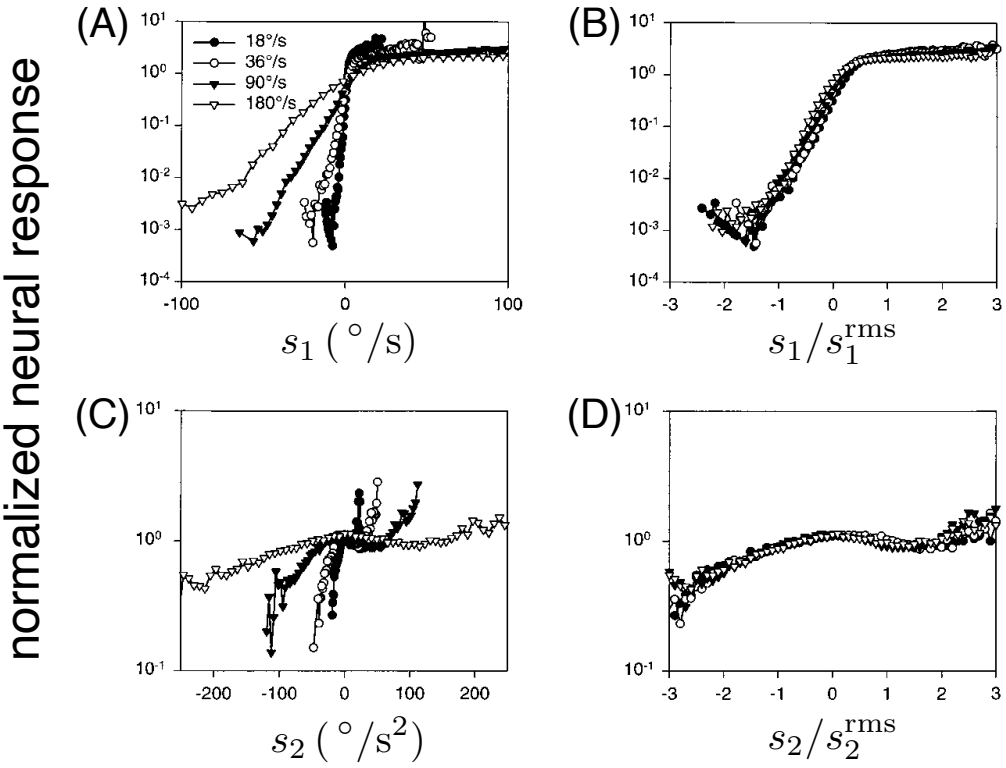

Figure 19: Adaptive rescaling in the responses of a motion–sensitive neuron in the fly visual system. Normalized neural response $F(s_1, s_2)$ from Eq (69) projected onto stimulus features s_1 in (A, B) and onto s_2 in (C, D). (A, C) Input/output functions are different when inputs are chosen from different distributions and plotted in physical units; inset in (A) indicates the standard deviation of the input velocities in the four distributions. (B, D) Input/output relations collapse in normalized units, confirming the prediction of Eq (72). Reproduced from Ref [155], with permission, and thanks to N Brenner and RR de Ruyter van Steveninck.

proportional to the standard deviation of the stimulus, $b = \lambda/s_{\mathrm{rms}}$. The real system is characterized by a particular value of λ, which depends on exactly how we normalize the filters $f_\mu(\tau)$; as we did things the real $\lambda = 1$. But we could imagine systems that achieve adaptive rescaling, stretching their input/output relation as the stimulus variance is increased, but the value of λ could be different. If we make a model of this behavior, we find that information transmission is maximized at the observed $\lambda = 1$: not only does the system rescale as predicted by our optimization principle, but the precise rescaling factor is optimal [155]. If we make a sudden shift in the input variance, then we can "catch" the neuron in an intermediate state, before adaptation, and in this state each spike really does transmit less information about the input [158]. The recovery of the information is very fast, perhaps as fast as the system can reliably infer that the input distribution has changed, although longer time scale dynamics help to encode the input variance, resolving ambiguities.

The same adaptive rescaling phenomena have been seen in the bird "field L," which is an analog of our auditory cortex [159]; in the rat "barrel cortex," which responds to whisker movements [160]; and many other systems. Related effects are seen in midbrain regions of mammalian auditory processing, and there is direct evidence that these adaptations improve information transmission by the population of neurons as a whole [161]. The dynamics of

these adaptation processes, as in the example of the fly [158], span many time scales, which is a separately fascinating subject [162]. In the retina one can make progress toward identifying the molecular mechanisms responsible for adaptation to the distribution [163, 164].

3.3 Positional error in the embryo

The gap gene expression levels $\{g_i\}$ encode information about position x along the anterior–posterior axis. In §2.3 we have discussed how to decode this information, but we haven't talked about how much information is present. In general this is hard to estimate, but a useful observation is that noise levels are small. Let's explore [120, 165].

Formally we have Eqs (29–31),

$$P(\{g_i\}|x) = \frac{1}{Z(x)} \exp\left[-\frac{1}{2}\chi^2(\{g_i\};x)\right], \tag{73}$$

$$\chi^2(\{g_i\};x) = \sum_{ij} \delta g_i(x) \left[\hat{C}^{-1}(x)\right]_{ij} \delta g_j(x), \tag{74}$$

$$Z(x) = \sqrt{(2\pi)^4 \det \hat{C}(x)}, \tag{75}$$

where as usual $\delta g(x)$ is the fluctuation around the mean value at x,

$$\delta g_i(x) = g_i(x) - \langle g_i(x)\rangle, \tag{76}$$

and $\hat{C}(x)$ is the covariance matrix of these fluctuations

$$\left[\hat{C}(x)\right]_{ij} = C_{ij}(x) = \langle \delta g_i(x) \delta g_j(x)\rangle . \tag{77}$$

Suppose that the concentrations $\{g_i\}$ are those found in a cell at position $x = x_{\rm true}$. Then it is convenient to expand $\chi^2(\{g_i\};x)$, and we'll assume that the covariance varies more slowly than the mean. So we start with

$$\langle g_i(x)\rangle = \langle g_i(x_{\rm true})\rangle + \left[\frac{d\langle g_i(x)\rangle}{dx}\right]_{x=x_{\rm true}} (x - x_{\rm true}) + \cdots , \tag{78}$$

which then implies

$$\chi^2(\{g_i\};x) = \chi^2(\{g_i\};x_{\rm true}) - 2A(x - x_{\rm true}) + B(x - x_{\rm true})^2 + \cdots , \tag{79}$$

$$A = \left[\sum_{ij} \frac{d\langle g_i(x)\rangle}{dx} \left[\hat{C}^{-1}(x)\right]_{ij} (g_j - \langle g_j(x)\rangle)\right]_{x=x_{\rm true}}, \tag{80}$$

$$B = \left[\sum_{ij} \frac{d\langle g_i(x)\rangle}{dx} \left[\hat{C}^{-1}(x)\right]_{ij} \frac{d\langle g_j(x)\rangle}{dx}\right]_{x=x_{\rm true}}. \tag{81}$$

If the distribution of positions $P(x)$ is smooth, then the conditional distribution $P(x|\{g_i\})$ becomes a Gaussian with a mean that is slightly shifted from $x_{\rm true}$ by terms related to the noise δg and a variance $\sigma_x^2 = 1/B$. All of these approximations are self–consistent if the values of σ_x that we compute in this way comes out to be small enough.

It is useful to think of these results as a generalization of error propagation, as in the discussion surrounding Fig 17. If the position x is encoded by a single variable g that has variance σ_g^2, then the effective variance in x is determined by

$$\frac{1}{\sigma_x^2} = \frac{1}{\sigma_g^2}\left[\frac{d\langle g\rangle}{dx}\right]^2 . \tag{82}$$

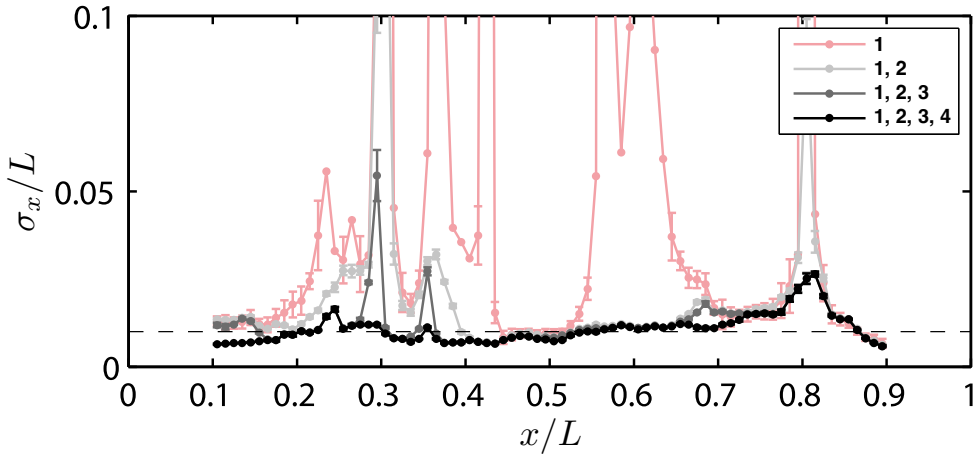

Figure 20: Effective positional noise levels from gap gene expression levels. Calculations of σ_x are done in the small noise limit, from Eq (84), using the data illustrated in Fig 11. As indicated in the inset, different curves correspond to including just g_1, the combinations $\{g_1, g_2\}$ and $\{g_1, g_2, g_3\}$ and finally all four gap genes $\{g_1, g_2, g_3, g_4\}$. With all the genes we find that the positional error is small along the entire anterior–posterior axis, and close to $\sigma_x/L = 0.01$ (dashed line). Reproduced from Ref [120], with thanks to JO Dubuis, G Tkačik, EF Wieschaus, and T Gregor.

Measurements are like springs that hold our estimate of the underlying variable close to its true value. As with the thermal fluctuations of a particle hanging from a spring, the variance is inversely proportional to the spring constant. If we make multiple independent measurements the spring constants should add, reducing the total variance,

$$\frac{1}{\sigma_x^2} = \sum_a \frac{1}{\sigma_{g_a}^2}\left[\frac{d\langle g_a\rangle}{dx}\right]^2 . \tag{83}$$

In fact we have

$$\frac{1}{\sigma_x^2} = \sum_{ij} \frac{d\langle g_i(x)\rangle}{dx}\left[\hat{C}^{-1}(x)\right]_{ij}\frac{d\langle g_j(x)\rangle}{dx}, \tag{84}$$

but this reduces to Eq (83) if we rotate into the eigenbasis of the covariance matrix $\hat{C}(x)$. Notice that in this small noise approximation, the effective positional noise σ_x depends on the true position but not on the particular concentrations $\{g_i\}$ that we happen to observe. Also, in contrast to the discussion of decoding in §2.3, this analysis is local, and does not address ambiguities; again, it makes sense in a small noise limit, where ambiguities are resolved.

Everything that we need to compute the effective positional error is contained in the data of Fig 11, where we see the mean expression levels $\langle g_i(x)\rangle$ and covariance matrices $\hat{C}(x)$ as functions of position. Although Eq (84) includes multiple genes, we can start with just one, corresponding to g_1 in Fig 11. We see that with this one gene the positional noise drops to $\sigma_x/L \sim 0.01$ in a window near the middle of the embryo, but is much larger outside this window. As we add more of the gap genes, the very largest values of σ_x are reduced dramatically, and we end up with a nearly uniform $\sigma_x/L \sim 0.01$ across the full length of the embryo. Among other things, this is small enough that our approximations above are self–consistent. One should also see this as quantifying the progressive improvement of the decoding maps as we add more genes in Fig 13.[30]

[30] Although the analysis of positional errors came before the full decoding maps, by roughly five years [10, 120].

There are at least two distinct points of interest in Fig 20. First, the scale of the positional errors, $\sigma_x/L \sim 0.01$. This is the same as the scale of positional errors in pair–rule stripe positions, and also in the placement of the "cephalic furrow" [166], which is the first macroscopic structural change that happens after the embryo is separated into discrete cells. The conventional perspective on this system has been that information flow from maternal inputs to gap genes to pair–rule genes entails a gradual refinement, with noisy inputs ultimately generating a precise and reproducible pattern [167, 168]. We see here that the precision visible at the pair–rule genes already is present in the gap genes, so information must be transformed and preserved rather than refined.

Second, the positional error is nearly uniform along the anterior–posterior axis. This is surprising because individual genes have complicated patterns of expression and noise, and provide precise information only in limited regions. Indeed, as we saw in §2.3, information carried by single gap genes is not only inhomogeneous but also ambiguous. Somehow the signal and noise (and even the covariances) of all four gap genes conspire to generate nearly constant precision.

To understand why the constancy of precision is so interesting, let's finish our project of estimating the information, in bits carried by $\{g_i\}$. We can write

$$I(\{g_i\}; x) = S[P_X(x)] - \langle S[P(x|\{g_i\})]\rangle \,, \tag{85}$$

where as before $S[P]$ is the entropy of the distribution P. But we have seen that $P(x|\{g_i\})$ is nearly Gaussian, which means immediately that

$$S[P(x|\{g_i\})] = \frac{1}{2}\ln\left[2\pi e\sigma_x^2\right] . \tag{86}$$

Further, we have seen that σ_x doesn't depend on the precise expression levels, but only on the position of the cell in which these expression levels are observed. Thus the average in Eq (85) can be written as an average over positions, so that

$$I(\{g_i\}; x) = -\int dx\, P_X(x)\ln P_X(x) - \frac{1}{2}\int dx\, P_X(x)\ln\left[2\pi e\sigma_x^2(x)\right] \tag{87}$$

$$= -\int dx\, P_X(x)\ln\left[P_X(x)\sqrt{2\pi e}\sigma_x(x)\right] , \tag{88}$$

which brings us back exactly to Eq (54). Although there are many dimensions to the output of this system, the fact that there is only one dimension at the input means that, at least in the low noise limit, we can get back to a picture very much like the one input, one output system that we started with in §3.1.

In particular, if we imagine that the embryo could adjust the distribution of cell positions, then the optimal information transmission would occur when

$$P_X(x) \propto \frac{1}{\sigma_x(x)} . \tag{89}$$

Since the real $P_X(x)$ is essentially uniform, optimal information transmission predicts that $\sigma_x(x)$ should be uniform, and this is what we see. Again, it requires considerable conspiracy among the four gap genes to achieve uniform positional error, so this seems like a significant success for the theory of optimal information transmission. There are two problems.

The first problem is that $\sigma_x(x)$ is not exactly uniform, but theory actually gives us a way to measure the significance of these deviations. With uniform $P_X(x) = 1/L$ the actual information transmitted is, substituting into Eq (88),

$$I = \frac{1}{L}\int_0^L dx\, \ln\left[\frac{L}{\sqrt{2\pi e}\sigma_x(x)}\right] . \tag{90}$$

On the other hand, if we are free to optimize $P_X(x)$ over the range $0 < x < L$ then from Eq (61) the maximum information possible given the measured $\sigma_x(x)$ becomes

$$I_{\rm max} = \ln\left[\frac{1}{\sqrt{2\pi e}}\int_0^L \frac{dx}{\sigma_x(x)}\right]. \tag{91}$$

If we plug in the results from our analysis of the experimental data, we find[31]

$$\frac{I}{I_{\rm max}} = 0.984 \pm 0.003\,. \tag{92}$$

Thus optimization of information transmission predicts a match between the positional noise levels and the distribution of cell positions, and this match is sufficiently good that it brings the embryo within 2% of the optimum [120].

The second problem is that it seems a bit weird to talk about redistributing cells along the embryo's axis. We don't really need to do this. The gap gene network output can be thought of not as providing information about position but rather about the concentrations of the maternal input molecules. The mapping between position and input concentration is something that could be different with different parameters of the relevant dynamics, different anchoring of the mRNA molecules, etc.. Thus we *can* imagine a family of embryos with different possible distributions of inputs to the gap gene network. But if the mapping between position and input concentration is deterministic and invertible, then information about concentration is the same as information about position, and all of the arguments here about matching go through.

To get a rough estimate of the total positional information conveyed by the gap genes we can use Eq (90) but with $\ln \to \log_2$ so the units are bits, and the approximate $\sigma_x/L \sim 0.01$, which gives $I \sim \log_2(100/\sqrt{2\pi e}) \sim 4.6$ bits. We can't see all of this because we are measuring only in the central 80% of the anterior–posterior axis ($0.1 < x/L < 0.9$); outside this region imaging becomes prone to systematic errors from the curvature of the embryo. But if we correct for this then the integral gives a result very close to the rough answer, and also very close to a more brute force integration over the four dimensional space $\{g_i\}$, which does not require any small noise or Gaussian approximations [120, 165].

The actual number of bits is not that much more than four, and there are four gap genes ... maybe it's just one bit per gene after all? The way to test this is to imagine that the readout mechanisms can only resolve two levels of expression, high and low; formally this means transforming

$$g_i \to \sigma_i = H(g_i - \theta_i), \tag{93}$$

where H is the step function and θ is a threshold that divides the two levels. If we know the thresholds then we can compute $I_\theta(\{\sigma_i\}; x)$, where the notation reminds us that the answer depends on the thresholds $\theta = \{\theta_i\}$. To be generous, we can optimize the thresholds, maximizing how much information this on/off description of gene expression can capture. The answer is that $I(\{\sigma_i\}; x) < 3$ bits [120]. Further, this information is distributed very inhomogenously, so that some "binary words" $\{\sigma_i\}$ point to x values that span $\sim 10\%$ of the anterior–posterior axis. Thus, to extract ~ 4 bits of positional information from four genes requires mechanisms that have much more than this capacity to read out the expression levels.

Recently we returned to the issue of readout precision [169, 170]. We know, as just explained, that reading each concentration with one bit of precision is not enough to extract all the available information. On the other hand, our discussion of decoding positional information from the gap genes (§2) assumed that cells had access to the true measured concentrations

[31] Estimating information theoretic quantities from real data can be challenging, and there are interesting theoretical questions about how best to do this, especially if you need control over the error bars (as in this case). See, for example, Appendix A.8 of Ref [57].

of each molecule, which surely is unrealistic. How can the embryo best trade bits of precision in the readout against the bits of relevant positional information that are preserved?

Formally we can imagine that the expression levels of the gap genes are mapped into some intermediate variable, such as the occupancy of binding sites along the DNA; let's call this intermediate variable C. Inevitably this mapping is noisy, and this means that the information which C can carry about $\{g_i\}$, $I(C;\{g_i\})$, is limited. Given this limitation, what mapping $\{g_i\} \to C$ will maximize the information about position $I(C;x)$? This defines a new optimization problem

$$\max_{P(C|\{g_i\})} [I(C;x) - TI(C;\{g_i\})] , \tag{94}$$

where T is a Lagrange multiplier and we make explicit that the mapping $\{g_i\} \to C$ is probabilistic; this is an example of the information bottleneck problem [171]. To extract all the available information, that is to have $I(C;x) \approx I(\{g_i\};x)$, requires readout mechanisms with a capacity $I(C;\{g_i\})$ of at least 8 bits [169]. Further, the most efficient mechanisms involve C being sensitive to combinations of the different expression levels; if we have separate sensors $g_i \to C_i$ they need to have vastly more capacity. Finally, the embryo is in a regime where the trading of $I(C;x)$ vs $I(C;\{g_i\})$ has a universal form [170]. If we take seriously the idea that C is something like the occupancy of binding sites, or the collective states of the "enhancers" into which these sites are grouped, then these arguments about trading bits provide a path to predict the molecular mechanisms that instantiate the optimal decoding strategies of §2.3.

Before leaving this topic I want to emphasize that Figs 18, 19, and 20 all are testing the *same* optimization principle. Thus we are making predictions about neurons that generate graded voltages, neurons that generate spikes, and networks of genes, all with the same theoretical idea and all connecting to experiment with no free parameters.

4 Network architecture

In the flow of positional information from the maternal inputs to the gap genes to the pair–rule genes, we have seen evidence for optimization in the distribution of inputs (§3.3) and in the processing of the outputs (§2.3). We now have to ask if we can apply optimization principles to the network itself. Doing this involves returning to the challenge laid out in the first lecture: realistic descriptions of the gap gene network require 50+ parameters, and are in some obvious sense very complicated. To approach optimization of such a complex network it will be useful to break off smaller pieces of the problem and gain intuition.

The idea that we can derive the functional behavior of a network from the optimization of information transmission is quite old, having its origins in the context of sensory information processing by the brain. Just a decade after Shannon's original work on information theory the National Physical Laboratory in the UK hosted a remarkable *Symposium on the Mechanization of Thought Processes*. Among other presentations, Horace Barlow spoke about "Sensory mechanisms, the reduction of redundancy, and intelligence" [172]. This was, I think, the first place where optimization principles for neural information processing were articulated.[32] It also seems worth emphasizing the ambition that one sees in these titles, both of Barlow's paper and for the symposium as a whole.

4.1 Linear filtering in neural networks

Let's dive in and think about a network in which a layer of input variables $\{x_i\}$ drives a layer of output variables $\{y_i\}$, as in Fig 21. The simplest possibility is that the transformation is a

[32] Barlow revisited these ideas at another conference a few years later, and this is the more widely cited version of his ideas about "efficient coding" [173]. Thinking about our modern publication habits, it seems worth noting that this work had impact on generations of scientists even though there is no regular journal article to cite.

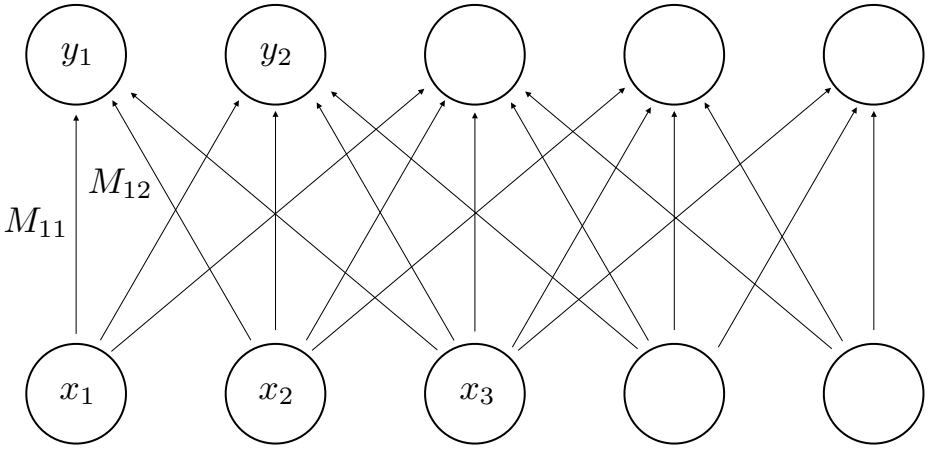

Figure 21: Transformation from inputs $\{x_i\}$ to outputs $\{y_i\}$, as in Eq (95). Only a fraction of possible connections M_{ij} are illustrated; fewer still are labelled. Not shown are the independent noise sources added to each output element.

noisy linear mapping, so that

$$y_i = \sum_j M_{ij} x_j + \zeta_i \,, \tag{95}$$

again to keep things simple we imagine the number of inputs and outputs are the same so $i = 1, 2, \cdots, N$ and similarly for j. Let the noise ζ_i be Gaussian and independent at every site i. In this case optimizing the network means choosing the transformation matrix $\hat{M}$ to maximize the mutual information between inputs and outputs. If the variance of the noise is fixed, then one can always increase the mutual information by increasing the magnitude of $\hat{M}$, which seems like cheating. To have a well defined problem we should bound the magnitudes, so that the scale of the noise has meaning in relation to the inputs; a conventional choice is to constrain the summed variances of the outputs. This leaves us with the optimization problem

$$\max_{\hat{M}} \left[I(\{y_i\}; \{x_i\}) - \mu \sum_{i=1}^{N} \langle y_i^2 \rangle \right] . \tag{96}$$

The structure of this problem depends on the distribution $P(\{x_i\})$ from which inputs are drawn.

The general version of Eq (96) is challenging. It is much easier if we can approximate $P(\{x_i\})$ as Gaussian. Then we have

$$P(\{x_i\}) = \frac{1}{\sqrt{(2\pi)^N}} \exp\left[-\frac{1}{2} \ln\det \hat{C}_x - \frac{1}{2} \sum_{i,j} x_i \left(\hat{C}_x^{-1}\right)_{ij} x_j \right] , \tag{97}$$

$$P(\{y_i\}|\{x_i\}) = \frac{1}{\sqrt{(2\pi)^N}} \exp\left[-\frac{N}{2} \ln\langle\zeta^2\rangle - \frac{1}{2\langle\zeta^2\rangle} \sum_i \left(y_i - \sum_j M_{ij} x_j \right)^2 \right] , \tag{98}$$

$$P(\{y_i\}) = \frac{1}{\sqrt{(2\pi)^N}} \exp\left[-\frac{1}{2} \ln\det \hat{C}_y - \frac{1}{2} \sum_{i,j} y_i \left(\hat{C}_y^{-1}\right)_{ij} y_j \right] , \tag{99}$$

the covariance matrices are

$$\left(\hat{C}_x\right)_{ij} = \langle x_i x_j \rangle \,, \tag{100}$$

$$\left(\hat{C}_y\right)_{ij} = \langle y_i y_j \rangle = M_{ik}\left(\hat{C}_x\right)_{kl} M_{jl} + \langle \zeta^2 \rangle \delta_{ij} \,, \tag{101}$$

$$\Rightarrow \hat{C}_y = \hat{M}\hat{C}\hat{M}^T + \langle \zeta^2 \rangle \mathbb{1} \,. \tag{102}$$

We can substitute into the mutual information

$$I(\{y_i\};\{x_i\}) \equiv \int d^N x \int d^N y \, P(\{y_i\}|\{x_i\}) P(\{x_i\}) \log\left[\frac{P(\{y_i\}|\{x_i\})}{P(\{y_i\})}\right] \tag{103}$$

$$= \left\langle \log\left[\frac{P(\{y_i\}|\{x_i\})}{P(\{y_i\})}\right] \right\rangle \tag{104}$$

$$= -\frac{N}{2}\ln\langle\zeta^2\rangle - \frac{1}{2\langle\zeta^2\rangle}\sum_i \left\langle \left(y_i - \sum_j M_{ij} x_j \right)^2 \right\rangle + \frac{1}{2}\ln\det \hat{C}_y + \frac{1}{2}\sum_{i,j}\langle y_i y_j\rangle \left(\hat{C}_y^{-1}\right)_{ij} \,. \tag{105}$$

Notice that, from Eq (95), we have

$$y_i - \sum_j M_{ij} x_j = \zeta_i \Rightarrow \frac{1}{2\langle\zeta^2\rangle}\sum_i \left\langle \left(y_i - \sum_j M_{ij} x_j \right)^2 \right\rangle = \frac{1}{2\langle\zeta^2\rangle}\sum_i \langle\zeta^2\rangle = \frac{N}{2} \,. \tag{106}$$

Similarly

$$\langle y_i y_j \rangle = \left(\hat{C}_y\right)_{ij} \Rightarrow \frac{1}{2}\sum_{i,j}\langle y_i y_j\rangle \left(\hat{C}_y^{-1}\right)_{ij} = \frac{1}{2}\sum_{i,j}\left(\hat{C}_y\right)_{ij}\left(\hat{C}_y^{-1}\right)_{ij} = \frac{N}{2} \,. \tag{107}$$

Thus the complicated looking summations cancel and we have

$$I(\{y_i\};\{x_i\}) = -\frac{N}{2}\ln\langle\zeta^2\rangle + \frac{1}{2}\ln\det\hat{C}_y = \frac{1}{2}\mathrm{Tr}\ln\left[\mathbb{1} + \frac{1}{\langle\zeta^2\rangle}\hat{M}\hat{C}\hat{M}^T\right] \,. \tag{108}$$

If the eigenvalues of $\hat{M}\hat{C}\hat{M}^T$ are λ_n, then the optimization problem in Eq (96) becomes

$$\max\left[\frac{1}{2}\sum_n \ln\left(1 + \lambda_n/\langle\zeta^2\rangle\right) - \mu\sum_n \lambda_n\right] \,. \tag{109}$$

The optimum is reached where all the λ_n are equal.

The correlation structure of the inputs usually is non–trivial, so the eigenvalues of the covariance matrix $\hat{C}_x$ are spread over some spectrum. Optimizing information transmission in the class of problems defined by Eq (95) means rearranging these inputs to "whiten" this spectrum.[33] Notice that if all eigenvalues of $\hat{M}\hat{C}_x\hat{M}^T$ are equal then $\hat{C}_y$ is proportional to the unit matrix and hence different signals y_i and $y_{j\neq i}$ are independent of one another. Thus the optimal $\hat{M}$ diagonalizes the covariance matrix of the input signals, transforming to principal components, and then rescales these so that they have equal variances.

A simple but important example of these ideas is color vision. Recall that in daylight our vision is based on three kinds of cones, each tuned to a different range of photon energies. We can think of these as sensitive approximately to red, green, and blue, though you should never say this in front of someone who actually studies color vision—"red" is a percept, not the label of a cell type; the convention is to describe the different cones as sensitive to long, medium, and

[33] The name comes from the fact that in truly white light all components of the spectrum have equal weight.

short wavelengths. It is an old idea that the three cone signals are processed in well defined combinations corresponding roughly to the percepts of luminance, red vs green, and blue vs yellow. The history of these ideas is complicated, with all sorts of heroic figures shouting at one another. Things were confusing in part because the "opponent process" theory seemed to involve four color axes (red, green, blue, and yellow) while the alternative "trichromatic" theory involved only three (which we now attach to the three types of cones). It seems to have been Schrödinger who sorted this out in 1925 [174]. We now know that the three cone signals indeed are grouped together in the three combinations already in the retina, and it is these combined signals that are transmitted to the brain [175].

In 1983, Buchsbaum and Gottschalk suggested that the transformation into luminance, red vs green, and blue vs yellow might be the solution to the optimization problem we are discussing here [176]. In order to see if this makes sense, we need to know the covariance matrix of the three cone signals. This depends strongly on the absorption spectra of the cone pigments, with large positive correlations arising simply because these spectra overlap. But the cone signal statistics also depend on the spectral composition of the images that the eye is looking at. To get at these, Ruderman, Cronin, and Chiao used a hyperspectral camera to take pictures in natural environments—woodlands, forests and rainforests, and a mangrove swamp [177]. In effect this camera does spectroscopy in each pixel of a digital image, and then these spectrally resolved images can be projected onto the known sensitivities of the individual cone pigments to estimate the photon capture in each of the three cones looking at the same point. In keeping with the discussion of adaptation above (§3.2) the cone signals x_i were taken as the log of the number of photon counts.[34]

The results obtained by Ruderman et al are remarkably simple and crisp. In our notation, the optimal matrix $\hat{M}$ takes the form

$$\hat{M} = \begin{bmatrix} \frac{1}{\sqrt{3}\sigma_\ell} & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}\sigma_{by}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}\sigma_{rg}} \end{bmatrix} \begin{bmatrix} 1.004 & 1.005 & 0.991 \\ 1.014 & 0.968 & -2.009 \\ 0.993 & -1.007 & 0.016 \end{bmatrix} . \tag{110}$$

We see that y_1 is the sum of the signals in the three cones, and the weights are equal in the second decimal place or better. Similarly y_3 is the difference between the signals in long and medium length cones, with almost no contribution from the short wavelength cones; this is the "red minus green" opponent channel. Finally y_2 combines the long and medium cone signals and subtracts the short wavelength signal, corresponding to "yellow minus blue." As with the luminance channel y_1, the combinations of cone signals in the two opponent channels are in integer ratios, with one percent accuracy. The signals along the different channels have standard deviations in the ratio $\sigma_\ell : \sigma_{by} : \sigma_{rg} \sim 1 : 0.2 : 0.02$. While we have a great appreciation for the subtleties of color, most of the variance in the images that we see is in the luminance channel.[35] Embarrassingly, I don't know if the combination of cone signals into neural signals has been measured precisely enough to test these predictions of integer ratios.

Staying with the visual system, we can think of the $\{x_i\}$ as signals from cones at different positions (now neglecting color). If the cones form a regular lattice, and the input visual signals are translationally invariant, then everything will be diagonal after a Fourier transform. Let's call $\mathbf{k}$ the spatial Fourier variable, and then

$$\tilde{y}(\mathbf{k}) = \tilde{M}(\mathbf{k})\tilde{x}(\mathbf{k}) + \tilde{\zeta}(\mathbf{k}) . \tag{111}$$

[34] We use our cones in bright light, so there is no worry about zero counts.

[35] In fact this might be a cautionary tale about dimensionality reduction. We can capture $\sim 95\%$ of the variance in what we see with just one dimension, corresponding to a completely greyscale world.

Further, the transformation $\tilde{M}(\mathbf{k})$ that maximizes information transmission will obey

$$|\tilde{M}(\mathbf{k})| \propto \frac{1}{\sqrt{S_x(\mathbf{k})}}, \tag{112}$$

where $S_x(\mathbf{k})$ is the power spectrum of the signals $\{x_i\}$ [178, 179]. We know that the spatial power spectrum of natural images is scale invariant [180],

$$S(\mathbf{k}) = \frac{A}{|\mathbf{k}|^{2-\eta}}, \tag{113}$$

so the prediction is $|\tilde{M}(\mathbf{k})| \propto |\mathbf{k}|^{1-\eta/2}$. The growth with $|\mathbf{k}|$ is cut off by the effects of noise in the inputs x_i, e.g. the random arrival of photons.

Given the statistical structure of natural images, optimization of information flow predicts that there will be zero gain for zero spatial frequency, $|\tilde{M}(\mathbf{k}=0)| = 0$. The transformation M_{ij} thus serves roughly as a spatial differentiator, so that the output y_i is large when the pattern $\{x_j\}$ includes a sharp edge, and is small when the $\{x_j\}$ are nearly uniform. Qualitatively these predictions are correct for neurons in the retina. The output cells, which carry information from eye to brain, have long been known to have "center–surround" receptive fields [181,182]: the response of these neurons is driven by the difference between the average light intensity in a small central region and the average over a larger surrounding region, and in many cases the two regions have equal weight so that the response is differentiating. Barlow understood that this sort of spatial differencing, also called "lateral inhibition" [183], would remove the redundancy between signals in neighboring photoreceptors, enhancing the transmission of information by limited numbers of action potentials along limited numbers of neurons [172, 173]; the analysis here translates this qualitative observation into equations.

These arguments about whitening apply also in the time domain. In the simplest example

$$y(t) = \int d\tau M(\tau) x(t-\tau) + \eta(t). \tag{114}$$

If η is white noise, $\langle \eta(t)\eta(t')\rangle = \mathcal{N}_0\delta(t-t')$, then the rate at which information about $x(t)$ is conveyed by $y(t)$ is [57, 148, 184]

$$R_{\text{info}} = \frac{1}{2}\int \frac{d\omega}{2\pi} \ln\left[1 + \frac{1}{\mathcal{N}_0}|\tilde{M}(\omega)|^2 S_x(\omega)\right], \tag{115}$$

where $S_x(\omega)$ is the power spectrum of the input signal

$$\langle x(t)x(t')\rangle = \int \frac{d\omega}{2\pi} e^{-i\omega(t-t')} S_x(\omega). \tag{116}$$

If we optimize R_{info} while holding fixed the output dynamic range $\langle y^2\rangle$, then we find an analog of Eq (112),

$$|\tilde{M}(\omega)| \propto \frac{1}{\sqrt{S_x(\omega)}}. \tag{117}$$

If input signals have scale invariant dynamics, $S_x(\omega) \sim 1/|\omega|$, then we predict $|\tilde{M}(\omega)| \sim \sqrt{|\omega|}$. This is weird, since it corresponds to the filter M taking half of a derivative.

Direct measurements of the transformation M between neurons are not so easy. Our linear models make more sense if the cells have graded voltage responses, as in the first stages of vision. In the fly retina, for example, both the photoreceptor cells and the next cells in line

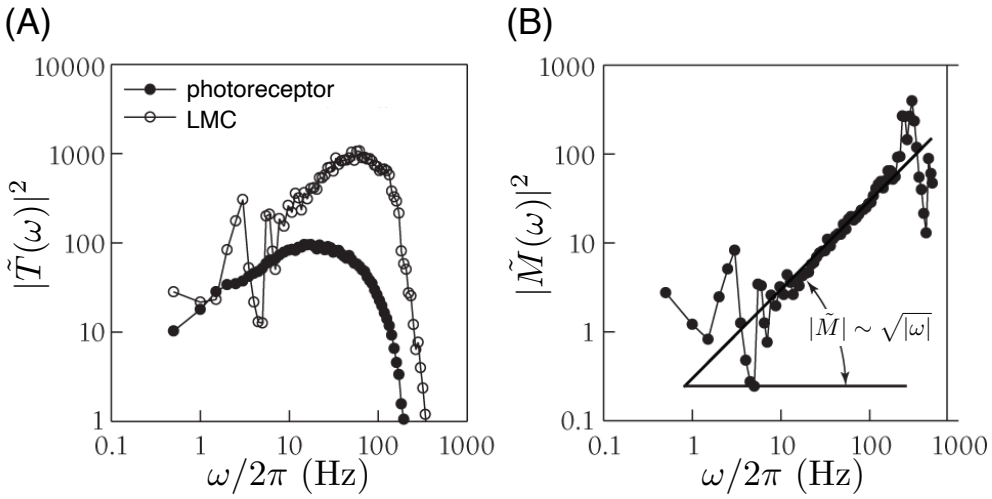

Figure 22: Filtering at the first synapse in fly vision. (A) Linear response of voltage to temporal variations in light intensity, as in Eq (118), for photoreceptors and large monopolar cells (LMC). (B) Effective transfer function between photoreceptors and LMCs, from Eq (119). My thanks to SB Laughlin and RR de Ruyter van Steveninck for sharing these data, from experiments described in Ref [185].

(large monopolar cells, LMCs from §3.2) have a large regime in which the average voltage $V(t)$ responds linearly to time variations in light intensity $I(t)$ around a background I_0,

$$\langle V(t)\rangle - V_0 = \frac{1}{I_0}\int d\tau\, T(\tau)[I(t-\tau)-I_0], \tag{118}$$

defining a transfer function $T(\tau)$; it is more convenient to think about the Fourier transform $\tilde{T}(\omega)$. These transfer functions can be measured in both the receptors and LMCs, with results in Fig 22A. Then we can take the ratio to estimate the filter across the synapse,

$$\tilde{M}(\omega) = \frac{\tilde{T}_{\rm LMC}(\omega)}{T_{\rm receptor}(\omega)}, \tag{119}$$

with results in Fig 22B. Strikingly, we really do see $|\tilde{M}(\omega)| \sim \sqrt{|\omega|}$ across a wide range of frequencies. Such fractional differentiation is much more widespread, perhaps serving to optimize the transmission of information about scale invariant signals throughout the brain [186, 187]. These ideas have distant but fascinating precursors [188].

4.2 Ingredients of a genetic network

The simplest example of a genetic network was shown schematically in Fig 3. To be more consistent with the notation below, let's describe this by saying that a single transcription factor (TF) at concentration c controls the expression level g of a target gene. We'll assume for simplicity that the system is in steady state.

The information that g provides about the input concentration c is

$$I(g;c) = \int dg \int dc P(g|c)P_{\rm in}(c)\log\left[\frac{P(g|c)}{P_{\rm out}(g)}\right], \tag{120}$$

where the distribution of outputs

$$P_{\rm out}(g) = \int dc\, P(g|c) P_{\rm in}(c)\,. \tag{121}$$

Consistent with the analysis of decoding in §2.3, we'll assume that the noise in the response of g to the input c is Gaussian, so that

$$P(g|c) = \frac{1}{\sqrt{2\pi\sigma_g^2(c)}} \exp\left[-\frac{(g-\bar{g}(c))^2}{2\sigma_g^2(c)}\right]. \tag{122}$$

We expect that the mean expression level $\bar{g}(c)$ has a sigmoidal dependence on the TF concentration, as in Fig 3, which we can write as

$$\bar{g}(c) = \frac{c^h}{K^h + c^h}\,, \tag{123}$$

$h > 0$ means that the TF activates expression and K is the concentration at which this effect is half–maximal. More quantitatively h is a measure of sensitivity. We can rationalize this form by imagining the h molecules of the TF bind cooperatively to sites along the DNA, and this binding influences transcription. To complete our description, and the calculation of the information transmission, we need to know the noise level $\sigma_g(c)$.

The molecules whose concentration is measured by g ultimately are made one at a time, and these are random events. So we expect $\sqrt{N}$ fluctuations in making N molecules. We have to be a bit careful, first because we have chosen units in which the maximum $\bar{g}$ is unity, and second because many proteins can be translated from a single mRNA molecule before it is degraded. Thus we can write the contribution of this counting noise as

$$\sigma^2_{g,\rm count}(c) = \frac{1}{N_{\rm max}} \bar{g}(c)\,, \tag{124}$$

where $N_{\rm max}$ is the maximum number of *independent* molecules being made. Probably this is the number of mRNAs, but we can be flexible.

In order to regulate gene expression, the TF molecules have to arrive at a small target along the DNA, where they bind to specific sequences. We can thus think of the regulatory mechanism as a small sensor of transcription factor concentration, with gene expression as the readout. The physical limits to sensing concentrations were first discussed by Berg and Purcell in the context of bacterial chemotaxis [56]. This remains one of the foundational papers of our field, so it is worth taking a detour to review their arguments, and some of the subsequent developments, leading up to Eq (136).

A sensor of linear dimension a sitting in a solution of molecules with concentration c will count $\bar{n} \sim ca^3$ molecules on average, as shown in Fig 23. But as molecules diffuse in and out of the sensitive volume, this molecule count will fluctuate by $\delta n \sim \sqrt{\bar{n}}$. Since concentration is proportional to molecule number, this means that estimates of concentration will have a fractional fluctuation

$$\left.\frac{\delta c}{c}\right|_1 \sim \frac{\delta n}{\bar{n}} \sim \frac{1}{\sqrt{ca^3}}\,, \tag{125}$$

where the subscript reminds us that this is based on one snapshot of molecule counts. We should be able to reduce the noise by averaging over time, but this works only if we make multiple *independent* measurements; just counting the same molecules again doesn't give a better estimate of the surrounding concentration. It takes a time $\tau_c \sim a^2/D$ for molecules in

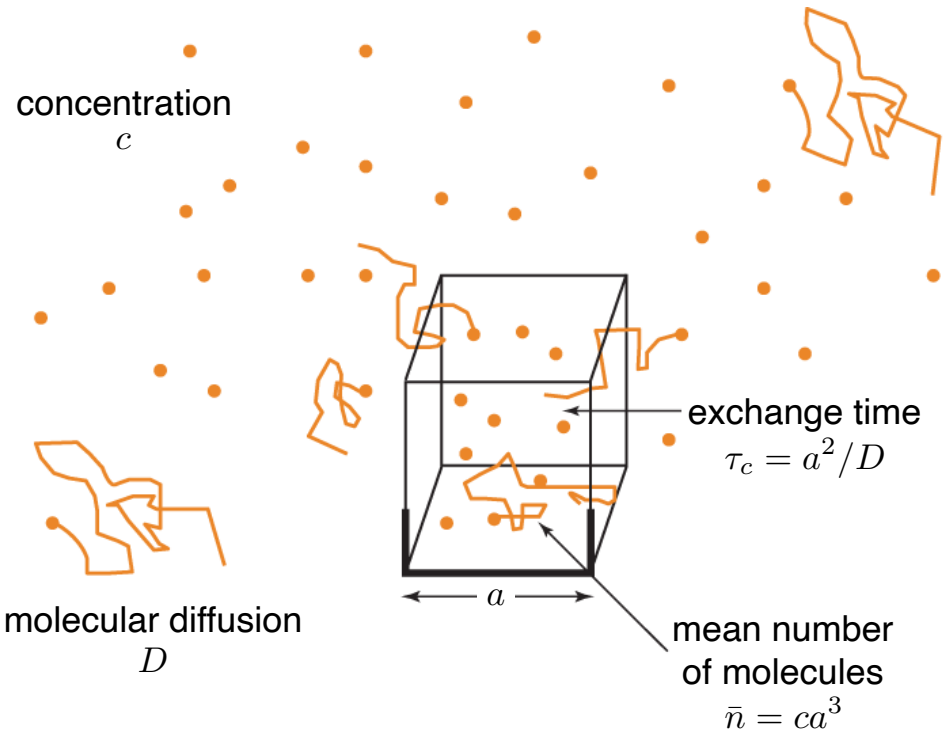

Figure 23: Understanding the physical limits to concentration measurements [56, 57]. A receptor of linear size a samples a volume $\sim a^3$ and thus counts a mean number of molecules $\bar{n} \sim ca^3$, where c is the concentration. Molecules move randomly in and out of the sensitive volume with a diffusion constant D, setting the correlation time $\tau_c \sim a^2/D$. Figure adapted from Ref [57], with permission.

the sensitive volume to exchange with the bulk solution via diffusion, so if we average over time $\tau_{\rm avg}$ we can make $\tau_{\rm avg}/\tau_c$ independent measurements, and thus reduce the noise

$$\frac{\delta c}{c} \to \left.\frac{\delta c}{c}\right|_1 \cdot \frac{1}{\sqrt{\tau_{\rm avg}/\tau_c}} \sim \frac{1}{\sqrt{Dac\tau_{\rm avg}}} \,. \tag{126}$$

We see that the different parameters of the problem combine in a very simple way. This is the "Berg–Purcell limit" to concentration sensing.

It's always good to check that the units work out:

$$Dac\tau_{\rm avg} = \left[\ell^2/t\right][\ell]\left[1/\ell^3\right][t] = [] \,, \tag{127}$$

so this is a dimensionless combination. Note that the 3 in $[1/\ell^3]$ is from three dimensions, so the answer must be different if we are trying to sense the concentration of molecules diffusing in a membrane, or if there is a significant contribution from proteins sliding along the length of the (one–dimensional) DNA molecule [189].

It is useful to note that $\sim Dac$ is the mean rate at which molecules arrive at a (three–dimensional) target of size a via diffusion. For example, if we have the chemical reaction $A + B \to AB$, we write the dynamics of the concentration as

$$\frac{d[AB]}{dt} = k_2[A][B] \,, \tag{128}$$

where k_2 is a "second order rate constant." This rate constant is bounded by the rate at which A and B can find each other, and this is $k_2 \sim Da$, where a is the size of the molecules. One can

make this precise by solving the diffusion equation with appropriate boundary conditions,[36] and develop more precise models in which the molecules have to approach one another at the correct orientation. Thus one can think of Dac as the rate at which molecules arrive at the sensor, and $Dac\tau_{\rm avg}$ as the mean number of molecules that our sensor counts. The Berg–Purcell limit then is the statement that there are $\sqrt{n}$ fluctuations in counting molecules, just as with counting photons from a conventional light source.

In this discussion the linear dimensions of the sensor a is a rough concept. In their original discussion, Berg and Purcell were thinking about a bacterium sensing the concentration of attractive or repulsive molecules in its environment, and the initial guess is that a is just the size of the bacterium itself. But this is weird, since what really happens is that molecules bind to particular receptors on the cell surface. These receptors are themselves of molecular dimensions—nanometers rather than the micron size of the whole cell—and receptors for any particular molecules cover only a tiny fraction of the cell's surface.

If there is just one receptor of linear dimension $a_{\rm r}$ then the Berg–Purcell argument should still work. If there are $N_{\rm r}$ of these receptors then it is plausible that noise is reduced by a factor $1/\sqrt{N_{\rm r}}$, which is equivalent to saying they act together as a single receptor with effective size $a_{\rm eff} \sim N_{\rm r} a_{\rm r}$. But if the number of receptors becomes large enough that the area occupied by the receptors becomes comparable to the area of the cell surface, $N_{\rm r} a_{\rm r}^2 \sim a_{\rm cell}^2$, then it seems like the whole cell should act as one big receptor and $a_{\rm eff} \to a_{\rm cell}$. How does this work?

In a fabulous bit of hand waving plus analogies, Berg and Purcell argued that the tendency of random walk trajectories to bounce along a surface leads to correlations in the encounters of individual molecules with nearby receptors. With $N_{\rm r}$ receptors scattered over the surface of the cell, this leads to an effective size

$$a_{\rm eff} \sim a_{\rm cell} \frac{N_r a_{\rm r}}{N_r a_{\rm r} + a_{\rm cell}} \,. \tag{129}$$

Notice that this saturates when $N_r a_{\rm r} \sim a_{\rm cell}$, at which the fractional coverage of the surface is $N_r a_{\rm r}^2 / a_{\rm cell}^2 \sim a_{\rm r}/a_{\rm cell} \sim 10^{-3}$. This is a dramatic effect, and matters enormously in the life of the cell, which can act as one big sensor even though only a small fraction of its surface is covered by any single receptor type.

The goal of Berg and Purcell's work was (in part) to compare the physical limits to concentration sensing with the performance of real bacteria as they navigate through chemical gradients. The result was that it would be physically impossible for cells to make decisions with the observed reliability if they were making spatial measurements, comparing concentrations at head vs. tail to see if they are moving in the right direction. Instead they must measure changing concentrations in time along the path taken as they swim. They need to average these time derivatives, effectively comparing the recent past with a longer term average, but the duration of this averaging is limited by the rotational Brownian motion of the cell itself. The result is a semi-quantitative theory of what bacteria must do in order to achieve their observed chemotactic performance, and all of these predictions proved to be correct.

One question is whether the Berg–Purcell arguments could be sharpened to give a more quantitative theory of chemotactic strategies, and this continues to be an interesting direction. More relevant to our discussion is whether these limits to concentration sensing are sufficiently general that they can be applied, for example, to the problem of transcriptional control. This matters because of a simple order–of–magnitude argument. Transcription factors function at concentrations of tens of nanoMolar,

$$c \sim 10\,{\rm nM} = 10 \times 10^{-9} \times (6 \times 10^{23}) \frac{1}{10^3\,({\rm cm})^3} \times \left(\frac{10^{-4}\,{\rm cm}}{\mu{\rm m}} \right)^3 \sim 6\,(\mu{\rm m})^{-3} \,. \tag{130}$$

[36] See Problem 54 in Ref [57].

Diffusion constants for proteins in the cytoplasm are $D \sim 1\,\mu\text{m}^2/\text{s}$, and the target to which these molecules bind has dimensions $a \sim 3\,\text{nm}$. The result is that

$$Dac \sim 2 \times 10^{-2} s^{-1} \Rightarrow \frac{\delta c}{c} \sim \frac{1}{\sqrt{Dac\tau_{\text{avg}}}} \sim \left(\frac{1\,\text{min}}{\tau_{\text{avg}}} \right)^{1/2} . \tag{131}$$

This suggests that reliable responses to ten percent differences in transcription factor concentration will require more than an hour of temporal integration. Neighboring cells in the fly embryo experience maternal inputs that differ by $\sim 10\%$ in concentration, and they reliably express different combinations of gap genes; this certainly takes less than one hour [125].

The conclusion from Eq (131) is not that transcriptional regulation definitely reaches the physical limits to concentration sensing, but rather that it operates in a regime where these limits are relevant. Even this more modest conclusion hinges on applying the Berg–Purcell ideas far from their origins; this is made more uncertain by the beautifully intuitive but non–rigorous nature of the original arguments.[37] Some years ago my colleagues and I started to worry about this [189–191], and by now the literature has grown substantially [192–199]. Here is what I think we know:

- One way to make the Berg–Purcell argument rigorous is to analyze the fluctuations in occupancy of a receptor binding site as it comes to equilibrium with molecules diffusing in the surrounding solution. These fluctuations are a form of thermal noise, subject to the fluctuation–dissipation theorem.
- Using the fluctuation–dissipation theorem we can see explicitly how diffusion among nearby receptors generates correlated noise and results with the flavor of Eq (129).
- Cooperativity among multiple binding events enhances sensitivity and reduces excess noise, but never below the bound set by Berg and Purcell.
- Analytic results from the fluctuation dissipation theorem can be reproduced by careful numerical simulation, though small discrepancies remain to be understood.
- Cells can push below the Berg–Purcell limit by factors ~ 2 with signal processing strategies that are more sophisticated than just averaging receptor occupancy.
- These strategies only work away from equilibrium, providing a path to connect energy dissipation and signaling accuracy.
- There are interesting generalizations to sensing time dependent concentrations, or the concentrations of multiple species by multiple receptors; results can be counterintuitive.

From all this work, the general conclusion is that Berg and Purcell got it right: Eq (126) defines a minimal noise level for sensing concentrations, up to factors of order unity that could also be seen as ambiguity in defining the size a of the detector.

What the Berg–Purcell limit tells us is that the random arrival of molecules at their target site generates an effective concentration noise with variance

$$\sigma_c^2 = \left(\frac{\delta c}{c} \right)^2 c^2 \sim \frac{c}{Da\tau_{\text{avg}}} . \tag{132}$$

[37] I never had the chance to discuss this with Purcell, but Berg was clear that their arguments were rough and that it would be nice to have something more rigorous.

This effective concentration noise will propagate through the genetic regulatory element, contributing to the variance in the output as

$$\sigma^2_{g,\mathrm{BP}}(c) \sim \left|\frac{d\bar{g}(c)}{dc}\right|^2 \sigma_c^2 . \tag{133}$$

Putting this together with the counting noise from Eq (124) we have

$$\sigma_g^2(c) = \sigma^2_{g,\mathrm{count}}(c) + \sigma^2_{g,\mathrm{BP}}(c) = \frac{1}{N_{\mathrm{max}}}\bar{g}(c) + \left|\frac{d\bar{g}(c)}{dc}\right|^2 \frac{c}{Da\tau_{\mathrm{avg}}} \tag{134}$$

$$= \frac{1}{N_{\mathrm{max}}}\left[\bar{g}(c) + (c/c_0)\left|\frac{d\bar{g}(c)}{d(c/c_0)}\right|^2\right], \tag{135}$$

where the natural units of concentration are $c_0 = N_{\mathrm{max}}/Da\tau_{\mathrm{avg}}$. From the numbers above, we see that real transcription factor concentrations are comparable to c_0, which is part of why the optimization of information transmission leads to interesting results.

Now we can make use of results from the previous lecture, §3.1. First we propagate the output noise σ_g back through the input/output relation, as in Fig 17, to obtain an effective input noise

$$\sigma_c^{\mathrm{eff}} = \left|\frac{d\bar{g}(c)}{dc}\right|^{-1} \sigma_g(c) = \frac{c_0}{h\sqrt{N_{\mathrm{max}}}} \frac{(c/c_0)}{\bar{g}(c)[1-\bar{g}(c)]} \left[\bar{g}(c) + \frac{h^2}{(c/c_0)}\bar{g}(c)[1-\bar{g}(c)]\right]^{1/2} . \tag{136}$$

Then we can work in the small noise limit to estimate the maximum information, from Eqs (60) and (61),

$$I_{\mathrm{max}} = \log_2\left[\frac{1}{\sqrt{2\pi e}}\int_0^{c_{\mathrm{max}}} \frac{dc}{\sigma_c^{\mathrm{eff}}}\right] \mathrm{bits}, \tag{137}$$

where we go back to conventional units and note explicitly that the transcription factor has some maximum concentration c_{max}. Playing with this a bit one can see that

$$I_{\mathrm{max}} = \frac{1}{2}\log_2\left[\frac{N_{\mathrm{max}}}{\sqrt{2\pi e}}\right] + F(h, K/c_0; c_{\mathrm{max}}/c_0). \tag{138}$$

Thus we can optimize the two parameters h and K, and the answer will depend on the maximum concentration c_{max}; again it is natural to express concentrations in units of c_0.

Figure 24A shows the results of this optimization at $c_{\mathrm{max}}/c_0 = 1$ [200]. Perhaps the most important conclusion is that the optimal parameters are perfectly sensible, not driven off to extreme values. This happens because of the interplay between the two components of the noise, counting at the output and Berg–Purcell at the input.

A natural generalization is to the case where the single transcription factor controls multiple genes, at expression levels $\{g_1, g_2, \cdots, g_K\}$, but to keep things simple these target genes do not interact. In Figure 24B–G we see the results when there are K = 5 targets and c_{max}/c_0 changes across a dynamic range of 30×. At small values of c_{max}/c_0 the optimal solution is for all K genes to have the same values of K and h, so that they are completely redundant copies of one another (Fig 24B). I have always found this result fascinating because redundancy often is taken as prima facie evidence against any information theoretic optimization principle. Here we see that redundancy actually is the result of such a principle, in the right regime. I don't know whether this is a path to understanding the appearance of redundancy in biological signaling more broadly, but it surely is an object lesson.

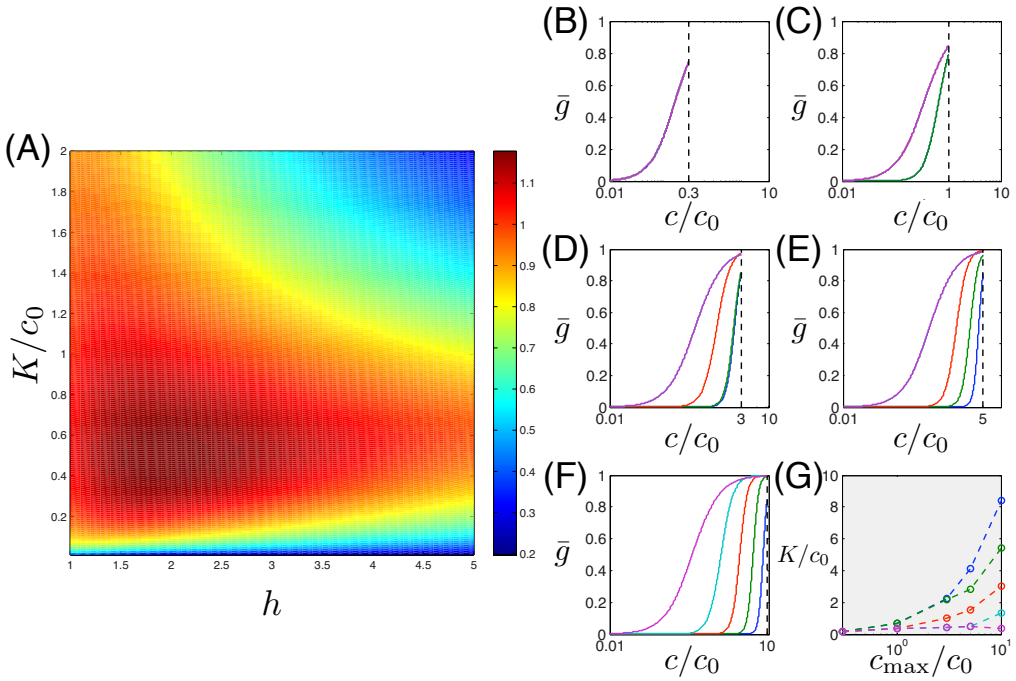

Figure 24: Optimizing information about the concentration of a single transcription factor. (A) A map of F from Eq (138) as a function of K and h, showing a single clear optimum at reasonable parameters ($c_{max}/c_0 = 1$). (B–F) Optimal input/output relations $\bar{g}_i(c)$ with five target genes, with increasing values of c_{max}/c_0 indicated by dashed vertical lines. At smaller c_{max}/c_0 the optimal solutions are redundant, so fewer distinct input/output relations are visible. (G) Optimal values of K/c_0, showing a sequence of bifurcations as c_{max}/c_0 increases and responses become distinct. Redrawn from Ref [200], with thanks to G Tkačik and AM Walczak.

As the maximum allowed concentration of the input transcription factor increases, the optimal strategy for information transmission changes and more of the target genes come to have distinguishable responses (Fig 24B–F). This happens through a series of bifurcations that we can visualize in a plot of K/c_0 vs c_{max}/c_0 (Fig 24G). Successive bifurcations add distinct target gene responses at higher concentration (larger K) and with steeper responses (larger h), until the set of input/output relations "tile" the full dynamic range of inputs.

The results of Fig 24 are just in the case where a single transcription factor activates a set of non–interacting target genes. I should admit that when we started thinking about these problems my colleagues and I thought that the path from these simple examples to something realistic would be quick, but we were wrong. We did learn about some pieces of the problem.

To begin, the pattern of responses from multiple targets in Fig 24 is redundant, because activation of the genes with large K allows us to infer that the genes with smaller K also are active. This redundancy can be reduced, and information transmission enhanced, by repressive interactions among the targets, as shown in Fig 25 [201]. This parallels the discussion of lateral inhibition in the retina, outlined in §4.1, and produces profiles of (mean) expression level vs input concentration that remind us of those shown by the gap genes. These benefits of repression in enhancing efficiency are seen even when there are no feedback loops.

The prototypical example of feedback is when a single target gene also can activate itself [202]. This system has a bistable regime at sufficiently strong self–activation, but the optimal parameter values are on the monostable side of this bifurcation. As the bifurcation is approached, however, there is critical slowing down. This emergent long time scale serves

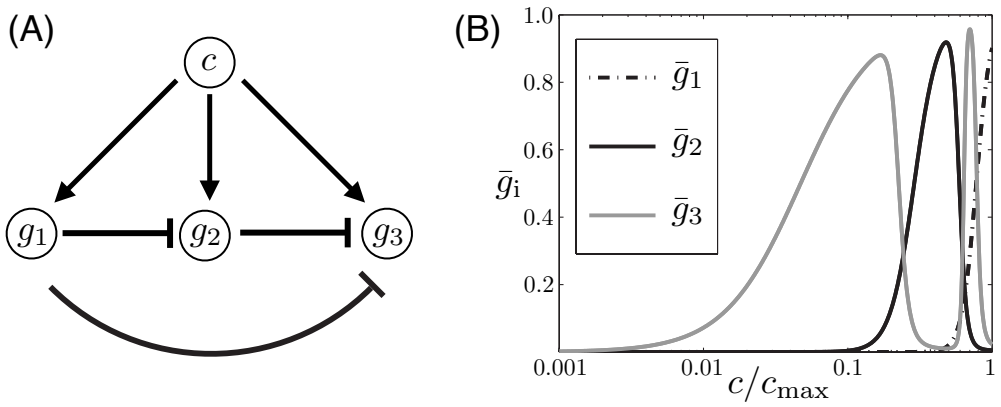

Figure 25: Redundancy reduction. (A) A single transcription factor (concentration c) activates multiple genes $\{g_1, g_2, g_3\}$ that repress one another. (B) Mean expression levels that optimize information transmission, with $c_{max}/c_0 = 10$. Redrawn from Ref [201], with thanks to G Tkačik and AM Walczak.

to reduce noise, but near criticality there is also a strong path for noise in the output to be injected back into the system. The closeness of the optimum to the critical point thus depends on the maximal concentration of output molecules N_{max}.

In many systems, including the fly embryo, multiple cells or nuclei can exchange proteins or mRNA through diffusion. This generates a spatial averaging that can reduce noise, adding to the effects of temporal averaging.[38] Importantly, because diffusion itself is noisy, this effect can't push the final noise level below the Poisson level of counting noise. In the embryo, meaningful variations in input and output are laid out in space, and of course sufficiently strong diffusion will degrade these patterns. The result is that there is an optimal diffusion constant that maximizes information transmission [204].

In Figures 24B–F we see that even as the different target genes acquire distinct input/output relations, the optimal solutions still do not make much use of the lowest accessible concentrations; this regime is simply too noisy to allow reliable information transmission. But suppose that the cell makes many mRNA molecules of some intermediate protein y, and that the protein c binds to these mRNA and represses translation, as in Fig 26; y can then act indirectly as a transcriptional repressor so that the whole path $c \to g$ is activating. Then the "measurement" of low concentrations is made in parallel at many sites, rather than at just one site along the DNA, and the response is largest and most reliable at the lowest concentrations. This suggests that a molecule which functions both as a transcriptional activator and a translational repressor could make better use of its full dynamic range. It requires a careful calculation to show that information transmission really can be larger even when we constrain the total number of molecules, but it works [205]. One of the primary maternal morphogens in the fly embryo belongs to a whole class of proteins that function as such dual transcription/translation regulators [206–209], and our best estimates put this system in the regime where dual regulation enhances information transmission. I found it remarkable that this almost baroque level of biological complexity can be derived as part of the solution to a fundamental physics problem faced by the cell.

Each of these pieces—repression to reduce redundancy, feedback to average over time, diffusion to average over space, and dual regulation—is an interesting problem by itself, and it was possible to make progress with a combination of analytic and (modest) numerical ap-

[38] This combination of spatial and temporal averaging seems to be necessary to understand how high levels of noise in the initiation of transcription yield low levels of noise in the concentration of the gap gene proteins [203].

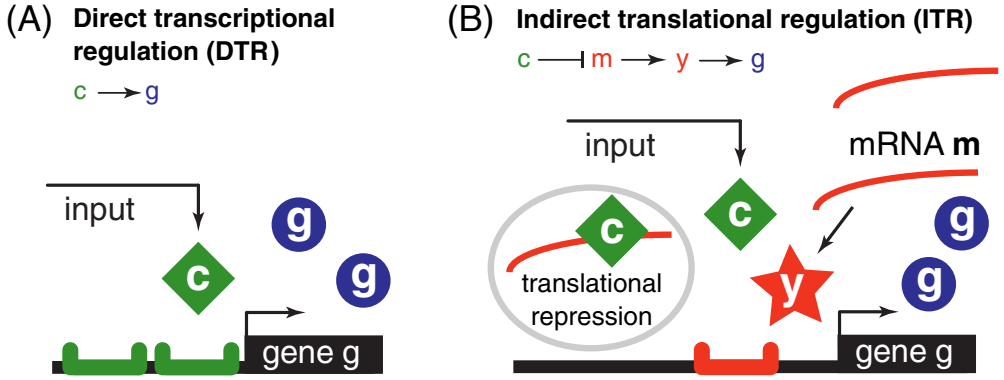

Figure 26: Schematic of direct transcriptional regulation (DTR) and indirect translational regulation (ITR). (A) In DTR, as above, activator TFs (green squares at concentration c) interact with binding sites along the DNA to activate expression of the regulated gene g. (B) In ITR scenario, input molecules (green squares) bind to mRNAs m of protein y (red chain) to make the mRNA inaccessible for translation (gray oval). Translation proceeds from unbound mRNA molecules, giving rise to proteins y (red stars). These proteins act as repressors for gene g; the overall mapping from $c \to g$ is thus activating in both scenarios. Redrawn from Ref [205], with thanks to TR Sokolowski, G Tkačik, and AM Walczak.

proaches. But all of these things are happening at once in the fly embryo, even just in the gap gene network. To put (most of) these pieces together requires a more sophisticated approach [210].

4.3 The gap genes, once more

Let's plunge right in. We have a collection of nuclei labeled $\mathrm{n} = 1, 2, \cdots, N$ along the anterior–posterior axis of the embryo, with a distance Δ from one to the next. Associated with each nucleus are the expression levels of the four gap genes $\{g_i(\mathrm{n})\}$, and these molecules can diffuse between neighbors with an effective diffusion constant D. As before we will keep track of one concentration for each species, not worrying about the separate dynamics of mRNA and proteins. We assume that each gap gene product has the same maximal synthesis rate $r_{\max}$ and the same decay time τ; these are simplifications but also supported by experiment. The gap genes are driven by maternal inputs $\{c_\alpha(\mathrm{n})\}$ with $\alpha = 1, 2, 3$, and we will assume that theese inputs are constant in time, which we know to be correct over the relevant time window for at least one of them [64]. With these assumptions, the dynamics of the gap gene expression levels obey a generalization of the first equation that we wrote down in these lectures [Eq (1)]:

$$\frac{dg_i(\mathrm{n},t)}{dt} = r_{\max} f_i\left(\{g_j(\mathrm{n},t)\};\{c_\alpha(\mathrm{n})\}\right) - \frac{1}{\tau} g_i(\mathrm{n},t) + \frac{D}{\Delta^2}\left[g_i(\mathrm{n}+1,t) - 2g_i(\mathrm{n},t) + g_i(\mathrm{n}-1,t)\right] + \eta_i(\mathrm{n},t), \tag{139}$$

where f_i is a separate regulation function for each gap gene, and $\eta_i(\mathrm{n},t)$ are noise terms. Three static inputs are driving four interacting genes, as in Fig 27, so there are $(3\times4)+(4\times4) = 28$ arrows, as noted in §1.1.

We take the maternal inputs $\{c_\alpha(\mathrm{n})\}$ as known, and fixed as we try to optimize the gap gene network itself. This certainly is fair for the input which is large at the anterior end, where we have accurate measurements in both live and fixed embryos establishing the peak

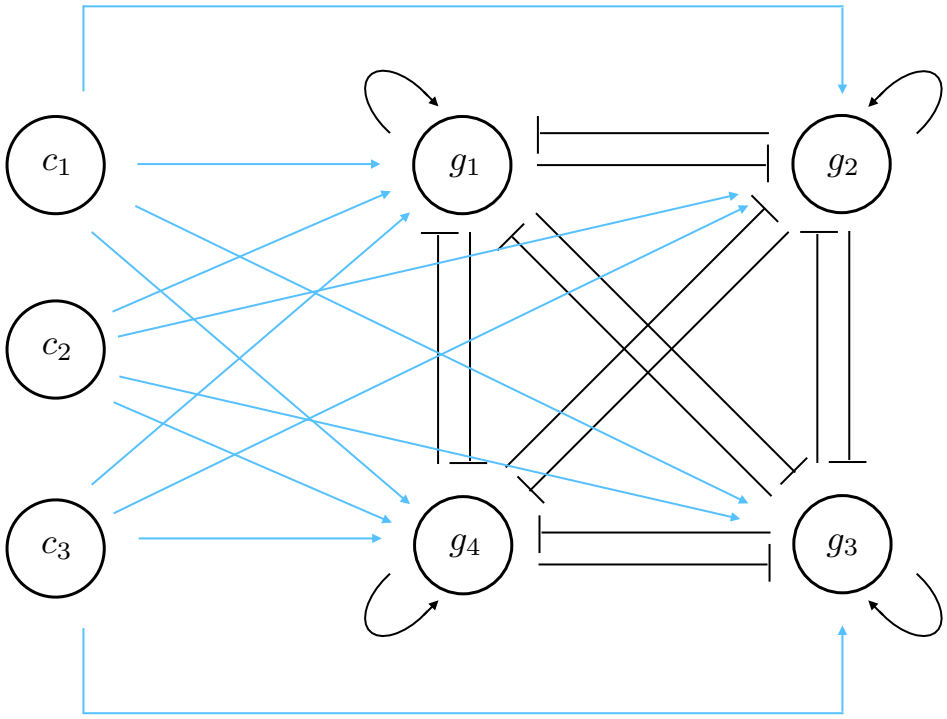

Figure 27: The network of four gap genes $\{g_i\}$ driven by three maternal inputs $\{c_\alpha\}$. Regulatory interactions $c_\alpha \to g_i$ are strictly feedforward (blue arrows). In contrast, interactions $g_i \to g_j$ can form feeback loops. It is thought that $c_\alpha \to g_i$ are largely activating while $g_i \to g_{j\neq i}$ are largely repressive (black blunt arrows), although there is evidence for self–activation $g_i \to g_i$ (curved black arrows). The schematic includes this consensus, which proves to be a feature of the optimized network, though it is not imposed.

absolute concentration and the approximate exponential decay with distance into the egg. We will assume that the input which is large at the posterior end is just a mirror image, and that the terminal inputs are symmetric with the same peak concentration but more rapid decays.

In order to proceed we need a model for the regulation functions. In particular we need to describe what happens as multiple regulatory arrows converge on a single target gene. We have taken a simple view inspired by allostery in proteins [211–214]. In a single large protein molecule, binding of a small molecule at one point on the surface can influence the binding of molecules far away; a classic example is the cooperative binding of four oxygen molecules to hemoglobin in our blood. Monod, Wyman, and Changeux (MWC) proposed that this happens because the protein can exist in two structures, and the small molecules have different binding energies to these two structures. In this picture there is no direct interaction between the binding events; all interactions are mediated through the protein [215]. In the simplest formulation these events occur at thermal equilibrium, even though one often is describing the activity of enzymes, ion channels, and other systems that evidently are not in equilibrium. The idea is that the rate of the events that we are interested in is proportional to the occupancy of some state(s), and this occupancy is well approximated by estimates from equilibrium statistical mechanics. There is a tradition of using such equilibrium arguments for transcriptional control as well [216, 217], although the validity of this quasi–equilibrium view remains an open question [218].

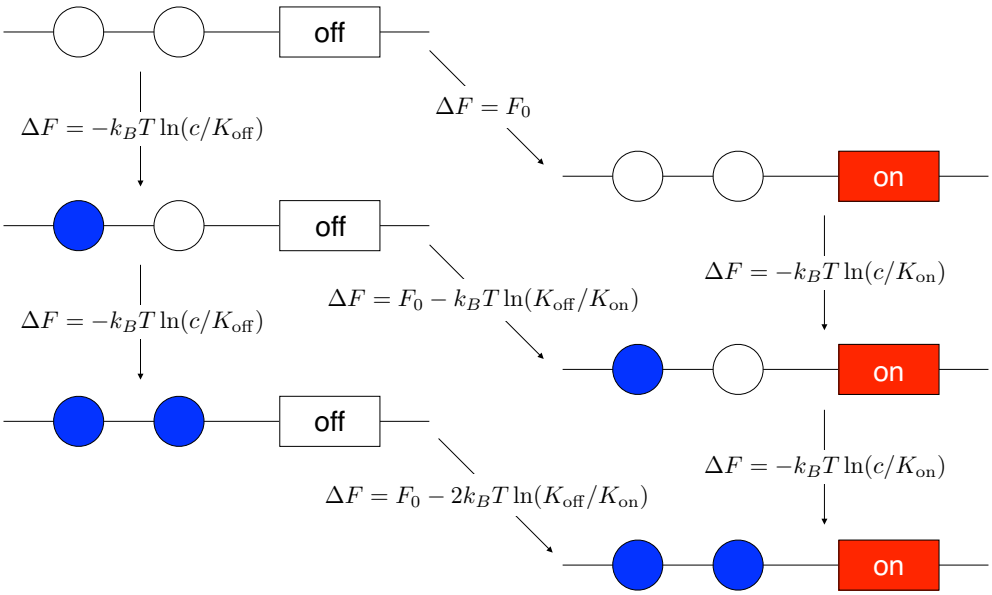

Figure 28: The Monod–Wyman–Changeux (MWC) model for regulation. Here there are two binding sites for the (blue) transcription factor (TF) protein, at concentration c. Binding at the two sites are independent events, but the binding constants $K_{\rm on}$ and $K_{\rm off}$ are different in the two states. If binding is stronger in the on state ($K_{\rm on} < K_{\rm off}$) then the free energy difference between off and on states increases as TFs bind.

Figure 28 schematizes the MWC model for regulation of a gene by a single transcription factor (TF). The system can be in "on" or "off" states, with transcription occurring in the on state, and there are two binding sites for the TF, which is at concentration c. Assuming the system is in equilibrium we can write the probability of being in the on state in terms of the free energy differences as shown, to give

$$P_{\rm on} = \frac{e^{-F_0/k_BT}(1+c/K_{\rm on})^2}{(1+c/K_{\rm off})^2 + e^{-F_0/k_BT}(1+c/K_{\rm on})^2} \tag{140}$$

$$= \frac{1}{1+\exp[-\mathcal{F}(c)]}, \tag{141}$$

$$\mathcal{F}(c) = -\frac{F_0}{k_BT} + 2\ln\left(\frac{1+c/K_{\rm on}}{1+c/K_{\rm off}}\right). \tag{142}$$

Notice that if binding is very tight in the on state and very weak in the off state, we should have $K_{\rm on} \ll c \ll K_{\rm off}$ and we can rewrite

$$P_{\rm on} = \frac{c^2}{c_{1/2}^2 + c^2}, \tag{143}$$

with $c_{1/2} = K_{\rm on} e^{F_0/2k_BT}$. We identify the probability of being in the on state with the regulation function f. Notice that factor of 2 in Eq (142) is counting the two binding sites.

One attractive feature of the MWC model is that it generalizes to having multiple TFs converge to regulate a single gene. Consider that gene i has $H_{\rm ij}$ binding sites for the protein encoded by gene j and $H_{\rm i\alpha}$ sites for the maternal input α. Then if we follow the same

equilibrium statistical mechanics arguments as above, we will find

$$f_i\left(\{g_j\};\{c_\alpha\}\right) = \frac{1}{1+\exp\left[-\mathcal{F}_i\left(\{g_j\};\{c_\alpha\}\right)\right]}, \tag{144}$$

$$\mathcal{F}_i\left(\{g_j\};\{c_\alpha\}\right) = -\frac{F_{i0}}{k_B T} + \sum_j H_{ij}\ln\left(1+\frac{g_j}{K_{ij}}\right) + \sum_\alpha H_{i\alpha}\ln\left(1+\frac{c_\alpha}{K_{i\alpha}}\right), \tag{145}$$

where we work in the approximation that one of the two binding constants is large (very weak binding). Note that by changing the sign of H we can have TFs act as both activators or repressors. This gives us a parameterization for the regulation functions in Eq (139), and as promised in the first lecture we have two parameters K_{ij} and H_{ij} for each pair of species that can interact.

The noise in Eq (139) has contributions from the input (Berg–Purcell) and output (counting) noise as in Eq (135).[39] There is also noise attached to the diffusion terms, and one has to be careful that there is independent noise in the fluxes $n \to n \pm 1$, not independent noise added to each site; this insures that diffusion noise does not violate conservation of molecules. All of these noise sources are white, as can be seen for example from the fact that variances in Eq (135) are inversely proportional to averaging times. Spectral densities are set by the absolute numbers or concentrations of molecules, as above, and these are known quite well from experiment; for details see Ref [210].

Now that we have all the ingredients in Eq (139), we could just simulate. But our goal is to optimize the parameters so as to maximize the information that expression levels $\{g_i\}$ provide about position or cellular identity n. It would not be enough to have a single solution of these stochastic differential equations, we need to know about the whole *distribution* of solutions. This quickly becomes intractable. Fortunately we know that in the real system noise levels are small and fluctuations are approximately Gaussian. This means that we can linearize Eq (139) in the small fluctuations around the mean, and find closed equations for the time evolution of the means $\langle g_i(n,t)\rangle$ and the covariance matrix $\langle \delta g_i(n,t)\delta g_j(n,t)\rangle$. Thus at a single setting of all the parameters $(\mathbf{H},\mathbf{K})$ we can do one (large) integration forward in time and find everything we need in order to evaluate the positional information at $t \sim 45$ min into nuclear cycle 14. We explore the 50+ dimensional parameter space by a version of simulated annealing [210].

The essential results of the optimization process are shown in Fig 29. We start with some random setting of all the adjustable parameters, and typically this will leave some of the gap genes fully on and some fully off, uniformly across the entire embryo, so that there is zero positional information (point 1 in Fig 29). As we explore and anneal we find parameter settings that allow first one then two, three, and all four gaps genes to be driven on and off by the full dynamic range of the maternal inputs (points 2, 3, and 4). Much of the time spent in optimization is required to converge from these patterns onto something richer and more informative, finally arriving at an optimum (point 5). This optimal network has spatial patterns of gene expression very similar to those of the real network, and the absolute magnitude of the positional information is close as well.

I want to emphasize that in deriving the patterns of gap gene expression that we see in the upper right of Fig 29, there are no free parameters—all are determined by the optimization principle. In fact if we set this principle aside and try to fit the model to the mean profiles, we don't do any better in matching the data. More subtly, these best fits convey substantially less than the maximum information because they correspond to parameter settings with excess noise. Thus in some way the optimization principle is getting us closer to the real network than standard model fitting. Also, the small differences between the optimum and the real

[39] To be fully realistic we can add a small extra noise with constant fractional variance, which seems to be necessary to match the data [210, 219].

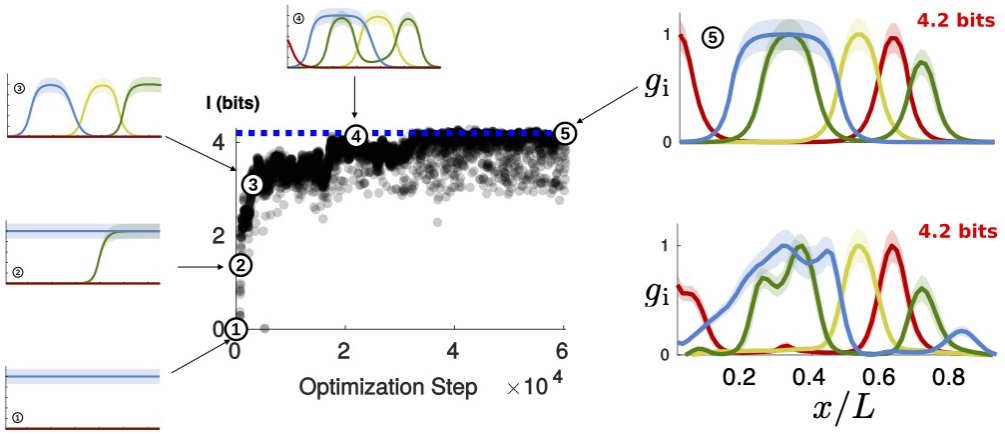

Figure 29: Maximizing positional information in the gap gene network. Random initial parameter settings produce zero positional information, with all gap genes fully on or off across the length of the embryo (1). Across ∼20,000 steps of optimization we find parameter settings that allow first one then two, three, and finally all four gap genes to be modulated by the dynamic range of maternal inputs (2 → 3 → 4). The majority of the optimization run then tunes the parameters to maximize the positional information, arriving at a setting (5) that behaves much like the real network. Redrawn from Ref [210], with thanks to T Gregor, TR Sokolowksi, and G Tkačik.

network must reflect a limitation in the class of models that we are considering, not a failure of optimization. There are a few more interesting features of the solution(s):

- The positional information seen in real networks really is at the edge of what this class of models can support, as seen by random sampling of the parameter space.
- Almost all of the possible interactions are realized in the optimal solution, and their signs agree with available data.
- We have constrained the maximum molecule counts and concentrations; other local optima differ in the mean utilization of these resources.
- We read out positional information at a moment in nuclear cycle 14 where the real expression levels are maximally informative, and do not constrain the dynamics, but the optimal networks have very slowly changing patterns, as with the real system.

We can also think of this as a "laboratory" in which to test alternative scenarios, making a change and then re–optimizing all the remaining parameters:

- We have taken the diffusion constant D in Eq (139) as known, but there is a broad maximum of information near this measured value.
- If we eliminate feedback loops we can still reach similar levels of positional information but only with unphysically large values of the H parameters; within realistic settings feedback is essential.
- If the same maximum number of molecules were spread across only three gap genes it would be impossible to achieve the same level of positional information; benefits of a fifth gap gene are minimal.
- Although cross–regulation within the gap gene network provides some resilience, eliminating any one of the maternal inputs results in significantly less positional information.

I suspect we are just scratching the surface. Perhaps the most intriguing observation is that different local optima have broadly consistent gap gene expression patterns, but with shifts and swaps similar to those seen in related species [220–222].

5 Conclusion

This Summer School celebrates a special moment in the long history of interactions between physics and biology. Just one generation back, physicists and biologists couldn't agree on much, but they could agree that searching for a theoretical physics of life was a waste of time. Physicists saw biology as too messy, and biologists saw the physicists' search for simplicity and universality as a poor match to the evident complexity and diversity of life. Much has changed.

I have emphasized that enormous progress in experiment has created new opportunities for theory. In particular, experiments have revealed that many living systems exhibit behaviors with a precision and reproducibility far beyond what once was imagined. In some cases this precision corresponds to functional behavior close to the limits of what is allowed by the laws of physics. Thus, aspects of early embryonic development can take their place alongside classical examples such as photon counting in vision and molecule counting in bacterial chemotaxis.

It is an old idea that evidence for near–optimal performance could be turned around and formulated as a theoretical principle from which aspects of mechanism and function can be derived. This idea has a checkered history. Many people have used optimization principles in the absence of any direct evidence for optimization, which is fine if the next experiments provide those direct tests. More subtly, when we formulate an optimization principle we often must search among a class of possible mechanisms in order to find the optimum, and too often this class has been woefully oversimplified. Again these simplifications are fine as starting points, but sometimes the more realistic calculations never come. The message is that seemingly simple and clear principles of physical optimization become challenging when we try to use them in the realistic context of life's complexity.

I don't know if what I have described here will survive the next rounds of experiments. But I am excited that we are implementing physical notions of optimization in realistic settings. We have asked how the embryo can best decode positional information contained in the *actual* expression levels of gap genes, and this makes detailed predictions for the distortion of the body plan in mutants; theory and experiment agree quantitatively, with no adjustable parameters (§2.3). We have asked how the embryo can match the statistics of maternal inputs to the *measured* signal and noise characteristics of the gap gene network, and the surprising uniformity of positional information along the anterior–posterior axis emerges as a prediction of this optimization; the real embryo is within two percent of the optimum (§3.3). Finally, we have asked how the architecture and parameters of a *reasonably realistic* model for the gap gene network can be tuned to maximize information transmission, and we find networks very much like the real network emerging as a result, with no free parameters; this optimization principle predicts many features of the network that we did not constrain, and provides a framework for exploring the interplay of chance and necessity (§4.3).

I want to end by saying a few words about what we are not doing. Much of the progress described here, and in other lectures at the Summer School, takes a phenomenological view, hoping to say something crisp about system level functions without digging into too much microscopic detail. There is another community that is much more focused on molecules, and the divide between these communities can be quite stark. To give one example, we probably know more about the functional dynamics of individual ion channels than about any other class of protein molecules, and we have a precise theoretical framework for how these molecular dynamics shape the electrical dynamics of single neurons. There are 100+ different kinds

of channels encoded in the genome, and single neurons express a cocktail of perhaps seven different kinds. But all of this molecular richness disappears rather abruptly once we start to talk about coding and computation in networks of neurons.

As physicists we are used to the idea that macroscopic phenomena are described by coarse–grained models, and that in these models many microscopic details have disappeared (see Approach 3 in §1.2). But in the examples we understand, from the inanimate world, we can do this coarse–graining explicitly to see which (relevant) elements of the model survive to influence macroscopic dynamics and which (irrelevant) details are erased. As far as I know nobody knows how to start with a realistic description of ion channels and coarse–grain to arrive at a model of neural networks. Further, we do know that some molecular details must matter, because particular classes of cells in the brain will switch from using one kind of channel to another at crucial moments in development, and different types of cells use different combinations of channels. We also know that not all macroscopic quantities are universal, and there is the danger that what is universal might not be relevant for the organism.

My concern about the gap between molecular and system level descriptions is not just the vague feeling that we are missing something. In many cases, including the fly embryo, some of the most powerful new experimental tools allow direct and reliable manipulation at the molecular level. The parameters of our network description, attached to the arrows in Fig 27, are themselves encoded (in part) in the DNA sequence of transcription factor binding sites. It now is possible to edit these sequences with single base pair precision, but we don't know how these manipulations relate to the circles and arrows, so it is hard to see how our current theories connect to an exploding set of new experiments. In truth we don't even know if we can start with sequences and do a systematic calculation to arrive at something like Eqs (139, 144, 145). A more complete physics of life will build these bridges.

Acknowledgments

Teaching in summer schools is a special pleasure, a stimulus to putting one's thoughts in order. My thanks to the organizers—Anne–Florence Bitbol, Thierry Mora, Ilya Nemenman, and Aleksandra Walczak—for this opportunity, and for years of friendship. Thanks also to the students, who provided a warm and enthusiastic welcome. The ideas presented here owe much to many wonderful collaborators and friends, as may be seen from the references. But this is a school, and acknowledging these debts also is an opportunity to provide advice to the students.

The idea that living systems approach fundamental physical limits to their performance has fascinated me from the first time I heard about photon counting in vision. On arriving in Groningen as a postdoctoral fellow forty years ago (!), I had the good fortune to meet Rob de Ruyter van Steveninck, who was in the office next door. The experiments he was doing on the fly visual system, combining measurements in the retina with measurements on motion–sensitive neurons deep in the brain, made it seem possible to connect general theoretical ideas about physical limits and optimization with detailed, quantitative experiments. We could not have predicted that those early discussions would turn into decades of continuing collaboration and friendship. We also have had the pleasure of being joined by many colleagues. Lesson for theorists: talk to the experimentalist next door, it literally can change your life.

When I moved to Princeton I wanted to take the ideas of signals, noise, and information that Rob and I had explored in neurons down to the molecular level; Sima Setayeshgar and I made some first theoretical efforts in this direction. David Tank had moved at the same time, and was excited that new fluorescence techniques might make it possible to measure signals on this scale with enough precision to say something meaningful. Eric Wieschaus joined our conversation, and while he was unclear on exactly what David and I wanted to do (we were

unclear as well), he explained why we should do it in flies. Eric ended by saying that if some physics student wanted to try, we should send the student to him to learn the relevant techniques. Thomas Gregor was working with me on a theory project, but decided to take Eric up on his offer, and this led to spectacular new experiments. In parallel, Curt Callan and I were working with another student, Gašper Tkačik, on purely theoretical questions about the optimization of information flow in transcriptional control. Luckily, Gašper and Thomas sat in the same office, and they made many of the crucial theory/experiment connections on their own. Again, we could not have predicted that these early discussions would form the basis for two decades of work, for collaborations with generations of new colleagues, and for valued friendships. Lesson for theorists: be ready for experimental developments which allow speculative theoretical ideas to become concrete.

Lesson from both stories: one experiment can test a theory or raise a specific theoretical question, but sometimes theory/experiment collaboration is a long term investment.

In addition to the handful of friends who were at the origins of these ideas, the specific things I discussed in these lectures involved collaborations with Naama Brenner, Julien Dubuis, Adrienne Fairhall, Dmitry Krotov, Geoff Lewen, Geoff Owen, Mariela Petkova, Marc Potters, Fred Rieke, Shiva Sinha, Thomas Sokolowski, and Aleksandra Walczak. This bland listing doesn't do justice to all the fun we have had together.

Funding information This work was supported in part by the US National Science Foundation through the Center for the Physics of Biological Function (PHY–1734030), and by Fellowships from the Simons Foundation and the John Simon Guggenheim Memorial Foundation.

References

[1] E. R. Harrison, *Darkness at night: A riddle of the Universe*, Harvard University Press, Cambridge, USA, ISBN 9780674192713 (1987).

[2] P. J. E. Peebles, *Principles of physical cosmology*, Princeton University Press, Princeton, USA, ISBN 9780691209814 (1993).

[3] P. W. Anderson, *Basic notions of condensed matter physics*, Benjamin/Cummings, Menlo Park, USA, ISBN 9780201328301 (1984).

[4] N. W. Timoféff-Ressovsky, K. G. Zimmer and M. Delbrück, *Über die Natur der Genmutation und der Genstruktur*, Nachr. Ges. Wiss. Gött. **1**, 189 (1935).

[5] P . R. Sloan and B. Fogel (eds.), *Creating a physical biology. The three-man paper and early molecular biology*, University of Chicago Press, Chicago, USA, ISBN 9780226767833 (2011).

[6] E. Schrödinger, *What is life?*, Cambridge University Press, Cambridge, UK (1944).

[7] J. D. Watson, T. A. Baker, S. P. Bell, A. Gann, M. Levine and R. Losick, *Molecular biology of the gene, 7th edition*, Pearson, Boston, USA, ISBN 9780321762436 (2014).

[8] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular biology of the cell, 4th edition*, Garland Science, New York, USA, ISBN 9780815332183 (2002).

[9] C. Nüsslein-Volhard and E. Wieschaus, *Mutations affecting segment number and polarity in Drosophila*, Nature **287**, 795 (1980), doi:10.1038/287795a0.

[10] M. D. Petkova, G. Tkačik, W. Bialek, E. F. Wieschaus and T. Gregor, *Optimal decoding of cellular identities in a genetic network*, Cell **176**, 844 (2019), doi:10.1016/j.cell.2019.01.007.

[11] E. Wieschaus and C. Nüsslein-Volhard, *The Heidelberg screen for pattern mutants of Drosophila: A personal account*, Annu. Rev. Cell Dev. Biol. **32**, 1 (2016), doi:10.1146/annurev-cellbio-113015-023138.

[12] P. A. Lawrence, *The making of a fly: The genetics of animal design*, Blackwell, Oxford, UK, ISBN 9780632030484 (1992).

[13] R. N. Cahn, *The eighteen arbitrary parameters of the Standard Model in your everyday life*, Rev. Mod. Phys. **68**, 951 (1996), doi:10.1103/RevModPhys.68.951.

[14] J. Bardeen, L. N. Cooper and J. R. Schrieffer, *Theory of superconductivity*, Phys. Rev. **108**, 1175 (1957), doi:10.1103/PhysRev.108.1175.

[15] K. G. Wilson, *The renormalization group: Critical phenomena and the Kondo problem*, Rev. Mod. Phys. **47**, 773 (1975), doi:10.1103/RevModPhys.47.773.

[16] G. Parisi, *Statistical field theory*, Addison-Wesley, Redwood City, USA, ISBN 9780201059854 (1988).

[17] J. Zinn-Justin, *Quantum field theory and critical phenomena*, Oxford University Press, Oxford, UK, ISBN 9780198509233 (1989), doi:10.1093/acprof:oso/9780198509233.001.0001.

[18] J. Cardy, *Scaling and renormalization in statistical physics*, Cambridge University Press, Cambridge, UK, ISBN 9780521499590 (1996), doi:10.1017/CBO9781316036440.

[19] R. B. Laughlin, *Anomalous quantum Hall effect: An incompressible quantum fluid with fractionally charged excitations*, Phys. Rev. Lett. **50**, 1395 (1983), doi:10.1103/PhysRevLett.50.1395.

[20] D. J. Gross and F. Wilczek, *Ultraviolet behavior of non-Abelian gauge theories*, Phys. Rev. Lett. **30**, 1343 (1973), doi:10.1103/PhysRevLett.30.1343.

[21] H. D. Politzer, *Reliable perturbative results for strong interactions?*, Phys. Rev. Lett. **30**, 1346 (1973), doi:10.1103/PhysRevLett.30.1346.

[22] C. G. Callan and D. J. Gross, *Bjorken scaling in quantum field theory*, Phys. Rev. D **8**, 4383 (1973), doi:10.1103/PhysRevD.8.4383.

[23] P. Newman, *Deep inelastic lepton-nucleon scattering at HERA*, Int. J. Mod. Phys. A **19**, 1061 (2004), doi:10.1142/S0217751X0401897X.

[24] Y. LeCun, Y. Bengio and G. Hinton, *Deep learning*, Nature **521**, 436 (2015), doi:10.1038/nature14539.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Attention is all you need*, in *Advances in neural information processing systems 31*, Curran Associates, Red Hook, USA, ISBN 9781510860964 (2017).

[26] T. Brown et al., *Language models are few-shot learners*, in *Advances in neural information processing systems 33*, Curran Associates, Red Hook, USA, ISBN 9781713829546 (2020).

[27] H. D. Block, *The perceptron: A model for brain functioning. I*, Rev. Mod. Phys. **34**, 123 (1962), doi:10.1103/RevModPhys.34.123.

[28] H. D. Block, B. W. Knight and F. Rosenblatt, *Analysis of a four-layer series-coupled perceptron. II*, Rev. Mod. Phys. **34**, 135 (1962), doi:10.1103/RevModPhys.34.135.

[29] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci. **79**, 2554 (1982), doi:10.1073/pnas.79.8.2554.

[30] D. J. Amit, *Modeling brain function*, Cambridge University Press, Cambridge, UK, ISBN 9780521361002 (1989), doi:10.1017/CBO9780511623257.

[31] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the theory of neural computation*, CRC Press, Boca Raton, USA, ISBN 9780201515602 (1991).

[32] P. Mehta and D. J. Schwab, *An exact mapping between the variational renormalization group and deep learning*, (arXiv preprint) doi:10.48550/arXiv.1410.3831.

[33] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher and D. J. Schwab, *A high-bias, low-variance introduction to machine learning for physicists*, Phys. Rep. **810**, 1 (2019), doi:10.1016/j.physrep.2019.03.001.

[34] G. Carleo, J. I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto and L. Zdeborová, *Machine learning and the physical sciences*, Rev. Mod. Phys. **91**, 045002 (2019), doi:10.1103/RevModPhys.91.045002.

[35] J. Kaplan et al., *Scaling laws for neural language models*, (arXiv preprint) doi:10.48550/arXiv.2001.08361.

[36] D. A. Roberts, S. Yaida and B. Hanin, *The principles of deep learning theory*, Cambridge University Press, Cambridge, UK, ISBN 9781009023405 (2022), doi:10.1017/9781009023405.

[37] L. Peliti (ed.), *Biologically inspired physics*, Springer, Boston, USA, ISBN 9781475794854 (1991), doi:10.1007/978-1-4757-9483-0.

[38] D. Nelson, T. Piran and S. Weinberg (eds.), *Statistical mechanics of membranes and surfaces*, World Scientific, Singapore, ISBN 9789813201491 (1989), doi:10.1142/0706.

[39] S. Ramaswamy, *Active matter*, J. Stat. Mech.: Theory Exp. 054002 (2017), doi:10.1088/1742-5468/aa6bc5.

[40] M. M. Desai, D. S. Fisher and A. W. Murray, *The speed of evolution and maintenance of variation in asexual populations*, Curr. Biol. **17**, 385 (2007), doi:10.1016/j.cub.2007.01.072.

[41] D. S. Fisher, *Leading the dog of selection by its mutational nose*, Proc. Natl. Acad. Sci. **108**, 2633 (2011), doi:10.1073/pnas.1100339108.

[42] R. A. Neher and O. Hallatschek, *Genealogies of rapidly adapting populations*, Proc. Natl. Acad. Sci. **110**, 437 (2012), doi:10.1073/pnas.1213113110.

[43] M. M. Desai, A. M. Walczak and D. S. Fisher, *Genetic diversity and the structure of genealogies in rapidly adapting populations*, Genetics **193**, 565 (2013), doi:10.1534/genetics.112.147157.

[44] D. S. Fisher, *Asexual evolution waves: Fluctuations and universality*, J. Stat. Mech.: Theory Exp. P01011 (2013), doi:10.1088/1742-5468/2013/01/P01011.

[45] S. F. Levy, J. R. Blundell, S. Venkataram, D. A. Petrov, D. S. Fisher and G. Sherlock, *Quantitative evolutionary dynamics using high-resolution lineage tracking*, Nature **519**, 181 (2015), doi:10.1038/nature14279.

[46] B. H. Good, M. J. McDonald, J. E. Barrick, R. E. Lenski and M. M. Desai, *The dynamics of molecular evolution over 60,000 generations*, Nature **551**, 45 (2017), doi:10.1038/nature24287.

[47] A. N. Nguyen Ba, I. Cvijović, J. I. Rojas Echenique, K. R. Lawrence, A. Rego-Costa, X. Liu, S. F. Levy and M. M. Desai, *High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast*, Nature **575**, 494 (2019), doi:10.1038/s41586-019-1749-3.

[48] M. Łuksza and M. Lässig, *A predictive fitness model for influenza*, Nature **507**, 57 (2014), doi:10.1038/nature13087.

[49] R. A. Neher, T. Bedford, R. S. Daniels, C. A. Russell and B. I. Shraiman, *Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses*, Proc. Natl. Acad. Sci. **113**, E1701 (2016), doi:10.1073/pnas.1525578113.

[50] F. Zanini, J. Brodin, L. Thebo, C. Lanz, G. Bratt, J. Albert and R. A. Neher, *Population genomics of intrapatient HIV-1 evolution*, eLife **4**, e11282 (2016), doi:10.7554/eLife.11282.001.

[51] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers and J. P. Sethna, *Perspective: Sloppiness and emergent theories in physics, biology, and beyond*, J. Chem. Phys. **143**, 010901 (2015), doi:10.1063/1.4923066.

[52] K. S. Brown, C. C. Hill, G. A. Calero, C. R. Myers, K. H. Lee, J. P. Sethna and R. A. Cerione, *The statistical mechanics of complex signaling networks: Nerve growth factor signaling*, Phys. Biol. **1**, 184 (2004), doi:10.1088/1478-3967/1/3/006.

[53] D. A. Rand, A. Raju, M. Sáez, F. Corson and E. D. Siggia, *Geometry of gene regulatory dynamics*, Proc. Natl. Acad. Sci. **118**, e2109729118 (2021), doi:10.1073/pnas.2109729118.

[54] S. Hecht, S. Shlaer and M. H. Pirenne, *Energy, quanta, and vision*, J. Gen. Physiol. **25**, 819 (1942), doi:10.1085/jgp.25.6.819.

[55] H. B. Barlow, *The size of ommatidia in apposition eyes*, J. Exp. Biol. **29**, 667 (1952), doi:10.1242/jeb.29.4.667.

[56] H. C. Berg and E. M. Purcell, *Physics of chemoreception*, Biophys. J. **20**, 193 (1977), doi:10.1016/S0006-3495(77)85544-6.

[57] W. Bialek, *Biophysics: Searching for principles*, Princeton University Press, Princeton, USA, ISBN 9780691138916 (2012).

[58] M. Nikolić, V. Antonetti, F. Liu, G. Muhaxheri, M. D. Petkova, M. Scheeler, E. M. Smith, W. Bialek and T. Gregor, *Scale invariance in early embryonic development*, (arXiv preprint) doi:10.48550/arXiv.2312.17684.

[59] O. Shimomura, F. H. Johnson and Y. Saiga, *Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, aequorea*, J. Cell. Comp. Physiol. **59**, 223 (1962), doi:10.1002/jcp.1030590302.

[60] F. H. Johnson, O. Shimomura, Y. Saiga, L. C. Gershman, G. T. Reynolds and J. R. Waters, *Quantum efficiency of cypridina luminescence, with a note on that of aequorea*, J. Cell. Comp. Physiol. **60**, 85 (1962), doi:10.1002/jcp.1030600111.

[61] D. C. Prasher, V. K. Eckenrode, W. W. Ward, F. G. Prendergast and M. J. Cormier, *Primary structure of the aequorea victoria green-fluorescent protein*, Gene **111**, 229 (1992), doi:10.1016/0378-1119(92)90691-H.

[62] M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward and D. C. Prasher, *Green fluorescent protein as a marker for gene expression*, Science **263**, 802 (1994), doi:10.1126/science.8303295.

[63] R. Y. Tsien, *Constructing and exploiting the fluorescent protein paintbox (Nobel lecture)*, Angew. Chem. Int. Ed. **48**, 5612 (2009), doi:10.1002/anie.200901916.

[64] T. Gregor, E. F. Wieschaus, A. P. McGregor, W. Bialek and D. W. Tank, *Stability and nuclear dynamics of the bicoid morphogen gradient*, Cell **130**, 141 (2007), doi:10.1016/j.cell.2007.05.026.

[65] J. O. Dubuis, R. Samanta and T. Gregor, *Accurate measurements of dynamics and reproducibility in small genetic networks*, Mol. Syst. Biol. **9**, 639 (2013), doi:10.1038/msb.2012.72.

[66] S. C. Little and T. Gregor, *Single mRNA molecule detection in Drosophila*, in *RNA detection: Methods in molecular biology*, Springer, New York, USA, ISBN 9781493972128 (2018), doi:10.1007/978-1-4939-7213-5_8.

[67] E. Lubeck and L. Cai, *Single-cell systems biology by super-resolution imaging and combinatorial labeling*, Nat. Methods. **9**, 743 (2012), doi:10.1038/nmeth.2069.

[68] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang and X. Zhuang, *Spatially resolved, highly multiplexed RNA profiling in single cells*, Science **348**, aaa6090 (2015), doi:10.1126/science.aaa6090.

[69] B. Zoller, S. C. Little and T. Gregor, *Diverse spatial expression patterns emerge from unified kinetics of transcriptional bursting*, Cell **175**, 835 (2018), doi:10.1016/j.cell.2018.09.056.

[70] D. R. Larson, D. Zenklusen, B. Wu, J. A. Chao and R. H. Singer, *Real-time observation of transcription initiation and elongation on an endogenous yeast gene*, Science **332**, 475 (2011), doi:10.1126/science.1202142.

[71] T. Lucas, T. Ferraro, B. Roelens, J. De Las Heras Chanes, A. M. Walczak, M. Coppey and N. Dostatni, *Live imaging of bicoid-dependent transcription in Drosophila embryos*, Curr. Biol. **23**, 2135 (2013), doi:10.1016/j.cub.2013.08.053.

[72] H. G. Garcia, M. Tikhonov, A. Lin and T. Gregor, *Quantitative imaging of transcription in living Drosophila embryos links polymerase activity to patterning*, Curr. Biol. **23**, 2140 (2013), doi:10.1016/j.cub.2013.08.054.

[73] P.-T. Chen, M. Levo, B. Zoller and T. Gregor, *Gene activity fully predicts transcriptional bursting dynamics*, (arXiv preprint) doi:10.48550/arXiv.2304.08770.

[74] H. Chen, M. Levo, L. Barinov, M. Fujioka, J. B. Jaynes and T. Gregor, *Dynamic interplay between enhancer-promoter topology and gene activity*, Nat. Genet. **50**, 1296 (2018), doi:10.1038/s41588-018-0175-z.

[75] L. Barinov, S. Ryabichko, W. Bialek and T. Gregor, *Transcription-dependent spatial organization of a gene locus*, (arXiv preprint) doi:10.48550/arXiv.2012.15819.

[76] E. D. Adrian, *The basis of sensation: The action of the sense organs*, W. W. Norton, New York, USA (1928).

[77] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol. **117**, 500 (1952), doi:10.1113/jphysiol.1952.sp004764.

[78] D. J. Aidley, *The physiology of excitable cells*, Cambridge University Press, Cambridge, UK, ISBN 9780521574211 (1998), doi:10.1017/CBO9781139171182.

[79] F. Rieke, D. Warland, R. de Ruyter van Steveninck and W. Bialek, *Spikes: Exploring the neural code*, MIT Press, Cambridge, USA, ISBN 9780262181747 (1997).

[80] R. Segev, J. Goodhouse, J. Puchalla and M. J. Berry, *Recording spikes from a large fraction of the ganglion cells in a retinal patch*, Nat. Neurosci. **7**, 1155 (2004), doi:10.1038/nn1323.

[81] A. M. Litke et al., *What does the eye tell the brain?: Development of a system for the large-scale recording of retinal output activity*, IEEE Trans. Nucl. Sci. **51**, 1434 (2004), doi:10.1109/TNS.2004.832706.

[82] O. Marre, D. Amodei, N. Deshmukh, K. Sadeghi, F. Soo, T. E. Holy and M. J. Berry, *Mapping a complete neural population in the retina*, J. Neurosci. **32**, 14859 (2012), doi:10.1523/JNEUROSCI.0723-12.2012.

[83] P. K. Campbell, K. E. Jones, R. J. Huber, K. W. Horch and R. A. Normann, *A silicon-based, three-dimensional neural interface: Manufacturing processes for an intracortical electrode array*, IEEE Trans. Biomed. Eng. **38**, 758 (1991), doi:10.1109/10.83588.

[84] J. J. Jun et al., *Fully integrated silicon probes for high-density recording of neural activity*, Nature **551**, 232 (2017), doi:10.1038/nature24636.

[85] J. E. Chung et al., *High-density, long-lasting, and multi-region electrophysiological recordings using polymer electrode arrays*, Neuron **101**, 21 (2019), doi:10.1016/j.neuron.2018.11.002.

[86] Y. Gong, *The evolving capabilities of rhodopsin-based genetically encoded voltage indicators*, Curr. Opin. Chem. Biol. **27**, 84 (2015), doi:10.1016/j.cbpa.2015.05.006.

[87] H. H. Yang and F. St-Pierre, *Genetically encoded voltage indicators: Opportunities and challenges*, J. Neurosci. **36**, 9977 (2016), doi:10.1523/JNEUROSCI.1095-16.2016.

[88] J. Platisa, H. Zeng, L. Madisen, L. B. Cohen, V. A. Pieribone and D. A. Storace, *Voltage imaging in the olfactory bulb using transgenic mouse lines expressing the genetically encoded voltage indicator ArcLight*, Sci. Rep. **12**, 1875 (2022), doi:10.1038/s41598-021-04482-3.

[89] S. Wenceslao Evans et al., *A positively tuned voltage indicator for extended electrical recordings in the brain*, Nat. Methods **20**, 1104 (2023), doi:10.1038/s41592-023-01913-z.

[90] L. Tian et al., *Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators*, Nat. Methods **6**, 875 (2009), doi:10.1038/nmeth.1398.

[91] L. Tian, S. A. Hires and L. L. Looger, *Imaging neuronal activity with genetically encoded calcium indicators*, in *Imaging in neuroscience*, Cold Spring Harbor Protocols Press, Cold Spring Harbor, USA, ISBN 9780879699383 (2012), doi:10.1101/pdb.top069609.

[92] Y. Zhang et al., *Fast and sensitive GCaMP calcium indicators for imaging neural populations*, Nature **615**, 884 (2023), doi:10.1038/s41586-023-05828-9.

[93] W. Denk, J. H. Strickler and W. W. Webb, *Two-photon laser scanning fluorescence microscopy*, Science **248**, 73 (1990), doi:10.1126/science.2321027.

[94] D. A. Dombeck, A. N. Khabbaz, F. Collman, T. L. Adelman and D. W. Tank, *Imaging large-scale neural activity with cellular resolution in awake, mobile mice*, Neuron **56**, 43 (2007), doi:10.1016/j.neuron.2007.08.003.

[95] A. Song, A. S. Charles, S. A. Koay, J. L. Gauthier, S. Y. Thiberge, J. W. Pillow and D. W. Tank, *Volumetric two-photon imaging of neurons using stereoscopy (vTwINS)*, Nat. Methods **14**, 420 (2017), doi:10.1038/nmeth.4226.

[96] S. Weisenburger et al., *Volumetric Ca^{2+} imaging in the mouse brain using hybrid multiplexed sculpted light microscopy*, Cell **177**, 1050 (2019), doi:10.1016/j.cell.2019.03.011.

[97] J. Demas, J. Manley, F. Tejera, K. Barber, H. Kim, F. Martínez Traub, B. Chen and A. Vaziri, *High-speed, cortex-wide volumetric recording of neuroactivity at cellular resolution using light beads microscopy*, Nat. Methods **18**, 1103 (2021), doi:10.1038/s41592-021-01239-8.

[98] C. D. Harvey, F. Collman, D. A. Dombeck and D. W. Tank, *Intracellular dynamics of hippocampal place cells during virtual navigation*, Nature **461**, 941 (2009), doi:10.1038/nature08499.

[99] M. B. Ahrens, M. B. Orger, D. N. Robson, J. M. Li and P. J. Keller, *Whole-brain functional imaging at cellular resolution using light-sheet microscopy*, Nat. Methods **10**, 413 (2013), doi:10.1038/nmeth.2434.

[100] J. P. Nguyen, F. B. Shipley, A. N. Linder, G. S. Plummer, M. Liu, S. U. Setru, J. W. Shaevitz and A. M. Leifer, *Whole-brain calcium imaging with cellular resolution in freely behaving caenorhabditis elegans*, Proc. Natl. Acad. Sci. **113**, E1074 (2015), doi:10.1073/pnas.1507110112.

[101] G. Nagel, T. Szellas, W. Huhn, S. Kateriya, N. Adeishvili, P. Berthold, D. Ollig, P. Hegemann and E. Bamberg, *Channelrhodopsin-2, a directly light-gated cation-selective membrane channel*, Proc. Natl. Acad. Sci. **100**, 13940 (2003), doi:10.1073/pnas.1936192100.

[102] X. Han and E. S. Boyden, *Multiple-color optical activation, silencing, and desynchronization of neural activity, with single-spike temporal resolution*, PLoS ONE **2**, e299 (2007), doi:10.1371/journal.pone.0000299.

[103] F. Zhang, M. Prigge, F. Beyrière, S. P. Tsunoda, J. Mattis, O. Yizhar, P. Hegemann and K. Deisseroth, *Red-shifted optogenetic excitation: A tool for fast neural control derived from Volvox carteri*, Nat. Neurosci. **11**, 631 (2008), doi:10.1038/nn.2120.

[104] F. Randi, A. K. Sharma, S. Dvali and A. M. Leifer, *Neural signal propagation atlas of caenorhabditis elegans*, Nature **623**, 406 (2023), doi:10.1038/s41586-023-06683-4.

[105] L. Meshulam, J. L. Gauthier, C. D. Brody, D. W. Tank and W. Bialek, *Coarse graining, fixed points, and scaling in a large population of neurons*, Phys. Rev. Lett. **123**, 178103 (2019), doi:10.1103/PhysRevLett.123.178103.

[106] L. Meshulam, J. L. Gauthier, C. D. Brody, D. W. Tank and W. Bialek, *Collective behavior of place and non-place neurons in the hippocampal network*, Neuron **96**, 1178 (2017), doi:10.1016/j.neuron.2017.10.027.

[107] K. Svoboda, C. F. Schmidt, B. J. Schnapp and S. M. Block, *Direct observation of kinesin stepping by optical trapping interferometry*, Nature **365**, 721 (1993), doi:10.1038/365721a0.

[108] E. A. Abbondanzieri, W. J. Greenleaf, J. W. Shaevitz, R. Landick and S. M. Block, *Direct observation of base-pair stepping by RNA polymerase*, Nature **438**, 460 (2005), doi:10.1038/nature04268.

[109] H. Ueno, T. Suzuki, K. Kinosita and M. Yoshida, *ATP-driven stepwise rotation of F_0F_1-ATP synthase*, Proc. Natl. Acad. Sci. **102**, 1333 (2005), doi:10.1073/pnas.0407857102.

[110] A. Cavagna, D. Conti, C. Creato, L. Del Castello, I. Giardina, T. S. Grigera, S. Melillo, L. Parisi and M. Viale, *Dynamic scaling in natural swarms*, Nat. Phys. **13**, 914 (2017), doi:10.1038/nphys4153.

[111] A. Cavagna, I. Giardina and T. S. Grigera, *The physics of flocking: Correlation as a compass from experiments to theory*, Phys. Rep. **728**, 1 (2018), doi:10.1016/j.physrep.2017.11.003.

[112] B. Qin, C. Fei, A. A. Bridges, A. A. Mashruwala, H. A. Stone, N. S. Wingreen and B. L. Bassler, *Cell position fates and collective fountain flow in bacterial biofilms revealed by light-sheet microscopy*, Science **369**, 71 (2020), doi:10.1126/science.abb8501.

[113] K. Copenhagen, R. Alert, N. S. Wingreen and J. W. Shaevitz, *Topological defects promote layer formation in Myxococcus xanthus colonies*, Nat. Phys. **17**, 211 (2020), doi:10.1038/s41567-020-01056-4.

[114] A. Warmflash, B. Sorre, F. Etoc, E. D. Siggia and A. H. Brivanlou, *A method to recapitulate early embryonic spatial patterning in human embryonic stem cells*, Nat. Methods **11**, 847 (2014), doi:10.1038/nmeth.3016.

[115] M. N. Shahbazi, E. D. Siggia and M. Zernicka-Goetz, *Self-organization of stem cells into embryos: A window on early mammalian development*, Science **364**, 948 (2019), doi:10.1126/science.aax0164.

[116] W. Bialek et al., *Physics of life*, National Academies Press, Washington DC, USA, ISBN 9780309274005 (2022), doi:10.17226/26403.

[117] J. Shlens, *A tutorial on principal component analysis*, (arXiv preprint) doi:10.48550/arXiv.1404.1100.

[118] S. T. Roweis and L. K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science **290**, 2323 (2000), doi:10.1126/science.290.5500.2323.

[119] J. P. Gergen, D. Coulter and E. F. Wieschaus, *Segmental pattern and blastoderm cell identities*, in J. G. Gall, *Gametogenesis and the early embryo*, Liss, New York, USA, ISBN 9780845115053 (1986).

[120] J. O. Dubuis, G. Tkačik, E. F. Wieschaus, T. Gregor and W. Bialek, *Positional information, in bits*, Proc. Natl. Acad. Sci. **110**, 16301 (2013), doi:10.1073/pnas.1315642110.

[121] L. McGough, H. Casademunt, M. Nikolić, Z. Aridor, M. D. Petkova, T. Gregor and W. Bialek, *Finding the last bits of positional information*, PRX Life **2**, 013016 (2024), doi:10.1103/PRXLife.2.013016.

[122] M. Ptashne, *A genetic switch: Phage and higher organisms*, Blackwell, Cambridge, USA, ISBN 9780865422094 (1992).

[123] P. V. Pedone, R. Ghirlando, G. M. Clore, A. M. Gronenborn, G. Felsenfeld and J. G. Omichinski, *The single Cys2-His2 zinc finger domain of the GAGA protein flanked by basic residues is sufficient for high-affinity specific DNA binding.*, Proc. Natl. Acad. Sci. **93**, 2822 (1996), doi:10.1073/pnas.93.7.2822.

[124] R. L. Winston, D. P. Millar, J. M. Gottesfeld and S. B. H. Kent, *Characterization of the DNA binding properties of the bHLH domain of deadpan to single and tandem sites*, Biochemistry **38**, 5138 (1999), doi:10.1021/bi982856a.

[125] T. Gregor, D. W. Tank, E. F. Wieschaus and W. Bialek, *Probing the limits to positional information*, Cell **130**, 153 (2007), doi:10.1016/j.cell.2007.05.025.

[126] W. Bialek, *Thinking about the brain*, in *Physics of bio-molecules and cells. Physique des biomolécules et des cellules*, Springer, Berlin, Germany, ISBN 9783540441328 (2002), doi:10.1007/3-540-45701-1_12.

[127] G. D. Field and F. Rieke, *Nonlinear signal transfer from mouse rods to bipolar cells and implications for visual sensitivity*, Neuron **34**, 773 (2002), doi:10.1016/S0896-6273(02)00700-6.

[128] F. Rieke, W. G. Owen and W. Bialek, *Optimal filtering in the salamander retina*, in *Advances in neural information processing systems 3*, Morgan Kaufmann Publishers, Cambridge, USA, ISBN 9781558601840 (1991).

[129] B. Hassenstein and W. Reichardt, *Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers Chlorophanus*, Z. Naturforsch. B **11**, 513 (1956), doi:10.1515/znb-1956-9-1004.

[130] W. Reichardt and T. Poggio, *Visual control of orientation behaviour in the fly: Part I. A quantitative analysis*, Quart. Rev. Biophys. **9**, 311 (1976), doi:10.1017/S0033583500002523.

[131] E. H. Adelson and J. R. Bergen, *Spatiotemporal energy models for the perception of motion*, J. Opt. Soc. Am. A **2**, 284 (1985), doi:10.1364/JOSAA.2.000284.

[132] J. P. H. van Santen and G. Sperling, *Elaborated Reichardt detectors*, J. Opt. Soc. Am. A **2**, 300 (1985), doi:10.1364/JOSAA.2.000300.

[133] N. J. Strausfeld, *Atlas of an insect brain*, Springer, Berlin, Heidelberg, Germany, ISBN 9783540073437 (1976).

[134] K. Hausen, *Motion sensitive interneurons in the optomotor system of the fly*, Biol. Cybern. **45**, 143 (1982), doi:10.1007/BF00335241.

[135] D. G. Stavenga and R. C. Hardie, *Facets of vision*, Springer, Berlin, Germany, ISBN 9783642740848 (1989), doi:10.1007/978-3-642-74082-4.

[136] R. R. de Ruyter van Steveninck and S. B. Laughlin, *The rate of information transfer at graded-potential synapses*, Nature **379**, 642 (1996), doi:10.1038/379642a0.

[137] R. R. de Ruyter van Steveninck and S. B. Laughlin, *Light adaptation and reliability in blowfly photoreceptors*, Int. J. Neur. Syst. **07**, 437 (1996), doi:10.1142/S0129065796000415.

[138] W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck and D. Warland, *Reading a neural code*, Science **252**, 1854 (1991), doi:10.1126/science.2063199.

[139] R. de Ruyter van Steveninck and W. Bialek, *Reliability and statistical efficiency of a blowfly movement-sensitive neuron*, Philos. Trans. R. Soc. Lond. B: Biol. Sci. **348**, 321 (1995), doi:10.1098/rstb.1995.0071.

[140] M. Potters and W. Bialek, *Statistical mechanics and visual signal processing*, J. Phys. I France **4**, 1755 (1994), doi:10.1051/jp1:1994219.

[141] S. R. Sinha, W. Bialek and R. R. de Ruyter van Steveninck, *Optimal local estimates of visual motion in a natural environment*, Phys. Rev. Lett. **126**, 018101 (2021), doi:10.1103/PhysRevLett.126.018101.

[142] J. E. Fitzgerald, A. Y. Katsov, T. R. Clandinin and M. J. Schnitzer, *Symmetries in stimulus statistics shape the form of visual motion estimators*, Proc. Natl. Acad. Sci. **108**, 12909 (2011), doi:10.1073/pnas.1015680108.

[143] R. Behnia, D. A. Clark, A. G. Carter, T. R. Clandinin and C. Desplan, *Processing properties of ON and OFF pathways for Drosophila motion detection*, Nature **512**, 427 (2014), doi:10.1038/nature13427.

[144] L. Wolpert, *Positional information and the spatial pattern of cellular differentiation*, J. Theor. Biol. **25**, 1 (1969), doi:10.1016/S0022-5193(69)80016-0.

[145] D. Krotov, J. O. Dubuis, T. Gregor and W. Bialek, *Morphogenesis at criticality*, Proc. Natl. Acad. Sci. **111**, 3683 (2014), doi:10.1073/pnas.1324186111.

[146] T. Tohme and W. Bialek, *A brief tutorial on information theory*, (arXiv preprint) doi:10.48550/arXiv.2402.16556.

[147] C. E. Shannon, *A mathematical theory of communication*, Bell Sys. Tech. J. **27**, 379 (1948), doi:10.1002/j.1538-7305.1948.tb01338.x.

[148] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley, Now York, USA, ISBN 9780471241959 (2005), doi:10.1002/047174882X.

[149] M. Mézard and A. Montanari, *Information, physics, and computation*, Oxford University Press, Oxford, UK, ISBN 9780198570837 (2009), doi:10.1093/acprof:oso/9780198570837.001.0001.

[150] S. Laughlin, *A simple coding procedure enhances a neuron's information capacity*, Z. Naturforsch. C **36**, 910 (1981), doi:10.1515/znc-1981-9-1040.

[151] S. M. Smirnakis, M. J. Berry, D. K. Warland, W. Bialek and M. Meister, *Adaptation of retinal processing to image contrast and spatial scale*, Nature **386**, 69 (1997), doi:10.1038/386069a0.

[152] L. F. Abbott and P. Dayan, *Theoretical neuroscience. Computational and mathematical modeling of neural systems*, MIT Press, Cambridge, USA, ISBN 9780262041997 (2001).

[153] E. De Boer and P. Kuyper, *Triggered correlation*, IEEE Trans. Biomed. Eng. **BME-15**, 169 (1968), doi:10.1109/TBME.1968.4502561.

[154] R. de Ruyter van Steveninck and W. Bialek, *Real-time performance of a movement-sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences*, Proc. R. Soc. Lond. B. **234**, 379 (1988), doi:10.1098/rspb.1988.0055.

[155] N. Brenner, W. Bialek and R. de Ruyter van Steveninck, *Adaptive rescaling maximizes information transmission*, Neuron **26**, 695 (2000), doi:10.1016/S0896-6273(00)81205-2.

[156] W. Bialek and R. de Ruyter van Steveninck, *Features and dimensions: Motion estimation in fly vision*, (arXiv preprint) doi:10.48550/arXiv.q-bio/0505003.

[157] N. C. Rust, O. Schwartz, J. A. Movshon and E. P. Simoncelli, *Spatiotemporal elements of macaque V1 receptive fields*, Neuron **46**, 945 (2005), doi:10.1016/j.neuron.2005.05.021.

[158] A. L. Fairhall, G. D. Lewen, W. Bialek and R. R. de Ruyter van Steveninck, *Efficiency and ambiguity in an adaptive neural code*, Nature **412**, 787 (2001), doi:10.1038/35090500.

[159] K. I. Nagel and A. J. Doupe, *Temporal processing and adaptation in the songbird auditory forebrain*, Neuron **51**, 845 (2006), doi:10.1016/j.neuron.2006.08.030.

[160] M. Maravall, R. S. Petersen, A. L. Fairhall, E. Arabzadeh and M. E. Diamond, *Shifts in coding properties and maintenance of information transmission during adaptation in barrel cortex*, PLoS Biol **5**, e19 (2007), doi:10.1371/journal.pbio.0050019.

[161] I. Dean, N. S. Harper and D. McAlpine, *Neural population coding of sound level adapts to stimulus statistics*, Nat. Neurosci. **8**, 1684 (2005), doi:10.1038/nn1541.

[162] B. Wark, B. N. Lundstrom and A. Fairhall, *Sensory adaptation*, Curr. Opin. Neurobiol. **17**, 423 (2007), doi:10.1016/j.conb.2007.07.001.

[163] K. J. Kim and F. Rieke, *Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells*, J. Neurosci. **21**, 287 (2001), doi:10.1523/JNEUROSCI.21-01-00287.2001.

[164] K. J. Kim and F. Rieke, *Slow Na^+ inactivation and variance adaptation in salamander retinal ganglion cells*, J. Neurosci. **23**, 1506 (2003), doi:10.1523/JNEUROSCI.23-04-01506.2003.

[165] G. Tkačik, J. O. Dubuis, M. D. Petkova and T. Gregor, *Positional information, positional error, and readout precision in morphogenesis: A mathematical framework*, Genetics **199**, 39 (2014), doi:10.1534/genetics.114.171850.

[166] F. Liu, A. H. Morrison and T. Gregor, *Dynamic interpretation of maternal inputs by the Drosophila segmentation gene network*, Proc. Natl. Acad. Sci. **110**, 6724 (2013), doi:10.1073/pnas.1220912110.

[167] A. Martinez Arias and P. Hayward, *Filtering transcriptional noise during development: Concepts and mechanisms*, Nat. Rev. Genet. **7**, 34 (2006), doi:10.1038/nrg1750.

[168] T. C. Lacalli, *Patterning, from conifers to consciousness: Turing's theory and order from fluctuations*, Front. Cell Dev. Biol. **10**, 871950 (2022), doi:10.3389/fcell.2022.871950.

[169] M. Bauer, M. D. Petkova, T. Gregor, E. F. Wieschaus and W. Bialek, *Trading bits in the readout from a genetic network*, Proc. Natl. Acad. Sci. **118**, e2109011118 (2021), doi:10.1073/pnas.2109011118.

[170] M. Bauer and W. Bialek, *Information bottleneck in molecular sensing*, PRX Life **1**, 023005 (2023), doi:10.1103/PRXLife.1.023005.

[171] N. Tishby, F. C. Pereira and W. Bialek, *The information bottleneck method*, (arXiv preprint) doi:10.48550/arXiv.physics/0004057.

[172] H. B. Barlow, *Sensory mechanisms, the reduction of redundancy, and intelligence*, in *NPL symposium on the mechanization of thought process*, London, UK (1959).

[173] H. B. Barlow, *Possible principles underlying the transformation of sensory messages*, in *Sensory communication*, MIT Press, Cambridge, USA, ISBN 9780262518420 (1961).

[174] E. Schrödinger, *Über das Verhältnis der Vierfarben- zur Dreifarbentheorie*, Sitzungsber. Kaiserl. Akad. Wiss. Wien. **134**, 471 (1925).

[175] K. R. Gegenfurtner and L. T. Sharpe, *Color Vision: From genes to perception*, Cambridge University Press, Cambridge, UK, ISBN 9780521004398 (2001).

[176] G. Buchsbaum and A. Gottschalk, *Trichromacy, opponent colours coding and optimum colour information transmission in the retina*, Proc. R. Soc. Lond. B. **220**, 89 (1983), doi:10.1098/rspb.1983.0090.

[177] D. L. Ruderman, T. W. Cronin and C.-C. Chiao, *Statistics of cone responses to natural images: Implications for visual coding*, J. Opt. Soc. Am. A **15**, 2036 (1998), doi:10.1364/JOSAA.15.002036.

[178] J. J. Atick and A. N. Redlich, *Towards a theory of early visual processing*, Neural Comput. **2**, 308 (1990), doi:10.1162/neco.1990.2.3.308.

[179] J. H. van Hateren, *Real and optimal neural images in early vision*, Nature **360**, 68 (1992), doi:10.1038/360068a0.

[180] D. L. Ruderman and W. Bialek, *Statistics of natural images: Scaling in the woods*, Phys. Rev. Lett. **73**, 814 (1994), doi:10.1103/PhysRevLett.73.814.

[181] H. B. Barlow, *Summation and inhibition in the frog's retina*, J. Physiol. **119**, 69 (1953), doi:10.1113/jphysiol.1953.sp004829.

[182] S. W. Kuffler, *Discharge patterns and functional organization of mammalian retina*, J. Neurophys. **16**, 37 (1953), doi:10.1152/jn.1953.16.1.37.

[183] H. K. Hartline, *Visual receptors and retinal interaction*, in *Nobel lectures, physiology or medicine 1963-1970*, Elsevier, Amsterdam, Netherlands, ISBN 9780444409942 (1972).

[184] C. E. Shannon, *Communication in the presence of noise*, Proc. IRE **37**, 10 (1949), doi:10.1109/JRPROC.1949.232969.

[185] S. B. Laughlin and R. R. de Ruyter van Steveninck, *Measurements of signal transfer and noise suggest a new model for graded transmission at an adapting retinal synapse*, J. Physiol. **494**, P19 (1996).

[186] B. N. Lundstrom, M. H. Higgs, W. J. Spain and A. L. Fairhall, *Fractional differentiation by neocortical pyramidal neurons*, Nat. Neurosci. **11**, 1335 (2008), doi:10.1038/nn.2212.

[187] B. N. Lundstrom, A. L. Fairhall and M. Maravall, *Multiple timescale encoding of slowly varying whisker stimulus envelope in cortical and thalamic neurons in vivo*, J. Neurosci. **30**, 5071 (2010), doi:10.1523/JNEUROSCI.2193-09.2010.

[188] J. Thorson and M. Biederman-Thorson, *Distributed relaxation processes in sensory adaptation*, Science **183**, 161 (1974), doi:10.1126/science.183.4121.161.

[189] G. Tkačik and W. Bialek, *Diffusion, dimensionality, and noise in transcriptional regulation*, Phys. Rev. E **79**, 051901 (2009), doi:10.1103/PhysRevE.79.051901.

[190] W. Bialek and S. Setayeshgar, *Physical limits to biochemical signaling*, Proc. Natl. Acad. Sci. **102**, 10040 (2005), doi:10.1073/pnas.0504321102.

[191] W. Bialek and S. Setayeshgar, *Cooperativity, sensitivity, and noise in biochemical signaling*, Phys. Rev. Lett. **100**, 258101 (2008), doi:10.1103/PhysRevLett.100.258101.

[192] J. S. van Zon, M. J. Morelli, S. Tănase-Nicola and P. R. ten Wolde, *Diffusion of transcription factors can drastically enhance the noise in gene expression*, Biophys. J. **91**, 4350 (2006), doi:10.1529/biophysj.106.086157.

[193] R. G. Endres and N. S. Wingreen, *Maximum likelihood and the single receptor*, Phys. Rev. Lett. **103**, 158101 (2009), doi:10.1103/PhysRevLett.103.158101.

[194] T. Mora and N. S. Wingreen, *Limits of sensing temporal concentration changes by single cells*, Phys. Rev. Lett. **104**, 248101 (2010), doi:10.1103/PhysRevLett.104.248101.

[195] K. Kaizu, W. de Ronde, J. Paijmans, K. Takahashi, F. Tostevin and P. R. ten Wolde, *The Berg-Purcell limit revisited*, Biophys. J. **106**, 976 (2014), doi:10.1016/j.bpj.2013.12.030.

[196] T. Mora, *Physical limit to concentration sensing amid spurious ligands*, Phys. Rev. Lett. **115**, 038102 (2015), doi:10.1103/PhysRevLett.115.038102.

[197] M. Carballo-Pacheco, J. Desponds, T. Gavrilchenko, A. Mayer, R. Prizak, G. Reddy, I. Nemenman and T. Mora, *Receptor crosstalk improves concentration sensing of multiple ligands*, Phys. Rev. E **99**, 022423 (2019), doi:10.1103/PhysRevE.99.022423.

[198] T. Mora and I. Nemenman, *Physical limit to concentration sensing in a changing environment*, Phys. Rev. Lett. **123**, 198101 (2019), doi:10.1103/PhysRevLett.123.198101.

[199] V. Ngampruetikorn, D. J. Schwab and G. J. Stephens, *Energy consumption and cooperation for optimal sensing*, Nat. Commun. **11**, 975 (2020), doi:10.1038/s41467-020-14806-y.

[200] G. Tkačik, A. M. Walczak and W. Bialek, *Optimizing information flow in small genetic networks*, Phys. Rev. E **80**, 031920 (2009), doi:10.1103/PhysRevE.80.031920.

[201] A. M. Walczak, G. Tkačik and W. Bialek, *Optimizing information flow in small genetic networks. II. Feed-forward interactions*, Phys. Rev. E **81**, 041905 (2010), doi:10.1103/PhysRevE.81.041905.

[202] G. Tkačik, A. M. Walczak and W. Bialek, *Optimizing information flow in small genetic networks. III. A self-interacting gene*, Phys. Rev. E **85**, 041903 (2012), doi:10.1103/PhysRevE.85.041903.

[203] S. C. Little, M. Tikhonov and T. Gregor, *Precise developmental gene expression arises from globally stochastic transcriptional activity*, Cell **154**, 789 (2013), doi:10.1016/j.cell.2013.07.025.

[204] T. R. Sokolowski and G. Tkačik, *Optimizing information flow in small genetic networks. IV. Spatial coupling*, Phys. Rev. E **91**, 062710 (2015), doi:10.1103/PhysRevE.91.062710.

[205] T. R. Sokolowski, A. M. Walczak, W. Bialek and G. Tkačik, *Extending the dynamic range of transcription factor action by translational regulation*, Phys. Rev. E **93**, 022404 (2016), doi:10.1103/PhysRevE.93.022404.

[206] J. Dubnau and G. Struhl, *RNA recognition and translational regulation by a homeodomain protein*, Nature **379**, 694 (1996), doi:10.1038/379694a0.

[207] R. Rivera-Pomar, D. Niessing, U. Schmidt-Ott, W. J. Gehring and H. Jacklë, *RNA binding and translational suppression by bicoid*, Nature **379**, 746 (1996), doi:10.1038/379746a0.

[208] D. Niessing, W. Driever, F. Sprenger, H. Taubert, H. Jäckle and R. Rivera-Pomar, *Homeodomain position 54 specifies transcriptional versus translational control by bicoid*, Mol. Cell **5**, 395 (2000), doi:10.1016/S1097-2765(00)80434-7.

[209] O. Johnstone and P. Lasko, *Translational regulation and RNA localization in Drosophila oocytes and embryos*, Annu. Rev. Genet. **35**, 365 (2001), doi:10.1146/annurev.genet.35.102401.090756.

[210] T. R. Sokolowski, T. Gregor, W. Bialek and G. Tkačik, *Deriving a genetic regulatory network from an optimization principle*, (arXiv preprint) doi:10.48550/arXiv.2302.05680.

[211] J. Monod, J.-P. Changeux and F. Jacob, *Allosteric proteins and cellular control systems*, J. Mol. Biol. **6**, 306 (1963), doi:10.1016/S0022-2836(63)80091-1.

[212] J. Monod, J. Wyman and J.-P. Changeux, *On the nature of allosteric transitions: A plausible model*, J. Mol. Biol. **12**, 88 (1965), doi:10.1016/S0022-2836(65)80285-6.

[213] M. F. Perutz, *Mechanisms of cooperativity and allosteric regulation in proteins*, Cambridge University Press, Cambridge, UK, ISBN 9780521386487 (1990).

[214] R. Phillips, *The molecular switch: Signaling and allostery*, Princeton University Press, Princeton, USA, ISBN 9780691200248 (2020).

[215] J. J. Hopfield, *Relation between structure, co-operativity and spectra in a model of hemoglobin action*, J. Mol. Biol. **77**, 207 (1973), doi:10.1016/0022-2836(73)90332-X.

[216] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev and R. Phillips, *Transcriptional regulation by the numbers: Models*, Curr. Opin. Genet. Dev. **15**, 116 (2005), doi:10.1016/j.gde.2005.02.007.

[217] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman and R. Phillips, *Transcriptional regulation by the numbers: Applications*, Curr. Opin. Genet. Dev. **15**, 125 (2005), doi:10.1016/j.gde.2005.02.006.

[218] B. Zoller, T. Gregor and G. Tkačik, *Eukaryotic gene regulation at equilibrium, or non?*, Curr. Opin. Syst. Biol. **31**, 100435 (2022), doi:10.1016/j.coisb.2022.100435.

[219] P. S. Swain, M. B. Elowitz and E. D. Siggia, *Intrinsic and extrinsic contributions to stochasticity in gene expression*, Proc. Natl. Acad. Sci. **99**, 12795 (2002), doi:10.1073/pnas.162041399.

[220] A. Crombach, M. A. García-Solache and J. Jaeger, *Evolution of early development in dipterans: Reverse-engineering the gap gene network in the moth midge Clogmia albipunctata (psychodidae)*, Biosystems **123**, 74 (2014), doi:10.1016/j.biosystems.2014.06.003.

[221] K. R. Wotton, E. Jiménez-Guri, A. Crombach, H. Janssens, A. Alcaine-Colet, S. Lemke, U. Schmidt-Ott and J. Jaeger, *Quantitative system drift compensates for altered maternal inputs to the gap gene network of the scuttle fly megaselia abdita*, eLife **4**, 04785 (2015), doi:10.7554/eLife.04785.

[222] Y. Goltsev, W. Hsiong, G. Lanzaro and M. Levine, *Different combinations of gap repressors for common stripes in anopheles and Drosophila embryos*, Dev. Biol. **275**, 435 (2004), doi:10.1016/j.ydbio.2004.08.021.