INFORMATION FLOW IN BIOLOGICAL NETWORKS

Gašper Tkačik

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF PHYSICS

September 2007

Abstract

Biology presents us with several interesting kinds of networks that transmit and process information. *Neurons* communicate by exchanging action potentials; *proteins* in a signaling pathway detect chemicals at the cell surface and conduct the information to the nucleus in a cascade of chemical reactions; and in gene regulation, *transcription factors* are responsible for integrating environmental signals and appropriately regulating their target genes.

Understanding the collective behavior of biological networks is not easy. These systems are inherently noisy and thus require a model of the mean dynamics as well as that of the noise; in addition, if we view the regulatory networks as information transmission devices, implemented by the "hardware" of chemical reactions, we need to describe them in a probabilistic, not deterministic (or noiseless), language. Unfortunately, connecting theory with experiment then becomes exponentially hard due to sampling problems as the number of interacting elements grows, and progress depends on finding some simplifying principle.

In this work two such principles are presented. In the first half I discuss a bottom-up approach and analyze the responses of a set of retinal ganglion cells to a naturalistic movie clip, and the activation states of proteins in a signaling cascade of immune system cells. The simplifying principle here is the idea that the distribution of activities over elements of the network is maximum entropy (or most random), but still consistent with some experimentally measured moments (specifically, pairwise correlations). The analogies between maximum entropy and Ising models are illustrated and connected to the previously existing theoretical work on spin-glass properties of neural networks.

In the second part a top-down approach is presented: viewing genetic regulatory networks as being driven to maximize the reliability of the information transmission between their inputs and outputs, I first examine the best performance of genetic regulatory elements achievable given experimentally motivated models of noise in gene regulation; and second, make the hypothesis that, in some systems at least, such optimization is beneficial for the organism and that its predictions are verifiable. Data from early morphogenesis in the fruit fly, *Drosophila melanogaster*, are used to illustrate these claims.

Acknowledgements

There is something remarkable about my advisors, William Bialek and Curtis Callan, apart from their sheer strength as scientific thinkers. Bill and Curt seem to enjoy life, and let younger colleagues take part in that. They lead by example in showing that being 'professorial' and having fun are not incompatible. It sounds like a simple idea, but I found it important. At times, even a strong combination of intellectual curiosity and a long-term view of one's scientific endeavors (in particular how they might be realized by obtaining a tenured faculty position) fail to fully motivate the pursuit of a career in science. At those moments, it becomes particularly important to be happy and to enjoy the company of the people around you.

I would also like to thank people that I had the pleasure of interacting with in the past few years: Justin, Sergey, Thomas, Noam, Elad, Greg and Stephanie. Postdoctoral fellows especially were happy to share some of their accumulated wisdom – both scientific and of things more general – with their younger, inexperienced friends.

I hereby acknowledge the support of the Princeton University, Burroughs-Wellcome Fund and Charlotte E. Procter fellowship.

Some parts of this thesis have appeared as preprints on the arXiv. I would like to separately thank my collaborators on the following projects:

Section 2.6	*Ising models for networks of real neurons* with E Schneidman, MJ Berry II and W Bialek	arXiv:q-bio.NC/0611072
Section 3.2	*The role of input noise in gene regulation* with T Gregor and W Bialek	arXiv:q-bio.MN/0701002
Section 4.6	*Information flow and optimization in transcriptional control* with CG Callan Jr and W Bialek	arXiv:0705.0313v1

Sonji.

Contents

List of Figures

Chapter 1

Introduction

Organisms are devices that have evolved to produce particular responses, beneficial for their own survival and reproduction, to the stimuli that they are exposed to in their natural environments. The proverbial encounter of a cheetah with its prey, the antelope, is a good example. While grazing, the antelope is on the lookout for its predator, and has to react quickly to the visual, auditory or olfactory signals that could indicate the cheetah's presence. There is a fatal cost for a "false negative" and potentially as high a cost for being slow; in comparison, making a "false positive" choice is a mere nuisance, and the worst that can happen is that the animal gets tired trying to escape from phantom cats lurking in the shadows.

The imaginary grassland scene where the antelope is being chased represents an interplay of several conceptual ingredients. First, there is the current state of the world and its rules – the probability of encountering a cheetah or of other events happening given the location or the time of day, physical laws governing the dynamics of the animals' bodies and so on; second, there are the detection and processing of environmental signals by the antelope's sensory system and its brain; and finally, there is the policy of action – (how) to flee or whether to continue grazing – and a possible, perhaps deadly, feedback. We have a clear intuition that for the antelope there is a limited set of "correct" ways to behave, and there are theoretical frameworks that embody this intuition, like optimal Bayesian decision making (MacKay, 2003) or game theory (von Neumann and Morgenstern, 1947). Recent work has incorporated two additional important notions into this overall picture. The first is the idea of learning and adaptation. Unlike textbook problems on inference, the underlying probabilities of various events are not initially known to the animal. The world is an orderly place because of its rules, which operate on many levels and induce correlations in time between events: knowing what has happened until now is informative about what will happen next (Bialek et al., 2001), and the animal will pay for the cost of information gathering and processing if it can extract the predictive part of the information that will guide its future actions (Tishby, 2007). The second idea is that there are physical limits to perception and signaling due to sources of noise that cannot be eliminated given other relevant constraints, and this noise must restrict the reliability of the organism's decision-making (Bialek, 1987, 2002). Putting everything together, one has a view of a complex – yet not chaotic – world from which the organisms constantly acquire information at some cost and with limited precision, and learn to use the predictive part of this information to perform actions, either in the blink of an eye or perhaps even years later.

While in principle there exists an information theoretic language based on the laws

of probability in which the above framework can be formalized (Shannon, 1948; Cover and Thomas, 1991), the practical goal of taking some organism and mapping the stimuli which act on it to its choice of possible behaviors is probably still unattainable. We run into difficulty as soon as we try to characterize quantitatively the relevant dimensions in the space of stimuli and responses and have to stop before even trying to understand the computations that implement the stimulus-response map, or the statistical structure of the inputs. For example, human "behavior" manifests itself on many levels, from low-level motor actions to complex social phenomena, and trying to label it, find the stimuli that elicit the corresponding responses, or sample enough stimulus-response pairs seems daunting. Even for very simple organisms like the extensively studied roundworm, *Caenorhabditis elegans* (an animal with 959 cells of which 302 are neurons), finding the relevant "features" in organism's behavior, i.e. trying to quantify and separate a stereotyped response from its natural variability or noise, can be a serious research undertaking (Stephens et al., 2007).

There are cases, however, where the behavior of the whole organism is in some way dependent on, and therefore limited by, a much smaller and perhaps manageable subsystem. For example, neurons at the periphery of the visual system that encode the sensory inputs are the single conduit through which all image-related information must flow to the brain, regardless of the complexity of downstream processing (Barlow, 1981). Alternatively, some behaviors of a yeast or bacterial cell can be orchestrated by networks of a few coupled genes that coherently respond to various experimentally controllable conditions, or so called *modules* (Ihmels et al., 2002; Slonim et al., 2006). Pinpointing the cases (and subsystems) where such "information bottlenecking"[1] occurs can be productive for many reasons. Firstly, the set of possible inputs and outputs of this, much smaller, subsystem might now be restricted enough in size to allow for quantitative description and thorough experimental sampling. Secondly, the subsystem itself will probably be smaller, and as we move along the length scale from cheetahs and antelopes past cells down to single molecules, we gain a new leverage: for example, physical laws now tell us how the neural spike propagates along the axon (Hodgkin and Huxley, 1952) or about the structural and functional interaction between the DNA and example regulatory proteins (Benoff et al., 2002; Kuhlman et al., 2007). In general, we will have the knowledge to model phenomenologically the response to an input (and the accompanying noise) from physical principles.

Apart from having a well-defined parametrization of the functionally relevant features of the inputs and outputs, and being smaller in terms of physical size, there is yet another way in which a subsystem can be "small". Biological computation often is implemented in a dynamical system composed of (relatively) simple nodes that interact and thus give rise to (possibly) non-trivial collective behaviors of its constituent elements (Hopfield, 1982; Hopfield and Tank, 1986; Hartwell et al., 1999). Smallness in this context refers to the number of effective constituent elements, and simplicity to the fact that the elements internally do not have many states or different behaviors. In this case we often represent the dynamical system graphically as a network (Alon, 2003), where every vertex denotes a constituent element, or a *node*, and we draw connections between the nodes based on some notion of correlation or interaction. Every node has an associated dynamical variable that we abstractly call *activity*, and this can mean either the spiking rate of a particular neuron, the expression level of a gene, or the activation/phosphorylation state of a particular signaling protein. We will define the concepts more precisely in the following chapters.

[1] There is a formal notion of an information bottleneck as a specific form of data compression which has a correspondence to the way we use the phrase here; see Tishby et al. (2000) for details.

In the first part of the thesis we discuss the problem of inferring the structure of the network from experiments that provide snapshots of the states of nodes. More precisely, we will assume that we have identified some of the interacting elements, but do not know the strengths of their interactions. Our approach will be data-driven, or bottom-up: starting with multiple joint measurements of the activities of a set of nodes, we will try to characterize their interactions. Clearly, we have many useful approaches at our disposal. We consider deterministic models, in which one postulates a set of differential equations describing the behavior of the node activities (and motivates it by some microscopic picture of how the system is assumed to work) and then proceeds to fit the model coefficients to the data; the connectivity map of the network that gives rise to the dynamical equations can even be iteratively updated for a better fit. Such approaches have been successfully applied in the study of circadian clocks and the oscillators responsible for the cell cycle, with the nodes representing the expression levels of various mutually (in)activating genes and protein activation levels (Chen et al., 2000; Leloup and Goldbeter, 2003); other systems have been studied as well in this context, see Tyson et al. (2001) for a review. Similarly, theoretical neuroscience has explored neurons from a dynamical systems perspective (Dayan and Abbott, 2001). Alternatively, genetic circuitry has been also modeled as a logical deterministic network with activities taking on a discrete set of possible values (Sanchez and Thieffry, 2001), or studied from a purely topological perspective (Shen-Orr et al., 2002).

We will soon see, however, that noise – in the sense of observed variability at constant input conditions – is important in the cases of interest to us, and that a probabilistic description is required. Only cases where systems are exposed to stimuli drawn from some specified ensemble will be considered, after the adaptation to the statistics of the ensemble has taken place. In this "steady-state" regime a noisy system constantly fluctuates around a fixed point, and each measurement can be considered as a draw from a stationary distribution that can depend on the stimulus. These assumptions, along with the desire to formulate a probabilistic model that does not incorporate any prior knowledge beyond the measured data, will lead us to *maximum-entropy models*, strongly analogous to the Ising models of statistical physics. Importantly, these models will be phenomenological, in contrast to microscopic ones, where the mechanistic treatment of an interaction between two elements is followed up to the whole network.[2] This distinction between microscopic and phenomenological models of biological networks is one that we will have to constantly keep in mind when interpreting our results.

In the second part of the thesis, we will take a top-down approach to study transcriptional control in genetic regulatory networks, that is, we will try to derive the properties of simple regulatory elements from a first principle. Again, noise will play an important role and will correspondingly motivate us to posit that the *information capacity* of genetic circuits is maximized, a variation on the idea in neural sensory coding where adaptation of neurons to the input signals is often explained in terms of such maximization. This hypothesis will generate testable predictions about the distribution of proteins used for signaling and will put bounds on the reliability of regulatory elements, which can be compared to measurements.

We will pick a biochemical network of signaling proteins, retinal ganglion neurons and transcriptional regulation during morphogenesis as examples, all of them being instances of biological systems where information flows through a small and rather restricted network

[2] In a microscopic model one would postulate, for example, Michaelis-Menten or similar kinetic equations for enzymatic reactions or integrate-and-fire equations for neurons, and try to learn about the collective behavior of the network by coupling the equations for constitutive elements together.

under our (and the experimentalist's) control, and where the inputs and outputs are relatively easily describable: for signaling proteins one stimulates the network with chemical perturbations, for neurons one projects the images onto the retina and records the outgoing spikes; for transcriptional regulation, one measures the concentration fields of fluorescently-tagged transcription factors using microscopy. If these subsystems are biologically essential for the whole organism, yet noisy in their implementation, we can hope to capture the signatures of an optimization principle at work. And finally – much in the same way as for the antelope's story, where information theory allows us to make statements about perception, computation and optimal behavior despite being unable to write down the corresponding detailed dynamical model – the same information theory here will help us understand the smaller networks and shift the focus from the prevailing question *'What is the set of microscopic interactions between constituent elements?'* to a more interesting *'How can a biological network collectively perform its computation?'*

The thesis is organized around three papers referenced on the Acknowledgments page (and in their corresponding sections) that have been reproduced here with as few changes as possible. I have tried to provide enough introductory material to make the reading smooth, at the expense of sometimes (but hopefully not often) repeating what was already said. Because the thesis covers what are considered to be distinct topics, some references for classic and review papers specific to particular chapters are cited there instead of in the Introduction. The data analysis methods and some computations, along with a substantial number of interesting side results referred to in the main text, are presented in the Methods Appendix.

Chapter 2

Building networks from data

In physical systems correlation functions are of interest because of two reasons. First, they are "natural" to compute in theory and involve taking derivatives of the log partition function with respect to the coupling constants. Second, they are connected to experimentally observable properties, such as scattering cross sections or susceptibilities, and their behavior often characterizes the macroscopic order in the system. In networks of genes or neurons the experiments usually amount to collecting snapshots of the instantaneous state of the system, and while it is possible to define and compute the correlations between the constituent elements, it is not immediately clear how to carry over our intuitions from statistical mechanics to biological networks.

We try to explore the issue by first presenting information theoretic tools that are required to deal formally with non-parametric probabilistic descriptions of the data; specifically, we will first show how to measure the "correlation" between nodes in a principled way, and will generalize this measure in several interesting directions. Then, the difference between correlations and interactions will be discussed, with special emphasis placed on distinguishing effective interactions of the phenomenological model from the real, physical interactions in the modeled system. Finally, two concrete networks will be studied: a set of interacting proteins in a signaling cascade of human immune response cells, and a group of ganglion neurons of the salamander retina.

2.1 Computing correlations

Let σ denote the activity variables defined for each of the nodes in a network. When we think of a correlation measure between two elements σ_i and σ_j, we usually mean either their covariance:

$$C_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle, \tag{2.1}$$

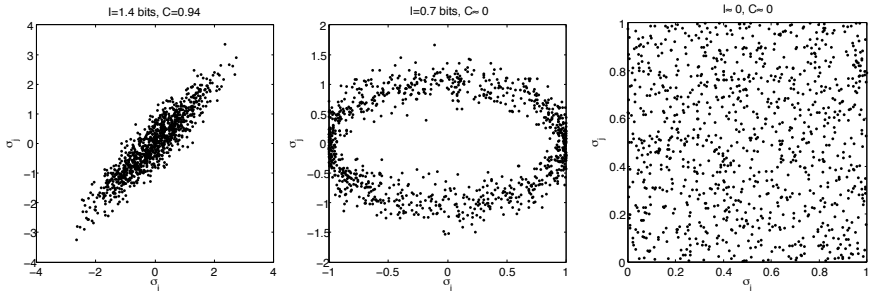
or its normalized version, the correlation coefficient:

$$R_{ij} = \frac{\langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle}{\sqrt{(\langle \sigma_i^2 \rangle - \langle \sigma_i \rangle^2)(\langle \sigma_j^2 \rangle - \langle \sigma_j \rangle^2)}}, \tag{2.2}$$

where σ_i is a measure of "activity" at node i, and can quite generally be either a discrete or continuous quantity. This measure of dependence is intuitive as it "interpolates" between the case where σ_i and σ_j are chosen independently from each other and $R_{ij} = 0$, and the

case in which they are perfectly linearly correlated and $|R_{ij}| = 1$. R can be taken as a measure of the goodness-of-fit if the model dependence is linear, i.e. $\sigma_j = A\sigma_i + B$.

Despite being conceptually appealing and easy to estimate from the data, correlation has at least two problems as a generic measure of dependency. Firstly, it does not capture non-linear relationships, as shown in Fig 2.1b; secondly, when σ take on discrete values that are not ordered (e.g. a set of possible multiple-choice responses on a test), the linear correlation loses its meaning, although the problem itself is well posed (e.g. What is the correlation between two answers on a multiple-choice test across respondents?).



2.1a: Linear correlation.

2.1b: Nonlinear correlation.

2.1c: No correlation.

Figure 2.1: Correlation coefficient and mutual information as measures of dependency. Left panel: the points drawn from a joint distribution that embodies linear dependence plus noise have both a high mutual information and high linear correlation. Middle panel: in case of nonlinear dependence, the correlation coefficient can be zero although the variables are clearly strongly correlated. Right panel: if the joint probability distribution is a product of factor distributions for both variables, then the correlation coefficient and the mutual information measures are zero.

There is an alternative way of defining dependency, or correlation, between two variables due to Shannon (Shannon, 1948; Cover and Thomas, 1991). Let us suppose that both σ_i and σ_j are drawn from a joint distribution $p(\sigma_i, \sigma_j)$. For argument's sake, suppose further that we do not know anything about the value of σ_i. Then the entropy of $p(\sigma_j)$:

$$S[p(\sigma_j)] = -\int d\sigma_j \, p(\sigma_j) \log_2 p(\sigma_j) \tag{2.3}$$

is a useful measure of uncertainty about the value of σ_j, and, as defined above, is a value measured in bits. This information-theoretic entropy is equivalent to physical entropy up to a multiplicative constant, and is defined up to an additive constant (connected to the finite resolution of σ) for continuous variables, with a straightforward generalization for discrete variables.

We have assumed that σ_i and σ_j have been drawn from an underlying joint distribution; in contrast to the case above, if we actually know something about σ_i, our uncertainty about σ_j might be reduced. The uncertainty in σ_j that remains if the value of σ_i is known is again defined by the (conditional) entropy:

$$S[p(\sigma_j|\sigma_i)] = -\int d\sigma_j \, p(\sigma_j|\sigma_i) \log_2 p(\sigma_j|\sigma_i). \tag{2.4}$$

We can now define the mutual information between elements σ_i and σ_j as:

$$I(\sigma_i; \sigma_j) = S[p(\sigma_j)] - \langle S[p(\sigma_j|\sigma_i)] \rangle_{p(\sigma_i)}, \tag{2.5}$$

where we write $I(\sigma_i;\sigma_j)$ as a shorthand for $I[p(\sigma_i,\sigma_j)]$, i.e. the mutual information is a functional of the joint distribution.

The mutual information is the average reduction in entropy about one variable if one has knowledge of the other, related, variable. This measure is symmetric in the exchange of the variables, which is manifest if we rewrite Eq (2.5) as:

$$I(\sigma_i;\sigma_j) = \int d\sigma_i \int d\sigma_j \, p(\sigma_i,\sigma_j) \log_2 \frac{p(\sigma_i,\sigma_j)}{p(\sigma_i)\,p(\sigma_j)}. \tag{2.6}$$

Mutual information is always positive and is measured here in bits; it is well-defined for continuous and discrete supports because it is a difference of two entropies. This measure also has a clear interpretation: saying that the variable σ_i has I bits of mutual information with variable σ_j means that there are $2^{I(\sigma_i;\sigma_j)}$ states of the variable σ_i that give rise to distinguishable states in σ_j.[1] One bit of mutual information at least conceptually means that there are two such values (or intervals) of σ_i, which separately map into two corresponding values (or intervals) of σ_j, without the possibility of confusion due to noise. Note that mutual information in no way specifies the nature of the dependency, as opposed to the correlation coefficient that specifically measures how good a *linear* model is; mutual information is a statistical measure that states how many bits, on average, one learns about σ_j by knowing σ_i, but says nothing about $\sigma_j = \sigma_j(\sigma_i)$, i.e. about the *actual value* that σ_j takes if σ_i is known.

Estimating the mutual information from the data by way of Eq (2.5) is notoriously difficult as one is making systematic errors due to undersampling,[2] but there are methods that explicitly use the scaling of entropy with the number of samples to correct for this bias to lowest order (Strong et al., 1998). We have adapted the method to large-scale estimation of many pairwise mutual information relations on a finite number of samples;[3] this so-called *direct method* will be used in the remainder of the thesis (Slonim et al., 2005a).

If we adopt the terminology used by Shannon to discuss information transmission, we speak of the communication channel between two variables, and regard σ_i as the source (or input), and σ_j as the output.[4] Equation (2.5) can then be read as follows: the maximum value for the mutual information would be the entropy of the output, $S[p(\sigma_j)]$, but because the communication channel is noisy, we have to subtract the so-called *noise entropy*, or $\langle S[p(\sigma_j|\sigma_i)]\rangle$. Clearly then the information is bounded from above by the output entropy, and is often normalized by $S[p(\sigma_j)]$ to give a value between zero and one.

[1] "Distinguishable" is meant as "distinguishable given noise": a certain value of σ_i corresponds to a distribution of values for σ_j, as described by the conditional distribution $p(\sigma_j|\sigma_i)$; for two different values of σ_i there will be two distributions of σ_j, and to be distinguishable, they must not substantially overlap.

[2] When *naive estimation* of the mutual information is performed, one takes N samples of data and bins them into a 2D histogram, from which a frequentist estimate of the joint distribution, $\tilde{p}(\sigma_i,\sigma_j)$, is created and used to compute I in Eq (2.6). Often there will be too few samples to have a good coverage of the histogram domain.

[3] The basic idea is that the mutual information will behave as $I(N) = I(\infty) + \alpha/N + \cdots$, where $I(\infty)$ is the correct information value that one would obtain at infinite sample size; $I(N)$ is the naive estimate at N samples, obtained by binning the data and estimating the entropies; and the term inversely proportional to N is the first-order bias. By taking many subsamples of the data at different fractions of total size N, one is able to estimate α and extrapolate to infinite sample size. The reader is directed to Slonim et al. (2005a) for details on extrapolation, binning etc.

[4] Because of the symmetry of information in its arguments, the designations "input" and "output" are arbitrary and acquire their meaning only when we map the information theoretic framework onto a physical system.

The concept of mutual information can be generalized in several ways. First, we can define multi-information among N variables, $\vec{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$, by extending Eq (2.6) as follows:

$$I(\vec{\sigma}) = \int d\vec{\sigma}\, p(\vec{\sigma}) \log_2 \frac{p(\vec{\sigma})}{\prod_i p_i(\sigma_i)}. \tag{2.7}$$

The multi-information quantifies how much "structure" or non-uniformity there is in the distribution, and is a general measure of dependency between N elements. Is there any scale to which this number can be compared, in a way similar to the comparison of the mutual information with the entropy of the output, as we have shown above? If the network consists of N nodes, with each node having Q possible states (in a discrete case), the entropy of the distribution must be smaller than $S_{\text{uniform}} = N \log_2 Q$, i.e. the entropy of the uniform distribution on the same support. We can go even a step further: we can imagine a *factor*, or *independent, distribution*, $p^{(1)}(\vec{\sigma}) = \prod_i p_i(\sigma_i)$, where $p_i(\sigma_i) = \sum_{\sigma_j, j\neq i} p(\vec{\sigma})$ – this is a distribution that is a product of single-element distributions for each of the N elements where every factor is just a marginal of the joint distribution over all other elements. The independent distribution is easy to compute and all its single-point statistics will match those of the full joint distribution. Because it has a product form, its entropy is simply the sum of entropies of individual factor distributions:

$$S\left[p^{(1)}(\vec{\sigma})\right] = -\sum_i \int d\sigma_i\, p_i(\sigma_i) \log_2 p_i(\sigma_i). \tag{2.8}$$

We will refer to the entropy of the independent distribution $p^{(1)}(\vec{\sigma})$ as the *independent entropy* and denote it as $S_{\text{ind}}[p^{(1)}(\vec{\sigma})]$ to remind ourselves that this entropy does not include any effects of the correlations and that the true entropy of the joint distribution must be lower. Furthermore, we see that the multi-information of Eq (2.7) can be rewritten as:

$$I(\vec{\sigma}) = S_{\text{ind}}[p^{(1)}(\vec{\sigma})] - S[p(\vec{\sigma})]. \tag{2.9}$$

The multi-information is therefore the reduction in entropy due to the correlations between N elements.[5] We see from Eq (2.9) that S_{ind} again provides a convenient normalization for the multi-information; when we introduce the maximum entropy models, we will also find out that there exists a unique decomposition of the multi-information into a sum of connected information terms that describe the successive reductions in entropy due to the presence of 2-body, 3-body, and up to N-body interactions.

Multi-information is a special case of the Kullback-Leibler (KL) divergence of the two distributions (Cover and Thomas, 1991):

$$D_{\text{KL}}(p \,||\, q) = \int dx\, p(x) \log_2 \frac{p(x)}{q(x)}. \tag{2.10}$$

The KL divergence is (almost) a distance metric on the space of distributions, but is not symmetric. If random variables x were really drawn from distribution $p(x)$, but we assumed they were drawn from $q(x)$ instead, and were to build an encoding for x using this (wrong) assumption, the KL divergence would measure the number of bits needed to encode x *in excess* of the optimal encoding achievable using the correct model, $p(x)$. The multi-information is related to the KL divergence as:

$$I(p(\vec{\sigma})) = D_{\text{KL}}\left(p(\vec{\sigma}) \,||\, p^{(1)}(\vec{\sigma})\right). \tag{2.11}$$

[5] If there are only two variables, the reader can easily verify that multi-information is just the mutual information and that all formulae for N-body system match the corresponding ones in the case of 2 elements.

Consistent with our intuition that the multi-information captures the dependency beyond the single-point statistics, Eq (2.11) states that the multi-information is the number of bits we need, in addition to the factor distribution model, to specify elements drawn from the full joint distribution $p(\vec{\sigma})$.

A symmetric version of the KL divergence is the Jensen-Shannon (JS) divergence, or:

$$r(x) = \frac{1}{2}\left(p(x) + q(x)\right), \tag{2.12}$$

$$D_{\mathrm{JS}}(p,q) = \frac{1}{2}\left(D_{\mathrm{KL}}(p\,||\,r) + D_{\mathrm{KL}}(q\,||\,r)\right). \tag{2.13}$$

The JS divergence is always between zero and one and is approximately the inverse of the number of samples that would have to be drawn from the distribution $p(x)$ to say with confidence that they do not come from $q(x)$ (Lin, 1991).

2.2 Networks from correlations

After having introduced the information theoretic measures of correlation, it is time to ask why it is advantageous to use those instead of easily computable linear correlation coefficients. In this section, I will discuss an interesting example where the analysis has benefited from the properties of the information measure that the linear correlation lacks. In parallel, we will also present one of the simplest, yet powerful and productive ways of understanding the collective behavior of networks of many elements – clustering.

Microarrays have enabled genome-wide surveys of the changes in gene expression (more precisely in mRNA levels) as the organism is subjected to different environments. Yeast, for example, can be grown both at reference (neutral) growth conditions and in a number of environments that are strongly perturbed away from the reference, usually by changing the nutrient media, pH, temperature or by adding various chemicals that stress the organism. Messenger RNA from reference yeast is then extracted and tagged with green fluorescent probes, while perturbed yeast is tagged in red; both are hybridized to the same DNA microarray, onto which segments of coding DNA have been spotted, such that each spot localizes the sequences from a single known gene. The red and green message reverse transcripts compete on the microarray for binding onto their complementary spots, and by reading the relative red/green intensity ratios for each spot (and thus for each one of the ≈ 6000 yeast genes) one can determine whether, for every condition, a certain gene is over- or under-expressed relative to the reference.

A standard data analysis technique takes these N genes exposed to M different conditions, and tries to group genes into clusters that behave "similarly" when the conditions are changed (Eisen et al., 1998). More precisely, a matrix of pairwise similarities is first constructed from the data, usually simply by calculating the matrix of correlation coefficients between the genes (and across the conditions). With this $N \times N$ matrix in hand, one employs a standard clustering algorithm[6] to partition N genes in $O(\sqrt{N})$ clusters; if the clustering is successful, the genes are probably co-regulated and form a *module*. Although clustering would not provide a description of *how* the co-regulation works, it does reduce the dimensionality of the problem by cutting its size from N elements down to several clusters, based on each gene's response to a changed environment and thus, probably, on its function.

[6]For example, K-means or hierarchical clustering; see Jain et al. (1999) for a short review.

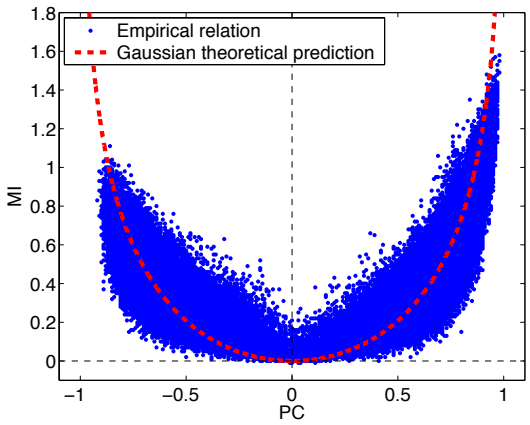


Figure 2.2: Linear (Pearson) correlation coefficient (PC) vs the mutual information (MI) in bits. Each point represents a pair of yeast Emergency Stress Response (ESR) genes; information and correlation are estimated across 173 conditions, and the measurements are the log light intensity ratios of microarray spots in the perturbed and reference condition.

We decided to reexamine both steps of the data analysis with information theoretic tools. The publicly available dataset of Gasch et al. (2000) comprising about $N = 900$ yeast genes was chosen; experimenters characterized their expression patterns in different conditions. Pairwise mutual information values $I(\sigma_i, \sigma_j)$ were calculated using the direct method and compared to the linear correlation coefficients, as shown in Fig 2.2. If every pairwise distribution $p(\sigma_i, \sigma_j)$ were a Gaussian, there would be an analytic relation between the mutual information and the correlation of the two variables, shown in red. Although this relation is generally observed, there is significant deviation; moreover, there are pairs of genes for which the correlation coefficient is small, but mutual information is significantly above zero, and for some datasets this effect can be more pronounced (not shown). These pairs are thus (wrongly) declared to be unrelated by the linear measure.

In addition to being sensitive to nonlinear correlations, the mutual information is also invariant to invertible reparametrizations of its arguments (because it is a functional of the probability distribution). It therefore does not matter if the information is estimated between the green/red intensity ratio, the log of the intensity ratio, or any affine transformation of it, including, of course, the change in units; this is in stark contrast to the linear correlation measure, and is a very desirable feature for biological data sets where the experimental error model (the likelihood of a measured signal given some physical event taking place on the array) is not well understood.

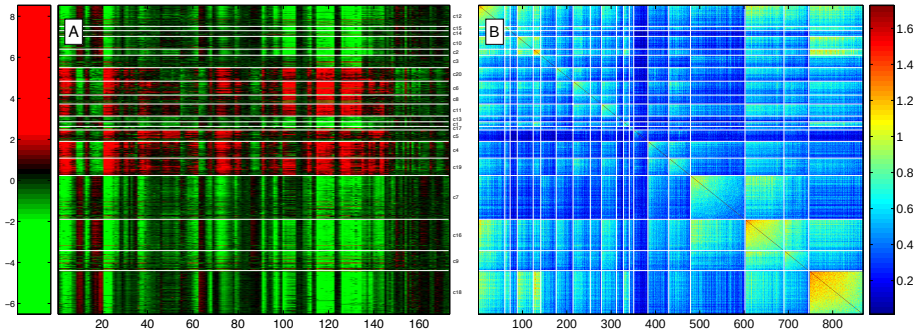
In the work of Slonim et al. (2005b) we have reformulated the problem of clustering: taking the pairwise similarity matrix obtained through information estimation as an input and partitioning elements into a smaller number of clusters, such that the elements within a cluster are more similar to each other on average than they are to the elements in other clusters. The main idea is to assign each of the σ_i elements to one of the clusters C by choosing the assignments $P(C|i)$ such the functional $\mathcal{F}$ is maximized:

$$\mathcal{F} = \langle s \rangle - TI(C;i) \tag{2.14}$$

$$\langle s \rangle = \sum_C P(C)s(C) = \sum_C P(C) \sum_{i_1,\ldots,i_r} P(i_1|C)\cdots P(i_r|C)I(i_1,\ldots,i_r), \tag{2.15}$$

where $I(i_1,\ldots,i_r)$ is a similarity measure between r elements. Here we take r to be 2 and I is therefore *pairwise* or mutual information, but the clustering formulation allows the generalization along the lines of multi-information in Eq (2.7). The functional $\mathcal{F}$ is a tradeoff, enforced by a tunable parameter T, between the desire to increase the similarity $\langle s \rangle$

between the elements of the same cluster (which is maximized if every element is placed in its own cluster), and decrease the description length when elements σ_i are replaced by clusters C (which is minimized if all elements are placed into the same cluster). Mutual information enters the computation as both the Shannon measure of channel capacity or compression in $I(C;i)$, which we want to minimize, and as a r-body element-wise similarity measure, $I(i_1,\ldots,i_r)$. Optimal assignments of elements i to clusters C are found at the stationary points of the functional $\mathcal{F}$. Figure 2.3a shows the raw data, and Fig 2.3b the mutual information matrix $I(\sigma_i,\sigma_j)$ reordered such that the genes in the same cluster are next to each other. The use of information theoretic formulation of clustering has helped us uncover interesting new structure in the dataset (Slonim et al., 2005a). Moreover, we demonstrate in the paper that the *generic* nature of mutual information as the similarity measure enabled us to construct a clustering method that works for data of different origins and statistical properties (gene expression, daily fluctuations in stock market prices, qualitative ratings of movies by fans) without any tuning parameters apart from the generic parameter T. Information-based clustering outperforms the other state-of-the art tools in its class.



2.3a: Raw data for ESR module.

2.3b: Clustered ESR genes.

Figure 2.3: Left panel: All yeast Emergency Stress Response (ESR) genes (vertical) in 173 microarray experiments (horizontal), sorted into 20 clusters (white separating lines). Color scale is over- or under-expression of a specific gene relative to the same gene in a reference condition. Right panel: The matrix of all pairwise information relations, sorted such that the genes belonging to the same cluster are consecutively listed. Blocks on the diagonal are thus intra-cluster similarity, and off-diagonal rectangles are inter-cluster similarity. Color scale is in bits.

Instead of discussing the results of the clustering project in detail, we emphasize that clustering is one of the simplest and scalable methods of understanding the collective behavior of a network. Consider the information matrix of Fig 2.3b as a matrix of weights between the nodes of a graph: the graph has strongly connected components that correspond to clusters (blocks on the diagonal of the information matrix) and these blocks are weakly coupled to other blocks. One might even threshold the information matrix and draw binary links in the graph whenever the similarity measure exceeds the threshold value, and some researchers have indeed taken this approach; cf. Remondini et al. (2005) for network inference by thresholding correlation coefficient values.

Clustering turns out to be an extremely powerful approach for several reasons. Firstly, in gene regulation we know that out of the whole set of genes (around 6000) for yeast, the total number of genes that regulate other genes – so called *transcription factors* – is on the order of a few percent (van Nimwegen, 2003). Although this, much smaller, group of

genes with regulatory power could conspire combinatorially and still regulate every other gene in a complicated individual fashion, many genes need to be up- or down-regulated together, because they act as enzymes in connected reaction pathways. This *coregulation* is the basis for the success of the clustering approach: coregulated genes cluster and cluster members are assumed to be regulated in identical ways by their (one or few) transcription factor(s). Instead of looking separately into regulatory regions of every single gene whose regulation we want to understand, we can look across the members of the cluster and hope that common elements in the regulatory regions will emerge above the random background sections of the genetic code (Kinney et al., 2007; Foat et al., 2006).

Although clustering is clearly productive as a first step in understanding genetic regulatory networks, it is not a generative model of the network. It reorders the nodes so that the structure (hopefully) becomes apparent, but does not give any prescription about *how* the activity of one gene influences the activity of the others – the only input to the clustering procedure is the mutual information, and we explicitly stated that information measures dependency without revealing anything about underlying functional relationships. Moreover, as we will soon see, understanding that elements σ_i and σ_j are correlated, which is the basis of clustering, tells us nothing about whether σ_i is really directly influencing σ_j;[7] in particular, in gene regulation, the genes are coregulated and are therefore coexpressed, and correlation does not imply causation or direct interaction. Despite being very practical, clustering leaves too many questions unanswered if we want to understand network behavior.

2.3 Computing interactions

Can we disentangle the mesh of correlations and separate the correlations caused by real underlying interactions from the correlations induced indirectly by other interactions, as is illustrated in Fig 2.4?

To start, we should recall a classic problem in statistical physics: we are given a lattice of Ising spins, and some specification of exchange couplings (interactions) – perhaps nearest neighbor only – and the exercise requires us to find the equilibrium correlation function between the spins, i.e. $\langle \sigma_i \sigma_{i+\Delta} \rangle$. In our case however, we will be dealing with network "reverse engineering". The exchange interactions themselves will be unknown, yet we will observe a mesh of correlations. The problem will then be to compute the exchange interactions, and the hope to find a network defined by the interactions to be *simpler* (for instance sparser) than the network of correlations.

Let us formulate the problem more precisely. The network consists of N nodes with activities σ_i, which, we will for now assume, can take on only two values, $\sigma_i \in \{-1, 1\}$.[8] Our data consist of patterns $\mathcal{D} = \left\{ \vec{\sigma}^1, \vec{\sigma}^2, \ldots, \vec{\sigma}^T \right\}$, i.e. there are a total of T simultaneous measurements of the activities at all nodes, while the network is in some stationary state. These samples can be thought of as "instantaneous" snapshots of the system or, in simulation, draws made during a Monte Carlo sampling run. From the samples we can estimate the moments at successively increasing orders: first order moments are N mean activity values, $\langle \sigma_i \rangle$; second moments are $N(N-1)/2$ correlations, $\langle \sigma_i \sigma_j \rangle$; and so on. Because the system is noisy, there will be fluctuations around the stationary state and not all T patterns are going to be equal. We expect some patterns to be more likely than the others,

[7] Let us leave "direct influence" as an intuition until the next section.

[8] This assumption will be relaxed later.

and the full description of the system as it rests near the minimum of an effective potential in an equivalent of a "thermal bath" must be contained in a joint probability distribution, $p(\sigma_1, \sigma_2, \ldots, \sigma_N)$. Getting a handle on this distribution is therefore our final goal, and as we will soon discover, computing successive approximations to it will give us the desired interactions that cause the observed correlations.

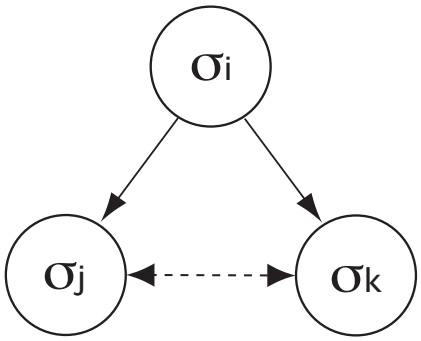


Figure 2.4: A small three-element section from a network of interacting nodes. Suppose that σ_i modulates the activity of σ_j and σ_k through some microscopic mechanism (denoted by thick lines). We can expect to observe strong correlations between σ_i and σ_j, and between σ_i and σ_k due to this direct influence. On the contrary, σ_j and σ_k are not directly coupled, but can still show significant correlation (dashed line) because of common control by σ_i.

Except for a very small number of network nodes there is no hope of directly sampling the distribution from the data. Its size grows exponentially in N and for a modest network of 10 binary nodes we would need to estimate $2^{10} \approx 1000$ parameters. To proceed, we clearly need a simplifying principle.

A commonly used procedure is called Bayesian network reconstruction (Friedman, 2004), and it is a method from the more general class of graphical models. One starts by assuming a specific (initial) factorization of the joint probability distribution over all nodes and represents it as a graph $\mathcal{G}^0$, as in Fig 2.5. Remembering that the activities are discrete variables, all conditional distributions in the factorization can be represented as probability tables with unknown entries that need to be fit from the data. Such fitting procedure can be performed in many ways, and one can evaluate the likelihood of the fit $\mathcal{L}(\mathcal{G}^n)$.[9] Of course, we have no prior knowledge of what the correct graph factorization of the initial distribution is, therefore a procedure is devised that wanders in the space of possible graph topologies and tries a likelihood on each, producing a sequence $\{\mathcal{L}(\mathcal{G}^0), \ldots, \mathcal{L}(\mathcal{G}^n), \mathcal{L}(\mathcal{G}^{n+1}), \ldots\}$.[10] The complexity of each graph, e.g. the number of links, is penalized and combined together with the fit likelihood into a scoring function. The goal is to find the factorization of the probability distribution with the best score. Presumably, we will then have discovered a simple graph that fits the data well.

There have been successful network reconstructions using this approach (Sachs et al., 2005). The key simplifying assumption that makes this approach feasible is that the graph of interactions is sparse, i.e. that there are many fewer real than potential interactions. Given such sparsity, the factorized probability distribution will have a far smaller number of unknown parameters than the full joint distribution, and there will be reasonable hope of fitting them from the data. The method allows interactions of arbitrary complexity (as many arrows converging on a single node as possible), but has some drawbacks. Firstly, there is

[9] Bayesian network reconstruction is an iterative procedure, and at n-th step, we are considering graph $\mathcal{G}^n$, hence the index.

[10] This will usually be some sort of gradient descent or simulated annealing procedure.

an exploding number of graph topologies over N elements, and no hope in exhaustively trying all of them; whatever algorithm one devises to explore the space of topologies, it can get stuck in local extrema of the scoring function. Secondly, due to computational constraints not all kinds of graphs can be explored – usually one has to exclude loops and this is a big handicap for biological systems where feedback plays a very important role. In statistical mechanics the "no loops" approximation on the lattice is the Bethe-Peierls approximation, in which one explicitly treats a chosen spin and a shell of nearest neighbors around it, while the rest of the lattice produces an effective molecular field which is determined self-consistently. Finally, because we are looking for a tradeoff between the best likelihood fit and the simplicity of the model, we have to (arbitrarily) decide how to penalize complex topologies. It is not *a priori* clear that one should simply minimize the number of links and disregard other features of the graph. In particular, we expect that for systems, in which collective effects are driven by the presence of weak interactions between lots of pairs, Bayesian method will perform poorly.

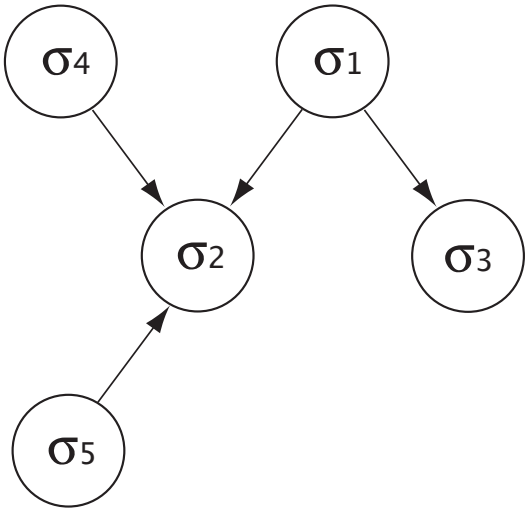


Figure 2.5: Bayesian factoring of the probability distribution over five nodes. This example graph $\mathcal{G}$ implies that the joint probability distribution can be written as follows: $p(\sigma_1, \ldots, \sigma_5) = p(\sigma_1)p(\sigma_4)p(\sigma_5)p(\sigma_3|\sigma_1)p(\sigma_2|\sigma_1, \sigma_4, \sigma_5)$.

Let us try to take a radically different route to the solution. As has been said, with a limited number of samples, T, we can successfully estimate several lowest-order moments of the distribution, for example, the means $\langle \sigma_i \rangle$ and covariances C_{ij} of Eq (2.1), or, in general, a set of mean values of some operators, $\langle \hat{O}_\mu(\vec{\sigma}) \rangle$. For any reasonable choice of the operators there is an infinite number of joint distributions over N elements with the same mean operator values. Nevertheless, there is only one distribution that also has maximum entropy, i.e. there is one distribution that is *as random as possible* but still satisfies those statistics that have been measured in an experiment (Jaynes, 1957). This is the distribution that we would like to find, and the maximum entropy principle embodies the idea that any structure (or constraint) in the distribution has to be the structure induced by the measurement (and not by explicit or hidden assumptions on our part). Formally then, we are looking for the extremum of the following functional:

$$\mathcal{L}[p(\vec{\sigma})] = S[p(\vec{\sigma})] - \sum_\mu g_\mu \langle \hat{O}_\mu(\vec{\sigma}) \rangle - \Lambda \int d\vec{\sigma}\, p(\vec{\sigma}) \tag{2.16}$$

$$= -\int d\vec{\sigma}\, p(\vec{\sigma}) \log_2 p(\vec{\sigma}) - \sum_\mu g_\mu \int d\vec{\sigma}\, p(\vec{\sigma}) \hat{O}_\mu(\vec{\sigma}) - \Lambda \int d\vec{\sigma}\, p(\vec{\sigma}). \tag{2.17}$$

The first term is the entropy of the distribution, and there are μ constraints enforced by their Lagrange multipliers g_μ:

$$\langle \hat{O}_\mu(\vec{\sigma}) \rangle_{p(\vec{\sigma})} = \langle \hat{O}_\mu \rangle_{\text{expt } \mathcal{D}}, \tag{2.18}$$

such that the average values of the operators over the sought-after distribution $p(\vec{\sigma})$ are equal to the averages over data patterns, $\mathcal{D} = \left\{ \vec{\sigma}^1, \ldots \vec{\sigma}^T \right\}$. It is easy to take the variation in Eq (2.16) and write the explicit form for the maximum entropy solution:

$$p(\vec{\sigma}) = \frac{1}{Z} \exp \left[\sum_\mu g_\mu \hat{O}_\mu(\vec{\sigma}) \right]. \tag{2.19}$$

We call Eq (2.19) the maximum entropy distribution with constraints $\hat{O}_\mu$ and we will encounter it many times in subsequent chapters. The solution has an exponential form and looks like a Boltzmann probability distribution with the factor $\beta = 1/k_B T = 1$. Indeed, in statistical mechanics, we can view the Boltzmann distribution, $p \propto \exp(-\beta H)$, as maximum entropy distribution that is constrained to reproduce the mean energy, $\langle H \rangle$. The temperature is the corresponding Lagrange multiplier. Usually we ask what the mean energy is of the system *given* some temperature:

$$\langle H(\beta) \rangle = \frac{\int d\vec{\sigma}\, H(\vec{\sigma}) e^{-\beta H(\vec{\sigma})}}{\int d\vec{\sigma}\, e^{-\beta H(\vec{\sigma})}}, \tag{2.20}$$

while in the maximum entropy modeling, we are interested the inverse question – what is the temperature that will make the expected energy of the system equal to the observed average.[11]

Operators that constrain the distribution can be arbitrary, but we can gain further insight by restricting ourselves to the moments of increasing orders (the variables are still binary for simplicity). If one chooses $\hat{O}_\mu = \sigma_\mu$, then the mean values, $\langle \sigma_i \rangle$, are constrained, and the maximum entropy distribution is the factor distribution:

$$p^{(1)}(\vec{\sigma}) = \frac{1}{Z} e^{\sum_\mu g_\mu \sigma_\mu} = \prod_\mu \frac{1}{Z_\mu} e^{g_\mu \sigma_\mu}, \tag{2.21}$$

$$Z_\mu = 2\cosh(g_\mu). \tag{2.22}$$

Mean field models look similar and it is instructive to pursue this analogy a little further. For a simple linear chain model where a spin couples to two of its nearest neighbors with exchange interactions J, the full distribution is given by:

$$p(\{\sigma_i\}) = \frac{1}{Z} \exp \left[\beta \left(h \sum_i \sigma_i + \frac{1}{2} J \sum_i \sigma_i \sigma_{i+1} \right) \right]. \tag{2.23}$$

[11] A lot of physicist's "staple" distributions are also intriguing when viewed as solutions to the maximum entropy problem. For example, the exponential distribution is the maximum entropy distribution with a constrained mean; the Gaussian distribution is a maximum entropy solution for continuous variables with constrained mean and covariance; and clearly, the uniform distribution is a non-constrained maximum entropy distribution (alternatively, the only constraint is normalization). More exotic distributions such as gamma, beta, Laplace etc. can also be seen as solutions to the constrained maximum entropy problem; things seem to be less clear for the Poisson distribution.

In the mean field approximation, we would approximate the Hamiltonian as follows:

$$H = h\sum_i \sigma_i + \frac{1}{2}J\sum_i \sigma_i\sigma_{i+1} \approx \sum_i \sigma_i \left(h + \frac{1}{2}zJ\langle\sigma\rangle\right) = \sum_i \tilde{h}_i(\langle\sigma\rangle)\sigma_i, \tag{2.24}$$

that is, we neglect correlated fluctuations and write the average of the product of spins as the product of averages (z stands for the number of nearest neighbors, $z = 2$ here). The new Hamiltonian is now one-body with effective magnetic field $\tilde{h}(\langle\sigma\rangle)$. Clearly, the mean-field Hamiltonian also makes the distribution factorize, such that

$$p^{(\mathrm{MFA})}(\vec{\sigma}) = \frac{1}{Z}e^{\sum_i \tilde{h}\sigma_i}. \tag{2.25}$$

One can easily calculate the expected value of σ_i in this model by solving

$$\langle\sigma\rangle = \tanh\left(\beta h + \frac{1}{2}\beta zJ\langle\sigma\rangle\right), \tag{2.26}$$

which is a well-known mean field equation for magnetization. Both models, namely the maximum entropy constrained by mean values and the mean field approximation, yield factorizable approximations to the joint distribution; however, for the maximum entropy problem we are given the mean values $\langle\sigma_i\rangle$, and we *exactly* reproduce them; for the mean-field approximation we are given the (microscopic) interactions and we compute *approximate* mean values by disregarding fluctuations. In general, the Lagrange multipliers of the factor maximum entropy distribution will not be the same as the effective magnetic fields of the mean field approximation; they would be equal only if the mean field approximation yielded the exact magnetization.

Returning now to the maximum entropy problem, we could continue constraining the maximum entropy distribution with correlation functions of higher and higher orders. If we were to fix both mean values and two-point correlations, the resulting distribution, Eq (2.19), would have an Ising form. Constraining the three-point correlations would induce a new term in the Hamiltonian of the form $\sum_{ijk} J_{ijk}\sigma_i\sigma_j\sigma_k$. There is clearly a "ladder," where higher and higher order constraints are imposed on the distribution, and as a result, better and better maximum entropy approximations are constructed. Let us call, then, $p^{(k)}(\vec{\sigma})$ a maximum entropy distribution consistent with correlations of order k and smaller, in line with our notation for the factor distribution, $p^{(1)}(\vec{\sigma})$. In an N-body system, the highest order of correlation is N, and $p^{(N)}(\vec{\sigma})$ must therefore be the exact joint distribution – at this order our approximation *is* the exact solution, with entropy equal to $S[p(\vec{\sigma})]$. Schneidman et al. (2003) have shown that this sequence of ever better maximum entropy approximations defines a unique decomposition of multi-information:

$$I[p(\vec{\sigma})] = \sum_{k=2}^{N} I^{(k)} \tag{2.27}$$

$$I^{(k)} = S[p^{(k-1)}(\vec{\sigma})] - S[p^{(k)}(\vec{\sigma})]. \tag{2.28}$$

In words, the *connected information of order k* is the difference of the entropies of the maximum entropy distribution consistent with correlations of order $k-1$ and one higher order. For example, connected information of the second order is the reduction of the entropy due to pairwise interactions; one creates the best factor (independent) model for the data and the best pairwise (two-body Ising) model for the data, and compares their entropies to see how much of the total structure in the joint distribution has been explained by purely pairwise terms.

2.4 Networks from interactions

Have we made any progress? In theory we could have looked forward to the scenario in which the only limitation came from the finite data size and that would put a bound on how far we could reliably sample in the space of correlations; once we collected the measured correlations, we would postulate the maximum entropy model of Eq (2.19) and solve the equations that determine all couplings, Eq (2.18). We now need only to interpret the result: since we have a generative model of the data, i.e. the probability distribution, we can calculate any new kind of average. Moreover, we can examine the couplings g_μ, conjugate to the constrained operators, and interpret these as *interactions* that cause and explain the observed correlations. For example, if the data were generated by a 1D Ising nearest-neighbor chain (but we did not know that), the correlation structure would appear complicated: each pair of spins would be correlated and correlation would have some complicated dependence on the spin-spin separation. Putting the correlation matrix into the maximum entropy calculation and reconstructing the interactions J_{ij} would, however, reveal a very simple picture – the only non-zero couplings would be between nearest neighbors and they would be all the same. We would have achieved our goal of disentangling the mesh of correlations. In (biologically realistic) situations which lack the perfect translational symmetry of an Ising chain, the mapping from correlations to interactions, $C_{ij} \to J_{ij}$, can be very non-trivial: if there is frustration, there could be interactions between pairs where there is no correlation, or vice-versa, or the signs of the correlation and interaction could be different. Nevertheless, there is only one solution for the above mapping, and it is, in principle, computable.

As is done in Bayesian network reconstruction, once we have computed the couplings and therefore the Hamiltonian that explains the data, we can draw a graphical model of the network with a link for each nonzero coupling $g_{i_1 i_2 \ldots i_l}$ connecting those l elements that are conjugate to it in the Hamiltonian.[12] These weighted links are undirected as there is generally no way of determining the "direction" of the interactions from an equilibrium model. Assumptions underlying maximum entropy reconstruction are also quite different from its Bayesian relative: whereas in the latter case we assume sparse a network of (arbitrarily complex) interactions, we assume an arbitrarily dense network of simple (low order, e.g. pairwise or triplet) interactions in the former case. To explain all $N(N-1)/2$ pairwise correlations one needs the full matrix of $N(N-1)/2$ exchange couplings J_{ij},[13] and therefore no discrete topology on the graph is assumed *a priori*. There is hence no problem of searching and scoring the space of topologies, no exclusion of graphs that include loops, and reduced dependence on the implementation details of the algorithm. The drawback is the *ab initio* exclusion of complex irreducible interactions between many nodes. Clearly, the real question to ask is about the approximation regime that is more suitable to biological systems, if a general answer exists at all.

In practice, unfortunately, the maximum entropy network reconstruction is made difficult by two problems. One is technical – solving coupling Eqs (2.18) is very hard. In essence, one needs to solve

$$\frac{\partial \log Z(\{g_\nu\})}{\partial g_\mu} = \langle \hat{O}_\mu \rangle_{\text{expt } \mathcal{D}}, \tag{2.29}$$

where Z is the partition function of the maximum entropy distribution in Eq (2.19). This

[12] Therefore, for instance, the graphical decomposition of the probability distribution plotted in Fig 2.5 would correspond to the Hamiltonian $H = \sum_i h_i \sigma_i + J_{13}\sigma_1\sigma_3 + J_{145}\sigma_1\sigma_4\sigma_5$ in the maximum entropy picture.

[13] For higher orders, there is similarly no restriction on the structure of, for example, three-point interactions, J_{ijk}.

set of equations is both nonlinear in couplings g and requires the evaluation of the partition function, $Z(\{g_\nu\})$, or effectively a complete solution of the statistical mechanics problem. The other problem concerns the identification of the nodes that are observed in the experiment. First, usually one will be able to take measurements of only a small subset of the nodes comprising the network and we will need to discuss how the hidden nodes influence models of visible nodes. Second, even if the all nodes were identified, there is an issue of "coarse-graining." Is a node with two states really an elementary, physical object that only has two states (a protein with two phosphorylation states), or is it in itself a complex with many states, but for which a two-state model might (or not) be a valid approximation? We will address both problems to some extent in the following sections that present two applications of the maximum entropy principle to network reconstruction.

2.5 Ising models for biochemical networks

The expression level of a specific gene in a regulatory network or an activation state of a specific protein in a signaling pathway depends both on the states of other interacting genes or proteins as well as on external stimuli. If the system is subjected to a wide variety of such stimuli, the patterns of correlated activity can provide us with an insight into the network structure. Recent advances in molecular biology techniques have enabled researchers to collect measurements of protein activation or gene expression levels that have not been averaged across populations of cells. Consequently, each cell selected from a pool exposed to the same condition provides an independent, simultaneous measurement of the relevant activation levels. In principle, therefore, one could attempt to construct a joint distribution of these levels at every external condition, as opposed to approaches that require the extraction of lysates from many cells and thus only allow access to the mean of the distribution. In practice, however, the data is usually badly undersampled and we are unable to explore the distribution without making prior simplyfing assumptions about its form.

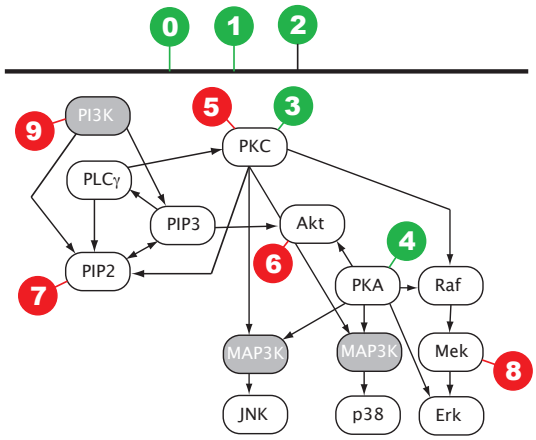


Figure 2.6: A diagram of MAP signaling network in human CD4 T-cells, reproduced and simplified from Sachs et al. (2005). Phosphorylation level of 11 white nodes was observed; red and green numbers indicate points of intervention (i.e. the change of external conditions C in which the network operates). These chemical interventions change the state of the whole network by locking the activity of the nodes on which they act into activated (green) or deactivated (red) state. Chemicals 0, 1 and 2 represent naturally occurring stimulatory agents; 0 and 1 are present in all C, while 2 is present in $C = 2$. The arrows represent experimentally verified chemical interactions; there are a number of known interactions through intermediaries that are known, but not plotted.

Here we present a maximum-entropy-based approach to biochemical network reconstruction following the steps outlined in previous sections. We assume that, given a set of N network nodes, their interactions can be well described as occuring only between pairs or perhaps triplets, and not as combinatorial interactions involving quadruplets or larger groups. Finding a distribution consistent with data that incorporates these simplifications is a well-known problem in machine learning that has a unique solution for which convergent algorithms exist (for small enough N) (Hinton and Sejnowski, 1986; Dudik et al., 2004). It has also been shown to have an appealing physical interpretation as an N-body system in thermal equilibrium whose Hamiltonian is written out in terms of one-, two-, three- etc. body potentials (Schneidman et al., 2003). These potentials parametrizing interactions at

increasing orders of complexity are our final goal, and to compute them we must be able to estimate the corresponding one-, two-, three- etc. body marginals from the data. Schneidman et al. (2006) have succesfully deployed the procedure at pairwise order to explain the observed correlations between the spiking of neurons in the vertebrate retina. Can protein-protein or genetic regulatory interactions be simplified to the same extent? Alternatively, does a failure of the pairwise model indicate where more complex interactions are dominating?

We tackle these questions on the set of 11 interacting proteins and phospholipids (jointly referred to as *biomolecules* here) in a signaling network of human primary immune system T cells. We use data from Sachs et al. (2005), where approximately 600 single-cell measurements of the activity level of biomolecules have been made for each of the 9 available conditions using flourescent cell cytometry [Methods A.1.1]. The network has been studied in detail and Fig 2.6 shows the conventionally accepted interactions, placing the observed proteins into their biological context.

A typical experiment to which maximum-entropy network reconstruction can be applied will yield a large number of simultaneous observations of N real-valued activation levels for each external stimulus. We first outline how to quantize each of the N series into Q discrete levels in a conceptually meaningful way that retains as much information as possible.[14] We illustrate the maximum entropy reconstruction by focusing first on each of the nine experimental conditions separately, and attempt to address the questions presented above. We then show how to formulate maximum entropy problem such that the network reconstruction takes advantage of all experimental conditions simultaneously; lastly, we examine if the results justify the conceptual picture of a network as a sparse graph.

2.5.1 Discovering structure in the data

We begin by examining the activation levels of single biomolecules across all conditions. When the raw activity values are histogramed, they frequently exhibit distinct peaks that correlate well with external stimuli (see Fig 2.7 for an example). This suggests that perhaps the scatter around the mean value at every condition is smaller than the spacing between mean values and hence the mean values can be proxies for what we would intuitively call discrete activity states. Moreover, if there are strong correlations between such "states" of all proteins in the system, then knowing that protein A is in state 1 should tell us something about the state of protein B. We would like to formalize this notion.

Suppose that we split the range of values taken on by the activity of protein i, x_i, into Q consecutive intervals, such that each value of x_i maps to a discrete value $\sigma_i \in \{0, \ldots, Q-1\}$. Every measurement is now denoted by a N-letter word with an alphabet of size Q, and the whole dataset is sequence of random draws from a probability distribution $p(\sigma_1, \ldots, \sigma_N|C)$, where external conditions C are chosen by the experimenters. Saying that the discretized activity levels for different proteins should be as informative of one another as possible is equivalent to making a statement that multi-information of the distribution [Eq (2.9)] marginalized over C, $I[\sum_C p(\sigma_1, \ldots, \sigma_N|C)p(C)]$, is maximized for some quantization assignment $x_i \to \sigma_i$.[15] The quantization method that finds such an assignment will be called

[14] Discretization can be regarded as a form of data compression; the original continuous data have some correlation structure, and as quantization maps the data into the discrete domain, we would like the structure to remain preserved. Note that we use the terms "quantization" and "discretization" interchangeably.

[15] Note that we need $p(C)$. This prior over experimental conditions C is chosen by the experimentalists, who collect 600 samples at each C.

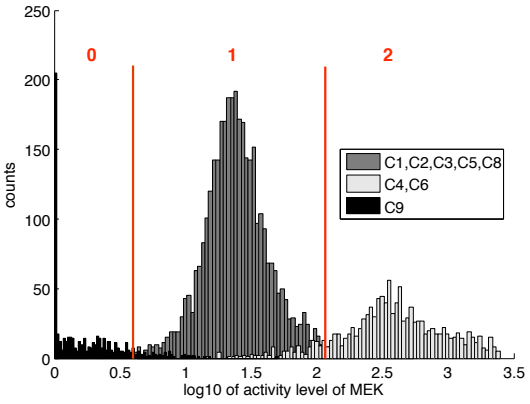


Figure 2.7: Histogram of values of the activity level of protein 2 (MEK) across all 9 experimental conditions. Three discrete activation levels can be discerned and they correlate well with the conditions (see legend). Each count in the histogram corresponds to a single cell measurement, with raw value being the light intensity proportional to the MEK activity and detected by the flow cytometer (FACS).

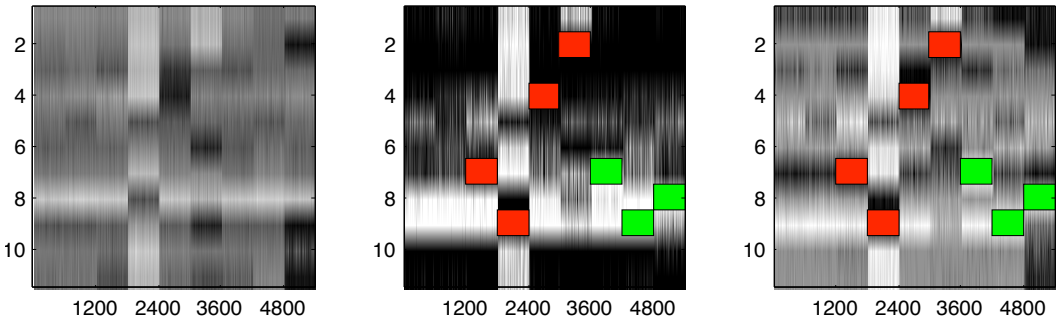


Figure 2.8: Left: original raw data for the activity level of 11 proteins (vertical axis) across different conditions (horizontal axis, 9 conditions with 600 samples each, arranged sequentially such that samples $1-600$ represent condition 1 etc). Middle: the raw data discretized into 2 levels such that the multi-information between proteins is maximized. Right: the data discretized into 3 levels such that the average mutual information is maximized [Methods A.1.2]. In experiments that involve chemical interventions [Methods A.1.1], the "red" activity level for proteins means that all samples there are forced to be deactivated (lowest state), and the "green" level means that all samples are activated (highest state). Due to the resolution of the printing process not all sample points can be seen.

maximum-multi-information quantization. We can find such assignment for $Q = 2$, because we (barely) have enough samples to calculate the entropy $S[p(\vec{\sigma})]$ of the full distribution, including small sample corrections [Methods A.1.2]. For $Q = 3$, we are completely undersampled and unable to maximize the multi-information, but we can find an assignment that maximizes the average pairwise information $\langle I^{(2)} \rangle = \langle S[p(\sigma_i, \sigma_j)] \rangle_{ij} - 2\langle S[p(\sigma_i)] \rangle_i$. Figure 2.8 shows the data discretized into two and three levels; note the high correlation between the discrete states and the external stimulus.[16]

Note that choosing a good quantization mapping is important because restricted sample size and computing power are limiting us to small number of discrete levels. We could have used the traditional quantization approach (Slonim et al., 2005a) that assigns a separate

[16]Irrespective of the way we discretize (i.e. either independently for each condition, or jointly for all conditions), there is significant information about the condition in the activity patterns of the nodes, $I(\vec{\sigma}; C)$; because of undersampling we can only estimate this number approximately to be between 1.5 and 2 bits (for reference, $\log_2 |C| = \log_2 9 \approx 3.2$ bits). Interestingly, both the pattern of fluctuations in each condition and the pattern of changes in the mean activities across conditions, reflecting two different quantization schemes, are similarly informative about the condition.

discrete level to each of the Q uniformly-partitioned quantiles of the data (i.e. each bin is equi-populated); the resulting distribution for each discretized variable is then uniform and the independent entropy, $S_{\rm ind}[p(\vec{\sigma})]$, of the joint distribution is maximized. We would, however, not have been taking into account the fact that experimental setup enforces equal priors $p(C)$ on conditions[17] and that it is therefore very likely that we see a very active or inhibited protein in only one out of all nine conditions; as a result we would recover all pairwise dependencies only when the number of quantization levels was approximately equal to the number of conditions. This is indeed what we see in Fig 2.9a, which compares the mutual information capture of quantization methods described above on the whole dataset.

In addition, the fact that we only can handle a small number of quantization levels forces us to trade dynamic range for fine-structure details of the distribution [Fig 2.9b], especially because certain interventions move some activation levels far away from their unperturbed values. We can, however, restrict ourselves to a single condition, quantize only the corresponding subset of the data, and try to infer the interactions from the local "fluctuations" around the steady state. In this case, applying either the traditional quantization or multi-information maximization quantization does not result in appreciable difference in the recovery of mutual information, indicating that the distribution at fixed condition is much less structured then the distribution of data pooled across conditions.

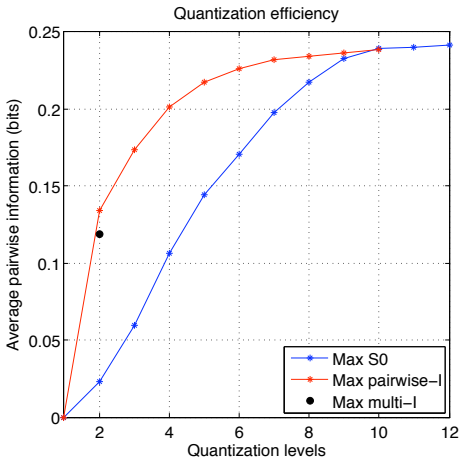
Why is there any need to discretize the data at all? First, we might have prior reasons for believing that there is a limited number of states that each node can have; for instance, the enzyme might be unphosphorylated (and thus inactive) or phosphorylated (and thus active).[18] In each condition, a fraction of proteins will be in either of the two states, and discretization takes this (and only this) gradation into account by discounting all remaining variability as noise. Furthermore, the discretized picture can capture complex behaviors of the system, e.g. the presence of multiple peaks in the joint distribution over activities, and there is evidence (at least in 2D cross-sections of the joint distribution that we can sample, e.g. Fig 2.9b) that there are such structures in the data. Conversely, we know that a maximum entropy distribution over N *continuous* variables with constraints on the first and second moments is a Gaussian, which is fully specified by the mean and covariance matrix. Moreover, the constraining covariance matrix is also (inverse of) the matrix that appears in the exponential form of the maximum entropy solution [Eq (2.19)] and the mapping from correlations to interactions is trivial: $C_{ij} \to J_{ij} \equiv J_{ij} = C_{ij}^{-1}$; multi-dimensional Gaussian is easy to understand and is a function with one peak, but unfortunately is not rich enough to describe the data.[19]

Having reduced the data to the binary representation in the most informative way

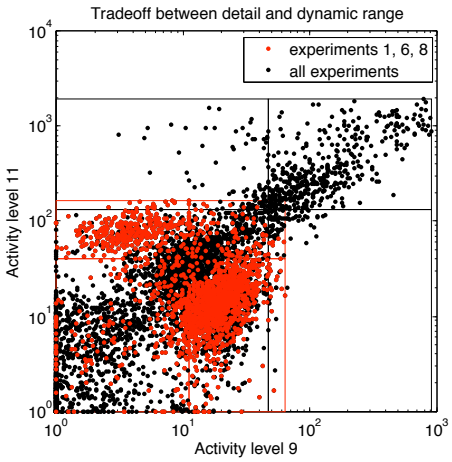
[17] These priors simply reflect the number of samples at each condition that get measured and have no relation to the set of conditions a typical cell experiences during its lifetime; this is why the distribution of all activation levels pooled over different stimuli is *not* the *natural* distribution of expression levels. The same reasoning holds in gene expression arrays: experimentally induced conditions which the yeast is exposed to are not a properly weighted ensemble of conditions that yeast sees during its life – they might reveal strong correlations useful for clustering input, but cannot be used to build a "natural" probability distribution of expression levels.

[18] Equivalently, one could have phosphorylation or some other chemical modification in multiple locations on the protein which would induce more than two levels of activity. There is ample evidence for such behavior in MAP signaling cascades (Kolch, 2000).

[19] A popular continuous model often assumed in machine learning that can account for multi-peaked structure in the data is the mixture-of-Gaussians. An interesting project for future work is the exploration of an alternative model, where one uses the maximum entropy models of real-valued activation levels, which constrain 1-D marginal histograms to capture their non-Gaussian distributions, and the two-point correlations $\langle x_i x_j \rangle$.



2.9a: Information capture by various discretization methods.



2.9b: Dynamic range and discretization.

Figure 2.9: Left panel: mutual information capture using various discretization methods as a function of the number of discrete states Q, data from all conditions is being discretized simultaneously. Blue line puts an equal number of samples into each of the Q bins (and thus maximizes S_0, the independent entropy of the distribution), while the red line greedily chooses binning boundaries so that the average mutual information between all pairs, $\langle I(\sigma_i;\sigma_j)\rangle_{ij}$, is maximized. As mentioned in the text, the blue quantization needs $Q \approx 9$ levels to fully capture all pairwise relations. For $Q = 2$ we are able to maximize the total multi-information, $I[p(\vec{\sigma})]$, and the result is shown as the black dot. The black dot has slightly smaller average pairwise information, but bigger total multi-information than either blue or red line; this is the binary quantization used in the subsequent analysis. Right panel: scatter plot of activity levels of proteins 9 and 11 across all (black) and certain subset (red) of conditions (see legend). If binary quantization is performed, then optimally quantizing over the whole dataset will set the bin boundaries at thin black lines and the "low" and "high" states of both proteins will appear positively correlated. However, this will ignore the red substructure. If, instead, only the red data subset is optimally quantized (and bin boundaries are at thin red lines), then the data will appear anti-correlated and we will have revealed the previously hidden substructure that, in the whole-data-quantization, was all lumped together in the "low/low" state. Note also that continuous maximum entropy models constrained by pairwise correlations (Gaussian distributions) cannot be used to model the multi-peaked structure seen in this example cross-section.

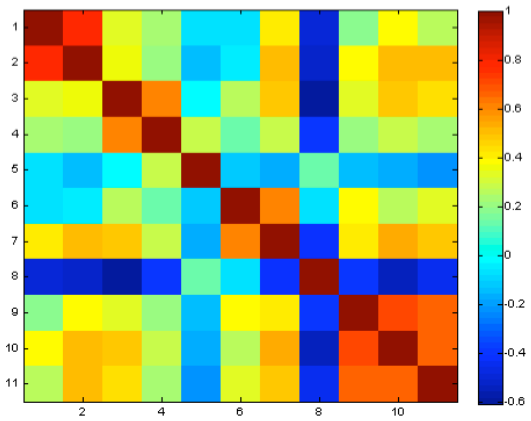


Figure 2.10: Correlation coefficient between log activity of 11 protein activation levels across experimental conditions. Almost all pairs appear statistically significantly correlated (the largest error bar on the correlation coefficient is $\approx 2 \cdot 10^{-2}$).

possible, we now turn our attention to pairwise correlations and show in Fig 2.10 the matrix of correlation coefficients between protein activation levels. The majority of pairs show strong correlations, and even weaker correlations are statistically significant. Based on the correlations alone the network would thus seem to be densely connected – but how does the connectivity map look in terms of exchange interactions J_{ij} of the maximum entropy distribution, Eq (2.19) and Eq (2.30)? We will approach the question systematically: firstly, we will examine data at each condition separately; then we will find a way to model multiple conditions at the same time.

2.5.2 Analyzing a single condition

The data collected for conditions 1 and 2 describe the activation levels of 11 biomolecules when the cells are exposed to their natural stimulatory signals. If we focus on each of the two conditions separately, we will be dealing with draws from two stationary distributions: applying the max-multi-info quantization discussed in previous section separately to each condition will produce binary words that represent fluctuations around the steady state in that condition. Because the nodes are functionally connected, the fluctuations are not independent, and must reflect local couplings between nodes near the given steady state. Can we learn something from the correlated fluctuations in the activities?

Having quantized the data into two levels and calculated the correlations and mean values, we write down the form of maximum entropy distribution consistent with these operators,[20] Eq (2.19):

$$p(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_N) = \frac{1}{Z} \exp\left\{ \sum_i h_i \tilde{\sigma}_i + \frac{1}{2} \sum_{ij} J_{ij} \tilde{\sigma}_i \tilde{\sigma}_j \right\}. \tag{2.30}$$

We proceed to calculate the *interaction map* J_{ij} and the magnetic fields h_i that explain the measured observables [Eq 2.29, Methods A.1.3].

Figure 2.11 shows interaction maps J_{ij} and magnetic fields for each condition's data quantized and analyzed separately. Interestingly, both condition 1 and 2 exhibit a similar pattern of interactions, with those of condition 1 being a subset of condition 2; moreover they

[20] Let's denote by a tilde over the activity variable, $\tilde{\sigma}$, the Ising model convention of naming two states -1 and 1; the corresponding states of σ are 0 and 1.

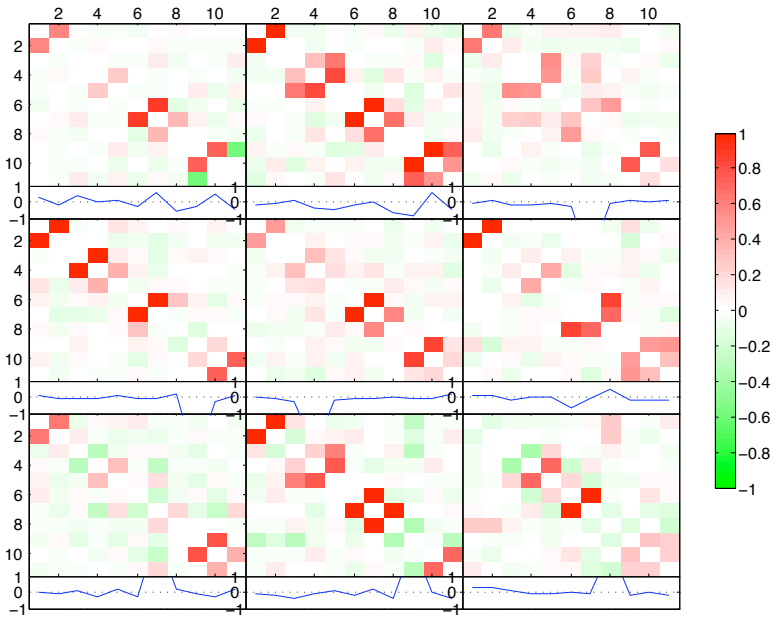
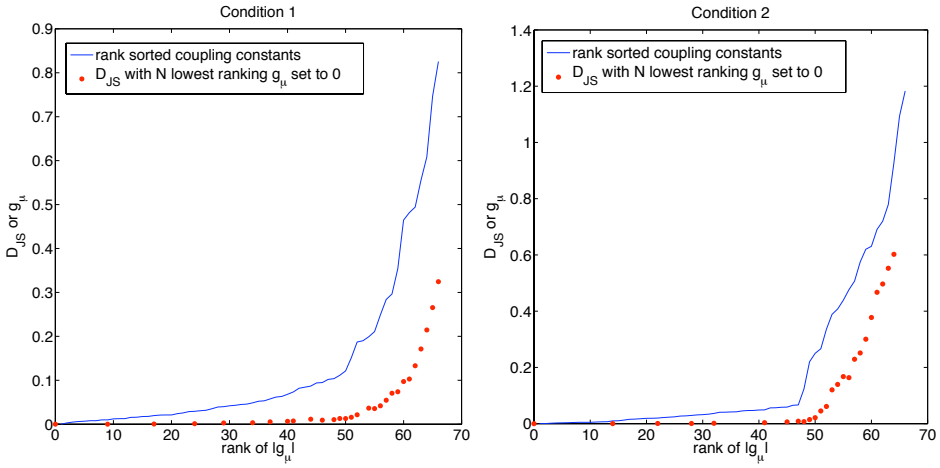


Figure 2.11: Interactions (color map) and magnetic fields (blue line) for all 9 external conditions, proceeding top down, left to right, computed (both quantization and maximum entropy reconstruction) separately for each condition. All interactions J_{ij} are drawn on the same scale, with red color indicating positive and green color indicating negative couplings. Note that since the data is requantized in each condition, and requantization amounts to the change in single-body marginals (averages) which are constrained by the magnetic fields, the magnetic fields are in this case a "side-effect" of the data analysis procedure. Conditions 1 and 2 represent cells exposed to the naturally occurring stimulatory chemical signals; other conditions represent environments where "intervention" chemicals – which are supposed to lock the activity states of certain nodes to either "on" or "off" – have been added to the stimulatory chemicals of condition 1.

also agree with the conventional map of interactions in Fig 2.6, except for the interaction between 10 and 11 (p38, JNK) in condition 2. A possible explanation for this interaction is the cross-talk in the MAPKK pathway upstream of p38 and JNK: unobserved biomolecules that couple pairs of observed proteins would induce effective interactions between them. In general, the interaction matrices are sparse, and most of the small coupling constants can be set to zero with minimal change to the distribution, as shown in Figs 2.12a and 2.12b.[21]



2.12a: Condition 1.

2.12b: Condition 2.

Figure 2.12: Sparsity of the interaction maps for both condition 1 (left) and condition 2 (right). Blue line plots the rank-ordered magnitudes of the couplings (both magnetic fields and exchange interactions) g_μ. Smallest n couplings of the original Ising model inferred from the data are set to zero and the resulting "pruned" Ising model is compared to the original. For this comparison we plot the Jensen-Shannon divergence as a function of n (red dots). For both conditions we see a sharp bend at first 40-50 couplings that can be zeroed at low D_{JS}; the remaining 11 magnetic fields and ~ 10 pairwise couplings are significant.

Note again that we are looking only at fluctuations around a naturally stimulated steady state. These fluctuations are much smaller then those induced by intervening chemicals, which is presumably why we detect only a subset of full interactions.

How much of the complexity of the true distribution is captured by the maximum entropy approximation? To answer this question we look at the fraction of the multi-information of the real distribution that is captured by the pairwise model. As Fig 2.13 demonstrates, in case of condition 2 it recovers almost all of the 2.8 bits of total information; for condition 1, however, the fraction is around 70 percent out of the total of 1.5 bits.[22] A further test of the pairwise model involves comparing the predictions about connected three-point correlations $\langle(\sigma_i - \bar{\sigma}_i)(\sigma_j - \bar{\sigma}_j)(\sigma_k - \bar{\sigma}_k)\rangle$ with values estimated from the data, as shown in Fig 2.14. As expected, the match between predictions and measurement is

[21] One starts with the smallest couplings and proceeds towards bigger ones by setting them to zero and calculating the Jensen-Shannon distance between such "pruned" and the original distributions. For conditions 1 and 2, if all exchange interactions but for the "skeleton" around the diagonal are set to 0, the Jensen-Shannon distance will be around 0.015, i.e. one would need on the order of 70 samples to distinguish the full maximum entropy from the pruned distribution.

[22] There might be larger systematic errorbars on experiments 1, 6 and 9 because the distribution seems considerably more uniform than for other experiments and we are low on samples.

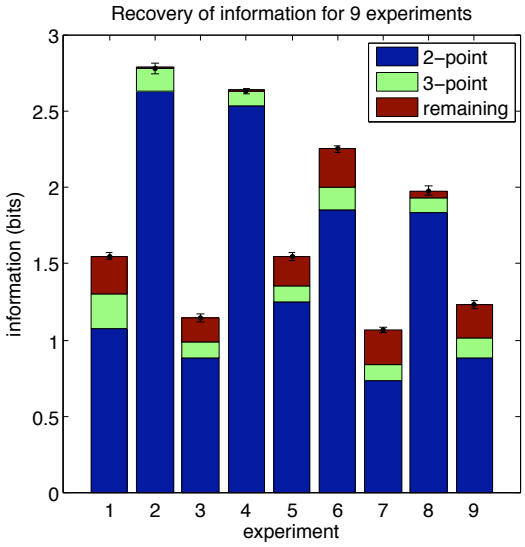


Figure 2.13: How much of the total multi-information in the distribution of activation levels does the pairwise model capture? For each of the 9 conditions on the horizontal axis, the data is discretized using max-multi-info method into 2 levels and a pairwise (or three-body) Ising model is constructed. The bar height is the total multi-information of the distribution [Eq (2.7)]. The blue (green) segment represents the information of the second (third) order, $I^{(2,3)}[p(\vec{\sigma})]$, of Eq (2.28). The error bars are entropy estimation errors from the *direct* estimation obtained by 100 repeated reestimations (Slonim et al., 2005a).

good in condition 2 (not shown), while for condition 1 we see a single three-point predicted correlation deviating strongly from its measured value. The corresponding biomolecules are 2, 3 and 4, namely PLCγ, PIP2 and PIP3, and they are suspected to form a feedback loop [Fig 2.6]. To ascertain that it is not only the observed correlation, but actually a true triplet interaction between the molecules that generates the discrepancy, we build a new maximum-entropy model consistent with three-point marginals. The corresponding Hamiltonian has the following form:

$$H = -\sum_i h_i \tilde{\sigma}_i - \frac{1}{2}\sum_{ij} J_{ij}\tilde{\sigma}_i\tilde{\sigma}_j - \frac{1}{6}\sum_{ijk} J_{ijk}\tilde{\sigma}_i\tilde{\sigma}_j\tilde{\sigma}_k \tag{2.31}$$

In the generalized Hamiltonian of Eq (2.31) the largest three-point interaction term is J_{345}. Moreover, in order to convincingly show that it really is J_{345} that fixes the offending three-point correlation (as opposed to all other triplet degrees of freedom in Eq (2.31)), we construct yet another maximum entropy model: a pairwise Ising system that constrains exactly one three-point marginal, $p(\sigma_3, \sigma_4, \sigma_5)$, and has a single three-point coupling, J_{345}. The agreement between prediction and observations is then restored up to third-order in correlations, at the cost of one additional underlying interaction. Experimentally it is also known that PLCγ hydrolyses its substrate PIP2 to produce PIP3; furthermore it is suspected that PIP3 can recruit PLCγ (Goodridge and Harnett, 2005; Kolch, 2000).

We believe that the described procedure generalizes. The theoretical foundation (Schneidman et al., 2003) provides a way of decomposing the total information of a given distribution into a sum of positive terms, each of which indicates the extent to which maximum-entropy models incorporating successively higher order marginals recover the total complexity. A failure to account for the total information with a simple model is diagnostic of complexity being unaccounted for in the model; to pinpoint the problem, one compares the prediction and measurement of next order correlations; hopefully, the failure is localized and not distributed through the network. If this is the case and fixing the failure requires the introduction of a single new interaction, we might believe that we have learned something new about the system.

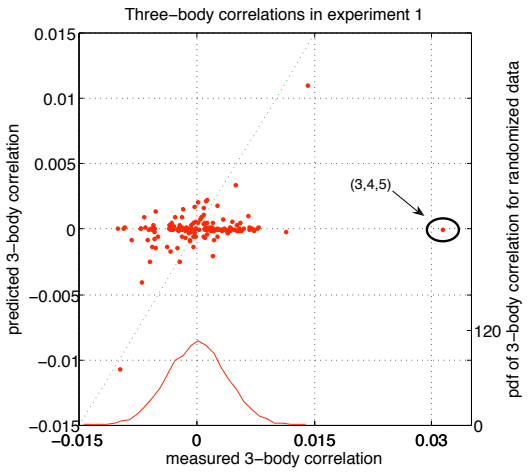


Figure 2.14: Comparison between three-point connected correlations measured from the data (horizontal axis) and predicted by the pairwise Ising model (vertical axis). Due to small sample effects, there are error bars in the correlation estimates; the distribution in red is the distribution of the three-point connected correlations in a shuffled dataset. We see that three-point correlations are small with the exception of several elements, and all but one agree with the model prediction. The mismatch is the three-body correlation between biomolecules 3, 4 and 5 (see main text).

No coupling	$H_0 = -\sum_i h_i \sigma_i - \frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j$
Coupled magnetic fields	$H_1 = H_0 - \frac{1}{2} \sum_{ik} h_i^{(k)} \sigma_i y_k$
Coupled fields and interactions	$H_2 = H_1 - \frac{1}{6} \sum_{ijk} J_{ij}^{(k)} \sigma_i \sigma_j y_k$

Table 2.1: Different models of incorporating conditional dependence on the external stimuli. The Ising model of Eq (2.30) is complemented by binary variables y_k that are "on" in condition k and "off" otherwise and therefore mimic the presence or absence of intervening chemicals. Dependence on conditions is achieved by coupling the 11 internal nodes σ_i to the condition nodes y_k; the coupling constants, when computed using the maximum entropy network reconstruction, will then parametrize the *condition-dependent magnetic fields* $h_i^{(k)}$ in H_1 and, additionally, *condition-dependent exchange interactions* $J_{ij}^{(k)}$ in H_2.

2.5.3 Combining multiple conditions

To extend the analysis to conditions of chemical intervention, we need to formulate the maximum entropy problem such that it will constitute a proper description of $p(\sigma_1, \ldots, \sigma_N | C)$ for various conditions C. It is clear that we cannot proceed without further assumptions about how a change in condition affects the system. Conceptually, we can view inhibition and activation chemicals that are added to the cell culture to induce the change in network behavior on the same footing as 11 measured proteins – the fact that these chemicals are "external" to the cell, while the flourescently-tagged ones are natural and "internal" is of no consequence for the method. We therefore extend the support of the probability distribution $\{\sigma_1, \ldots, \sigma_N\}$ with external variables $C = \{y_1, \ldots, y_K\}$ that speficy the condition, such that the concatenated dataset defines a probability distribution $p(\sigma_1, \ldots, \sigma_N, y_1, \ldots, y_K)$. For each measured sample in condition C_i we know what values to assign to variables y_i: $y_i = 1$ if during C_i a given intervention chemical was present, and -1 otherwise [Methods A.1.1]. Introduction of the new variables allows us to make our intuition about the influence of the condition C on parts of the network precise: we couple the terms in the Ising model of Eq (2.30) that are assumed to vary upon change in C, to variables y. The immediate cases of interest are the described in Table 2.1 and below.

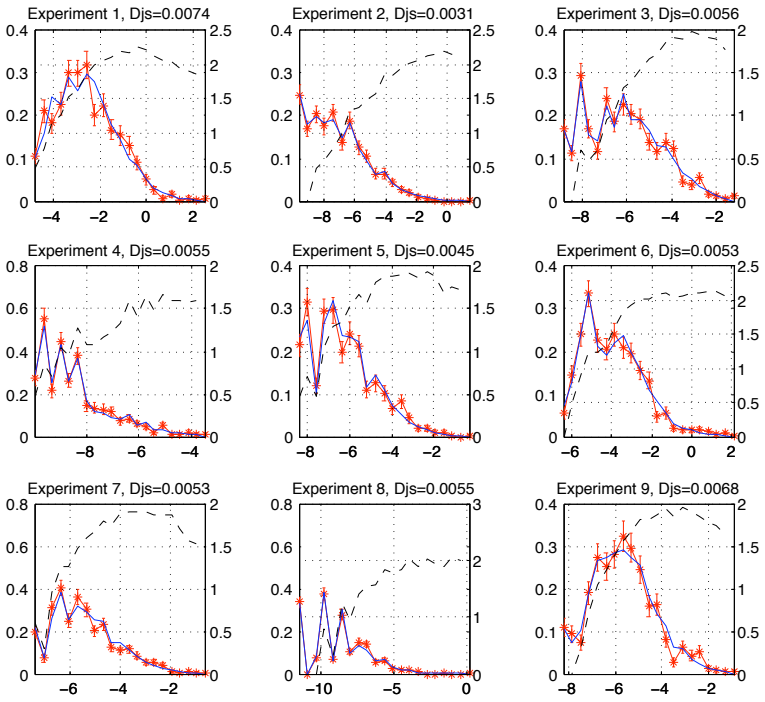


Figure 2.15: Once the maximum entropy model has been solved from the correlations and the Hamiltonian is known, we can calculate the energy histograms over the real data samples (red) and compare them to the Boltzmann density of states prediction (blue), shown for each condition separately. Black dashed line (on right vertical axis) shows log base 10 of the number of possible *distinct* patterns in a given energy bin. Condition 1 has the worst agreement with the pairwise prediction, specifically in the states with the low energy.

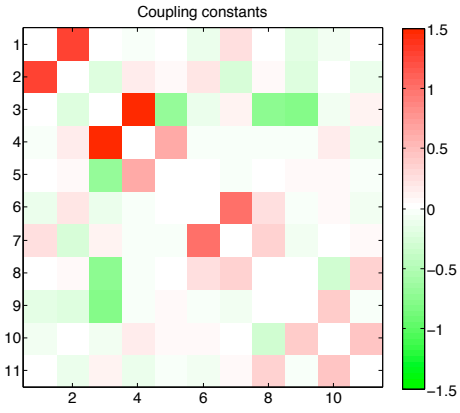
To construct a distribution that corresponds to a Hamiltonian with no y-couplings, we constrain only the pairwise marginals of 11 observed biomolecules, quantized together to maximize the multi-information. This is equivalent to pooling all data together and building a single model to account for all conditions, an approach that should perform rather disastrously.[23] On the other extreme there is the Hamiltonian H_2 of Table 2.1 that couples both J and h terms to y variables and corresponds to maximum-entropy distribution that constrains all three-point marginals of the form $p(\sigma_i, \sigma_j, y_k)$ – this is the model with maximum number of free parameters, where changing a condition can completely change both the interactions J and magnetic fields h. The middle ground is the model that assumes that the interaction map of J is independent of the condition, and all inter-condition variation is subsumed into $h_i^{(k)} y_k \sigma_i$ terms that set the strength of influence of external chemical k on the protein activation level σ_i. When the chemical k is present, y_k is always 1, and the condition-independent Hamiltonian H_0 magnetic fields are transformed by $h_i \to h_i + h_i^{(k)}$. The equivalent maximum entropy problem constrains marginals of the form $p(\sigma_i, \sigma_j)$ and $p(\sigma_i, y_k)$.[24] The assumptions of this model are consistent to what is thought to occur biologically, that is, that the intervening chemicals change the activation state of the signaling proteins, but do not affect the nature of biochemical interactions between them.

Figure 2.16a shows the results for the model H_1 of Table 2.1 in which external variables couple to magnetic fields. Data has been quantized into binary levels to maximize the multi-information and all 9 conditions have been used. Because we chose a discrete representation in which certain activation states correlate strongly with external conditions, only a small number (~ 10) of patterns have considerable weight in each condition, and we can plot their measured against the predicted frequency for each condition separately [Fig 2.17]. Despite only having 600 samples for each plot, the agreement is reasonable with average Jensen-Shannon divergence of 0.036 (not corrected for sample size). The interaction matrix is still relatively sparse and captures more of the expected interactions, especially between JNK, p38 and other proteins. There are missing links between PKA, PKC and RAF and MEK. Histograms of MEK [Fig 2.7] and to some extent PKA exhibit three activation levels that cannot be reflected in the binary quantization, and we would hope to recover the interactions using a finer quantization.

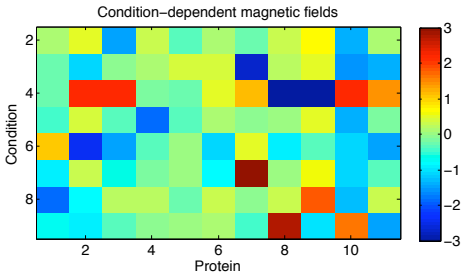
Condition-dependent magnetic fields $h_i^{(k)}$ are, for each intervention condition except $C4$, in agreement with expectations – the field that couples to the perturbed chemical has the largest value and the correct sign (first 4 interventions are inhibitions while the other 3 are activations). However, each intervention also perturbs to some extent other activation levels and, moreover, a model in which each y_k couples *only* to the σ_i that it is supposed to influence, produces much higher divergence values (data not shown). Note that the intervention in condition 4 specifically seems to affect activation levels other than that of PKC.

[23] Which it does. The Ising model description of the activation levels pooled over conditions is good; however the conditional distributions, $p(\vec{\sigma}|C)$, are very different from condition to condition, with $\langle D_{JS} \rangle_C \approx 0.35$, and one cannot expect that one distribution is a good model for all C.

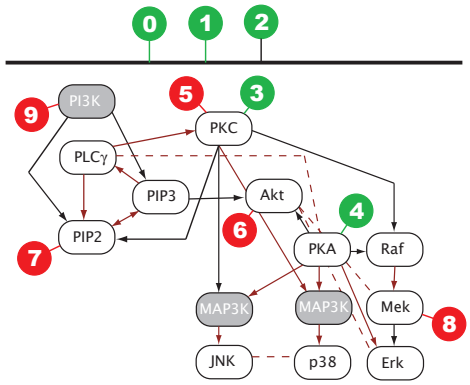
[24] Note that in constraining two-point marginals $p(\sigma_i, y_k)$ we implicitly constrain one-point marginals $p(\sigma_i)$, yielding h_i, corresponding to the magnetic fields in H_0, and $p(y_k)$, yielding magnetic fields for y variables. Values of h_i are reflective of the quantization (e.g. in case of equi-populated bins h_i=0, as there is no inherent 'bias' for σ_i being in a -1 or $+1$ state). Values of magnetic fields for y_k variables recount the priors $p(C)$ over conditions, and as such only carry information about the experimental setup and not the biological system itself.



2.16a: The interaction map for the network of 11 signaling proteins.



2.16b: Condition dependent magnetic fields.



2.16c: Thresholded interaction map for 11 signaling proteins.

Figure 2.16: Exchange interactions (left) and condition dependent magnetic fields (right) using the Hamiltonian H_1 from Table 2.1. The interaction map is fixed from condition to condition, but the magnetic fields change; the inferred condition-dependent magnetic fields mostly match the expectation that the biggest value in condition C couples to the protein that has been influenced by the chemical intervention in condition C (interventions are protein 7 in condition 3 and 7, 9 in condition 4 and 8, 4 in condition 5, 2 in condition 6 and 8 in condition 9) [Fig 2.6, Methods A.1.1]. Comparison between known map of protein-protein interactions and interaction map J_{ij} thresholded at 0.25 to define links (bottom): brown solid line – known map and reconstructed interactions overlap; brown dashed line – link is predicted by the maximum entropy model; black line – link is not predicted by the maximum entropy model.

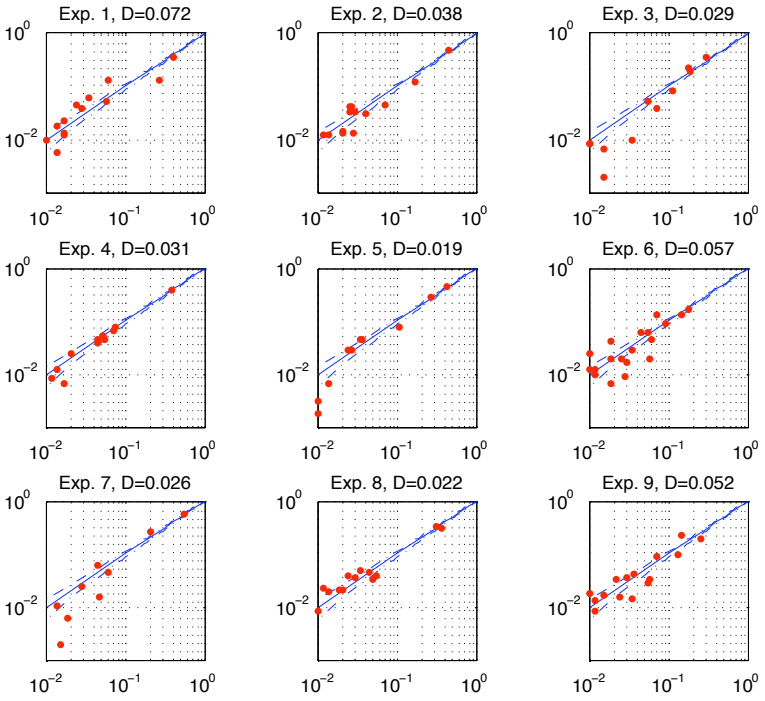


Figure 2.17: Observed and expected frequencies of activity states of 11 proteins over 9 conditions (top-to-bottom, left-to-right). With 11 proteins, there are $2^{11} = 2048$ possible activation states of the network. Because the whole dataset is quantized into binary symbols at the same time using max-multi-info scheme, each condition is dominated by a small number of frequent binary patterns (while most of the others have zero or near zero frequency), allowing us to get estimates for their frequencies from only 600 data points per condition. Observed frequency on horizontal, and Ising prediction using H_1 from Table 2.1 on vertical axis; equality with counting error shown in blue.

2.5.4 Discussion

Several lessons can be learned from our first application of the maximum entropy to the real biological dataset. On the technical side, we were lucky to have dealt with less than 20 binary variables and thus had exact algorithms for finding the interactions. On the other hand, the sample size was severely limiting. There were two clear ways of improving the fidelity of our models (either by increasing the number of discretization levels Q [Methods A.1.3] or by increasing the order of correlations captured) but we could not take either while remaining convinced that small-sample systematics would not be corrupting the results. Only $Q = 2$ models are therefore presented here.

On the modeling side, we devised a method for quantizing the data that retains as much information in the distribution as possible, a huge improvement over naive quantization schemes with equi-populated bins [Fig 2.9a]. We saw that having a limited dynamic range in terms of discrete levels can induce unwanted effects by missing finer structure in the data [Fig 2.9b]. Nonetheless, we could verify that our results do not qualitatively depend on the quantization by choosing random quantization schemes and performing a reconstruction on each [Fig A.1a]. Finally, we showed how to make the model parameters dependent on external conditions in a controllable way.

The data are examined in two ways. In the analysis of a single condition, one observes the fluctuations around the steady state and tries to infer the network structure from correlated fluctuations, independently from condition to condition. We learned that pairwise models here capture most of the information, and where they fail, they fail so as to suggest how the model has to be modified to correct for the failure, in particular with the addition of a three-point interaction in Fig 2.14. Because of small data size it was impossible to directly sample the distribution, but there are projections from the pattern space into a space of lower dimensions, e.g. energy in Fig 2.15, that we can use to compare data with predictions reasonably well. The interaction maps across 9 conditions look very similar: they are sparse and there exists a "skeleton" structure of common interactions across conditions which nicely corresponds to what is known from previous work (more precisely, the single-condition interaction maps of conditions 1 and 2 are a subset of microscopic interactions in Fig 2.6). We understand that if an interaction is seen in one condition, it could disappear in the other (e.g. when that activation level is at saturation in one of the two conditions); having the interaction change the sign (for example in Fig 2.11, conditions 1 and 2, proteins 10 and 11) is harder to understand and could point to a general difficulty with network reconstruction, namely the presence of hidden nodes. The interaction that changes sign could be "renormalized" between the two conditions by being coupled to a node that we do not observe and that has a different value in both conditions. Such a possibility nicely illustrates the point that our interactions are *effective and phenomenological* and not necessarily aligned with the microscopic reactions of phosphorylation and dephosphorylation. There might be cross-talk between the nodes, or common intermediary chemicals in the pathway (see Fig 2.6 pathways above p38 and JNK, for instance), or hidden nodes; or perhaps one chemical has many *different* phosphorylation states. In all these cases we can expect the reconstructed exchange interactions to deviate from the microscopic picture while still being a good model for the data, and therefore comparison with the known and verified arrows in the interaction diagram of Fig 2.6 cannot be the gold standard of validity. The agreement is nevertheless good, presumably because the network *is* sparse and seems mostly non-frustrated.

Various conditions induce very different activation patterns and the chemical interven-

tions are *not* small perturbations. The separation between the mean values of activation levels can be similar or even larger than their spread around the mean [Fig 2.7]; although we miss some of the structure by quantizing in only $Q = 2$ levels, this is the best that we can do. Then, according to the prescription of Table 2.1, we can build a maximum entropy model that incorporates the external, perturbing, chemicals on the same footing as the internal chemicals, and assume that the presence of an inhibitory or excitatory drug will change the activation level of its target protein, but will leave the interaction structure intact. Under such assumptions the maximum entropy yields one interaction map in addition to condition-dependent magnetic fields [Fig 2.16a] and the model accounts reasonably well for the data [Fig 2.17] (see divergence values at various conditions). Because the chemical perturbations are very strong and the dynamic range of quantization limited, there is only a small number of distinct patterns with nonzero probability in the data. The interaction map has more structure now, which is close but not the same as the "conventional wisdom" for the interactions in the MAP system. As a task for the future, it would be interesting to see how the network reconstruction performs on a simulated system, where one assumes the reaction equations and a noise model for a certain number of microscopic processes and reconstructs back the interaction maps from the simulated data.

Our second example will deal with the structure of responses of the retinal ganglion cells when they are shown a naturalistic movie clip. The problem of data quantization will fortunately be absent, because there will exist a natural quantization into 2 bins (whether the neuron spikes or does not), and we will have two orders of magnitude more data to analyze. We will nevertheless face the technical challenges related to computing maximum entropy solutions on more than 20 nodes, and conceptual difficulties because we will have good reasons to believe that the set of observed nodes is much smaller than the set of truly interacting nodes.

2.6 Ising models for networks of real neurons[25]

Physicists have long explored analogies between the statistical mechanics of Ising models and the functional dynamics of neural networks (Hopfield, 1982; Amit, 1999). Recently it has been suggested that this analogy can be turned into a precise mapping (Schneidman et al., 2006): In small windows of time, a single neuron i either does ($\sigma_i = +1$) or does not ($\sigma_i = -1$) generate an action potential or "spike" (Rieke et al., 1997); if we measure the mean probability of spiking for each cell ($\langle\sigma_i\rangle$) and the correlations between pairs of cells ($C_{ij} = \langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle$), then the maximum entropy model consistent with these data is *exactly* the Ising model

$$P(\{\sigma_i\}) = \frac{1}{Z}\exp\left[\sum_{i=1}^{N} h_i\sigma_i + \frac{1}{2}\sum_{i\neq j}^{N} J_{ij}\sigma_i\sigma_j\right], \tag{2.32}$$

where the magnetic fields $\{h_i\}$ and the exchange couplings $\{J_{ij}\}$ have to be set to reproduce the measured values of $\{\langle\sigma_i\rangle\}$ and $\{C_{ij}\}$. We recall that maximum entropy models are the least structured models consistent with known expectation values (Jaynes, 1957; Schneidman et al., 2003); thus the Ising model is the minimal model forced upon us by measurements of mean spike probabilities and pairwise correlations.

The surprising result of Schneidman et al. (2006) is that the Ising model provides a very accurate description of the combinatorial patterns of spiking and silence in retinal ganglion cells as they respond to natural movies, despite the fact that the model explicitly discards all higher order interactions among multiple cells. This detailed comparison of theory and experiment was done for groups of $N \sim 10$ neurons, which are small enough that the full distribution $P(\{\sigma_i\})$ can be sampled experimentally. Here we extend these results to $N = 40$, and then argue that the observed network is typical of an ensemble out of which we can construct larger networks. Remarkably, these larger networks seem to be poised very close to a critical point, and exhibit other collective behaviors which should become visible in the next generation of experiments.

To be concrete, we consider the salamander retina responding to naturalistic movie clips, as in the experiments of Schneidman et al. (2006); Puchalla et al. (2005). Under these conditions, pairs of cells within $\sim 200\,\mu$m of each other have correlations drawn from a homogeneous distribution; the correlations decline at larger distance.[26] This correlated patch contains $N \sim 200$ neurons, of which we record from $N = 40$;[27] experiments typically run for ~ 1 hr.[28]

The central problem is to find the magnetic fields and exchange interactions that reproduce the observed pairwise correlations. It is convenient to think of this problem more generally: We have a set of operators $\hat{O}_\mu(\{\sigma_i\})$ on the state of the system, and we consider

[25] This section appeared on the arXiv as Tkačik et al. (2006).

[26] The correlations between pairs are drawn from a single distribution that depends neither on the location nor on the separation between the neurons, as long as they are within $\sim 200\,\mu$m. These are approximate statements; for details see Puchalla et al. (2005).

[27] Alternative recording methods can capture more cells at low density (Mathieson et al., 2004), or fewer cells at higher density (Segev et al., 2004).

[28] Some experimental details (Schneidman et al., 2006): The visual stimulus consists of a 26.2 s movie that was projected onto the retina 145 times in succession; using $\Delta\tau = 20$ ms quantization this yields 1310 samples per movie repeat, for a total of 189950 samples. The effective number of independent samples is smaller because of correlations across time; using bootstrap error analysis we estimate $N_{\rm samp} \sim 7\cdot 10^4$ [Methods A.2.1].

a class of models

$$P(\{\sigma_\mathrm{i}\}|\mathbf{g}) = \frac{1}{Z(\mathbf{g})} \exp\left[\sum_{\mu=1}^{K} g_\mu \hat{O}_\mu(\{\sigma_\mathrm{i}\})\right]; \tag{2.33}$$

our problem is to find the coupling constants $\mathbf{g}$ that generate the correct expectation values, which is equivalent to solving the equations $\partial \ln Z(\mathbf{g})/\partial g_\mu = \langle \hat{O}_\mu(\{\sigma_\mathrm{i}\})\rangle_\mathrm{expt}$. Up to $N \sim 20$ cells we can solve exactly, but this approach does not scale to $N = 40$ and beyond. For larger systems, this "inverse Ising problem" or Boltzmann machine learning, as it is known in computer science (Hinton and Sejnowski, 1986), is a hard computational problem rarely encountered in physics, where we usually compute properties of the system given a known model of the interactions.

Given a set of coupling constants $\mathbf{g}$, we can estimate the expectation values $\langle \hat{O}_\mu \rangle_\mathbf{g}$ by Monte Carlo simulation. Increasing the coupling g_μ will increase the expectation value $\langle \hat{O}_\mu \rangle$, so a plausible algorithm for learning $\mathbf{g}$ is to increase each g_μ in proportion to the deviation of $\langle \hat{O}_\mu \rangle$ (as estimated by Monte Carlo) from its target value (as estimated from experiment). This is not a true gradient ascent, since changing g_μ has an impact on operators $\langle \hat{O}_{\nu\neq\mu} \rangle$, but such an iteration scheme does have the correct fixed points; heuristic improvements include a slowing of the learning rate with time and the addition of some 'inertia', so that we update g_μ according to

$$\Delta g_\mu(t+1) = -\eta(t)\left[\langle \hat{O}_\mu \rangle_{\mathbf{g}(t)} - \langle \hat{O}_\mu \rangle_\mathrm{expt}\right] + \alpha \Delta g_\mu(t), \tag{2.34}$$

where $\eta(t)$ is the time–dependent learning rate and α measures the strength of the inertial term.[29]

Figure 2.18 shows the success of the learning algorithm by comparing the measured pairwise correlations to those computed from the inferred Ising model for 40 neurons [Methods A.2.2]. To verify that the pairwise Hamiltonian captures essential features of the data, we predict and then check statistics that are sensitive to higher order structure: the probability $P(K)$ of patterns with K simultaneous spikes, connected triplet correlations and the distribution of energies [Methods A.2.3]. The model overestimates the significant 3–point correlations by about 7% and generates small deviations in $P(K)$; most notably it underestimates the no–spike pattern, $P_\mathrm{expt}(K=0) = 0.550$ vs. $P_\mathrm{Ising}(K=0) = 0.502$. These deviations are small, however, and it seems fair to conclude that the pairwise Ising model captures the structure of the $N = 40$ neuron system very well. Smaller groups of neurons for which exact pairwise models are computable also show excellent agreement with the data[30] (Schneidman et al., 2006).

It is surprising that pairwise models work well both on $N = 40$ neurons and on smaller subsets of these: not observing σ_χ will induce a triplet interaction among neurons $\{\sigma_\alpha, \sigma_\beta, \sigma_\gamma\}$ for any triplet in which there were pairwise couplings between σ_χ and all triplet members. Moreover, comparison of the parameters in $\mathbf{g}^{(40)}$ with their corresponding averages from different subnets $\mathbf{g}^{(20)}$ leaves the exchange interactions almost unchanged,

[29] The learning rate $\eta(t)$ was decreased as $O(1/t)$ or slower according to a custom schedule; $\alpha = 0.9$ for a network of $N = 120$ neurons and 0 otherwise. An initial approximate solution for $\mathbf{g}$ was obtained by contrastive divergence (CD) Monte Carlo (Hinton, 2002) for 40 neurons for which we have the complete set of patterns needed by CD. The Hamiltonian was rewritten such that J_ij was constraining $(\sigma_\mathrm{i} - \langle\sigma_\mathrm{i}\rangle_\mathrm{expt})(\sigma_\mathrm{j} - \langle\sigma_\mathrm{j}\rangle_\mathrm{expt})$, and we found that this removed biases in the reconstructed covariances.

[30] Exact pairwise models at $N = 20$ exhibit similar but smaller systematic deviations from the data, suggesting that the deviations are not due to convergence problems [Methods A.2.4].

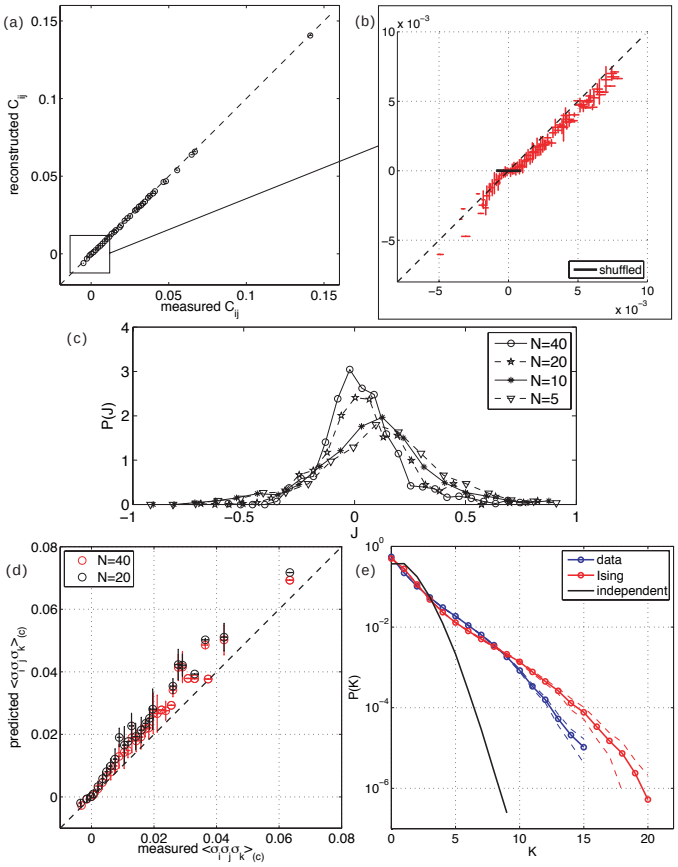


Figure 2.18: (a) Precision of the Ising model learned via Eq (2.34): measured covariance elements are binned on the x-axis and plotted against the corresponding reconstructed covariances on y-axis; vertical error bars denote the deviation within the bin and horizontal error bars denote the bootstrap errors on covariance estimates [Methods A.2.1]. (b) Zoom–in for small $C_{\rm ij}$, with scale bar representing the distribution of covariances from shuffled data. Not shown are the reconstructions of the means, $\langle\sigma_{\rm i}\rangle$, which are accurate to better than 1%. (c) Distribution of coupling constants $J_{\rm ij}$. (d) Measured vs predicted connected three–point correlations for 40 neurons (red) and exact solution for a 20 neuron subset (black). (e) Probability of observing K simultaneous spikes, compared to the failure of the independent model (black line). Dashed lines show error estimates.

while magnetic fields change substantially [Methods A.2.5]. To explain both phenomena, we examine the flow of the couplings under decimation. Specifically, we include three–body interactions, isolate terms related to spin $\sigma_{\rm n}$, sum over $\sigma_{\rm n}$, expand in $J_{\rm in}, J_{\rm ijn}$ up to $O(\sigma^4)$, and then identify renormalized couplings [Methods A.2.6]:

$$h_{\rm i} \rightarrow h_{\rm i} + \omega\tilde{J}_{\rm in} + \textstyle\sum_{\rm j}\beta_{\rm ij} + \mathcal{O}(\gamma,\delta), \tag{2.35}$$

$$J_{\rm ij} \rightarrow J_{\rm ij} + \beta_{\rm ij} + \mathcal{O}(\gamma,\delta), \tag{2.36}$$

$$J_{\rm ijk} \rightarrow J_{\rm ijk} + \mathcal{O}(\gamma,\delta) \tag{2.37}$$

where $\tilde{J}_{\rm in} = J_{\rm in} - \sum_{\rm j} J_{\rm ijn}$, $\beta_{\rm ij} = \tilde{J}_{\rm in}\tilde{J}_{\rm jn}(1-\omega^2) + \omega J_{\rm ijn}$ and $\omega = \tanh(h_{\rm n} - \sum_{\rm i} J_{\rm in} + \frac{1}{2}\sum_{\rm ij} J_{\rm ijn})$. The terms $\gamma, \delta \propto (1-\omega^2)$ originate from terms with 3 and 4 factors of σ, respectively. The key point is that neurons spike very infrequently (on average in $\sim 2.4\%$ of the bins) and so $\langle\sigma_{\rm i}\rangle \approx -1$, in which case ω is approximately the hyperbolic tangent of the mean field at site n and is close to -1. If pairwise Ising is a good model at size N, and couplings are small

enough to permit expansion, then at size $(N-1)$ the corrections to pairwise terms, as well as J_{ijk}, are suppressed by $1-\omega^2$. This could explain the dominance of pairwise interactions: it is not that higher order terms are intrinsically small, but the fact that spiking is rare means that they do not have much chance to contribute. Thus, the pairwise approximation is more like a Mayer cluster or virial expansion than like simple perturbation theory.

We test these ideas by selecting 100 random subgroups of 10 neurons out of 20; for each, we compute the exact Ising model from the data, as well as applying Eqs (2.35–2.37) 10 times in succession to decimate the network from 20 cells down to the chosen 10. The resulting three–body interactions J_{ijk} have a mean and standard deviation ten times smaller than the pairwise J_{ij}. If we ignore these terms, the average Jensen–Shannon divergence (Lin, 1991) between this probability distribution and the best pairwise model for the $N = 10$ subgroups is $\overline{D}_{JS} = 9.3 \pm 5.4 \times 10^{-4}$ bits, which is smaller than the average divergence between either model and the experimental data and means that $\gg 10^3$ samples would be required to distinguish reliably between the two models. Thus, sparsity of spikes keeps the complexity in check.

Given a model with couplings $\mathbf{g}$, we can explore the statistical mechanics of models with $\mathbf{g} \to \mathbf{g}/T$. In particular, this exercise might reveal if the actual operating point ($T = 1$) is in any way privileged. Tracking the specific heat vs T also gives us a way of estimating the entropy at $T = 1$, which measures the capacity of the neurons to convey information about the visual world; we recall that $S(T=1) = \int_0^1 C(T)/T\, dT$, and the heat capacity can be estimated by Monte Carlo from the variance of the energy, $C(T) = \langle(\delta E)^2\rangle/T^2$ [Methods A.2.7].

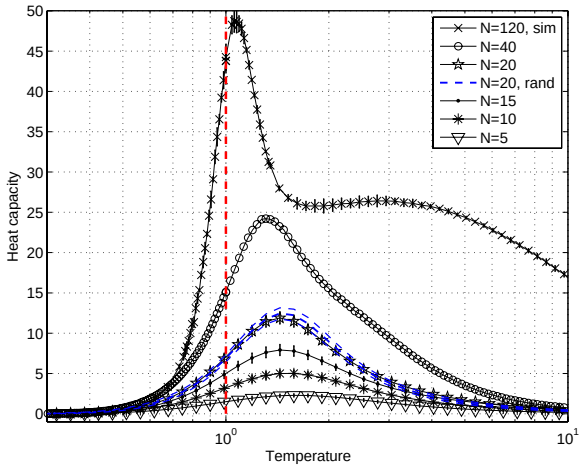


Figure 2.19: $C(T)$ for systems of different sizes. Ising models were constructed for 400 subnetworks of size 5, 180 of size 10, 90 of size 15 and 20, 1 full network of size 40 (all from data), and 3 synthetic networks of size 120; vertical error bars are standard deviations across these examples. The mean of the heat capacity curve and the 1 sigma envelope for Ising models of randomized networks are shown in blue dashed lines.

Figure 2.19 shows the dependence of heat capacity on temperature at various system sizes. We note that the peak of the heat capacity moves towards the operating point with increasing size. The behavior of the heat capacity $C(T)$ is diagnostic for the underlying density of states, and offers us the chance to ask if the networks we observe in the retina are typical of some statistical ensemble. One could generate such an ensemble by randomly choosing the matrix elements J_{ij} from the distribution that characterizes the real system, but models generated in this way have wildly different values of $\langle\sigma_i\rangle$. An alternative is to consider that these expectation values, as well as the pairwise correlations C_{ij}, are drawn independently out of a distribution, and then we construct Ising model consistent with

these randomly assigned expectation values. Figure 2.19 shows $C(T)$ for networks of 20 neurons constructed in this way,[31] and we see that, within error bars, the behavior of these randomly chosen systems resembles that of real 20 neuron groups in the retina.

Armed with the results at $N = 20$, we generated several synthetic networks of 120 neurons by randomly choosing once more out of the distribution of $\langle\sigma_i\rangle$ and C_{ij} observed experimentally [32]. The heat capacity $C_{120}(T)$ now has a dramatic peak at $T^* = 1.07 \pm 0.02$, very close to the operating point at $T = 1$. If we integrate to find the entropy, we find that the independent entropy of the individual spins, $S_0(120) = 17.8 \pm 0.2$ bits, has been reduced to $S(120) = 10.7 \pm 0.2$ bits. Even at $N = 120$ the entropy deficit or multi–information $I(N) = S_0(N) - S(N)$ continues to grow in proportion to the number of pairs ($\sim N^2$) [Fig A.7], continuing the pattern found in smaller networks (Schneidman et al., 2006). Looking in detail at the model, the distribution of J_{ij} is approximately Gaussian $\overline{J} = -0.016 \pm 0.004$ and $\sigma_J = 0.61 \pm 0.04$; 53% of triangles are frustrated (46% at $N = 40$), indicating the possibility of many stable states, as in spin glasses (Mezard et al., 1987). We examine these next.

At $N = 40$ we find 4 local energy minima $(\mathcal{G}_2, \cdots, \mathcal{G}_5)$ in the observed sample that are stable against single spin flips, in addition to the silent state $\mathcal{G}_1$ ($\sigma_i = -1$ for all i) [Methods A.2.8]. Using zero–temperature Monte Carlo, each configuration observed in the experimental data is assigned to its corresponding stable state. Although this assignment makes no reference to the visual stimulus, the collective states $\mathcal{G}_\alpha$ are reproducible across multiple presentations of the same movie [Fig 2.20a], even when the microscopic state $\{\sigma_i\}$ varies substantially [Fig 2.20b].

At $N = 120$, we find a much richer structure:[33] the Gibbs state now is a superposition of thousands of $\mathcal{G}_\alpha$, with a nearly Zipf–like distribution [Fig 2.20c]. The entropy of this distribution is 3.4 ± 0.3 bits, about a third of the total entropy. Thus, a substantial fraction of the network's capacity to convey visual information would be carried by the collective state, that is by the identity of the basin of attraction, rather than by the detailed microscopic states [Methods A.2.9].

To summarize, the Ising model with pairwise interactions continues to provide an accurate description of neural activity in the retina up to $N = 40$. Although correlations among pairs of cells are weak, the behavior of these large groups of cells is strongly collective [Methods A.2.10], and this is even clearer in larger networks that were constructed to be typical of the ensemble out of which the observed network has been drawn. In particular, these networks seem to be operating close to a critical point. Such tuning might serve to maximize the system's susceptibility to sensory inputs, as suggested in other systems (Duke and Bray, 1999; Eguiluz et al., 2000; Camalet et al., 2000); by definition operating at a peak of the specific heat maximizes the dynamic range of log probabilities for the different microscopic states, allowing the system to represent sensory events that occur with a wide range of likelihoods.[34] The observed correlations are not fixed by the anatomy of the

[31] Not all combinations of means and correlations are possible for Ising variables. After each draw from the distribution of $\langle\sigma_i\rangle$ and C_{ij}, we check that all 2×2 marginal distributions are in $[0, 1]$, and repeat if needed. Once the whole synthetic covariance matrix is generated, we check (e.g. using Kolmogorov–Smirnov) that the distribution of its elements is consistent with the measured distribution.

[32] Learning the couplings $\mathbf{g}$ was slow, but eventually converged: C_{ij} converged to within 10% for the largest quartile of elements by absolute value, and within 15% for the largest half, without obvious systematic biases.

[33] One run of $2 \cdot 10^7$ independent samples ($\tau_{\mathcal{G}}^{-1} \approx 5 \cdot 10^2$flips) is collected; for each sample ZTMC is used to determine the appropriate basin; we track $5 \cdot 10^4$ lowest energy stable states and keep detailed statistics for 10^3 lowest [Methods A.2.9].

[34] In contrast, simple notions of efficient coding require all symbols to be used with equal probability. But

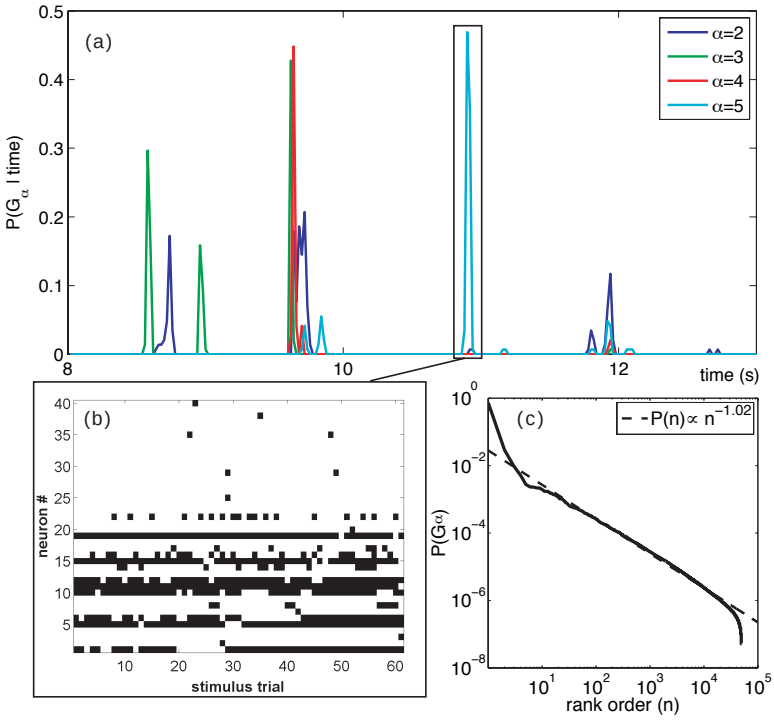


Figure 2.20: (a) Probability that the 40 neuron system is found in a configuration within the basin of each nontrivial ground state $\mathcal{G}_\alpha$, as a function of time during the stimulus movie; $P(\mathcal{G}_\alpha|t) = 0.4$ means that the retina is in that basin on 40% of the 145 repetitions of the movie. (b) All unique patterns assigned to $\mathcal{G}_5$ at $t = 10.88 - 10.92\,\mathrm{s}$. (c) Zipf plot of the rank ordered frequencies with which the lowest lying $5 \cdot 10^4$ stable states are found in the simulated 120 neuron system.

retina or by the visual input alone, but reflect adaptation to the statistics of these inputs (Smirnakis et al., 1997); it should be possible to test experimentally whether these adaptation processes preserve the tuning to a critical point as the input statistics are changed. Finally, the transition from $N = 40$ to $N = 120$ opens up a much richer structure to the configuration space, suggesting that the representation of the visual world by the relevant groups of $N \sim 200$ cells may be completely dominated by collective states that are invisible to experiments on smaller systems.

since states of the visual world occur with wildly varying probabilities, this (and related notions of efficiency) require codes that are extended over time. If the brain is interested directly in how surprised it should be by the current state of its inputs, then it might be more important to maximize the dynamic range for instantaneously representing this (negative log) likelihood.

2.6.1 Extension to stimulus dependent models

In the previous section we explored the distribution of spike pattern responses that the retina emits when it is shown a naturalistic movie. Our primary goal was to characterize the output binary "vocabulary" without any regard for the inputs, s, that this vocabulary encodes. By studying solely the words of the output, we observed how the number of accessible patterns, $2^{S(N)}$, where $S(N)$ is the entropy of the network of N neurons, increases with network size [Fig A.7]; we furthermore pointed out that information could possibly be encoded in collective states corresponding to the local energy minima of the Ising model, instead of in the precise microscopic patterns of spiking and silence. Last but not least, we justified our analysis of the correlations in retinal output by pointing out that brain itself does not have independent access to the stimuli: it only "sees" the spikes streaming in along the optic nerve, and therefore all the information that it can extract about the world must be present in $p(\vec{\sigma})$.

Here we try to incorporate the dependence on the input into the maximum entropy framework by building on the ideas of conditional dependence in biochemical network and with the aim of extracting explicit encoding conditional probability distribution, $p(\vec{\sigma}|s)$.

A lot of work has been done in characterizing the responses of a single neuron to time-dependent stimuli (Rieke et al., 1997; Dayan and Abbott, 2001). One usually looks for a description of the neural processing in the form of a spatio-temporal receptive field $\mathcal{R}(\mathbf{r}, t)$, such that the probability of spiking at time t is related to the stimulus $s(\mathbf{r}, t)$ as follows:

$$p(\text{spike}|s) = \mathcal{F}\left(\int d\mathbf{r} \int dt' \, \mathcal{R}(\mathbf{r}, t') s(\mathbf{r}, t - t')\right), \tag{2.38}$$

where $\mathcal{F}$ is some nonlinear function, for instance a sigmoid or half-wave rectification. The receptive fields for neurons in the retina (as opposed to higher processing centers such as V1) factorize, such that $\mathcal{R}(\mathbf{r}, t) = \mathcal{R}_{\mathbf{r}}(\mathbf{r})\mathcal{R}_t(t)$; the first factor is the spatial receptive field, and the second factor the temporal receptive field.[35] While this simple connection between the stimulus and the response of a single neuron performs reasonably for a single neuron, it has been shown in the work of Schneidman et al. (2006) and Shlens et al. (2006) and subsequent analysis of Section 2.6, that understanding the correlation structure between the neurons is crucial for capturing the properties of the *population code*. It is with this in mind that we now proceed to create maximum entropy models of the *joint distribution* of (spikes, stimulus), which simultaneously addresses both the issue of input dependence and the neuron-to-neuron interactions.

The problem will be simplified here because the stimulus used in the experiment will be uniform in space and the spatial receptive field will therefore be unimportant. We will try to compute the temporal receptive fields, or spike-triggered-averages (STA), jointly with neuron interactions. In the same way that the condition C was represented in the case of biochemical networks by a set of condition variables y_k, $C \equiv \{y_1, y_2, \ldots, y_K\}$, we will use a set of binary variables *in addition* to N variables describing spiking, in order to encode the stimulus, s. The response at time t is a convolution of the kernel over the past stimulation, as in Eq (2.38), and therefore the encoding of the stimulus that determines response at time t must contain some of the stimulus history. We detail the procedure below.

[35] The temporal receptive field is, in the case of Gaussian white noise stimulus ensemble, proportional to the so called spike-triggered average, or STA, i.e. the time course of the stimulus preceding the spike.

Stimulus encoding

The stimulus, i.e. the instantaneous light intensity that is projected onto the retina, and spike trains are discretized into bins of size Δt.

In the case of the stimulus, the average of the input light intensity is computed for the period of Δt in duration and that average is discretized into Q_S levels. For biochemical networks we devised a special quantization scheme that retained dependencies between different protein activation levels; here, in contrast, we deal with a single random time series, and we discretize it such that the bin boundaries cut the domain into intervals containing the same number of data points (Slonim et al., 2005a) . Thus, the real valued raw light intensities are mapped to:

$$s(t) \to s^t, s^t \in \{0, \ldots, Q_S - 1\}\,. \tag{2.39}$$

In case of the spike train, σ_i^t is equal to 1 if there was at least one spike of neuron i in time bin $[t, t + \Delta t]$, and -1 otherwise.

We want a probabilistic model of population behavior,

$$p\left(\{\sigma_i\}, \vec{s}\right) = \frac{1}{Z} \exp\left\{ \sum_i h_i^0 \sigma_i + \frac{1}{2} \sum_{i,j} J_{ij} \sigma_i \sigma_j + \sum_i \sum_{\alpha=-\alpha_0}^{-K+1-\alpha_0 \log_2 Q_S} \sum_{\beta=1} h_i^{\alpha,\beta} \tilde{s}^{\alpha,\beta} \sigma_i + \cdots \right\}, \tag{2.40}$$

where $\tilde{s}^t$ is the binary encoding of s^t, and the index β runs over all $\log_2 Q_S$ bits of this encoding.[36] The last term of Eq (2.40) describes the coupling of each spin σ_i to the past history of the stimulus, extending back in time for K time bins, starting α_0 time bins in the past relative to spike/silence in the current bin. In particular, α_0 could be 0, which would couple instantaneous spiking to the stimulus at the same moment, but we can generally leave α_0 as a free parameter. The $\cdots$ terms describe couplings among stimulus variables and are uninteresting for this discussion.[37]

Here we can observe a nice analogy with position weight matrices used to describe the transcription-factor / DNA interaction energy (Berg and von Hippel, 1987), see also Section 3.3. A position weight matrix (PWM) is a linear filter that takes a short piece of DNA sequence and returns a corresponding scalar, or binding energy, that is *a sum* of independent energy contributions from every letter at each position in the sequence. In the situation we are considering here, the "sequence" is the time-ordered sequence of stimulus code words, $\tilde{s}^{\alpha,\beta}$, and one is computing the inner product between it and the *stimulus-dependent magnetic fields*, $h_i^{\alpha,\beta}$, to yield a scalar quantity. The probability of spiking is then given by the Boltzmann distribution, just as in the case of DNA binding, where either a sharp threshold (a step function) or a smooth one (Fermi distribution) are applied to the energy of the DNA site to determine whether the site is bound or unbound [Eq (3.18)].

Let's suppose that we have discretized the stimulus into a sequence

$$\{\ldots, s_{-K+1}, s_{-K+2}, \ldots, s_{-1}, s_0, \ldots\},$$

where each s_i can take on Q_S different, discrete values. We are claiming that each neuron, at time $t = 0$, looks at the past history of K such samples – essentially a K-letter, $\log_2(Q_S)$

[36] This is done just to reduce the stimulus encoding problem to the two-state Ising model.

[37] Since the stimulus will be a random Gaussian-intensity spatially-uniform flicker with a fixed frequency, these terms will be all close to zero, with the possible exception of small couplings between the neighboring time bins, as time-averaging in Δt windows mixes some signal from the previous bin into the next one.

bit word – and creates a dot product between the "weight matrix" and the sequence:

$$h_{\text{eff}}(\vec{s}) = \sum_{\alpha=0}^{-K+1} \mathcal{M}_{\alpha,s_\alpha} = \sum_{\alpha=0}^{-K+1} \sum_{\beta} h^{\alpha,\beta} \tilde{s}^{\alpha,\beta}, \tag{2.41}$$

where $\mathcal{M}$ is a properly arranged matrix of dimension $K \times \log Q_S$ of stimulus-dependent magnetic fields from the Hamiltonian in Eq (2.40). Such energy matrix $\mathcal{M}$ must exist, because it is simply a different way to rewrite a linear function of the stimulus from Eq (2.40). To illustrate, consider a trivial network of a single neuron exposed to a given stimulus. Its probability of firing will be, according to Eq (2.40):

$$p(\sigma|\vec{s}) = \frac{1}{Z'} \exp\left\{h^0\sigma + h_{\text{eff}}(\vec{s})\sigma\right\} \tag{2.42}$$

In the simplest model above, the average probability of firing is a sigmoidal function $\langle\sigma\rangle = \tanh(h_0 + h_{\text{eff}}(\vec{s}))$. In fact, our general N-body setup has in this case been reduced to the known single-neuron problem of linear filter inference: the linear filter is embodied in $h_{\text{eff}}(\vec{s})$, and the firing rate is a simplified version of Eq (2.38). It is unclear what happens when the network is extended to two or more neurons: how much of the spiking of these two neurons is explained by their own sensitivity to the stimulus and how much by their mutual coupling?

If all practical problems could be addressed, it seems that this approach could combine both the inherent couplings between the neurons and the dependence on stimulus history for each neuron. If inherent couplings really are a property of the network and not of the stimulus, they should remain constant for the same set of neurons regardless of the stimulus, or at least for different time slices of the same recording. Because it incorporates the stimulus, this formulation naturally yields a testable prediction for spike trains of the neural population, given the stimulus. Finally, relative sizes of the h^0, J and the distribution of the stimulus-dependent magnetic fields h_i^α describe the importance weight carried by three very different functional processes (inherent spiking, coupling with other neurons, stimulus sensitivity) that together determine the spiking probability of each neuron.

Application

The data consists of simulaneous recordings from 20 neurons, exposed to spatially homogenous light-intensity flicker at 120Hz, chosen with a Gaussian prior on intensities, such that the contrast variance is about 30 percent. I would like to acknowledge and thank Greg Schwartz of Michael Berry's lab at Princeton University, for sharing the data (Schwatz, 2006).

Both the neural response and the stimulus time series are discretized into $\Delta t = 25$ms time bins, which span approximately 3 different and independent light intensity draws from the stimulus time series (120 Hz corresponds to 8.3ms intervals of constant intensity). We focus on 10 neurons only, and reserve 10 bits per time bin for description of the stimulus: the first 2 bits represent the intensity at $t_0 - \Delta t$, the second two represent the intensity at $t_0 - 2\Delta t$, and so on, stretching 5 time bins into the past.

The resulting data is a binary matrix of 20 time series (10 spike series and 10 stimulus series), with a total of 84236 samples. The maximum entropy pairwise ansatz produces the distribution $p(\{\sigma\}, \vec{s})$. Since the stimulus is random, the pairwise model of the marginal distribution $p(\vec{s})$ will be unimportant, however, both the encoding distribution, $p(\{\sigma\}|\vec{s})$,

and the spike marginal, $p(\{\sigma\})$, can be extracted from the joint and they fit the data well [Fig 2.21].

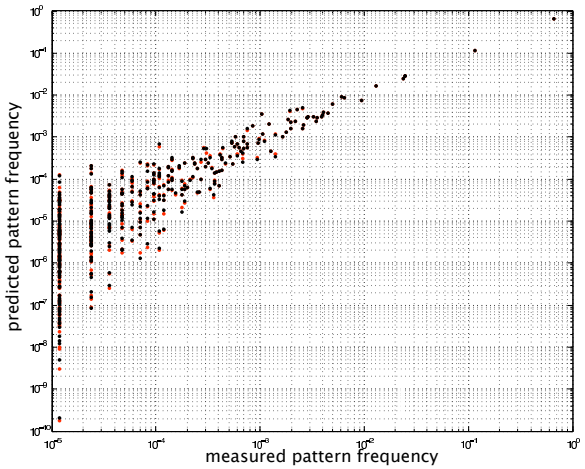


Figure 2.21: Maximum entropy models of 2^{10} spike patterns for 10 neurons. Predicted frequency of the 10-neuron pattern on vertical axis, observed frequency on horizontal axis. Red – model conditioned on the stimulus, black – stimulus independent model. $D_{JS}(\text{red}, \text{data}) = 0.0076$, $D_{JS}(\text{black}, \text{data}) = 0.0078$, $S = 2.45$, $S_0 = 2.82$, $S^{(2)} = 2.48$ (all data).

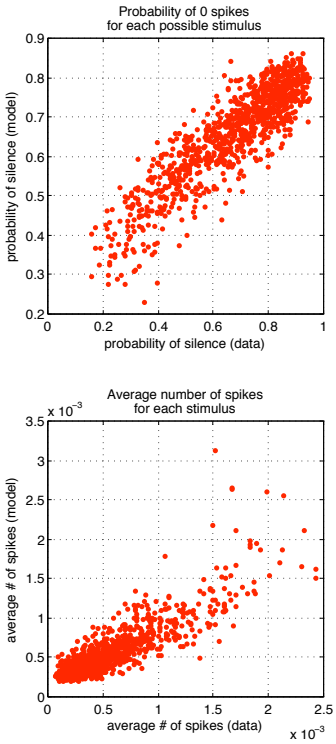


Figure 2.22: Top: the probability of silence for each of the 1024 stimuli estimated from the data, plotted against the same probability given by the condition-dependent Ising model. Bottom: measured and predicted number of spikes for each possible stimulus. We use a joint model of spiking and stimulus for 10 neurons, with 10 bits reserved for stimulus encoding.

There is not enough data to sample conditional responses and make frequency-frequency conditional plots similar to Fig 2.21, but some projections of the distributions are made in Fig 2.22. We can compare the conditionally dependent probability of silence of the joint spike-stimulus maximum entropy model with the prediction of conditionally independent neurons, i.e. the maximum entropy model assuming $\langle \sigma_i \sigma_j \rangle$ are not constrained to reproduce measured correlations and thus $J_{ij} = 0$. The latter model systematically underestimates the probability of silence by about 17 percent on average, while the model with coupling

overestimates it by about 11 percent – including coupling therefore helps, but there is still room for improvement. In general we observe an agreement between the maximum entropy prediction and experiment despite significant scatter, which partly must also be due to our crude representation of the stimulus.

How well does the model describe the responses of single neurons? We can compute the predicted spike triggered average ($\mathbf{w}_i$) for each neuron i:

$$\mathbf{w}_i = \sum_{\vec{s}} p(\vec{s}\,|\,\sigma_i = 1)\vec{s}, \tag{2.43}$$

$$p(\vec{s}\,|\,\sigma_i) = p(\sigma_i, \vec{s})/p(\sigma_i), \tag{2.44}$$

$$p(\sigma_i, \vec{s}) = \sum_{\sigma_j, j\neq i} p(\{\sigma\}, \vec{s}). \tag{2.45}$$

The figures that compare the spike triggered averages constructed from the Ising model and the spike triggered averages computed with the traditional reverse-correlation methods from the data for each neuron independently, are shown in Fig 2.23; given the crudeness of stimulus representation the match is quite remarkable. A calculation similar to Eq (2.43) allows one to condition the STA on alternative triggers (such as a pair of neurons firing at the same time).

How should we visualize the corresponding terms in the Hamiltonian? For each neuron, the dependence on the stimulus is reflected by the sum of the form:

$$H_i = -\cdots - \sum_{\alpha,\beta} h_i^{\alpha,\beta}\sigma_i\tilde{s}^{\alpha,\beta}$$

where α is the time index that extends K bins back into the past, and, at each time instant, there is $\log_2(Q_S)$ bits, indexed by β, available for description of the light intensity $s(t)$. In the case at hand we have 2 bits per timebin, so that we can encode 4 grayscale levels.[38]

The energy matrix form of the stimulus coupling in the Hamiltonian would allow the neuron, for each time bin, to be funnily sensitive to the intensity: a hypothetical neuron might "prefer" light level 3 and "dislike" levels 1, 2 and 4. Although this is not forbidden by our model, we know that the temporal kernel is linear at each point in time and is essentially the scalar multiplier for the stimulus intensity at each time bin. A relevant question is then to assess if our energy matrix is also linear and not "combinatorial" at each bin. If this is so, then one coefficient (the linear slope) is sufficient to describe the response in that bin. Figure 2.24 shows that this seems to be a good approximation.

If such an approximation can be made and the energy contribution conditioned on the stimulus intensity in a given time bin is proportional to the intensity, there exists a spike-triggered-average equivalent plot in energy space, which is displayed in Fig 2.25.

By inspecting the magnitudes of various terms in the Hamiltonian we see that inherent magnetic field and the stimulus dependent magnetic field range over the same order of magnitude (1 in natural units), while the couplings are a bit smaller (but since the neurons are mostly silent, the effective contribution from the other neurons is close to $h_i^{\text{coupling}} \approx -\sum_j J_{ij}$, which is also order 1); this confirms our intuition that it is a bad approximation to leave out the neuron-to-neuron couplings when considering the population code, even if the

[38] In terms of raw data all intensities below intensity 155 (mean 141) in arbitrary units are assigned level $1 = (0,0)$, everything between 155 and 175 (mean 166) is assigned level $2 = (0,1)$, between 175 and 186 (mean 180) is level $3 = (1,0)$ and everything above 186 (mean 212) is level $4 = (1,1)$. These divisions capture quartiles of data and have been used to map the discrete values back into the raw intensity space.

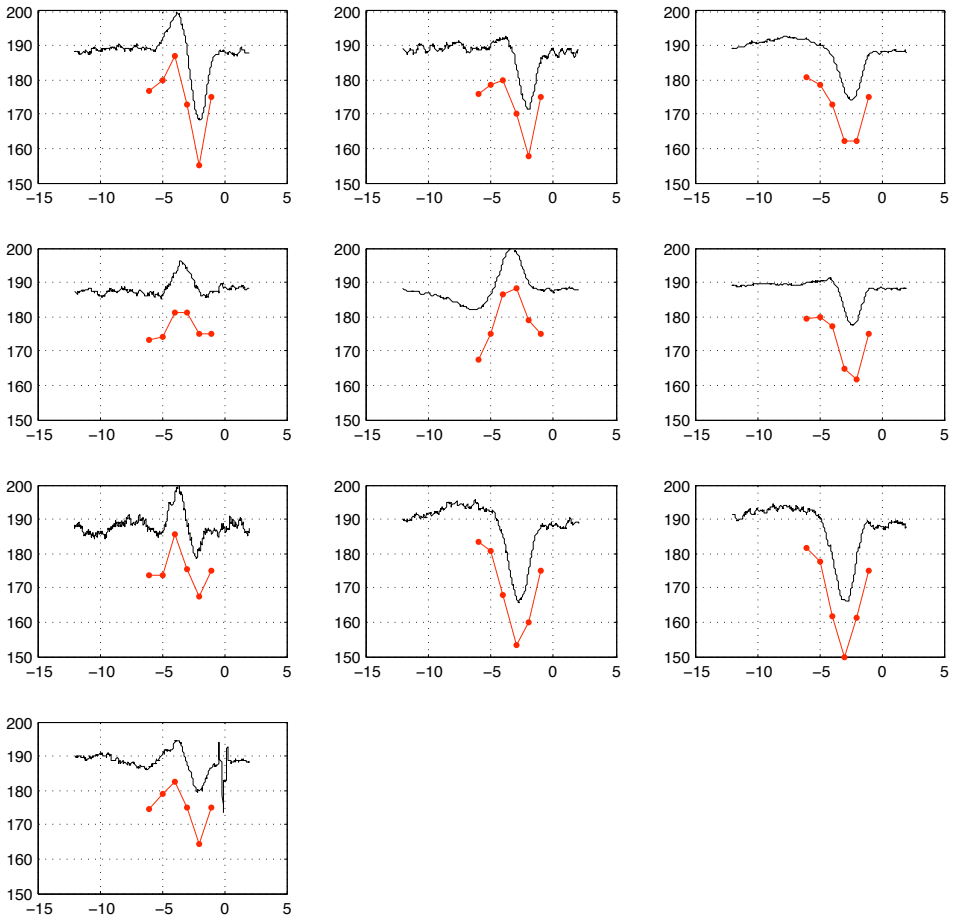


Figure 2.23: Spike triggered averages for 10 neurons. Time is depicted on horizontal axis for a spike occurring at $t = 0$, and each unit corresponds to $\Delta t = 25$ms. Vertical axis contains arbitrary intensity units. Black line shows single-neuron reverse correlation calculations of STA from white noise stimuli. Red line shows Ising model predictions for STA $\mathbf{w}$ [Eq (2.43)], based on marginalizing the Ising model distribution of spiking and stimulus for 10 neurons, $p(\vec{\sigma}, \vec{s})$, over all but one neuron. The first timebin preceding the spike is always equal to the stimulus average, since we are computing a model that is coupled only to stimulus preceding spikes by more than 1 bin.

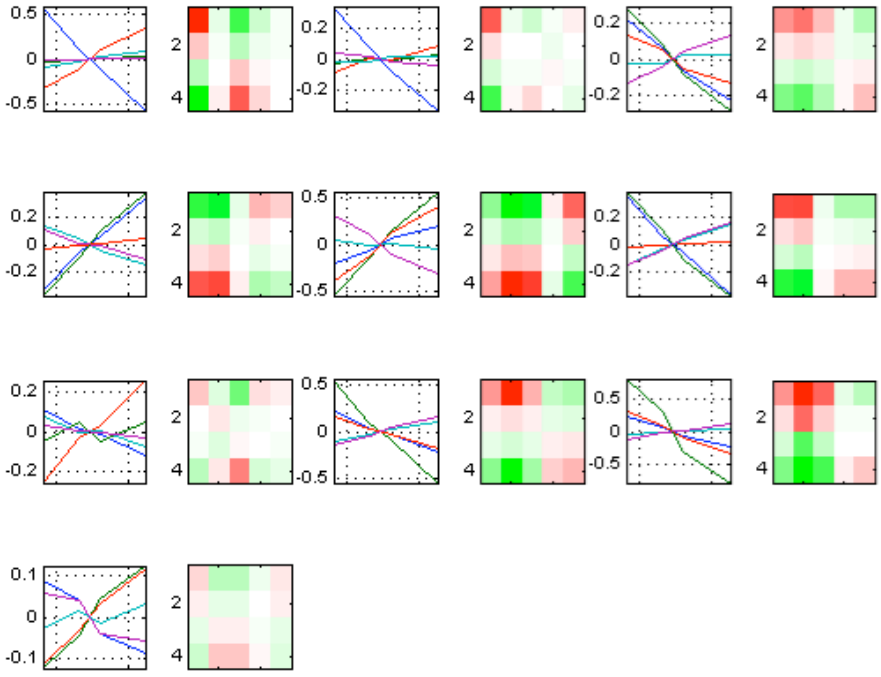


Figure 2.24: Linearity in the neuron-stimulus coupling, for 10 neurons, listed top-to-bottom, left-to-right. For each neuron we plot two panels. Left panels show the linear relation between the energy contribution (vertical axis) and the stimulus intensity (horizontal axis), for each of the $K = 5$ bins in time separately (each of the 5 lines is determined by 4 points, corresponding to the discretization of light intensity into 4 bins). Right panels show energy contributions to stimulus-dependent magnetic fields, $\mathcal{M}_{\alpha s}$, plotted in the matrix form. Five bins in time (α) are on the horizontal axis, with the left-hand bin being closest to the spike, and the intensity increases downwards (4 levels for s). The statement that neuron is linear in each time bin corresponds to the observation that in the energy matrix the colors change in each column from red-to-green or from green-to-red along the vertical direction, linearly and without alternating.

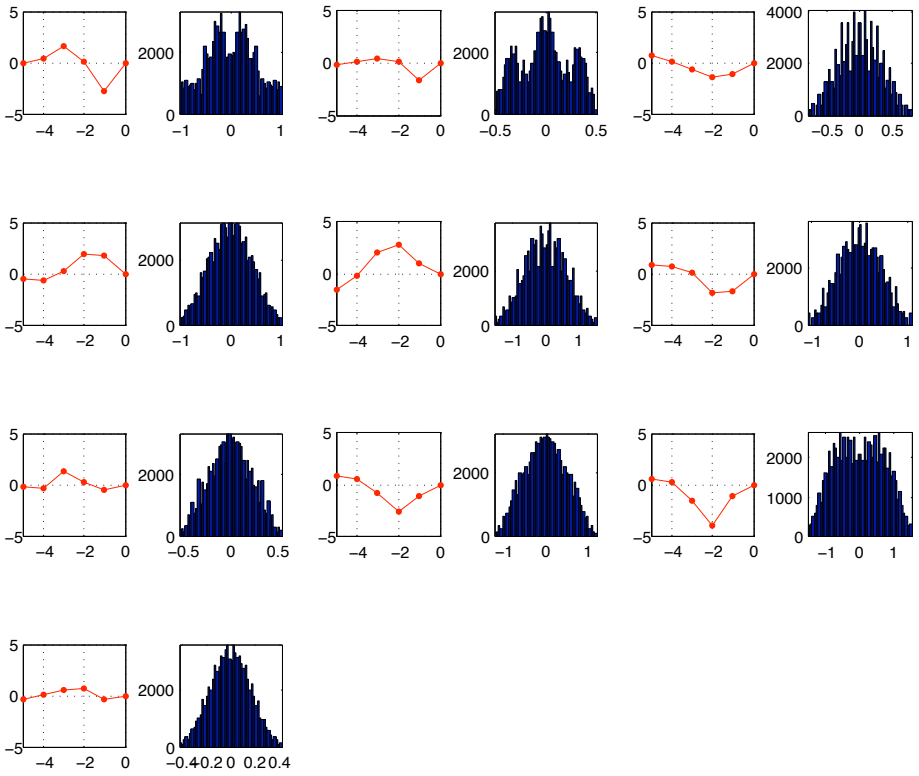


Figure 2.25: For each of the 10 neurons and assuming that the energy contribution in each time bin is linearly proportional to light intensity, this figure shows the energy STA-equivalent. For each stimulus $I(t)$ (raw intensity units), the energy contribution for neuron i will be given by $h_i = \sum_t I(t)\chi_i(t)$, where $\chi_i(t)$ are, for each t and each neuron i, the slopes of the response functions of Fig 2.24. $\chi_i(t)$ are only relevant up to an additive constant and are plotted as functions of time in red; qualitatively, their shape is similar to STA. In other words, χ are the kernels with which one convolves the stimulus, $I(t)$, to get the time-dependent energy contribution. The histograms show the exact (calculated from the weight matrices and stimuli) distributions of stimulus-dependent magnetic fields that each neuron experiences.

Spike entropy	2.45 bits
Stimulus entropy	≈ 10 bits
Spike multi-information	0.37 bits
Fraction of I captured by Ising model, $I^{(2)}$, on spike trains	0.92
Channel capacity with $p(\vec{\sigma}\|\vec{s})$ and the optimal choice of $p(\vec{s})$	0.80 bits

Table 2.2: Information-theoretic quantities for a system of 10 neurons collectively driven by a spatially homogenous flicker. Spike entropy is estimated from the data; stimulus is random and is therefore uniform in its encoding variables.

stimulus dependence is separately accounted for in the Hamiltonian.[39] To make precise the statements about the relative influences of the stimulus energy contributions and inherent inter-neuron couplings, we would need to tighten the approximations we made in the way we encoded the stimulus; one possibility that is currently being explored is to present the retina with repeats of the same movie segment, and use time index within the movie segment as the proxy for the stimulus. In this case the magnetic fields acting on the neurons literary are directly time dependent, and can capture the behavior of the neurons without assuming which features specifically they are sensitive to (and how far back in past one needs to look).[40]

Although the maximum entropy model is formulated for the joint (spikes, stimulus) distribution, the random stimulus prior, $p(\vec{s}) = \sum_\sigma p(\{\sigma\}, \vec{s})$, does not tell us anything new; therefore it is meaningful to extract the encoder distribution $p(\{\sigma\}\,|\vec{s})$ only. We can make limited progress if we look for such distribution on stimulus space, $p(\vec{s})$, that will maximize the channel capacity of the encoder system, bearing in mind that the encoder distribution was deduced in an experiment where the retina was adapted to a uniform, full-field flicker stimulus and not its natural ensemble. In this specific case, we use Blahut-Arimoto algorithm (Blahut, 1972) to compute the optimal distribution of 1024 possible stimulus patterns; the results are shown in Table 2.2. The most frequently used stimulus patterns in an optimal code are shown in Fig 2.27, and the maximum information transmission that can be sustained by a system of N neurons is plotted in Fig 2.26. From Table 2.2 we see that the maximum information transmission would be about a third of the output entropy, a relatively low coding efficiency. Furthermore, if the stimulus distribution is not optimal but uniform – close to what is actually being shown in the experiment – the information transmission is only about a third of the capacity, or about 9 bits per second for 10 neurons. It would be extremely interesting to know if, by increasing the number of neurons N, one can increase the fraction of the entropy that is used to convey information due to the hypothesized error correcting nature of the code discussed in the preceding section; or to see how the encoding kernel adaptation affects the capacity of the population code.

[39] Alternatively this means that the correlation effects are not driven purely by the exposure to a common stimulus.

[40] At this time we are just becoming able to carry out the computations required to construct models with time-dependent fields.

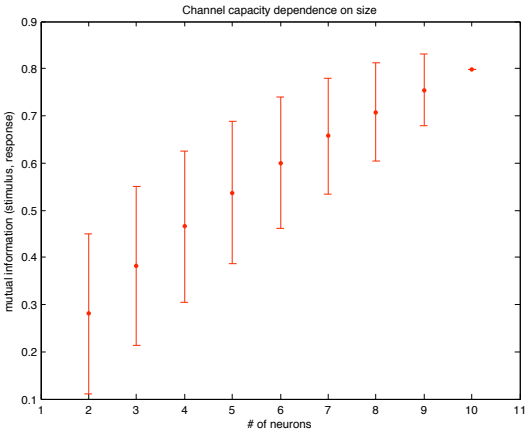


Figure 2.26: Optimal channel capacity between the simultaneous spiking of 10 neurons and the stimulus, as a function of the number of neurons in the network. For each possible subgroup of n neurons out of 10, we find the pairwise Ising model, calculate the encoding distribution $p(\vec{\sigma}|\vec{s})$, and then compute the optimal input alphabet $p(\vec{s})$ and the corresponding channel capacity. Error bars are plotted by exhaustively choosing every possible group of n neurons out of 10, and computing the deviation of the channel capacities.

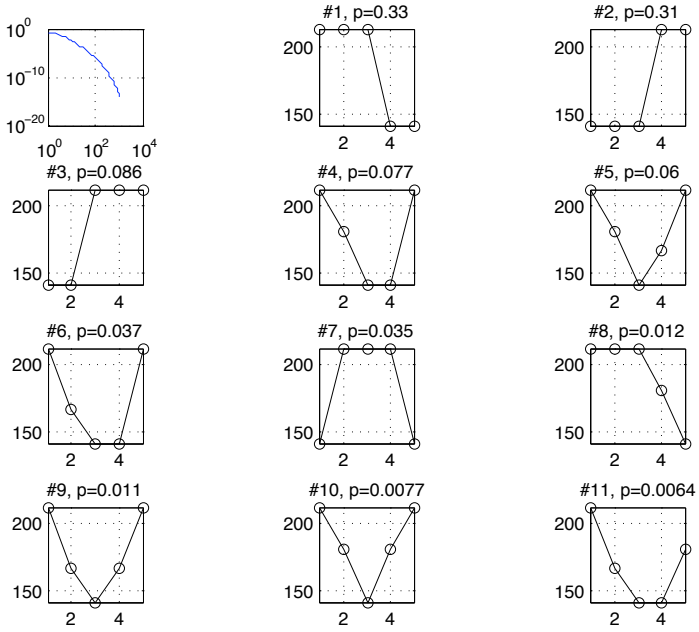


Figure 2.27: First panel: the rank ordered optimal probability distribution of 1024 stimuli that maximize channel capacity of the system of 10 neurons. The first 10 codewords account for 0.96, and the first 20 for 0.99 of the total weight. Remaining panels: codewords most frequently used in capacity achieving distribution, sorted by their probability. Horizontal axis is time in $\Delta t = 25\,\text{ms}$ increments, vertical axis is the stimulus intensity in arbitrary units.

2.7 Summary

In this chapter we attempted to build a picture of the collective behavior of N interacting elements, starting from a number of joint experimental measurements of their activities. We showed how to systematically create probability distributions that describe patterns of activities of the nodes when the network is in stationary state, and have further demonstrated how these maximum entropy distributions can be interpreted graphically as networks of interactions. The maximum entropy models constrained by pairwise marginals were later extended to include the dependence of activities on a (parametrized) set of external conditions (intervening chemicals or light intensity). Two very different systems were examined with this maximum entropy toolkit. In each, we were faced with a fresh set of problems, but in the end finished by observing two diametrically opposite network structures.

A signaling network of interacting proteins was small, with experimentally observed activities of 11 nodes. We were severely undersampled and therefore had to devise ways of encoding the experimental data as efficiently as possible before computing the corresponding Ising model, in itself an easy task for the case at hand. We found *sparse*, mostly unfrustrated and mutually similar interaction maps across all external conditions,[41] and argued that pairwise models either work well or fail cleanly in a localized way that indicates where the model must be supplemented with a higher-order interaction. Interestingly, interaction maps created from correlated fluctuations in a single experiment turned out to be quite similar to the interaction map describing the correlations between the activity changes under conditions of chemical intervention.[42] We were careful not to interpret the inferred interactions as microscopic chemical reactions between signaling proteins, because we have incomplete access to all nodes and internal states in the network, and so our reconstructed interactions reflect "averaging over" the states of the invisible parts of the network; this should also be a caveat when strong interventions are used to experimentally confirm or disprove microscopic interactions in vivo.[43] It actually seems that, for our analysis at least, using *small* perturbations would be much more revealing as it would only slightly change the activity patterns and – given the restriction on the dynamic range we face because of the quantization and sampling issues – would allow us to observe simultaneously the change in local fluctuations as well as the shift in the mean activation due to the interventions. In addition, if the interventions were small, we could expect their effects to add and one could test this additivity assumption by verifying that the effects of two simultaneously intervening chemicals are captured by adding the corresponding stimulus dependent magnetic fields.

A completely different set of issues arose when we analyzed the network of neurons. We could use $\approx 2 \cdot 10^5$ samples of naturally discretized data, and previous work has confirmed that the pairwise maximum entropy models work remarkably well on groups of up to 15 neurons. When we attempted to push the size of the analyzed network to the whole dataset of 40 and beyond, we were faced with technical problems of computing the maximum entropy distributions and estimating the information theoretic quantities, and finally had

[41] This is also probably the reason why a Bayesian network reconstruction assumptions are valid for this dataset.

[42] A guess for why this is so might be that the energy landscape of the system is simple and although the intervention chemicals shift the mean activity values, the "forces" that determine the dynamics around the stable state look the same no matter which state one is in; this could also explain why the joint activity/condition distribution, where conditions only couple to magnetic fields, is a good model for the data.

[43] The popular approach of using strong interventions probably works here also because the underlying network is sparse.

to resort to large Monte Carlo simulations. Nevertheless, we were able to reconstruct the interaction map for 40 neurons. In contrast to the signaling network, the neurons are densely connected, which gives rise to weak pairwise correlations, but ever stronger collective effects as the network grows larger. The pairwise model is very good but not perfect; however, the failure seems to be distributed throughout the network,[44] contrary to the signaling protein example, where a single higher-order interaction was required or a better fit. By constructing a simulated network of 120 neurons, which we believe would behave thermodynamically in a similar way to the smaller measured networks, we pushed the system into a regime where correlations start to dominate, $C_{\rm ij} \approx 1/N$,[45] and the entropy of the system significally decreases from its independent value. The system could use the correlation structure in the spike trains to perform error correction, if, as we suggest, the information is encoded in the identities of local energy attractors and not in the exact microscopic patterns of spiking. Such encoding would be rich, with a language-like distribution of "word" frequencies that span a wide range of probabilities, perhaps to allow for the encoding of very rare and very frequent stimuli; we suggest that the system could actively tune itself to this wide dynamic range. In this system, similarly to the biochemical network, we cannot observe all nodes; moreover, here the retina is indeed driven by the movie and the interactions represent the effect of both the underlying connectivity and excitation by the stimulus. We concluded the neural case by trying to account explicitly for the stimulus dependence in case where the stimulus can be easily encoded, with some success. In particular, it seems as if it were possible to study, for example, two neurons by simultaneously determining their receptive fields *and* their pairwise coupling. Despite being currently just a working idea for future explorations, the possibility is intriguing as it simultaneously incorporates both stimulus and non-trivial population coding, and consequently enables us to compute all the relevant information theoretic quantities for the neuronal population.

[44] For example, most of the three-point predicted interactions deviate systematically, although by a small amount, from their measured values.

[45] For a simple system in which all spins couple with the same exchange interaction and are exposed to the same external field, it is easy to show that susceptibility (per spin) is $\chi_1 = 1 - \langle\sigma\rangle^2 + (N-1)C$, where C is the connected pairwise correlation and N the number of neurons; in thermodynamic limit χ_1 should be intensive, and therefore the contribution of the last term would have to be negligible. In the scaling examined here (which is not the TD limit), C is held fixed as N is increased, and we watch out for new collective behavior as $C \sim 1/N$.

Chapter 3

Genetic regulatory networks

The second half of this thesis represents an attempt to derive and predict some of the properties of (genetic regulatory) networks with a top-down approach, starting with a theoretical principle and working out its observable consequences. Accordingly, the principle of maximization of information capacity will be examined in Chapter 4. To proceed, however, one must first understand the noise in the biological system of interest, and we take up this task here, in Chapter 3. By its end we should be able to define (and parametrize) *a class of biologically relevant models for noise in transcriptional regulation*, which will be used subsequently for computation of channel capacities.

In this chapter we focus specifically on genetic networks, mainly because of the amount of research that has already been invested into characterizing their signal transmission properties. As we will demonstrate shortly using the data from fruit fly development, there are precision methods available today that reveal the detailed properties of the noise in gene regulation and, despite extensive existing literature on the topic, can yield surprisingly new results.

3.1 Signal transduction in gene regulation

The *central dogma* of molecular biology puts into words the conceptual scheme by which the hereditary information is stored, read out and expressed into protein on a cellular level (Alberts et al., 2002). DNA, the information carrying molecule whose integrity is actively maintained by the cell, encodes this information in a linearly ordered sequence of base pairs. There are four possible types of bases attached to the sugar-phosphate DNA backbone, namely *adenine, cytosine, thymine and guanine*, and they constitute a four-letter genetic alphabet, $\{A, C, T, G\}$. When needed, the information is *transcribed* by the transcription apparatus, which is essentially a complex of proteins that can attach to DNA at particular places, called *promoter sequences*. After binding, this machinery can slide along the length of the gene coding region on the DNA and produce a messenger RNA (mRNA) transcript, carrying an inverted copy of the DNA original.[1] The message only has a limited lifetime that is oftentimes actively controlled by the cell; in eukaryotes it is exported from the nucleus into the cytoplasm for further processing. In the *translation* step that ensues, special RNA/protein complexes called *ribosomes* bind to the starter regions of the message, and, for each of the $4^3 = 64$ possible triplets of base pairs of the message

[1] Letter A on the DNA original maps into a T equivalent of the message and vice versa; the same complementarity holds for the C and G pair.

(or *codons*) that they encounter as they "read" the message sequence, they attach one of the 20 amino-acids to the growing protein chain undergoing assembly (or terminate the process). After the synthesis is complete, the ribosome falls off the message and the finished polypeptide starts to fold into its active 3D conformation that will allow it to perform its structural or enzymatic role.

While this picture might be intuitively simple, it almost exclusively focuses on the process of *production* of proteins from the DNA, rather than on *regulation*, i.e. the molecular mechanisms that determine when, and how much, of each protein the cell will make. The genes are clearly not being constantly transcribed and translated into protein.[2] We find a clear demonstration of the importance of regulation when we recall that the cells of higher organisms, despite sharing the same DNA, differentiate into morphologically and functionally distinct types during development, and later give rise to various organs. It is due to the cells having expressed different sets of genes in a regulated way during their development history that this commitment to a specific cell fate occurs. Alternatively, both complex and simpler organisms, including bacteria, respond effectively to certain signals because they dynamically measure their environment and act by "switching" the relevant genes on or off. The example of the yeast Emergency Stress Response module from Chapter 2 provides a striking and genome-wide illustration of such a coherent response evoked by genetic regulation.

There are many ways in which the final amount of a certain protein might be regulated, and Fig 3.1 shows several kinds of regulatory mechanisms available to the cell. At a very early stage of gene expression the cell might make the relevant protein-coding segment of the DNA physically inaccessible to the transcriptional apparatus by packing it into highly condensed chromatin structures (in eukaryotic cells only). If unpacked, the *affinity* of the apparatus for the DNA can be drastically modified by the presence or absence of *transcription factors*, proteins that bind specific short sequences of DNA in order to facilitate or prevent transcription. It is this *transcriptional regulation* that we think of when we discuss turning some genes "off" or "on" and it will be the focus of our further exploration. Nevertheless, the regulation repertoire has by no means been exhausted yet. After the message is produced, it will be processed (*splicing* will remove certain nonsense stretches but sometimes also select among functional variants of the protein) and modified (chemically tagged for transport, physical localization, destruction etc). Finally, the protein itself can be regulated, either by having its lifetime controlled (e.g. through active degradation mechanism called *ubiquitination*) or by having its activity tuned through various chemical modifications (e.g. phosphorylation by activating enzymes).

The detailed study of transcriptional regulation started with the work of Jacob and Monod (1961), for which they were awarded the Nobel prize in 1965. The basic organization of prokaryotic DNA gradually emerged: genes (or in general *coding regions* that end up being transcribed/translated into proteins) comprise the majority of the DNA; in front of each gene or set of coregulated genes there is a noncoding piece of the DNA that contains operators and promoters. The promoter is the target binding site of the transcriptional apparatus (called the *RNA polymerase* or RNAP in prokaryotes); operators contain target binding sites for transcription factors (TFs). There are millions of short sequences of around 10 to 20 base-pairs to which a transcription factor could presumably bind in a bacterial genome, yet it will most often be found on its operator, because of the extremely favorable

[2] This unregulated and constant gene expression, known as *constitutive expression*, does happen for a few essential housekeeping genes.

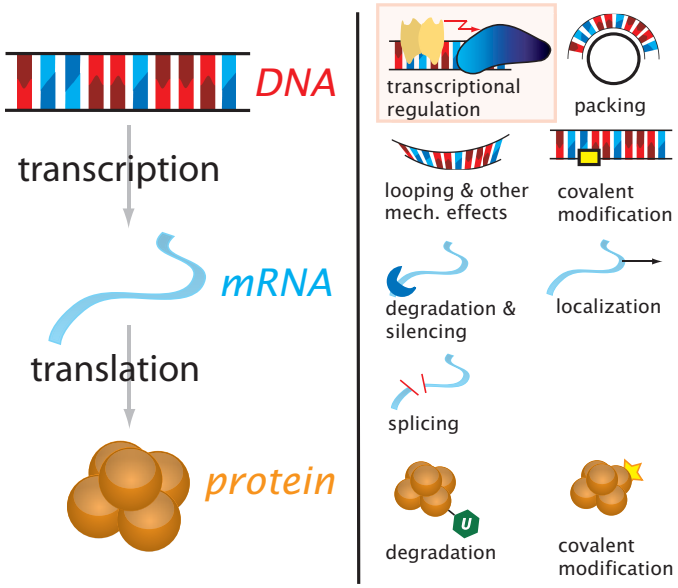


Figure 3.1: Information in the cell flows from the DNA to messenger RNA to proteins. Regulatory processes intervene at various stages and control the final amount of the protein expressed. Transcription factors bind operators on the DNA and influence the rate of transcription by recruiting the RNA polymerase complex, usually through favorable energetic interactions; packing of the DNA onto the histone spools can stop transcription by making genes inaccessible. Alternative mechanisms involve mechanical deformation of the DNA, as in protein-induced DNA looping, or covalent modifications of the DNA (methylation). On the messenger level, silencing and other controlled degradation mechanisms modulate the quantity of mRNA present; expression can also be varied by controlling the nuclear export of mRNA or by localizing the message in other ways; splicing can produce alternative protein versions from the same gene. Finally, proteins can be degraded in a highly controlled way, and can have their activity modified through covalent modification by other signaling enzymes.

sequence-dependent interaction energy that this particular TF has with its operator. The occupancy of the operator is determined by the amount of transcription factor, and it in turn determines the occupancy of the nearby promoter region by the RNA polymerase complex. The interaction between the RNAP and the TF is usually by means of a direct physical interaction by touching, which can either enhance the equilibrium occupancy of the promoter or reduce it (for example by TF physically obscuring the RNAP binding site). The situation in eukaryotes is biologically more complicated (usually involving what resembles a combinatorial code in the *enhancer* regions that are equivalent to the operators in bacteria) but the basic ideas remain similar.

In a seminal series of papers starting in 1987, Berg and von Hippel outlined the basic physics involved in the first step of transcriptional regulation, namely the interaction of the TF and the DNA (Berg and von Hippel, 1987). More specifically, they have identified three problems that need to be addressed satisfactorily by physical models of transcriptional regulation. In current terminology, they can be briefly summarized as follows:

- **The specificity problem.** The question here is about how the transcrption factor recognizes its operator site on the DNA. What is the physical mechanism by which a TF gains the binding energy difference between the specific operator site and the genomic background, such that the equilibrium probability of being on the operator

is high, despite an overwhelming number of non-specific, or distractor, sites? A good model should take into account also that the genomic background has some distribution of energies, rather than a single non-specific value. See Bintu et al. (2005) for equilibrium statistical mechanics calculations for various regulatory scenarios, Djordjevic et al. (2003) for the discussion of non-specific binding in the models of transcriptional regulation and binding site discovery, and Bakk and Metzler (2004) for estimates of the fraction of non-specifically bound CI / Cro proteins in *E. coli* inferred from the data. Kinney et al. (2007) have shown how to infer models of TF–DNA interaction from high-throughput data without making most of the usual assumptions.

- **The noise problem.** When the mean equilibrium occupancy of the operator is reached, the instantaneous occupancy nevertheless fluctuates in time as the TF binds and unbinds from the specific site. The mean occupancy is read out by the transcriptional machinery and it ultimately sets the mean level of gene expression. Similarly, the fluctuations in occupancy get transmitted to the output and are one of the contributors to the noise in gene expression. It turns out that there is a tradeoff – at least in simple models – between the specificity of the TF-DNA binding and the noise magnitude: highly specific TFs have large binding energies and stay on the binding sites for a long time before unbinding, thus triggering an increase in output noise, because these slow fluctuations are not effectively averaged away by downstream protein production steps. The noise problem deals with equilibrium fluctuations in the expression of the gene under transcriptional control, and their biological impact.

- **The search problem.** The search problem examines how a TF finds the binding site embedded in the DNA background within a time window consistent with available measurements.[3] If the cytoplasmic, or 3D, diffusion of TF is too slow and 1D diffusion along the arc-length of the DNA is the main method by which TFs translocate on the DNA, and if the binding energy landscape of the DNA is rough, then the total search time could be completely dominated by those periods when the TF stays ineffectively stuck on relatively strong, yet non-functional sites on the DNA (traps). To a large extent the negative impact of traps is affected by the mode of diffusion (either 1D along the DNA or 3D in solution) and by the energy landscape of the binding sites on the DNA (especially by the variance in the distribution of the non-specific site energies). The resolution of the search problem is to show that there exists an optimal combination of 1D and 3D diffusion strategies which can explain the observed short search times. See Halford and Marko (2004); Slutsky and Mirny (2004) for recent theoretical discussions, or Wang et al. (2006) for measurements.

As stated, our final goal is to compute the information capacity of simple regulatory elements, for which the characterization of noise is needed [Eq (2.5)]. We are therefore concerned with the second problem above, that is the *noise problem*, and the bulk of this chapter is devoted to the noise analysis in transcription factor binding, transcription and translation. While we do not discuss the other two problems listed above in detail, the mechanisms that have been put forward as their potential solutions in the literature can influence the noise behavior as well; this chapter therefore ends with the estimates on how

[3]These search times can be on the order of a minutes or less; if the diffusion limited on-rate, $k_+ = 4\pi Da$, is postulated, where D is the TF diffusion constant and a is the linear dimension of the target site, then for certain transcription factors like *lac* such an on-rate would be too slow to account for experimental observations.

the noise calculations have to be modified by including the non-specific binding sites and the 1D sliding along the DNA.

3.2 Input and output noise in transcriptional regulation[4]

3.2.1 Introduction

A number of recent experiments have focused attention on noise in gene expression (Elowitz et al., 2002; Ozbudak et al., 2002; Blake et al., 2003; Raser and O'Shea, 2004; Rosenfeld et al., 2005; Pedraza and van Oudenaarden, 2005; Golding et al., 2005; Newman et al., 2006; Bar-Even et al., 2006). The study of noise in biological systems more generally has a long history, with two very different streams of thought. On the one hand, observations of noise in behavior at the cellular or even organismal level give us a window into mechanisms at a much more microscopic level. The classic example of using noise to draw inferences about biological mechanism is perhaps the Luria–Delbrück experiment (Luria and Delbrück, 1943), which demonstrated the random character of mutations, but one can also point to early work on the nature of chemical transmission at synapses (Fatt and Katz, 1950, 1952) and on the dynamics of ion channel proteins (Lecar and Nossal, 1971a,b; Stevens, 1972; Conti et al., 1975). On the other hand, noise limits the reliability of biological function, and it is important to identify these limits. Examples include tracking the reliability of visual perception at low light levels down to the ability of the visual system to count single photons (Hecht et al., 1942; Barlow, 1981), the implications of channel noise for the reliability of neural coding (Verveen and Derksen, 1965, 1968; Schneidman et al., 1998), and the approach of bacterial chemotactic performance to the limits set by the random arrival of individual molecules at the cell surface (Berg and Purcell, 1977).

After demonstrating that one can observe noise in gene expression, most investigators have concentrated on the mechanistic implications of this noise. Working backward from the observation of protein concentrations, one can try to find the components of noise that derive from the translation of messenger RNA into protein, or the components that arise from noise in the transcription and degradation of the mRNA itself. At least in some organisms, a single mRNA transcript can give rise to many protein molecules, and this "burst" both amplifies the fluctuations in mRNA copy number and changes their statistics, so that even if the number of mRNA copies obeys the Poisson distribution the number of protein molecules will not (Paulsson, 2004). This discussion parallels the understanding that Poisson arrival of photons at the retina generates non–Poisson statistics of action potentials in retinal ganglion cells because each photon triggers a burst of spikes (Barlow et al., 1971). Recent large scale surveys of noise in eukaryotic transcription have suggested that the noise in most protein levels can be understood in terms of this picture, so that the fractional variance in the number of proteins g_i expressed from gene i is given by

$$\eta_\mathrm{i}^2 \equiv \frac{\langle(\delta g_\mathrm{i})^2\rangle}{\langle g_\mathrm{i}\rangle^2} = \frac{b}{\langle g_\mathrm{i}\rangle}, \tag{3.1}$$

where $b \sim 10^3$ is the burst size, and is approximately constant for all genes (Bar-Even et al., 2006).

The mechanistic focus on noise in transcription vs translation perhaps misses the functional role of gene expression as part of a regulatory network. Almost all genes are subject to transcriptional regulation, and hence the expression level of a particular protein can be viewed as the cell's response to the concentration of the relevant transcription factors. Seen in this way, transcription and translation are at the "output" side of the response, and the

[4]This section appeared on the arXiv as Tkačik et al. (2007a).

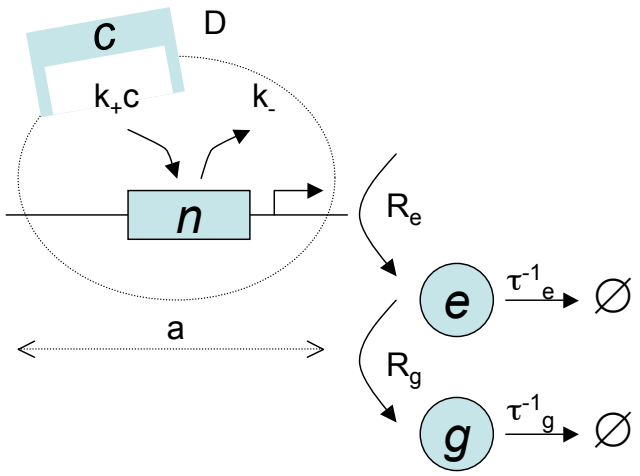


Figure 3.2: A simple model for transcriptional regulation. Transcription factor is present at an average concentration c, diffusing freely with diffusion constant D; it can bind to the binding site of linear dimension a and the fractional occupancy of this site is $n \in [0, 1]$. Binding occurs with a second order rate constant k_+, and unbinding occurs with a first order rate constant k_-. When the site is bound, the mRNA are transcribed at rate R_e and degraded with rate τ_e^{-1}, resulting in a number of transcripts e. Proteins are translated from each mRNA molecule with rate R_g and degraded with rate τ_g^{-1}, resulting in a copy number g.

binding of transcription factors to their targets along the genome is at the "input" side [Fig 3.2]. Noise can arise at both the input and output, and while fluctuations in transcription factor concentration could be viewed as an extrinsic source of noise (Elowitz et al., 2002; Swain et al., 2002), there will be fluctuations in target site occupancy even at fixed transcription factor concentration (Bialek and Setayeshgar, 2005; Walczak et al., 2005; van Zon et al., 2006). There is a physical limit to how much the impact of these input fluctuations can be reduced, essentially because any physical device that responds to changes in concentration is limited by shot noise in the diffusive arrival of the relevant molecules at their target sites (Berg and Purcell, 1977; Bialek and Setayeshgar, 2005, 2006).

In this chapter we revisit the relative contributions of input and output noise. Input noise has a clear signature, namely that its impact on the output protein concentration peaks at an intermediate value of the input transcription factor concentration. The analogous signature was essential, for example, in identifying the noise from random opening and closing of individual ion channels in neurons (Sigworth, 1977, 1980). Perhaps surprisingly, we show that this signature is easily obscured in conventional ways of plotting the data on noise in gene expression. Recent experiments on the regulation of Hunchback expression by Bicoid in the early *Drosophila* embryo (Gregor, 2005; Gregor et al., 2006a) are consistent with the predicted signature of input noise, and (although there are caveats) a quantitative analysis of these data supports a dominant contribution of diffusive shot noise. We discuss what experiments would be required to test this conclusion more generally. We begin, however, by asking whether any simple global model such as Eq (3.1) can be consistent with the imbedding of gene expression in a network of regulatory interactions.

3.2.2 Global consistency

Consider a gene i which is regulated by several transcription factors. In steady state, the mean number of these proteins in the cell will be a function of the copy numbers of all the relevant transcription factors:

$$\langle g_{\rm i} \rangle = f_{\rm i}(g_1, g_2, \cdots, g_{\rm K}) \tag{3.2}$$

If the copy numbers of the transcription factors fluctuate, this noise will propagate through the input/output relation f (Pedraza and van Oudenaarden, 2005; Hooshangi et al., 2005), so that

$$\langle(\delta g_{\rm i})^2\rangle = \sum_{\mu=1}^{\rm K}\sum_{\nu=1}^{\rm K}\frac{\partial f_{\rm i}}{\partial g_\mu}\frac{\partial f_{\rm i}}{\partial g_\nu}\langle\delta g_\mu\delta g_\nu\rangle + \langle(\delta g_{\rm i})^2\rangle_{\rm int}, \tag{3.3}$$

where we include the intrinsic noise $\langle(\delta g_{\rm i})^2\rangle_{\rm int}$ that occurs at fixed transcription factor levels.

If the noise in gene expression is dominated by the processes of transcription and translation, and if the transcription factors are not regulating each other, then the correlations between fluctuations in the copy numbers of different proteins will be very small, so we expect that

$$\langle\delta g_\mu\delta g_\nu\rangle = \delta_{\mu\nu}\langle(\delta g_\mu)^2\rangle. \tag{3.4}$$

This allows us to simplify the propagation of noise in Eq (3.3) to give

$$\langle(\delta g_{\rm i})^2\rangle = \sum_{\mu=1}^{\rm K}\left(\frac{\partial f_{\rm i}}{\partial g_\mu}\right)^2\langle(\delta g_\mu)^2\rangle + \langle(\delta g_{\rm i})^2\rangle_{\rm int}. \tag{3.5}$$

If, as in Eq (3.1), we express the noise in protein copy number as a fractional noise η, then this becomes

$$\eta_{\rm i}^2 = \sum_{\mu=1}^{\rm K}\left(\frac{\partial\log f_{\rm i}}{\partial\log g_\mu}\right)^2\eta_\mu^2 + \eta_{\rm i,int}^2. \tag{3.6}$$

In particular, this means that there is a minimum level of noise,

$$\eta_{\rm i}^2 \geq \sum_{\mu=1}^{\rm K}\left(\frac{\partial\log f_{\rm i}}{\partial\log g_\mu}\right)^2\eta_\mu^2. \tag{3.7}$$

But if the fractional variance in protein copy number has a simple, global relation to the mean copy number, as in Eq (3.1) (Bar-Even et al., 2006), then this simplifies still further:

$$\frac{b}{\langle g_{\rm i}\rangle} \geq \sum_{\mu=1}^{\rm K}\left(\frac{\partial\log f_{\rm i}}{\partial\log g_\mu}\right)^2\frac{b}{\langle g_\mu\rangle} \tag{3.8}$$

$$\Rightarrow 1 \geq \sum_{\mu=1}^{\rm K}\left(\frac{\partial\log f_{\rm i}}{\partial\log g_\mu}\right)^2\frac{\langle g_{\rm i}\rangle}{\langle g_\mu\rangle}. \tag{3.9}$$

Since the proteins labeled by the indices μ represent transcription factors, usually present at low concentrations, and the protein i is a regulated gene—such as a structural or metabolic protein—but not a transcription factor itself, one expects that $\langle g_{\rm i}\rangle/\langle g_\mu\rangle \gg 1$. But then we have

$$\sum_{\mu=1}^{\rm K}\left(\frac{\partial\log f_{\rm i}}{\partial\log g_\mu}\right)^2 \ll 1. \tag{3.10}$$

Since this inequality constrains the sum of squares of terms, each must be much smaller than one. This means that when we make a small change the concentration of any transcription factor, the response of the regulated gene must be much less than proportional. In this sense, the assumption of a simple global description for the level of noise in gene expression, Eq (3.1), leads us to the conclusion that transcriptional "regulation" can't really be very

effective, and this must be wrong. Notice that this problem is independent of the burst size b, and hence doesn't depend on whether the noise is dominated by transcription or translation.

Our conclusion from the inequality in Eq (3.10) is that we should re–examine the original hypothesis about noise [Eq (3.1)]. An alternative is that this hypothesis is correct, but that there are subtle correlations among all the protein copy number fluctuations of all the different transcription factors. If we want the global output model to be correct, these correlations would have to take on a very special form—different transcription factors regulating a single gene would have to be correlated in a way that matches their impact on the expression of that gene—which seems implausible but would be very interesting if it were true.

3.2.3 Sources of noise

Figure 3.2 makes clear that the concentration of a protein can fluctuate for many reasons. The processes of synthesis and degradation of the protein molecules themselves are discrete and stochastic, as are the synthesis and degradation of mRNA molecules; together these constitute the "output noise" which has been widely discussed. But if we are considering a gene whose transcription is regulated, we need a microscopic model for this process. For the case of a transcriptional activator, there are binding sites for the transcription factors upstream of the regulated gene, and when these sites are occupied transcription proceeds at some rate, but when the site is empty transcription is inhibited. Because there are only a small number of relevant binding sites (in the simplest case, just one), the occupancy of these sites must fluctuate, and this random switching is an additional source of noise. In addition, the binding of transcription factors to their target sites along the genome depends on the concentration in the immediate neighborhood of these sites, and this fluctuates as molecules diffuse into and out of the neighborhood.

All of the different processes described above and schematized in Fig 3.2 can be analyzed analytically using Langevin methods, and the predictions of this analysis can be tested against detailed stochastic simulations. The details of the analysis are given in the Methods section [Methods A.3]. Notice that variations in cell size, protein sorting in cell division, fluctuations in RNA polymerase and ribosome concentrations, and all other extrinsic contributions to the noise are neglected.

When the dust settles, the variance in protein copy number σ_g^2 can be written as a sum of three terms, which correspond to the output, switching, and diffusion noise. To set the scale, we express the copy number as a fraction of its maximum possible mean value, g_0, which is reached at high concentrations of the transcriptional activator. In these units, we find

$$\left(\frac{\sigma_g}{g_0}\right)^2 = \frac{1+R_g\tau_e}{g_0}\bar{g} + \frac{(1-\bar{g})^2\,\bar{g}}{k_-\tau_g} + \frac{(1-\bar{g})^2\,\bar{g}^2}{\pi Dac\tau_g} \tag{3.11}$$

where $\bar{g} = \langle g\rangle/g_0$ is the protein copy number expressed as a fraction of its maximal value, c is the concentration of the transcription factor, and other parameters are as explained in Fig 3.2.

The first term in Eq (3.11) is the output noise and has a Poisson–like behavior, with variance proportional to the mean, but the proportionality constant differs from 1 by $R_g\tau_e$, i.e. the burst size or the number of proteins produced per mRNA (Paulsson, 2004). This is just the simple model of Eq (3.1), with $b = 1 + R_g\tau_e$.

The second term in Eq (3.11) originates from binomial "switching" as the transcription factor binding site occupation fluctuates, and is most closely analogous to the noise from random opening and closing of ion channels. This term will be small for unbinding rates k_- that are fast compared to the protein lifetime, but might be large for factors that take a long time to equilibrate or that form energetically stable complexes on their promoters.

The third term in Eq (3.11) arises because the diffusive flux of transcription factor molecules to the binding site fluctuates at low input concentration c; in effect the receptor site "counts" the number of molecules arriving into its vicinity during a time window τ_g, and this number is of the order $\sim Dac\tau_g$. This argument is conceptually the same as that for the limits to chemoatractant detection in chemotaxis, as discussed by Berg and Purcell (1977). It can be shown that this is a theoretical noise floor that cannot be circumvented by using sophisticated "binding site machinery" as long as this machinery is contained within a region of linear size a (Bialek and Setayeshgar, 2005, 2006). For example, cooperative binding to the promoter or promoters with multiple internal states will modify the binomial switching term, but will leave the diffusion noise unaffected if we express it as an effective noise in transcription factor concentration σ_c such that

$$\sigma_{\rm g} = \left|\frac{\partial g}{\partial c}\right|\sigma_c. \tag{3.12}$$

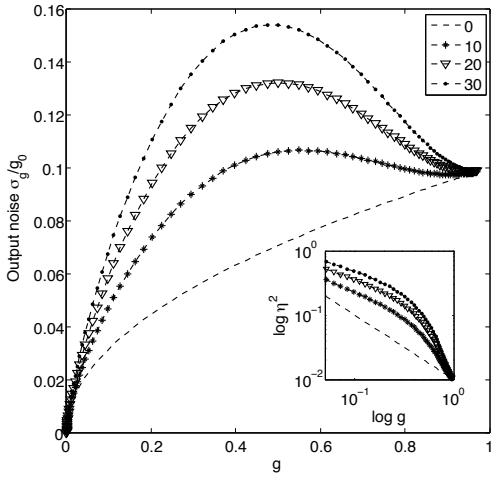


Figure 3.3: Expression noise as a function of the mean. The standard deviation of the protein concentration $\sigma_{\rm g}/g_0$ is plotted against the mean protein concentration $\bar{g} = \langle g\rangle/g_0$, from Eq (3.14) with $h = 5$. In all cases the output noise term has a strength $\alpha = 0.01$, and the different curves are indexed by the ratio of input noise to output noise $\beta/\alpha = 0, 10, 20, 30$. In the absence of input noise, the noise level is a monotonic function of the mean, but input noise contributes a peak near the point of half maximal expression $\bar{g} = 0.5$. In the inset, we show the same results plotted as a fractional noise variance $\eta^2_{\rm g}$ vs the mean [Eq (3.15)], on a logarithmic scale, and we see that the prominent peak has become just an inflection. For most of the dynamic range of means, the contribution of input noise is to increase the fractional variance without substantial changes in the slope of the double–log plot, so that we can confuse input noise with a larger level of output noise, especially if we remember that real data will be scattered due to measurement errors.

Although cooperativity does not change the effective concentration noise due to diffusion, it does reduce the relative significance of the switching noise (Bialek and Setayeshgar, 2006). Since we will discuss a system which is strongly cooperative, in much of what follows

we neglect the switching noise term and focus on the output noise and diffusion noise. Then the generalization to multisite, cooperative regulation is straightforward [Methods A.3.2]. We expect that cooperative effects among h transcription factors generate a sigmoidal dependence of expression on the transcription factor concentration, so that

$$\bar{g} = \frac{c^h}{c^h + K_d^h}, \tag{3.13}$$

where h is called the Hill coefficient, and K_d is the concentration required for half maximal activation. We can invert this relationship to write the concentration c, which is relevant for the diffusive noise, as a function of the mean fractional expression level $\bar{g}$. Substituting back into Eq (3.11), and neglecting the switching noise, we obtain

$$\left(\frac{\sigma_g}{g_0}\right)^2 = \alpha\,\bar{g} + \beta\,\bar{g}^{2-1/h}(1-\bar{g})^{2+1/h}, \tag{3.14}$$

where α and β are combinations of parameters that measure the strength of the output and diffusion noise, respectively. If we express the variance in fractional terms, this becomes

$$\eta_g^2 = \alpha\frac{1}{\bar{g}} + \beta\bar{g}^{-1/h}(1-\bar{g})^{2+1/h}. \tag{3.15}$$

The global output noise model of Eq (3.1) corresponds to $\beta = 0$ (no input noise) and $b = \alpha g_0$. Figure 3.3 shows the predicted noise levels for different ratios of output to input noise (β/α).

For very highly cooperative, essentially switch–like systems, we can take the limit $h \to \infty$ to obtain

$$\left(\frac{\sigma_g}{g_0}\right)^2 = \alpha\,\bar{g} + \beta\,\bar{g}^2(1-\bar{g})^2 \tag{3.16}$$

$$\eta_g^2 = \alpha\frac{1}{\bar{g}} + \beta(1-\bar{g})^2. \tag{3.17}$$

In particular, if we explore only expression levels well below the maximum ($\bar{g} \ll 1$), then the diffusion noise just add a constant β to the fractional variance. Thus, diffusion noise in a highly cooperative system could be confused with a global or even extrinsic noise source.

3.2.4 Signatures of input noise

Input noise arises from fluctuations in the occupancy of the transcription factor binding sites. Thus, if we go to very high transcription factor concentrations, where all sites are fully occupied, or to very low concentrations, where the sites are never occupied, the fluctuations must vanish. These limits correspond, in the case of a transcriptional activator, to maximal and minimal expression levels, respectively. Thus, the key signature of input noise is that it must be largest at some intermediate expression level, as shown in Fig 3.3.

The claim that many genes have expression noise levels which fit the global output noise model of Eq (3.1) would seem to contradict the prediction of a peak in the noise as a function of the mean. But if we plot the predictions of the model with input noise as a fractional variance vs mean, the prominent peak disappears (inset to Fig 3.3). In fact, over a large dynamic range, the input noise seems just to increase the magnitude of the

fractional variance while not making a substantial change in the slope of $\log(\eta_g^2)$ vs $\log(\langle g\rangle)$. Confronted with real data on a system with significant input noise, we could thus fit much of those data with the global output noise model but with a larger value of b. There is, of course, a difference between input and output noise, even when plotted as $\log(\eta_g^2)$ vs $\log(\langle g\rangle)$, namely a rapid drop in noise level as we approach maximal expression. But this effect is confined to a narrow range, essentially a factor of two in mean expression level. As we discuss below, there are variety of reasons why this might not have been seen in the data of Bar-Even et al. (2006).

Recent experiments on the precision of gene expression in the early *Drosophila* embryo provide us with an opportunity to search for the signatures of input noise (Gregor, 2005; Gregor et al., 2006a). The embryo contains a spatial gradient of the protein Bicoid (Bcd), translated from maternal mRNA, and this protein is a transcription factor which activates, among other genes, *hunchback*. Looking along the anterior–posterior axis of the embryo one thus has an array of nuclei that experience a graded range of transcription factor concentrations. Using antibody staining and image processing methods, it thus is possible to collect thousands of points on a scatter plot of input (Bicoid concentration) vs. output (Hunchback protein concentration); since even in a single embryo there are many nuclei that have the same Bcd concentration, one can examine both the mean Hunchback (Hb) response and its variance; data from Gregor et al. (2006a) are shown in Fig 3.4.

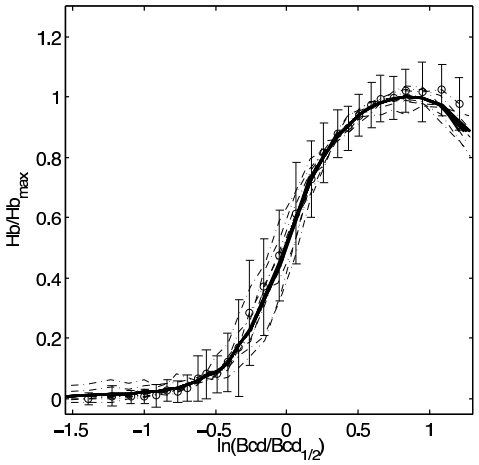


Figure 3.4: The input–output relation for Bicoid regulation of Hunchback expression, redrawn from Gregor et al. (2006a). Dashed curves show mean expression levels in different embryos, thick black line is the mean across all embryos, and points with error bars show the mean and standard deviation of Hb expression at a given Bcd concentration in one embryo.

The mean response of Hb to Bcd is fit reasonably well by Eq (3.13) with a Hill coefficient $h = 5$ (Gregor et al., 2006a), and in Fig 3.5 we replot the noise in this response as a function of the mean. The peak of expression noise near half maximal expression—the signature of input noise—is clearly visible. More quantitatively, we find that the data are well fit by Eq (3.14) with the contribution from output noise ($\alpha \approx 1/380$) much smaller than that from input noise ($\beta \approx 1/2$). We also consider the same model with $h \to \infty$, and this fully switch–like model, although formally still within error bars, systematically deviates from the data. Finally we consider a model in which diffusion noise is absent, but we include the switching noise from Eq (3.11), which generalizes to the case of cooperative binding (see Methods). Interestingly, this model has the same number of parameters as the diffusion noise model, but does a significantly poorer job of fitting the data. While the fit can be improved further by adding a small background to the noise, we emphasize that Eq (3.14) correctly captures the non–trivial shape of the noise curve with only two parameters. Because input noise falls to zero at maximal expression, the sole remaining noise at that

point is the output noise, and this uniquely determines the parameter α. The strength of the input noise (β) then is determined by the height of the noise peak, and there is no further room for adjustment. The *shape* of the peak is predicted by the theory with no additional parameters, and the different curves in Fig 3.5 demonstrate that the data can distinguish among various functional forms for the peak.

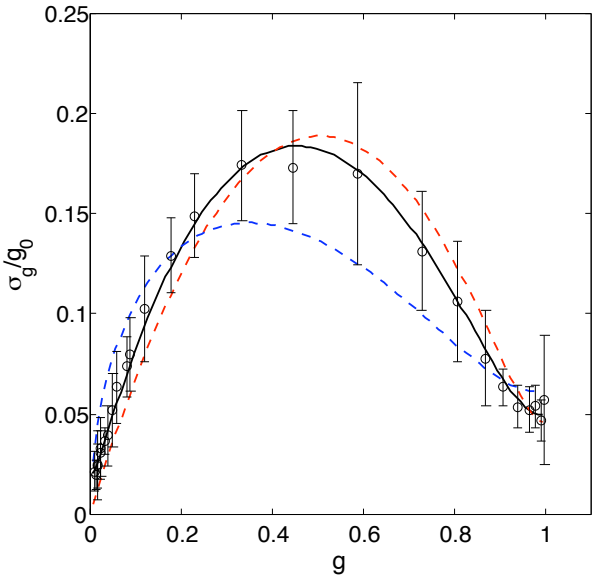


Figure 3.5: Standard deviation of Hunchback expression as a function of the mean (points with error bars), replotted from Gregor et al. (2006a). The black line is a fit of combined output and diffusion noise contributions, from Eq (3.14) with $h = 5$, and the dashed red line is with $h \to \infty$, from Eq (3.16). In contrast, the dashed blue line is the best fit of combined output and switching noise contributions. Although both diffusion and switching noise produce a peak at intermediate expression levels, the shapes of the peaks are distinguishable, and the data favor the diffusion noise model.

Are the parameters α and β that fit the Bcd/Hb data biologically reasonable? The fact that diffusive noise dominates at intermediate levels of expression ($\beta \gg \alpha$) is the statement that the Hunchback expression level provides a readout of Bcd concentration with a reliability that is close to the physical limit set by diffusional shot noise, as was argued in Gregor et al. (2006a) based on the magnitude of the noise level and estimates of the relevant microscopic parameters that determine β. The dominance of diffusive noise over switching noise presumably is related to the high cooperativity of the Bcd/Hb input/output relation (Bialek and Setayeshgar, 2006).

The parameter α measures the strength of the output noise and thus depends on the absolute number of Hb molecules and on the number proteins produced per mRNA transcript. If this burst size is in the range $R_g\tau_e \sim 1 - 10$, then our fit predicts the maximum expression level of Hb corresponds to $g_0 = 700 - 4000$ molecules in the nucleus. Given the volume of the nuclei at this stage of development ($\sim 140\,\mu\text{m}^3$; see Gregor et al. (2006a,b)), this is a concentration of $8 - 48\,\text{nM}$. Although we don't have independent measurements of the absolute Hunchback concentration, this is reasonable for transcription factors, which typically act in the nanoMolar range (Ptashne, 1992; Pedone et al., 1996; Ma et al., 1996; Burz et al., 1998; Winston et al., 1999; Zhao et al., 2002), and can be compared with the maximal nuclear concentration of Bcd, which is $55 \pm 3\,\text{nM}$ (Gregor et al., 2006a). Larger burst sizes would predict larger maximal expression levels, or conversely measurements of absolute expression levels might give suggestions about the burst size for translation in the early *Drosophila* embryo.

3.2.5 Discussion

In the process of transcriptional regulation, the (output) expression level of regulated genes acts as a sensor for the (input) concentration of transcription factors. The performance of this sensor, and hence the regulatory power of the system, is limited by noise. While changes in the parameters of the transcriptional and translational apparatus can change the level of output noise, the input noise is determined by the physical properties of the transcription factor and its interactions with the target sites along the genome. Ultimately, there is a lower bound on this input noise level set by the shot noise in random arrival of the transcription factors at their targets, in much the same way that any imaging process ultimately is limited by the random arrival of photons.

Input and output noise seem to be so different that it is hard to imagine that they could be confused experimentally. Some of the difficulty, however, can be illustrated by plotting the results from the Bcd/Hb experiments of Gregor et al. (2006a) in the form which has become conventional in the study of gene expression noise, as a fractional variance vs mean expression level [Fig 3.6]. The signature of input noise, so clear in Fig 3.5, now is confined to a narrow range ($\sim \times 2$) near maximal expression. In contrast, over more than a decade of expression levels the noise level is a good fit to $\eta_g^2 \propto \langle g \rangle^{-\gamma}$, with $\gamma = 1.04$ being very similar to the prediction of the global noise model ($\gamma = 1$) in Eq (3.1). The departures from power–law behavior are easily obscured by global noise sources, experimental error, or by technical limitations that lead to the exclusion of data at the very highest expression levels, as in Bar-Even et al. (2006).

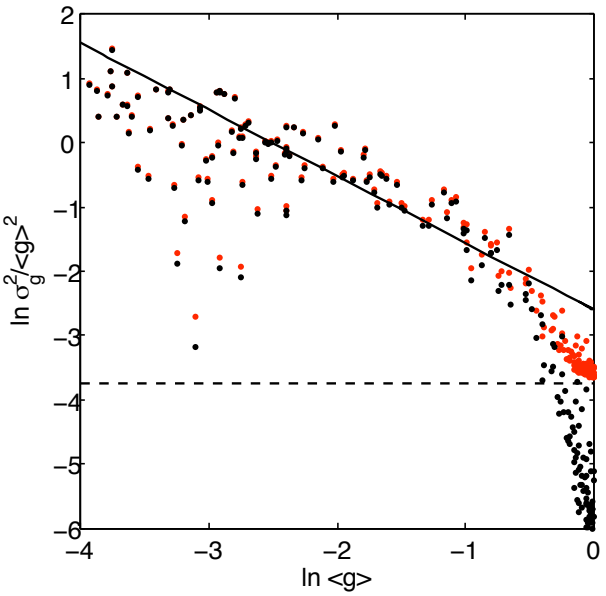


Figure 3.6: Logarithmic plot of fractional variance vs the mean expression level for Hunchback, replotted from Gregor et al. (2006a). Each black point represents the noise level measured across nuclei that experience the same Bcd concentration within one embryo, and results are collected from nine embryos. The solid line shows a fit to $\eta_g^2 \propto \langle g \rangle^{-\gamma}$ in the region below half maximal mean expression; we find a good fit, with $\gamma = 1.04$, despite the fact that these data show a clear signature of input noise when plotted in Fig 3.5. Dashed line indicates the global noise floor suggested in Bar-Even et al. (2006), and red points show the raw data with this variance added. Although the input noise still appears as a drop in fractional noise level near maximal mean expression, this now is quite subtle and easily obscured by experimental errors.

The lesson from this analysis of the Bicoid/Hunchback data is that the signatures of input noise are surprisingly subtle. In this system, however, the behavior near half maximal expression is exactly the most relevant question biologically, since this is where the "deci-

sion" is made to draw a boundary, as a first step in spatial patterning. In other systems, the details of noise in this region of expression levels might be less relevant for the organism, but it is only in this region that different sources of noise are qualitatively distinguishable, as is clear from Fig 3.6. Thus, unless we have independent experiments to measure some of the parameters of the system, we need experimental access to the full range of expression levels and hence, implicitly, to the full dynamic range of transcription factor concentrations, if we want to disentangle input and output noise.

The early *Drosophila* embryo is an attractive model system precisely because the organism itself generates a broad range of transcription factor concentrations, and conveniently arranges these different samples along the major axes of the embryo. A caveat is that since we don't directly control the transcription factor concentration, we have to measure it. In particular, in order to measure the variance of the output (Hunchback, in the present discussion) we have to find many nuclei that all have the same input transcription factor (Bicoid) concentration. Because the mean output is a steep function of the input, errors in the measurement of transcription factor concentration can simulate the effects of input noise, as discussed in Gregor et al. (2006a). Thus, a complete analysis of input and output noise requires not only access to a wide range of transcription factor concentrations, but rather precise measurements of these concentrations.

Why are the different sources of noise so easily confused? If noise is dominated by randomness in a single step of the translation process, then the number of protein molecules will obey the Possion distribution, and the variance in copy number will be equal to the mean. But if we can't actually turn measurements of protein level into molecule counts, then all we can say is that the variance will be *proportional* to the mean. If the dominant noise source is a single step in transcription, then the number of mRNA transcripts will obey the Poisson distribution, and the variance of protein copy numbers still will be proportional to the mean, but the proportionality constant will be enhanced by the burst size. The same reasoning, however, can be pushed further back: if, far from maximal expression, the dominant source of noise is the infrequent binding of a transcriptional activator (or dissociation of a repressor) to its target site, then the variance in protein copy number still will be proportional to the mean. Thus, the proportionality of variance to mean implies that there is some single rare event that dominates the noise, and by itself doesn't distinguish the nature of this event.

If noise is dominated by regulatory events, then the number of mRNA transcripts should be drawn from a distribution broader than Poisson. In effect the idea of bursting, which amplifies protein relative to mRNA number variance, applies here too, amplifying the variance of transcript number above the expectations from the Poisson distribution. Transcriptional bursting has in fact been observed directly (Golding et al., 2005), although it is not clear whether this arises from fluctuations in transcription factor binding or from other sources.

Previous arguments have made it plausible that input noise is significant in comparison to the observed variance of gene expression (Bialek and Setayeshgar, 2005), and we have shown here that models which assign all of the noise to common factors on the output side are inconsistent with the embedding of gene expression in a regulatory network. The signatures of input noise seem clear, but can be surprisingly subtle to distinguish in real data. We have argued that the Bicoid/Hunchback system provides an example in which input noise is dominant, and further that the detailed form of the variance vs mean supports a dominant role for diffusion rather than switching noise. Although there are caveats, this is consistent with the idea that, as with other critical biological processes (Barlow, 1981; Berg and Purcell, 1977; Bialek, 1987, 2002), the regulation of gene expression can operate

with a precision limited by fundamental physical principles.

3.3 Alternative noise sources

3.3.1 Effects of non-specific binding sites

We start by briefly introducing the statistical mechanical framework in which the TF–DNA interaction is usually discussed.[5] Different sites on the DNA – where by a site we mean a consecutive sequence of L base pairs starting at a particular location in the genome – to which a transcription factor can bind are viewed as distinct energy states of the TF–DNA system. The interaction energy is a complicated function of the relative position of the protein with respect to the DNA and carries contributions from several sources: the main contribution is usually taken to be the direct, or sequence-dependent, energy, which arises from specific hydrogen contacts between the amino acid residues of the TF and the DNA bases; to get favorable energetic contribution, the TF and DNA therefore have to be in a proper spatial alignment. There are also indirect contributions to the binding energy: the non-specific attractive electrostatic interaction, the mechanical energy of deformation if the TF bends the DNA and so on, which here simply redefine the zero of the sequence dependent energy.

Once the energy E_ν of a site ν is known, the probability of that site being bound, or the site *occupancy*, is easy enough to calculate:

$$n_\nu = \frac{1}{1 + e^{E_\nu - \mu}}, \tag{3.18}$$

where all energies have been expressed in units of $k_B T$; μ is the chemical potential of the *free* transcription factors in the cytoplasm, and is proportional to the logarithm of their concentration c.[6] By inspecting the equation for the mean occupancy [Eq (A.12)] and comparing it to Eq (3.18), we recognize that $e^{E_\nu - \mu} = K_d^\nu / c = k_-^\nu / k_+^\nu c$, where k_- and k_+ are the on- and off- rates for TF binding and unbinding at site ν and K_d^ν is its equilibrium binding constant. Here the dynamical and equilibrium pictures connect.

What remains to be determined is the dependence of the energy of the site, E_μ, on the site sequence, $\vec{\sigma}_\mu = \{\sigma(1), \sigma(2), \ldots, \sigma(L)\}_\mu$, where $\sigma_\mu(i)$ is one of the four letters of the genetic alphabet. We can always write this energy as a series:

$$E(\vec{\sigma}) = \sum_i \epsilon_{i\sigma(i)} + \frac{1}{2} \sum_{ij} J_{ij\sigma(i)\sigma(j)} + \cdots . \tag{3.19}$$

Here, the lowest order approximation is written as an energy matrix ϵ (of dimension $L \times 4$), which parametrizes the linear contribution to the binding energy: a base $\sigma_\mu(i)$ at position i of the site μ adds an amount to the energy that is independent of bases at other positions. Note that this "expansion" is not a systematic expansion in some small parameter; in principle the dominant contribution could come from any order in the series, although in practice this does not seem to be the case (Robinson et al., 1998).

Berg and von Hippel (1987) presented a simple argument by which the entries of ϵ can be computed if we know a set of sites $\vec{\sigma}_\mu$ to which the transcription factor binds strongly,

[5] Equilibrium calculations would by convention precede the noise calculations, but as modeling (and inferring from the data) of the interaction energy between DNA and protein is a whole separate field and such models were not necessary for simple computations of the input and output noise, the issue was postponed until now.

[6] The free concentration is the total concentration minus the number of transcription factors bound over all sites. We will return to this point shortly.

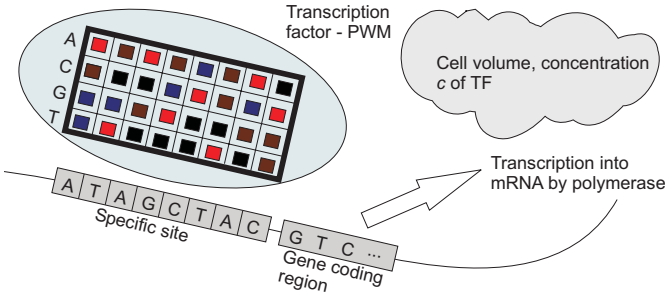


Figure 3.7: A depiction of the transcription factor binding to its specific site on the DNA. The energy of the interaction is described by the position weight matrix, with red (black) elements denoting favorable (unfavorable) interactions. In the example presented here, the site on the DNA exactly corresponds to the consensus sequence of the TF (red entries in the matrix), and the binding energy is the sum of the corresponding red elements.

under some assumptions.[7] In the limit where the occupancies are not saturated, i.e. the concentration is small compared to the relevant equilibrium binding constants, K_d^μ, the matrix ϵ is given by:

$$\epsilon_{ib} = -\log \frac{N_i^b}{\max_b N_i^b}, \tag{3.20}$$

$$N_i^b = N_0 + \sum_\mu \delta_{\sigma_\mu(i),b}, \tag{3.21}$$

where N_i^b is the count of the number of times the base $b = \{A, C, G, T\}$ appears at position i in the set of known binding sites $\vec{\sigma}_\mu$, and N_0 is the so-called "pseudo-count" regularization parameter (usually 1). Kinney et al. (2007) have later relaxed some of the assumptions required by the Berg and von Hippel construction, at the expense of increasing the amount of data needed to do proper inference of ϵ; in addition they argue that if the Berg – von Hippel assumptions do not hold, the energy matrix and the matrix computed from counts in Eq (3.20), also called *position weight matrix or PWM*, are no longer simply related. High-throughput essays like chromatin immunoprecipitation (ChIP) or protein binding microarrays (PBM) can provide datasets big enough for such analyses.

Mean response

Let the system be composed of N binding sites (with density $\rho = N/V$, where the volume of cell nucleus is V) indexed by μ, with kinetic parameters k_+^μ and k_-^μ and occupancies m_μ. The site equilibrium dissociation constants are denoted by $K_\mu = k_-^\mu / k_+^\mu$. The functional binding site embedded in this background has a dissociation constant K_d and we give it a reference binding energy of 0, with the nonspecific site μ having a relative energy of E_μ in units where $k_B T = 1$ with respect to this reference, so that $K_\mu = K_d \exp(E_\mu)$. Then the

[7] Assumptions are: the specific sites have approximately the same affinity; the genetic background, or the set of nonspecific sites, is random; positions within the binding site are independent; the selection that acts on the sites in order to preserve their function is a linear function of the energy.

relation between the total and free concentration must be:

$$c_t = c_f + \frac{1}{V}\sum_{\mu}\frac{c_f}{c_f + K_\mu} \tag{3.22}$$

$$\approx c_f + \rho\int dE\, p(E)\frac{c_f}{c_f + K_d\exp(E)}, \tag{3.23}$$

where we have replaced the sum over the sites with the integral over the distribution $p(E)$ of binding energies relative to the reference (functional) binding site. If the system is in the weak binding regime with $c_f \ll \langle K_\mu\rangle$, then the fractional occupancy under the integral can be approximated with linear response, and therefore:

$$c_t \approx c_f\left\{1 + \frac{\rho}{K_d}\int dE\, e^{-E}p(E)\right\} \tag{3.24}$$

If the nucleus were bathed in an extremely high concentration of TF, all non-specific sites would be occupied and therefore we have to get $c_t = c_f + \rho$ in the saturated limit of Eq (3.23).

We will assume that the distribution of the binding energies, $p(E)$, is a Gaussian with mean $\bar{E}$ and variance σ^2. For simple physical models of DNA-TF interaction, for instance the energy matrix model [first term of Eq (3.19)], where each position in the binding site contributes independently to the binding energy, the energies on a random genome will indeed be normally distributed. To evaluate the integral in Eq (3.24) we remember that for Gaussian distributions $\langle\exp(ikx)\rangle = \exp\left(ik\langle x\rangle - \frac{1}{2}k^2\sigma^2\right)$ and therefore (with $x = iE$ and $k = \beta = 1$):

$$c_t = c_f\left\{1 + \frac{\rho}{K_d}\exp\left(-\bar{E} + \frac{1}{2}\sigma^2\right)\right\} \tag{3.25}$$

We are interested in examining how the specificity of the functional binding site, K_d, influences the relation between the total and free concentration. At first sight it appears that the relation is linear, and if the affinity of the nonspecific binding sites, $K_{\rm NS} = K_d e^{\bar{E}}$ and σ are held fixed, K_d cancels out of the expression Eq (3.25), and the spread of the distribution effectively just renormalizes the mean nonspecific affinity. Alternatively, if $\sigma = 0$, we get the simplest model that makes all the sites except for the functional site have the same affinity, $K_{\rm NS}$. We can, however, also expect σ^2 to scale with $\bar{E}$: this is evident if we consider the simplest model that, for each position in the binding site, assigns energy 0 if that position matches the specific functional site, and assigns energy Δ otherwise. In such a model the expected discrimination energy and the variance under the assumed random genetic background, of the binding site of length L, will be

$$\bar{E} = \frac{3}{4}\Delta L \tag{3.26}$$

$$\sigma_E^2 = \frac{3}{16}\Delta^2 L \tag{3.27}$$

Since $\bar{E}$ is related to K_d at constant $K_{\rm NS}$, and we assume $\sigma_E^2 \propto \bar{E}$, we get the dependence on K_d of the relation between the total and free concentrations in Eq (3.25):

$$c_t = c_f\left\{1 + \frac{\rho}{K_{\rm NS}}\left(\frac{K_{\rm NS}}{K_d}\right)^{\eta}\right\}, \tag{3.28}$$

where $\eta > 0$ is a parameter that depends on the proportionality constant between the mean energy and its variance. The important message here is that even though the distribution of nonspecific site energies has a finite width, when K_d is increased and the specificity of the functional site is thus decreased, the number of TFs stuck to the nonspecific sites, or $c_t - c_f$, will not start to increase in a super-linear fashion for K_d greater than some critical K_d^*. On the contrary, a linear relationship between the total and free concentrations is expected to hold, even when weak-binding limit is no longer valid, although possibly with a large proportionality constant; e.g. there are (perhaps typical) cases when 90% of the TF molecules can be nonspecifically bound to the DNA, cf. Bakk and Metzler (2004).

Fluctuations

The fluctuations in the occupancy of the binding sites are coupled to each other by diffusion in a limited geometry, such that there are no fluxes through the volume boundaries and the total number of particles (those diffusing in free solution and those bound on the binding sites) is held fixed. The equations for the system are as follows:

$$\frac{dm_\mu}{dt} = k_+^\mu c(\mathbf{x}_\mu)(1-m_\mu) - k_-^\mu m_\mu + \xi_\mu \tag{3.29}$$

$$\frac{\partial c(\mathbf{x},t)}{\partial t} = D\nabla^2 c(\mathbf{x},t) - \sum_\mu \frac{dm_\mu}{dt}\delta(\mathbf{x}-\mathbf{x}_\mu) + \eta(\mathbf{x},t). \tag{3.30}$$

Here, Eq (3.29) describes the binding and unbinding to the site at location $\mathbf{x}_\mu$, and Eq (3.30) is a diffusion equation for the free concentration with the additional terms that take into account the possibility that the molecule gets absorbed to or released from a binding site, cf. Eq (A.7). We linearize and Fourier-transform:

$$(-i\omega + 1/\tau_\mu)\,\delta m_\mu = k_+(1-\bar{m}_\mu)\delta c(\mathbf{x}_\mu) + \xi_\mu \tag{3.31}$$

$$\left(-i\omega + k^2 D\right)\delta c_{\mathbf{k}} = i\omega \sum_\mu \delta m_\mu e^{i\mathbf{k}\mathbf{x}_\mu} + \eta(\mathbf{k},\omega). \tag{3.32}$$

The terms ξ_μ and $\eta(\mathbf{x},t)$ represent Langevin noise, and they are normalized as follows:

$$\langle \xi_\mu(t)\xi_\mu(t')\rangle = \left(k_-^\mu m_\mu + k_+^\mu c(1-m_\mu)\right)\delta(t-t') \tag{3.33}$$

$$\langle \eta(\mathbf{k},\omega)\eta^*(\mathbf{k}',\omega')\rangle = 4Dk^2\bar{c}\delta(\mathbf{k}-\mathbf{k}')\delta(\omega-\omega'). \tag{3.34}$$

Expressing δm_μ from Eq (3.31) and inserting it into Eq (3.32), we get a coupled system of equations for the modes of concentration fluctuations:

$$\left(-i\omega + k^2 D\right)\delta c_{\mathbf{k}} = \eta(\mathbf{k},\omega) + \sum_\mu e^{i\mathbf{k}\mathbf{x}_\mu}\frac{i\omega}{-i\omega+1/\tau_\mu}\left(k_+^\mu(1-\bar{m}_\mu)\frac{1}{V}\sum_{k'}\delta c_{\mathbf{k}'}e^{-i\mathbf{k}'\mathbf{x}_\mu} + \xi_\mu\right), \tag{3.35}$$

where we have written $\delta c(\mathbf{x}_\mu) = \frac{1}{V}\sum_{\mathbf{k}'}\delta c_{\mathbf{k}'}e^{-i\mathbf{k}'\mathbf{x}_\mu}$. To get the power spectrum at low frequency $S_c(\omega\to 0) = \lim_{\omega\to 0}\sum_{\mathbf{k}}\langle\delta c_{\mathbf{k}}(\omega)\delta c_{\mathbf{k}}^*(\omega)\rangle$ we have to treat separately two cases, namely $\mathbf{k}=0$ and all others. In case of non-zero mode, the only surviving term after the limit is taken is the noise η term, from which it follows that:

$$\delta c_{\mathbf{k}} = \frac{\eta(\mathbf{k},\omega)}{Dk^2}, k \neq 0. \tag{3.36}$$

On the other hand, the conservation of total particle number gives us for zero mode, in the small ω limit:

$$\langle \delta c_0 \delta c_0^* \rangle = \frac{1}{\left(1 + \frac{1}{V}\sum_\mu k_+^\mu \tau_\mu (1-\bar{m}_\mu)\right)^2} \left\{ \langle \eta \eta^* \rangle_{\mathbf{k}=0} + \sum_\mu \tau_\mu^2 \langle \xi_\mu \xi_\mu^* \rangle \right. \tag{3.37}$$

$$+ \frac{1}{V^2} \sum_{\mu,\nu,\mathbf{k}'\neq 0,\mathbf{k}''\neq 0} k_+^\mu (1-\bar{m}_\mu) k_+^\nu (1-\bar{m}_\nu) \times \tag{3.38}$$

$$\left. \times \tau_\mu \tau_\nu \frac{4\bar{c}\delta(\mathbf{k}'-\mathbf{k}'')}{Dk'^2} e^{-i\mathbf{x}_\mu \mathbf{k}' + i\mathbf{x}_\nu \mathbf{k}''} \right\}. \tag{3.39}$$

Contribution from diffusion noise η is 0 at zero mode. The summation does not extend over $k = 0$, because this term has been absorbed into the denominator. Collapsing the double sum into the single sum using the delta function, the last term can be written separately as a contribution when $\mu = \nu$ and otherwise, as follows:

$$\langle \delta c_0 \delta c_0^* \rangle = \frac{1}{\left(1 + \frac{1}{V\bar{c}}\sum_\mu \bar{m}_\mu (1-\bar{m}_\mu)\right)^2} \left\{ 2\sum_\mu \tau_\mu \bar{m}_\mu (1-\bar{m}_\mu) + \right. \tag{3.40}$$

$$+ \frac{2}{\pi D a \bar{c}} \sum_\mu \bar{m}_\mu^2 (1-\bar{m}_\mu)^2 + \tag{3.41}$$

$$\left. + \frac{4}{\bar{c}} \sum_{\mu,\nu} \bar{m}_\mu (1-\bar{m}_\mu) \bar{m}_\nu (1-\bar{m}_\nu) \frac{1}{V} \sum_{\mathbf{k}\neq 0} \frac{e^{i\mathbf{k}(\mathbf{x}_\mu - \mathbf{x}_\nu)}}{Dk^2} \right\}. \tag{3.42}$$

In the above expression, a is the size of the binding site, i.e. the corresponding k-cutoff in sum over momenta is π/a. Assuming that the position of the binding site is not correlated with its affinity, we can rewrite the full noise spectrum at zero frequency in terms of site averages:

$$\langle |\delta c|^2 \rangle = \frac{1}{V^2} \sum_{\mathbf{k}} \langle |\delta c_{\mathbf{k}}|^2 \rangle \tag{3.43}$$

$$= \frac{\bar{c}}{\pi D a} + \frac{1}{(1 + \frac{N}{V\bar{c}} \langle \bar{m}(1-\bar{m}) \rangle_m)^2} \left\{ \frac{2N}{V^2} \langle \tau \bar{m}(1-\bar{m}) \rangle_m + \right. \tag{3.44}$$

$$\left. + \frac{\bar{c}}{\pi D a} \frac{2N}{(V\bar{c})^2} \langle \bar{m}^2 (1-\bar{m})^2 \rangle_m + \frac{\bar{c}}{\pi D b} \frac{N^2}{(V\bar{c})^2} \langle \bar{m}(1-\bar{m}) \rangle_m^2 \right\} \tag{3.45}$$

Angular brackets with subscript m, i.e. $\langle \cdots \rangle_m$ stand for an average over all sites. Length scale b depends on the geometry of the binding sites, where it is assumed that the position of the site is uncorrelated with its affinity:

$$\frac{1}{4\pi b} = \left\langle \frac{1}{V} \sum_{\mathbf{k}} \frac{e^{i\mathbf{k}\mathbf{x}}}{k^2} \right\rangle_m = \frac{1}{4\pi} \left\langle \frac{1}{|\mathbf{x}|} \right\rangle_{\mathbf{x}} \tag{3.46}$$

We can rewrite the result in terms of fractional fluctuations in the input concentration

at the position of the specific site:

$$\left(\frac{\sigma_c}{\bar{c}}\right)^2 = \frac{1}{\pi D a \bar{c} \tau}\left(1 + \frac{1}{N}\frac{\Gamma_1^2}{(1+\Gamma_0)^2} + \frac{a}{b}\frac{\Gamma_0^2}{(1+\Gamma_0)^2}\right) + + \frac{1}{N}\frac{\Gamma_2^2}{(1+\Gamma_0)^2} \quad (3.47)$$

$$\Gamma_0 = \frac{N}{V\bar{c}}\langle \bar{m}(1-\bar{m})\rangle \quad (3.48)$$

$$\Gamma_1 = \frac{N}{V\bar{c}}\sqrt{\langle \bar{m}^2(1-\bar{m})^2\rangle} \quad (3.49)$$

$$\Gamma_2 = \frac{N}{V\bar{c}}\sqrt{2\langle \frac{\tau_m}{\tau}\bar{m}(1-\bar{m})\rangle} \quad (3.50)$$

We recognize the diffusion input noise contribution as the leading term in the parenthesis of Eq (3.47); the other terms are contributed by the nonspecific sites and vanish if $N = 0$. Moreover, we can put reasonable bounds on the new terms: the fractions of Γ-averages in Eq (3.47) are at most of order 1, and their magnitude is therefore upper-bounded by their prefactors. These, however, are clearly small: $1/N$ is the inverse of the number of nonspecific sites, a/b is the ratio between the receptor size a and a typical distance between the sites b, and $\tau_m/\tau \ll 1$, because the nonspecific site occupancies presumably equilibrate on a much shorter timescale than the integrating time (minutes or more). The conclusion here is that the nonspecific sites will not significantly increase the diffusion contribution to the input noise.

3.3.2 Effects of TF diffusion along the DNA

We have already computed the noise power spectrum in fractional occupancy n of the specific transcription binding site located at position x_0 on the DNA due to the binomial and diffusion flux fluctuations [Eq (A.34)]:

$$S_n(\omega \to 0) = \frac{2\bar{n}(1-\bar{n})^2}{k_-} + \frac{2\bar{n}^2(1-\bar{n})^2}{2\pi D a \bar{c}}. \tag{3.51}$$

The result is derived in the low frequency (long integration time) limit, assuming an average bulk concentration $\bar{c}$ of free transcription factor molecules in the cell, which bind and unbind to the specific site in a one-step process, as described by Eqs (A.7, A.8). In particular, the diffusion constant D that enters into these equations is the bulk, or 3D, diffusion constant for the stochastic motion of TF molecules in the cytoplasm.

Noise in Eq (3.51) has two contributions: the first one arises from the fact that the binding and unbinding state is a binary process, which has a binomial variance – this term would exist even if the local concentration at x_0 were perfectly fixed; the second contribution is a consequence of concentration fluctuations at x_0 around the mean $\bar{c}$.

Here, we analyze the case where the transcription factor can attach non-specifically to the DNA, slide for a certain length along its contour (while being nonspecifically bound to it and performing a 1D random walk), and either bind to the specific site, or dissociate from the DNA back into bulk solution. The situation is illustrated in Fig 3.8.

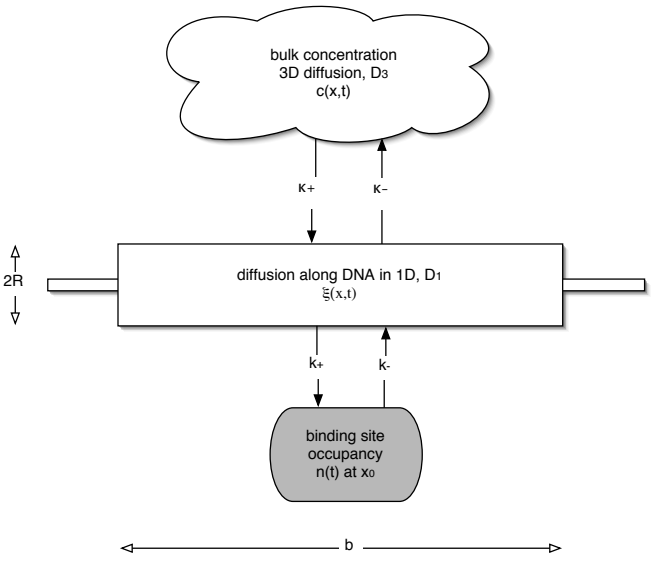


Figure 3.8: The transcription factors can either be free in solution at concentration c, or they can enter a region on the DNA where they diffuse by sliding. The effective 1D concentration is denoted by ξ and is position and time dependent. The specific binding site on the DNA is at location x_0; k_+ and k_- are the on- and off-rates for transition from/to 1D "solution" ξ. The effective radius of the DNA molecule is R and the "sliding length", or average distance along the contour covered in the 1D random walk before dissociation, is denoted by b.

We describe the system by the following set of equations:

$$\frac{dn}{dt} = k_+\xi(x_0,t)(1-n) - k_- n \tag{3.52}$$

$$\frac{\partial \xi(x,t)}{\partial t} = D_1 \frac{\partial^2 \xi(x,t)}{\partial^2 x} - \frac{dn}{dt}\delta(x-x_0) + \tag{3.53}$$

$$+ \kappa_+ \int dy\, dz\, c(\mathbf{x},t)\delta(y)\delta(z) - \kappa_- \xi(x,t) \tag{3.54}$$

$$\frac{\partial c(\mathbf{x},t)}{\partial t} = D_3 \nabla^2 c(\mathbf{x},t) - \kappa_+ c(\mathbf{x},t)\delta(y)\delta(z) + \kappa_- \xi(x,t)\delta(y)\delta(z) \tag{3.55}$$

We have assumed that DNA is stretched along the x-axis and that it is an infinitely thin molecule – as we will see soon, we will need to regularize the Dirac-delta functions (see Fig 3.8 for explanation of the length symbols R, b etc). ξ is a function of only one variable, x, while c is a function of all three spatial coordinates. κ_- is the rate at which TF dissociates from the non-specific binding mode into the bulk; κ_+ is the corresponding on-rate per unit length.

We need to linearize and Fourier-transform the equations, which is straightforward, apart from Eq (3.55), where we have a product of $c(x,y,z,t)\delta(y)\delta(z)$ that turns into a convolution of Fourier-transforms. This effectively couples the equation for wave-mode $\mathbf{k}$ with other wave-modes, making problem difficult to solve analytically. However, we can proceed as follows. First, the delta functions are really just approximations for finite-size regions and we are really trying to take the Fourier-transform of

$$c(x,y,z,t)H_R(y)H_R(z)$$

Here $H_R(y)$ is a Heaviside function that is 1, if its argument is in the interval $[-R,R]$, and zero outside. We write:

$$\begin{aligned} c(x,y,z,t)H_R(y)H_R(z) &= \int\int\int \frac{dk_x}{2\pi}\frac{dk_y}{2\pi}\frac{dk_z}{2\pi} c(k_x,k_y,k_z,t)e^{-ik_x x - ik_y y - ik_z z} \times \\ &\times \frac{1}{(2\pi)^2}\int\int dk'_y dk'_z H_R(k'_y)H_R(k'_z)e^{-ik'_y y - ik'_z z} \\ &= \frac{1}{(2\pi)^5}\int dk'_y dk'_z d^3\mathbf{k} c(k_x, k_y - k'_y, k_z - k'_z, t)e^{-i\mathbf{kx}}H_R(k'_y)H_R(k'_z) \\ &\approx \int \frac{d^3\mathbf{k}}{(2\pi)^5}\left(\frac{2\pi}{R}\right)^2 \cdot c(\mathbf{k},t). \end{aligned} \tag{3.56}$$

The approximation we make is that since $H_R(y)$ goes to 0 for absolute y larger than R, its Fourier transform, $H_R(k'_y)$ has to go to zero for $|k'_y| > \frac{\pi}{R}$. In the region where it is not 0, we assume that the integrand can be evaluated at $k'_y = k'_z = 0$ in the first approximation, integrated over the allowed range for primed momenta. The Fourier transform of the whole term is then simply $\approx \frac{1}{R^2}c_{\mathbf{k}}(t)$.

Using this result we can write down the complete linearized and transformed set of Eqs (3.54, 3.55):

$$-i\omega\delta c_{\mathbf{k}} = -k^2 D_3 \delta c_{\mathbf{k}} - \frac{\kappa_+}{R^2}\delta c_{\mathbf{k}} + \kappa_-\delta\xi \tag{3.57}$$

$$-i\omega\delta\xi = -k_x^2 D_1\delta\xi + i\omega\delta n e^{ik_x x_0} - \kappa_-\delta\xi + \kappa_+\int \delta c_{\mathbf{k}}\frac{dk_y dk_z}{(2\pi)^2}. \tag{3.58}$$

Here, $\delta\xi$ is function of k_x only. Let us evaluate the last term of the second equation, by expressing it from the first equation and integrating:

$$\delta c_{\mathbf{k}} = \kappa_-\delta\xi\frac{1}{-i\omega + k^2 D_3 + \frac{\kappa_+}{R^2}} \tag{3.59}$$

$$\int \frac{dk_y dk_z}{(2\pi)^2}\delta c_{\mathbf{k}} = \kappa_-\delta\xi\frac{1}{(2\pi)^2}\int_0^\infty \frac{2\pi k_r dk_r}{-i\omega + (k_x^2 + k_r^2)D_3 + \kappa_+/R^2} \tag{3.60}$$

$$= \frac{\kappa_-\delta\xi}{4\pi D_3}\ln\left\{1 + \left(\left(\frac{k_x}{\Lambda}\right)^2 + \frac{\kappa_+}{D_3(R\Lambda)^2}\right)^{-1}\right\}. \tag{3.61}$$

In the second step, we transformed the integration over k_y and k_z into a radial integration over k_r, where $k_r^2 = k_y^2 + k_z^2$. In addition, we discarded the $i\omega$ term, because we are calculating the noise in the small ω limit. In the third step, we had to introduce ultra-violet cutoff at $\Lambda = \frac{\pi}{R}$.

This result can be plugged back into the equation for $\delta\xi(x_0) = \int \frac{dk_x}{2\pi}\delta\xi(k_x)e^{-ik_x x_0}$:

$$\delta\xi(x_0) = i\omega\delta n \int \frac{dk}{2\pi} \frac{1}{k^2 D_1 + \kappa_- \left(1 - \frac{\kappa_+}{4\pi D_3}\ln\{\cdots\}\right)}. \tag{3.62}$$

Before we insert the result into Fourier transform of Eq (3.52), let us briefly recapitulate how one obtains the noise power spectrum, starting with the linearized equation for occupancy responding to the thermal fluctuation δF, as in Bialek and Setayeshgar (2005):

$$\begin{aligned} -i\omega\delta n &= -k_-\delta n - k_+\bar{\xi}\delta n + k_+(1-\bar{n})\delta\xi(x_0) - \beta k_-\bar{n}\delta F & (3.63)\\ \delta n &= \frac{\beta k_-\bar{n}\delta F}{-i\omega + k_- + k_+\bar{\xi} - k_+(1-\bar{n})i\omega/(\pi\Lambda D_1)I}, & (3.64)\end{aligned}$$

where I is a rewritten form of integral in Eq (3.62):

$$\begin{aligned} I(\alpha,\beta) &= \int_0^\infty \frac{dt}{t^2 + \beta\left(1 - \alpha\ln\{1 + (t^2 + 4\alpha/\pi)^{-1}\}\right)}, & (3.65)\\ \alpha &= \frac{\kappa_+}{4\pi D_3}, & (3.66)\\ \beta &= \frac{\kappa_-}{\Lambda^2 D_1}. & (3.67)\end{aligned}$$

Observe that in the calculation with 3D diffusion only, Eq (3.64) would have exactly the same form with different dimensional parameters (rates, diffusion constants) standing next to the integral I. Therefore, we can use the same expression for noise power spectrum without rederiving it here, by simply replacing the case of 3D integral with combined 1D/3D integral, cf. Eq (3.51):

$$S_n(\omega\to 0) = \frac{2\bar{n}(1-\bar{n})}{k_+\bar{\xi} + k_-} + \frac{2\bar{n}^2(1-\bar{n})^2}{\pi\Lambda D_1\bar{\xi}} I(\alpha,\beta). \tag{3.68}$$

We can use the equilibrium conditions to eliminate the average 1D concentration ξ and estimate the rates from dimensionality arguments (see the text below for the explanation):

$$\begin{aligned} \bar{n} &= \frac{k_+\bar{\xi}}{k_+\bar{\xi} + k_-}, & (3.69)\\ \kappa_+\bar{c} &= \kappa_-\bar{\xi}, & (3.70)\\ \kappa_+ &\approx 4\pi D_3, & (3.71)\\ b^2 &\approx D_1\kappa_-^{-1}, & (3.72)\end{aligned}$$

where b is the typical sliding length along the DNA before the TF dissociates into the solution. This finally allows us to rewrite the result in a more intuitive form:

$$S_n(\omega\to 0) = \frac{2\bar{n}(1-\bar{n})^2}{k_-} + \frac{\bar{n}^2(1-\bar{n})^2}{D_3 R\bar{c}}\beta I(\alpha,\beta). \tag{3.73}$$

The result looks similar to the pure 3D diffusion case: the noise due to concentration fluctuations has its length scale a (receptor size in pure 3D case) replaced with R (effective DNA cross-section in mixed 1D/3D case), and the noise contribution term gets multiplied by the "structure factor" $\beta I(\alpha, \beta)$. What is the meaning of parameters α and β?

Parameter $\alpha = \frac{\kappa_+}{4\pi D_3}$ is close to 1 for diffusion-limited approach to DNA. Imagine that the area that TF attempts to hit in order to stick non-specifically to the DNA is a cylindrical segment of DNA with radius R and length b. Length scale b is the average 1D diffusion length, $b^2 = D_1 t_{\text{diff}} \approx D_1 \kappa_-^{-1}$. If we were treating the cylindrical DNA segment as a sphere of radius b, then Smoluchowski limit $\tilde{\kappa}_+ = 4\pi D_3 b$ would aply. In the first approximation, the on rate κ_+ (note that κ_+ is the rate per unit length of the DNA, which effectively is b) would then be $4\pi D_3$. Since the DNA segment is a cylinder and not a sphere, this reasoning is not exact, and an effective length scale $\approx (R^2 b)^{1/3}$ would probably be a better approximation. Regardless of the exact geometrical factors, however, it turns out that α has a smaller effect on the result than β.

What is β? Written out in terms of b defined above, β is:

$$\beta = \frac{\kappa_- R^2}{\pi^2 D_1} = \left(\frac{R}{\pi b}\right)^2. \tag{3.74}$$

We see that β is approximately the square of the ratio between the cross-section of the 1D cylinder (the "target" that 3D diffusion has to hit) and the average "sliding length" along the DNA. R must be of order of several nanometers; while b is, at DNA stacking length of $a = 0.3\,\text{nm}$ per base-pair and 100 bp average diffusion length (Halford and Marko, 2004; Slutsky and Mirny, 2004), around $b = 10 - 100\,\text{nm}$. It is therefore not unreasonable to assume that the factor β could be as low as $\beta \sim 10^{-3} - 10^{-2}$, and the corresponding decrease in noise variance relative to pure 3D diffusion, $\beta\, I(\alpha, \beta)$, is shown in Fig 3.9.[8]

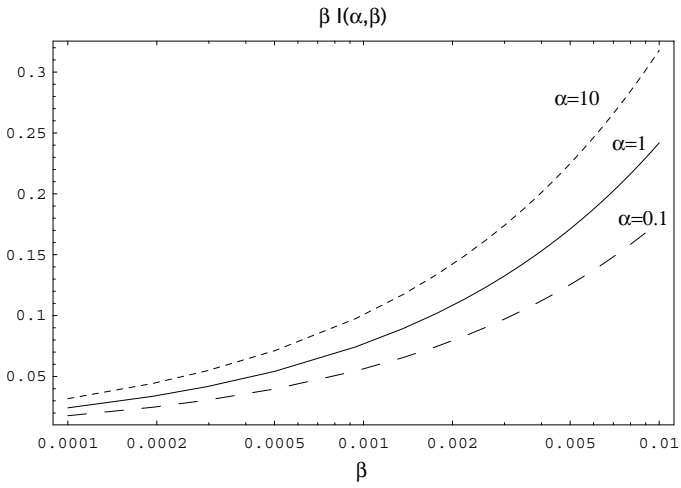


Figure 3.9: The relative decrease in noise variance, compared to the pure 3D diffusion model, as a function of parameters α and β. Three values for α are shown, spanning two orders of magnitude; β covers the relevant range if typical 1D diffusion length is as expected from search time optimality arguments (order hundred base pairs).

[8] An alternative diffusion noise calculation is possible if we only had 1D diffusion, and no coupling to the 3D cytoplasmic bath in the model. In that case, interestingly, we cannot take the limit $\omega \to 0$ in the power spectrum, but the noise variance is finite; it is suppressed relative to pure 3D diffusion by a factor $\sim \frac{R}{b}\sqrt{k_- \tau}$, which could be of order one for biologically relevant parameters.

3.4 Summary

In this chapter we have analyzed the gene expression noise in an activator, one of basic building blocks of genetic regulatory networks. Starting with a simplified mechanistic picture of transcriptional regulation and using the Langevin approximations, we were able to decompose the total noise in protein levels into the input and output contributions. The input part arises from the fluctuations in transcription factor binding site occupancy and the diffusive flux of TF molecules towards it, and the output part from the stochastic production of mRNA and protein. Comparison with precise measurements of the noise in the fruit fly Bicoid-Hunchback system supports the claim that the input noise can significantly contribute to the total noise. A further argument that this conclusion is valid more generally is provided by simple theoretical considerations. These show that global noise models, where output noise is the sole major contributor to the total noise, are inconsistent with the embedding of such elements into a regulatory network. Recent experiments that claim consistency with the global noise models might have wrongly assigned part of the input noise to the output, and such misattribution could easily occur if the experiment does not probe the full range of input TF concentrations.

We have further shown that the presence of non-specific binding modifies the relation between free and total transcription factor concentrations significantly, but probably still linearly; and that such sites do not appreciably increase the noise on the input side. In contrast, if transcription factor executes the binding site search in an optimal combination of 1D and 3D diffusion, the diffusive contribution to the input noise can change considerably.

In the process of analyzing the fruit fly data we have also created a family of biologically plausible noise models, defined by a small number of parameters, and differing in the relative contributions of various noise sources. In the following and last chapter of the thesis we will examine regulatory elements similar to the one dissected here, by systematically exploring the space of the elements' noise models and computing their corresponding information capacities.

Chapter 4

Building networks that transmit information

4.1 Introduction

Networks of interacting genes coordinate complex cellular processes, such as responding to stress, adapting the metabolism to a varying diet, maintaining the circadian cycle or producing an intricate spatial arrangement of differentiated cells during development. The success of such regulatory modules is at least partially characterized by their ability to produce reliable, stereotyped responses to repeated stimuli or changes in environment, and to perform the genetic computations reproducibly, either on a day-by-day or generation timescale. In doing so the regulatory elements are confronted by noise arising from physical processes that implement such genetic computations, and this noise ultimately traces its origins back to the fact that the state variables of the system are concentrations of chemicals and "computations" are really reactions between chemical species, usually present at low copy numbers.

It is useful to picture the regulatory module as a device that given some input computes an output, which in our case will be a set of expression levels of regulated genes. Sometimes the inputs to the module are easily identified, such as when they are the actual chemicals that a system detects and responds to, for example chemoattractant molecules, hormones or transcription factors. There are cases, however, when it is beneficial to think about the inputs on a more abstract level: in early patterning we talk of positional information and think of the regulatory module as trying to produce a different gene expression footprint at each spatial location; alternatively, circadian clocks generate distinguishable gene expression profiles corresponding to various phases of the day. Regardless of whether we regard the input as a physical concentration of some transcription factor or perhaps a position within the embryo, and whether the computation is complicated or as simple as an inversion produced by a repressor, we want to quantify its reliability in the presence of noise, and ask what the biological system can do to maximize it.

If we make many observations of a genetic regulatory element in its natural conditions we are collecting a sample drawn from a distribution $p(\mathcal{I}, \mathcal{O})$, where $\mathcal{I}$ describes the state of the input and $\mathcal{O}$ the state of the output. Saying that the system is able to produce a reliable response $\mathcal{O}$ across the spectrum of naturally occurring input conditions $p(\mathcal{I})$ amounts to saying that the dependency – either linear or strongly non-linear – between the input and output is high, i.e. far from random. Shannon has shown how to associate a unique measure,

the mutual information I of Eq (2.5), with the notion of dependency between two quantities drawn from a joint distribution:

$$I(\mathcal{I},\mathcal{O}) = \iint d\mathcal{I}\, d\mathcal{O}\, p(\mathcal{I},\mathcal{O}) \log_2 \frac{p(\mathcal{I},\mathcal{O})}{p(\mathcal{I})p(\mathcal{O})} \tag{4.1}$$

The resulting quantity is a measure in bits and is essentially the logarithm of the number of states in the input that produce distinguishable outputs given the noise. A device that has one bit of capacity can be thought of as an "on-off" switch, two bits correspond to four distinguishable regulatory settings and so on. Although the input is usually a continuous quantity, such as nutrient concentration or phase of the day, the noise present in the regulatory element corrupts the computation and does not allow the arbitrary resolution of a real-valued input to propagate to the output; instead, the mutual information tells us how precisely different inputs are distinguishable to the organism.

Experimental or theoretical characterization of the joint distribution, $p(\mathcal{I},\mathcal{O})$, for a regulatory module can be very difficult if the inputs and outputs live in a high-dimensional space. We can proceed, nevertheless, by remembering that the building blocks of complex modules are much simpler, and finally must reduce to the point where a single gene is controlled by transcription factors that bind to its promoter region and tune the level of its expression. While taking a simple element out of its network will not be illuminating about how the network as a whole behaves in general – especially if there are feedback loops – there may be cases where the information flow is "bottlenecked" through a single gene, and its reliability will therefore limit that of the network. In addition, the analysis of a simple regulatory element will provide directions for taking on more complicated systems.

Our aim in this chapter is therefore to try to understand the reliability of a simple genetic regulatory element, that is, of a single activator or repressor transcription factor controlling the expression level of its downstream gene. We will identify the concentration c of the transcription factor as an input, $\mathcal{I} = \{c\}$, and the expression level of the downstream gene g as the output, $\mathcal{O} = \{g\}$. The regulatory element itself will be parametrized by input/output kernel, $p(g|c)$, i.e. the distribution (as opposed to a "deterministic" function $g = g(c)$ in case of a noiseless system) of possible outputs given that the input is fixed to some particular level c. For each such kernel, we will then compute the maximum amount of information, $I(c;g)$, that can be transmitted through it, and examine how it depends on the properties of the kernel.

4.2 Maximizing information transmission

Our central idea is the realization that the input/output kernel of a simple regulatory element, $p(g|c)$, is determined by the biophysics of transcription factor-DNA interaction, transcription and translation, whereas the distribution of inputs, $p(c)$, that the cell uses during its "typical" lifetime, is free for the cell to change. The cell's transcription factor expression footprint is its representation of the environment and internal state, and the form of this representation can be the target of adaptation or evolutionary processes. Together, the input/output kernel and the distribution of inputs define the joint distribution, $p(c,g) = p(g|c)\,p(c)$, and consequently the mutual information of Eq (4.1) between the input and the output, $I(c;g)$.

Maximizing the information between the inputs and outputs, which corresponds to our notions of reliability in representation and computation, will therefore imply a specific

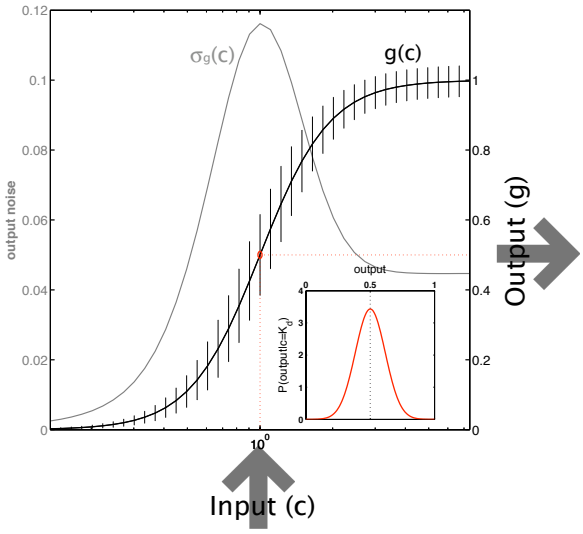


Figure 4.1: A schematic diagram of a simple regulatory element. Each input is mapped to a mean output according to the input/output relation (thick sigmoidal black line). Because the system is noisy, the output fluctuates about the mean. This noise is plotted in gray as a function of the input and shown in addition as error bars on the mean input/output relation. Inset shows the probability distribution of outputs at half saturation, $p(g|c = K_d)$ (red dotted lines); in this simple example we assume that the distribution is Gaussian and therefore fully characterized by its mean and variance.

matching between the given input/output kernel and the distribution of inputs, $p(c)$, that is being optimized. If one believes that a specific regulatory element has been tuned for maximal information transmission, then the optimal solution for the inputs, $p^*(c)$, and the resulting optimal distribution of outputs, $p^*(g) = \int dc\, p(g|c)p^*(c)$, become experimentally verifiable predictions. If, on the other hand, the system is not really maximizing information transmission, then the largest capacity achievable with a given kernel and its optimal input distribution, $I[p(g|c), p^*(c)]$, can still be regarded as a (hopefully revealing) upper bound on the true information capacity of the system.

During the past decades the measurements of regulatory elements have focused on recovering the mean response of a gene under the control of a transcription factor that had its activity modulated by experimentally adjustable levels of inducer or inhibitor molecules. Typically, a sigmoidal response is observed with a single regulator, as in Fig 4.1, and more complicated regulatory "surfaces" are possible when there are two or more simultaneous inputs to the system (Setty et al., 2003). In our notation, these experiments measure the conditional average over the distribution of outputs, $\bar{g}(c) = \int dg\, g\, p(g|c)$. Efforts to characterize the noise in gene expression were renewed by theoretical work of Swain et al. (2002) that has shown how to separate intrinsic and extrinsic components of the noise, i.e. the noise due to the stochasticity of the observed regulatory process in a single cell, and the noise contribution that arises because typical experiments make many single-cell measurements and the internal chemical environments of these cells differ across the population. Consequently, the work exploring the noise in gene expression, or $\sigma_g^2(c) = \int dg\, (g - \bar{g})^2 p(g|c)$, has begun to accumulate, on both the experimental and biophysical modeling side.

4.2.1 Small noise approximation

We start by showing how the optimal distributions can be computed analytically if the input/output kernel is Gaussian and the noise is small, and proceed by presenting the exact numerical solution later. Let us assume then that the first and second moments of the conditional distribution are given, and write the input/output kernel as a set of Gaussian

distributions $\mathcal{G}(g;\bar{g}(c),\sigma_g(c))$, or explicitly:

$$p(g|c) = \frac{1}{\sqrt{2\pi\sigma_g^2(c)}} \exp\left\{ - \frac{[g-\bar{g}(c)]^2}{2\sigma_g^2(c)} \right\}, \tag{4.2}$$

where both the mean response, $\bar{g}(c)$, and the noise, $\sigma_g(c)$, depend on the input, as illustrated in Fig 4.1.

We rewrite the mutual information between the input and the output of Eq (4.1) in the following way [Eq (2.5)]:

$$\begin{aligned} I(c;g) &= \int dc\, p(c) \int dg\, p(g|c) \log_2 p(g|c) - \\ &- \int dc\, p(c) \int dg\, p(g|c) \log_2 p(g). \end{aligned} \tag{4.3}$$

The first term can be evaluated exactly for Gaussian distributions, $p(g|c) = \mathcal{G}(g;\bar{g}(c),\sigma_g(c))$. The integral over g is just the calculation of the (negative of the) entropy of the Gaussian, and the first term therefore evaluates to $-\langle S\left[\mathcal{G}(g;\bar{g},\sigma_g]\right\rangle_{p(c)} = -\frac{1}{2}\langle \log 2\pi e\sigma_g^2(c)\rangle_{p(c)}$.

In the second term of Eq (4.3) the integral over g can be viewed as calculating $\langle \log_2 p(g)\rangle$ under the distribution $p(g|c)$. For an arbitrary continuous function $f(g)$ we can expand the integrals with the Gaussian measure around the mean:

$$\begin{aligned} \langle f(g)\rangle_{\mathcal{G}(g;\bar{g},\sigma_g)} &= \int dg\, \mathcal{G}(g) f(\bar{g}) + \\ &+ \int dg\, \mathcal{G}(g) \frac{\partial f}{\partial g}\Big|_{\bar{g}} (g-\bar{g}) + \\ &+ \frac{1}{2} \int dg\, \mathcal{G}(g) \frac{\partial f^2}{\partial g^2}\Big|_{\bar{g}} (g-\bar{g})^2 + \cdots \end{aligned} \tag{4.4}$$

The first term of the expansion simply evaluates to $f(\bar{g})$. The series expansion would end at the first term if we were to take the small noise limit, $\lim_{\sigma_g\to 0} \mathcal{G}(g;\bar{g},\sigma_g) = \delta(g-\bar{g})$. The second term of the expansion is zero because of symmetry, and the third term evaluates to $\frac{1}{2}\sigma_g^2 f''(\bar{g})$. We apply the expansion of Eq (4.4) and compute the second term in the expression for the mutual information, Eq (4.3), with $f(g) = \log_2 p(g)$. Taking only the zeroth order of the expansion, we get

$$I(c;g) = -\int dc\, p(c) \left[\log\sqrt{2\pi e}\sigma_g(c) + \log p(\bar{g}(c)) \right]; \tag{4.5}$$

we can rewrite the probability distributions in terms of $\bar{g}$, using $p(c)\,dc = p(\bar{g})\,d\bar{g}$. The optimal solution is obtained by taking a variational derivative with respect to $p(\bar{g})$ and enforcing the normalization through a Lagrange multiplier; the solution is

$$p^*(\bar{g}) = \frac{1}{Z} \cdot \frac{1}{\sigma_g(\bar{g})}. \tag{4.6}$$

By inserting the optimal solution, Eq (4.6), into the expression for mutual information, Eq (4.3), we get the explicit result for the capacity:

$$I_{\rm opt}(c;g) = \log_2 \left[\frac{Z}{\sqrt{2\pi e}} \right], \tag{4.7}$$

where Z is the normalization of the optimal solution in Eq (4.6):

$$Z = \int_0^1 \frac{d\bar{g}}{\sigma_g(\bar{g})}. \tag{4.8}$$

If we were to include the second-order term of Eq (4.4), the approximate solution for the optimal distribution of expression levels would become:

$$p^*(\bar{g}) = \frac{1}{Z} \cdot \frac{1}{\sigma_g(\bar{g})} e^{\frac{1}{2}\sigma_g^2 (\log_2 \sigma_g)''}. \tag{4.9}$$

The optimization with respect to the distribution of inputs, $p(c)$, has led us to the result for the optimal distribution of *mean outputs*, Eq (4.6). We had to assume that the input/output kernel is Gaussian and that the noise is small, and we refer to this result as the small-noise approximation (SNA) for channel capacity. Note that in this approximation only the knowledge of the noise in the output as a function of mean output, $\sigma_g(\bar{g})$, matters for capacity computation and the direct dependence on the input c is irrelevant. Note also that for big enough noise the normalization constant Z will be small compared to $\sqrt{2\pi e}$, and the small-noise capacity approximation of Eq (4.7) will break down by predicting negative information values.

4.2.2 Large noise approximation

Simple regulatory elements usually have a monotonic, saturating input/output relation, as shown in Fig 4.1, and (at least) a shot noise component whose variance scales with the mean. If the noise strength is increased, the information transmission capacity must drop and, even with the optimally tuned input distribution, eventually yield only a bit or less of capacity. Intuitively, the best such noisy system can do is to utilize only the lowest and highest achievable input concentrations, and ignore the continuous range in between. Thus, the mean responses will be as different as possible, and the noise at low expression will also be low because it scales with the mean. More formally, if only $\{c_{\min}, c_{\max}\}$ are used as inputs, then the result is either $p(g|c_{\min})$ or $p(g|c_{\max})$; the optimization of channel capacity reduces to finding $p(c_{\min})$, with $p(c_{\max}) = 1 - p(c_{\min})$ subsequently given by the normalization condition. This problem can be solved either by assuming that each of the two possible input concentrations produces their respective Gaussian output distributions, and maximizing information for $p(c_{\min})$; or simplifying even further and assuming that each of the two possible inputs, "min" and "max", maps into two possible outputs, "on" and "off", and that "min" input might be misunderstood as "on" output and vice versa with probabilities given by the output distribution overlaps, as shown schematically in Fig 4.2.

In the latter case we can use the analytic formula for the capacity of the binary asymmetric channel. If η is the probability of detecting an "off" output if "max" input was sent, and ξ is a probability of receiving an "off" output if "min" input was sent, and $H(\,\cdot\,)$ is a binary entropy function:

$$H(p) = -p \log_2 p - (1-p)\log_2(1-p), \tag{4.10}$$

then the capacity of such asymmetric channel is (Silverman, 1955):

$$I(c;g) = \frac{-\eta H(\xi) + \xi H(\eta)}{\eta - \xi} + \log_2(1 + 2^{\frac{H(\xi)-H(\eta)}{\eta-\xi}}). \tag{4.11}$$

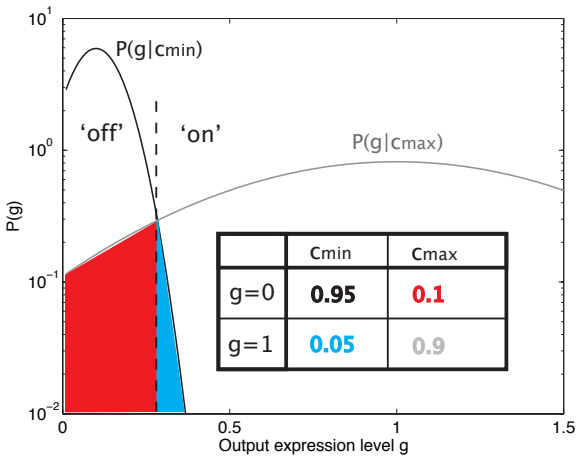


Figure 4.2: An illustration of the large noise approximation. We consider distributions of the output at minimal ($c_{\min}$) and full ($c_{\max}$) induction as trying to convey a single binary decision, and construct the corresponding encoding table (inset). The capacity of such an asymmetric binary channel is degraded from the theoretical maximum of 1 bit, because the distributions overlap (blue and red). For unclipped Gaussians the optimal threshold is at the intersection of two alternative pdfs.

Because this approximation reduces the continuous distribution of outputs to only two choices, "on" or "off", it can underestimate the true channel capacity and is therefore a lower bound.

4.2.3 Exact solution

The information between the input and output in Eq (4.3) can be maximized numerically for any input/output kernel, $P(g|c)$, if the variables c and g are discretized, making the solution space that needs to be searched, $p(c_i)$, finite. One possibility is to use a gradient descent-based method and make sure that the solution procedure always stays within the domain boundaries $\sum_i p(c_i) = 1, p(c_j) \geq 0$ for every j. Alternatively, a procedure known as Blahut-Arimoto algorithm has been derived solely for the purpose of finding optimal channel capacities (Blahut, 1972). Both methods yield consistent solutions, but we prefer to use the second one because of faster convergence and convenient inclusion of constraints on the cost of coding [Methods A.4.1].

One should be careful in interpreting the results of such an optimization and worry about the artifacts introduced by discretization of input and output domains. After discretization, the formal optimal solution is no longer required to be smooth and could, in fact, be composed of a sum of Dirac-delta function spikes. On the other hand, the real, physical concentration c cannot be tuned with arbitrary precision in the cell; it is a result of noisy gene expression, and even if this noise source were removed, the *local* concentration at the binding site is still subject to fluctuations caused by randomness in diffusive flux [Section 3.2]. The Blahut-Arimoto algorithm is completely agnostic as to which (physical) concentrations belong to which bins after concentration has been discretized, and so it could assign a lot of probability weight into bin c_i and zero weight in the neighboring bin c_{i+1} that might represent a concentration change of less than σ_c (i.e. the scale of local concentration fluctuations) from c_i. It is clear that such a solution is physically unrealizable.

One way to address this problem is to include a term in the functional that represents a smoothness constraint on the scale of $\sigma_c(\bar{c})$. The other way is to let the procedure find the spiky solution, but interpret it not as a real, "physical" concentration, but rather as the distribution of concentrations that the cell attempts to generate, c^*. In this case, however, the limited resolution $\sigma_c(\bar{c})$ must be referred to the output as the effective noise in gene expression, $\sigma_g = \sigma_c|\partial g/\partial c|$. The optimal solution $p(c^*)$ is therefore the distribution of the

levels that the cell would use if it had infinitely precise control over choosing various c^*, but the physical concentrations are obtained by convolving this optimal result $p(c^*)$ with a Gaussian of width $\sigma_c(c^*)$. Both of these approaches are presented in Methods A.4.1; as we point out next in the discussion of signals and noise, we use the second method of referring all noise to the output explicitly in the noise model. The third alternative would be to take the finite resolution of the input into account when the input axis is discretized, by matching the sizes of the bins to the input precision.

4.3 A model of signals and noise

If enough data were available, one could directly sample $P(g|c)$ and proceed by calculating the optimal solutions as described previously. Here we start, in contrast, by assuming a Gaussian model [Eq (4.2)] in which the mean, $\bar{g}(c)$, and the output variance, $\sigma_g(c)$, are functions of the transcription factor concentration, c. Our goal for this section is to build an effective microscopic model of transcriptional regulation and gene expression, and therefore define both functions with a small number of biologically interpretable parameters;[1] later we plan to vary those and thus systematically observe the changes in information capacity.

In the simplest picture, the interaction of the TF with the promoter site consists of binding with a (second order) rate constant k_+ and unbinding at a rate k_-. The equilibrium occupancy of the site is [Eq (3.13)]:

$$n = \frac{c^h}{c^h + K_d^h}, \tag{4.12}$$

where the Hill coefficient, h, captures the effects of cooperative binding, and $K_d = k_-/k_+$ is the equilibrium constant of binding. The mean expression level g is then [Eq (A.14)]:

$$g(c) = g_0\bar{g} = g_0 \begin{cases} n & \text{activator} \\ 1-n & \text{repressor} \end{cases}, \tag{4.13}$$

where $\bar{g}$ has been normalized to vary between 0 and 1, and g_0 is the maximum expression level. In what follows we will assume the activator case, where $\bar{g} = n$, and present the result for the repressor in the end.

The fluctuations in occupancy have a (binomial) variance $\sigma_n^2 = n(1-n)$ [Eq (A.33)] and a correlation time $\tau_c = 1/(k_+c^h + k_-)$ [Eq (A.31)]. If the expression level of the target gene is effectively determined by the average of the promoter site occupancy over some window of time $\tau_{\rm int}$, then the contribution to variance in the expression level due to the "on-off" promoter switching will be [Eq (A.46)]:

$$\left(\frac{\sigma_g}{g_0}\right)^2 = \sigma_n^2\frac{\tau_c}{\tau_{\rm int}} = \frac{n(1-n)}{(k_+c^h + k_-)\tau_{\rm int}} = \frac{n(1-n)^2}{k_-\tau_{\rm int}}, \tag{4.14}$$

where in the last step we use the fact that $k_+c^h(1-n) = k_-n$.

At low TF concentrations the arrival times of single transcription factor molecules to the binding site are random events. As we argue in Chapter 3, recent measurements (Gregor et al., 2006a) seem to be consistent with the hypothesis that this variability in diffusive flux

[1] The discussion here briefly recapitulates the results of Section 3.2 and focuses on those features of the noise model that will be needed subsequently.

contributes an additional noise term (Bialek and Setayeshgar, 2005; Tkačik et al., 2006), similar to the Berg-Purcell limit to chemoattractant detection in chemotaxis. The noise in expression level due to fluctuations in the binding site occupancy, or the total *input* noise, is therefore a sum of this diffusive component [Eq (A.46)] and the switching component of Eq (4.14):

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\rm input} = \frac{n(1-n)^2}{k_-\tau_{\rm int}} + \frac{h^2(1-n)^2n^2}{\pi Dac\tau_{\rm int}}, \tag{4.15}$$

where D is the diffusion constant for the TF and a is the receptor site size, $a \sim 3\,$nm for a typical binding site on the DNA.

To compute the information capacity in the small noise limit using the simple model developed so far we need the constant Z from Eq (4.8), which is defined as an integral over expression levels. As both input noise terms are proportional to $(1-\bar{g})^2$, the integral must take the form:

$$Z \propto \int_0^1 \frac{d\bar{g}}{(1-\bar{g})\,F(\bar{g})}, \tag{4.16}$$

where $F(\bar{g})$ is a function that approaches a constant as $\bar{g} \to 1$. Strangely, we see that this integral diverges near full induction ($\bar{g} = 1$), which means that the information capacity also diverges.

Naively we expect that modulations in transcription factor concentration are *not* especially effective at transmitting regulatory information once the relevant binding sites are close to complete occupancy. More quantitatively, the sensitivity of the site occupancy to changes in TF concentration, $\partial n/\partial c$, vanishes as $n \to 1$, and hence small changes in TF concentration will have vanishingly small effects. Our intuition breaks down, however, because in thinking only about the mean occupancy we forget that even very small changes in occupancy could be effective if the noise level is sufficiently small. As we approach complete saturation, the variance in occupancy decreases, and the correlation time of fluctuations becomes shorter and shorter; together these effects cause the standard deviation as seen through an averaging time $\tau_{\rm int}$ to decrease faster than $\partial n/\partial c$, and this mismatch is the origin of the divergence in information capacity. Of course the information capacity of a physical system can't really be infinite; there must be an extra source of noise (or reduced sensitivity) that becomes limiting as $n \to 1$.

The noise in Eq (4.15) captures only the *input* noise, i.e. the noise in the protein level caused by the fluctuations in the occupancy of the binding site. In contrast, the *output* noise arises even when the occupancy of the binding site is fixed (for example, at full induction), and originates in the stochasticity in transcription and translation. The simplest model postulates that when the activator binding site is occupied with fractional occupancy n, mRNA molecules are synthesized in a Poisson process at a rate R_e that generates $R_e\tau_e n$ mRNA molecules on average during the lifetime of a single mRNA molecule, τ_e. Every message is a template for the production of proteins, which is another Poisson process with rate R_g. If the integration time is larger than the lifetime of single mRNA molecules, $\tau_{\rm int} \gg \tau_e$, the mean number of proteins produced is $g = R_g\tau_{\rm int}R_e\tau_e n = g_0 n$, and the variance associated with both Poisson processes is [Eq (A.36)]

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\rm output} = \frac{1 + R_p\tau_e}{g_0}\, n, \tag{4.17}$$

where $b = R_g\tau_e$ is the burst size, or the number of proteins synthesized per mRNA.

Parameter	Value	Description
α	$(1+b)/g_0$	Output noise strength
β	$h^2/\pi DaK_d\tau_{\text{int}}$	Diffusion input noise strength
γ	$(k_-\tau_{\text{int}})^{-1}$	Switching input noise strength
h		Cooperativity (Hill coefficient)

Table 4.1: Gaussian noise model parameters. Note that if burst size $b \gg 1$, then the output noise is determined by the average number of mRNA molecules, $\alpha \sim (\langle \text{mRNA} \rangle)^{-1}$. Note further that if the on-rate is diffusion limited, i.e. $k_+ = 4\pi Da$, then both input noise magnitudes, β and γ, are proportional and decrease with increasing k_-, or alternatively, with increasing K_d.

We can finally put the results together by adding the input noise Eq (4.15) and the output noise Eq (4.17), and expressing both in terms of the normalized expression level $\bar{g}(c)$:

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\text{act}} = \alpha\,\bar{g} + + \beta(1-\bar{g})^{2+\frac{1}{h}}\bar{g}^{2-\frac{1}{h}} + \gamma\bar{g}(1-\bar{g})^2, \tag{4.18}$$

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\text{rep}} = \alpha\,\bar{g} + + \beta(1-\bar{g})^{2-\frac{1}{h}}\bar{g}^{2+\frac{1}{h}} + \gamma\bar{g}^2(1-\bar{g}), \tag{4.19}$$

with the relevant parameters $\{\alpha, \beta, \gamma, h\}$ explained in Table 4.1. Note that both repressor and activator cases differ only in the shape of the input noise contributions (especially for low cooperativity h). Note further that the output noise increases monotonically with mean expression $\bar{g}$, while the input noise peaks at the intermediate levels of expression [Section 3.2.4]. To make the examination of the parameter space in the next section feasible, we set $\gamma = 0$; models with switching noise instead of diffusive noise produce qualitatively similar results [Figs A.17a, A.17b].

4.4 Results

4.4.1 Capacity of simple regulatory elements

Having at our disposal both a simple model of signals and noise and a numerical way of finding the optimal solutions given an arbitrary input-output kernel [Methods A.4.1], we are now ready to examine the channel capacity as a function of the noise parameters from Table 4.1. Our first result [Fig 4.3] concerns the simplest case of an activator with no cooperativity, $h = 1$; for this case, the noise in Eq (4.18) simplifies to:

$$\left(\frac{\sigma_g}{\bar{g}}\right)^2 = \alpha\bar{g} + \beta(1-\bar{g})^3\bar{g}. \tag{4.20}$$

Here we have assumed that there are two relevant sources of noise, i.e. the output noise (which we parametrize by α and plot on the horizontal axis) and the input diffusion noise (parametrized by β, vertical axis). Each point of the information plane in Fig 4.3a therefore represents a system characterized by a Gaussian noise model, Eq (**??**), with variance given above.

As expected, the capacity increases most rapidly when the origin of the information plane is approached approximately along its diagonal, whereas along each of the edges one of the two noise sources effectively disappears, leaving the system dominated by either output or input noise alone. We pick two illustrative examples, the blue and the red systems of Figs 4.3b and 4.3c, that have realistic noise parameters. The blue system has, apart for the decreased cooperativity ($h = 1$ instead of $h = 5$), the characteristics of the *bicoid-hunchback* regulatory element in *Drosophila melanogaster* [Section 3.2]; the red system has the same (dominant output noise) characteristics as those recently measured for about 40 yeast genes by Bar-Even et al. (2006). We would like to emphasize that both the small-noise approximation for capacity, which is easily computable from measured noise at various induction levels, and the exact solution predict that these realistic systems are capable of transmitting more than 1 bit of regulatory information and that they, indeed, could transmit up to about 2 bits.

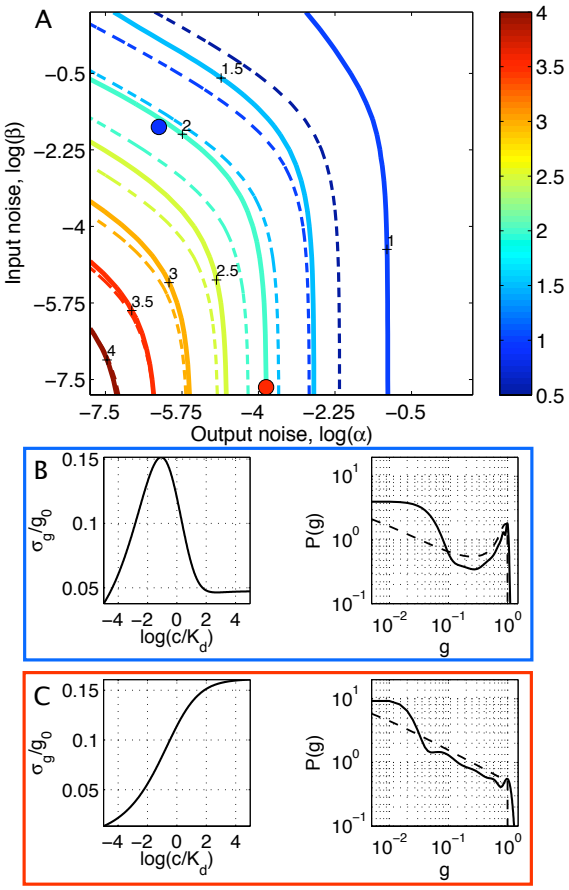


Figure 4.3: Information capacity (color code, in bits) as a function of input and output noise using the activator input-output relation with Gaussian noise given by Eq (4.20) and no cooperativity ($h = 1$). Panel A shows the exact capacity calculation (thick line) and the small noise approximation (dashed line). Panel B displays the details of the blue dot in the information plane: the noise in the output is shown as the function of the input, with a peak being characteristic of the dominant input noise contribution; also shown is the exact solution (thick black line) and the small-noise approximation (dashed black line) to the optimal distribution of output expression levels. Panel C similarly displays details of the system denoted by red dot in information plane; here the output noise is dominant and both approximate and exact solutions for the optimal distribution of outputs show a trend monotonically decreasing with the mean output.

A closer look at the overall agreement between the small-noise approximation (dashed lines in Fig 4.3a) and the exact solution (thick lines) shows that the small-noise approximation underestimates the true capacity, consistent with our remark that for large noise the approximation will incorrectly produce negative results; at the 2-bit information contour the approximation is about $\sim 15\%$ off but improves as the capacity is increased.

Indeed, in the high noise regime we are making yet another approximation, the validity of which we now need to examine. In our discussion about the models of signals and noise we assumed that we can talk about the fractional occupancy of the binding site and the

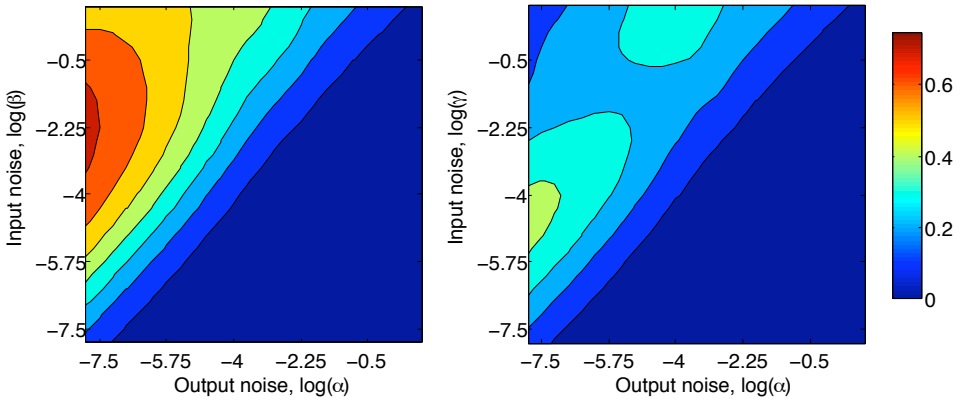


Figure 4.4: Difference in the information transmission capacity between the repressors and activators (color code in bits). Left panel shows $I_{\rm rep}(h=1) - I_{\rm act}(h=1)$, with the noise model that includes output (α) and input diffusion noise (β) contributions (see Fig 4.3 for absolute values of $I_{\rm act}(h=1)$). Right panel shows $I_{\rm rep} - I_{\rm act}$ for the noise model that includes output noise (α) and input switching noise (γ) contributions. This latter difference does not depend on cooperativity (see Fig A.17b for the corresponding absolute information values).

continuous concentrations of mRNA, transcription factors and protein, instead of counting these species in discrete units, and that noise can effectively be treated as Gaussian. Both of these assumptions are the cornerstones of the Langevin approximation for calculating the noise variance [Methods A.4.5]. If parameters α and β actually arise due to the underlying microscopic mechanisms described in the section on signals and noise and schematized in Fig 3.2, we expect that at least for some large-noise regions of the information plane the discreteness in the number of mRNA molecules will become important and the Langevin approximation will fail. In such cases (a much more time-consuming) exact calculation of the input-output relations using the Master equation is possible for some noise models; in Fig A.18 we show that in the region where $\log \alpha > -2$ the channel capacities calculated with Gaussian kernels can be overestimated by $\sim 10\%$ or more; there the Langevin calculation gives the correct second moment, but misses the true shape of the distribution. We can nevertheless conclude that Langevin approximation provides a good analytic framework for the analysis of information capacity in the biologically relevant region of parameter space.

Is there any difference between activators and repressors in their capacity to convey information about the input? We concluded Section 4.3 on the noise models with separate expressions for activator noise, Eq (4.18), and repressor noise, Eq (4.19); focusing now on the repressor case, we recompute the information plane in the same manner as we did for the activator in Fig 4.3a, and display the difference between the capacities of the repressor and activator with the same noise parameters in Fig 4.4. As expected, the biggest difference occurs above the main diagonal, where the input noise dominates over the output noise. In this region the capacity of the repressor can be bigger by as much as third than that of the corresponding activator. Note that as $h \to \infty$, the activator and repressor noise expressions become indistinguishable and the difference in capacity vanishes for the noise models with output and input diffusion noise contributions [Eqs (4.18, 4.19)]. The difference between the two modes of regulation is still present but less striking in an alternative model with output and input switching noise contributions.

The behavior of the regulatory element can be conveniently visualized in Fig 4.5 by plotting a cut through the information plane along its main diagonal. Moving along this

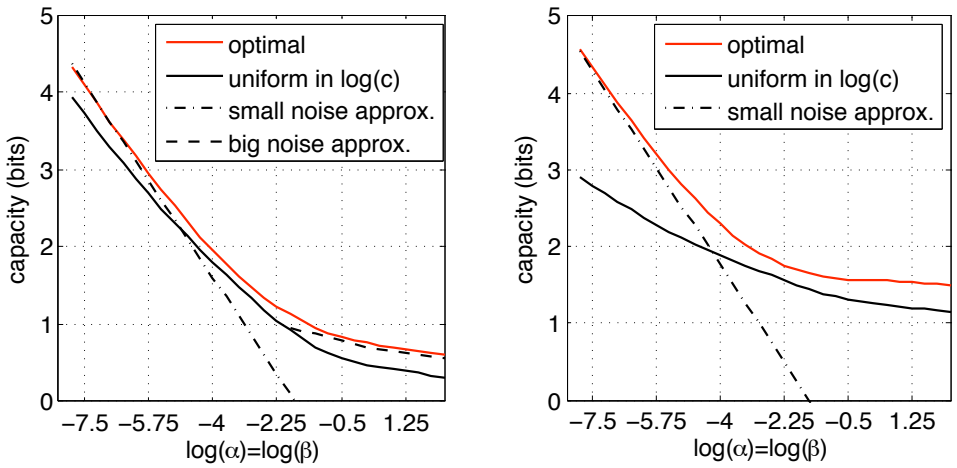


Figure 4.5: Comparison of exact channel capacities and various approximate solutions. For both panels (no cooperativity, $h = 1$, on the left; strong cooperativity, $h = 3$, on the right) we take a cross-section through the information plane in Fig 4.3 along the main diagonal, where the values for noise strength parameters α and β are equal. The exact solution is shown in red. By moving along the diagonal of the information plane one changes both input and output noise by the same multiplicative factor s, and since, in small-noise approximation, $I_{\mathrm{SNA}} \propto \log Z \propto \int \sigma_g(\bar{g})^{-1}\, dg$, that factor results in an additive change in capacity by $\log_2 s$. We can use the large noise approximation lower bound on capacity for the case $h = 1$, in the parameter region where capacities fall below 1 bit.

cut corresponds to scaling the total noise of the system up or down by a multiplicative factor, and allows us to observe the overall agreement between the exact solution and small- and large-noise approximations. In addition we point out the following interesting features of Fig 4.5 that will be examined more closely in subsequent sections.

Firstly, there is only a small regime where the capacity is below one bit and the large noise approximation can be applied. With higher cooperativity this regime disappears, suggesting that a biological implementation of a reliable binary channel could be relatively straightforward if our noise model is appropriate. In addition, there are distributions not specifically optimized for the input/output kernel that nevertheless achieve considerable capacities: we illustrate this idea by using input distributions that are uniform in $\log(c/K_d)$ in Fig 4.5 (thick black line) and interpret it as an indication that the maximum in capacity cannot be very sharp with respect to small perturbations of the optimal solution, $p(c^*)$. We revisit this idea more systematically in the next section.

Secondly, it can be seen from Fig 4.5 that at small noise the cooperativity has a minor effect on the channel capacity, which is unexpected at first because the functional form of the noise explicitly depends on cooperativity, Eq (4.18), and additionally, the shape of the mean response $\bar{g}(c)$ strongly depends on h. We recall, however, that mutual information $I(c;g)$ is invariant to any invertible reparametrization of either g or c. In particular, changing the cooperativity or the value of the equilibrium binding constant, K_d, in theory only results in an invertible change in the input variable c, and therefore the change in the steepness or midpoint of the mean response must not have any effect on $I(c;g)$. This argument breaks down in the high noise regime, because reparametrization invariance would work only if the input concentration could extend over the whole positive interval, from zero to infinity. The substantial difference between capacities of cooperative and non-cooperative systems

in Fig 4.5 at *low capacity* stems from the fact that in reality the cell (and our computation) is limited to a finite range of concentrations, $c \in [c_{\min}, c_{\max}]$, instead of the whole positive half-axis, $c \in [0, \infty)$. We explore the issue of limited input dynamic range further in the next section.

Finally, we draw attention to the simple linear scaling of the channel capacity with the logarithm of the total noise strength, as explained in the caption of Fig 4.5, when small noise approximation is valid. In general, increasing the number of input and output molecules by a factor of four will decrease the relative input and output noise by a factor of $\sqrt{4} = 2$, and therefore, in the small noise approximation, increase the capacity by $\log_2 2 = 1$ bit. If there is no cost that needs to be paid by the cell to make more transcription factor and output protein molecules, then scaling the noise along the horizontal axis of Fig 4.5 is directly related to the scaling of the total number of signaling molecules used by the regulatory element. If there are metabolic or time costs to making more molecules, our optimization needs to be modified appropriately, and we present the relevant computation in the section on the costs of coding.

4.4.2 Cooperativity, dynamic range and the tuning of solutions

So far we have assumed that the computed optimal input distributions are biologically realizable. For instance, the range of the allowed input concentrations was not constrained explicitly to a narrow band of several-fold change around K_d; neither was any special attention paid to the absolute value of K_d, because we argued that it would only modify the units on the concentration axis; nor did we examine what are the possible consequences of the requirements on the smoothness of optimal distributions, or discuss their "fine-tuning." Relying on simplifications of this sort or ignoring constraints to which a biological system must inevitably be subjected amounts to making assumptions that do not necessarily hold in nature, and the goal of this section is to study how the information transmission is affected if such idealizations are relaxed.

We start by considering the impact on channel capacity of changing the allowed total dynamic range to which the input concentration is restricted. Figure 4.6 displays, in the left plot, the capacity as a function of the dynamic range (where we talk about a "10-fold range" if $c \in \left[\frac{1}{5}K_d, 5K_d\right]$), output noise and cooperativity. The main feature of the plot is the difference between low and high cooperativity cases at each noise level; regardless of cooperativity the total information at infinite dynamic range would saturate at approximately the same level (which depends on the output noise magnitude). However, highly cooperative systems manage to reach a high fraction (80% or more) of their saturated information transmission levels even at reasonable dynamic ranges of about 10 to 20-fold, whereas low cooperativity systems require a much bigger dynamic range for the same capacity. The decrease in capacity with decreasing dynamic range is a direct consequence of the nonlinear relationship between the concentration and occupancy, Eq (4.12), and for low cooperativity systems in particular means not being able to fully shut down or fully induce the promoter. This contrasts sharply with the theoretical noise model with infinite concentration range in which, at zero input concentration, the noise is always zero, $\sigma_g(c = 0) = 0$, and therefore this "zero input" constitutes a "letter" of the input alphabet that can be perfectly discriminated from all other inputs. Although we do not discuss it explicitly in our models of signals and noise, leaky expression (i.e. non-zero output when the concentration of activators is zero) will have a similar effect of degrading capacity to that of reducing the available input dynamic range.

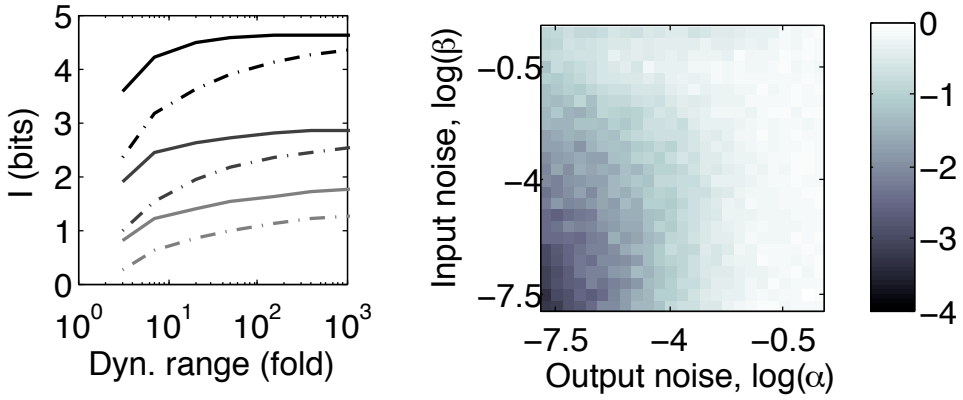


Figure 4.6: Effects of imposing realistic constraints on the space of allowed input distributions. Left panel shows the change in capacity if the dynamic range of the input around K_d is changed. The regulatory element is a repressor with either no cooperativity (dash-dot line) or high cooperativity, $h = 3$ (thick line), and with output noise α only. We plot three high-low cooperativity pairs for three different choices of the output noise magnitude (high noise in light gray, $\log \alpha \approx -2.5$; medium noise in dark gray, $\log \alpha \approx -5$; low noise in black, $\log \alpha \approx -7.5$). Right panel shows the sensitivity of channel capacity to perturbations in the optimal input distribution (grayscale indicates the number of bits of capacity lost per unit Jensen-Shannon divergence between the optimal and suboptimal input distribution – see main text and Fig A.19).

We conclude this section by discussing how precisely tuned the resulting optimal distributions have to be to take full advantage of the regulatory element's capacity. For each point in the information plane of Fig 4.3a the optimal input distribution $p^*(c)$ is perturbed many times to create an ensemble of suboptimal inputs $p_i(c)$ [Methods A.4.6]. For each $p_i(c)$, we compute, first, its distance away from the optimal solution by means of Jensen-Shannon divergence, $d_i = D_{\rm JS}(p_i, p^*)$; next, we use the $p_i(c)$ to compute the suboptimal channel capacity I_i. A scatter plot of many such pairs (d_i, I_i) obtained with various perturbations $p_i(c)$ for each system of the information plane characterizes the sensitivity of the optimal solution for that system; the main feature of such a plot [Fig A.19] is the linear (negative) slope that describes how many bits in channel capacity are lost for each unit of Jensen-Shannon distance away from the optimal solution. These slopes are shown in grayscale in the right plot of Fig 4.6 for the whole information plane. We note that for systems with high capacity the linear relationship between the the divergence d_i and capacity I_i provides a better fit than for systems with small capacity. Most importantly, the figure not only shows that high capacity solutions are more sensitive to deviations from the optimal solution, but also that achieving 1 bit of capacity does not require much tuning – if we take the linear slopes seriously and try to extrapolate as $D_{\rm JS} \to 1$ to ask how much channel capacity remains if one were to use very "un-tuned" solutions, we see that this value is about a bit for most of the information plane.

4.4.3 Non-specific binding and the costs of higher capacity

Real regulatory elements must balance the pressure to convey information reliably with the cost of maintaining the cell's internal state, represented by the expression levels of transcription factors. The fidelity of the representation is increased (and the fractional fluctuation in their number is decreased) by having more molecules "encode" a given state. On the other hand, making or degrading more transcription factors puts a metabolic burden

on the cell, and frequent transitions between various regulatory states could involve too large time lags as, for example, the regulation machinery attempts to keep up with a changed environmental condition, by accumulating or degrading the corresponding TF molecules. In addition, the output genes themselves that get switched on or off by transcription factors and therefore "read out" the internal state must not be too noisy, otherwise the advantage of maintaining precise transcription factor levels is lost.

Suppose that there is a cost to the cell for each molecule of output gene that it needs to produce, and that this incremental cost per molecule is independent of the number of molecules already present. Then, on the output side, the cost must be proportional to $\langle g \rangle = \int dg\, g\, p(g)$. We remember that in optimal distribution calculations g is expressed as relative to the maximal expression, such that its mean is between zero and one. To get an absolute cost in terms of the number of molecules, this normalized $\bar{g}$ therefore needs to be multiplied by the inverse of the output noise strength, α^{-1}, as the latter scales with g_0 [Table 4.1]. The contribution of the output cost is thus $\propto \alpha^{-1}\bar{g}$.

On the input side, the situation is similar: the cost must be proportional to $K_d\langle\tilde{c}\rangle = K_d \int d\tilde{c}\,\tilde{c}\,p(\tilde{c})$, where our optimal solutions are expressed, as usual, in dimensionless concentration units, $\tilde{c} = c/K_d$. In either of the two input noise models (i.e. diffusion or switching input noise), with diffusion constant held fixed, $K_d \propto \beta^{-1}$ or $K_d \propto \gamma^{-1}$.

Before continuing we need to make a careful distinction between the total concentration of the input transcription factors, c_t, and the free concentration c_f, diffusing in solution in the nucleus. We imagine the true binding site embedded in a pool of non-specific binding sites – perhaps all other short fragments of DNA – and there being an ongoing competition between one functional site (with strong affinity) and large number of weaker non-specific sites. If these non-specific sites are present at concentration ρ in the cell, and have affinities drawn from some distribution $p(K)$, the relationship between the free and the total concentration of the input is [Eq (3.23)]:

$$c_t = c_f + \rho \int dK\, p(K)\, \frac{c_f}{c_f + K}. \tag{4.21}$$

Importantly, the concentration that enters all information capacity calculations is the *free* concentration c_f, because it directly determines both the promoter occupancy in Eq (4.12) as well as diffusive noise; on the other hand, the cell can influence the free concentration only by producing more or less of the transcrption factor, i.e. by varying (and paying for) the *total* concentration. Since our costs are determined only up to a proportionality constant, the important question is whether or not the relation between c_t and c_f in Eq (4.21) is close to linear. In Section 3.3.1 we have shown that a linear relation between total and free concentrations is a reasonable approximation, and that no new appreciable noise is generated by the presence of nonspecific binding sites. The bottom line is that the effect of the nonspecific sites is restricted to changing the proportionality factor between the average free concentration of the input molecules and their metabolic cost, but nothing else.

Collecting all our thoughts on the costs of coding, we can write down the "cost functional" as the sum of input and output cost contributions:

$$\langle \mathcal{C}[p(c)] \rangle = \frac{v_1}{\beta} \int dc\, p(c)\, c + \frac{v_2}{\alpha} \int dc\, p(c) \int dg\, p(g|c)\, g, \tag{4.22}$$

where v_1 and v_2 are proportional to the unknown costs per molecule of input or output, respectively, and α and β are noise parameters of Table 4.1. This ansatz captures the

intuition that while decreasing noise strengths will increase information transmission, it will also increase the cost. Instead of maximizing the information without regard to the cost, the new problem to extremize is:

$$\mathcal{L}[p(c)] \quad = \quad I[p(c)] - \Phi\langle \mathcal{C}[p(c)]\rangle - \Lambda \int dc\, p(c), \tag{4.23}$$

and the Lagrange multiplier Φ has to be chosen so that the cost of the resulting optimal solution $\langle \mathcal{C}[p^*(c)]\rangle$ equals some predefined cost C_0 that the cell is prepared to pay.

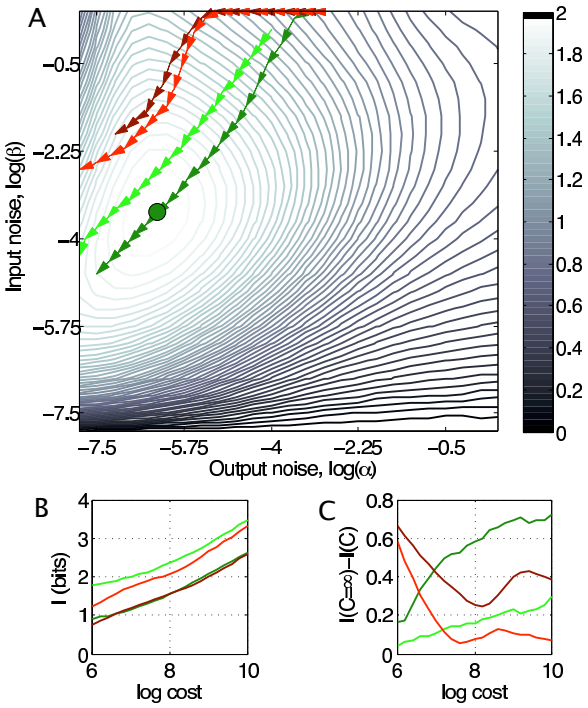


Figure 4.7: The effects of metabolic or time costs on the achievable capacity of simple regulatory elements. Contours in panel A show the information plane for non-cooperative activator from Fig 4.3, with the imposed constraint that the average total (input + output) cost is fixed to some C_0; as the cost is increased, the optimal solution (green dot) moves along the arrows on a dark green line (the contours change correspondingly, not shown). Light green line shows activator with cooperativity $h = 3$, dark and light red lines show repressors without and with cooperativity ($h = 3$). For each flow line in the information plane, panel B shows the cost vs capacity plot, while panel C shows the difference between the constrained and unconstrained capacities along the trajectories.

We now wish to recreate the information plane of Fig 4.3, while constraining the total cost of each solution to C_0. To be concrete and pick the value for the cost and proportionality constants in Eq (4.22), we use the estimates from *Drosophila* noise measurements in Gregor et al. (2006a) and the analysis of Section 3.2, which assign to the system denoted by a blue dot in Fig 4.3a, the values of ~ 800 Bicoid molecules of input at K_d, and a maximal induction of $g_0 \sim 4000$ Hunchback molecules if the burst size b were 10. Figure 4.7a is the information plane for an activator with no cooperativity, as in Fig 4.3, but with the cost limited to an average total of $C_0 \sim 7000$ molecules of input and output per nucleus. There is now one optimal solution denoted by a green dot; if one tries to choose a system with lower input or output noise, the cost constraint forces the input distribution, $p(c)$, and the output distribution, $p(g)$, to have very low probabilities at high induction, consequently limiting the capacity.

Clearly, a different system will be optimal if another total allowed cost C_0 is selected. The dark green line on the information plane in Fig 4.7a corresponds to the flow of the optimal solution for an activator with no cooperativity if the allowed cost is increased, and the corresponding cost-capacity curve is shown in Fig 4.7b. The light green line is the trajectory of the optimal solution in the information plane of the activator system with

cooperativity $h = 3$, and the dark and light red trajectories are shown for the repressor with $h = 1$ and $h = 3$, respectively. We note first that the behavior of the cost function is quite different for the activator (where low input implies low output and therefore low cost; and conversely high input means high output and also high cost) and the repressor (where input and output are mutually exclusively high or low and the cost is intermediate in both cases). Secondly, we observe that the optimal capacity as a function of cost is similar for the activators and repressors [Fig 4.7b] in sharp contrast to the comparison of Fig 4.4, where repressors would enable higher capacities. Thirdly, we note in the same figure that increasing the cooperativity at fixed noise strength β brings a substantial increase, of almost a bit over the whole cost range, in the channel capacity, in agreement with our observations about the interaction between capacity and the dynamic range [Fig 4.6]. The last and perhaps the most significant conclusion is that even with input distributions matched to maximize the transmission at a fixed cost, the capacity still only scales roughly linearly with the logarithm of the number of available signaling molecules, and this fact must ultimately be limiting in a single regulatory element.

4.5 Discussion

We have tried to analyze a simple regulatory element as an information processing device. As a result we find that one cannot discuss an element in isolation from the statistics of the input that it is exposed to. Yet in cells the inputs are often transcription factor concentrations that "encode" the state of various genetic switches, from those responsible for cellular identity to those that control the rates of metabolism and cell division, and the cell exerts control over these concentrations. While it *could* use different distributions to represent various regulatory settings, we argue that the cell *should* use the one distribution that allows it to make the most of its genetic circuitry – the distribution that maximizes the dependency, or mutual information, between inputs and outputs. Mutual information can then be seen both a measure of how well the cell is doing by using *its* encoding scheme, and the best it could have done using the *optimal* scheme, which we can compute; comparison between the optimal and measured distributions gives us a sense of how close the organism is to the achievable bound. Moreover, mutual information has absolute units, i.e. bits, that have a clear interpretation in terms the number of discrete distinguishable states that the regulatory element can resolve [Methods A.4.3, Fig A.16b]. This last fact helps clarify the ongoing debates about what is the proper noise measure for genetic circuits, and in what context a certain noise is either "big" or "small" (as it is really a function of the inputs). Information does not replace the standard *noise-over-the-mean* measure – noise calculations or measurements are still necessary to compute the element's capacity – but does give it a functional interpretation.

We have considered a class of simple parametrizations of signals and noise that can be used to fit measurements for several model systems, such as *bicoid-hunchback* in the fruit fly, *lac* in *Escherichia coli* and a number of yeast genes. We find that the capacities of these realistic elements are generally larger than 1 bit, and can be as high as 2 bits. By simple inspection of optimal output distributions in Figs 4.3b or 4.3c it is difficult to tell anything about the capacity: the distribution might look bimodal yet carry more than one bit, or might even be a monotonic function without any obvious structure, indicating that the information is encoded in the graded response of the element. When the noise is sufficiently high, on the other hand, the optimal strategy is that of achieving one bit of capacity and

only utilizing maximum and minimum achievable levels of transcription factors for signaling. The set of distributions that achieve capacities close to the optimal one is large, suggesting that perhaps one-bit switches are not difficult to implement biologically.

Finally, we discussed how additional biophysical constraints can modify the optimal capacity. By assuming a linear cost model for signaling molecules and a limited input dynamic range, the capacity and cost couple in an interesting way and the maximization principle allows new questions to be asked. For example, increasing the cooperativity reduces the cost, as we have shown; on the other hand, it increases the sensitivity to fluctuations in the input, because the input noise strength β is proportional to h^2 [Table 4.1]. In a given system we could therefore predict the optimal effective cooperativity, if we knew the real "cost per molecule" [Methods A.4.2]. Further work is needed to tease out the consequences of cost (if any) from experimental data.

The principle of information maximization clearly is not the only possible lens through which regulatory networks are to be viewed. One can think of examples where only a single bit needs to be conveyed, but it has to be done reliably in a fluctuating environment, perhaps by being *robust* to the changes in outside temperature. It seems both concepts, that of maximal information transmission and the robustness to fluctuations in certain auxiliary variables that also influence the noise, could be included into the same framework, but the issue needs further work. Alternatively, consider signaling systems where there are constraints on the *dynamics*, something that our analysis has ignored by only looking at steady state behavior; for example, the chemotactic system of *Escherichia coli* has to perfectly adapt in order for the bacterium to be able to climb the attractant gradients. All these examples, however, assume some knowledge about the behavior over and above the ability to define the element, its inputs and its outputs: they imply what is essential for the proper functioning of that specific module, and this knowledge can be viewed (and perhaps later formally included) as introducing additional constraint to which the basic information transmission is subjected.

We emphasize that the kind of analysis carried out here is not restricted to a single regulatory element. As was pointed out in the introductory Section 4.1, the inputs $\mathcal{I}$ and the outputs $\mathcal{O}$ of the regulatory module can be multi-dimensional, and the module could implement complex internal logic with multiple feedback loops. It seems that especially in such cases, when our intuition about the noise – now a function of multiple variables – starts breaking down, the information formalism could prove to be helpful. Although the solution space that needs to be searched in the optimization problem grows exponentially in the inputs, there are biologically relevant situations that nevertheless appear tractable: for example, when there are multiple readouts of the same input, or combinatorial regulation of a single output by a pair of inputs; in addition, knowing that the capacities of a single input-output chain are on the order of a few bits also means that only a small number of distinct input levels for each input need to be considered. Some cases therefore appear amenable to biophysical modeling approaches, and in the next section we will apply the theory presented here to the regulation of Hunchback expression by the Bicoid TF in *Drosophila melanogaster*.

4.6 Information flow and optimization in transcriptional regulation – Morphogenesis in the early *Drosophila* embryo[2]

Cells control the expression of genes in part through transcription factors, proteins which bind to particular sites along the genome and thereby enhance or inhibit the transcription of nearby genes [Fig 4.8]. We can think of this transcriptional control process as an input/output device in which the input is the concentration of transcription factor and the output is the concentration of the gene product. Although this qualitative picture has been with us for roughly forty years (Jacob and Monod, 1961), only recently have there been quantitative measurements of *in vivo* input/output relations and of the noise in output level when the input is fixed (Elowitz et al., 2002; Ozbudak et al., 2002; Blake et al., 2003; Setty et al., 2003; Raser and O'Shea, 2004; Rosenfeld et al., 2005; Pedraza and van Oudenaarden, 2005; Golding et al., 2005; Kuhlman et al., 2007; Gregor et al., 2006a). Because these input/output relations have a limited dynamic range, noise limits the "power" of the cell to control gene expression levels. In this section, we quantify these limits and derive the strategies that cells should use to take maximum advantage of the available power. We show that, to make optimal use of its regulatory capacity, cells must achieve the proper quantitative matching among the input/output relation, the noise level, and the distribution of transcription factor concentrations used during the life of the cell. We test these predictions against recent experiments on the Bicoid and Hunchback morphogens in the early *Drosophila* embryo (Gregor et al., 2006a), and find that the observed distributions have a nontrivial structure which is in good agreement with theory, with no adjustable parameters. This suggests that, in this system at least, cells make nearly optimal use of the available regulatory capacity and transmit substantially more than the simple on/off bit that might suffice to delineate a spatial expression boundary.

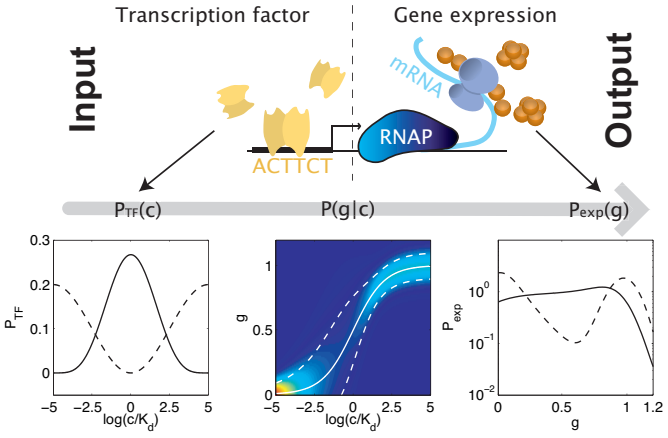


Figure 4.8: Transcriptional regulation of gene expression. The occupancy of the binding site by transcription factors sets the activity of the promoter and hence the amount of protein produced. The physics of TF–DNA interaction, transcription and translation processes determine the conditional distribution of expression levels g at fixed TF concentration c, $P(g|c)$, shown here as a heat map with red (blue) corresponding to high (low) probability. The mean input/output relation is shown as a thick white line, and the dashed lines indicate ± one standard deviation of the noise around this mean. Two sample input distributions $P_{TF}(c)$ (lower left) are passed through $P(g|c)$ to yield two corresponding distributions of outputs, $P_{\rm exp}(g)$ (lower right).

[2]This section appeared on the arXiv as Tkačik et al. (2007b).

Gene expression levels (g) change in response to changes in transcription factor (TF) concentration (c). These changes often are summarized by an input/output relation $\bar{g}(c)$ in which the mean expression level is plotted as a function of TF concentration [Fig 4.8]. The average relationship is a smooth function but, because of noise, this does not mean that arbitrarily small changes in input transcription factor concentration are meaningful for the cell. The noise in expression levels could even be so large that reliable distinctions can only be made between (for example) "gene on" at high TF concentration and "gene off" at low TF concentration. To explore this issue, we need to quantify the number of reliably distinguishable regulatory settings of the transcription apparatus, a task to which Shannon's mutual information (Shannon, 1948; Cover and Thomas, 1991) is ideally suited. While there are many ways to associate a scalar measure of correlation or control with a joint distribution of input and output signals, Shannon proved that mutual information is the only such quantity that satisfies certain plausible general requirements, independent of the details of the underlying distributions. Mutual information has been successfully used to analyze noise and coding in neural systems (Rieke et al., 1997), and it is natural to think that it may be useful for organizing our understanding of gene regulation; see also Ziv et al. (2006).

Roughly speaking, the mutual information $I(c;g)$ between TF concentration and expression level counts the (logarithm of the) number of distinguishable expression levels achieved by varying c. If we measure the information in bits, then

$$I(c;g) = \int dc\, P_{TF}(c) \int dg\, P(g|c) \log_2 \left[\frac{P(g|c)}{P_{\rm exp}(g)} \right], \tag{4.24}$$

where $P_{TF}(c)$ is the distribution of TF concentrations the cell generates in the course of its life, $P(g|c)$ is the distribution of expression levels at fixed c, and $P_{\rm exp}(g)$ is the resulting distribution of expression levels,

$$P_{\rm exp}(g) = \int dc\, P(g|c) P_{TF}(c). \tag{4.25}$$

The distribution, $P(g|c)$, of expression levels at fixed transcription factor concentration describes the physics of the regulatory element itself, from the protein/DNA interaction, to the rates of protein synthesis and degradation; this distribution describes both the mean input/output relation *and* the noise fluctuations around the mean output. The information transmission, or regulatory power, of the system is not determined by $P(g|c)$ alone, however, but also depends on the distribution, $P_{TF}(c)$, of transcription factor "inputs" that the cell uses, as can be seen from Eq (4.24). By adjusting this distribution to match the properties of the regulatory element, the cell can maximize its regulatory power.

Matching the distribution of inputs to the (stochastic) input/output relation of the system is a central concept in information theory (Cover and Thomas, 1991), and has been applied to the problems of coding in the nervous system. For sensory systems, the distribution of inputs is determined by the natural environment, and the neural circuitry can adapt, learn or evolve (on different times scales) to adjust its input/output relation. It has been suggested that maximizing information transmission is a principle which can predict the form of this adaptation (Barlow, 1961; Laughlin, 1981; Atick and Redlich, 1990; Brenner et al., 2000). In transcriptional regulation, by contrast, it seems more appropriate to regard the input/output relation as fixed and ask how the cell might optimize its regulatory power by adjusting the distribution of TF inputs.

It is difficult to make analytic progress in the general calculation of mutual information, but there is a simple and plausible approximation. The expression level at a fixed TF concentration c has a mean value $\bar{g}(c)$, which we can plot as an input/output relation [Fig 4.8]. Let us assume that the fluctuations around this mean are Gaussian with a variance $\sigma_g^2(c)$ which will itself depend on the TF concentration. Formally this means that

$$P(g|c) = \frac{1}{\sqrt{2\pi\sigma_g^2(c)}} \exp\left\{ -\frac{[g-\bar{g}(c)]^2}{2\sigma_g^2(c)} \right\}. \tag{4.26}$$

Further let us assume that the noise level is small. Then we can expand all of the relevant integrals as a power series in the magnitude of σ_g [Eq (4.4)]:

$$\begin{aligned} I(c;g) &= -\int d\bar{g}\, \hat{P}_{\rm exp}(\bar{g}) \log_2 \hat{P}_{\rm exp}(\bar{g}) \\ &\quad -\frac{1}{2}\int d\bar{g}\, \hat{P}_{\rm exp}(\bar{g}) \log_2[2\pi e \sigma_g^2(\bar{g})] + \cdots , \end{aligned} \tag{4.27}$$

where $\cdots$ are terms that vanish as the noise level decreases and $\hat{P}_{\rm exp}(\bar{g})$ is the probability distribution for the average levels of expression. We can think of this as the distribution that the cell is "trying" to generate, and would generate in the absence of noise:

$$\hat{P}_{\rm exp}(\bar{g}) \equiv \int dc\, P_{TF}(c)\delta[\bar{g}-\bar{g}(c)] \tag{4.28}$$

$$= P_{TF}(c=c_*(\bar{g})) \left| \frac{d\bar{g}}{dc} \right|^{-1}_{c=c_*(\bar{g})}, \tag{4.29}$$

where $c_*(\bar{g})$ is the TF concentration at which the mean expression level is $\bar{g}$; similarly, by $\sigma_g(\bar{g})$ we mean $\sigma_g(c)$ evaluated at $c = c_*(\bar{g})$.

We now can ask how the cell should adjust these distributions to maximize the information being transmitted. In the low-noise approximation summarized by Eq (4.27), maximizing $I(c;g)$ poses a variational problem for $\hat{P}_{\rm exp}(\bar{g})$ whose solution has a simple form:

$$\hat{P}^*_{\rm exp}(\bar{g}) = \frac{1}{Z}\cdot\frac{1}{\sigma_g(\bar{g})} \tag{4.30}$$

$$Z = \int d\bar{g}\, \frac{1}{\sigma_g(\bar{g})}. \tag{4.31}$$

This result captures the intuition that effective regulation requires preferential use of signals that have high reliability or low variance—$\hat{P}^*_{\rm exp}(\bar{g})$ is large where σ_g is small. The actual information transmitted for this optimal distribution can be found by substituting $\hat{P}^*_{\rm exp}(\bar{g})$ into Eq (4.27), with the result $I_{\rm opt}(c;g) = \log_2\left(Z/\sqrt{2\pi e}\right)$.

Although we initially formulated our problem as one of optimizing the distribution of *inputs*, the low noise approximation yields a result [Eq (4.30)] which connects the optimal distribution of *output* expression levels to the *variances* of the same quantities, sampled across the life of a cell as it responds to natural variations in its environment. To the extent that the small noise approximation is applicable, data on the variance vs mean expression thus suffice to calculate the maximum information capacity; details of the input/output

relation, such as its degree of cooperativity, do not matter except insofar as they leave their signature on the noise.

Recent experiments provide the data for an application of these ideas. Elowitz and coworkers have measured gene expression noise in a synthetic system, placing fluorescent proteins under the control of a lac–repressible promoter in *E. coli* (Elowitz et al., 2002). Varying the concentration of an inducer, they determined the intrinsic variance of expression levels across a bacterial population as a function of mean expression level. Their results can be summarized as $\sigma_g^2(\bar{g}) = a\bar{g} + b\bar{g}^2$, where the expression level g is normalized to have a maximum mean value of 1, and the constants are $a = 5 - 7 \times 10^{-4}$ and $b = 3 - 10 \times 10^{-3}$. Across most of the dynamic range ($\bar{g} \gg 0.03$), the small noise approximation should be valid and, as discussed above, knowledge of $\sigma_g(\bar{g})$ alone suffices to compute the optimal information transmission. We find $I_{\rm opt}(c;g) \sim 3.5$ bits: rather than being limited to on/off switching, these transcriptional control systems could in principle specify $2^{I_{\rm opt}} \sim 10 - 12$ distinguishable levels of gene expression (see Section A.4.3)! It is not clear whether this capacity, measured in an engineered system, is available to or used by *E. coli* in its natural environment. The calculation does demonstrate, however, that optimal information transmission values derived from real data are more than one bit, but perhaps small enough to provide significant constraints on regulatory function.

When the noise is not small, no simple analytic approaches are available. On the other hand, so long as $P(g|c)$ is known explicitly, our problem is equivalent to one well–studied in communication theory, and efficient numerical algorithms are available for finding the input distribution $P_{TF}(c)$ that optimizes the information $I(c;g)$ defined in Eq (4.24) [Methods A.4.1] (Blahut, 1972). In general we must extract $P(g|c)$ from experiment and, to deal with finite data, we will assume that it has the Gaussian form of Eq (4.26). $P(g|c)$ then is completely determined by measuring just two functions of c: the mean input/output relation $\bar{g}(c)$ and the output variance $\sigma_g^2(c)$. The central point is that, in the general case, solving the information optimization problem requires *only* empirical data on the input/output relation and noise.

The initial events of pattern formation in the embryo of the fruit fly *Drosophila* provide a promising testing ground for the optimization principle proposed here. These events depend on the establishment of spatial gradients in the concentration of various morphogen molecules, most of which are transcription factors (Wolpert, 1969; Lawrence, 1992). To be specific, consider the response of the *hunchback* (Hb) gene to the maternally established gradient of the transcription factor Bicoid (Bcd) (Driever and Nusslein-Volhard, 1988b,a, 1989; Struhl et al., 1989). A recent experiment reports the Bcd and Hb concentrations in thousands of individual nuclei of the *Drosophila* embryo, using fluorescent antibody staining (Gregor et al., 2006a); the results can be summarized by the mean input/output relation and noise level shown in Fig 4.9. These data can be understood in some detail on the basis of a simple physical model as in Section 3.2 (Tkačik et al., 2007a), but here we use the experimental observations directly to make phenomenological predictions about maximum available regulatory power and optimal distribution of expression levels.

Given the measurements of the mean input/output relation $\bar{g}(c)$ and noise $\sigma_g(c)$ shown in Fig 4.9, we can calculate the maximum mutual information between Bcd and Hb concentrations by following the steps outlined above; we find $I_{\rm opt}(c;g) = 1.7$ bits. To place this result in context, we imagine a system that has the same mean input/output relation, but the noise variance is scaled by a factor F, and ask how the optimal information transmission depends on F. This is not just a mathematical trick: for most physical sources of noise, the relative variance is inversely proportional to the number of molecules, and so scaling

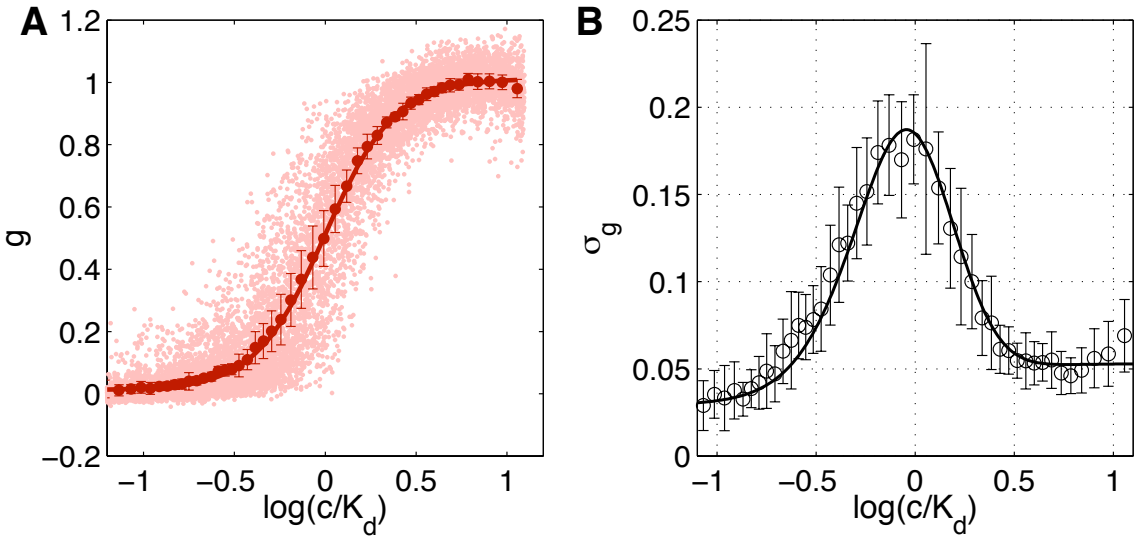


Figure 4.9: The Bcd/Hb input/output relationship in the *Drosophila melanogaster* syncitium at early nuclear cycle 14 (Gregor et al., 2006a). (a) Each point marks the Hb (g) and Bcd (c) concentration in a single nucleus, as inferred from immunofluorescent staining; data are from $\sim 11 \cdot 10^3$ individual nuclei across 9 embryos. Hb expression levels g are normalized so that the maximum and minimum mean expression levels are 1 and 0 respectively; small errors in the estimate of background fluorescence result in some apparent expression values being slightly negative. Bcd concentrations c are normalized by K_d, the concentration of Bcd at which the mean Hb expression level is half maximal. For details of normalization across embryos, see Gregor et al. (2006a). Solid red line is a sigmoidal fit to the mean g at each value of c, and error bars are $\pm$ one s.e.m.. (b) Noise in Hb as a function of Bcd concentration; error bars are $\pm$ one s.d. across embryos, and the curve is a fit from Tkačik et al. (2007a), cf. Section 3.2.

the expression noise variance down by a factor of ten is equivalent to assuming that all relevant molecules are present in ten times as many copies. We see in Fig 4.10 that there is a large regime in which the regulatory power is well approximated by the small noise approximation. In the opposite extreme, at large noise levels, we expect that there are (at best!) only two distinguishable states of high and low expression, so that our problem approaches the asymmetric binary channel (Silverman, 1955). The exact result interpolates smoothly between these two limiting cases with the real system ($F = 1$) lying closer to the small noise limit, but deviating from it significantly.

In the embryo, maximizing information flow from transcription factor to target gene has a very special meaning. Cells acquire "positional information," and thus can take actions which are appropriate to their position in the embryo, by responding to the local concentration of morphogen molecules (Wolpert, 1969). In the original discussions, "information" was used colloquially. But in the simplest picture of *Drosophila* development (Lawrence, 1992; Rivera-Pomar and Jäckle, 1996), information in the technical sense really does flow from physical position along the anterior–posterior axis to the concentration of the primary maternal gradients (such as Bcd) to the expression level of the gap genes (such as Hb). Maximizing the mutual information between Bcd and Hb thus maximizes the positional information that can be carried by the Hb expression level.

More generally, rather than thinking of each gap gene as having its own spatial profile, we can think of the expression levels of all the gap genes together as a code for the position

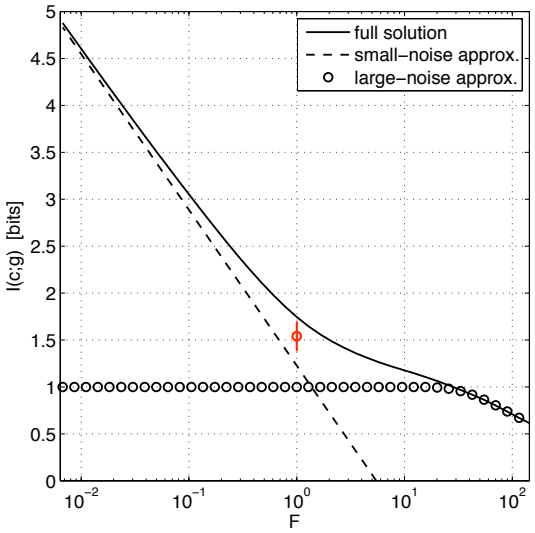


Figure 4.10: Optimal information transmission for the Bcd/Hb system as a function of the noise variance rescaling factor F. $1/F$ is approximately equal to the factor by which the number of input and output signaling molecules has to be increased for the corresponding gain in capacity. Dashed and dotted curves show the solutions in the small-noise and large-noise approximations, respectively. The real system, $F = 1$, lies in an intermediate region where neither the small-noise nor the large-noise approximation are valid. Measured information $I_{\rm data}(c;g)$ shown in red (errorbar is s.d. over 9 embryos).

of each cell. In the same way that the four bases (two bits) of DNA must code in triplets in order to represent arbitrary sequences of 20 amino acids, we can ask how many gap genes would be required to encode a unique position in the $N_{\rm rows} \sim 100$ rows of nuclei along the anterior–posterior axis. If the regulation of Hb by Bcd is typical of what happens at this level of the developmental cascade, then each letter of the code is limited to less than two bits ($I_{\rm opt} = 1.7$ bits) of precision; since $\log_2(N_{\rm rows})/I_{\rm opt} = 3.9$, the code would need to have at least four letters. It is interesting, then, to note that there are four known gap genes—*hunchback, krüppel, giant* and *knirps* (Rivera-Pomar and Jäckle, 1996)—which provide the initial readout of the maternal anterior–posterior gradients.

Instead of plotting Hunchback expression levels vs either position or Bcd concentration, we can ask about the *distribution* of expression levels seen across all nuclei, $P_{\rm exp}(g)$, as shown in Fig 4.11. The distribution is bimodal, so that large numbers of nuclei have near zero or near maximal Hb, consistent with the idea that there is an expression boundary—cells in the anterior of the emrbyo have Hb "on" and cells in the posterior have Hb "off." But intermediate levels of Hunchback expression also occur with nonzero probability, and the overall distribution is quite smooth. We can compare this experimentally measured distribution with the distribution predicted if the system maximizes information flow, and we see from Fig 4.11 that the agreement is quite good. The optimal distribution reproduces the bimodality of the real system, hinting in the direction of a simple on/off switch, but also correctly predicts that the system makes use of intermediate expression levels. From the data we can also compute directly the mutual information between Bcd and Hb levels, and we find $I_{\rm data}(c;g) = 1.5 \pm 0.15$ bit, or $\sim 90\%$ (0.88 ± 0.09) of the theoretical maximum.

The agreement between the predicted and observed distributions of Hunchback expression levels is encouraging. We note, however, some caveats. Bicoid has multiple targets and many of these genes have multiple inputs (Ochoa-Espinosa et al., 2005), so to fully optimize information flow we need to think about a more complex problem than the single input, single output system considered here. Measurement of the distribution of expression levels requires a fair sampling of all the nuclei in the embryo, and this was not the intent of the experiments of Gregor et al. (2006a). Similarly, the theoretical predictions depend somewhat on the behavior of the input/output relation and noise at low expression levels, which are

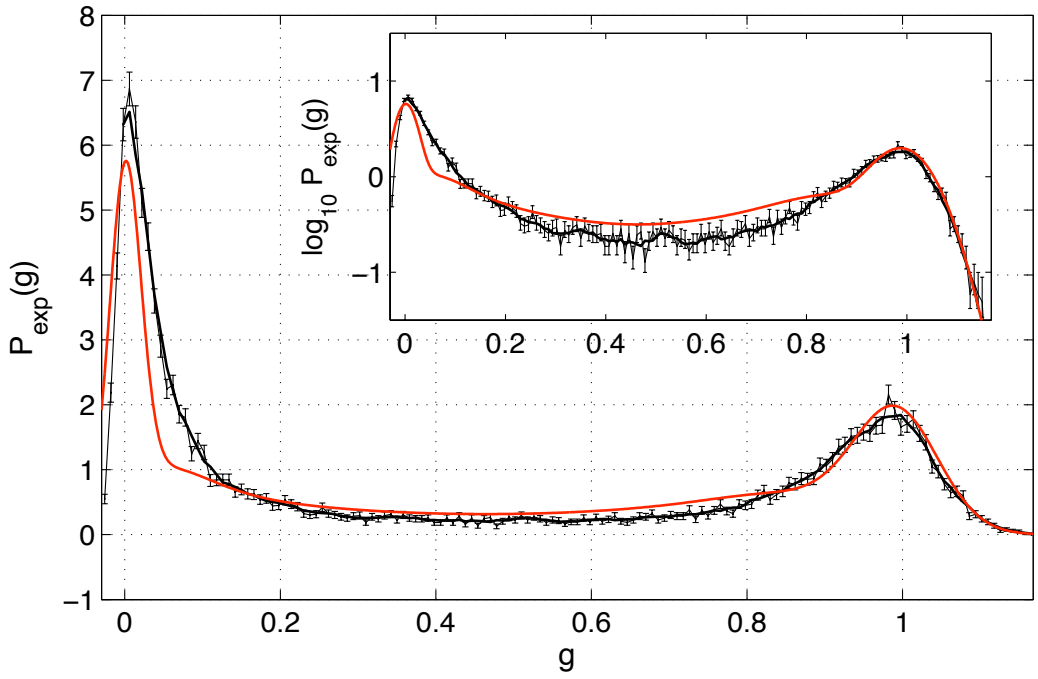


Figure 4.11: The measured (black) and optimal (red) distributions of Hunchback expression levels. The measured distribution is estimated from data of Gregor et al. (2006a), by making a histogram of the g values for each data point in Fig 4.9. The optimal solution corresponds to the capacity of $I_{\rm opt}(c;g) = 1.7\,$bits. The same plot is shown on logarithmic scale in the inset.

difficult to characterize experimentally, as well as the (possible) deviations from Gaussian noise. A complete test of our theoretical predictions will thus require a new generation of experiments.

In summary, the functionality of a transcriptional regulatory element is determined by a combination of its input/output relation, the noise level, and the dynamic range of transcription factor concentrations used by the cell. In parallel to discussions of neural coding (Laughlin, 1981; Brenner et al., 2000), we have suggested that organisms can make maximal use of the available regulatory power by achieving consistency among these three different ingredients; in particular, if we view the input/output relation and noise level as fixed, then the distribution of transcription factor concentrations or expression levels is predicted by the optimization principle. Although many aspects of transcriptional regulation are well studied, especially in unicellular organisms, these distributions of protein concentrations have not been investigated systematically. In embryonic development, by contrast, the distributions of expression levels can literally be read out from the spatial gradients in morphogen concentration. We have focused on the simplest possible picture, in which a single input transcription factor regulates a single target gene, but nonetheless find encouraging agreement between the predictions of our optimization principle and the observed distribution of the Hunchback morphogen in *Drosophila*. We emphasize that our prediction is not the result of a model with many parameters; instead we have a theoretical principle for what the system ought to do so as to maximize its performance, and no free parameters.

Chapter 5

Conclusion

In this work we discussed two approaches to understanding biological networks, or dynamical stochastic systems that process and transmit information in cells or organisms. The first approach, i.e. building maximum entropy distributions with correlation constraints, is data-driven and allows us to treat the network as a set of interacting nodes for which we can compute the map of phenomenological interactions in a principled way. By applying this method to the measurements of neural responses in retinal ganglion cells and activation patterns in a protein signaling cascade, we have learned a number of new things summarized below.

Firstly, despite an a priori possibility of finding very complex interactions, where k elements jointly cooperate and influence the network state, the two data sets are very satisfactorily explained by assuming that all interactions only happen between pairs of elements; this is a huge simplification in model complexity. Secondly, the network structure of two experiments is very different: in neurons we observe a dense mesh of all pairs interacting, which causes weak but significant correlation between every pair, in addition to strong collective effects once the network is big enough; in signaling proteins we observe a skeleton of a few strong interactions that explain the whole correlation structure. Thirdly, the pairwise interaction approximation is probably valid in neurons because one of the two possible states at a node (spiking vs non-spiking) is very rare; in contrast, in proteins we see at least one instance where a higher order (triplet) interaction has to be added to account for the data. Finally, in neurons we hypothesize about the role of collective states once the network approaches the "critical" size of a few hundred components, as being one of providing the error-correction capability to the population code.

In the second part of the thesis we introduced a new way of looking at building blocks of genetic regulatory networks, namely transcriptional regulatory elements. These noisy genetic components can be understood as information transmission devices in the sense defined by Shannon, and as such they can be tuned to achieve a well-defined maximum of the "regulatory power" when their properties are matched to the statistical properties of the incoming signals. In the case of genetic regulatory elements, the incoming signals are transcription factor concentrations, because this "chemical language" is the way in which the cell stores a representation of its state and the state of its environment. If such maximum of regulatory power, or information capacity, has been selected for during the course of evolution, our theory generates testable predictions about the distributions of TF concentrations in the cell; analysis of the data in the case of the fruit fly development seems to support the hypothesis that information transmission is maximized. Even in the absence of

such detectable optimization, however, information theory for genetic networks offers three substantial benefits when formulating questions about the network's functionality. Firstly, it defines a scalar measure for the power of the regulatory element in conjunction with its natural ensemble of inputs, that is a universal yard (or bit) stick for various regulatory elements. Secondly, it defines, in the same units, the limits to what the same regulatory element could achieve if it is perfectly matched to the input ensemble, and therefore sets a theoretical upper bound on the information transmission. Finally, it provides an interpretation of information capacity of a regulatory element in terms of biologically relevant concepts: for example, the number of distinguishable regulatory settings for a genetic switch, or the precision of spatial partitioning of the embryo for patterning systems during development.

There are a number of issues not addressed here in detail. In the phenomenological picture of interactions we do not discuss time dynamics and therefore the interactions between nodes are undirected. It seems possible to extend the maximum entropy modeling to encapsulate (at least a "one-timestep") dynamics, but this has not been done yet. Moreover, when the maximum entropy formulation is extended to cover the dependence on the stimuli, the problems of how to encode and include stimulus features in more general and complex cases are not worked out. When information transmission in the genetic network is discussed, a single regulatory element is extracted from the network, thus avoiding the problems posed by feedback loops (or topology in general) and issues pertaining to global information flow optimization versus local, single element, optimization. All these complications probably arise to some extent even in the fruit fly example considered here. Ideas about incorporating constraints other than maximal information transmission, such as the metabolic cost or robustness to parameter variations, are not worked out completely.

On the upside, it is easy to see the outlines of a more comprehensive theoretical program here. We are thinking in particular about approaching the first steps of some complex genetic program, for instance embryonic patterning, with the information transmission framework: information flows from position to morphogen gradients (via the physical process of gradient establishment) and from there to the gap genes (via the morphogen readout mechanisms). Given the upcoming generation of quantitative experiments exploring the early development, it seems conceivable that both approaches to understanding networks could be applied in parallel to great benefit. A successful theory of biological information processing in such a system requires three components: a model – almost by necessity phenomenological due to the complicated nature of elementary biophysical interactions – of the measured data; predictive power that validates the model on unobserved data, e.g. on knockout and other mutant perturbations; and most importantly, an answer to *how* the system achieves what it is supposed to do, that is, assign unique identities to nuclei that start out being the same. Phenomenological analysis of the data would uncover the structure of interactions between the genes; a stochastic dynamical model can then be postulated or inferred from the data (if they are abundant enough) and verified. When the agreement with measurements is satisfactory, such model, analyzed with information theoretic tools, would result in a prediction for its information capacity – a quantity for which we can develop a functional and intuitive understanding because it counts the number of distinguishable nuclear identities that *could* be conferred by the regulatory network performing at its best.

Appendix A

Methods

A.1 The biochemical network

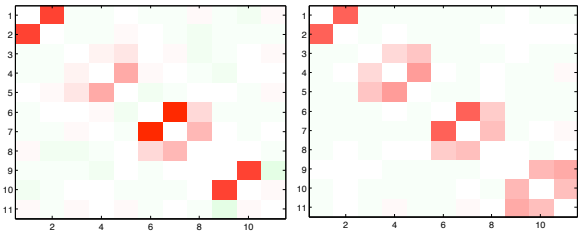
1. Interpreting data

The dataset of Sachs et al. (2005) consists of the simultaneous measurements of the activity level of $N = 11$ biomolecules (proteins and phospholipids) in the MAP cascade of human CD4+ T cells: RAF, MEK, PLCγ, PIP2, PIP3, ERK, AKT, PKA, PKC, p38 and JNK, which we number in this order. The cells were treated with flourescent antibodies recognizing specific phosphorylated forms of the biomolecules, and multicolor flow cytometry was performed to collect ≈ 700 single-cell samples at every condition in steady state. There are 9 conditions in total: the first 2 represent treatment with stimulatory agents ($S1$, $S2$) that activate the cells through their surface receptors; the remaining conditions utilize various additional combinations of non-natural intervention (inhibitory or activating) chemicals that interfere at (presumably known) points in the pathway: condition 3 ($C3 = S1 + I1$) inhibits AKT, $C4 = S1 + I2$ inhibits PKC, $C5 = S1 + I3$ inhibits PIP2, $C6 = S1 + I4$ inhibits MEK, $C7 = S1 + I5$ activates AKT, $C8 = S1 + I6$ activates PKC and $C9 = S1 + I7$ activates PKA. To be consistent with the experimenter's procedure of preprocessing the data when conditions were combined, we ignored raw values deviating more than 3σ from the mean and randomly drew a sample of exactly 600 measurements per condition from the remaining data before the analysis.

2. Quantizing data

Quantization that maximizes the multi-information is a hard computational problem: simulated annealing Metropolis Monte Carlo has been used to choose the best way to partition each of the 11 data series into 2 bins such that the multi-information among all of the 11 data series is maximal. The Monte Carlo moves consisted of increasing or decreasing the quantization boundary by a decile of data points in each data series independently. A custom annealing schedule was used that kept the ratio of rejected moves approximately constant. 100 quantizations were done, all of which converged to within the order of estimation error in entropy and gave essentially the same results on the whole dataset; the best run was chosen for further analysis. Following the analysis by Sachs et al. (2005), during quantization in the conditions that involve intervention ($C3 - C9$), the levels of perturbed biomolecules were set to high for activated species or to low for inhibited species.

Quantization that maximizes mutual information (which is the method to choose when



A.1a: Condition 1. A.1b: Condition 2.

Figure A.1: Average exchange interactions for conditions 1 and 2, calculated separately for each condition. 100 quantizations with random quantization boundaries are done for each condition, and the maximum entropy model is computed for every such quantization. The magnetic fields vary wildly from quantization to quantization because they constrain the single-element means (and those depend on the quantization choice); the average magnetic field over random quantizations is zero within error bars for all proteins. On the other hand the exchange interactions add coherently to yield the average interaction maps plotted above.

one is too undersampled for max-multi-info quantization, for instance for quantization into three bins) was performed by first quantizing data into 10 equipopulated bins and successfully coalescing neighboring bins in a greedy approach that minimizes the loss of average mutual information at each step until all the data has been quantized into 2 or 3 levels (see Supplementary Information of Sachs et al. (2005)). Small sample correction was used for all estimates, following the direct method (Slonim et al., 2005a).

Note that the main features of the interaction maps can be recovered even if complicated quantization schemes are not used. We could, for example, make quantizations in which the boundary between state "off" and "on" is chosen randomly and separately for each protein at decile boundaries, and repeat many times the maximum entropy reconstruction with these quantizations. The average of the interaction maps over random quantizations for conditions 1 and 2 is shown in Figs A.1a,A.1b, with clear similarities to Fig 2.11. By using the max-multi-info quantization one simply hopes to extract as much of significant correlations from the data as possible given the limited dynamic range of the discrete alphabet.

3. Constructing maximum entropy distributions

In the introduction to the chapter we only discussed how to compute maximum entropy models in which the activities could take on two distinct states. Let us generalize this to Q states here.[1] In other words, we are looking for $p(\sigma_1, \ldots, \sigma_N)$ such that $S[p(\vec{\sigma})]$ is maximized, subject to constraints:

$$p(\sigma_i = q_k, \sigma_j = q_l) = m_{ij}^{kl} \tag{A.1}$$

where $p(\sigma_i, \sigma_j)$ are two-point marginals of the wanted distribution, $q_k, q_l \in \{0, \ldots, Q-1\}$ and m_{ij}^{kl} are $N(N-1)/2$ measured marginal tables of dimension $Q \times Q$. This is a constrained

[1] We find that the dataset of Sachs et al. (2005) is too small to estimate the quality of maximum entropy reconstructions at $Q = 3$ (with max-mutual-info quantization, verification by measuring the three-point correlation functions and entropy estimations similar to Fig 2.13.

variational problem with an analytic solution:

$$p(\sigma_1, \cdots, \sigma_N) = \frac{1}{Z} \exp\left\{ \sum_i h_i(\sigma_i) + \sum_{i<j} J_{ij}(\sigma_i, \sigma_j) \right\} \tag{A.2}$$

Our task is then to solve constraint equations Eq (A.1), with probability distribution given by Eq (A.2), and with unknowns h, J. In this case, where the quantization level Q is larger then 2, the coupling J is not a single number, but is a matrix for each pair (i, j), and can be in general regarded as a pairwise "potential" that depends on the levels of both interacting elements.

It is easy to see why, in the binary case where $\sigma_i = \{-1, 1\}$, the general ansatz of Eq (A.2) can be rewritten into the Ising form [Eq (2.30)]. Each pairwise constraint is a 2×2 marginal probability table m with one normalization constraint and consequently 3 free parameters. Suppose then that these three independent parameters are m_{01}, m_{10} and m_{11}. Constraining the mean of the first spin, $\langle \sigma_i \rangle$, by the magnetic field h_i, is equivalent to constraining the linear combination $1 \cdot (m_{10} + m_{11}) + 0 \cdot (m_{00} + m_{01})$, and similarly for spin j. Constraining $\langle \sigma_i \sigma_j \rangle$ by the exchange coupling J_{ij} is equal to constraining $1 \cdot m_{11} + 0 \cdot (m_{00} + m_{01} + m_{10})$. The Lagrange multipliers of the Ising model and Lagrange multipliers that constrain the two-by-two marginal table in Eq (A.1) are consequently trivially related.

Note that the parametrization of Eq (A.2) contains more parameters than are needed to uniquely specify the distribution, because there are consistency constraints between elements of all marginal probability tables. To remove this ambiguity at $Q = 3$, we can select the following gauge for each i and each pair (i, j): $h_i(1) = 0$, $\sum_{q_i} V_{ij}(q_i, q_j) = \sum_{q_j} V_{ij}(q_i, q_j) = 0$.

Maximum entropy distributions of binary variables were computed using a custom Matlab code that solves a set of Eqs (A.1) with the distribution parametrized by Eq (A.2); the solution is accelerated by analytic evaluation of derivatives. For larger problems (i.e. $Q = 3$) and the cases where conditions were combined in one reconstruction, we use L1-regularized maximum entropy algorithm by Dudik et al. (2004). Regularization parameters of this procedure do not have large impact on the solution and are determined by computing the variance in constrained marginals in 20 random draws of half of the data at each condition.

A.2 Network of retinal ganglion cells

1. Bootstrap errors. Because the 40-neuron sample consists of responses to movie repeats, bootstrap errors are estimated by repeatedly taking a random half of the repeats[2] and estimating deviations in covariance and means. Figure A.2 shows the process of error estimation. Out of 145 repeats, first 25 are discarded to remove the systematic variation (adaptation of neurons apparent as a drift in the mean firing rate). Then, bootstrap replicas are generated by selecting, with replacement, $1, 2, \cdots, 64$ random repeats out of total of remaining 120, using 200 resamplings, and calculating $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$ over those subsamples. The deviation in the means and covariances is taken to be the sample error in the corresponding statistics, and is extrapolated to the full dataset. The upper right panel in Fig A.2 shows that this extrapolation procedure is reliable, with expected scaling (which is noticably less exact if bootstrapping is done without replacement). Effectively, this means we treat the movie repeats as independent draws from some underlying probability distribution that generates the whole 1310-sample-long spike train. Having so extrapolated the errors in means and correlations to the whole dataset or, alternatively, to one repeat, we compute the predictions for the errors in the mean and correlation.

For example, let the estimated mean of the spin σ_i be $\hat{\sigma}_i$. Because the spins are binary, the variance associated with the mean estimation is $(\operatorname{Std} \sigma_i)^2 = \langle \sigma_i \sigma_i \rangle - \langle \sigma_i \rangle^2 = 1 - \langle \sigma_i \rangle^2$ per draw, and is decreased to $\operatorname{Std} \hat{\sigma}_i = \frac{\operatorname{Std} \sigma_i}{\sqrt{N_{\text{eff}}}}$ if N_{eff} independent samples are measured.

By comparing the sample error in a statistic (inter-repeat deviation, on the left-hand side of the expression above) with the expected total deviation of the statistics (computable from the mean, on the right-hand side of the expression above) we can figure out the effective sample size, N_{eff}. For the true dataset with preserved repeat structure, the effective number of samples per repeat is $N_{\text{eff}} \approx 3500$. If timebins are randomly permuted in the dataset, so that the repeat structure is destroyed, and the bootstrap estimation is performed again, the $N_{\text{eff}}^{\text{rand}} \approx 1300$. The number of total samples in a repeat is 1310.

The interpretation of these results is as follows: the random reshuffling of the time bins generates a homogenous data set in which the partition into "repeats" is arbitrary. Therefore intra-repeat and inter-repeat variance in the estimates must be consistent, and indeed they are (see the black line fit in the lower panel in Fig A.2). In contrast, for real data set, the inter-repeat variance is too small given the statistics to which the Ising model was fit. This is due to the fact that, within each repeat, the response is driven by the stimulus and is not stationary, which makes the intra-repeat variance big compared to inter-repeat variance; the smallness of the latter basically says that the system gives reliable response to the same repeated stimulus. Because the inter-repeat variance is so small, it appears as if one needed approximately 2.7 times more samples within each repeat to make the observed variances equal. Turning this around, for a given intra-repeat variance, we really have just about 54 **independent** effective repeats of the data, instead of nominal 145.

2. Reconstruction precision. For 40 neurons we solve for $\{h_i, J_{ij}\}$ using Contrastive Divergence Monte Carlo (one-step learning (Hinton, 2002)) followed by gradient descent learning Eq (2.34) with $\alpha = 0$. For 120 neurons we use Eq (2.34) with the momentum $\alpha = 0.9$. The set of constrained operators consisted in this case of means and covariances C_{ij} instead of means and products $\langle \sigma_i \sigma_j \rangle$; the latter change is made because otherwise a small error in the reconstruction of a mean systematically influences a large number of C_{ij}, with consequences for the thermodynamic quantities of interest. See Fig A.3 for

[2]Note that this is not the same as taking any random half of the data.

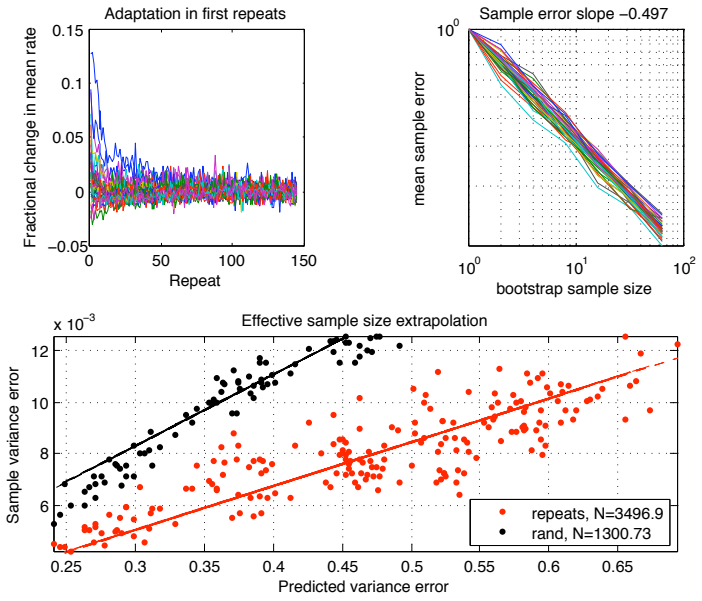


Figure A.2: Upper left: non-stationarity in the statistics in the first ~ 20 repeats of the movie. Upper right: bootstrap error extrapolation for all means and covariances of 40 neurons with bootstrap sample size (N_r repeats) on x-axis (200 repeats for each bootstrap sample). The extrapolation is reliable and estimated errors decrease as $1/\sqrt{N_r}$, as expected. Lower panel: scatterplot of total deviation in each correlation, $\sqrt{1-\langle\sigma_i\sigma_j\rangle^2}$ against the per-repeat sampling error in the same correlation. Red dots show the scatterplot when the repeat structure of the data is preserved; black dots show the time-axis reshuffled, so that the time structure is destroyed. The fits are used to calculate N_{eff}, the effective number of samples in a given repeat.

reconstruction precision. Monte Carlo simulation was implemented in C/C++, and the learning algorithm in Matlab; Matlab drove parallel instances of Monte Carlo simulations on the cluster. Order 10^6 independent samples (at least $2N$, but sometimes more, spin flips were made before a sample was drawn from MCMC) were taken for the computation of operator averages at each step of the learning iteration.

3. Energy histograms. If we take the Hamiltonian of the pairwise model that is computed from the correlations in the data, we can evaluate the energy of all patterns observed in the data set; in addition, we can *calculate* (or simulate with Monte Carlo for bigger cases) what the expected distribution of the energies would be. Figs A.4a and A.4c show energy histograms for 40 and 20 neurons, respectively. In addition to the probability of observing K simultaneous spikes [Fig 2.18] this is yet another projection of 40 (20) dimensional sample space down to one scalar dimension. We see the expected disagreement in the ground state frequency (related to the $P(K=0)$ disagreement of the firing curve), but also a more marked discrepancy in 40-neuron MC energy histogram in the tail.

4. Comparison with $n = 20$ neuron reconstructions. To decide whether the systematic deviations in the case of three-point correlations and the firing curve are a consequence of imperfect MC reconstruction or a real effect, we perform the same analysis on a pairwise maxent model of first 20 out of 40 neurons, for which we have a convergent, deterministic algorithm. Figure A.5a shows, as expected, that the reconstruction has converged to within sampling limits, and that MC simulations for computing back the observables given the fully converged coupling constants work without bias.

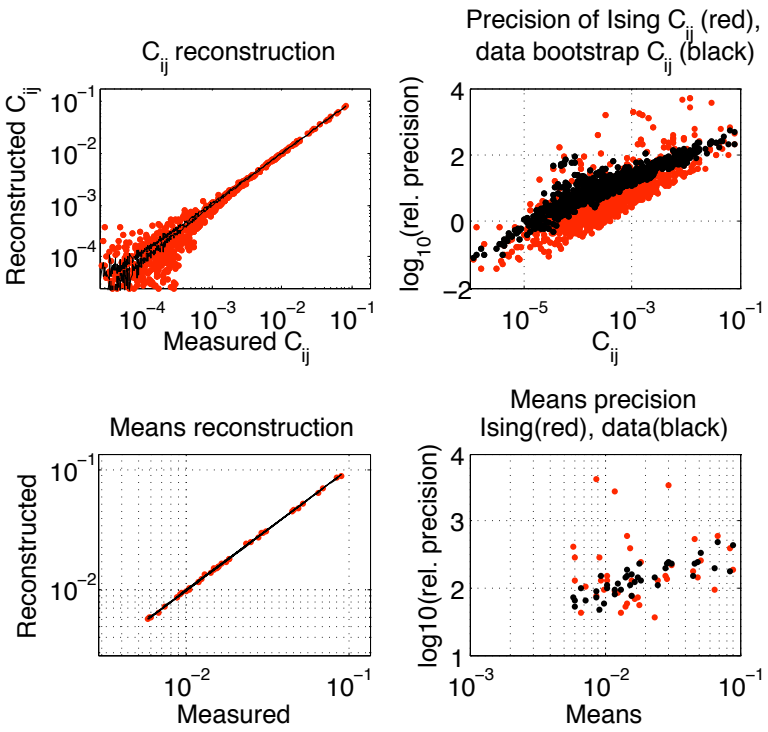
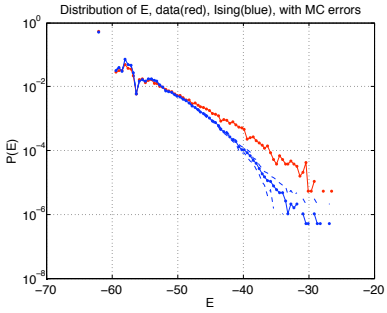


Figure A.3: Reconstruction precision of Monte Carlo (MC) on full data set of 40 neurons. Upper left: measured covariance vs reconstructed covariance. Reconstruction is averaged over 5 MC runs of 190k samples each. Black lines represent error estimates obtained by taking a random selection of a half of the samples from the real data, and estimating covariances 5 times. Lower left: the same plot for mean firing rates. Upper right: a plot of precision of reconstructed variances, $-\log_{10}\left(\frac{C^{\mathbf{g}}_{ij}-C^{\mathrm{expt}}_{ij}}{C^{\mathrm{expt}}_{ij}}\right)$, where C^{expt}_{ij} is obtained from the data and $C^{\mathbf{g}}_{ij}$ is the average of 5 MC runs, in red; in black, bootstrap error of the covariance element estimates plotted in the same manner for 5 half-size subsamples of the data. Lower right: the same plot for mean firing rates.

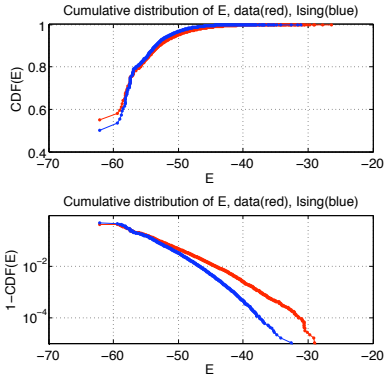
Figure A.5b shows that firing curve deviates with the same systematics as in 40 neuron case. Here, the probabilities of silence in the data and Ising model are $P_D(0) = 0.621$ vs $P_M(0) = 0.599$, a deviation of about one-half of the size of that for 40 neurons. In addition, the deviations between data and predictions for K simultaneous spikes are in the same direction as in the 40-neuron case.

Furthermore, the triplet correlations show the same deviation as in 40-neuron case, in which the three-point correlations are over-estimated by the pairwise model [Fig 2.18]. I have again checked that this is not a consequence of faulty MC implementation, as we have the ability to explicitly sum the partition function and compute the three-point correlations exactly: MC and explicit calculations agree with no systematics (not plotted). The consistency of results between 20 and 40 neurons indicates that although the 40-neuron reconstruction does not perfectly converge to the data, this fact is *not* responsible for disagreements in the three-point correlations and the firing curve; on the contrary, it is probably an indication that while second-order Ising is a good model that captures most of the features of the data, it does not capture all of them.

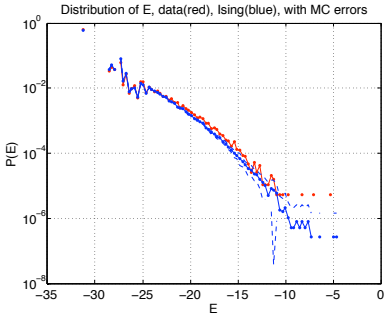
5. Reconstruction of 40-neuron coupling constants from 20-neuron subnetworks



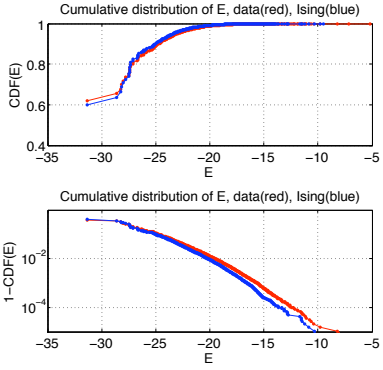
A.4a: Energy, 40 neurons



A.4b: Energy, 40 neurons, cumulative

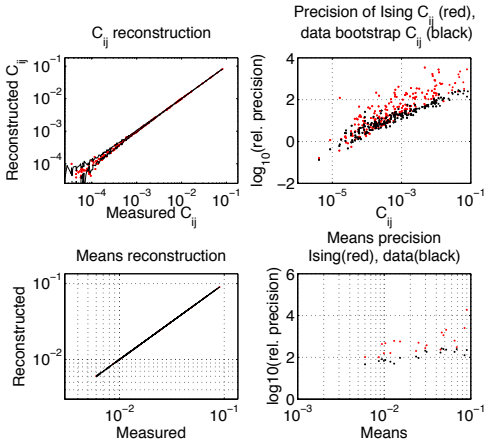


A.4c: Energy, 20 neurons

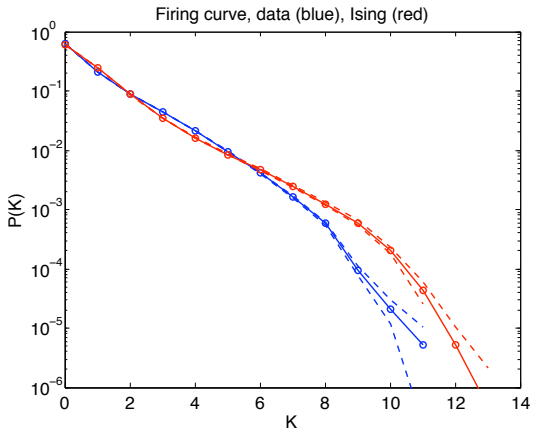


A.4d: Energy, 20 neurons, cumulative

Figure A.4: Energy histograms for 20 and 40 neurons. The MC-reconstructed couplings are used to calculate the energy of each sample in the real dataset and the histogram of such energies is plotted in red. 5 MC simulations of 190k samples each are used to generate the corresponding blue curve, and the dashed blue lines describe the error estimates. The line is broken in several places because the density of states is low at low energies (e.g. ground state is far away from the bulk).



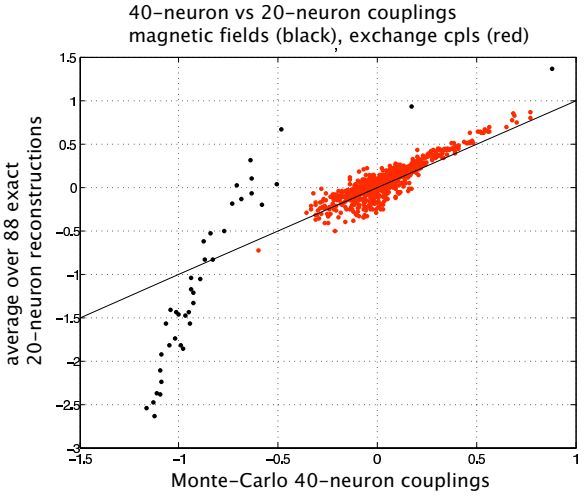
A.5a: Reconstruction precision for the first 20 neurons.



A.5b: Firing curve for first 20 neurons.

Figure A.5: Reconstruction precision (a) and the probability $P(K)$ of observing K simultaneous spikes (b) for the first 20 neurons of the dataset. Computed using L-1 regularized maximum-entropy algorithm of Dudik et al. (2004), a convergent and deterministic learning procedure for $\{h_i, J_{ij}\}$. Although the observables are fit to within better than bootstrap error estimates, there is nevertheless a systematic deviation both in the firing curve and in the triplet correlations [Fig 2.18], showing that in the 40-neuron case the incomplete convergence of the Monte-Carlo reconstruction is not the cause of the disagreement of the model and the high-order statistical features in the data.

Figure A.6: Comparison between the 40-neuron couplings computed in a MC reconstruction (horizontal axis), with the corresponding average couplings reconstructed in 88 20-neuron subnetworks (vertical axis). Magnetic fields are shown in black and exchange interactions in red. Each 20-neuron coupling appears in several 20-neuron subnetworks, and the figure shows the average of that coupling across all network instances where it appears; the error bars in pairwise couplings across reconstructions are usually on the order of 20 percent for bigger couplings.



Here we explore the agreement between coupling constants reconstructed by doing maximum entropy on 88 subnetworks of size 20, using exact calculation, and coupling constants obtained by full 40-neuron MC. Comparison of couplings reconstructed exactly from 20-neurons with MC 40-neuron reconstructions shows interesting features: A) magnetic fields follow a linear trend and are by a factor of ≈ 2.2 bigger for 20-neuron case; B) pairwise couplings exhibit the same slope, but are offset by approximately a constant relative to the 40-neuron case, see Fig A.6. The increase in 20-neuron magnetic fields is likely to be ac-

counted for by the mean-field of the unobserved 20-neurons; the change in pairwise coupling between two arbitrary neurons N1 and N2 is a consequence of (unobserved) N3 that couples to both N1 and N2. In this case, it could also be expected that the three-point couplings (induced by unobserved elements) are smaller in 40-neuron case and that the disagreement in three-point correlations is somewhat smaller for the bigger system [Fig 2.18].

6. Decimation flow of coupling constants. We would like to study the behavior of the mapping $\{h, J\}_N \to \{h, J\}_{N-1}$, i.e. the flow of coupling constants when we stop keeping track of the last, or N-th, neuron (e.g. suppose this is the neuron we cannot observe). The idea is to start out with a probability distribution defined on N neurons (we decide to keep the three-point interaction term):

$$p^{(N)}(\{\sigma_1, \ldots, \sigma_N\}) = \frac{1}{Z} \exp\left(\sum h_i \sigma_i + \frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j + \frac{1}{6} \sum_{ijk} J_{ijk} \sigma_i \sigma_j \sigma_k \right)$$

We are looking for the probability distribution $p^{(N-1)}$ of the same form, but defined over $N-1$ neurons, and the relation between the two is marginalization:

$$p^{(N-1)}(\sigma) = \sum_{\sigma_N} p^{(N)}(\sigma)$$

In general, the marginalization cannot hold exactly while the forms of both distributions are fixed to an Ising form with two- or three-point interactions. However, our experiments on real data show that a two-point Ising is a good description for any size, from 2 neurons to 40, of the observed spike train statistics. We might be hopeful, therefore, that the interactions of level three and higher will be negligible.

Since the spiking is rare, let's switch first into sparse representation $\sigma_i \to 2\sigma_i - 1$, with the new $\sigma_i \in \{0, 1\}$. Then, the marginalization reads:

$$\begin{aligned} p^{(N-1)}(\sigma) &= \frac{1}{Z'} \exp\left(-H^{(N-1)(\sigma)}\right) \times \\ &\times \sum_{\sigma_N} \exp\{h_N(2\sigma_N - 1) + \\ &+ \sum_i J_{iN}(2\sigma_N - 1)(2\sigma_i - 1) + \frac{1}{2} \sum_{ij} J_{ijN}(2\sigma_i - 1)(2\sigma_j - 1)(2\sigma_N - 1) \Bigg\} \end{aligned}$$

This equation is now exactly summed over σ_N; the terms resulting from the summation are expanded (up to fourth order) in σ_i. If we define renormalized $\tilde{J}_{iN} = J_{iN} - \sum_j J_{ijN}$ (use renormalized two-point J from now on without a tilde) , and introduce factors $\theta^{(N)} =$

$\exp\left(2h_N - 2\sum_i J_{iN} + \sum_{ij} J_{ijN}\right)$ and $\omega^{(N)} = \frac{\theta^{(N)}-1}{\theta^{(N)}+1}$, such expansion reads:

$$\begin{aligned}
p^{(N-1)}(\sigma) &= \frac{1}{Z'}\exp(-H^{(N-1)}(\sigma))(\theta+1)\times \\
&\times \left\{1 + 2\omega\sum_{iN} J_{iN}\sigma_i + 2\omega\sum_{ijN} J_{ijN}\sigma_i\sigma_j + 2\sum_{ij} J_{iN}J_{jN}\sigma_i\sigma_j + \right. \\
&+ 4\sum_{ijk} J_{iN}J_{jkN}\sigma_i\sigma_j\sigma_k + 2\sum_{ijkl} J_{ijN}J_{klN}\sigma_i\sigma_j\sigma_k\sigma_l + \\
&+ \frac{8\omega}{6}\sum_{ijk} J_{iN}J_{jN}J_{kN}\sigma_i\sigma_j\sigma_k + 4\omega\sum_{ijkl} J_{iN}J_{jN}J_{klN}\sigma_i\sigma_j\sigma_k\sigma_l + \\
&+ \left.\frac{2^4}{24}\sum_{ijkl} J_{iN}J_{jN}J_{kN}J_{lN}\sigma_i\sigma_j\sigma_k\sigma_l\right\}
\end{aligned}$$

Since the distribution for $N-1$ neurons also has to take the Ising form, we require that the expansion in curly braces above be an expansion of a term of the form (sums over repeated indices are understood, but notice that sums are taken in the expression above as well as below also over equal indices $i = j = k \cdots$):

$$f = \exp\left\{\alpha_i\sigma_i + \beta_{ij}\sigma_i\sigma_j + \gamma_{ijk}\sigma_i\sigma_j\sigma_k + \delta_{ijkl}\sigma_i\sigma_j\sigma_k\sigma_l + \cdots\right\}$$

Matching terms order-by-order we identify:

$$\begin{aligned}
\alpha_i &= 2\omega J_{iN} \\
\beta_{ij} &= 2J_{iN}J_{jN}(1-\omega^2) + 2\omega J_{ijN} \\
\gamma_{ijk} &= 4(1-\omega^2)\left(J_{iN}J_{jkN} - \frac{2}{3}\omega J_{iN}J_{jN}J_{kN}\right) \\
\delta_{ijkl} &= (1-\omega^2)\left(2J_{ijN}J_{klN} - 6\omega J_{iN}J_{jN}J_{klN} - \right. \\
&- \left. 2\omega J_{ijN}J_{kN}J_{lN} + \frac{4}{3}(3\omega^2-1)J_{iN}J_{jN}J_{kN}J_{lN}\right)
\end{aligned}$$

Now that the expression for decimated probability distribution has been collapsed back to the exponential form, we switch into $\sigma_i \in \{-1, 1\}$ representation and contract the sums on equal summation indices, since $\sigma_i^2 = 1$. This yields the final mapping for the coupling constants as the system is reduced by marginalizing over N-th neuron:

$$h_i \rightarrow h_i + \frac{1}{2}\alpha_i + \frac{3}{8}\sum_{jk}\gamma_{ijk} + \frac{4}{16}\sum_{jkl}\delta_{ijkl} + 3\sum_j \tilde{\gamma}_{ijj} - 2\tilde{\gamma}_{iii} \tag{A.3}$$

$$\frac{1}{2}J_{ij} \rightarrow \frac{1}{2}J_{ij} + \frac{1}{4}\beta_{ij} + \frac{3}{8}\sum_k \gamma_{ijk} + \frac{6}{16}\sum_{kl}\delta_{ijkl} + \frac{6}{16}\sum_k \delta_{ijkk} - \frac{2}{16}\delta_{ijjj} \tag{A.4}$$

$$\frac{1}{6}J_{ijk} \rightarrow \frac{1}{6}J_{ijk} + \frac{1}{8}\gamma_{ijk} + \frac{4}{16}\sum_l \delta_{ijkl} \tag{A.5}$$

where $\tilde{\gamma} = \frac{1}{8}\gamma_{ijk} + \frac{4}{16}\sum_l \delta_{ijkl}$.

The form of Eqs (A.3, A.4, A.5) explains both why pairwise Ising model can successfully describe less than N neurons if it works on N neurons; and the scaling of the coupling

constants seen in Fig A.6. If parameter $|\omega|$ is close to 1, factors of $1-\omega^2$ are found in the expressions for β, γ and δ, and they will tend to make β, γ, δ close to zero; the only sizeable contribution is therefore to α, which renormalizes magnetic fields; and would be to β from three-point coupling terms. However, as we postulate (and show on real data) that two-point Ising is a good description for N neurons, these 3-point couplings are 0 for N neurons and can grow only through γ and δ terms, which again are suppressed as $1-\omega^2$.

If $\theta \to 0$, then $\omega \to -1$, which happens when $h_N - \sum_i \tilde{J}_{iN}$ tends to be very negative. Because we can rewrite $h_N - \sum_i \tilde{J}_{iN} \approx h_N + \sum_i \tilde{J}_{iN}\langle\sigma_i\rangle$, we see that this limit is actually a requirement that in mean field treatment, the effective bias experienced by spin N must be such that the $\langle\sigma_N\rangle \approx -1$.

7. Calculation of the entropy for 40 neurons. The entropy of the 40-neuron system can be estimated in two similar ways: in method A, one starts with magnetic fields that reproduce the observed 1-point marginals and zero pairwise couplings (in this state we know how to calculate the entropy), and then magnetic fields and couplings are varied to their final values. In the process $S = S_0 - \sum_\mu \int d\langle O_\mu\rangle g_\mu$.[3] This process yields a multi-information of $I = 0.689$ bits. As a sanity check, for a 20-neuron system, it yields $I = 0.368$ bits, compared to the exact value (by state enumeraton) of $I = 0.369$ bits.

Method B for estimating the entropy involves heating up the system from low temperature (where it is in the ground state at $S_0 = 0$) to the final state with $T = 1$, using the fact that heat capacity is $C = T\frac{\partial S}{\partial T}$ and $C = \frac{1}{k_B T^2}\langle(\delta E)^2\rangle$, which can be evaluated either analytically (for 20 neurons) or in a MC simulation by sampling. The entropy for 40 neurons is thus $S = 5.466$, independent entropy $S_0 = 6.1534$ and multi-information $I = 0.6868$, all in bits, in excellent agreement with method A.

There exists an entropy bound, or Ma entropy estimate (Strong et al., 1998) that we can use for binary patterns discussed here. On the 40-neuron dataset this estimate gives $S = 4.66$ bits, and on 20-neuron dataset $S = 3.36$ bits (for this case, sample-sized corrected direct estimate from the data is $S = 3.65$ bits and maximum entropy model is $S = 3.69$ bits).

Figure A.7 shows the scaling of the entropy and multi-information with the size of the system. Notice that the expected exponents for the scaling are 1 for entropy and 2 for multi-information (in the limit of weak coupling). The point at which the curves would intersect, yielding a zero-entropy frozen system, is at $N_c \approx 300$ (with the entropy starting to decrease at about $N_c \approx 200$), cf. Schneidman et al. (2006). This point cannot be achieved as entropy does not decrease with increasing size, and the real system must either be composed from a smaller number of elements, or the extrapolation loses its validity at $N < N_c$. For fits based on up to 20 neurons only, the exponents are 0.96 for the entropy and 2.05 for the multi-information, and the intersection happens at about 200 neurons (Schneidman et al., 2006).

8. Ground states. 40-neuron system has 5 1-spin-flip-stable ground states $\mathcal{G}_\alpha$ found in the dataset, depicted (with their corresponding energies) in Fig A.8a. These states were found by taking the real data sample and examining the stability of each distinct state.[4] If such state can be perturbed by a single spin flip so that the energy is lowered, this is done until a local energy minimum is reached. Then the complete 190k data series can be projected onto its corresponding ground states, as shown in Figs A.8b, A.9.

[3] A quick way to show this is to take the derivative of the entropy with respect to coupling g_μ, and write the entropy as $S = \log Z - \sum_\mu g_\mu \langle\hat{O}_\mu\rangle$.

[4] There are probably some stable states that have not been discovered, especially if they have high energies.

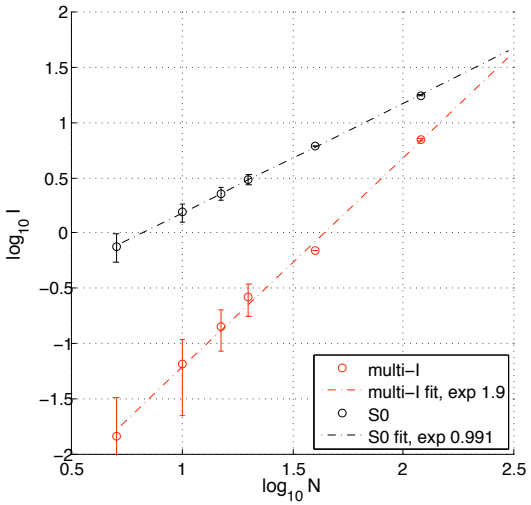


Figure A.7: Scaling of entropy and multi-information with system size. The fit exponents are shown in the legend. There is 1 data point for 40 neurons, around 100 reconstructed subnetworks for 20 and 15 neurons, 200 for 10 neurons and 400 for 5 neurons. The data point at $N = 120$ is an average over 3 instances of the synthetic network model, obtained by drawing correlations and means from their measured distributions. The intersection of the independent entropy and multi-information is around $N_c \approx 300$.

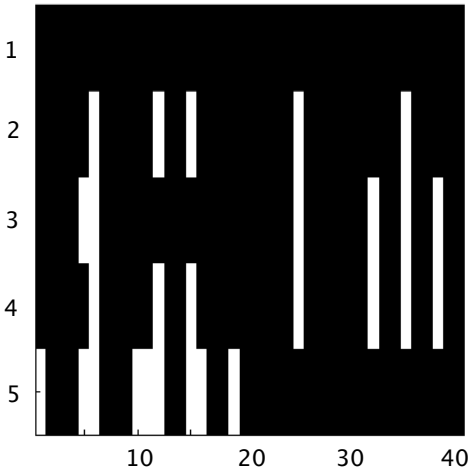
For each of the ground states we can try to estimate the size of the "basin" of states that are assigned to it. The number of samples for each state but the state 1 (silence) is small: out of 189950 total samples, 658 (383 distinct, ground state appears 70 times, with approximately 10-20 other states appearing around 10 times, and others being single repeats) correspond to state 2, 295 (113 distinct, ground state appears around 40 times, and there are around 20 states appearing around 5 times, others being single or double repeats) to state 3, 154 (57 distinct, ground state appears 20 times, there are around 5 other states appearing 10-15 times, and the remaining are single repeats) to state 4 and 235 (152 distinct, ground state appears around 25 times, there are order of 20 states apparing 3-5 times, the remaining are single repeats) to state 5. One can try to use the Ma estimate of entropy of these states (i.e. $S[P(\{\sigma_i\}\,|\mathcal{G}_\alpha)]$) from Strong et al. (1998), to get 7.5 bits for state 2, 5.1 bits for state 3, 4 bits for state 4, 5.2 bits for state 5 and 4.6 bits for silence (which is similar to total Ma entropy estimate for the whole sample).

The information conveyed by the five possible ground state assignments about the time (which stands as a proxy for the stimulus) is small, $I(\mathcal{G}_\alpha; t) \approx 0.046$ bit. This is bounded from above by the entropy in the ground state assignment distribution, which is very skewed by the frequent occurence of the trivial ground state, $S[p(\mathcal{G}_\alpha)] \approx 0.07$ bit. The brain thus learns a lot on the rare occasion when a specific non-trivial state happens, but most frequently it does not learn much from such a small piece of the retina with this *stable-state code*.

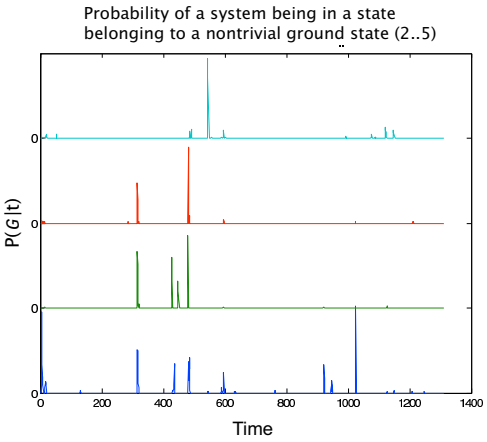
9. Properties of low-lying stable states For a network of 120 neurons we can generate a large sample ($10 \cdot 10^6$ independent samples, drawn from a long MCMC run where K spin-flips are performed between successive draws), and identify lowest lying stable states by zero-temperature Monte Carlo (ZTMC starts with the pattern and iterates Eq (A.6) until no more spins are flipped):

$$\sigma_i \leftarrow \text{sign}\left[\sum_j J_{ij}\sigma_j + h_i\right] \tag{A.6}$$

For lowest 1000 stable states, arranged by energy, we want to compute several local statistics, such as magnetization within the basin, entropy of the basin, average energy etc. For this, every MC sample in a long simulation run is assigned to its basin using ZTMC,



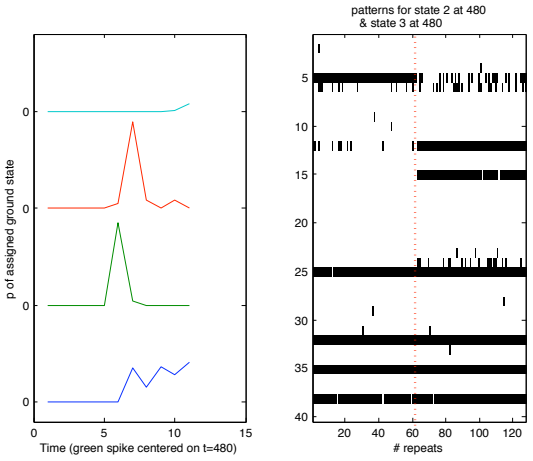
A.8a: Stable states of the 40-neuron system.



A.8b: Projection of the real data series onto the stable states.

Figure A.8: Left: firing (white) and silence (black) patterns of the non-trivial stable states. Index α of the stable state $\mathcal{G}_\alpha$ runs along the vertical axis, and the neuron index $i = 1, \ldots, 40$ along the horizontal. Right: projection of the real data series onto the stable states, averaged over 145 repeats. Time is plotted on the horizontal axis in units of $\Delta t = 20\,\text{ms}$; vertical axis shows the probability of the system being in a certain state that corresponds to one of the four non-trivial stable states.

Figure A.9: Left: zoom-in from Fig A.8b for the time course (horizontal axis, units of $\Delta t = 20\,\text{ms}$) of probability of observing non-trivial stable states $\mathcal{G}_\alpha$ (vertical axis). Right: patterns across repeats in which state 2 appears at time-bin 480 (left half of the right-hand figure) followed by appearance of state 3 in the time-bin 481 (right half of the right-hand figure).



and we hope to collect enough patterns in each basin to make the local statistics feasible to estimate. First, we have to determine K, the minimum number of spin-flips in the Metropolis procedure that breaks the Markov chain correlations with the previous sample. We start the Markov Chain in stable states that we have pre-identified, and measure the average number of spin-flips needed to depart from that stable state and fall into the basin of another stable state. Figure A.10 shows this distribution in lower right corner: we set $K = 1000$ as the number of spin flips between sample draws. With this value K we can accumulate enough samples to estimate the local average energy of basin of attraction $\mathcal{G}_i$, as well as a naive estimate of the entropy in that basin. As a consistency check we then plot the log (empirical MC) probability of observing a pattern from basin $\mathcal{G}_i$ against the free

energy $F(\mathcal{G}_i) = \langle E \rangle_{\mathcal{G}_i} - S(\mathcal{G}_i)$ and find a linear relationship with a slope 1.07 between both quantities, thereby confirming that our local estimates are reasonably good and properly sampled (note that we only track 1000 lowest states out of many more, but the lowest 1000 account for most of the probability weight [Fig 2.20]).

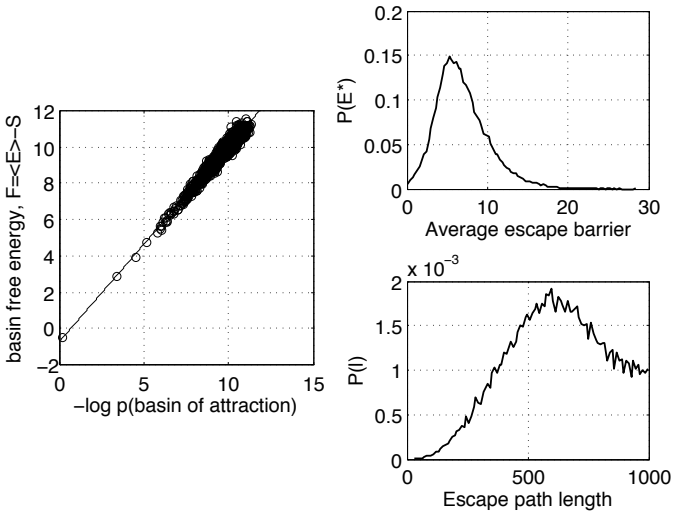


Figure A.10: Properties of the lowest-lying 1000 stable states. Upper right: the average energy barrier for escape from a stable state to a nearby stable state, averaged across the "from" and "to" states. For each of the 1000 states studied, we attempted 1000 escapes to a neighboring state in the pattern space. The path from the stable state to the separatrix between the two states is also measured and histogrammed in lower right plot; from here we pick $K \approx 1000$ as the number of flips between samples. Left: for each energy basin we can estimate the log probability of finding a state from that basin (horizontal axis) and compare this to the free energy of the basin, $F(\mathcal{G}_i)$, with a satisfactory linear match (black line).

10. Redundancy. How much information does a group of neurons contain about another neuron in the network or about the identity of the basin of attraction?

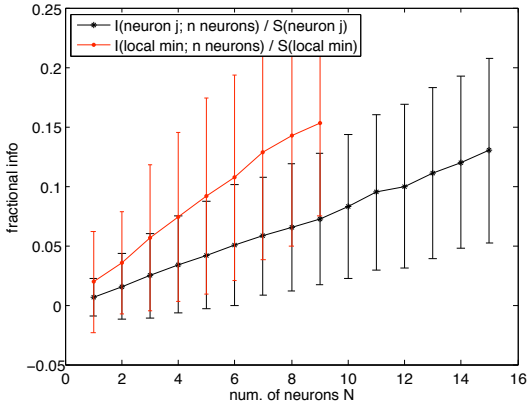


Figure A.11: The scaling of information between a group of neurons about another neuron (black) and a group of neurons about the identity of the basin of attraction (red). We repeatedly take groups of N neurons (horizontal axis) and compute the desired information, error bars are across repeats. The information is normalized by the output entropy, as shown in the legend.

Figure A.11 shows the scaling of these information quantities for small enough subgroups of 120 neurons, for which we can generate enough MC samples using the synthetic model. One learns approximately twice as fast as the number of neurons grows about the stable state (with extrapolated "full knowledge" at 60 neurons, or one half of the total network

size) in comparison with learning about the state of the single neuron (extrapolated "full knowledge" at 120 neurons, i.e. the network is close to error-correcting for single neuron fluctuations). Alternatively, it is easier to infer the collective state of the network than the state of a single neuron from a subgroup of neurons, a property that might be desirable for reliable and robust downstream decoding of a population code.

A.3 Input and output noise in transcriptional regulation

1. Langevin noise model

We consider a simplified model of regulated gene expression, as schematized in Fig 3.2:

$$\partial_t c = D\nabla^2 c(\mathbf{x},t) - \dot{n}\,\delta(\mathbf{x}-\mathbf{x}_0) + \mathcal{S} - \mathcal{D} \quad (A.7)$$

$$\dot{n} = k_+ c(\mathbf{x_0},t)(1-n) - k_- n + \xi_n \quad (A.8)$$

$$\dot{e} = R_e n - \tau_e^{-1} e + \xi_e \quad (A.9)$$

$$\dot{g} = R_g e - \tau_g^{-1} g + \xi_g. \quad (A.10)$$

Equation (A.7) describes the diffusion of the transcription factor that can be absorbed to or released from a binding site on the DNA located at $\mathbf{x}_0$. These transcription factors are produced at sources $\mathcal{S}$ and degraded at sinks $\mathcal{D}$, which can both be spatially distributed and can also contribute to the noise in c. Equation (A.8) describes the dynamics of the binding site occupancy; binding occurs with a second order rate constant k_+ and unbinding with a first order rate constant k_-, and the dissociation constant of the site is $K_d = k_-/k_+$. The Langevin term ξ_n induces stochastic (binomial) switching between occupied and empty states of the site. Equations (A.9) and (A.10) describe the production and degradation of mRNA and protein, respectively, and include Langevin noise terms associated with these birth and death processes.

This seems a good place to note that, while conventional, the assumption that transcription and translation are simple one step processes seems a bit strong. We hope to return to this point at another time.

Our goal is to compute the variance in protein copy number, $\sigma_g^2(\bar{c})$. For simplicity we will assume that the transcription factors are present at a fixed total number in the cell and that they do not decay, $\mathcal{S} = \mathcal{D} = 0$. We will see that even with this simplification, where the overall concentration of transcription factors does not fluctuate, we still get an interesting noise contribution from the randomness associated with diffusion in Eq (A.7).

Our basic strategy is to find the steady state solution of the model, and then linearize around this to compute the response of the variables $\{n, e, g\}$ to the various Langevin forces $\{\xi_n, \xi_e, \xi_g\}$. In the linear approximation, the steady states are also the mean values:

$$c = \bar{c} \quad (A.11)$$

$$\langle n\rangle = \frac{k_+\bar{c}}{k_+\bar{c}+k_-} = \frac{\bar{c}}{\bar{c}+K_d} \quad (A.12)$$

$$\langle e\rangle = R_e\tau_e\langle n\rangle \quad (A.13)$$

$$\langle g\rangle = R_g\tau_g\langle e\rangle = g_0\langle n\rangle, \quad (A.14)$$

where $g_0 = R_e\tau_e R_g\tau_g$ is the maximum mean expression level. Notice that what we have called $\bar{g} = \langle g\rangle/g_0$ is just the mean occupancy, $\langle n\rangle$, of the transcription factor binding site.

Small departures from steady state are written in a Fourier representation:

$$c(\mathbf{x},t) = \bar{c} + \int \frac{d\omega}{2\pi} \int \frac{d^3k}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} e^{-i\omega t} \delta\hat{c}(\mathbf{k},\omega) \tag{A.15}$$

$$n = \langle n\rangle + \int \frac{d\omega}{2\pi} e^{-i\omega t} \delta\hat{n}(\omega) \tag{A.16}$$

$$e = \langle e\rangle + \int \frac{d\omega}{2\pi} e^{-i\omega t} \delta\hat{e}(\omega) \tag{A.17}$$

$$g = \langle g\rangle + \int \frac{d\omega}{2\pi} e^{-i\omega t} \delta\hat{g}(\omega). \tag{A.18}$$

Similarly, each of the Langevin terms is written in its Fourier representation,

$$\xi_\mu = \int \frac{d\omega}{2\pi} e^{-i\omega t} \hat{\xi}_\mu(\omega), \tag{A.19}$$

where $\mu = n, e, g$.

As a first step we use the Fourier representation to solve Eq (A.7) for $\delta c(\mathbf{x_0}, t)$ that we need to substitute into Eq (A.8) for the binding site occupancy:

$$\delta c(\mathbf{x_0},t) = \int \frac{d\omega}{2\pi} e^{-i\omega t} \delta\tilde{c}(\mathbf{x_0},\omega) \tag{A.20}$$

$$\delta\tilde{c}(\mathbf{x_0},\omega) = i\omega\delta\hat{n}(\omega) \int \frac{d^3k}{(2\pi)^3} \frac{1}{-i\omega + D|\mathbf{k}|^2} \tag{A.21}$$

$$= \frac{i\omega\delta\hat{n}(\omega)}{\pi D a}. \tag{A.22}$$

The integral over $\mathbf{k}$ in Eq (A.21) is divergent at large $|\mathbf{k}|$ (ultraviolet). This arises, as explained in (Bialek and Setayeshgar, 2005), because we started with the assumption that the binding reaction occurs at a point—the delta function in Eq (A.7). In fact our description needs to be coarse grained on a scale corresponding to the size of the binding site, so we introduce a cutoff so that $|\mathbf{k}| \leq k_{\max} = 2\pi/a$, where a is the linear size of the binding site.

Linearizing Eq (A.8) for the dynamics of the site occupancy, we have

$$-i\omega\delta\hat{n}(\omega) = -(k_+\bar{c} + k_-)\delta\hat{n}(\omega) + k_+(1 - \langle n\rangle)\delta\tilde{c}(\mathbf{x_0},\omega) + \hat{\xi}_n(\omega). \tag{A.23}$$

Substituting our result for $\delta\tilde{c}(\mathbf{x}_0,\omega)$ from Eq (A.22), we find

$$-i\omega\delta\hat{n}(\omega) = -(k_+\bar{c} + k_-)\delta\hat{n}(\omega) + \tag{A.24}$$

$$+ \; k_+(1 - \langle n\rangle)\frac{i\omega\delta\hat{n}(\omega)}{\pi D a} + \hat{\xi}_n(\omega) \tag{A.25}$$

$$-i\omega\left[1 + \frac{k_+(1-\langle n\rangle)}{\pi D a}\right]\delta\hat{n}(\omega) = -(k_+\bar{c} + k_-)\delta\hat{n}(\omega) + \hat{\xi}_n(\omega) \tag{A.26}$$

$$\delta\hat{n}(\omega) = \frac{\hat{\xi}_n(\omega)}{-i\omega(1+\Sigma) + (k_+\bar{c} + k_-)} \tag{A.27}$$

where $\Sigma = k_+(1 - \langle n\rangle)/(\pi D a)$. The linearization of Eqs (A.9) and (A.10) takes the form

$$-i\omega\delta\hat{e}(\omega) = -\frac{1}{\tau_e}\delta\hat{e}(\omega) + R_e\delta\hat{n}(\omega) + \hat{\xi}_e(\omega) \tag{A.28}$$

$$-i\omega\delta\hat{g}(\omega) = -\frac{1}{\tau_g}\delta\hat{g}(\omega) + R_g\delta\hat{e}(\omega) + \hat{\xi}_g(\omega) \tag{A.29}$$

Each Langevin term is independent, and each frequency component ω is correlated only with the component at $-\omega$, defining the noise power spectrum $\langle\hat{\xi}_\mu(\omega)\hat{\xi}_\mu(-\omega')\rangle = 2\pi\delta(\omega - \omega')\mathcal{N}_\mu(\omega)$ for $\mu = n, e, g$. Solving the three linear equations, Eqs (A.27–A.29), we can find the power spectrum of the protein copy number fluctuations,

$$\mathcal{S}_g(\omega) = \frac{\mathcal{N}_g}{\omega^2 + 1/\tau_g^2} + R_g^2\frac{\mathcal{N}_e}{(\omega^2 + 1/\tau_g^2)(\omega^2 + 1/\tau_e^2)} + \tag{A.30}$$

$$+ R_g^2R_e^2\frac{\mathcal{N}_n}{(\omega^2 + 1/\tau_g^2)(\omega^2 + 1/\tau_e^2)[(1+\Sigma)^2\omega^2 + 1/\tau_c^2]}, \tag{A.31}$$

where $1/\tau_c = k_+\bar{c} + k_-$. This form has a very intuitive interpretation: each Langevin term represents a noise source; as this noise propagates from the point where it enters the dynamical system to the output, it is subjected both to gain of each successive stage (prefactors R), and to filtering by factors of $\mathcal{F}_\tau = (\omega^2 + 1/\tau^2)^{-1}$.

The total variance in protein copy number is given by an integral over the spectrum,

$$\langle(\delta g)^2\rangle \equiv \sigma_g^2 = \int\frac{d\omega}{2\pi}\mathcal{S}_g(\omega), \tag{A.32}$$

and the noise power spectra of the Langevin terms associated with the mRNA and protein dynamics have the simple forms $\mathcal{N}_e(\omega) = 2R_e\langle n\rangle$ and $\mathcal{N}_g(\omega) = 2R_g\langle e\rangle$, respectively. The spectrum $\mathcal{N}_n(\omega)$ is more subtle. One way to derive it is to realize that since there is only one binding site and this site is either occupied or empty, the total variance of δn must be given by the binomial formula,

$$\langle(\delta n)^2\rangle = \langle n\rangle(1 - \langle n\rangle). \tag{A.33}$$

Starting with Eq (A.27) and the analog of Eq (A.32), we can use this condition to set the magnitude of $\mathcal{N}_n$. Alternatively, we can use the fact that binding and unbinding come to equilibrium, and hence the fluctuations in n are a form of thermal noise, like Brownian motion or Johnson noise, and hence the spectrum $\mathcal{N}_n$ is determined by the fluctuation–dissipation theorem (Bialek and Setayeshgar, 2005). The result is that

$$\mathcal{N}_n = \frac{2}{\tau_c}(1+\Sigma)\langle n\rangle(1 - \langle n\rangle). \tag{A.34}$$

For simplicity we consider the case where the protein lifetime τ_g is long compared with all other time scales in the problem. Then we can approximate Eq (A.31) as

$$\mathcal{S}_g(\omega) \approx \frac{1}{\omega^2 + 1/\tau_g^2}\left[\mathcal{N}_g + (R_g\tau_e)^2\mathcal{N}_e + (R_g\tau_eR_e\tau_c)^2\mathcal{N}_n\right]. \tag{A.35}$$

Substituting the forms of the individual noise spectra $\mathcal{N}_\mu$ and doing the integral over ω [Eq (A.32)], we find the variance in protein copy number

$$\sigma_g^2 = \tau_g[R_g\langle e\rangle + (R_g\tau_e)^2R_e\langle n\rangle] + \frac{\tau_g}{\tau_c}(R_g\tau_eR_e\tau_c)^2(1+\Sigma)\langle n\rangle(1 - \langle n\rangle). \tag{A.36}$$

We notice that the first term in this equation is $R_g\tau_g\langle e\rangle$, which is just the mean number of proteins $\langle g\rangle$ from Eq (A.14). The second term

$$\tau_g(R_g\tau_e)^2R_e\langle n\rangle = R_g\tau_g(R_e\tau_e\langle n\rangle)(R_g\tau_e) \tag{A.37}$$

$$= R_g\tau_g\langle e\rangle(R_g\tau_e) \tag{A.38}$$

$$= R_g\tau_e\langle g\rangle. \tag{A.39}$$

Thus, the first two terms together contribute $(1 + R_g\tau_e)\langle g\rangle$ to the variance, and this corresponds to the output noise term in Eq (3.14).

The third term in Eq (A.36) contains the contribution of input noise to the variance in protein copy number. To simplify this term we note that the steady state of Eq (A.8) is equivalent to

$$k_+\bar{c}(1 - \langle n\rangle) = k_-\langle n\rangle. \tag{A.40}$$

Thus we can write

$$\frac{1}{\tau_c} \equiv k_+\bar{c} + k_- \tag{A.41}$$

$$= k_- \left[\frac{\langle n\rangle}{1 - \langle n\rangle} + 1\right] = \frac{k_-}{1 - \langle n\rangle}. \tag{A.42}$$

The term we are interested in is

$$\frac{\tau_g}{\tau_c}(R_g\tau_e R_e\tau_c)^2 \times (1 + \Sigma)\langle n\rangle(1 - \langle n\rangle) = (R_g\tau_g R_e\tau_e)^2\frac{\tau_c}{\tau_g}(1 + \Sigma)\langle n\rangle(1 - \langle n\rangle) \tag{A.43}$$

$$= g_0^2\frac{1}{k_-\tau_g}(1 + \Sigma)\langle n\rangle(1 - \langle n\rangle)^2 \tag{A.44}$$

$$= g_0^2\frac{\langle n\rangle(1 - \langle n\rangle)^2}{k_-\tau_g} + g_0^2\frac{1}{k_-\tau_g}\frac{k_+(1 - \langle n\rangle)}{\pi Da}\langle n\rangle(1 - \langle n\rangle)^2 \tag{A.45}$$

$$= g_0^2\frac{\langle n\rangle(1 - \langle n\rangle)^2}{k_-\tau_g} + g_0^2\frac{\langle n\rangle^2(1 - \langle n\rangle)^2}{\pi Da\bar{c}\tau_g}, \tag{A.46}$$

where in the last step we once again use Eq (A.40) to rewrite the ratio k_+/k_- in terms of $\langle n\rangle$. We recognize the two terms in this result as the switching and diffusion terms in Eq (3.14).

2. Cooperativity

To generalize this analysis of noise to cooperative interactions among transcription factors it is useful to think more intuitively about the two terms in Eq (A.46), corresponding to switching and diffusion noise. Consider first the switching noise.

We are looking at a binary variable n such that the number of proteins is $g_0\, n$. The total variance in n must be $\langle(\delta n)^2\rangle = \langle n\rangle(1 - \langle n\rangle)$ [Eq (A.33)]. This noise fluctuates on a time scale τ_c, so during the lifetime of the protein we see $N_s = \tau_g/\tau_c$ independent samples. The current protein concentration is effectively an average over these samples, so the effective variance is reduced to

$$\langle(\delta n)^2\rangle_{\rm eff} = \frac{1}{N_s}\langle n\rangle(1 - \langle n\rangle) = \frac{\tau_c}{\tau_g}\langle n\rangle(1 - \langle n\rangle). \tag{A.47}$$

Except for the factor of g_0 that converts n into g, this is the first term in Eq (A.46).

Now if h transcription factors bind cooperatively, we can still have two states, one in which transcription is possible and one in which it is blocked. For the case of activation, which we are considering here, the active state corresponds to all binding sites being filled, and so the rate at which the system leaves this state, k_-, shouldn't depend on the concentration of the transcription factors. The rate at which the system enters the active state does depend on concentration, but this doesn't matter, because with only two states we must always have an analog of Eq (A.40), which allows us to eliminate the "on rate" in

favor of k_- and $\langle n \rangle$. The conclusion is that the first term in Eq (A.46), corresponding to switching noise, is unchanged by cooperativity as long as the system is still well approximated as having just two states of transcriptional activity that depend on the potentially many more states of binding site occupancy.

For the diffusion noise term we use the ideas of Berg and Purcell (1977) and Bialek and Setayeshgar (2005, 2006). Diffusion noise should be thought of as an effective noise in the measurement of the concentration c, with a variance

$$\frac{\sigma_c^2}{\bar{c}^2} \sim \frac{1}{\pi D a \bar{c} \tau_g}, \tag{A.48}$$

where again we identify the protein lifetime as the time over which the system averages. For the system with a single binding site,

$$\langle n \rangle = \frac{\bar{c}}{\bar{c} + K_d}, \tag{A.49}$$

so that

$$\frac{\partial \langle n \rangle}{\partial c} = \frac{1}{\bar{c}} \langle n \rangle (1 - \langle n \rangle). \tag{A.50}$$

The noise in concentration, together with this sensitivity of n to changes in the concentration, should contribute a noise variance

$$\langle (\delta n)^2 \rangle_{\rm eff} = \left| \frac{\partial \langle n \rangle}{\partial c} \right|^2 \sigma_c^2 = \frac{\langle n \rangle^2 (1 - \langle n \rangle)^2}{\pi D a \bar{c} \tau_g}. \tag{A.51}$$

This is (up to the factor of g_0) the second term in Eq (A.46). Now the generalization to cooperative interactions is straightforward. If we have

$$\langle n \rangle = \frac{\bar{c}^h}{\bar{c}^h + K_d^h}, \tag{A.52}$$

then

$$\frac{\partial \langle n \rangle}{\partial c} = \frac{h}{\bar{c}} \langle n \rangle (1 - \langle n \rangle). \tag{A.53}$$

Since the effective noise in concentration is unchanged (Bialek and Setayeshgar, 2006), the only effect of cooperativity is to multiply the second term in Eq (A.46) by a factor of h^2.

Thus, in the expression [Eq (3.14)] for the variance of protein copy number, cooperativity has no effect on the switching noise but actually increases the diffusion noise by a factor of h^2. When written as a function of the mean copy number and the transcription factor concentration, this leaves the functional form of the variance fixed, only changing the coefficients. The overall effect it to make the contribution of diffusion noise more important. One way to say this is that, when we refer the noise in copy number back to the input, cooperativity causes the equivalent concentration noise to become closer to the limit Eq (A.48) set by diffusive shot noise (Bialek and Setayeshgar, 2006).

Gregor et al. (2006a) also consider the possibility that noise is reduced by averaging among neighboring nuclei. This does not change the form of any of the noise terms, but does change the microscopic interpretation of the coefficients α and β. For example, averaging for a time τ_g over N nuclei is equivalent to having one nucleus with an averaging time $N\tau_g$.

A.4 Information flow in transcriptional regulation

A.4.1 Finding optimal channel capacities

If we treat the kernel $p(g|c)$ on a discrete (c, g) grid we can easily choose $p(c)$ so as to maximize the mutual information $\mathrm{I}(c; g)$ between the TF concentration and the expression level. The problem can be stated in terms of the following variational principle:

$$\mathcal{L}[p(c)] = \sum_{c,g} p(g|c)p(c) \log_2 \frac{p(g|c)}{p(g)} - \Lambda \sum_c p(c) \tag{A.54}$$

where the multiplier Λ enforces normalization of $p(c)$, and $p(g)$ is a function of the unknown distribution, since $p(g) = \sum_c p(g|c)p(c)$. The solution $p^*(c)$ of this problem achieves the maximum capacity $\mathrm{I}(c; g)$ of the channel.

The original idea behind the Blahut-Arimoto approach (Blahut, 1972) for finding optimal $p^*(c)$ was to understand that the maximization of Eq (A.54) using variational objects $p(c_i)$ is equivalent to the following maximization:

$$\max_{p(c)} \mathcal{L}[p(c)] \sim \max_{p(c)} \max_{p(c|g)} \mathcal{L}'[p(c), p(c|g)], \tag{A.55}$$

where

$$\mathcal{L}'[p(c), p(c|g)] = \sum_{g,c} p(c)p(g|c) \log \frac{p(c|g)}{p(c)} - \Lambda \sum_c p(c). \tag{A.56}$$

In words, finding the extremum in variational object $p(c)$ is equivalent to a double maximization of a modified Lagrangean, where both $p(c)$ and $p(c|g)$ are treated as independent variational objects. The extremum of the modified Lagrangean is achieved exactly when the consistency condition $p(c|g) = \frac{p(g|c)p(c)}{\sum_c p(g|c)p(c)}$ holds. This allows us to make an iterative algorithm, where Eq (A.56) is first solved for $p(c)$ given some guess for $p(c|g)$:

$$p(c) = \frac{1}{Z} \exp \left\{ \sum_g p(g|c) \log p(c|g) \right\}, \tag{A.57}$$

and the guess is then updated with the new $p(c)$.

Let us suppose that each input signal c carries some metabolic or time cost to the cell. Then we can introduce a cost vector $v(c)$ that assigns a cost to each codeword c, and require of the solution the following:

$$\sum_c p(c)v(c) \leq C_0 \tag{A.58}$$

where C_0 is the maximum allowed total expense. The constraint can be introduced into the functional [Eq (A.54)] through appropriate Lagrange multiplier; the same approach can be extended to introduce the cost of coding for the output words, $\sum_g \sum_c p(g|c)p(c)\tilde{v}(g)$, because it reduces to an additional "effective" cost for the input, $v(c) = \sum_g p(g|c)\tilde{v}(g)$.

We demonstrate next how to include a smoothness constraint into the functional; the degree of "smoothness" of the resulting input distribution $p(c^*)$ will then be controllable through an additional Lagrange multiplier, and both ways of computing the capacity explained in the main text – that of referring the limited input resolution $\sigma_c(\bar{c})$ to the noise in the output, and that of including it as a smoothness constraint on the input distribution –

will be possible within a single framework. For the capacities in the paper we assume that all relevant noise has already been included in the explicit output noise model.

By analogy to the field theories in which kinetic energy terms of the form $\int |\nabla f(x)|^2 \, dx$ constrain the gradient, we write the following functional:

$$\mathcal{L}[p(c)] = \mathrm{I}(c;g) - \Lambda_0 \sum_c p(c) - \tag{A.59}$$

$$- \Phi_1 \sum_c p(c) v_1(c) - \Phi_2 \sum_g p(g) v_2(g) - \tag{A.60}$$

$$- \Theta \sum_c \left(\frac{\Delta p}{\Delta c} \sigma(c) \right)^2 . \tag{A.61}$$

Eq (A.59) maximizes the capacity with respect to variational objects $p(c)$ while keeping the distribution normalized by Lagrange multiplier Λ_0; Eq (A.60) imposes cost $v_1(c)$ on input symbols and cost $v_2(g)$ on output symbols; finally, Eq (A.61) limits the derivative of the resulting solution. The difference operator Δ is defined for an arbitrary function $f(c)$:

$$\Delta f(c) = f(c_{i+1}) - f(c_i). \tag{A.62}$$

$\sigma(c)$ assigns different weights to different ranges of input c; there is arbitrariness in the selection of scale Θ, as it can be absorbed into $\sigma(c)$. This construction can be seen as placing limits on the resolution of the input, in the following way. If the input cannot be precisely controlled, but has an uncertainty of $\sigma(c)$ at mean input level c, we require that the optimal probability distribution must not change much as the input fluctuates on the scale $\sigma_c(c)$, or in other words,

$$|\delta p| = \left| \frac{\Delta p}{\Delta c} \sigma(c) \right| \ll 1. \tag{A.63}$$

The term in Eq (A.61) constrained by Lagrange multiplier Θ can be seen as the sum of squares of such variations for all possible values of the input.

By differentiating the functional with respect to $p(c_i)$ we get the following equation:

$$0 = \sum_g p(g|c_i) \log p(c_i|g) - \log p(c_i) - \Lambda - \Phi_1 v_1(c_i) - \Phi_2 \sum_g p(g|c_i) v_2(g) + \tag{A.64}$$

$$+ \Theta \left\{ [p(c_{i+1}) - p(c_i)] \frac{\sigma^2(c_i)}{(c_{i+1} - c_i)^2} - [p(c_i) - p(c_{i-1})] \frac{\sigma^2(c_{i-1})}{(c_i - c_{i-1})^2} \right\}. \tag{A.65}$$

Let us denote by $F(c, p(c)) = \Delta \frac{\Delta p}{(\Delta c)^2} \sigma(c)^2$ the term in braces. The solution for $p(c)$ is given by:

$$p(c) = \frac{1}{Z} \exp \left\{ \sum_g p(g|c) \log p(c|g) - \Phi_1 v_1(c) - \Phi_2 \sum_g p(g|c) v_2(g) + \Theta F \right\} \tag{A.66}$$

We can now continue to use the Blahut-Arimoto trick of pretending that $p(c|g)$ is an independent variational object, and that $p(c)$ has to be solved with $p(c|g)$ held fixed; however, even in that case, Eq (A.66) is an implicit equation for $p(c)$ which needs to be solved by

numerical means. The complete iterative prescription is therefore as follows:

$$p^n(g) = \sum_c p(g|c)p^n(c) \tag{A.67}$$

$$p^n(c|g) = \frac{p(g|c)p^n(c)}{p^n(g)} \tag{A.68}$$

$$p^{n+1}(c) = \frac{1}{Z}\exp\left\{\sum_g p(g|c)\log p^n(c|g) - \Phi_1 v_1(c) - \right. \tag{A.69}$$

$$\left. - \Phi_2 \sum_g p(g|c)v_2(g) + \Theta F(c, p^{n+1}(c))\right\} \tag{A.70}$$

To reiterate, Eq (A.70) has to be solved on its own by numerical means as the variational objects for iteration $(n+1)$ appear both on its left- and right-hand side. The input and output costs of coding are neglected if one sets $\Phi_1 = \Phi_2 = 0$; likewise, smoothness constraint is ignored for $\Theta = 0$, in which case Eq (A.70) is the same as in the original Blahut-Arimoto derivation and gives the value of $p^{n+1}(c)$ explicitly.

If we take a fixed input-output relation $P(g|c)$ and vary Lagrange multiplier Θ, we trace out the so-called rate distortion (RD) curve, shown in Fig A.12; this curve is parametrized by Θ and plots the achievable channel capacity as the function of the smoothness, which is the term conjugate to multiplier Θ in Eq (A.61). Solutions that are forced to be smoother have a smaller capacity, and in the extreme case one would obtain the capacity for distribution that is uniform in the input. Figure A.12 (left) shows that smoothing will remove features in the low concentration region, where the gene expression is turned off regardless of precise shape of the distribution, at practically no cost in capacity;[5] in contrast, the shape of the input near $c/K_d \approx 1$ is important and is preserved if the smoothing is not too strong. Figure A.12 (right) shows a case of very high cooperativity. It is clear that as $h \to \infty$, the input-output relation becomes a step function, and as such would be able to transmit at most 1 bit of information. For h large but not infinite, the system has to move in a controlled manner from zero occupancy to full occupancy within a very small concentration window; the optimal distribution must therefore be steep there, which is in direct contradiction with the imposed smoothing. In particular, the optimal distribution is "tuned" to the input-output relation, and if (in case of extreme smoothing) the uniform input distribution were used (black dashed line), the capacity would fall considerably from its unconstrained maximum [Fig A.12].

A.4.2 Channel capacity at constrained expense

Here we take a look at how the optimization principle can be used to make non-trivial predictions about the regulatory element, especially when metabolic cost of coding is a real concern. Suppose we take an input-output activator kernel with variable cooperativity (parametrized by Hill coefficient h), that has the noise parameters of the blue system of Fig 4.3b. We assume explicitly that the diffusive noise strength scales with h^2 as expected from theoretical considerations, see Table 4.1. Conceptually, increasing the cooperativity increases the sensitivity to fluctuations in the input (especially at half-maximal induction),

[5] This is a consequence of the degeneracy of the input/output kernel. For high cooperativity, the input/output relation is flat in the regions $c \ll K_d$ and $c \gg K_d$, and the local shape of the $p(c)$ there must be irrelevant for the channel capacity. It is not irrelevant for the smoothness penalty constraint, however.

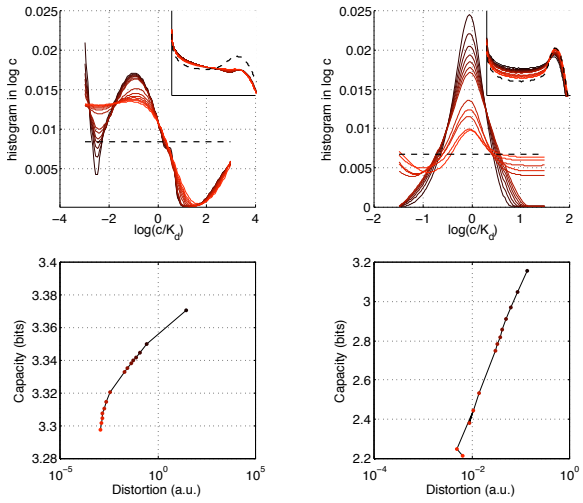


Figure A.12: Imposing smoothness constraint on the solutions. Left panel shows the optimal input concentrations of the input-output relation $P(g|c)$ for an activator system with $h = 2$ cooperativity whose output noise characteristics mimic that of the red system presented in Fig 4.3c. As the Lagrange multiplier conjugate to the smoothness term is increased, the solutions lose sharp features (and the color in the plot changes from black to red). The corresponding output distributions are plotted in the inset, and the rate-distortion curve is shown below the distributions. We see that the sharp features of $p^*(c)$ in the $c \ll K_d$ region are smoothed out with no appreciable decrease in the channel capacity because the input/output kernel is flat (degenerate) there. Right panel shows similar computations carried out for a highly cooperative $h = 15$ kernel of with the same noise characteristics as the blue system of Fig 4.3b. Here the decrease in capacity is much greater, because the input is forced to be precisely tuned to the extremely sharp input/output relation at $c = K_d$ on one hand, but also constrained to have small derivative in $p(c)$ on the other.

and thus reduces the channel capacity; on the other hand, increased cooperativity allows the system to explore the full dynamic range of promoter occupancy at a smaller cost in the input. Consequently there should be an optimal cooperativity $h^*(C_0)$ that depends on the total expense for coding inputs and outputs. This kind of tradeoff is illustrated in Fig A.13.

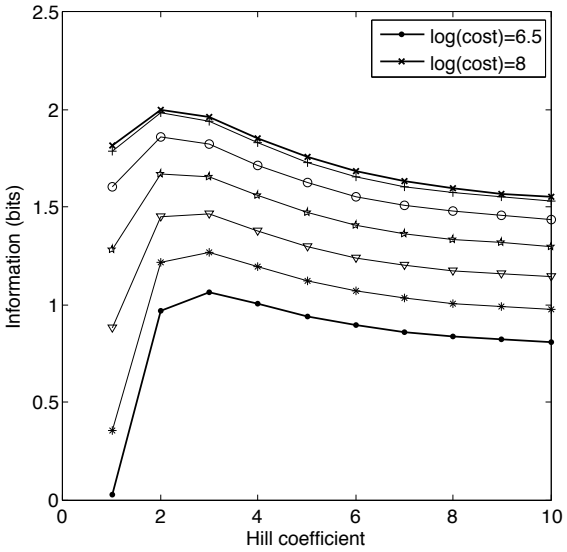


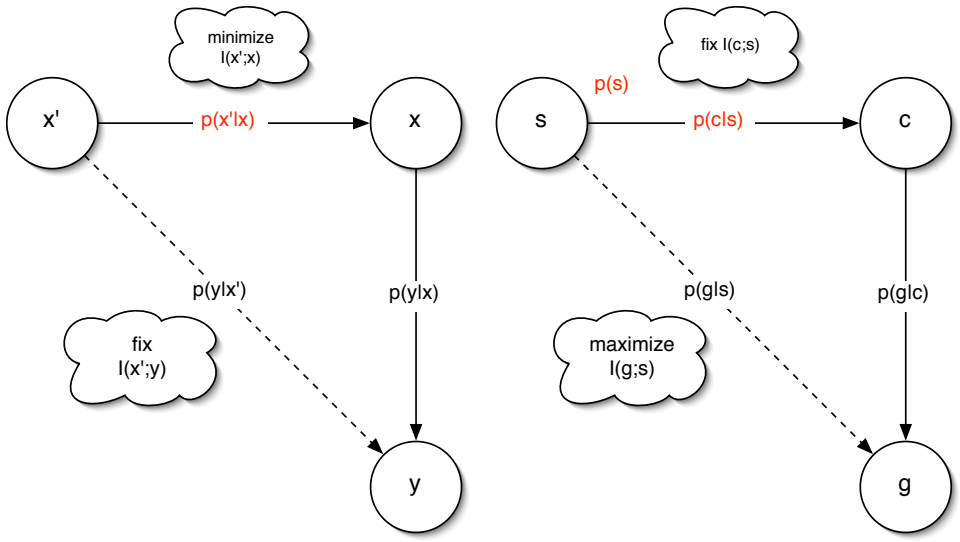
Figure A.13: Information capacity of the blue system of Fig 4.3b in which cooperativity (horizontal axis) is being changed. Different curves correspond to different allowed values of the total expense, Eq (4.22), for coding inputs and outputs (the logarithm of the allowed cost increases by 0.25 for each curve between thick black curves indicated in the legend). Qualitatively similar results are obtained as the ratio of costs of input and output are changed (i.e. the peak at low $h > 1$). Note that although the blue system is similar to the *Drosophila* Bicoid/Hunchback system studied in the next section, we do not recover as optimal the observed Hill coefficient $h = 5$, presumably because we do not have the correct expense model or because the element in the fruit fly does not operate in isolation.

A.4.3 Capacity as the number of distinguishable discrete states

Here we discuss if it is possible to use constructively the interpretation of information capacity $I(c; g)$ between the input c and output g, namely that there are $2^{I(c;g)}$ levels in the input that have distinguishable outputs. In other words, despite the fact that both input and output are continuous variables in our problem, we are looking to find the smallest set of discrete c_i and the related optimal discrete distribution $p(c_i)$, such that the channel capacity remains undegraded compared to its maximal value when the whole continuous range in c is used. This can also be seen as compressing the continuous range of input concentrations onto a discrete subset of concentrations.

To tackle this task, we introduce the notion of abstract signal s, i.e. a quantity that describes some abstract form of control over the concentration c, as is shown in Fig A.14b. The new "handle" s defines a Markov Chain $s \to c \to g$. As before, we want to maximize the information between the ultimate input, s, and the final output, g; additionally, we will now also limit the amount of information that can flow between s and c. This compound channel will be called the information bottleneck (IB) channel.

The situation is similar, but not equal to the information bottleneck problem. There, we look for a probabilistic mapping of a true signal x into the compressed form x', such that x' retains a fixed amount of information with y, as shown in Fig A.14a. Note that



A.14a: Information bottleneck problem. A.14b: Information bottleneck channel.

Figure A.14: Left: information bottleneck problem. We are given the joint distribution $p(x, y)$, and would like to compress signal x into x', so that the mutual information $I(x'; x)$ is minimized. The compressed signal x' has to retain a fixed amount of information about the relevant feature y. Right: information bottleneck channel. Diagram of the explicit influence of signal control s over c, which couples to the expression level through known $p(g|c)$. The unknown quantities to be optimized for are in red (the distribution of signals and the conditional $p(c|s)$). We are maximizing the information flow through the system at fixed information flow through the first segment.

there is more variational freedom in the new IB channel formulation than in the classical IB problem, because we also have to compute the optimal signal distribution $p(s)$.

We expect that as the constraint between the signal s and concentration c is increased, c will become more and more "compressed" and will start to take non-zero values only on smaller and smaller support until only a few discrete states in c remain. If this can happen without significant loss of total capacity, $I(s;g)$, we will have found the discrete states that we refer to when explaining the concept of mutual information. Clearly, if the constraint is increased still further, the total capacity will become degraded. How this happens is an interesting question in itself, because the IB compound channel can be seen as a problem of optimally coupling together two channels, i.e. $I(s;c)$ and $I(c;g)$; we will not pursue this – more general – issue further.

For the IB channel the variational problem can be stated as follows:

$$\begin{aligned}
\mathcal{L}[p(s),p(c|s)] &= I(s;g) - \beta I(s;c) - \Phi_1 \sum_c p(c)v(c) - && \text{(A.71)}\\
&- \sum_{c,s} \lambda_s p(c|s) - \lambda_0 \sum_s p(s) && \text{(A.72)}\\
I(s;g) &= \sum_{s,g} p(g|s)p(s)\log\frac{p(g|s)}{p(g)}, && \text{(A.73)}\\
I(s;c) &= \sum_{s,c} p(c|s)p(s)\log\frac{p(c|s)}{p(c)}. && \text{(A.74)}
\end{aligned}$$

Constraint β is conjugate to the constrained mutual information between s and c; constraint Φ_1 is conjugate to the total expense $\mathcal{C}$, and there are $|s|+1$ normalization constraints $\{\lambda_0,\lambda_s\}$. Furthermore, $p(g|s)=\sum_c p(g|c)p(c|s)$, and $p(c)$ and $p(g)$ are implicit functions of both $p(s)$ and $p(c|s)$ through marginalization.

Let us first evaluate $\frac{\delta\mathcal{L}}{\delta p(c|s)}$. The first term can be computed as follows:

$$\begin{aligned}
\frac{\delta I(s;g)}{\delta p(c|s)} &= \sum_g p(g|c)p(s)\log\frac{p(g|s)}{p(g)} + && \text{(A.75)}\\
&+ \sum_{s',c',g} p(g|c')p(c'|s')p(s')\frac{p(g)}{\sum_{c''}p(g|c'')p(c''|s')}\left\{\frac{p(g|c)}{p(g)}\delta_{ss'} - \right. && \text{(A.76)}\\
&- \left.\frac{p(g|s')}{p^2(g)}p(g|c)p(s)\right\} && \text{(A.77)}\\
&= \sum_g p(g|c)p(s)\log\frac{p(g|s)}{p(g)}. && \text{(A.78)}
\end{aligned}$$

The contributions in curly braces that originate by taking derivative under the logarithm cancel out. Similarly,

$$\frac{\delta I(s;c)}{\delta p(c|s)} = p(s)\log\frac{p(c|s)}{p(c)}. \qquad \text{(A.79)}$$

Inserting both results into Eq (A.72) gives us the first set of equations:

$$\sum_g p(g|c)p(s)\log\frac{p(g|s)}{p(g)} - \beta p(s)\log\frac{p(c|s)}{p(c)} - \lambda_s - \Phi_1 p(s)v(c) = 0. \qquad \text{(A.80)}$$

Now let's take the derivatives with respect to $p(s)$: $\frac{\delta \mathcal{L}}{\delta p(s)} = 0$:

$$\frac{\delta I(s;g)}{\delta p(s)} = \sum_g p(g|s) \log \frac{p(g|s)}{p(g)} - \tag{A.81}$$

$$- \sum_{g,s'} p(g|s')p(s') \frac{p(g)}{p(g|s')} \frac{p(g|s')}{p^2(g)} \sum_{c'} p(g|c')p(c'|s) \tag{A.82}$$

$$= \sum_g p(g|s) \log \frac{p(g|s)}{p(g)} - 1. \tag{A.83}$$

Similar calculation yields the result for the derivative of $I(s;c)$. Absorbing constants into λ_0, we can write down the next set of equations:

$$\sum_g p(g|s) \log \frac{p(g|s)}{p(g)} - \beta \sum_c p(c|s) \log \frac{p(c|s)}{p(c)} - \Phi_1 \sum_c p(c|s)v(c) = \Lambda_0. \tag{A.84}$$

In total, Eq (A.80) and Eq (A.84) define $|s| + |s| \cdot |c|$ equations; in addition, we have the marginalization and normalization consistency conditions.

We can now generate an iterative scheme in a way similar to the modified Blahut-Arimoto algorithm presented above:[6]

$$p^n(s|g) = \frac{p^n(g|s)p^n(s)}{p^n(g)}, \tag{A.85}$$

$$p^n(s|c) = \frac{p^n(c|s)p^n(s)}{p^n(c)}. \tag{A.86}$$

Finally, this gives the complete iterative scheme for new value $p^{n+1}(s)$:

$$p^{n+1}(s) = \frac{p^n(s)}{Z} \exp \left\{ \frac{1}{1-\beta} \sum_g p^n(g|s) \log \frac{p^n(g|s)}{p^n(g)} \right. \tag{A.87}$$

$$\left. - \frac{\beta}{1-\beta} \sum_c p^n(c|s) \log \frac{p^n(c|s)}{p^n(c)} - \frac{\Phi_1}{1-\beta} \sum_c p^n(c|s)v(c) \right\}. \tag{A.88}$$

From Eq (A.80) we can similarly express the other unknown $p^{n+1}(c|s)$, as follows:

$$p^{n+1}(c|s) = \frac{p^n(c)}{Z} \exp \left\{ \frac{1}{\beta} \sum_g p^n(g|c) \log \frac{p^n(g|s)}{p^n(g)} - \frac{\Phi_1}{\beta} v(c) \right\}. \tag{A.89}$$

Eqs (A.88, A.89), together with consistency equations form the iterative scheme. The scheme needs both initial conditions, $p^0(s)$ and $p^0(c|s)$, and the values of the Lagrange constraints β and Φ_1 in order to work, and is carried out until desired convergence is achieved in variational objects. At given pair (β, Φ_1), the resulting distributions are constrained by $(I(s;c), \mathcal{C}(c))$, where $\mathcal{C}$ is the expense of the distribution $p(c)$. The rate-distortion curve in our case is a plot of $I(s;g)$ against $I(s;c)$.

As a simple example consider the binary symmetric channel, which is fully characterized by the "switching" parameter p. The transition matrix for this channel can be written as:

$$p(g|c) = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}.$$

[6] We give no proof of convergence for these iterative equations.

We use this channel but constrain the input c by modifying it only through s. Figure A.15 shows how the total capacity through the IB channel depends on the capacity of the first, bottleneck, segment.

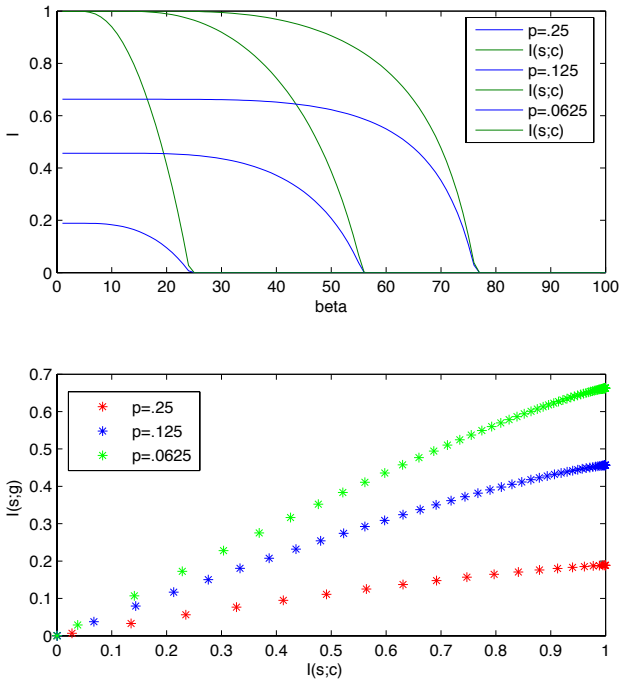
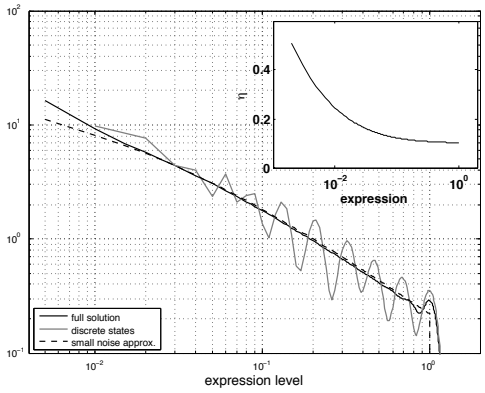
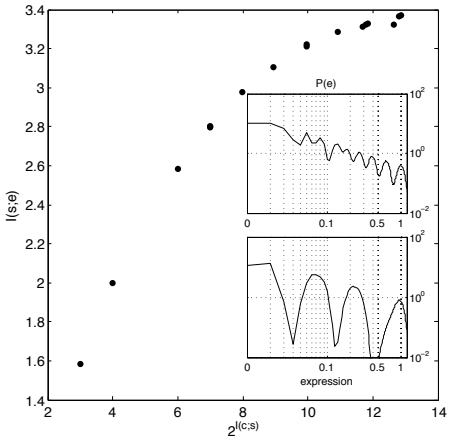


Figure A.15: Upper panel: as constraint β varies between 0 and 1, the constrained capacity $I(s;c)$ is plotted in green for three decreasing values of p, and the total channel capacity is plotted in blue. Lower panel: full channel capacity vs first segment capacity for three p values; this is the rate-distortion curve for the upper plot. The algorithm has been executed with 3 different values for p, $p = \{0.25, 0.125, 0.0625\}$, with 10 iterations for each p value, zero cost function, cardinality of set of s equal to 10, stopping condition equal to step-by-step change below 10^{-7} in variational objects, and random initialization. When the constraint is inactive (the first segment can transmit maximum 1 bit of capacity), the resulting total channel capacities are equal to the full capacities of the symmetric binary channel, i.e $I_{max}(p) = 1 + p\log_2 p + (1-p)\log_2(1-p)$, which gives values $\{0.1887, 0.4564, 0.6627\}$ for different p, consistent with numerical results (the points where the RD curve intersects the $I(s;c) = 1$ bit line).

For a more realistic application of the information bottleneck channel we analyze the information transmission through a *lac*-like repressor element with the noise parameters measured by Elowitz et al. (2002). Figure A.16a shows the exact numerical solution and the small-noise approximation for the optimal distribution of output gene expression levels, along with the distribution of outputs computed with IB channel, where the compression in the first segment of the IB channel was used to reduce the support of the input to just a few discrete TF concentrations. The resulting output distributions has distinguishable peaks. Figure A.16b shows the rate-distortion curve and two sample output distributions when the IB channel constraint is increased and only $I(s;c)$ bits are allowed to flow through the first segment.



A.16a: Optimal output distributions.



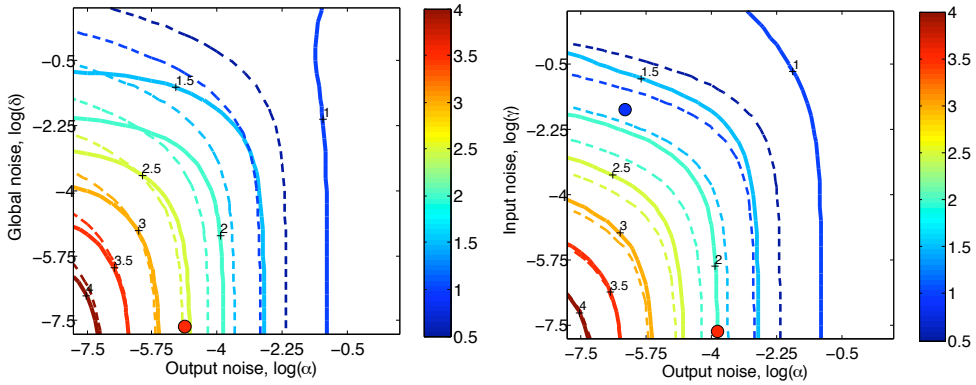
A.16b: Discrete states.

Figure A.16: Left: The optimal distribution of expression levels, $p(g)$, black line, for the system with *lac* noise parameters as measured by Elowitz et al. (2002); the fractional noise is shown in the inset. Note the solution in gray, computed using the information bottleneck channel. Here we constrain the size of the support, $|s|$ just enough such that the total capacity is not yet degraded compared to the unconstrained solution, and we see the emergence of $\sim 9-10$ distinguishable peaks in the output. Right: as the capacity of the first segment of the IB channel, $I(s;c)$, is varied by tuning the Lagrange multiplier β of Eq (A.72), we trace out the rate-distortion curve. $2^{I(s;c)}$ is shown on the horizontal axis; this is the number of discrete states in the input. The total IB channel capacity is plotted on the vertical axis. Two insets show output probability distributions for two choices of β: in the top inset there are about nine distinguishable levels of the input, while in the bottom we only see four (consequently the total capacity of the channel is 2 bits).

A.4.4 Solutions with alternative input-output relations

In this section we present channel capacities for two alternative noise models. We use activator kernels with no cooperativity and consider the models with output and switching input noise [Fig A.17b] or a model with output and global noise [Fig A.17a]. In contrast to the input diffusion noise, discussed in Section 4.4, changing cooperativity does not affect the shapes of these kernels, and the sole effect of making the regulatory element cooperative is to achieve a higher dynamic range in occupancy at the same cost for input transcription factor concentrations.

A comparison with the results for the diffusive input noise in Section 4.4 shows that, at the same numerical noise strength (e.g. $\beta = \gamma$ or $\beta = \delta$), the diffusive noise is less limiting than the alternative two models, at least at $h = 1$. It should also be noted that as cooperativity is increased, the diffusive noise strength grows in proportion to h^2, while γ and δ do not change.



A.17a: Output and global noise.

A.17b: Output and switching noise.

Figure A.17: Left: Information contours (color scale in bits, thick lines – full solution, dashed lines – small noise approximation) for noise models where varying amounts of output noise (horizontal axis) and global noise (vertical axis) are present. See Table 4.1 for the definition of α. The global noise δ adds a term $\sigma_g^2 = \delta \cdot g^2$ to the total noise. Right: the same plot for models with output noise (horizontal axis) and input switching noise (vertical axis). For illustration purposes, the dots have been replotted here in the same position as in Fig 4.3.

A.4.5 Validity of Langevin approximations

Langevin approximation assumes that the fluctuations of a quantity around its mean are Gaussian and proceeds to calculate their variance. For the calculation of exact channel capacity we must calculate the exact input-output relation, $P(g|c)$. Even if Langevin approach ends up giving the correct variance as the function of the input, $\sigma_g(c)$, the shape of the distribution might be far from Gaussian. We expect such a failure when the number of mRNA is very small: the distribution of expression levels might be then multi-peaked, with peaks corresponding to $b, 2b, 3b, \ldots$ proteins, where b is the burst size (number of proteins produced per transcript lifetime).

In the model used in Eq (4.18), parameter $\alpha = (1 + b)/g_0$ determines the output noise; $g_0 = b\bar{m}$, where $\bar{m}$ is the average number of transcripts produced during the integrating

time (i.e. the longest averaging timescale in the problem, for example the protein lifetime or cell doubling time). If $b \gg 1$, then the output noise α is effectively determined only by the number of transcripts, $\alpha \approx 1/\bar{m}$. We should therefore be particularly concerned what happens as $\bar{m}$ gets small.

Our plan is therefore to solve for $P(g|c)$ exactly by finding the stationary solution of the Master equation in the case where the noise consists of the output and switching input contributions. In this approach, we explicitly treat the fact that the number of transcribed messages, designated by m, is discrete. We start by calculating $P_i(m|c,t)$. The state of the system is described index i, which can be 0 or 1, depending on whether the promoter is bound by the activator or not, respectively. Normalization requires that for each value of c:

$$\sum_{i=0,1} \sum_{m} P_i(m|c,t) = 1. \tag{A.90}$$

The time evolution of the system is described by the following set of equations for an activator:

$$\begin{aligned}
\frac{\partial P_0(m|c,t)}{\partial t} &= R\left(P_0(m-1|c,t) - P_0(m|c,t)\right) \\
&- \frac{1}{\tau}\left(mP_0(m|c,t) - (m+1)P_0(m+1|c,t)\right) \\
&- k_- P_0(m|c,t) + k_+ c P_1(m|c,t), \qquad \text{(A.91)} \\
\frac{\partial P_1(m|c,t)}{\partial t} &= -\frac{1}{\tau}\left(mP_1(m|c,t) - (m+1)P_1(m+1|c,t)\right) \\
&+ k_- P_0(m|c,t) - k_+ c P_1(m|c,t), \qquad \text{(A.92)}
\end{aligned}$$

where τ is the integrating time, k_- is the rate for switching into the inactive state (off-rate of the activator), k_+ is the second-order on-rate, and R_e is the rate of mRNA synthesis. These constants combine to give $\bar{m} = R_e\tau$ and the input switching noise strength $\gamma = (k_-\tau)^{-1}$, see Table 4.1. This set of equations is supplemented by appropriate boundary conditions for $m = 0$. To find the steady state distribution $P(m|c, t \to \infty) = P(m|c)$, we set the left-hand side to zero and rewrite the set of equations (with high enough cutoff value of $m_{\max}$) in matrix form:

$$\mathbf{M}(c)\mathbf{p}(c) = \mathbf{b}$$

where $\mathbf{p} = (P_0(0|c), P_1(0|c), P_0(1|c), P_1(1|c), \cdots)$ and $\mathbf{b} = (0, 0, \cdots, 0, 1)$. Matrix $\mathbf{M}$ (of dimension $2(m_{\max}+1)+1$ rows and $2(m_{\max}+1)$ columns) contains, in its last row, only ones, which enforces normalization. The resulting system is a non-singular band-diagonal system that can be easily inverted. The input-output relation for the number of messages is given by taking $P(m|c) = P_0(m|c) + P_1(m|c)$.

Having found the distribution for the number of transcripts we then convolve it another Poisson process, $P(g|\langle g\rangle = b\,m)$, i.e. $P(g|c) = \sum_m P(m|c)P(g|\langle g\rangle = b\,m)$. Finally, the result is rediscretized such that mean expression $\bar{g}$ runs from 0 to 1.

Note that the Langevin approximation only depends on the combination of the burst size b and the mean number of transcripts $\bar{m}$ through α; in contrast, the Master equation solution depends on both b and $\bar{m}$ independently. The generalization of this calculation to repressors or Hill-coefficient-type cooperativity is straightforward.

Figure A.18c shows that the Langevin approximation yields correct second moments of the output distribution; however, Gaussian distributions themselves are, for large burst sizes and small number of messages, inconsistent with the exact solutions, as can be seen

in Fig A.18a. In the opposite limit where the number of messages is increased and burst size kept small [Fig A.18b], normal distributions are an excellent approximation. Despite these difficulties the information capacity calculated with either Gaussian or Master input-output relations differs by at most 12% over a large range of burst sizes b and values for α, illustrated by Fig A.18d.

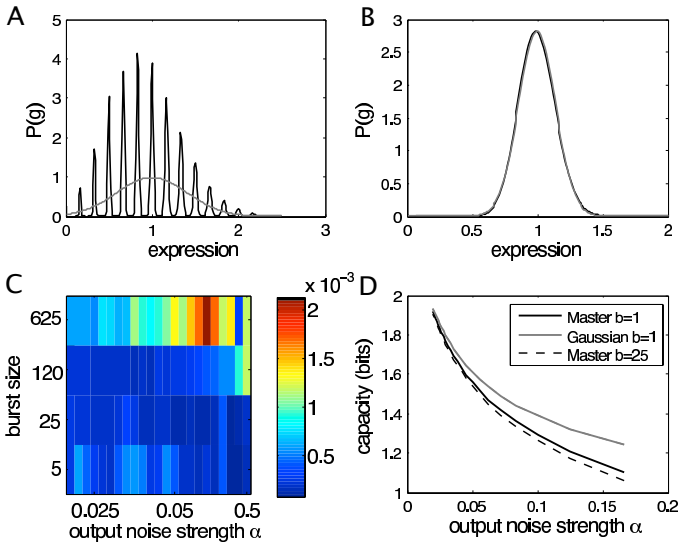


Figure A.18: Exact solutions (black) for input-output relations, $P(g|c)$, compared to their Gaussian approximations (gray). Panel A shows the distribution of outputs at maximal induction, $P(g|c_{\max})$ for a system with a large burst size, $b = 5^4$ and a large output noise $\alpha = \frac{1}{6}$ (i.e. the average number of messages is 6, as is evident from the number of peaks, each of which corresponds to a burst of translation at different number of messages). Panel B shows the same distribution for smaller output noise, $b = 5^2$ and $\alpha = \frac{1}{50}$; here Gaussian approximation performs well. Both cases are computed with switching noise parameter $\gamma = \frac{1}{50}$, and cooperativity of $h = 2$. Panel C shows in color-code the error made in computing the standard deviation of the output given c; the error measure we use is the maximum difference between the exact and Gaussian results over the full range of concentrations: $\max_c \text{abs}\,[\sigma_g(c)_{\text{Master}} - \sigma_g(c)_{\text{Gaussian}}]$. As expected the error decreases with decreasing output noise. Panel D shows that the capacity is overestimated by using an approximate kernel, but the error again decreases with decreasing noise as Langevin becomes an increasingly good approximation to the true distribution. In the worst case the approximation is about 12% off. Gaussian computation only depends on α and not separately on burst size, so we plot only one curve for $b = 1$.

A.4.6 Fine-tuning of optimal distributions

To examine the sensitivity to the perturbations in the optimal input distributions for Fig 4.6 we need to generate an ensemble of perturbations. We pick an *ad hoc* prescription, whereby the optimal solution is taken, and we add to it 5 lowest sine / cosyne waves on the input domain, each with a weight that is uniformly distributed on some range. The range determines whether the perturbation is small or not. The resulting distribution is clipped to be positive and renormalized. This choice was made to induce low-frequency perturbations (high frequency perturbations mostly just average out because the kernel is smooth). Then, for an ensemble of 100 such perturbations, $p_i(c)$, $i = 1, \ldots, 100$, and for every system of the information plane in Fig 4.3, the divergence of the perturbed input distribution to the true solution, $d_i = D_{\text{JS}}(p_i(c), p^*(c))$, is computed, as well as the channel capacity, $I_i =$

$I[p(g|c), p_i(c)]$. Figure A.19 plots the (d_i, I_i) scatter plots for 3×3 representative systems with varying amounts of output (β) and input (α) noise, taken from Fig 4.3 uniformly along the horizontal and vertical axes.

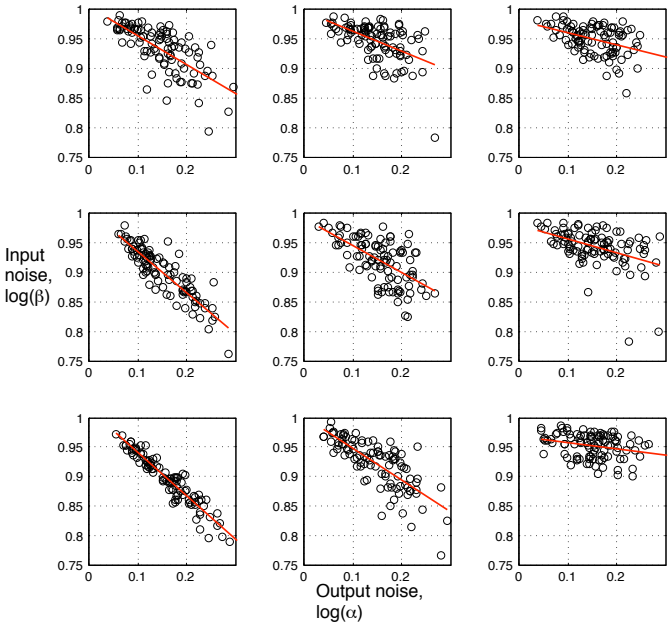


Figure A.19: Robustness of the optimal solutions to perturbations in the input distribution. Activator systems with no cooperativity are plotted; their parameters are taken from an uniformly spaced, 3×3 grid of points in the information plane of Fig 4.3, such that the output noise increases along the horizontal edge of the figure and the input noise along the vertical edge. Each subplot shows a scatter plot of 100 perturbations from the ideal solution; the Jensen-Shannon distance from the optimal solution, d_i, is plotted on the horizontal axis and the channel capacity (normalized to maximum when there is no perturbation), $I_i/I_{\max}$, on the vertical axis. Red lines are best linear fits.

Figure A.19 shows that as we move towards systems with higher capacity (lower left corner), perturbations to the optimal solution that are at the same distance from the optimum as in the low capacity systems (upper right corner), will cause greater relative loss (and therefore an even greater absolute loss) in capacity. As expected, higher capacity systems must be better tuned, but even for the highest capacity system considered, a perturbation of around $d_{\mathrm{JS}} \approx 0.2$ will only cause an average 15% loss in capacity.

Bibliography

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell, Fourth Edition*. Garland Science, New York.

Alon, U. (2003). Biological networks: The tinkerer as an engineer. *Science*, 301:1866.

Amit, D. J. (1999). *Modeling Brain Function: the World of Attractor Neural Networks*. Cambridge University Press, Cambridge.

Atick, J. J. and Redlich, A. N. (1990). Towards a theory of eary visual processing. *Neural Comp*, 2:208.

Bakk, A. and Metzler, R. (2004). Nonspecific binding of the OR repressors CI and Cro of bacteriophage lambda. *J Theor Biol*, 231:525.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nature*, 38:636.

Barlow, H. B. (1961). *Possible principles underlying the transformation of sensory messages in Sensory Communication*. MIT Press, Cambridge.

Barlow, H. B. (1981). Critical limiting factors in the design of the eye and visual cortex. *Proc R Soc Lond Ser B*, 212:1.

Barlow, H. B., Levick, W. R., and Yoon, M. (1971). Responses to single quanta of light in the retinal ganglion cells of the cat. *Vision Res Suppl*, 3:87.

Benoff, B., Yang, H., Lawson, C. L., Parkinson, G., Liu, J., Blatter, E., Ebright, Y. W., Berman, H. M., and Ebright, R. H. (2002). Structural basis of transcription activation: the CAP-alpha/CTD-DNA complex. *Science*, 297:1562.

Berg, H. C. and Purcell, E. M. (1977). Physics of chemoreception. *Biophys J*, 20:193.

Berg, O. G. and von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193:723.

Bialek, W. (1987). Physical limits to sensation and perception. *Ann Rev Biophys Biophys Chem*, 16:455.

Bialek, W. (2002). *Thinking about the brain in Physics of biomolecules and cells: Les Houches Session LXXV*. EDP Sciences and Springer-Verlag, Les Ulis, Berlin.

Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity and learning. *Neural Comput*, 13:2409.

Bialek, W. and Setayeshgar, S. (2005). Physical limits to biochemical signaling. *Proc Natl Acad Sci USA*, 102:10040.

Bialek, W. and Setayeshgar, S. (2006). Cooperativity, sensitivity and noise in biochemical signaling. *arXiv*, q-bio.MN/0601001.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15:116.

Blahut, R. E. (1972). Computation of channel capacity and rate–distortion functions. *IEEE Trans Info Th*, 4:460.

Blake, W. J., Kaern, M., Cantor, C. R., and Collins, J. J. (2003). Noise in eukaryotic gene expression. *Nature*, 422:633.

Brenner, N., Bialek, W., and de Ruyter van Steveninck, R. (2000). Adaptive rescaling optimizes information transmission. *Neuron*, 26:695.

Burz, D. S., Pivera-Pomar, R., Jackle, H., and Hanes, S. D. (1998). Cooperative DNA-binding by Bicoid provides a mechanism for threshold dependent gene activation in the Drosophila embryo. *EMBO J*, 17:5998.

Camalet, S., Duke, T. A. J., Julicher, F., and Prost, J. (2000). Auditory sensitivity provided by self-tuned critical oscillation of hair cells. *Proc Natl Acad Sci USA*, 97:3183.

Chen, K. C., Csikasz-Nagy, A., Gyorffy, B., Val, J., Novak, B., and Tyson, J. J. (2000). Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol Biol Cell*, 11:369.

Conti, F., de Felice, L. J., and Wanke, E. (1975). Potassium and sodium ion current noise in the membrane of the squid giant axon. *J Physiol (Lond)*, 248:45.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley, New York.

Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT Press, Cambridge.

Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13:2381.

Driever, W. and Nusslein-Volhard, C. (1988a). The Bicoid protein determines position in the Drosophila embryo. *Cell*, 54:95.

Driever, W. and Nusslein-Volhard, C. (1989). The Bicoid protein is a positive regulator of Hunchback transcription in the early Drosophila embryo. *Nature*, 337:138.

Driever, W. and Nusslein-Volhard, V. (1988b). A gradient of Bicoid protein in Drosophila embryos. *Cell*, 54:83.

Dudik, M., Phillips, S. J., and Schapire, R. E. (2004). Performance guarantees for regularized maximum entropy density estimation. Proceedings of the 17th Annual conference on learning theory.

Duke, T. A. J. and Bray, D. (1999). Heightened sensitivity of a lattice of membrane receptors. *Proc Natl Acad Sci USA*, 96:10104.

Eguiluz, V. M., Ospeck, M., Choe, Y., Hudspeth, A. J., and Magnasco, M. O. (2000). Essential nonlinearities in hearing. *Phys Rev Lett*, 84:5232.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95:14863.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297:1183.

Fatt, P. and Katz, B. (1950). Some observations on biological noise. *Nature*, 166:597.

Fatt, P. and Katz, B. (1952). Spontaneous subthreshold activity at motor nerve endings. *J Physiol (Lond)*, 117:109.

Foat, B. C., Morozov, A. V., and Bussemaker, H. J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22:e141.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303:799.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in response of yeast cells to environmental changes. *Mol Biol Cell*, 11:4241.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, 123:1025.

Goodridge, H. S. and Harnett, M. M. (2005). Introduction to immune cell signalling. *Parasitology*, 130:S3.

Gregor, T. (2005). Biophysics of early embryonic development. Dissertation, Princeton University.

Gregor, T., Tank, D. W., Wieschaus, E. F., and Bialek, W. (2006a). Probing the limits to positional information. submitted Cell.

Gregor, T., Wieschaus, E. F., McGregor, A. P., Bialek, W., and Tank, D. W. (2006b). Stability and nuclear dynamics of the Bicoid morphogen gradient. submitted Cell.

Halford, S. E. and Marko, J. F. (2004). How do site-specific DNA-binding proteins find their targets. *Nucl Acids Res*, 32:3040.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402:C47.

Hecht, S., Shlaer, S., and Pirenne (1942). Energy, quanta, and vision. *J Gen Physiol*, 25:819.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput*, 14:1771.

Hinton, G. E. and Sejnowski, T. J. (1986). *Learning and relearning in Boltzmann machines in Parallel distributed processing: explorations in the microstructure of cognition*, volume 1. MIT Press, Cambridge.

Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol*, 117:500.

Hooshangi, S., Thiberge, S., and Weiss, R. (2005). Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proc Natl Acad Sci USA*, 102:3581.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA*, 79:2554.

Hopfield, J. J. and Tank, D. W. (1986). Computing with neural circuits: a model. *Science*, 233.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat Gen*, 31:370.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31:264.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys Rev*, 106:620.

Kinney, J. B., Tkačik, G., and Jr, C. G. C. (2007). Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA*, 104:501.

Kolch, W. (2000). Meaningful relationships: the regulation of Ras/Raf/MEK/ERK pathway by protein interactions. *Biochem J*, 351:289.

Kuhlman, T., Zhang, Z., Jr, M. H. S., and Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of Escherichia coli. *Proc Natl Acad Sci USA*, 104:6043.

Laughlin, S. B. (1981). A simple coding procedure enhances a neuron's information capacity. *Z Naturforsch*, 36c:910.

Lawrence, P. A. (1992). *The making of a fly: the genetics of animal design.* Blackwell, Oxford.

Lecar, H. and Nossal, R. (1971a). Theory of threshold fluctuations in nerves. I: Relationships between eletctrical noise and fluctuations in axon firing. *Biophys J*, 11:1048.

Lecar, H. and Nossal, R. (1971b). Theory of threshold fluctuations in nerves. II: Analysis of various sources of membrane noise. *Biophys J*, 11:1068.

Leloup, J. C. and Goldbeter, A. (2003). Toward a detailed computational model for the mammalian circadian clock. *Proc Natl Acad Sci USA*, 100:7051.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Trans Info Th*, 37:145.

Luria, S. E. and Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28:491.

Ma, X., Yuan, D., Diepold, K., Scarborough, T., and Ma, J. (1996). The Drosophila morphogenetic protein Bicoid binds DNA cooperatively. *Development*, 122:1195.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge.

Mathieson, K., Kachiguine, S., Adams, C., Cunningham, W., Gunning, D., O'Shea, V., Smith, K. M., Chichilnisky, E. J., Litke, A. M., Sher, A., and Rahman, M. (2004). Large-area microelectrode arrays for recording of neural signals. *IEEE Trans Nucl Sci*, 51:2027.

Mezard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin Glass Theory and Beyond*. World Scientific, Singapore.

Newman, J. R., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L., and Weissman, J. S. (2006). Single-cel proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441:840.

Ochoa-Espinosa, A., Yucel, G., Kaplan, L., Pare, A., Pura, N., Obersten, A., Papatsenko, D., and Small, S. (2005). The role of binding site cluster strength in Bicoid–dependent patterning in Drosophila. *Proc Natl Acad Sci USA*, 102:4960.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nat Genet*, 31:69.

Paulsson, J. (2004). Summing up the noise in gene networks. *Nature*, 427:415.

Pedone, P. V., Ghirlando, R., Clore, G. M., Gronenbron, A. M., Felsenfeld, G., and Omchinski, J. G. (1996). 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 USA*, 93:2822.

Pedraza, J. M. and van Oudenaarden, A. (2005). Noise propagation in gene networks. *Science*, 207:1965.

Ptashne, M. (1992). *A genetic switch. Second edition: Phage lamnda and higher organisms*. Cell Press, Cambridge.

Puchalla, J. L., Schneidman, E., Harris, R. A., and II, M. J. B. (2005). Redundancy in the population code of the retina. *Neuron*, 46:493.

Raser, J. M. and O'Shea, E. K. (2004). Control of stochasticity in eukaryotic gene expression. *Science*, 304:1811.

Remondini, D., O'Connel, B., Intrator, N., Sedivy, J. M., Neretti, N., Castellani, G. C., and Cooper, L. N. (2005). Targeting c-Myc-activated genes with a correlation method: detection of global changes in large gene expression network dynamics. *Proc Natl Acad Sci USA*, 102:6902.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1997). *Spikes: exploring the neural code*. MIT Press, Cambridge.

Rivera-Pomar, R. and Jäckle, H. (1996). From gradients to stripes in Drosophila embryogenesis: Filling in the gaps. *Trends Gen*, 12:478.

Robinson, K., McGuire, A. M., and Church, G. M. (1998). A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. *J Mol Biol*, 284:241.

Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., and Elowitz, M. B. (2005). Gene regulation at a single cell level. *Science*, 307:1962.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-protein signaling networks derived from multiparameter single-cell data. *Science*, 308:523.

Sanchez, L. and Thieffry, D. (2001). A logical analysis of the Drosophila gap-gene system. *J Theor Biol*, 211:115.

Schneidman, E., Freedman, B., and Segev, I. (1998). Ion channel stochasticity may be critical in determinng the reliability and precision of spike timing. *Neural Comp*, 10:1679.

Schneidman, E., II, M. J. B., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007.

Schneidman, E., Still, S., II, M. J. B., and Bialek, W. (2003). Network information and connected correlations. *Phys Rev Lett*, 91:238701.

Schwatz, G. (2006). personal communication.

Segev, R., Goodhouse, J., Puchalla, J., and II, M. J. B. (2004). Recording spikes from a large fraction of the ganglion cells in a retinal patch. *Nat Neurosci*, 7:1154.

Setty, Y., Mayo, A. E., Surette, M. G., and Alon, U. (2003). Detailed map of the cis-regulatory input function. *Proc Natl Acad Sci USA*, 100:7702.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Sys Tech J*, 27:379–423, 623–656.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, 31:64.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M., and Chichilnisky, E. J. (2006). The structure of multi-neuron firing patterns in primate retina. *J Neurosci*, 26:8254.

Sigworth, F. J. (1977). Sodium channels in nerve apparently have two conductance states. *Nature*, 270:265.

Sigworth, F. J. (1980). The variance of sodium current fluctuations at the node of Ranvier. *J Physiol (Lond)*, 207:97.

Silverman, R. (1955). On binary channels and their cascades. *IEEE Trans Info Th*, IT-1:19.

Slonim, N., Atwal, G. S., Tkačik, G., and Bialek, W. (2005a). Estimating mutual information and multi-information in large networks. *arXiv*, cs.IT/0502017.

Slonim, N., Atwal, G. S., Tkačik, G., and Bialek, W. (2005b). Information-based clustering. *Proc Natl Acad Sci USA*, 51:18297.

Slonim, N., Elemento, O., and Tavazoie, S. (2006). Ab-initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Mol Syst Biol*, 2:1.

Slutsky, M. and Mirny, L. A. (2004). Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys J*, 87:4021.

Smirnakis, S. M., II, M. J. B., Warland, D. K., Bialek, W., and Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial scale. *Nature*, 386:69.

Stephens, G. J., Johnson-Kerner, B., Bialek, W., and Ryu, W. S. (2007). Dimensionality and dynamics in the behavior of C. elegans. *arXiv*, 0705.0313.

Stevens, C. F. (1972). Inferences about membrane properties from electrical noise measurement. *Biophys J*, 12:1028.

Strong, S. P., Koberle, R., de Ruyter van Staveninck, R., and Bialek, W. (1998). Entropy and information in neural spike trains. *Phys Rev Lett*, 80:197.

Struhl, G., Struhl, K., and Macdonald, P. M. (1989). The gradient morphogen Bicoid is a concentration–dependent transcriptional activator. *Cell*, 57:1259.

Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA*, 99:12795.

Tishby, N. (2007). Optimal adaptation and predictive information. In *Proceedings of the Cosyne 2007 conference*.

Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv*, physics/0004057v1.

Tkačik, G., Gregor, T., and Bialek, W. (2007a). The role of input noise in transcriptional reglation. *arXiv*, q-bio.MN/0701002.

Tkačik, G., Jr, C. G. C., and Bialek, W. (2007b). Information flow and optimization in transcriptional control. *arXiv*, arXiv:0705.0313v1.

Tkačik, G., Schneidman, E., II, M. J. B., and Bialek, W. (2006). Ising models for networks of real neurons. *arXiv*, q-bio.NC/0611072.

Tyson, J. J., Chen, K., and Novak, B. (2001). Network dynamics and cell physiology. *Nat Rev Mol Cell Biol*, 2:908.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet*, 19:479.

van Zon, J. S., Morelli, M. J., Tanase-Nicola, S., and ten Wolde, P. R. (2006). Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophys J*, 91:14350.

Verveen, A. A. and Derksen, H. E. (1965). Fluctuation in membrane potential of axons and the problem of coding. *Kybernetik*, 2:152.

Verveen, A. A. and Derksen, H. E. (1968). Fluctuation phenomena in nerve membrane. *Proc IEEE*, 56:906.

von Neumann, J. and Morgenstern, O. (1947). *Theory of Game and Economic Behavior*. Princeton University Press, Princeton.

Walczak, A. M., Sasai, M., and Wolynes, P. G. (2005). Self-consistent proteomic field theory of stochastic gene switches. *Biophys J*, 88:828.

Wang, Y. M., Austin, R. H., and Cox, E. C. (2006). Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys Rev Lett*, 97:048302.

Winston, R. L., Millar, D. P., Gottesfeld, J. M., and Kent, S. B. (1999). Characterization of the DNA binding properties of the bHLH domain of the Deadpan to single and tandem sites. *Biochemistry*, 38:5138.

Wolpert, L. (1969). Positional information and the spatial pattern of cellular differentiation. *J Theor Biol*, 25:1.

Zhao, C., York, A., Yang, F., Forsthoefel, D. J., Dave, V., Fu, D., Zhang, D., Corado, M. S., Small, S., Seeger, M. A., and Ma, J. (2002). The activity of the Drosophila morphogenetic protein Bicoid is inhibited by a domain located outside its homeodomain. *Development*, 129:1669.

Ziv, E., Nemenman, I., and Wiggins, C. (2006). Optimal signal processing in small stochastic biochemical networks. *arXiv*, q-bio.MN/0612041.