

# Information theory and adaptation

Ilya Nemenman\*

*Departments of Physics and Biology and Computational and Life  
Sciences Strategic Initiative, Emory University, Atlanta, GA 30322*

(Dated: November 24, 2010)

*To appear as Chapter 5 of Quantitative Biology: From Molecular to Cellular Systems, ME Wall, ed. (Taylor and Francis, 2011). In this Chapter, we ask questions (1) What is the right way to measure the quality of information processing in a biological system? and (2) What can real-life organisms do in order to improve their performance in information-processing tasks? We then review the body of work that investigates these questions experimentally, computationally, and theoretically in biological domains as diverse as cell biology, population biology, and computational neuroscience.*

## I. LIFE IS INFORMATION PROCESSING

All living systems have evolved to perform certain tasks in specific contexts. There are a lot fewer tasks than there are different biological solutions that the nature has created. Some of these problems are universal, while the solutions may be organism-specific. Thus a lot can be understood about the structure of biological systems by focusing on understanding of *what* they do and *why* they do it, in addition to *how* they do it on molecular or cellular scales. In particular, this way we can uncover phenomena that generalize across different organisms, thus increasing the value of experiments and building a coherent understanding of the underlying physiological processes.

In this Chapter, we will take this point of view while analyzing what it takes to do one of the most common, universal functions performed by organisms at all levels of organization: signal or information processing and shaping of a response (these are variously known in different contexts as learning from observations, signal transduction, regulation, sensing, adaptation, etc.) Studying these types of phenomena poses a series of well-defined, physical questions: How can organisms deal with noise, whether extrinsic or generated by intrinsic stochastic fluctuations within molecular components of information processing devices? How long should the world be observed before a certain inference about it can be made? How is the internal representation of the world made and stored over time? How can organisms ensure that the information is processed fast enough for the formed response to be relevant in the ever-changing world? How should the information processing strategies change when the properties of the environment surrounding the organism change? In fact, such “information processing” questions have been featured prominently in studies on all scales of biological complexity, from learning phenomena in animal behavior [1–7], to analysis of neural computation in small and large animals [8–16], and to molecular information processing circuits [17–25], to name just a

few.

In what follows, we will not try to embrace the unembraceable, but will instead focus on just a few questions, fundamental to the study of signal processing in biology: What is the right way to measure the quality of information processing in a biological system? and What can real-life organisms do in order to improve their performance in these tasks?

### The main questions addressed in this review:

- What is the right way to measure the quality of information processing in a biological system?
- What can real-life organisms do in order to improve their performance in these tasks?

The field of study of biological information processing has undergone a dramatic growth in the recent years, and it is expanding at an ever growing rate. There are now entire conferences devoted to the related phenomena (perhaps the best example is *The International q-bio Conference on Cellular Information Processing*, <http://q-bio.org>, held yearly in Santa Fe, NM, USA). Hence, in this short chapter, we have neither an ability, nor a desire to provide an exhaustive literature review. Instead the reader should keep in mind that the selection of references cited here is a biased sample of important results in the literature, and I apologize profusely to my friends and colleagues who find their deserving work omitted in this overview.

## II. QUANTIFYING BIOLOGICAL INFORMATION PROCESSING

In the most general context, a biological system can be modeled as an input-output device, cf. Fig. 1 that observes a time-dependent state of the world  $s(t)$  (where  $s$  may be intrinsically multidimensional, or even formally infinite dimensional), processes the information, and initiates a response  $r(t)$  (which can also be very large dimensional). In some cases, in its turn, the response changes

---

\*Electronic address: [ilya.nemenman@emory.edu](mailto:ilya.nemenman@emory.edu)

the state of the world and hence influences the future values of  $s(t)$ , making the whole analysis so much harder [26]. In view of this, analyzing the information processing means quantifying certain aspects of the mapping  $s(t) \rightarrow r(t)$ . In this section, we will discuss the properties that this quantification should possess, and we will introduce the quantities that satisfy them.

### A. What is needed?

One typically tries to model molecular or other physiological *mechanisms* of the response generation. For example, in well-mixed biochemical kinetics approaches, where  $s(t)$  may be a ligand concentration, and  $r(t)$  may be an expression level of a certain protein, we often write

$$\frac{dr(t)}{dt} = F_a(r, s, h) - G_a(r, s, h) + \eta_a(r, s, h, t), \quad (1)$$

where the nonnegative functions  $F_a$  and  $G_a$  stand for the production/degradation of the response, influenced by the level of the signal  $s$ , and  $\eta$  is a random forcing due to the intrinsic stochasticity of chemical kinetics at small molecular copy numbers [27]. The subscript  $a$  stands for the values of adjustable parameters that define the response (such as various kinetic rates, concentrations of intermediate enzymes, etc.), which themselves can change, but on time scales much slower than the dynamics of  $s$  and  $r$ . In addition,  $h$  stands for the activity of other, hidden cellular state variables, which change according to their own dynamics, similar to Eq. (1). This dynamics can be written for many diverse biological information processing systems, including the neural dynamics, where  $r$  will stand for the firing rate of a neuron induced by the stimulus [28].

Importantly, because of the intrinsic stochasticity in Eq. (1), and because of the effective randomness introduced by the state of the hidden variables, the mapping between the stimulus and the response is non-deterministic, and it is summarized in the probability distribution  $P[\{r(t)\} | \{s(t)\}, \{h(t)\}, a]$ , or, marginalizing over  $h$ ,  $P[\{r(t)\} | \{s(t)\}, a] \equiv P_a[\{r(t)\} | \{s(t)\}]$ . In addition,  $s(t)$  itself is not deterministic either: other agents, chaotic dynamics, statistical physics effects, and, at a more microscopic level, even quantum mechanics

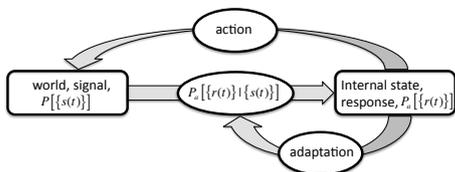


FIG. 1: Biological information processing and interactions with the world. In this review we leave aside the feedback action between the organism internal state and the state of the world and focus on the signal processing and the adaptation arrows.

conspire to ensure that  $s(t)$  can only be specified probabilistically. Therefore, a simple mapping  $s \rightarrow r$  is replaced by a joint probability distribution (note that we will drop the index  $a$  in the future where it doesn't cause ambiguities)

$$\begin{aligned} P[\{r(t)\} | \{s(t)\}, a] P[\{s(t)\}] \\ = P[\{r(t)\}, \{s(t)\} | a] \equiv P_a[\{r(t)\}, \{s(t)\}]. \end{aligned} \quad (2)$$

Hence the measure of the quality of the biological information processing must be a *functional* of this joint distribution.

**Biological information processing is almost always probabilistic.**

Now consider, for example, a classical system studied in cellular information processing: the *E. coli* chemotaxis (see Chapter 15 in this book) [29]. This bacterium is capable of swimming up gradients of various nutrients. In this case, the signal  $s(t)$  is the concentration of such extracellular nutrients. The response of the system is the activity levels of various internal proteins, like *cheY*, *cheA*, *cheB*, *cheR*, etc., which combine to modulate the cellular motion through the environment. It is possible to write the chemical kinetics equations that relate the stimulus to the response accurately enough and eventually produce the sought after conditional probability distribution  $P_a[\{r(t)\} | \{s(t)\}]$ . However, are the ligand concentrations the variables that the cell “cares” about? In this system, it is reasonable to assume that all protein expression states that result in the same intake of the catabolite are functionally equivalent. That is, the goal of the information processing performed by the cell likely is not to serve as a passive transducer of the signal into the response [24, 25], but to make an active computation that extracts only the part of the signal that is relevant to making behavioral decisions. We will denote such *relevant* aspects of the world as  $e(t)$ . For example, for the chemotactic bacterium,  $e(t)$  can be the maximum nutrient intake realizable for a particular spatiotemporal pattern of the nutrient concentration.

In general,  $e$  is not a subset of  $s$ , or vice versa, and instead the relation between  $s$  and  $e$  is also probabilistic,  $P[\{e(t)\} | \{s(t)\}]$ , and hence the relevant variable, the signal, and the response form a Markov chain:

$$\begin{aligned} P[\{e(t)\}, \{s(t)\}, \{r(t)\}] = \\ P[\{e(t)\}] P[\{s(t)\} | \{e(t)\}] P[\{r(t)\} | \{s(t)\}]. \end{aligned} \quad (3)$$

The quantity we are seeking to characterize the biological information processing must respect this aspect of the problem. Therefore, its value must depend explicitly on the choice of the relevance variable: a computation resulting in the same response will be either “good” or “bad” depending on what this response is used for. In other words, one needs to know what the problem is before saying if a solution is good or bad.

**It is impossible to quantify the information processing without specifying the purpose of the device; that is, the relevant quantity that it is supposed to compute.**

## B. Introducing the right quantities

The question of how much can be inferred about a state of a variable  $X$  from measuring a variable  $Y$  has been answered by Claude Shannon over sixty years ago [30]. Starting with basic, uncontroversial axioms that a measure of information must obey, he derived that the uncertainty in a state of a variable is given by

$$S[X] = - \sum_x P(x) \log P(x) = - \langle \log P(x) \rangle_{P(x)}, \quad (4)$$

which we know now as the *Boltzmann-Shannon entropy*. Here  $\langle \dots \rangle_P$  denotes averaging over the probability distribution  $P$ . When the logarithm in Eq. (4) is binary (which we always assume in this Chapter), then the unit of entropy is a *bit*: one bit of uncertainty about a variable means that the latter can be in one of two states with equal probabilities.

Observing the variable  $Y$  (a.k.a. *conditioning* on it) changes the probability distribution of  $X$ ,  $P(x) \rightarrow P(x|y)$ , and the difference between the entropy of  $X$  prior to the measurement and the average conditional entropy tells how informative  $Y$  is about  $X$ :

$$I[X; Y] = S[X] - \langle S[X|Y] \rangle_{P(y)} \quad (5)$$

$$= - \langle \log P(x) \rangle_{P(x)} + \left\langle \langle \log P(y|x) \rangle_{P(x|y)} \right\rangle_{P(y)} \quad (6)$$

$$= - \left\langle \log \frac{P(x, y)}{P(x)P(y)} \right\rangle_{P(x, y)}. \quad (7)$$

The quantity  $I[X; Y]$  in Eq. (7) is known as *mutual information*. As entropy, it is measured in bits. Mutual information of one bit means that specifying the variable  $Y$  provides us with the knowledge to answer one yes/no question about  $X$ .

Entropy and information are additive quantities. That is, when considering entropic quantities for time series data defined by  $P[\{x(t)\}]$ , for  $0 \leq t \leq T$ , the entropy of the entire series will diverge linearly with  $T$ . Therefore, it makes sense to define entropy and information rates [30]

$$\mathcal{S}[X] = \lim_{T \rightarrow \infty} \frac{S[x(0 \leq t < T)]}{T}, \quad (8)$$

$$\mathcal{I}[X; Y] = \lim_{T \rightarrow \infty} \frac{I[\{x(0 \leq t < T)\}; \{y(0 \leq t < T)\}]}{T} \quad (9)$$

which measure the amount of uncertainty in the signal and the reduction of this uncertainty by the response per unit time.

Entropy and mutual information possess some simple, important properties [31]:

1. Both quantities are non-negative,  $0 \leq S[X]$  and  $0 \leq I[X; Y] \leq \min(S[X], S[Y])$ .
2. Entropy is zero if and only if (*iff*) the studied variable is not random. Further, mutual information is zero *iff*  $P(x, y) = P(x)P(y)$ , that is, there are no any kind of statistical dependences between the variables.
3. Mutual information is symmetric,  $I[X; Y] = I[Y; X]$ .

4. Mutual information is well defined for continuous variables; one only needs to replace sums by integrals in Eq. (7). On the contrary, entropy formally diverges for continuous variables (any truly continuous variable requires infinitely many bits to be specified to an arbitrary accuracy), but many properties of entropy are also exhibited by the *differential entropy*,

$$S[X] = - \int_x dx P(x) \log P(x), \quad (10)$$

which measures the entropy of a continuous distribution relative to the uniformly distributed one. In this Chapter,  $S[X]$  will always mean the differential entropy if  $x$  is continuous and the original entropy otherwise.

5. For a Gaussian distribution with a variance of  $\sigma^2$ ,

$$S = 1/2 \log \sigma^2 + \text{const}, \quad (11)$$

and, for a bivariate Gaussian with a correlation coefficient of  $\rho$ ,

$$I[X; Y] = -1/2 \log(1 - \rho^2). \quad (12)$$

Thus entropy and mutual information can be viewed as generalizations of more familiar notions of variance and covariance.

6. Unlike entropy, mutual information is invariant under reparameterization of variables. That is

$$I[X; Y] = I[X'; Y'] \quad (13)$$

for all invertible  $x' = x'(x)$ ,  $y' = y'(y)$ . That is,  $I$  provides a measure of statistical dependence between  $X$  and  $Y$  that is independent of our subjective choice of the measurement device [101].

## C. When the relevant variable is unknown: The value of information about the world

One of the most fascinating properties of mutual information is the Data Processing Inequality [31].

Suppose three variables  $X$ ,  $Y$ , and  $Z$  form a Markov chain,  $P(x, y, z) = P(x)P(y|x)P(z|y)$ . In other words,  $Z$  is a probabilistic transformation of  $Y$ , which, in turn, is a probabilistic transformation of  $X$ . Then it can be proven that

$$I[X; Z] \leq \min(I[X; Y], I[Y; Z]). \quad (14)$$

That is, *you cannot get new information about the original variable by further transforming the measured data*; any such transformation cannot increase the information.

Together with the fact that mutual information is zero *iff* the variables are completely statistically independent, the Data Processing Inequality suggests that if the variable of interest that the organism cares about is unknown to the experimenter, then analyzing the mutual information between the entire input stimulus (sans noise) and the response may serve as a good proxy. Indeed, due to the Data Processing Inequality, if  $I[S; R]$  is small, then  $I[E; R]$  is also small for any mapping  $S \rightarrow E$  of the signal into the relevant variable, whether deterministic,  $e = e(s)$ , or probabilistic,  $P(e|s)$ . In many cases, such as [16, 21, 32], this allows us to stop guessing which calculation the organism is trying to perform and to put an upper bound on the efficiency of the information transduction, whatever an organism cares about. However, as was recently shown in the case of chemotaxis in *E. coli*, when  $e$  and  $s$  are substantially different (resource consumption rate vs. instantaneous surrounding resource concentration), maximizing  $I[S; R]$  is not necessarily what organisms do [25].

**Information about the outside world is the upper bound on information about any of its features.**

Another reason to study the information about the outside world comes from the old argument that relates information and game theory [33]. Namely, consider a zero-sum probabilistic betting game (think of a roulette without the zeros, where the red and the black are two equally likely outcomes, and betting on the right outcome doubles one's investment, while betting on the wrong one leads to a loss of the bet). Then the logarithmic growth rate of one's capital is limited from above by the mutual information between the outcome of the game and the betting strategy. This was recently recast in the context of population dynamics in fluctuating environments [34–36]. Suppose the environment surrounding a population of genetically identical organisms fluctuates randomly with no temporal correlations among multiple states with probabilities  $P(s)$ . Each organism, independently of the rest, may choose among a variety of phenotypic decisions  $d$ , and the log-growth rate depends on the pairing of  $s$  and  $d$ . Evolution is supposed to maximize this rate, averaged over long times. However, the current state of the environment is not directly known, and the organisms may need to respond probabilistically. While the short-term gain would suggest choosing the response

that has the highest growth rate for the most probable environment, the longer term strategy would require bet-hedging [37], with different individuals making different decisions.

Suppose an individual now observes the environment and gets an imperfect internal representation of it,  $r$ , with the conditional probability of  $P(r|s)$ . What is the value of this information? Under very general conditions, this information about the environment can improve the log-growth rate by as much as  $I[S; R]$  [34]. In more general scenarios, the maximum log-growth advantage over uninformed peers needs to be discounted by the cost of obtaining the information, by the delay in getting it [35], and, more trivially, by the ability of the organism to utilize it. Therefore, while these brief arguments are far from painting a complete picture of relation between information and natural selection, it is already clear that maximization of the information between the surrounding world and the internal response to it is not an academic exercise, but is directly related to fitness and will be selected for by evolution.[102]

**Information about the outside world puts an upper bound on the fitness advantage of an individual over uninformed peers.**

It is now well known that probabilistic bet hedging is the strategy taken by bacteria for survival in the presence of antibiotics [38, 39] and for genetic recombination [40–42]. In both cases, cell division (and hence population growth) must be stopped either to avoid DNA damage by antibiotics, or to incorporate newly acquired DNA into the chromosome. Still, a small fraction of the cells choose not to divide even in the absence of antibiotics to reap the much larger benefits if the environment turns sour (these are called the *persistent* and the DNA uptake *competent* bacteria for the two cases, respectively). However, it remains to be seen in an experiment if real bacteria can reach the maximum growth advantage allowed by the information-theoretic considerations. Another interesting possibility is that cancer stem cells and mature cancer cells also are two probabilistic states chosen to hedge bets against interventions of immune systems, drugs, and other surveillance mechanisms [43].

#### D. Time dependent signals: Information and Prediction

In many situations, such as persistence in the face of antibiotics treatment mentioned above, an organism settles into a certain response much faster than the environment has a chance to change again. In these cases, it is sufficient to consider the same-time mutual information between the signals and the responses, as in [21],  $I[s(t); r(t)] = I[S; R]$ , which is what we've been doing up to now.

More generally, formation of any response takes time,

which may be comparable to time scales of changes of the stimuli. What are the relevant quantities to characterize biological information processing in such situations? Traditionally, one either considers delayed informations [44],

$$I_\tau[S; R] = I[s(t); r(t + \tau)], \quad (15)$$

where  $\tau$  may be chosen as  $\tau = \arg \max_{t'} I[s(t); r(t + t')]$ , or studies information rates, as in Eq. (9). The first choice measures the information between the stimulus and the response most constrained by it; typically this would be the response formed a certain characteristic signal transduction time after the stimulus occurrence. The second approach looks at correlations between all possible pairs of stimuli and responses separated by different delays.

While there are plenty of examples of biological systems where one or the other of these quantities is optimized, both of these approaches are insufficient.  $I_\tau$  doesn't represent all of the gathered information since bits at different moments of time are not independent of each other. Further, it does not take into the account that temporal correlations in the stimulus allow to predict it, and hence the response may be formed even before the stimulus occurs. On the other hand, the information rate does not distinguish among predicting the signal, knowing it soon after it happens, or having to wait for  $T \rightarrow \infty$  in order to be able to estimate it from the response.

To avoid these pitfalls, one can consider the information available to an organism that is relevant for specifying not all of the stimulus, but only of its future. Namely, we can define the *predictive information* about the stimulus available from observation of a response to it of a duration  $T$ ,

$$I_{\text{pred}}[R(T); S] = I[\{r(-T \leq t \leq 0)\}; \{s(t > 0)\}]. \quad (16)$$

This definition is a generalization of the one used in [45], which had  $r(t) = s(t)$ , and hence calculated the upper bound on  $I_{\text{pred}}$  over all possible sensory schemes  $P[\{r(t)\}|\{s(t)\}]$ .

All of the  $I_{\text{pred}}$  bits are available to be used instantaneously, and there is no specific delay  $\tau$  chosen a priori and semi-arbitrarily. The predictive information is nonzero only to the extent that the signal is temporally correlated, and hence the response to its past values can say something about its future. Thus focusing on predictability may resolve a traditional criticism of information theory that bits don't have an intrinsic meaning and value, and some are more useful than the others: since any action takes time, only those bits have value that can be used to predict the stimulus at the time of action, that is, in the future [45, 46].

**Predictive information allows to assign an objective value to information: only those bits are useful that can be used to guide future responses.**

The notion of predictive information is conceptually appealing, and there is clear experimental and computational evidence that behavior of biological systems, from bacteria to mammals, is consistent with attempting to make better predictions (see, for example, [16, 47–52] for just some results). However, even almost ten years after  $I_{\text{pred}}$  was first introduced, it still remains to be seen experimentally if optimizing predictive information is one of the objectives of biological systems, and whether population growth rates in temporally correlated environments can be related to the amount of information available to predict them. Some of the reasons for the relative lack of progress may be practical considerations that estimation of informations among nonlinearly related multidimensional variables [16, 53, 54] or extracting the predictive aspects of the information [55] from empirical data is hard, while for simple Gaussian signals and responses with finite correlation times, optimization of predictive information reduces to a much more prosaic matching of Wiener extrapolation filters [56].

### III. IMPROVING INFORMATION-PROCESSING PERFORMANCE

Understanding the importance of information about the outside world and knowing which quantities can be used to measure it, we are faced with the next question: How can the available information be increased in view of the limitations imposed by the physics of the signal and of the processing device, such as stochasticity of molecular numbers and arrival times, or energy constraints?

#### A. Strategies for Improving The Performance

We start with three main theorems of information theory due to Shannon [30]. In the *source coding theorem*, he proved that to record a signal without losses, one needs only  $\mathcal{S}$ , the signal entropy rate, bits per unit time. In the *channel coding theorem*, he showed that the maximum rate of errorless transmission of information through a channel specified by  $P[\{r(t)\}|\{s(t)\}]$  is given by  $C = \max_{P(\{s(t)\})} \mathcal{I}[R; S]$ , which is called the channel capacity. Finally, the *rate distortion theorem* calculates the minimum size of the message that must be sent error-free in order to recover the signal with an appropriate mean level of some pre-specified distortion measure. None of these theorems considers the time delay before the message can be decoded, and typically one would need to wait for very long times and accumulate long message sequences to reach the bounds predicted by the theorems since, for example, responses long time away from a certain signal may still carry some information about it.

Leaving aside the complication of dynamics, which one may hope to solve some day using the predictive information ideas, these theorems tell us exactly what an or-

ganism can do to optimize the amount of information it has about the outside world. First, one needs to compress the measured signal, removing as many redundancies as possible. There is evidence that this happens in a variety of signaling systems, starting with the classical Refs. [47, 57–59]. Second, one needs to encode the signal in a way that allows the transmitted information to approach the channel capacity limit by remapping different values of signal into an intermediate signaling variable whose states are easier to transmit without an error. Again, there are indications that this happens in living systems [22, 60–64]. Finally, one may choose to focus only on important aspects of the signal, such as communicating changes in the signal, or thresholding its value [16, 29, 65, 66].

If the references in the previous paragraph look somewhat thin, it is by choice since neither of these approaches are unique to biology, and, in fact, most artificial communication systems use them: a cell phone filters out audio frequencies that are irrelevant to human speech, compresses the data, and then encodes it for sending with the smallest possible errors. A lot of engineering literature discusses these questions [31], and we will not touch them here anymore. What makes biological systems unique is an ability to improve the information transmission by modifying their own properties in the course of their life. This adjusts the  $a$  in  $P_a[\{r(t)|\{s(t)\}]$ , and hence modifies the conditional probability distribution itself. This would be equivalent to a cell phone being able to change its physical characteristics on the fly. Unfortunately, as the recent issues with the iPhone antenna have shown, human engineered systems are no match to biology in this regard: they are largely incapable of adjusting their own design if the original turns out to be flawed.

**Unlike most artificial systems, living organisms can change their own properties to optimize their information processing.**

The property of changing one’s own characteristics in response to the observed properties of the world is called *adaptation*, and the remainder of this section will be devoted to its overview. In principle, we make no distinction whether this adaptation is achieved by natural selection or by physiological processes that act on much faster time scales (comparable to the typical signal dynamics), and sometimes the latter may be as powerful as the former [21, 67]. Further, we note that adaptation of the response probability distribution and formation of the response itself are, in principle, a single process of formation of the response on multiple time scales. Our ability to separate it into a fast response and a slow adaptation (and hence much of the discussion below) depends on existence of two well-separated time scales in the signal and in the mechanism of the response formation. While such clear separation is possible in some cases, it is harder in others, and especially when the time scales of the signal and the fast response may be changing themselves.

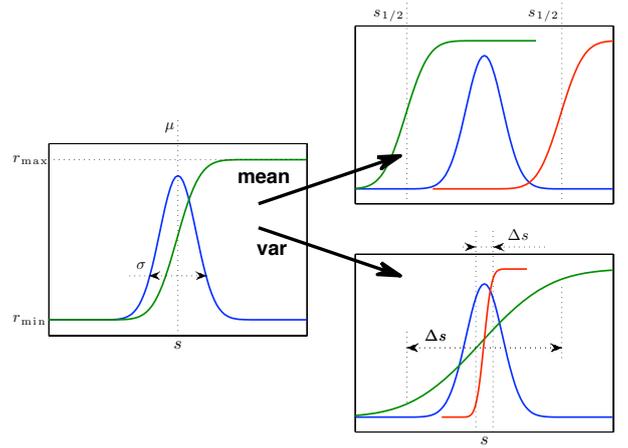


FIG. 2: Parameters characterizing response to a signal. Left panel: the probability distribution of the signal,  $P(s)$  (blue), and the best-matched steady state dose-response curve  $r_{ss}$  (green). Top right: mismatched response midpoint. Bottom right: mismatched response gain.

Cases without a clear separation of scales raise a variety of interesting questions, but we will leave them aside for this discussion.

### B. Three Kinds of Adaptation in Information Processing

We often can linearize the dynamics, Eq. (1), to get the following equation describing formation of small responses

$$\frac{dr}{dt} = f[s(t)] - kr + \eta(t, r, s). \quad (17)$$

Here  $r$  may be an expression of an mRNA following activation by a transcription factor  $s$ , or the firing rate of a neuron following stimulation. In the above expression,  $f$  is the response activation function, which depends on the current value of the signal;  $k$  is the rate of the first-order relaxation or degradation; and  $\eta$  is some stochastic process representing the intrinsic noisiness of the system. In this case,  $r(t)$  depends on the entire history of  $\{s(t')\}$ ,  $t' < t$ , and hence carries some information about it as well.

For quasi-stationary signals (that is, the correlation time of the signal,  $\tau_s \gg 1/k$ ), we can write the steady state dose-response (or firing rate, or ...) curve

$$r_{ss} = f[s(t)]/k, \quad (18)$$

and this will be smeared by the noise  $\eta$ . A typical monotonic sigmoidal  $f$  is characterized by only a few large-scale parameters: the range,  $f_{\min}$  and  $f_{\max}$ ; the argument  $s_{1/2}$  at the mid-point value  $(f_{\min} + f_{\max})/2$ ; and the width of the transition region,  $\Delta s$  (see Fig. 2). If

the mean signal  $\mu \equiv \langle s(t) \rangle_t \gg s_{1/2}$ , then, for most signals,  $r_{ss} \approx f_{\max}/k$  and responses to two typical different signals  $s_1$  and  $s_2$  are indistinguishable as long as

$$\left. \frac{dr_{ss}(s)}{ds} \right|_{s=s_1} (s_2 - s_1) < \sigma_\eta/k, \quad (19)$$

where  $\sigma_\eta/k$  is the precision of the response resolution expressed through the standard deviation of the noise. Similar situation happens when  $\mu \ll s_{1/2}$  and  $r_{ss} \approx f_{\min}/k$ . Thus, to reliably communicate information about the signal,  $f$  should be tuned such that  $s_{1/2} \approx \mu$ . If a real biological system can perform this adjustment, we call this *adaptation to the mean* of the signal, *desensitization*, or *adaptation of the first kind*. If  $s_{1/2}(\mu) = \mu$ , then the adaptation is *perfect*. This kind of adaptation has been observed experimentally and predicted computationally in a lot more systems than we can list here, including phototransduction, neural and molecular sensing, multi-state receptor systems, immune response, and so on, with active work persisting to date (see, e.g., Refs. [29, 64, 68–76] for a very incomplete list of references on the subject). For example, the best studied adaptive circuit in molecular biology, the control of chemotaxis of *E. coli* (see Chapter 15), largely produces adaptation of the first kind [77, 78]. Further, a variety of problems in synthetic biology are due precisely to the mismatch between the typical protein concentration of the input signal and the response function that maps this concentration into the rate of mRNA transcription or protein translation (cf. [79] and Chapter 4 in this book). Thus there is an active community of researchers working on endowing these circuits with proper adaptive matching abilities of the first kind.

Consider now the quasi-stationary signal taken from the distribution with  $\sigma \equiv (\langle (s(t))^2 \rangle_t - \mu^2)^{1/2} \gg \Delta s$ . Then the response to most of the signals is indistinguishable from the extremes, and it will be near the midpoint  $\sim (r_{\max} + r_{\min})/2$  if  $\sigma \ll \Delta s$ . Thus, to use the full dynamic range of the response, a biological system must tune the width of the sigmoidal dose-response curve to  $\Delta s \approx \sigma$ . We call this *gain control*, *variance adaptation*, or *adaptation of the second kind*. Experiments show that a variety of systems exhibit this adaptive behavior as well [80], especially in the context of neurobiology [11, 81], and maybe even of evolution [82].

These matching strategies are well known in signal processing literature under the name of histogram equalization. Surprisingly, they are nothing but a special case of optimizing the mutual information  $I[S; R]$ , as has been shown first in the context of information processing in fly photoreceptors [8]. Indeed, for quasi-steady state responses, when noises are small compared to the range of the response, the arrangement that optimizes  $I[S; R]$  is the one that produces  $P(r) \propto 1/\sigma_{r|s}$ . In particular, when  $\sigma_\eta$  is independent of  $r$  and  $s$ , this means that each  $r$  must be used equiprobably, that is,  $f^*(s) = \int_{-\infty}^s P(s') ds'$ . Adaptation of the first and the second kind follows from these considerations immediately. In more complex cases, when the noise variance

is not small or not constant, derivation of the optimal response activation function cannot be done analytically, but numerical approaches can be used instead. In particular, in transcriptional regulation of the early *Drosophila* embryonic development, the matching between the response function and the signal probability distribution has been observed for nonconstant  $\sigma_{r|s}$  [22]. However, we caution the reader that, even though adaptation *can* have this intimate connection to information maximization, and it is essentially omni-present, the number of systems where the adaptive strategy has been analyzed quantitatively to show that it results in optimal information processing is not that large.

We now relax the requirement of quasi-stationarity and return to dynamically changing stimuli. We rewrite Eq. (17) in the frequency domain,

$$r_\omega = \frac{[f(s)]_\omega + \eta_\omega}{k + i\omega}, \quad (20)$$

which shows that the simple first order (or linearized) kinetics performs low pass filtering of the nonlinearly transformed signal [18, 19]. As discussed long ago by Wiener [56], for given temporal correlations of the stimulus and the noise (which we summarize here for simplicity by correlation times  $\tau_s$  and  $\tau_\eta$ ), there is an optimal cut-off frequency  $k$  that allows to filter out as much noise as possible without filtering out the signal. Change of the parameter  $k$  to match the temporal structure of the problem is called the *time scale adaptation* or *adaptation of the third kind*. Just like the first two kinds, time scale adaptation also can be related to maximization of the stimulus-response mutual information by means of a simple observation that minimization of the quadratic prediction error of the Wiener filter is, under certain assumptions, equivalent to maximizing information about the signal, cf. Eq. (12).

This adaptation strategy is difficult to study experimentally since (a) detection of variation of the integration cutoff frequency  $k$  potentially requires observing the adaptation dynamics on very long time scales, and (b) prediction of optimal cutoff frequency requires knowing the temporal correlation properties of signals, which are far from trivial to measure (see, e.g., Ref. [83] for a review on literature on analysis of statistical properties of natural signals). Nonetheless, experimental systems as diverse as turtle cones [84], rats in matching foraging experiments [3], mice retinal ganglion cells [85], and barn owls adjusting auditory and visual maps [86] show adaptation of the filtering cutoff frequency in response to changes in the relative time scales and/or the variances of the signal and the noise. In a few rare cases, including fly self-motion estimation [13] and *E. coli* chemotaxis [87] (numerical experiment), it turned out to be possible to show that the time scale matching not only improves, but optimizes the information transmission.

**The three kinds of adaptation (to the mean, to the variance, and to the time scale of change of the signal) can all be related to maximization of the stimulus-response information.**

Typically one considers adaptation as a phenomenon different from redundancy reduction, and we have accepted this view. However, there is a clear relation between the two mechanisms. For example, adaptation of the first kind can be viewed as subtracting out the mean of the signal, stopping its repeated, redundant transmission and allowing to focus on the non-redundant, changing components of the signal. As any redundancy reduction procedure, this may introduce ambiguities: a perfectly adapting system will respond in the same fashion to different stimuli, preventing unambiguous identification of the stimulus based on the instantaneous response. Knowing statistics of responses on the scale of adaptation itself may be required to resolve the problem. This interesting complication has been explored in a few model systems [13, 85].

### C. Mechanisms of Different Adaptations

The three kinds of adaptation we consider here can all be derived from the same principle of optimizing the stimulus-response mutual information, and evolution can achieve all of them. However, the mechanisms behind these adaptations on physiological, non evolutionary time scales and their mathematical descriptions can be substantially different, as we describe below.

The adaptation of the first kind has been studied extensively. On physiological scales, it is implemented typically using negative feedback loops or incoherent feedforward loops, as illustrated in Fig. 3. In all of these cases, the fast activation of the response by the signal is then followed by a delayed suppression mediated by a memory node. This allows the system to transmit changes in the signal, and yet to desensitize and return close (and sometimes perfectly close) to the original state if the same excitation persists. This response to *changes* in the signal earns adaptation of the first kind the name of *differentiating filter*. In particular, the feedback loop in *E. coli* chemotaxis [29, 77] or yeast signaling [88] can be represented as the feedback topologies in the figure (see Chapter 15), and different models of *Dictyostelium* adaptation include both feedforward and feedback designs [68, 89].

The different network topologies have different sensitivities to changes in the internal parameters, different tradeoffs between the sensitivity to the stimulus change and the quality of adaptation, and so on. However, fundamentally they are similar to each other. This can be seen by noting that since the goal of these adaptive system is to keep the signal within the *small* transition region between the minimum and the maximum activation of the response, it makes sense to linearize the dynamics of the networks near the mean values of the signal and

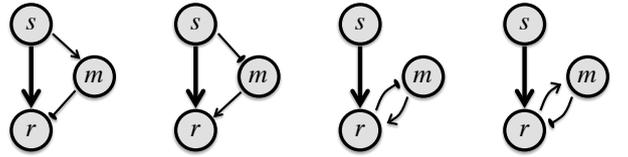


FIG. 3: Different network topologies able to perform adaptation to the mean. The nodes are labeled:  $s$  - signal,  $r$  - response, and  $m$  - memory. Sharp arrows indicate activation/excitation and blunt ones stand for deactivation/suppression. The thickness of arrows denotes the speed of action (faster action for thicker arrows).

the corresponding response. Defining  $\xi = s - \bar{s}$ ,  $\zeta = r - \bar{r}$ , and  $\chi = m - \bar{m}$ , one can write, for example, for the the feedback topologies in Fig. 3

$$\frac{d\zeta}{dt} = -k_{\zeta\zeta}\zeta + k_{\zeta\xi}\xi - k_{\zeta\chi}\chi + \eta_{\zeta}. \quad (21)$$

$$\frac{d\chi}{dt} = k_{\chi\zeta}\zeta - k_{\chi\chi}\chi + \eta_{\chi}, \quad (22)$$

where  $\eta_{\cdot}$  are noises, and the coefficients  $k_{**}$  are positive for the fourth topology, and some of them change their signs for the third. Doing the usual Fourier transform of these equations (see Ref. [71] for a very clear, pedagogical treatment) and expressing  $\zeta$  in terms of  $\xi$ ,  $\eta_{\zeta}$ , and  $\eta_{\chi}$ , we see that it is only the product of  $k_{\chi\zeta}k_{\zeta\chi}$  that matters for the properties of the filter, Eq. (21, 22). Hence both the feedback topologies in Fig. 3 are essentially equivalent in this regime. Furthermore, as argued in [68, 90], a simple linear transformation of  $\zeta$  and  $\chi$  allows to recast the incoherent feedforward loops (the two first topologies in Fig. 3) into a feedback design, again arguing that, at least in the linear regime, the differences among all of these organizations are rather small from the mathematical point of view.[103]

The reason why we can make so much progress in the analysis of adaptation to the mean is that the mean is a linear function of the signal, and hence it can be accounted for in a linear approximation. Different network topologies differ in their actuation components (that is, how the measured mean is then fed back into changing the response generation), but averaging a linear function of the signal over a certain time scale is the common description of the sensing component of essentially all adaptive mechanisms of the first kind.

**Adaptation to the mean can be analyzed linearly, and many different designs become similar in this regime.**

Variance and time scale adaptations are fundamentally different. While the actuation part for them is not any more difficult than for adaptation to the mean, adapting to the variance requires averaging the square or another nonlinear function of the signal to sense its current variance, and estimation of the time scale of the

signal requires estimation of the spectrum or of the correlation function (both are bilinear averages). Therefore, maybe it is not surprising that the literature on mathematical modeling of mechanisms of these types of adaptation is rather scarce. While functional models corresponding to a bank of filters or estimators of environmental parameters operating at different time scales can account for most of the experimentally observed data about changes in the gain and in the scale of temporal integration [3, 11, 13, 85, 91], to our knowledge, these models largely have not been related to non-evolutionary, mechanistic processes at molecular and cellular scales that underlie them.

The largest inroads in this direction have been achieved when integration of a nonlinear function of a signal results in an adaptation response that depends not just on the mean, but also on higher order cumulants of the signal, effectively mixing different kinds of adaptation together. This may be desirable in the cases of photoreception [71] and chemosensing [80], where the signal mean is unalienably connected to the signal or the noise variances (e.g. the standard deviation of brightness of a visual scene scales linearly with the background illumination, while the noise in the molecular concentration is proportional to the square root of the latter). Similarly, mixing means and variances allows the budding yeast to respond to *fractional* rather than additive changes of a pheromone concentration [92]. In other situations, like adaptation by a receptor with state-dependent inactivation properties, similar mixing of the mean signal with its temporal correlation properties to form an adaptive response may not serve an obvious purpose [73].

**We know very little about physiological mechanisms of adaptation of the second and the third kind.**

In a similar manner, integration of a strongly nonlinear function of a signal may allow a system to respond to signals in a gain-insensitive fashion, effectively adapting to the variance without a true adaptation. Specifically, one can threshold the stimulus around its mean value and then integrate it to count how long it has remained positive. For any temporally correlated stimulus, the time since the last mean-value crossing is correlated to the instantaneous stimulus value (it takes long time to reach high stimulus values), and this correlation is independent of the gain. It has been argued that adaptation to the variance in fly motion estimation can be explained at least in part by this non-adaptive process [81]. Similar mechanisms are easy to implement in molecular signaling systems as well [93].

#### IV. WHAT'S NEXT?

It is clear beyond that information theory has an important role in biology. It is a mathematically correct construction for analysis of signal processing systems. It

provides a general framework to recast adaptive processes on scales from evolutionary to physiological in terms of a (constrained) optimization problem. Sometimes it even makes (correct!) predictions about responses of living systems following exposure to various signals.

So, what's next for information theory in the study of signal processing in living systems?

The first, and the most important problem that still remains to be solved is that many of the stories we mentioned above are incomplete. Since we never know for sure which specific aspect of the world,  $e(t)$ , an organism cares about, and the statistics of signals are hard to measure in the real world, an adaptation that seems to optimize  $I[S; R]$  may be an artifact of our choice of  $S$  and of assumptions about  $P(s)$ , but not a consequence of the quest for optimality by an organism. For example, the time scale of filtering in *E. coli* chemotaxis [87] may be driven by the information optimization, or it may be a function of very different pressures. Similarly, a few standard deviations mismatch between the cumulative distribution of light intensities and a photoreceptor response curve in fly vision [8] can be a sign of an imperfect experiment, or it can mean that we simply got (almost) lucky, and the two curves nearly matched by chance. It is difficult to make conclusions based on one data point!

Therefore, to complete these and similar stories, the information arguments must be used to make predictions about adaptations in novel environments, and such adaptations must be observed experimentally. This has been done in some contexts in neuroscience [2, 11, 13, 94], but molecular sensing lags behind. This is largely because evolutionary adaptation, too slow to observe, is expected to play a major role here, and because careful control of dynamic environments, or characterization of statistical properties of naturally occurring environments [83] needed for such experiments is not easy. New experimental techniques, such as microfluidics [95] and artificially sped up evolution [96] are about to solve these problems, opening the proverbial doors wide open for a new class of experiments.

The second important research direction, which will require combined progress in experimental techniques and mathematical foundations, is likely going to be the return of dynamics. This has had a revolutionary effect in neuroscience [10], revealing responses unimaginable for quasi-steady-state stimuli, and dynamical stimulation is starting to take off in molecular systems as well [64, 97]. How good are living systems in filtering out those aspects of their time-dependent signals that are not predictive and are, therefore, of no use? What is the evolutionary growth bound when signals change in a continuous, predictive fashion? None of these questions have been touched yet, whether theoretically or experimentally.

Finally, we need to start building mechanistic models of adaption in living systems that are more complex than a simple subtraction of the mean. How are the amazing adaptive behaviors of the second and the third kind achieved in practice on physiological scales? Does it even

make sense to distinguish the three different adaptations, or can some molecular or neural circuits achieve them all? How many and which parameters of the signal do neural and molecular circuits estimate and how? Some of these questions may be answered if one is capable of probing the subjects with high frequency, controlled signals [98], and the recent technological advances will be a gamechanger as well.

Overall, studying biological information processing over the next ten years will be an exciting pastime!

## Acknowledgments

I am grateful to Michael E. Wall for asking me to write this book chapter. I would like to thank Sorin Tanase Nicola, and H. G. E. Hentschel for insightful comments about the manuscript.

- 
- [1] SKP Dayan. Acquisition in autoshaping. *Advances in Neural Information Processing Systems 12*, page 24, 2000.
- [2] CR Gallistel and J Gibbon. Time, rate, and conditioning. *Psychol Rev*, 107(2):289–344, 2000.
- [3] CR Gallistel, T Mark, A King, and P Latham. The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *J Exp Psychol: Anim Behav Process*, 27(4):354–72, 2001.
- [4] CR Gallistel, S Fairhurst, and P Balsam. The learning curve: implications of a quantitative analysis. *Proc Natl Acad Sci USA*, 101(36):13124–31, 2004.
- [5] L Sugrue, G Corrado, and W T Newsome. Matching behavior and the representation of value in the parietal cortex. *Science*, 304(5678):1782–7, 2004.
- [6] P Balsam and CR Gallistel. Temporal maps and informativeness in associative learning. *Trends Neurosci*, 32(2):73–8, 2009.
- [7] P Balsam, M Drew, and CR Gallistel. Time and associative learning. *Compar Cognition Behavior Rev*, 5:1–22, 2010.
- [8] S Laughlin. A simple coding procedure enhances a neuron’s information capacity. *Z Naturforsch, C, Biosci*, 36:910–2, 1981.
- [9] S Laughlin, R de Ruyter van Steveninck, and Anderson J. The metabolic cost of neural information. *Nat Neurosci*, 1:36–41, 1998.
- [10] F Rieke, D Warland, R de Ruyter van Steveninck, and W Bialek. *Spikes: Exploring the Neural Code*. MIT Press, 1999.
- [11] N Brenner, W Bialek, and R de Ruyter van Steveninck. Adaptive rescaling optimizes information transmission. *Neuron*, 26:695, 2000.
- [12] P Reinagel and R Reid. Temporal coding of visual information in the thalamus. *J Neurosci*, 20:5392–400, 2000.
- [13] A Fairhall, G Lewen, W Bialek, and R de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412:787, 2001.
- [14] R Liu, S Tzonev, S Rebrik, and K Miller. Variability and information in a neural code of the cat lateral geniculate nucleus. *J Neurophysiol*, 86:2789–806, 2001.
- [15] J Victor. Binless strategies for estimation of information from neural data. *Phys Rev E*, 66:51903, 2002.
- [16] I Nemenman, GD Lewen, W Bialek, and R de Ruyter van Steveninck. Neural coding of natural stimuli: information at sub-millisecond resolution. *PLoS Comput Biol*, 4(3):e1000025, 2008.
- [17] S Forrest and S Hofmeyr. Immunology as information processing. In L Segel and I Cohen, editors, *Design Principles for the Immune System and Other Distributed Autonomous Systems*. Oxford UP, 2000.
- [18] A Arkin. Signal processing by biochemical reaction networks. In J Walleczek, editor, *Self-Organized Biological Dynamics and Nonlinear Control: Toward Understanding Complexity, Chaos and Emergent Function in Living Systems*. Cambridge UP, 2000.
- [19] M Samoilov, A Arkin, and J Ross. Signal processing by simple chemical systems. *J Phys Chem A*, 106:10205–21, 2002.
- [20] B Andrews and P Iglesias. An information-theoretic characterization of the optimal gradient sensing response of cells. *PLoS Comput Biol*, 3(8):e153, 2007.
- [21] E Ziv, I Nemenman, and C H Wiggins. Optimal signal processing in small stochastic biochemical networks. *PLoS One*, 2(10):e1077, 2007.
- [22] G Tkacik, C Callan, and W Bialek. Information flow and optimization in transcriptional regulation. *Proc Natl Acad Sci USA*, 105:12265–70, 2008.
- [23] F Tostevin and P R ten Wolde. Mutual information between input and output trajectories of biochemical networks. *Phys Rev Lett*, 102:218101, 2009.
- [24] A Celani and M Vergassola. Bacterial strategies for chemotaxis response. *Proc Natl Acad Sci USA*, 107:1391–6, 2010.
- [25] N Wingreen. Why are chemotaxis receptors clustered but other receptors aren’t? In *The Fourth International q-bio Conference on Cellular Information Processing*. Center for Nonlinear Studies, LANL, Santa Fe, NM, 2010.
- [26] S Still. Information-theoretic approach to interactive learning. *EPL-Europhys Lett*, 85:28005, 2009.
- [27] J Paulsson. Summing up the noise in gene networks. *Nature*, 427:415–8, 2004.
- [28] P Dayan and L Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, 2005.
- [29] H Berg. *E. coli in motion*. Springer, 2004.
- [30] CE Shannon and W Weaver. *A mathematical theory of communication*. University of Illinois Press, Urbana, 1949.
- [31] T Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.
- [32] S Strong, R Koberle, R de Ruyter van Steveninck, and W Bialek. Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80(1):197–200, 1998.
- [33] J Kelly. A new interpretation of information rate. *Information Theory, IRE Transactions on*, 2:185–9, 1956.

- [34] C Bergstrom and M Lachmann. Shannon information and biological fitness. In *Information Theory Workshop, 2004*, pages 50–4. IEEE, 2005.
- [35] E Kussell and S Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, 309:2075–8, 2005.
- [36] M Donaldson-Matasci, M Lachmann, and C Bergstrom. Phenotypic diversity as an adaptation to environmental uncertainty. *Evo Ecology Res*, 10:493–515, 2008.
- [37] J Seger and H Brockmann. *What is bet hedging?*, volume 4. Oxford UP, 1987.
- [38] N Balaban, J Merrin, R Chait, L Kowalik, and S Leibler. Bacterial persistence as a phenotypic switch. *Science*, 305:1622–5, 2004.
- [39] E Kussell, R Kishony, N Balaban, and S Leibler. Bacterial persistence: a model of survival in changing environments. *Genetics*, 169:1807–14, 2005.
- [40] H Maamar, A Raj, and D Dubnau. Noise in gene expression determines cell fate in bacillus subtilis. *Science*, 317:526–9, 2007.
- [41] T Cagatay, M Turcotte, M Elowitz, J Garcia-Ojalvo, and G Süel. Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell*, 139:512–22, 2009.
- [42] CS Wylie, A Trout, D Kessler, and H Levine. Optimal strategy for competence differentiation in bacteria. *PLoS Genet*, 6:e1001108, 2010.
- [43] N Bowen, L Walker, L Matyunina, S Logani, K Totten, B Benigno, and J McDonald. Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells. *BMC Med Genomics*, 2:71, 2009.
- [44] A Arkin and J Ross. Statistical construction of chemical reaction mechanisms from measured time-series. *J Phys Chem*, 99:9709, 1995.
- [45] W Bialek, I Nemenman, and N Tishby. Predictability, complexity, and learning. *Neural Comput.*, 13:2409–63, 2001.
- [46] F Creutzig, A Globerson, and N Tishby. Past-future information bottleneck in dynamical systems. *Phys Rev E*, 79:041925, 2009.
- [47] M Srinivasan, S Laughlin, and A Dubs. Predictive coding: a fresh view of inhibition in retina. *Proc Royal Soc London B*, 216:427–59, 1982.
- [48] G Schwartz, R Harris, D Shrom, and MJ Berry. Detection and prediction of periodic patterns by the retina. *Nat Neurosci.*, 10:552–4, 2007.
- [49] T Hosoya, S Baccus, and M Meister. Dynamic predictive coding by the retina. *Nature*, 436:71–77, 2005.
- [50] M Vergassola, E Villermaux, and B Shraiman. ‘info-taxis’ as a strategy for searching without gradients. *Nature*, 445:406–9, 2007.
- [51] I Tagkopoulos, Y-C Liu, and S Tavazoie. Predictive behavior within microbial genetic networks. *Science*, 320:1313–7, 2008.
- [52] A Mitchell, G Romano, B Groisman, A Yona, E Dekel, M Kupiec, O Dahan, and Y Pilpel. Adaptive prediction of environmental changes by microorganisms. *Nature*, 460(7252):220–4, 2009.
- [53] I Nemenman, F Shafee, and W Bialek. Entropy and inference, revisited. *Advances in neural information processing systems*, 2002.
- [54] L Paninski. Estimation of entropy and mutual information. *Neural computation*, 15:1191–253, 2003.
- [55] N Tishby, F Pereira, and W Bialek. The information bottleneck method. In B Hajek and RS Sreenivas, editors, *Proc 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–77. U Illinois, 1999.
- [56] N Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, 1964.
- [57] H Barlow. Sensory mechanisms, the reduction of redundancy, and intelligence. In D Blake and A Utley, editors, *Proc Symp Mechanization of Thought Processes*, volume 2, page 53774. HM Stationery Office, London, 1959.
- [58] H Barlow. Possible principles underlying the transformation of sensory messages. In W Rosenblith, editor, *Sensory Communication*, page 21734. MIT Press, Cambridge, MA, 1961.
- [59] J Atick and A Redlich. Toward a theory of early visual processing. *Neural Comp*, 2:30820, 1990.
- [60] C Marshall. Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell*, 80:17985, 1995.
- [61] W Sabbagh Jr, L Flatauer, A Bardwell, and L Bardwell. Specificity of map kinase signaling in yeast differentiation involves transient versus sustained mapk activation. *Molecular Cell*, 8:68391, 2001.
- [62] G Lahav, N Rosenfeld, A Sigal, N Geva-Zatorsky, A Levine, M Elowitz, and U Alon. Dynamics of the p53-mdm2 feedback loop in individual cells. *Nat Genet*, 36:147–50, 2004.
- [63] L Ma, J Wagner, J Rice, W Hu, A Levine, and G Stolovitzky. A plausible model for the digital response of p53 to dna damage. *Proc Natl Acad Sci USA*, 102:14266–71, 2005.
- [64] P Hersen, M McClean, L Mahadevan, and S Ramanathan. Signal processing by the hog map kinase pathway. *Proc Natl Acad Sci USA*, 105:7165–70, 2008.
- [65] C-Y Huang and J Ferrel. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci USA*, 93:10078, 1996.
- [66] N Markevich, J Hock, and B Kholodenko. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J Cell Biol*, 164:354, 2004.
- [67] A Mugler, E Ziv, I Nemenman, and C Wiggins. Quantifying evolvability in small biological networks. *IET Syst Biol*, 3:379–87, 2009.
- [68] P Iglesias. Feedback control in intracellular signaling pathways: Regulating chemotaxis in dictyostelium discoideum. *European J Control*, 9:227–36, 2003.
- [69] R Normann and I Perlman. The effects of background illumination on the photoresponses of red and green cells. *J. Physiol.*, 286:491, 1979.
- [70] D MacGlashan, S Lavens-Phillips, and K Miura. Ige-mediated desensitization in human basophils and mast cells. *Front Biosci*, 3:d746–56, 1998.
- [71] P Detwiler, S Ramanathan, A Sengupta, and B Shraiman. Engineering aspects of enzymatic signal transduction: Photoreceptors in the retina. *Biophys J*, 79:2801, 2000.
- [72] C Rao, J Kirby, and A Arkin. Design and diversity in bacterial chemotaxis: a comparative study in escherichia coli and bacillus subtilis. *PLoS Biol*, 2:E49, 2004.
- [73] T Friedlander and N Brenner. Adaptive response by

- state-dependent inactivation. *Proc Natl Acad Sci USA*, 106:22558–63, 2009.
- [74] D Muzzey, C Gomez-Uribe, J Mettetal, and van Oudenaarden A. A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell*, 138:160–71, 2009.
- [75] A Anishkin and S Sukharev. State-stabilizing interactions in bacterial mechanosensitive channel gating and adaptation. *J Biol Chem*, 284:19153–7, 2009.
- [76] V Belyy, K Kamaraju, B Akitake, A Anishkin, and S Sukharev. Adaptive behavior of bacterial mechanosensitive channels is coupled to membrane mechanics. *J Gen Physiol*, 135:641–52, 2010.
- [77] U Alon, M Surette, N Barkai, and S Leibler. Robustness in bacterial chemotaxis. *Nature*, 397:168–71, 1999.
- [78] C Hansen, R Enders, and N Wingreen. Chemotaxis in escherichia coli: a molecular model for robust precise adaptation. *PLoS Comput Biol*, 4:e1, 2008.
- [79] H Salis, E Mirsky, and C Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotech*, 27:946–50, 2009.
- [80] R Endres, O Oleksiuk, C Hansen, Y Meir, V Sourjik, and N Wingreen. Variable sizes of escherichia coli chemoreceptor signaling teams. *Mol Syst Biol*, 4:211, 2008.
- [81] A Borst, V Flanagan, and H Sompolinsky. Adaptation without parameter change: Dynamic gain control in motion detection. *Proc Natl Acad Sci USA*, 102:6172, 2005.
- [82] N Kashtan and U Alon. Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci USA*, 102:13773–8, 2005.
- [83] P Reinagel and S Laughlin. Editorial: Natural stimulus statistics. *Network: Comput Neura Syst*, 12:237–40, 2001.
- [84] D Baylor and A Hodgkin. Changes in time scale and sensitivity in turtle photoreceptors. *J Physiol (Lond)*, 242:729–58, 1974.
- [85] B Wark, A Fairhall, and F Rieke. Timescales of inference in visual adaptation. *Neuron*, 61:750–61, 2009.
- [86] E Knudsen. Instructed learning in the auditory localization pathway of the barn owl. *Nature*, 417:322–8, 2002.
- [87] B Andrews, T-M Yi, and P Iglesias. Optimal noise filtering in the chemotactic response of escherichia coli. *PLoS Comput Biol*, 2(11):e154, Nov 2006.
- [88] RC Yu, CG Pesce, A ColmanLerner, L Lok, D Pincus, E Serra, M Holl, K Benjamin, A Gordon, and R Brent. Negative feedback that improves information transmission in yeast signalling. *Nature*, 456:75561, 2008.
- [89] L Yang and P Iglesias. Positive feedback may cause the biphasic response observed in the chemoattractant-induced response of dictyostelium cells. *Syst Control Lett*, 55:329–37, 2006.
- [90] E Sontag. Remarks on feedforward circuits, adaptation, and pulse memory. *IET Syst Biol*, 4:39–51, 2010.
- [91] M DeWeese and A Zador. Asymmetric dynamics in optimal variance adaptation. *Neural Comp*, 10:1179–202, 1998.
- [92] S Paliwal, P Iglesias, K Campbell, Z Hilioti, A Groisman, and A Levchenko. Mapk-mediated bimodal gene expression and adaptive gradient sensing in yeast. *Nature*, 446(7131):46–51, 2007.
- [93] I Nemenman. Gain control in molecular information processing: Lessons from neuroscience. *Physical Biol*, 2010. Submitted.
- [94] I Witten, E Knudsen, and H Sompolinsky. A hebbian learning rule mediates asymmetric plasticity in aligning sensory representations. *J Neurophysiol*, 100:1067–79, 2008.
- [95] J Melin and S Quake. Microfluidic large-scale integration: the evolution of design rules for biological automation. *Annu Rev Biophys Biomol Struct*, 36:213–31, 2007.
- [96] F Poelwijk, D Kiviet, D Weinreich, and S Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445:383–6, 2007.
- [97] J Mettetal, D Muzzey, C Gomez-Uribe, and A van Oudenaarden. The frequency dependence of osmo-adaptation in saccharomyces cerevisiae. *Science*, 319:482–4, 2008.
- [98] I Nemenman. Fluctuation-dissipation theorem and models of learning. *Neural Comp*, 17:2006–33, 2005.
- [99] JW Gibbs. *The Scientific Papers of J. Willard Gibbs*, volume 1. Ox Bow Press, 1993.
- [100] S Frank. Natural selection maximizes fisher information. *J Evolutionary Biol.*, 22:231–44, 2009.
- [101] From these properties, it is clear that mutual information is in some sense a “nicer”, more fundamental quantity. Indeed, even the famous Gibbs paradox in statistical physics [99] is related to the fact that entropy of continuous variables is ill-defined. Therefore, it is a pity that standard theoretical developments make entropy a primary quantity and derive mutual information from it. We believe that it should be possible to develop an alternative formulation of information theory with mutual information as the primary concept, without introducing entropy at all.
- [102] Interestingly, it was recently argued [100] that natural selection, indeed, serves to maximize the information that a population has about its environment, providing yet another evidence for the importance of information-theoretic considerations in biology.
- [103] Sontag has recently considered the case where the linear term in the feedforward and/or feedback interactions is zero, and the leading coupling term in the dynamics of  $\zeta$  is bilinear; there the distinctions among the topologies are somewhat more tangible [90].