

4.3 Does biology care about bits?

Section last updated
April 28, 2009.

The question for this section has been with us almost since Shannon's original work. The usual view of information theory is as a theory for communication, with its most sophisticated developments in the context of error correcting codes, but this misses much that is of relevance to the natural (as opposed to the engineered) world. In this section we'll review old ideas about the connection of information to gambling, and see how closely related ideas have reappeared in thinking about the life strategies of bacterial populations. Then we'll step back and try to look more generally at the connections among information, biological function and evolutionary fitness, and argue that evolution really can select for biological mechanisms that are efficient in an information theoretic sense. In following this path, we will wrestle with the question of when information is relevant, and how bits relate to perhaps more familiar metric notions of accuracy or performance. As in other sections of the course, we will see these ideas in many different contexts, from protein structure to gene expression in bacteria to sensorimotor behavior in primates.

To start, consider the following game.⁶ I will flip a coin, and you bet on whether it will come up heads or tails. If you get it right, I double your money. If you're wrong, you lose what you bet. If this is a fair coin, so that heads and tails each come up half the time, there really isn't anything to analyze, what happens is "just chance." But if you know, for example, that this is a biased coin, and that the probability of heads really is 60%, you might be tempted to put all of your money on heads. On average, if you bet one dollar you will receive $2 \times (0.6) = 1.2$ dollars in return, which sounds good. Indeed, if we play only once then this is what you should do if you want to maximize your expected return.

But what happens if we are going to play repeatedly, which you might think is a better metaphor for life? Now if you put all your money on heads, there is a 40% chance that, in one flip, you'll lose it all. Suppose that instead you put a fraction f of your money on heads and a fraction $1 - f$ on tails. If we introduce a binary variable $n = 1$ for heads and $n = 0$ for tails, then after one flip your winnings change by a factor

$$G = 2 \times [fn + (1 - f)(1 - n)]. \quad (4.86)$$

More generally, on the i^{th} flip your winnings will change by a factor

$$G_i = 2 \times [fn_i + (1 - f)(1 - n_i)], \quad (4.87)$$

⁶This may seem like a strange topic for a physics course, but please bear with me!

where n_i marks what happens on the i^{th} flip: heads on the i^{th} flip, $n_i = 1$, and tails on the i^{th} flip, $n_i = 0$. Then after N successive flips you will have a gain

$$G_{\text{total}}(N) = 2^N \prod_{i=1}^N [fn_i + (1-f)(1-n_i)], \quad (4.88)$$

where we are assuming that you consistently put a fraction f of your accumulated winnings down as a bet on heads, and the remaining fraction on tails.

To keep going, we want to write the product in Eq (4.88) as the exponential of a sum. It's useful to notice that, because n_i is either 0 or 1, we have

$$fn_i + (1-f)(1-n_i) = \exp [n_i \ln(f) + (1-n_i) \ln(1-f)]. \quad (4.89)$$

This means that we can write the total gain

$$\begin{aligned} G_{\text{total}}(N) &= 2^N \prod_{i=1}^N [fn_i + (1-f)(1-n_i)] \\ &= 2^N \prod_{i=1}^N \exp [n_i \ln(f) + (1-n_i) \ln(1-f)] \end{aligned} \quad (4.90)$$

$$= e^{N\Lambda}, \quad (4.91)$$

where

$$N\Lambda = N \ln 2 + \sum_{i=1}^N [n_i \ln(f) + (1-n_i) \ln(1-f)] \quad (4.92)$$

Written this way, Λ define a rate of exponential growth fro your winnings. But Λ depends not only on your betting strategy, summarized by the fraction f that you put on heads, but also on the sequence of heads and tails that come up in the game. The key point is that, if we play *many* times, so we can think about the limit $N \rightarrow \infty$, this dependence on the details of the flips goes away.

We recall that, for any well behaved random variable, the average over N observations must approach the mean computed from the probability distribution as N becomes large. In the present case, if n_i is a binary variable that takes the value $n_i = 1$ with probability p and $n_i = 0$ with probability $1-p$, then as N becomes large we should have

$$\frac{1}{N} \sum_{i=1}^N n_i \rightarrow p, \quad (4.93)$$

and similarly

$$\frac{1}{N} \sum_{i=1}^N (1 - n_i) \rightarrow 1 - p. \quad (4.94)$$

We can use this to evaluate the long term growth of your winnings, simplifying the results of Eq (4.92):

$$\frac{1}{N} \ln G_{\text{total}}(N) \equiv \Lambda(f) \quad (4.95)$$

$$= \frac{1}{N} \left\{ N \ln 2 + \sum_{i=1}^N [n_i \ln(f) + (1 - n_i) \ln(1 - f)] \right\} \quad (4.96)$$

$$= \ln 2 + \left(\frac{1}{N} \sum_{i=1}^N n_i \right) \ln(f) + \left(\frac{1}{N} \sum_{i=1}^N (1 - n_i) \right) \ln(1 - f) \quad (4.97)$$

$$\rightarrow \ln 2 + p \ln(f) + (1 - p) \ln(1 - f), \quad (4.98)$$

where again p is the probability that the coin will come up heads. But now we see that, since the growth rate $\Lambda(f)$ has a simple dependence on f , you can maximize your winnings by choosing the correct strategy!

To maximize the growth rate $\Lambda(f)$, as usual we differentiate and set the result to zero:

$$\begin{aligned} \Lambda(f) &= \ln 2 + p \ln(f) + (1 - p) \ln(1 - f) \\ \frac{d\Lambda(f)}{df} &= p \frac{1}{f} + (1 - p)(-1) \frac{1}{1 - f}; \end{aligned} \quad (4.99)$$

$$\begin{aligned} \left. \frac{d\Lambda(f)}{df} \right|_{f=f_{\text{opt}}} &= 0 \\ \Rightarrow 0 &= p \frac{1}{f_{\text{opt}}} + (1 - p)(-1) \frac{1}{1 - f_{\text{opt}}} \end{aligned} \quad (4.100)$$

$$\frac{1 - p}{1 - f_{\text{opt}}} = \frac{p}{f_{\text{opt}}}, \quad (4.101)$$

or more simply $f_{\text{opt}} = p$. This is an interesting result: you maximize the rate at which your winnings will grow by “matching” the fraction of your resources that you bet on heads to the probability that the coin will come up heads, and similarly for tails.

Problem 46. Check that $f_{\text{opt}} = p$ is a maximum, and not a minimum, of the growth rate $\Lambda(f)$.

Problem 47. If we bet only once, then in this simple game the maximum mean payoff is obtained by betting on the most likely outcome. On the other hand, as we play many times—more precisely, in the limit that we play infinitely many times—what we have seen is that a sort of matching strategy, or “proportional gambling” maximizes the growth rate. Explore the crossover between these limits. You might start with some simple simulations, and then see if you can make analytic progress, perhaps saying something about the leading $1/N$ corrections at large N . I am leaving this deliberately vague and open ended, hoping that you will play around.

Something even more interesting happens when we evaluate the optimal growth rate, that is $\Lambda_{\text{opt}} = \Lambda(f_{\text{opt}})$:

$$\Lambda_{\text{opt}} = \Lambda(f = p) \tag{4.102}$$

$$= \ln 2 + p \ln(p) + (1 - p) \ln(1 - p) \tag{4.103}$$

$$= \ln 2 - [-p \ln(p) - (1 - p) \ln(1 - p)]. \tag{4.104}$$

These terms should be starting to look familiar. The term $\ln 2$ is of course the entropy for a binary variable (heads/tails) if you don’t know anything about what to expect, and hence the two alternatives are equally likely. In contrast, the term in brackets,

$$-p \ln(p) - (1 - p) \ln(1 - p),$$

is the entropy of a binary variable if you know that the two alternatives come up with probabilities p and $1 - p$. Thus the optimal growth rate is the difference in entropy between what might happen with an arbitrary coin and what you know will happen with this coin. In other words, *the maximum rate at which your winnings can grow in a simple gambling game is equal to the information that you have about the outcome of a single coin flip.*

This connection between information theory and gambling was discovered in the 1950s by Kelly, who was searching for some interpretation of Shannon’s work that didn’t refer to the process of communication [Kelly 1956]. Obviously what we have worked out here is a very simple and special case, and we need to do much more in order to claim that the connection is general. But before launching into this let me emphasize something about Kelly’s result. At some intuitive level, we can all agree that if we know more

about the outcome of the coin flip (or the horse race, or the stock market, or ...) then we should be able to make more money. In a very general context, Shannon proved that “know more” should be quantified by various entropy-like quantities, but it’s not obvious that the knowledge measured by Shannon’s bits is actually the useful knowledge when it comes time to make a bet. Even if bits are the right measure, the connection between information and the growth of winnings could have been much more vague; you could imagine, for example, that the growth rate is bounded by some function of the information, and that this bound might or might not be realizable with feasible strategies. In contrast to these pessimistic alternatives, Kelly showed that the maximum growth rate *is* the information, and his proof is constructive so we actually know how to achieve this maximum. This really is quite astonishing.

Let’s try to generalize what we have done. Suppose that on each trial i , there are many possible outcomes, $\mu = 1, 2, \dots, K$; we’ll write $n_i^{(\mu)} = 1$ if on the i^{th} trial the outcome is μ , and $n_i^{(\mu)} = 0$ otherwise. Further, let’s say that you bet a fraction of your assets f_μ on each of outcome μ , and if μ actually happens then each dollar bet on this outcome becomes g_μ dollars; all money bet on things that don’t happen is lost. If you need an example of this sort of multi-outcome betting game, think of a horse race in which you get something back only if you pick the winner. We’ll assume that the different outcomes occur with probability p_μ , but we won’t assume anything about the relationship between these odds and the payoffs g_μ .

Having defined all the factors, it should be clear that the gain after N trials, analogous to Eq (4.88), is

$$G_{\text{total}}(N) = \prod_{i=1}^N \left[\sum_{\mu=1}^K f_\mu g_\mu n_i^{(\mu)} \right]. \quad (4.105)$$

Now we can follow the same steps as before:

$$\ln G_{\text{total}}(N) = \sum_{i=1}^N \ln \left[\sum_{\mu=1}^K f_\mu g_\mu n_i^{(\mu)} \right] \quad (4.106)$$

$$= \sum_{i=1}^N \sum_{\mu=1}^K n_i^{(\mu)} \ln(f_\mu g_\mu) \quad (4.107)$$

$$\frac{1}{N} \ln G_{\text{total}}(N) = \sum_{\mu=1}^K \left[\frac{1}{N} \sum_{i=1}^N n_i^{(\mu)} \right] \ln(f_\mu g_\mu) \quad (4.108)$$

$$\rightarrow \Lambda(\{f_\mu\}) = \sum_{\mu=1}^K p_\mu \ln(f_\mu g_\mu). \quad (4.109)$$

We want to maximize the growth rate Λ , subject to the normalization condition that the fractions of our assets placed on all the options add up ($\sum_\mu f_\mu = 1$), so we introduce a Lagrange multiplier α and find the maximum of the function

$$\tilde{\Lambda}(\{f_\mu\}) = \sum_{\mu=1}^K p_\mu \ln(f_\mu g_\mu) - \alpha \left[\sum_{\mu=1}^K f_\mu - 1 \right]. \quad (4.110)$$

The equations for the maximum are, as usual,

$$\left. \frac{\partial \tilde{\Lambda}(\{f_\mu\})}{\partial f_\mu} \right|_{\{f_\mu\}=\{f_\mu^{\text{opt}}\}} = 0 \quad (4.111)$$

$$\Rightarrow 0 = \frac{p_\mu}{f_\mu^{\text{opt}}} - \alpha, \quad (4.112)$$

$$f_\mu^{\text{opt}} = \frac{p_\mu}{\alpha}; \quad (4.113)$$

since we want $\sum_\mu f_\mu = 1$ and we know that $\sum_\mu p_\mu = 1$, we must have $\alpha = 1$, and the answer is

$$f_\mu^{\text{opt}} = p_\mu. \quad (4.114)$$

Substituting, we find the maximum growth rate

$$\Lambda_{\text{opt}} = \sum_{\mu=1}^K p_\mu \ln(p_\mu g_\mu). \quad (4.115)$$

The first interesting thing is that we recover from the simpler heads/tails problem the idea of proportional gambling [Eq (4.114)]: you maximize the rate at which your winnings will grow by “matching” the fraction of your resources that you bet on each horse in the race to the probability that this horse will win. Strangely, this is independent of the rewards or gains as expressed in the parameters $\{g_\mu\}$.

The second point is that we can see what it means for the odds in a horse race to be truly fair. If our opponent in this game (the track operators) set the returns in inverse proportion to the probability that each horse wins, $g_\mu = 1/p_\mu$, then the maximum growth rate of our winnings, Λ_{opt} , is exactly zero.

This notion of fairness leads us to an information theoretic interpretation of Λ_{opt} . Notice that we have done our calculation on the assumption that we have perfect knowledge of the distribution $\{p_\mu\}$. Perhaps the track operators have less knowledge, and so they set the odds *as if* the distribution were something else, which we can call $\{q_\mu\}$. More generally, we can define

$$q_\mu = \frac{1}{Z} \frac{1}{g_\mu}, \quad (4.116)$$

with Z chosen so that $\sum_\mu q_\mu = 1$. If $Z = 1$, then the payoffs $\{g_\mu\}$ are fair in the distribution $\{q_\mu\}$, while if $Z < 1$ the track operators are keeping something for themselves (as they are wont to do). Then we can see that

$$\Lambda_{\text{opt}} = -\ln Z + \sum_{\mu=1}^K p_\mu \ln \left(\frac{p_\mu}{q_\mu} \right). \quad (4.117)$$

You should recognize the second term as the Kullback–Leibler divergence between the probability distributions $\mathbf{p} \equiv \{p_\mu\}$ and $\mathbf{q} \equiv \{q_\mu\}$,

go back and check that we discuss D_{KL} !

$$D_{\text{KL}}(\mathbf{p}||\mathbf{q}) \equiv \sum_{\mu=1}^K p_\mu \ln \left(\frac{p_\mu}{q_\mu} \right). \quad (4.118)$$

We recall that the KL divergence measures the cost of coding signals with the wrong distribution. Knowing the real distribution \mathbf{p} we can encode the outcomes of the horse race in a binary string using an average of $-\sum_\mu p_\mu \log_2 p_\mu$ bits of space per race, but if we think the distribution is \mathbf{q} the best we can do is $-\sum_\mu p_\mu \log_2 q_\mu$ bits per race, and the difference is the KL divergence $D_{\text{KL}}(\mathbf{p}||\mathbf{q})$. The fact that $D_{\text{KL}} > 0$ reminds us that the shortest codes are built from accurate (if possibly implicit) knowledge of the underlying probabilities.

Equation (4.117) shows us that better knowledge of the probability distribution doesn't just allow us to make shorter codes. The amount by which we can compress the data describing the sequence of winners in the horse race is exactly the amount by which our winnings can grow. More precisely, if we can build a shorter code than the one built implicitly by the track operators, then we will gain exactly in proportion to this shortening. Thus, in this context, we literally get paid for constructing more efficient representations of the data (!).

In this calculation, we have connected the growth rate of winnings to the efficiency with we can represent data. Maybe this isn't quite as compelling as a direct connection to how much information we have about the outcome

of the game, which is where we started in the case of coin flips, so let's see if we can do better. Imagine that, on each trial i , we have access to some signal x_i that tells us something about the likely outcome. More precisely, when we observe x_i , the probability that the outcome will be μ on trial i is not p_μ but rather some conditional probability $p(\mu|x_i)$; if the signals x are themselves chosen from some distribution $P(x)$, then for consistency we must have

$$p_\mu = \int dx P(x)p(\mu|x). \quad (4.119)$$

To use the extra information provided by the signal x , you will adjust your strategy to bet a fraction $f_\mu(x_i)$ on the outcome μ given that you have 'heard' x_i . How does the extra information provided by x improve your winnings?

To compute the growth of winnings in the presence of extra information, we proceed along the same lines as before, to find the analog of Eq (4.109):

$$\Lambda[\{f_\mu(x)\}] = \int dx P(x) \sum_{\mu=1}^K p(\mu|x) \ln[f_\mu(x)g_\mu]. \quad (4.120)$$

Now we need to maximize this, choosing strategies that are defined by the *functions* $f_\mu(x)$, where for each x we have the constraint that $\sum_\mu f_\mu(x) = 1$. Once again the solution to this optimization problem is proportional gambling, but now the proportions are conditioned on your knowledge, so that the analog of Eq (4.114) becomes

$$f_\mu^{\text{opt}}(x) = p(\mu|x). \quad (4.121)$$

This determines the optimal growth rate,

$$\Lambda_{\text{opt}} = \int dx P(x) \sum_{\mu=1}^K p(\mu|x) \ln[p(\mu|x)g_\mu]. \quad (4.122)$$

Problem 48. Fill in the steps leading to the derivation of $\Lambda[\{f_\mu(x)\}]$ in Eq (4.120) and the consequences of optimizing this functional, Eq's (4.121) and (4.122).

The important result is the gain in growth rate that is possible by virtue of having access to the signal x , that is the difference between Λ_{opt} in Eq (4.122) and Eq (4.115):

$$\Delta\Lambda_{\text{opt}} = \int dx P(x) \sum_{\mu=1}^K p(\mu|x) \ln[p(\mu|x)g_{\mu}] - \sum_{\mu=1}^K p_{\mu} \ln[p_{\mu}g_{\mu}] \quad (4.123)$$

$$\begin{aligned} &= \int dx P(x) \sum_{\mu=1}^K p(\mu|x) \ln[p(\mu|x)g_{\mu}] \\ &\quad - \int dx P(x) \sum_{\mu=1}^K p(\mu|x) \ln[p_{\mu}g_{\mu}] \end{aligned} \quad (4.124)$$

$$= \int dx P(x) \sum_{\mu=1}^K p(\mu|x) \ln \left[\frac{p(\mu|x)}{p_{\mu}} \right]. \quad (4.125)$$

We see that the details of the payoffs g_{μ} drop out, and that *the gain in growth rate is exactly the mutual information between the signal x and the outcomes μ .*

Once again we see that information translates directly into the (increased) rate at which capital can grow. Thus, the abstract measure of information has a clear impact on very down to earth measures of performance in a real world task. But, beyond metaphor,⁷ what does this have to do with life?

The most direct connection between life and gambling is through the phenomenon of persistence. Many bacteria have two distinct lifestyles. In one, they grow quickly in most environments, but are very susceptible to being killed by antibiotics. In the other, they grow very slowly, but survive the antibiotics. This is almost exactly the horse race—if the bacterium bets correctly, it grows, but if it bets incorrectly it dies (or grows at rates far below what is possible). Absent any direct measurements on the environment, a population of genetically identical bacteria will maximize its growth rate by a form of proportional gambling, so that even in a healthy person, not taking antibiotics, we should see that some of the resident bacteria persist in a state of slow growth and (eventual) antibiotic resistance; the fraction of bacteria in this states reflects the population’s estimate of the probability that they will encounter the hostile environment of antibiotics. We also see that gaining information about the environment opens the possibility of faster growth, in precise proportion to the information gained.

⁷Life is a gamble, etc..

In a world of two alternatives, there is not much information to gain. There are examples of bacteria that choose among a wider variety of lifestyles, and these phenomena (including the simple example of two alternatives) are called ‘phenotypic switching.’ In the approximation that for each environment there is only one phenotype which grows, phenotypic switching is exactly horse racing problem studied by Kelly. For more on these connections, see Bergstrom & Lachmann (2005) and Kussell & Leibler (2005).

The example of phenotypic switching makes a nice map back to the early work about gambling, but is perhaps still a bit too simple. Let’s try to be more general. Imagine a bacterium that lives in an environment with one kind of nutrient, but the concentration of this nutrient is fluctuating (slowly, for the moment, so we don’t have to worry about dynamics). In order to actually make use of this nutrient, the bacterium must express the relevant enzymes involved in metabolism. Let’s simplify and assume that there is one nutrient or substrate at concentration s and one relevant gene at expression level g . The bacterium will then grow at some rate $r(s, g)$ that depends both on the state of the world (s) and on its internal state (g).

The growth rate of the bacterium is a compromise between two effects. On the one hand, growth requires metabolism of the available nutrient, and so growth should be faster if there is either more nutrient or more enzyme. On the other hand, making the enzyme itself takes resources, and this should slow the growth; in the limit of small nutrient concentrations, this cost can become dominant, and growth would stop if the cell tried to make too much enzyme. This scenario is shown schematically in Fig 4.2.

need to be sure we have discussed this before!

Problem 49. In fact the schematic in Fig 4.2 is based on a simple model. Suppose that growth is precisely proportional to the rate at which the enzyme degrades the substrate. In a Michaelis–Menten kinetic scheme for the enzyme, this means that the rate of degradation (in molecules per second) will be

$$V = V_{\max} g \frac{s_{\text{free}}}{K + s_{\text{free}}}, \quad (4.126)$$

where g is the number of copies of the enzyme molecule, V_{\max} is the maximum rate at which the enzyme can run, s_{free} is the concentration of the substrate free in solution, and K is the ‘Michaelis constant’ that sets the scale for half-saturation of the enzyme. The total substrate concentration is the sum of that free in solution and bound to the enzyme,

$$s = s_{\text{free}} + \frac{1}{\Omega} g \frac{s_{\text{free}}}{K + s_{\text{free}}}, \quad (4.127)$$

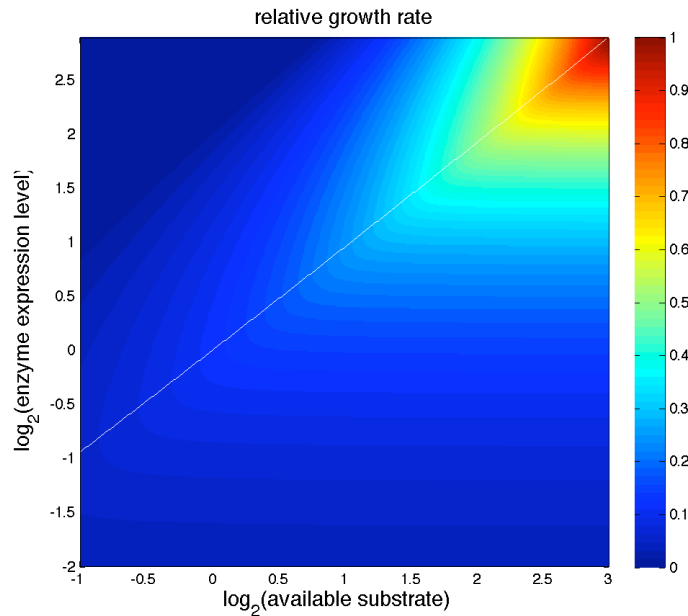


Figure 4.2: A schematic of bacterial growth rate as a function of available substrate concentration and enzyme expression level. The growth rate is a compromise between metabolizing the substrate (faster growth at both higher substrate and higher expression levels) and the cost of making the enzyme (slower growth at higher expression levels). The thin white line traces the optimal setting of expression level as a function of substrate availability.

where Ω is the cell volume. If the growth rate is proportional to the metabolic rate, less a correction for the cost of making the enzymes, we should have

$$r(s, g) = \alpha g \frac{s_{\text{free}}}{K + s_{\text{free}}} - \beta g. \quad (4.128)$$

Solve for s_{free} to rewrite $r(s, g)$ explicitly in terms of s . Then show that by proper choice of units, there is only one arbitrary parameter. What is the meaning of this remaining parameter? Make some reasonable choices, and plot your own version of Fig 4.2.

Imagine a bacterium whose life is governed by Fig 4.2. As the available substrate concentration fluctuates, one possibility is that all bacteria carefully adjust their enzyme expression levels to achieve optimal growth rate under each condition. An extreme alternative is that different bacteria in the population choose their expression levels at random out of some

distribution, and hope that some of them by chance have made very good choices, much as in the proportional gambling scenario. In the first case, the expression level carries an enormous amount of information about the concentration of available substrate—indeed, if we imagine that the optimum is traced perfectly, then knowing the expression level would tell us the exact substrate concentration, and this represents an infinite amount of information (!). In contrast, the gambling strategy involves no correlation of the internal and external states, and hence no information is conveyed. Evidently, the average growth rate across an ensemble of environment will be larger if the bacteria can adjust their expression levels perfectly, but maybe this is so obvious as not to be interesting. We know that there is some average growth rate which can be achieved with no information about the outside world, and that an infinite amount of information would allow the population to grow faster. What happens in between?

The mutual information between the internal state g and the external world s can be written as

$$I(g; s) = \int ds P(s) \int dg P(g|s) \log_2 \left[\frac{P(g|s)}{P(g)} \right]. \quad (4.129)$$

Evidently we can make $I(g; s)$ as small as we like by letting $P(g|s)$ approach $P(g)$. But suppose that we want to maintain some average growth rate in the ensemble of environments defined by $P(s)$? This average growth rate is

$$\langle r \rangle = \int ds P(s) \int dg P(g|s) r(s, g). \quad (4.130)$$

Now it seems clear that not all conditional distributions $P(g|s)$ are consistent with a given $\langle r \rangle$. To make this precise, we need to show that there is a minimum value of $I(g; s)$ consistent with $\langle r \rangle$.

The problem we have is a constrained minimization, so as usual we introduce a Lagrange multiplier and minimize

$$\mathcal{F}[P(g|s)] \equiv I(g; s) - \lambda \langle r \rangle - \int ds \mu(s) \int dg P(g|s), \quad (4.131)$$

where the second set of Lagrange multipliers $\mu(s)$ enforces normalization of the distributions $P(g|s)$ at each value of s . Finding the minimum in this case is straightforward. The key step is to evaluate the derivative of the information with respect to the conditional distribution:

$$\frac{\delta I(g; s)}{\delta P(g|s)} = \frac{\delta}{\delta P(g|s)} \int ds P(s) \int dg P(g|s) \log_2 \left[\frac{P(g|s)}{P(g)} \right] \quad (4.132)$$

$$\begin{aligned}
&= P(s) \log_2 \left[\frac{P(g|s)}{P(g)} \right] \\
&\quad + \frac{1}{\ln 2} P(s) P(g|s) \cdot \frac{1}{P(g|s)} \\
&\quad - \frac{1}{\ln 2} \int ds' P(s') P(g|s') \frac{1}{P(g)} \frac{\delta P(g)}{\delta P(g|s)} \quad (4.133)
\end{aligned}$$

$$\begin{aligned}
&= P(s) \log_2 \left[\frac{P(g|s)}{P(g)} \right] \\
&\quad + \frac{1}{\ln 2} P(s) \\
&\quad - \frac{1}{\ln 2} P(g) \frac{1}{P(g)} P(s) \quad (4.134)
\end{aligned}$$

$$= P(s) \log_2 \left[\frac{P(g|s)}{P(g)} \right], \quad (4.135)$$

which is nice because all the messy bits cancel out. Now we can solve our full problem:

$$0 = \frac{\delta \mathcal{F}[P(g|s)]}{\delta P(g|s)} \quad (4.136)$$

$$\begin{aligned}
&= \frac{\delta}{\delta P(g|s)} \left[I(g; s) - \lambda \int ds P(s) \int dg P(g|s) r(s, g) \right. \\
&\quad \left. - \int ds \mu(s) \int dg P(g|s) \right] \quad (4.137)
\end{aligned}$$

$$= P(s) \log_2 \left[\frac{P(g|s)}{P(g)} \right] - \lambda P(s) r(s, g) - \mu(s) \quad (4.138)$$

$$\log_2 \left[\frac{P(g|s)}{P(g)} \right] = \lambda r(s, g) + \frac{\mu(s)}{P(s)} \quad (4.139)$$

$$P(g|s) = \frac{1}{Z(s)} P(g) \exp \left[\tilde{\lambda} r(s, g) \right], \quad (4.140)$$

where $\tilde{\lambda} = \lambda \ln 2$, and $Z(s) = \exp[\ln 2 \mu(s) / P(s)]$ is a normalization constant,

$$Z(s) = \int dg P(g) \exp \left[\tilde{\lambda} r(s, g) \right], \quad (4.141)$$

and of course we must obey

$$P(g) = \int ds P(s) P(g|s). \quad (4.142)$$

Notice also that we can write the information and average growth rate as derivatives, much as in thermodynamics:

$$I(g; s) = \lambda \langle r \rangle - \int ds(s) \log_2 Z(s), \quad (4.143)$$

$$\langle r \rangle = \int ds(s) \frac{d \ln Z(s)}{d \tilde{\lambda}}. \quad (4.144)$$

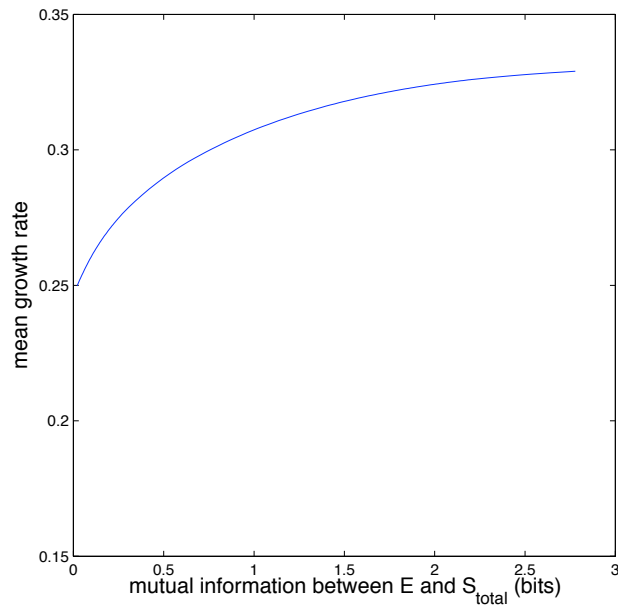


Figure 4.3: Mean growth rate as a function of the mutual information between expression levels and substrate availability for the system in Fig 4.2. We assume that the (log) substrate is chosen from a distribution that is uniform over the 16-fold range shown in Fig 4.2, and then we solve for the optimal $P(g|s)$ using Eq's (4.140–4.141).

The exponential form of the optimal solution in Eq (4.140) helps our intuition. At small λ , the distribution $P(g|s)$ is almost the same as $P(g)$, so that very little information is conveyed between internal and external states. In contrast, as $\tilde{\lambda}$ becomes large, the distribution $P(g|s)$ becomes sharply peaked around the value expression level $g_{\text{opt}}(s)$ that maximizes the growth rate. Increasing λ should trace out a curve of mean growth rate vs. information, and this is shown in Fig 4.3. Importantly, we see from the derivation that this curve represents the maximum mean growth rate achievable given a certain amount of mutual information, or alternatively

the minimum amount of information required to achieve a certain mean growth rate, $I_{\min}(\langle r \rangle)$.

Problem 50. The precise form of the relationship between the mean growth rate and the minimum information depends, of course, on details of the function $r(s, g)$. Show that the behavior at large values of the minimum information is more nearly universal. To do this, develop an asymptotic expansion at large values of λ ,

$$\begin{aligned} P(g|s) &= \frac{1}{Z(s)} P(g) \exp \left[\tilde{\lambda} r(s, g) \right] \\ &\approx \frac{1}{Z(s)} P(g) \exp \left[\tilde{\lambda} r(s, g_{\text{opt}}(s)) + \frac{\tilde{\lambda}}{2} \frac{\partial^2 r(s, g)}{\partial g^2} \Big|_{g=g_{\text{opt}}(s)} (g - g_{\text{opt}}(s))^2 \right], \end{aligned} \tag{4.145}$$

and use this expansion to evaluate $Z(s)$, from which you can calculate $I_{\min}(\langle r \rangle)$. Can you generalize your discussion to the case where there are many substrates and many genes to control?

In the problem of horse races, or phenotypic switching, information translated directly into a growth rate. Here we see that, more generally, there is a minimum amount of information needed to achieve a given average growth rate. In both of these cases, information is necessary and permissive, but not sufficient. Thus organisms *can* grow faster if they gather and represent more information, but this is not guaranteed—they might make poor use of the information, and fail to reach the bound on their growth rate.

We have focused here on achieving a certain average growth rate, but it should be clear that the whole discussion can be transposed to other domains. For example, if I ask you to point at a target that can appear at random in your visual field, and reward you in proportion to how close you come to the exact position of the target, then in order to collect a certain level of average reward your brain must represent some minimum amount of information about the target location. Quite generally, we can imagine plotting some “biological” measure of performance—probability of catching a mate, nutritional value extracted from picking fruit, growth rate, happiness, ... —versus the amount of information that the organism has about the relevant variables. This “information/fitness” plane will be divided by a curve which separates the possible from the impossible, since without a certain minimum level of information, higher fitness is impossible.

More to talk about here ... maybe for now just make a list:

- In the information theory literature, the sort of bounds we are computing here go by the name of rate–distortion curves. For example, if we measure image quality by some complicated perceptual metric, then it still is true that to have images of a certain quality, on average, we will need to transmit a minimum number of bits.
- An interesting connection of rate–distortion theory to biology is for protein structure. If I want to describe protein structures with high precision, I need to tell you where every atom is located. But if sequence determines structure, then to some accuracy I just need to tell you the amino acid sequence, which is at most $\log_2(20)$ bits per amino acid, and many fewer per atom. There is even an argument that there needs to be a compact (small number of bits) representation with high accuracy in order to make folding rapid. I am not sure how to make this rigorous, but it’s interesting.
- In constructing a rate distortion curve, we implicitly define some bits as being more relevant than others. Thus if I need to match my state to that of the environment, presumably some environmental variables need to be tracked more accurately than others; since the rate distortion curves gives the minimum number of bits, I need to get this right and put the precision (extra bits) in the right place. This is important, because it means that we have a framework for assigning value to bits.
- We can define value for bits more generally by asking for information about something in particular. For example, of all the data we collect, the only part we can use to guide our actions (and eventually collect rewards, reproduce, etc.) is the part that has predictive power, since by the time we act we are already in the future. Thus we can ask how to squeeze, out of all the bits we collect, only those bits which are relevant for prediction. Given the statistical structure of the world, we will find that we need a minimum number of bits about the past if we want to know, on average, a given number of bits about the future.
- In a very different direction, cells in a developing embryo need information about their physical position in order to do the right thing. This information is transmitted by various chemical signals, changes in the concentration of “morphogen” molecules. Presumably, making a precise spatial pattern with some level of complexity requires a minimum number of bits. Again, I don’t know exactly how to make this rigorous.