Information processing by living cells: concepts, notes and ideas

Frank J. Bruggeman*

Systems Bioinformatics, Amsterdam Institute for Molecules, Medicines & Systems, VU University, De Boelelaan 1087, 1081 HV, Amsterdam, The Netherlands

Motivation – The concept of information is central to all of cell biology. Yet, we often use the term in a vague manner in biology. Quantitative definitions of information, developed in fields outside of biology, have recently been applied to systems biology, in single-cell systems biology in particular. This document is about this development.

Background – Biologists use the term information in different contexts. The DNA of an organism is said to contain the 'information' for the synthesis of all proteins that it can express. We reason about the outputs of environment-sensing signalling networks as if they contain 'information' about the status of the environment, which is then used by the cell to express a phenotype that leads to a competitive fitness (or not). We know that intracellular, molecular control systems, such as feedback circuits, respond to particular metabolite or proteins concentrations; this 'information' they then use to carry out some control action, e.g. to (in)activate gene expression or protein activities.

Problem statement – In all these cases, the term 'information' is used in a casual, qualitative sense. This is in some cases problematic. For instance, when one aims to compare two genotypes, where one is impaired in a signalling capacity and the other not. Or, when the information processing capacity of an organism is to be quantified; for instance, to determine how accurately it can track the state of the environment. Alternatively, we may wish to answer how much information an organism's DNA encodes. Single cells experience inevitable fluctuations in their molecular make up; so one can wonder to what extent random fluctuations in signalling circuits activity distorts their information processing capacity. Finally, since cells only have limited information about their environmental and internal states, and experience information scrambling noise, cells have a finite information processing capacity; likely bounded by general principles from physics and statistical inference.

Aim – We require quantitative, measurable definitions to study the information processing by living cells to answer such questions. This document discusses some of those proposed in the scientific literature.

I. GETTING A FEELING FOR 'INFORMATION'

In everyday life we use the word information very often, in a casual manner. This is general not problematic, we have a good intuitive, common-sense understanding of its meanings.

For instance, books contain information. We can learn things or read stories. We have the impression that thicker books tend to contain more information. Information is likely additive according to our experiences.

We make predictions from information. For instance, we require the current information about local weather – pressures and

* Corresponding author: f.j.bruggeman@vu.nl

temperature – to initialise computer models of the weather in order to calculate (predict) future weather with a certain accuracy. We can improve those model predictions when we have more fine grained, improved information about local weather; say, from weather stations that are placed in 10×10 km grid instead of 100×100 km.

Any definition, or description, carries information. If I would like to describe an eating plate to you then I have to give you several pieces of information: its shape, size, color and material. Having more information about the plate, so knowing its shape and color, allows you to distinguish it better from other eating plates.

A more informative description has more discriminative value. Information therefore allows you to make predictions and infer the likelyhood of something being the thing you are looking for. With information, you can filter out alternatives.

In case of the eating plate, one could say that I have complete information if I can make an identical one, given its description alone. Being able to distinguish it from all other plates may require less information. For instance, it may have an unique scratch that, together with only its shape, makes it distinguishable from all other plates. So when one considers information that is minimally required depends a bit on the purpose of its use.

Codes carry information. Codes are like alphabets, with it you can make sentences and descriptions. With the 26 letter alphabet, I can write this text – assuming that you have an understanding of the English language.

One can think of much simpler codes, like a dice. The one shown in Figure 1 contains very simple information, six instructions. Morse code (`https://en.wikipedia.org/wiki/Morse_code`) is another code, not based on writing but on sound.

Living cells gather information about the state of their environment in order to adapt their phenotype accordingly, and attain a competitive fitness value, which increases their survival (persistence) prospects. So that they do not go extinct.

Figure 1: Six identical dice are shown, each displaying another side, so that you know all sides of a single one of them.

II. INFORMATION ENTROPY, OR SHANNON ENTROPY, AS A MEASURE FOR INFORMATION

A simple code can be made from a list of 0 and 1's and can be used to introduce the concept of information entropy. If this list is three numbers long then we can encode 8 different options, i.e. 2^3. Since this is a code, the first encoded piece of information corresponds to, for instance, 'male' or 'female', the second to 'child' or 'adult' and the third to 'club member' or 'not a club member'. An example code is,

$$\{\text{male, adult, club member}\} = \{1, 0, 1\} = 101.$$

Another perspective to uniquely identify one of the 8 options is to think of this problem as having to answer three YES/NO questions in a row.

The information entropy is denoted by H with bits as units. It is greatest if each encoded piece of information – each of the three 0 or 1's – has the highest uncertainty, which occurs if each number occurs with a probability 1/2. So, the information entropy is

linked to the probability of a certain informative message.

In our example, we have eight possible messages. The information entropy is highest when they all have equal probability $1/2 \times 1/2 \times 1/2 = \left(\frac{1}{2}\right)^3$. Since we have no prior information about the chances of 0 and 1 occurring in the message, getting to know them reveals most information.

Since we maximally can learn about three numbers then intuitively one could postulate that amount of information that we obtain is '3'. This indeed how the information entropy H works, it will turn out that it equals 3 (bits) in our example.

The information entropy is linked to the probability for the occurrence of all pieces of independent information in it – in this case each of the three numbers is a piece of information. The general equation for the information entropy – or Shannon entropy, named after the person who first defined it – is,

$$H_X = \sum_{i=1}^{N} p_i \log_2 p_i = \langle \log_2 p_i \rangle \qquad (1)$$

In this context X is random variable, which equals a certain code with a certain probability with $p_X(X = i) = p_i$.

So, in our example we have eight codes $i = 1.., 8$,

$$\mathbf{X} = \{000, 100, 010, 001, 110, 101, 011, 111\} = \{1, 2, ..., 8\},$$

and $p_X = \{p_1, p_2, ..., p_8\}$ such that

$$H_X(p_1, p_2, ..., p_8) = -\sum_{i=1}^{8} p_i \log_2 p_i$$

. If all codes are equally likely then we have,

$$H(p_1, p_2, ..., p_8) = -\sum_{i=1}^{8} \frac{1}{8} \log_2 \frac{1}{8} = \sum_{i=1}^{8} \frac{1}{8} \log_2 8 = \sum_{i=1}^{8} \frac{1}{8} \log_2 2^3 = \sum_{i=1}^{8} \frac{1}{8} 3 \log_2 2 = 3 \text{ bits}.$$

Note that the unit of information is a 'bit'. A bit equals $\log_2 2$, which equals 1.

Each number in the code therefore contributes 1 bit of information if each of them occurs with probability 1/2; so they occur with maximal uncertainty and most information is obtained when you learn them.

The information entropy is maximal when each element of X is equally likely. This can be seen from the simple relation,

$$H(p_1, p_2) = p_1 \log_2 p_1 + (1 - p_1) \log_2 (1 - p_1)$$

The optimal value of p_1, denoted by p_1^*, can be determined from,

$$\frac{d}{dp_1} H(p_1)\Big|_{p_1 = p_1^*} = 0,$$

and equals $p_1^* = \frac{1}{2}$ bits. This generalises to $H(p_1, p_2, ..., p_i..., p_n)$ which is maximal for $p_i = \frac{1}{n}$ bits. So information entropy is maximal when the codes are all maximally unlikely.

Summarising, associated with information entropy are the following statements:

1.

III. ENTROPY IN THERMODYNAMICS

A. Thermodynamic entropy

Now we shift gears and focus on entropy as it is used thermodynamics, to obtain a better understanding of what information means and how it can play a role in the natural sciences. We should do this with caution, however; entropy in thermodynamics is based on actual physical processes, while information entropy, and its use in statistics and information theory, is purely based on probability distributions. That these different disciplines use the same mathematical relation (with the same name) does not mean that those disciplines are deeply related and that the interpretations of entropy are interchangeable. The concept of information is therefore considered by some not to be applicable to thermodynamics; subtlety is needed when applying information concepts to thermodynamics. Some do it blindly, others with caution or not at all. We will take caution.

Entropy, like information, can be understood in a probabilistic manner, using probabilities. A state of a system is generally more improbable when more alternative states exist. Then, more information is required to rule one state out from all other states.

Let's assume that all states are equally probable and that only six of them exist. So, we are considering the scenario of the dice show in Figure 1. In this case, with $p_X = Prob(State = X) = 1/6$, the total number of states equals $1/p_X = 6$.

Distinguishing the state 'wash dishes' requires ruling it out from 5 alternative options, which requires less information than ruling it out from 10 alternative options. Think of the books example, two books contains more information than one and a dice with 10 sides carries more information than 6.

Information is linked to probability and probability is linked to the number of alternatives. Let's make this a bit more concrete by considering an explicit example.

Consider 2 balls, *red* and *blue*, and two boxes, $left = L$ and $right = R$. Balls end up in the boxes and a single box can contain two balls. So the system is made up out of boxes and balls. A *system state* of the system is a particular placement of its two balls, without reference to the colours of the balls. The system therefore has only three states, shown in Figure 2: i. 2 balls in L, ii. 2 balls in R, and iii. 1 in R and in L.

Note that when one considers the colours of the balls, four states exist: i. Red and Blue in L, ii. Red and Blue in R, iii. Red in L and Blue in R and iv. Blue in L and Red in R. We shall refer to these four states as the *microscopic* states, as they underlie the system states, mentioned in the previous paragraph.

The microscopic states contain more information than the macroscopic states, as the colours of the balls are considered. Sometimes the microscopic states are called microscopic realisations of a system state. A system state is sometimes called a macroscopic state. When we move from microscopic states to macroscopic states we therefore loose information. This captures by the entropy of the macroscopic state, which is therefore sometimes referred to as the 'missing information'.

Which of the system states is more probable? Is finding one ball in each box more or less probable than finding two balls in a single box? The answer is: that the system state of two balls in a single box, in L or R, is less probable as this is only achieved by one microscopic state, while a single ball in two boxes is achieved by microscopic states; the red ball can be either in the left or the right box. Here we assumed that all microscopic state are equally probable.

A macroscopic state can in principle be achieved by more than one microscopic state. A description of a macroscopic state there-

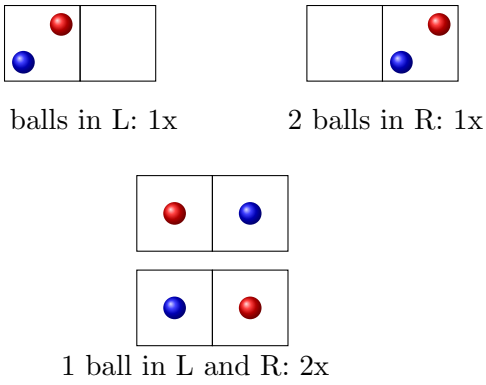

Figure 2: Number of ways to place two balls in two boxes.

fore requires less information than of a microscopic state.

In terms of the eating plate example, macroscopic states consider only shape, size and colour while the microscopic state consider those four features and also the material of which the plate is made. Macroscopic states that can be achieved by more microscopic states are more probable than those with less microscopic states associated with them – assuming that all microscopic states are equally likely.

The number of microscopic states that give rise to the same system state, or macroscopic state, is called the degeneracy of the system state, denoted by W_X for macroscopic state X.

On perspective on the degeneracy of a macroscopic state is that it is a measure of the *missing information* about its microscopic state. We miss more information about a system if the macroscopic state it is in can be achieved by more microscopic states.

Since the entropy, H_X of a macroscopic state of a system, X, is proportional to the logarithm of its degeneracy, $\ln W_X$ or $\log_Y X$ with base Y, the entropy of a macroscopic state can be linked to information. The information interpretation of entropy has a long history in physics. It is not accepted by all physicists, although its acceptance is growing.

Let's dig deeper. The probability to observe 1 ball in each box, so one state of the system, equals $p_{L=1,R=1} = \frac{\text{its microscopic realisations}}{\text{all microscopic realisations}} = \frac{2}{4}$, while having two balls in a single box, another system state, has probability $p_{L=2,R=0} = p_{L=0,R=2} = \frac{1}{4}$.

The total number of microscopic states of the system W equals the sum of the degeneracies of its macroscopic states: $W = W_{L=2,R=0}+W_{L=1,R=1}+W_{L=0,R=2} = 1+2+1 = 4$. Note that in this case the probability for an event is given by the binomial distribution,

$$p_{L=l,R=r} = \binom{l+r}{l} p_L^l (1-p_L)^r = W_{L=l,R=r} p_L^l (1-p_L)^r \tag{2}$$

With $p_L = p_R = (1-p_L) = \frac{1}{2}$; it is equally likely for a ball to be in the left or right box.

Equation 2 indicates that $p_X \propto W_X$: the most probable macroscopic state indeed has the highest degeneracy, it can be achieved by the greatest number of microscopic states.[**?**].

The second law of thermodynamics states any closed system – one without energy or mass exchange with its surroundings – will always relax to a state of equilibrium where the entropy is highest; so, that is the state with the greatest degeneracy. Below we shall see that this corresponds to the lowest energy state of the system, in the equilibrium no net work can be carried out by the system.

Note that if we would consider the colour of the balls, in addition to their placement

in the boxes, all states are equally likely, all have the same degeneracy and entropy. So, the essential aspect of the use of degeneracy, and entropy, is that we are characterising the states of a system of which we have limited information; if we knew all its microscopic states than we would have all information but we generally lack this luxury. To make this more clear let's consider the gas molecules in a volume and calculate the entropy of all its state and define its most probably state.

We consider a volume V that we partition in small volume elements with volume v, (with $v \ll V$) those volume elements are bounded in size by the mean free path of the moving gas molecules. Say we have N gas molecules that move randomly in the volume. The probability p that a single gas molecule is in a particular volume element equals $\frac{v}{V}$. The macroscopic state of the system is given by the numbers of molecules n_i in its volume elements, labelled with i, of which $\frac{V}{v} = \frac{1}{p}$ exist; so $i = 1..\frac{V}{v}$.

A macroscopic state is parameterised by the occupancy of the volume elements, $n_1, n_2, ..., n_{V/v}$ under the constraint $N = n_1 + n_2 + ... + n_{V/v}$. The degeneracy of a macroscopic state with microscopic state $n_1, n_2, ..., n_{V/v}$ equals (with $N = \sum_{i=1}^{V/v} n_i$)

$$W(n_1, n_2, ..., n_i, ..., n_{V/v}) = \frac{N!}{n_1! n_2! ..., n_i!, ...n_{V/v}!} \tag{3}$$

When we use Stirling's approximation $x! = \left(\frac{x}{e}\right)^x$, valid for large x, we obtain,

$$\begin{aligned} W(n_1, n_2, ..., n_{V/v}) &= \frac{\left(\frac{N}{e}\right)^N}{\left(\frac{n_1}{e}\right)^{n_1} \left(\frac{n_2}{e}\right)^{n_2} ... \left(\frac{n_{V/v}}{e}\right)^{n_{V/v}}} \\ &= \frac{(N)^N}{(n_1)^{n_1} (n_2)^{n_2} ... \left(n_{V/v}\right)^{n_{V/v}}} \\ &= \frac{1}{p_1^{n_1} p_2^{n_2} ... p_{V/v}^{n_{V/v}}}, \end{aligned} \tag{4}$$

with $p_i = \frac{n_i}{N}$ as the probability to find n_i molecules in the ith volume element. The entropy $S(n_1, n_2, ..., n_{V/v})$ of the macroscopic state obeys,

$$\begin{aligned} S(n_1, n_2, ..., n_{V/v}) &\propto \ln W(n_1, n_2, ..., n_{V/v}) \\ &= -\sum_{i=1}^{V/v} n_i \ln p_i \quad \Rightarrow \\ \frac{S(n_1, n_2, ..., n_{V/v})}{N} &\propto -\sum_{i=1}^{V/v} p_i \ln p_i \end{aligned} \tag{5}$$

The last relation is often a more useful relation to determine the entropy of a macroscopic state than from its degeneracy.

The second law of thermodynamics states that the entropy of a closed system is maximised in its final equilibrium state and since this applies to the system of gas molecules, we can apply this law. We therefore want to maximise $S(n_1, n_2, ..., n_{V/v})$ under the constraint $N = \sum_{i=1}^{V/v} n_i$, which we achieve by formulating the langrange function,

$$\mathcal{L}(p_1, p_2, ...p_{V/v}) = -\sum_{i=1}^{V/v} p_i \ln p_i - \lambda(\sum_{i=1}^{V/v} p_i - 1) \tag{6}$$

and setting its derivative with respect to the unknowns, the occupancies of the volume elements, to zero,

$$\forall i: \ \frac{\partial}{\partial p_i} \mathcal{L}(p_1, p_2, ...p_{V/v})\Big|_{p_i = p_i^{opt}} = -1 - \ln p_i^{opt} - \lambda = 0 \ \Rightarrow \ p_i^{opt} = e^{-\lambda - 1}. \tag{7}$$

Since $\sum_{i=1}^{V/v} p_i = 1$ we find that $\frac{V}{v}\left(e^{-\lambda-1}\right) = 1$ such that $\lambda = -\ln \frac{v}{V} - 1$ and $p_i^{opt} = \frac{v}{V}$. So

all optimal probabilities are equal and the N gas molecules are homogeneously distributed over the volume elements. This is the state with highest degeneracy and entropy. The degeneracy in this state equals $W^{opt} = \frac{1}{\left(p_1^{opt}\right)^{n_1}\left(p_2^{opt}\right)^{n_2}\ldots\left(p_{V/v}^{opt}\right)^{n_{V/v}}} = \frac{1}{\left(\frac{v}{V}^N \frac{v}{V}\right)^{\frac{V}{v}}} = \frac{1}{(p^{opt})^N}$ such that $S^{opt} \propto -N \ln p^{opt} = -N \ln \frac{v}{V}$. Alternatively we could have used that $\frac{S^{opt}}{N} \propto -\sum_{i=1}^{V/v} p_i^{opt} \ln p_i^{opt} = -\frac{V}{v}\frac{v}{V} \ln \frac{v}{V} = -\ln \frac{v}{V}$.

So the optimal degeneracy equals $W^{opt} = N \ln \frac{V}{v}$ which implies that entropy depends on the level of the coarse graining of the system. This makes absolute sense, when we allow more volume elements then more microscopic states exist – so unless the volume of the volume elements are strictly defined or the degeneracy is independent of the number of boxes, the degeneracy (and entropy) would be independent of the level of description. This has lead to some debate in physics and I am in no position to make a valid argument other than saying that the relation $W^{opt} = N \ln \frac{V}{v}$ implies that the coarse graining matters.

In terms of the concept of missing information the previous means that if we fine grain the microscopic description more, so we require more information to discern microscopic states and we keep the macroscopic state the same then the missing information at the level of the macroscopic state about the microscopic state is increased and, hence, the entropy of the macroscopic state is increased.

B. Energy and entropy, something to think about

I often notice that energy is a concept that is hard to crasp, it has an intuitive and unintuitive aspect. Consider the container with gas molecules again and assume that its temperature is that of its environment. The average speed with gas molecules move in it is only dependent on the temperature. Colissions of the molecules with the wall of the container, leads to the wall molecules moving faster, such that those in the container move slower and the temperature of the container would decrease; alternatively, if the wall is hotter than the air in the container, container molecules gain energy, and move faster, when colliding with the wall molecules. In this manner temperature differences are dissipated and thermodynamic equilibrium is maintained. (This is beautifully explained in the first chapter of Feynman's lectures on physics.)

We have to take a probabilistic perspective because temperature fluctuations occur spontaneously, because the number of molecules in a volume element spontaneously fluctuates due to random movement and collisions of gas molecules. To see this consider again N molecules in a volume V with volume elements v. The probability to find molecules in a particular element is given by,

$$Prob(N_i = n_i) = \binom{N}{n_i} \left(\frac{v}{V}\right)^{n_i} \left(1 - \frac{v}{V}\right)^{N-n_i} \tag{8}$$

The mean number of molecules equals $\langle n_i \rangle = N\frac{v}{V}$. The important thing is that the variance is not zero, $\langle \delta^2 n_i \rangle = N\frac{v}{V}(1 - \frac{v}{V}) \approx \langle n_i \rangle$ (when volume elements are small). So fluctuations in the number of molecules of volume elements occur, leading to fluctuations in the number of collisions and temperature fluctuations. And, therefore, the velocity of the gas molecules becomes a probabilistic aspect of the container – it should be treated as a ran-

dom variable. This is what Boltzmann did for the first time.

The kinetic energies that gas molecules have on average equals $k_B T$. This is the microscopic energy so to speak. Because of this energy the gas molecules move. The system has also energy, which is independent of the kinetic energy of the gas molecules, but solely dependent on their location. This is the form of energy that is more difficult to understand. With this energy the system can perform work. This energy is however a completely probabilistic understanding of energy, actually best done in terms of entropy. Energy is a confusing term in some argumentations about the system.

To make this clear, imagine an extreme scenario. Let's place all the N in the upper left corner of the box. Clearly this macroscopic state only has a single microscopic state and this macroscopic state therefore has the smallest degeneracy of 1. This state is therefore very improbably and the system will move spontaneously to macroscopic states that are more degenerate – have higher entropies, more microscopic realisations; in agreement with the second law of thermodynamics. Someone, or some system, has put energy into the system – work was performed on the system –, by placing all the molecules in one volume element.

So energy input means here to put the system in a very improbable state, move it therefore from thermodynamic equilibrium, into a non-equilibrium state. The system will now spontaneously display dynamics to a more probable macroscopic state, with higher entropies, until the macroscopic state with the highest entropy is reached and thermodynamic is (re-)established.

During this relaxation to thermodynamic equilibrium, from a non-equilibrium state, the system can perform work. To see this, imagine placing a movable wall in the middle of the container. Since all molecules were placed in the upper left corner of the container, all molecules are on the left of the movable wall and not on the right of it. When the molecules start moving, which they do independent on one another so they randomly spread over the left side of the movable wall until they start colliding with it. On the right hand side of the wall no molecules occur, so no collisions occur, and therefore a net force is exerted on the wall in the right direction. The wall will therefore move to the right, the system performs work. This is continued until the system has reached its highest entropy state, which corresponds to the movable wall completely displaced to the right of the container, touching its right wall.

The energy required to place all molecules into the upper left corner, in a single volume element, equals the total amount of information that is required to know all the positions of the gas molecules plus the energy to move (potential energy; overcoming the gravitational force). Since, $S = k_b \ln W$ and the starting, equilibrium state has $W = \left(\frac{V}{v}\right)^N$ and the final state has $W = 1$, the change in entropy equals $\Delta S = S_{final} - S_{start} = k_b \ln 1 - k_b \ln \left(\frac{V}{v}\right)^N = k_b \ln \frac{v}{V} N = N k_b \ln \frac{v}{V}$, which equals the energy to place N molecules in a single volume element of size v in the entire volume V.

So, energy input of the system means placing it in a state of lower entropy; so the energy is actually a entropic driving force. This is a different sort of energy than that associated with the movement of molecules due to their kinetic energy with is proportional to temperature. Energy is, therefore, in this case, much better understood in a probabilistic manner, using entropy. This is the understanding of entropy in thermodynamics.

C. Information entropy and thermodynamic entropy, putting it all together

It might have become a little confusing by now. Information entropy, entropy in thermodynamics, system energy and entropy,

nothing much about biology, while we were initially after the use of the information concept in biology. In particular, how well single cells can obtain information from their environment and adapt their phenotype accordingly, in order to obtain a competitive fitness and maximise their survival prospects. It turns out that what we have learned before is actually very useful for understanding the information concept in cell biology and this is how many scientists are doing it at the moment.

We will make the relations between biological information processing and entropy in words first, before we introduce a new concept 'mutual information' in the next section which is based in information entropy and captures how well single cells can process information and what this means. We will start with focussing on single receptor molecule, for which we assume to understand the kinetics; an example for which we do not have to use the concept of information entropy and then a example of a cell signalling network where we do have to invoke the information concept, because we lack complete understanding of the network kinetics.

D. A single receptor in a cell; how good can it sense?

A bacterium in some medium does not have receptors for all the chemical components in the medium. Those that it does sense, it does so only to some limited degree. To see this consider the following simple argumentations. We have already seen that variance in the number of molecules in a volume element equals the mean number of molecules. If the bulk concentration of this molecule, the ligand which will bind to the receptor, is c then this number equals $n = cv$ with v as the volume of the volume element. So the noise in the molecule number in this volume equals,

$$\frac{\langle \delta^2 n \rangle}{\langle n \rangle^2} = \frac{1}{n} = \frac{1}{cv}.$$

Those fluctuations contribute to the fluctuations in the bound state of the receptor, which become larger when the measurement volume (of the size of the catalytic site of the receptor) becomes smaller.

The receptor is the instrument that the cell uses to estimate the concentration c, it does this from the concentration of the ligand-bound receptor, which relates to c,

$$RL = \frac{c}{c + K}$$

This corresponds to the binding reaction $R + L \underset{k^- \times RL}{\overset{k^+ \times R \times c}{\rightleftharpoons}} RL$ with $K = \frac{k^-}{k^+}$. If the number of ligand molecules in the volume element, i.e. n, would not fluctuate then c would be constant and the fluctuations in RL would only be due to fluctuations in the binding and unbinding equilibrium.

So the question boils down to how well the mean value of c can be inferred by the cell from the values of RL, recorded over some time interval T, which we shall call the measurement time, taking into account that the molecular numbers in the receptor volume fluctuate.

If the measurement interval time T is finite then the estimated value of c, which we denote by $\hat{c}$, can be expected to have converged to the correct one. To understand this think of a dice, 3 throws will not lead to the average value of the dice 3.5, but a 100 or so throws would,

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} x_i = 3.5,$$

with x_i as the thrown value of the dice in the ith throw.

One way to describe the convergence of the estimated to the true value with the number of throws is to determine the variance in the estimated value $\langle \delta^2 \hat{c} \rangle$ after a certain number of throws. This captures the effect that

if you would throw a dice twice four times, then the mean value after each of the four throws would vary. This variation is captured by the variance we are talking about. (So we are not talking about the variance in c due to its stochastic fluctuations, due to the random diffusion of ligand molecules in the measurement value.) The variation in the estimated value we expect to converge to zero after infinite time.

Let's calculate the convergence of the variance in a mean estimate – the uncertainty in the measurement – to zero as function of the number of samples,

$$\langle \delta^2 \hat{x} \rangle = \left\langle \delta^2 \frac{x_1 + x_2 + ...x_i + ... + x_N}{N} \right\rangle = \frac{1}{N^2} \sum_{i=1}^{N} \langle \delta^2 x \rangle = \frac{1}{N} \langle \delta^2 x \rangle.$$

This equation tells us that the variance in the estimate after one measurement equals the variance in the variable, x, due to stochastic diffusion and kinetics effects. Taking more samples reduces the variance in the estimate in proportion to 1 over the number of samples. If x would not fluctuate than clearly no error would occur in the determination of its value from a single sample.

The estimated ligand concentration, $\hat{c}$ depends on the concentration of $\hat{RL}$ as $\hat{c} = c(\hat{RL})$ with RL being a random variable (a concentration that fluctuates). Given the relation squared error in the estimate, i.e. $\frac{\langle \delta^2 \hat{c} \rangle}{c^2}$, equals

$$\frac{\langle \delta^2 \hat{c} \rangle}{c^2} = \frac{1}{c^2} \left(\frac{\partial c}{\partial RL} \right)^2 \langle \delta^2 \hat{RL} \rangle = \frac{1}{N} \frac{1}{RL^2 (1 - RL)^2} \langle \delta^2 RL \rangle$$

Here we used the following relation:

$$c = \frac{KRL}{1 - RL} \;\Rightarrow\; \frac{\partial c}{\partial RL} = \frac{K}{(RL - 1)^2} \tag{9}$$

What remains to be determined is a physical interpretation of N; so how should we interpret the number of measurements that a receptor takes? Also we need to determine $\langle \delta^2 RL \rangle$ given the reaction $R + L \rightleftharpoons RL$. If the concentration of L is fixed, and equal to c, – which is not we assumed it now to not make it overly complicated – and we have copy of the receptor, such that $1 = R + RL$, then

$$Prob(RL = n | n = \{0, 1\}) = \begin{pmatrix} 1 \\ n \end{pmatrix} p^n (1 - p)^{1-n},$$

with $p = \frac{k^+ c}{k^+ c + k^-} = \frac{c}{c + K}$ with $K = \frac{k^-}{k^+}$. For this distribution we know that $\langle RL \rangle = np =$

$\frac{c}{c+K}$ and $\langle \delta^2 RL \rangle = np(1-p) = p(1-p) = \langle RL \rangle (1 - \langle RL \rangle) = RL(1-RL)$ (abuse of notation in last step). Given this we find for the squared error in the estimate of c,

$$\frac{\langle \delta^2 \hat{c} \rangle}{c^2} = \frac{1}{N} \frac{1}{RL(1-RL)}$$

The number of samples taken by the receptor after a measurement interval T equals $N = \frac{T}{\tau}$ with τ as the measurement time. Because the sample number concerns samples of independent events we can define the measurement time as the time that it takes for measurement volume to have re-equilibrated with its surrounding environment. After time the concentration in the measurement volume no longer correlates with the concentration a measurement time interval ago; you cannot predict this previous value from the current value, due to the intrinsic randomness of diffusion. The time scale τ of refreshment of the volume with new molecules, leading to the loss of the correlation, is given by the net loss rate of RL, given $\frac{dRL}{dt} = k^+ c(1 - RL) - k^- RL = k^+ c - (k^+ c + k^-) RL$ at steady state, $\tau = 1/(k^+ c + k^-)$. The value of k^- can be solved from $k^+ c(1 - RL) - k^- RL = 0$. We approximate the binding rate constant k^+ by a diffusion-limited rate constant, i.e. $k^+ = 4\pi D s$, with s as the reaction radius. As the final equation, we obtain,

$$\begin{aligned} \frac{\langle \delta^2 \hat{c} \rangle}{c^2} &= \frac{1}{N} \frac{1}{RL(1-RL)} = \frac{\tau}{T} \frac{1}{RL(1-RL)} = \frac{1}{T(k^+ c + k^-)} \frac{1}{RL(1-RL)} \\ &= \frac{1}{4\pi D s c (1-RL)} \frac{1}{T} = \frac{c+K}{4\pi D s c K} \frac{1}{T} \end{aligned} \tag{10}$$

The interpretation of this equation is:

-

Finally, if M receptors are contained in a cell then the squared error is reduced, because M times more independent samples are taken in time T,

$$\frac{\langle \delta^2 \hat{c} \rangle}{c^2} = \frac{1}{4\pi D s c (1-RL)} \frac{1}{TM} = \frac{c+K}{4\pi D s c K} \frac{1}{TM} \tag{11}$$

E. Characterisation of the errors associated with an input-output relation of a molecular circuit

We will view a molecular circuit from a very coarse grained perspective, only by considering the concentration of its input and output molecules. One of the strengths of the information theory approach is that we do not need to consider the exact networks structure and the kinetic properties of components, including that of their interactions.

So we are considering the coarse grained model,

$$S \xrightarrow{Network} O, \tag{12}$$

with 'S' as the signal, or input, of the network and 'O' its output.

According to intuition, if O is independent of S, changes in O and S will not correlate with each other and not information can be gained from O about S and also not from S about O. So, O carries information about S if their values correlate, the possible values that S can take reduces when you know the value of S. Given the information about O's value you can predict the value of S with greater accuracy.

If values of O and S are 1-to-1, such that for each value of S only a single value of O occurs, and vice versa, then you have complete information about S when you know the value of O, and vice versa. This for instance the case when $O(S) = 5\frac{S^3}{2+S^3}$. In this case, the relationship between O and S is deterministic, you can predict their values, given the value of the other, with probability 1.

The relation between S and O will rarely by deterministic in biology. In single cells networks will in general introduce noise such that the value of O will fluctuate at constant S. Then O may take on values at $S = s_1$ that is also displays when $S = s_2$ ($s_1 \neq s_2$), which reduces the probability that you predict the correct S value from the values of O alone. Now S and O are imperfect correlated.

If the correlation coefficient between S and O, ρ_{SO}, equals zero, no information transfer occurs from S to O by the network. If it equals 1 (or -1) then the network does therefore not cause any information loss. The 'mutual information' between S and O, the information that S carries about O and vice versa, is therefore maximal when the correlation coefficient is 1 (or -1) and zero if the correlation coefficient is zero.

The measure for the mutual information between S and O, the information that S carries about O and vice versa, we shall denote it by $I(O;S)$. If S and O carry information about each other their value then their value are related to each other by some function,

$$O(S) = f(S) + \text{noise}. \tag{13}$$

The noise reduces the correlation between S and O. The kinetics of the network determines the function $f(S)$ and the noise properties.

The equation for the mutual information $I(O;S)$ equals,

$$\begin{aligned} I(O;S) &= \sum_{o,s} p(o,s)\log_2\frac{p(o,s)}{p(o)p(s)} = H(O) - \underbrace{H(O|S)}_{\text{noise entropy}} = H(O) + H(S) - H(O,S) \geq 0 \\ &= D_{KL}(p(o,s)||p(o)p(s)) \end{aligned} \tag{14}$$

with 'H' as the information entropy defined above,

$$H_X = -\sum_i p_i \log_2 p_i.$$

$H(O)$ captures the amount of information in O, when it is higher, it can take many value and the current value of O is harder to predict accurately. $H(O|S)$ captures the information in O when you the value of S, if S and O correlate strongly then the information in O is greatly reduced when you know S. The mutual information between is therefore high when $H(O|S)$ is low relative to $H(O)$ then you have learned a lot about the value of O when you know the value of S. The joint entropy $H(O,S)$ is defined as,

$$H(O,S) = -\sum_{o,s} p(o,s)\log_2 p(o,s) \tag{15}$$

and the conditional entropy as,

$$\begin{aligned} H(O|S) &= \sum_s p(s) H(O|s) \\ &= -\sum_s p(s) \sum_o p(o|s) \log_2 p(o|s) \\ &= H(O,S) - H(S) \end{aligned} \tag{16}$$

For many applications it is useful to define the differential entropy for continuous random variables,

$$H_X = -\int f(x) \log_2 f(x) dx$$

, with $X \sim f(x)$ and $Prob(X = x) = f(x)dx$. If $X \sim NormalDistribution(\langle x \rangle, \langle \delta^2 x \rangle = \sigma_X^2) =: N(\langle x \rangle, \langle \delta^2 x \rangle)$ we find for the information entropy of this distribution,

$$\begin{aligned} H_X &= -\Big\langle \log_2 \frac{e^{\frac{-(x-\langle x \rangle)^2}{2\sigma_X^2}}}{\sqrt{2\pi\sigma_X^2}} \Big\rangle = -\Big\langle \log_2 e^{\frac{-(x-\langle x \rangle)^2}{2\sigma_X^2}} \Big\rangle + \log_2 \sqrt{2\pi\sigma_X^2} \\ &= -\Big\langle \frac{\ln 2}{\frac{-(x-\langle x \rangle)^2}{2\sigma_X^2}} \Big\rangle + \log_2 \sqrt{2\pi\sigma_X^2} = -\Big\langle \ln 2 \frac{2\sigma_X^2}{-(x-\langle x \rangle)^2} \Big\rangle + \log_2 \sqrt{2\pi\sigma_X^2} \\ &= 2\ln 2 + \log_2 \sqrt{2\pi\sigma_X^2} = \frac{\ln 2}{\ln e^{1/2}} + \log_2 \sqrt{2\pi\sigma_X^2} = \log_2 e^{1/2} + \log_2 \sqrt{2\pi\sigma_X^2} \\ &= \log_2 \sqrt{2\pi e \sigma_X^2} = \frac{1}{2} \log_2 (2\pi e \sigma_X^2). \end{aligned} \tag{17}$$

This result indicates that the entropy of a normal distribution is higher when its variance is greater. This corresponds to the intuition that the entropy of a probability distribution is higher when the random variable can take on more values. This is in agreement with what we found earlier, e.g. with a uniform discrete distribution $H_X(p_1, p_2, ..., p_N) = H_X(1/N, 1/N, ..., 1/N) = -\sum_{i=1}^N p_i \log_2 p_i = \log_2 N$. When the variance of a random variable is higher it take on more values with relatively-significant probabilities.

For continuous random variables the mutual information is defined as,

$$I(O;S) = \int\int f_{OS}(o,s) log_2 \frac{f_{OS}(o,s)}{f_S(s) f_O(o)} dods. \tag{18}$$

For instance, we consider the case that O and S are jointly distributed according to a bivariate normal distribution, i.e.

$$\{O, S\} \sim N(\{\langle s \rangle, \langle o \rangle\}, \{\sigma_o, \sigma_s, \rho\}) =: f_{OS}(o,s),$$

with ρ as the correlation coefficient between O and S defined as $\frac{\langle \delta s \delta o \rangle}{\sigma_o \sigma_s}$ with σ_o as the standard deviation (square root of the variance, $\langle \delta^2 o \rangle$). The following marginal probability distributions hold,

$$\begin{aligned} f_O(o) &= \int f_{OS}(o,s) ds = N(\langle o \rangle, \sigma_o) \\ f_S(S) &= \int f_{OS}(o,s) do = N(\langle s \rangle, \sigma_s) \end{aligned}$$

For this example the mutual information equals,

$$I(O;S) = -\frac{1}{2} \log_2 (1 - \rho^2). \tag{19}$$

Also this equation agrees with intuition O and S contain more information about each other when their correlation coefficient is high.

Another model is the following,

$$\begin{aligned} s &\sim N(\langle s \rangle, \sigma_s) \\ o &= f(s) + \epsilon \sigma_o, \ \epsilon \sim N(0,1) \end{aligned} \tag{20}$$

with

a. $f(s) = as$ such that $o \sim N(a\langle s\rangle, a^2\sigma_s^2 + \sigma_o^2)$, because $var(as) = a^2 var(s)$ then $H(o) = \frac{1}{2}\log_2(2\pi e(a^2\sigma_s^2 + \sigma_o^2)$ and $H(o|s) = \frac{1}{2}\log_2(2\pi e\sigma_o^2)$ such that,

$$I(O;S) = H(O) - H(O|S) = \frac{1}{2}\log_2\frac{2\pi e(a^2\sigma_s^2+\sigma_o^2)}{2\pi e\sigma_o^2} = \frac{1}{2}\log_2\left(1+\frac{a^2\sigma_s^2}{\sigma_o^2}\right) = \frac{1}{2}\log_2(1+SNR)$$

We defined the signal-to-noise ratio as $SNR = \frac{a^2\sigma_s^2}{\sigma_o^2}$. So if the variation in the output due to changes in S exceeds the variation in the output, given a fixed signal concentration, so $a^2\sigma_2^2 > \sigma_o^2$ then the mutual information between O and S, with high certainty the true S state can be inferred from knowing the value of O. Note is therefore low. The SNR equals $\frac{\langle\delta^2\langle O|S\rangle\rangle}{\langle\langle\delta^2 O|S\rangle\rangle}$ with $\langle\delta^2 O\rangle = \langle\delta^2\langle O|S\rangle\rangle + \langle\langle\delta^2 O|S\rangle\rangle$ with the first terms as the signal variation and the second term as the noise.

b. If

$$S \xrightarrow{\text{network 1, } x=f(S)} X(S) \xrightarrow{\text{network 2, } o=g(x)} O(S) \tag{21}$$

such that $P(O) = P(O|X)P(X|S)P(S)$ then $I(X;S) \geq I(O;S)$ which is known as the 'data-processing inequality'.

F. Mutual information maximisation

1. *Assuming negligible noise entropy (histogram equilisation)*

Understanding how the mutual information between O and S can be maximised, by varying the structure and the kinetics of molecular circuits, given suitable constraints is a relevant question for cell biology.

If the noise entropy is low, or fairly constant, then the mutual information (eq. 14) is maximised by entropy maximisation of $H(O) = -\int g_O(o)\log_2 g_O(o)do$ with $o \sim f_O(o)$. In order to do so we have to consider the constraint $\int f_O(o)do = 1$ such that we are faced with the following optimisation problem, phrased in terms of the Langrage function, $\mathcal{L}(f_O(o))$,

$$f_O^{opt}(o) = \arg\max_{f_O(o)}\left(\underbrace{-\int f_O(o)\log_2 f_O(o)do - \lambda\left(\int_0^{o_{max}} f_O(o)do - 1\right)}_{\mathcal{L}(f_O(o))}\right)$$

$$0 = \frac{d\mathcal{L}}{df_O(o)}\Big|_{f_O(o)=f_O^{opt}(o)} = \frac{\ln 2}{(\ln f_O^{opt}(o))^2} - \frac{\ln 2}{\ln f_O^{opt}(o)} - \lambda$$

$$\Rightarrow f_O^{opt}(o) = e^{\frac{\sqrt{\log(2)}\sqrt{4\lambda+\log(2)}-\log(2)}{2\lambda}}$$

$$1 = \int_0^{o_{max}} f_O^{opt}(o)do = e^{\frac{\sqrt{\log(2)}\sqrt{4\lambda+\log(2)}-\log(2)}{2\lambda}} o_{max}$$

$$\Rightarrow f_O^{opt}(o) = \frac{1}{o_{max}} \quad \text{(Continuous random variable, uniform distribution)} \tag{22}$$

So the mutual information $I(O;S)$ is maximised, when the noise entropy $H(O|S)$ is

negligible or constant, if the probability distribution of O is a uniform distribution. However, if the noise entropy is negligible, the distribution of O results from the input/output relation $o = g(s)$ and $s \sim f_S(s)$. The probability distribution of a transformed random variable, i.e. the random variable S is transformed by $g(s)$ leading to a new random variable O, can be determined from the relation,

$$f_O(o)do = f_S(s)ds \;\Rightarrow \frac{do}{ds} = \frac{f_S(s)}{f_O(o)} = \left|\frac{dg}{ds}\right| = \left|g'(s)\right| \;\Rightarrow g'(s) = f_S(s)o_{max}$$
$$g(s) = g(0) + \int_0^s g'(s)ds = g(0) + o_{max}\int_0^s f_S(x)dx \tag{23}$$

(Since $\frac{f_S(s)}{f_O(o)} > 0$ we have enforce positivity of $\frac{dg}{ds}$; at the end of the first line we drop the sign restriction since $f_S(s) > 0$ and we assume $o_{max} > 0$.) So the input/output relation that maximises the mutual information, under the current assumptions, is given by the cumulative probability distribution of the signal.

If $s \sim f(s) = N(\langle s\rangle, \sigma_S)$ then the optimal input/output relation equals,

$$o = g(s) = \frac{1}{2}o_{max}\left(Erf\Big[\frac{s-\langle s\rangle}{\sqrt{2}\sigma_S}\Big] + Erf\Big[\frac{\langle s\rangle}{\sqrt{2}\sigma_S}\Big]\right)$$

which is hard to distinguish from the sigmoidal function $g(s) = o_{max}\frac{s^n}{s^n+K_s}$ with as Hill coefficient $n = \frac{d}{d\ln s}\ln\frac{\theta(s)}{1-\theta(s)}$ (where $\theta(s) = \frac{g(s)}{o_{max}}$) and K_s is the solution of $g(K_s) = \frac{1}{2}o_{max}$.

In this approximation the mutual information equals the entropy of the output distribution.

2. *Mutual information maximisation, small noise limit*

We start by rewriting the mutual information in a useful format (using the relation $f_{OS}(o,s)dsdo = f_{O|S}(o|s)f_S(s)dsdo = f_{S|O}(s|o)f_O(o)dsdo$),

$$\begin{aligned}
I(O;S) &= \int\int f_{OS}(o,s)\log_2\frac{f_{OS}(o,s)}{f_S(s)f_O(o))}dsdo \\
&= \int\int f_{OS}(o,s)\log_2\frac{f_{OS}(o|s)f_S(s)}{f_S(s)f_O(o))}dsdo \\
&= \int\int f_{OS}(o,s)\log_2\frac{f_{OS}(o|s)}{f_O(o)}dsdo \\
&= \int\int f_{OS}(o,s)\log_2 f_{OS}(o|s)dsdo - \int\int f_{OS}(o,s)\log_2 f_O(o)dsdo \\
&= \int\int f_{O|S}(o|s)f_S(s)\log_2 f_{O|S}(o|s)dsdo - \int f_O(o)\log_2 f_O(o)do \\
&= -\int f_O(o)\log_2 f_O(o)do + \int f_S(s)\int f_{O|S}(o|s)\log_2 f_{O|S}(o|s)dods \\
&= \underbrace{-\int f_O(o)\log_2 f_O(o)do}_{\text{Entropy of output distribution, } H_O} \underbrace{-\int f_S(s)H_{O|S}ds}_{\langle H_{O|S}\rangle}
\end{aligned} \tag{24}$$

In order to determine the conditional entropy $H_{O|S}$ we define the distribution of output values at a fixed signal value,

$$O|S \sim N(g(s), \sigma_{o|s}(s)) \tag{25}$$

with $g(s)$ as the input/output relationship, for instance $g(s) = \frac{s^n}{K^n+s^n}$. The entropy of this distribution equals $H_{O|S}(o|s) = \frac{1}{2}\log_2(2\pi e\sigma_O^2(s)) = \frac{1}{2}\frac{\ln 2\pi e\sigma_O^2(s)}{\ln 2}$. Now we are going to make the small noise approximation, we going to assume that the following is true,

$$f_O(o)do = f_S(s)ds, \tag{26}$$

which means that the probability to observe S is equal to the probability of observing $O = g(s)$ so the noise in the variance in O at fixed S is negligible. So effectively we are saying the variation in S, σ_S^2 is much greater than $\sigma_{O|S}^2(s)$ such that $O(s) = g(s) + \epsilon\sigma_{O|S}(s) \approx g(s)$ with $\epsilon \sim N(0,1)$. So, in the small noise approximation, we say that o is a deterministic function of s while s is a random variable, i.e. $s \sim f_S(s)$ such that

$$\begin{aligned} f_O(o)g'(s)ds &= f_S(s)ds \\ g'(s) &= \frac{\partial g(s)}{\partial s} \\ do &= g'(s)ds \\ s &= g^{-1}(o) \end{aligned} \tag{27}$$

Now we arrive at

$$\begin{aligned} I(O;S) &= -\int f_O(o)\log_2 f_O(o)do - \int f_S(s)H_{O|S}ds \\ &= -\int f_O(o)\log_2 f_O(o)do - \int f_O(o)g'(s)H_{O|S}\frac{do}{g'(s)} \\ &= -\int f_O(o)\log_2 f_O(o)do - \int f_O(o)H_{O|S}do \\ &= -\int f_O(o)\log_2 f_O(o)do - \frac{1}{2\ln 2}\int f_O(o)\ln 2\pi\sigma_O^2(s)do \end{aligned} \tag{28}$$

Now the question is for which $f_O(o)$ is the mutual information between O and S maximised, given that $\langle o|s\rangle = g(s)$ and the noise in O is small at a fixed S, so $\frac{\sigma_{O|S}^2(s)}{\langle O|S\rangle^2} \approx 0$. We will find the optimal $f_O(o)$, denoted by $f_O^*(o)$ using the langrange multiplier method,

$$\mathcal{L} = I(O;S) - \lambda\Big(\int f_O(o)do - 1\Big), \quad \frac{\partial\mathcal{L}}{\partial f_O(o)}\Big|_{f_O(o)=f_O^*(o)} = 0 \tag{29}$$

(Note that this is a derivative wrt to a function. Likely this works the same as taking the derivative with respect to a variable.) Thus,

$$0 = \frac{\partial}{\partial f_O(o)}\Big(- \int f_O^*(o)\log_2 f_O^*(o)do - \frac{1}{2\ln 2}\int f_O^*(o)\ln 2\pi\sigma_O^2(s)do\Big)$$
$$0 = -\frac{1}{\ln 2}(\ln f_O^*(o)+1) - \frac{1}{2\ln 2}\ln[2\pi e\sigma_O^2(s)] - \lambda$$
$$f_O^*(o) = e^{-1-\frac{1}{2}\ln[2\pi e\sigma_O^2(s)]-\lambda\ln 2} = \frac{e^{-1-\lambda\ln 2}}{\sqrt{2\pi e}}\frac{1}{\sigma_O(s)} = \frac{e^{-1-\lambda\ln 2}}{\sqrt{2\pi e}}\frac{1}{\sigma_O(g^{-1}(o))} = \frac{1}{Z}\frac{1}{\sigma_O(g^{-1}(o))} \quad (30)$$

Where Z is a constant and defined such that $\int f^*(o)do = 1$ therefore,

$$1 = \int f^*(o)do = \int \frac{1}{Z}\frac{1}{\sigma_O(g^{-1}(o))}do$$
$$Z = \int \frac{1}{\sigma_O(g^{-1}(o))}do \quad (31)$$

We can also determine λ from the constraint that $\int f^*(o)do = 1$ then,

$$1 = \frac{e^{-1-\lambda\ln 2}}{\sqrt{2\pi e}}\int \frac{1}{\sigma_O(g^{-1}(o))}do$$
$$e^{-1-\lambda\ln 2} = \frac{\sqrt{2\pi e}}{\int \frac{1}{\sigma_O(g^{-1}(o))}do}$$
$$\lambda = -\frac{\ln\frac{\sqrt{2\pi e}}{\int\frac{1}{\sigma_O(g^{-1}(o))}do}}{\ln 2} - \frac{1}{\ln 2} \quad (32)$$

Now note that $\sigma_O(g^{-1}(o)) = \sigma_O(s)$ and that in the small noise approximation $\sigma_O^2(s) = g'(s)^2\sigma_S^2$ with σ_S as a constant; so, (I am not sure whether this is correct),

$$\int \frac{1}{\sigma_O(g^{-1}(o))}do =$$
$$\int \frac{1}{\sigma_O(s)}g'(s)ds =$$
$$\int \frac{1}{g'(s)\sigma_S}g'(s)ds = \frac{1}{\sigma_S} \quad (33)$$

Therefore,

$$\lambda = -\frac{\ln\frac{\sqrt{2\pi e}}{\int\frac{1}{\sigma_O(g^{-1}(o))}do}}{\ln 2} - \frac{1}{\ln 2}$$
$$= -\log_2\sqrt{2\pi e\sigma_S^2} - \frac{1}{\ln 2} \quad (34)$$

Now $f_O^*(o)$ becomes,

$$f_O^*(o) = \frac{e^{-1-\lambda\ln 2}}{\sqrt{2\pi e}}\frac{1}{\sigma_O(g^{-1}(o))}$$
$$= \frac{e^{-1-(-\log_2\sqrt{2\pi e\sigma_S^2}-\frac{1}{\ln 2})\ln 2}}{\sqrt{2\pi e}}\frac{1}{\sigma_O(g^{-1}(o))}$$
$$= \frac{\frac{1}{2}\ln(2\pi e\sigma_S^2)}{\sqrt{2\pi e}}\frac{1}{\sigma_O(g^{-1}(o))}$$
$$= \frac{\ln 2H(S)}{\sqrt{2\pi e}}\frac{1}{\sigma_O(g^{-1}(o))} \quad (35)$$

Here we assumed that $s \sim N(\langle s\rangle, \sigma_S^2)$.

By rewriting the optimal output distribution we can come up with an explanation why it is the optimal one,

$$f_O^*(o)do = \frac{1}{Z}\frac{1}{\sigma_O(g^{-1}(o))}$$
$$= \frac{1}{Z}\frac{1}{g'(s)\sigma_S}$$
$$= \frac{1}{Z\sigma_S}\frac{1}{g'(s)}do \quad (36)$$

When the slope of the input/output relation is high at some s value then the probability of the associated o should be low.

3. *Still to add*

1. channel capacity $C = \max_{\text{parameters}} I(O;S)$, a channel cannot transform more signal entropies exceeding C
2. rate distortion function, $C = \max I(O;S|D \le D^*)$
3. mutual information maximisation, analytical results

IV. THE FITNESS VALUE OF INFORMATION; HAVING HIGHER MUTUAL INFORMATION IS OFTEN SELECTED FOR

V. FINAL REMARKS AND THOUGHTS