# BE 159: Signal Transduction and Mechanics in Morphogenesis

Justin Bois

Caltech

Winter, 2016

# 3 Introduction to information theory

Gap genes are morphogens: they are expressed differentially in different cells in a developing embryo. They give global positioning cues to the developing embryo; their expression levels are read out to ensure cells have the correct fate in the proper part of the embryo.

In *Drosophila*, the major gap genes are *hunchback*, *krüppel*, *giant*, and *knirps*. The question addressed in the Dubuis, et al. paper is: How much information about position in an embryo is contained in the levels of gene expression of the *Drosophila* gap genes?

In order to address this question, we need a formal definition of what is meant by *information* and how it is measured. In this lecture, we come to this definition. Note that the notation we use in this lecture is a little esoteric, but is chosen to match the notation in the Dubuis, et al. paper.

## 3.1 From desiderata to information

The quantification of information rests of the theory of probability. This makes intuitive sense. Say event $i$ happens with probability $P_i$. If $i$ is very probable and we observe it, we haven't learned much. For example, if we observe that the current pope is Catholic, we haven't learned much about popes. But if $i$ is very improbable and we observe it, we have learned a lot. If we observe a pig flying, we have learned something new about nature.

To codify this in mathematical terms, we might think that the information gained by observing event $i$ should scale like $1/P_i$, since more rare events give higher information.

Now, say we observe two *independent* events, $i$ and $j$. Since they are totally independent, the information garnered from observing both should be the sum of the information garnered from observing each. We know that the probability of observing both is $1/P_iP_j$. But

$$\frac{1}{P_i} + \frac{1}{P_j} \neq \frac{1}{P_iP_j}. \tag{3.1}$$

So, our current metric of information does not satisfy this addibility requirement. However,

$$\log\frac{1}{P_i} + \log\frac{1}{P_j} = \log\frac{1}{P_iP_j}. \tag{3.2}$$

So, we choose $\log(1/P_i) = -\log P_i$ as a measure of information. We are free to choose the base of the logarithm, and it is traditional to choose base 2. The units of information are then called *bits*.

Now, saw we have an ensemble of events. Then the average information we get from observing a events (i.e., the level of surprise) is

$$S[P_i] = -\sum_i P_i \log_2 P_i. \tag{3.3}$$

This is called the *Shannon information* or **Shannon entropy**. It has its name because of its relation to the same quantity in statistical thermodynamics. We will not delve into that in this course.

Let's look at the Shannon entropy another way. Say we know all of the $P_i$'s. How much knowledge do we know about what events we might observe? If the probability distribution is flat, not much. Conversely, if it is sharply peaked, we know a lot about what event we will observe. In the latter case, observing an event does not give us more information beyond what we already knew from the probabilities. So, $S[P_i]$ **is a measure of ignorance**. It tells us how uncertain or unbiased we are ahead of an observation. This will be crucial for defining how much we learn through observation. In the context of the gap genes, it will help us quantify how much we learn about position be observing a gene expression level.

I pause to note that we shortcutted our way into this definition of entropy by using some logic and the desire that independent events add. A more careful derivation was done in 1948 by Claude Shannon. He showed that the function we wrote for the entropy is the only function that satisfies three desiderata about measurements of ignorance.

1. Entropy is continuous in $P_i$.
2. If all $P_i$ are equal, entropy is monotonic in $n$, the number of event we could observe.
3. Entropy satisfies a composition law; grouping of events does not change the value of entropy.

The derivation is beautiful, but we will not go into it here.

Finally, let's look at a quick example. Say we have four possible outcomes. Consider the case where all outcomes are equally likely. We are therefore maximally ignorant about what we will observe, so the entropy should be large. Let's compute it.

$$S\left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right] = -4\left(\frac{1}{4}\log_2\frac{1}{4}\right) = 2 \text{ bits}. \tag{3.4}$$

Now, let's say that the first event is more likely. Now, we have some bias.

$$S\left[\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - 2\left(\frac{1}{8}\log_2\frac{1}{8}\right) = \frac{7}{4} \text{ bits}. \tag{3.5}$$

Indeed, the entropy is greater for the uniform distribution.

## 3.2 Mutual information

With the definition of entropy in hand, we can move on to the big question at hand for the gap genes. Say we have two ensembles, $X$ and $Y$. How much does knowing $Y$ tell us about $X$? I.e., how much information about $X$ do we get from knowing $Y$?

To answer this question, we first need to extend our definitions of probability to include *conditional probability* and *joint probability*. The joint probability, $P(x,y)$, is the probability of observing both $x$ and $y$. The conditional probability, $P(x \mid y)$ is the probability of observing $x$ *given* that we have observed $y$. If $P_x(x)$ is the probability of observing $x$ regardless of whether or not we have observed $y$, then

$$P(x,y) = P(y \mid x)P_x(x). \tag{3.6}$$

We can swap $x$ and $y$ to get

$$P(x,y) = P(y,x) = P(x \mid y)P_y(y). \tag{3.7}$$

From this, we can show that

$$P_x(x) = \sum_y P(x \mid y)P_y(y). \tag{3.8}$$

This result is derived as

$$\begin{aligned}\sum_y P(x,y) = \sum_y P(x \mid y)P_y(y) = \sum_y P(y,x) = \sum_y P(y \mid x)P_x(x) \\ = P_x(x)\left[\sum_y P(y \mid x)\right] = P_x(x),\end{aligned} \tag{3.9}$$

where we have use the fact that the bracketed term is equal to one.

Similarly to how we've defined conditional probability, we can define *conditional entropy* of $X$ given an event $y$. Note that we can only talk about entropy of ensembles, not of events, hence the capital $X$.

$$S[P(X \mid y)] = -\sum_x P(x \mid y)\log_2 P(x \mid y). \tag{3.10}$$

The conditional entropy of $X$ given $Y$ is the average of the above, analogously to how entropy itself is defined.

$$S[P(X \mid Y)] = -\sum_y P_y(y)\sum_x P(x \mid y)\log_2 P(x \mid y). \tag{3.11}$$

Now, we can answer the question of how much information is gained about $X$ from knowing $Y$. It is the *loss of ignorance*, or how much entropy we lost, when we learned $Y$. This is called the **mutual information** of $X$ and $Y$, $I_{y\to x}$.

$$I_{y\to x} = S[P_x(x)] - S[P(X \mid Y)]. \tag{3.12}$$

We can write out the expression for this in terms of the event probabilities.

$$\begin{aligned} I_{y\to x} &= -\sum_x P_x(x)\log_2 P_x(x) + \sum_y P_y(y)\sum_x P(x\mid y)\log_2 P(x\mid y) \\ &= -\sum_x\sum_y P_y(x)P(x\mid y)\log_2 P_x(x) + \sum_x\sum_y P_y(y)P(x\mid y)\log_2 P(x\mid y) \\ &= \sum_x\sum_y P_y(y)P(x\mid y)\log_2\frac{P(x\mid y)}{P_x(x)}. \end{aligned} \tag{3.13}$$

Going from the first line to the second, we have inserted the expression for $P_x(x)$ given in equation (3.8) and brought the summation symbols to the front of the summands. Now, using the fact that $P_y(y)P(x\mid y) = P(x,y)$, we can write

$$\begin{aligned} I_{y\to x} &= \sum_x\sum_y P(x,y)\log_2\frac{P(x\mid y)}{P_x(x)} \\ &= \sum_x\sum_y P(x,y)\log_2\frac{P(x,y)}{P_x(x)P_y(y)}. \end{aligned} \tag{3.14}$$

Written this way, we see that $x$ and $y$ are totally interchangeable, so

$$I_{y\to x} = I_{x\to y}. \tag{3.15}$$

This is why it is called mutual information.

The most useful for of the mutual information for our purposes (including two expressions for the interchangeability) is

$$\begin{aligned} I_{y\to x} &= \sum_x\sum_y P_y(y)P(x\mid y)\log_2\frac{P(x\mid y)}{P_x(x)} \\ &= \sum_x\sum_y P_x(x)P(y\mid x)\log_2\frac{P(y\mid x)}{P_y(y)}. \end{aligned} \tag{3.16}$$

## 3.3 Entropy, mutual information, and continuous distributions

So far, we have used only discrete probabilities. What if, e.g., $x$ is a continuous variable and then $P(x\mid y)$ and $P_x(x)$ are continuous *distributions*. Remember, distributions have *units*, the inverse of whatever the units of $x$ are. Looking at the definition

of entropy and conditional entropy, we are in trouble because we would have to take the logarithm of a quantity with units. We can't do that. It is quite difficult to define entropy in a continuum; that is a whole other subject.

This is not a problem for the mutual information, though. We always take the log of probability distribution with the same units, so the units cancel. Therefore, in the mutual information definition, we can just replace the sums with integrals. For example, if $x$ is continuous and $y$ discrete, we have

$$I_{y\to x} = \sum_y \int \mathrm{d}x\, P_y(y) P(x \mid y) \log_2 \frac{P(x \mid y)}{P_x(x)}. \tag{3.17}$$

## 3.4 Information in the gap genes

We are now ready to consider how much information the gap genes give about position. This is exactly $I_{g\to x}$, where $g$ is the gene expression level and $x$ is position. As we have derived, we know this is an appropriate metric for information about $x$ provided by $g$.

$$\begin{aligned} I_{g\to x} &= \int \mathrm{d}g \int \mathrm{d}x\, P_g(g) P(x \mid g) \log_2 \frac{P(x \mid g)}{P_x(x)} \\ &= \int \mathrm{d}g \int \mathrm{d}x\, P_x(x) P(g \mid x) \log_2 \frac{P(g \mid x)}{P_g(g)}. \end{aligned} \tag{3.18}$$

The expression on the first line is not very useful because we have no way of measuring $P(x \mid g)$. However, we *can* measure $P(g \mid x)$. We take an image of the embryo with a gap gene fluorescently labeled. We know the position along the embryo for each pixel, and we can get the gene expression level (in arbitrary units) by the fluorescence intensity. We do this over and over again and we can construct an empirical probability distribution of gene expression levels for each position $x$. This is exactly $P(g \mid x)$.

We still have to know $P_x(x)$ and $P_g(g)$ to compute the mutual information. First $P_x(x)$ is what we know about position along the embryo *before* we have learned anything about the gene expression levels. A given cell does not "know" where it is without the gap gene compass, so it is equally likely to be anywhere in the embryo. Thus, $P_x(x)$ is just a uniform distribution over the length of the embryo. If we use fractional embryo length as our length metric, $P_x(x) = 1$. Finally, we can compute $P_g(g)$ using a continuous version of equation (3.8).

$$P_g(g) = \int \mathrm{d}x\, P(g \mid x) P_x(x). \tag{3.19}$$

So, we did it! We can take experimental measurements and compute the information that the gap genes can give about position. Of course, if there are multiple genes, say

4 of them as is the case with the gap genes considered in the Dubuis, et al. paper, we have to integrate over all of them.

$$I_{\{g\}\to x} = \int \mathrm{d}g_1 \cdots \int \mathrm{d}g_4 \int \mathrm{d}x\, P_x(x) P(\{g\} \mid x) \log_2 \frac{P(\{g\} \mid x)}{P_{\{g\}}(\{g\})}. \tag{3.20}$$

## 3.5 Information through a noisy channel

As a final example that is relevant in cell signaling, we will consider information flow through a channel. We will consider a *binary symmetric channel*. A signal $x$ comes through a channel and results in readout $y$. The value of $x$ can be zero or one, as can the value of $y$. The problem is that the channel is noisy. There is a probability $f$ that the value of $y$ will be different than the value of $x$. How much information does knowing $y$ tell us about $x$?

We write out the probabilities and then compute the mutual information.

$$P(y = 0 \mid x = 0) = 1 - f \tag{3.21}$$

$$P(y = 0 \mid x = 1) = f \tag{3.22}$$

$$P(y = 1 \mid x = 0) = f \tag{3.23}$$

$$P(y = 1 \mid x = 1) = 1 - f. \tag{3.24}$$

The mutual information is

$$I_{x\to y} = \sum_x \sum_y P_x(x) P(y \mid x) \log_2 \frac{P(y \mid x)}{P_y(y)} \tag{3.25}$$

Now, say $P_x(x) = \{1/2, 1/2\}$. Then,

$$P_y(y) = \sum_x P(y \mid x) P_x(x) = \frac{1}{2}(1-f) + \frac{1}{2} f = \frac{1}{2}, \tag{3.26}$$

for both $y = 0$ and $y = 1$. Then,

$$\begin{aligned} I_{x\to y} &= \frac{1}{2}(1-f)\log_2(2(1-f)) + \frac{1}{2} f \log_2(2f) \\ &\quad + \frac{1}{2}(1-f)\log_2(2(1-f)) + \frac{1}{2} f \log_2(2f) \\ &= 1 + f\log_2 f + (1-f)\log_2(1-f). \end{aligned} \tag{3.27}$$

So, if $f = 0$ (meaning there is no noise), we get one bit of information. This makes sense, knowing $y$ tells us exactly what $x$ was, and $x$ can only take on two values; hence one bit. Similarly, if $f = 1$, $y$ is wrong all the time, so we again can deduce what $x$ is. We get one bit of information for $f = 1$ as well. If $f = 1/2$, we have no idea what $x$ is given $y$. In this case, the mutual information evaluates to zero. For intermediate values of $f$ we get some information between zero and one bit.