How does an organism extract relevant information from transcription factor concentrations?

Bauer, M.S.

DOI
10.1042/BST20220333

Publication date
2022

Document Version
Final published version

Published in
Biochemical Society Transactions

Biochemical Society Transactions (2022)
https://doi.org/10.1042/BST20220333

PORTLAND PRESS

Review Article

How does an organism extract relevant information from transcription factor concentrations?

Marianne Bauer[1,2,3]

[1]Bionanoscience Department, Delft University of Technology, van der Maasweg 9, 2629 Delft, The Netherlands; [2]Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544, U.S.A.; [3]Lewis–Sigler Institute for Integrative Genomics Princeton University, Princeton, NJ 08544, U.S.A.

Correspondence: Marianne Bauer (m.s.bauer@tudelft.nl)

How does an organism regulate its genes? The involved regulation typically occurs in terms of a signal processing chain: an externally applied stimulus or a maternally supplied transcription factor leads to the expression of some downstream genes, which, in turn, are transcription factors for further genes. Especially during development, these transcription factors are frequently expressed in amounts where noise is still important; yet, the signals that they provide must not be lost in the noise. Thus, the organism needs to extract exactly relevant information in the signal. New experimental approaches involving single-molecule measurements at high temporal precision as well as increased precision in manipulations directly on the genome are allowing us to tackle this question anew. These new experimental advances mean that also from the theoretical side, theoretical advances should be possible. In this review, I will describe, specifically on the example of fly embryo gene regulation, how theoretical approaches, especially from inference and information theory, can help in understanding gene regulation. To do so, I will first review some more traditional theoretical models for gene regulation, followed by a brief discussion of information-theoretical approaches and when they can be applied. I will then introduce early fly development as an exemplary system where such information-theoretical approaches have traditionally been applied and can be applied; I will specifically focus on how one such method, namely the information bottleneck approach, has recently been used to infer structural features of enhancer architecture.

Theoretical approaches for gene regulation

The discovery that *Escherichia coli* could switch from growing on glucose to lactose depending on lactose's presence in the environment showed that cells can respond to environmental stimuli by gene regulation [1]. As the lac-operon was one of the most impressive early discoveries in gene regulation, many early models for gene regulation, also in the context of development, were based on it [2]: the lac gene is regulated by transcription factors whose binding sites are in direct proximity to the gene's promoter (see Figure 1). The area around the promoter contains binding sites for transcription factor molecules that can facilitate or block binding of the polymerase and, therefore, activate or repress the expression of the gene [3, 4]. A model for this assumes that h transcription factors of concentration s bind to the binding sites (cooperatively, i.e. at the same time), and that the mean concentration of expressed output (assuming ergodicity) corresponds to the averaged probability of the bound state. The chemical master equation for this process reads,

$$\frac{\mathrm{d}P(0,t)}{\mathrm{d}t} = k_{off} - (k_{off} + k_{on}s^h)P(0,t), \tag{1}$$

where k_{on} and k_{off} are the rate constants for binding and unbinding and $P(0,t)$ the probability that

Received: 10 July 2022
Revised: 16 August 2022
Accepted: 17 August 2022

Version of Record published:
16 September 2022

Biochemical Society Transactions (2022)
https://doi.org/10.1042/BST20220333

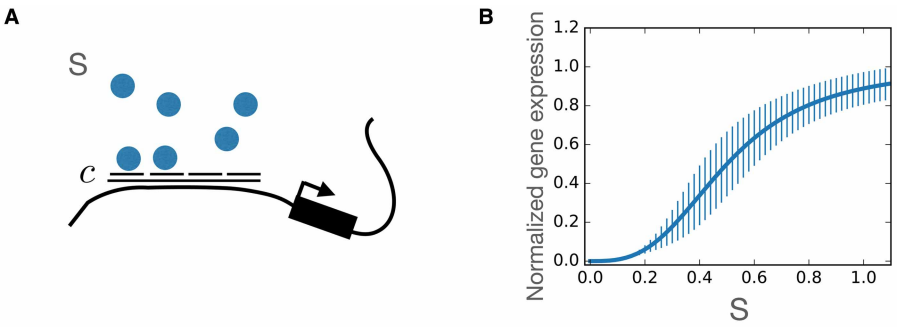

Figure 1. Gene regulation as a continuous function of number of bound transcription factors.
Left: Sketch for binding sites for transcription factors (blue) of concentration S close to a gene's promoter; right: expression as depending on S can be modeled by a Hill-function.

the site is free at time t. Then, in steady state, the mean presence of the bound state is

$$\bar{c} = \lim_{T\to\infty} \frac{1}{T} \int_0^T (1 - P(0,t))\, \mathrm{d}t = \frac{s^h}{k_{off}/k_{on} + s^h}. \tag{2}$$

This is a Hill-function (see Figure 1) [5, 6]: gene expression increases sigmoidally with transcription factor concentration and the steepness of the increase is given by h.

Another canonical model for gene regulation is based on sharp thresholds: this idea originated from Wolpert [7] concerning the question of how cells can express different genes given different transcription factor concentrations. Here, a gene is expressed (at maximal level) when the concentration of a particular input transcription factor is higher than a certain value, and not expressed ($g = 0$) when the transcription factor concentration is lower than this value; for a transcription factor gradient in development, for multiple genes, this corresponds to the so-called 'French flag model', see Figure 2 [8, 9]. This response to gene expression based on a sharp threshold value of concentration can be phrased mathematically as

$$C = H(s - \theta_1), \tag{3}$$

where H is the Heaviside stepfunction, defines as $H(x) = 0$ if $x < 0$ and $H(x) = 1$ if $0]] >$; for multiple thresholds, this works analogously, i.e. the amount of expressed genes varies in discrete units or states shown in

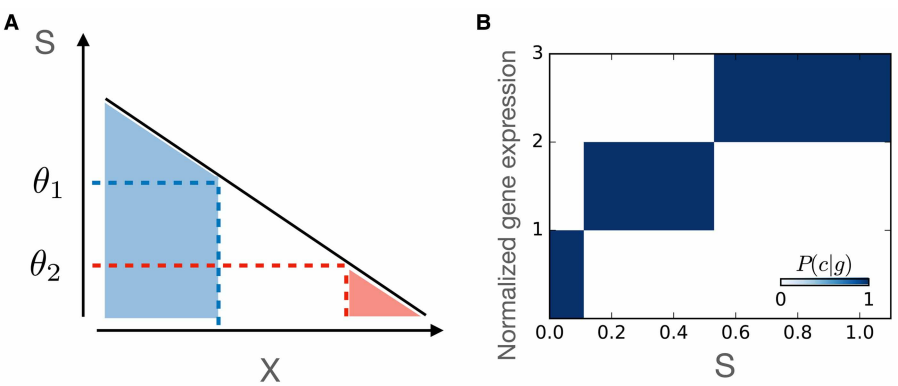

Figure 2. Gene regulation as a threshold-like response to varying transcription factor concentrations *S*.
Left: French flag model where regulatory response is sharply different if S exceeds a particular threshold (here for thresholds θ_1 and θ_2; right: gene expression of a single gene as a function of S for these two thresholds.

Figure 2 (right). Mechanistically, one could assume that such thresholds can be implemented in terms of a very steep Hill's function, with $h \to \infty$.

The advantage of these models is their simplicity, or their usefulness as a 'limiting case': for example, the analysis of the graph-theoretical models has shown that the Hill function from equation (2) gives the steepest possible slope (or threshold) of all various individual combinations of transcription factor binding [10, 11]. Thus, both of these models are still frequently used for understanding gene expression [12–14].

Yet, one important change in thinking about gene regulation around the early 2000s was the focus on noise and stochasticity of gene expression [15]. This stochasticity is a consequence of both stochastic promoter bursting and the limited number of transcription molecules which bind to the binding site region in a limited amount of time.

On the theoretical side, calculations for noise originate from work in 1977 on chemotaxis (which involves the sensing of a molecular gradient by receptors) [16]: Berg and Purcell argued that the signal-to-noise ratio $\delta c/c$ with which a 1D receptor of size A which measures for time τ can infer the concentration s of a freely diffusing signal molecule (diffusion constant D):

$$\left(\frac{\delta c}{c}\right)^2 = \frac{2}{Das\tau(1-p)}, \tag{4}$$

where p describes the occupation at the binding site region (the term thus implies that binding can not occur when the binding site is already fully occupied). Here, we use c in order to make explicit that this denotes the cell's estimate of the signal s.

Remarkably, this limit presents still a lower bound for noise in binding in equilibrium. More general work in the early 2000s [17], which included account the noise of binding and unbinding of the molecules, showed that the signal-to-noise ratio of inferring c changes to [17, 18]

$$\left(\frac{\delta c}{c}\right)^2 = \frac{2}{Das\tau(1-p)} + \frac{2}{k_{on}s(1-p)\tau}; \tag{5}$$

the first term corresponds to the diffusion-limited contribution (c.f. 4), and the second term to noise from binding events. This signal-to-noise ratio is higher than equation (4). Similarly, extensions towards cooperative binding [19] did not yield a lower bound. Clever readout [20] or spending energy (likely during gene regulation) [21] can lower this bound. Nevertheless, from a modeling perspective, the equilibrium case is frequently preferred as it depends less on the details of the model.

While the Berg–Purcell bound can be applied to the Hill-function model with a single binding site, generalization to more binding sites or more complicated mechanisms is difficult. The thresholded model does not incorporate noise at all: this is a key shortcoming of this intuitive model, especially as it has been suggested that increased cooperativity (i.e. a more threshold-like mechanism) may raise the noise by increasing the correlation time of the input noise, impeding noise averaging [19, 22].

The above discussion already shows that it is difficult to calculate both the mean and noise of gene expression in a model bottom-up. In addition, there are a series of experimental insights since the early work on gene regulation, which make the situation even more complex.

Gene regulation now

In eukaryotic organisms, the regulatory architecture is different from the lac-operon: genes can be regulated by one or more promoters, as well as several regions with binding sites for transcription factors (the so-called enhancers) which can be several kilobasepairs away from the promoter or the gene [23] (sketch in Figure 3). These enhancers frequently have binding sites for a larger number of different transcription factors, some of which have pioneering activities that make the chromatin accessible.

Modifications of these models to incorporate the more complex regulatory landscape of individual transcription factor binding have, for example, been made by the so-called 'thermodynamic models' for transcription [24, 25]. Here, the probability of the downstream gene to turn on and off depends on a partition function, which takes into account the probability of various combinations of bound states of transcription factors to binding site regions close to the DNA given the binding energies; different such combinations can lead to

Biochemical Society Transactions (2022)
https://doi.org/10.1042/BST20220333

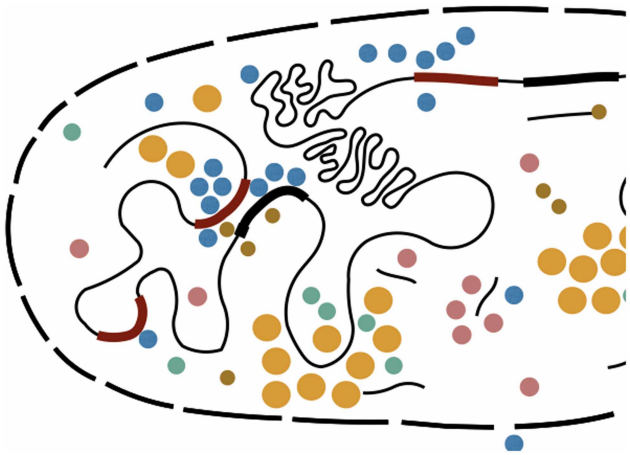

Figure 3. Sketch of gene regulatory environment: several enhancers (dark red) can regulate a gene (dark black); protein concentrations can be inhomogeneous.

different levels of gene expression. Recently, how the binding of transcription factors is affected when other transcription factors are already bound has been investigated by graph-theoretical models for transcription factor binding [11]. Finally, 'kinetic' models for gene regulation have taken seriously the possibility that not the thermodynamic steady state, but a series of non-equilibrium reactions are responsible for gene regulation [26–28]; these kinetic models are particularly important with the recent trend to investigate the importance of pioneer transcription factors which make chromatin accessible in the first place [29–34]. While these models present a significant progress, incorporating the effects of the joint activity of several enhancer elements is difficult. In addition, the calculation of noise outside a strict thermodynamic framework is difficult and highly parameter specific.

The situation is further complicated by the idea that transcription may involve a topological change in the genome that changes enhancer-promoter distances [34–37]. One additional complication is derived from the recent research focus on cellular compartmentalization, which means that the concentrations of transcription factors may vary across the cell [38, 39]. This is especially topical now as liquid–liquid phase separation (LLPS) has recently been implied to also affect transcription [40–43]. While LLPS is being established as a mechanism for cellular compartmentalization when the numbers of involved proteins are large, to what extent it affects gene regulation is still intensely debated [44]: especially in development, concentrations of some transcription factor peak of order 10 000 molecules per cell [30, 45–47]; thus, even if only ca 50 inhomogeneities or droplets are observed, they would need to contain less than the LLPS-typical numbers of 100s of molecules per droplet if only these transcription factors make up the droplet; this makes the applicability of the mechanism difficult. Nevertheless, the fact that transcription factors are likely inhomogeneously distributed is gaining prominence in the field [29, 48, 49].

These heterogeneous transcription factor distributions matter from the modeling perspective: frequently, transcription factor concentrations are only available averaged across the entire cell, but the local concentration of the transcription factor close to its binding site is required for the model (see equation 2). If these concentrations are unknown, estimating parameters for more specific models might lead to flawed conclusions. Similarly, calculating the noise of binding at the binding site regions is almost impossible when neither the number of transcription factors nor the mechanism for their accumulation around the binding site is known.

Overall, the added experimental complexities mean that although many advances have been made regarding modeling the regulation of individual genes in specific developmental time periods, an overall conceptual picture is still lacking. Such a conceptual picture is nevertheless important: conceptual understanding can help predict whether a particular gene may have many enhancers, where they might be located, or what binding site arrangements can detrimentally change expression.

Thus, in the following section, I will introduce a 'top-down' approach to complement to the 'bottom-up' mechanistic models; this approach is based on data and attempts to infer structural features necessary for precise gene regulation from these data.

Sensing approach to gene regulation

A complex system where it has been similarly complicated to draw up simple models due to the large number of functional elements involved is a net of neurons, such as the brain. One can think of the Hill functions, or more specific molecular schemes, as being like the Hodgkin–Huxley model for the electrical dynamics of neurons [50]. An alternative is to take the result that these dynamics generate action potentials or 'spikes', and ask how these spikes represent information of relevance to the organism [51]. The hope is that there are principles governing this representation without reference to molecular details. This question of 'reading the code' thus makes use of ideas from statistical physics or network theory and also from signal processing [52–55]. This approach has had considerable success in the neural context, and we can hope that something similar will help us think about information flow through transcriptional regulation.

One particularly exciting approach here is to treat the interpretation of the transcription factor concentration (s) as a (combinatorial) sensing problem. Such 'efficient' sensing approaches have been successful in neuroscience, for example, concerning olfaction [56] or concerning photoreceptors [57]. A crucial starting point for information optimization in neuronal systems was the work by Laughlin [51, 58]: he argued that photoreceptors are assigned such that they pick up on the most informative part of the signal, so that they can extract the most possible information given a limited number of receptors. This structural knowledge is important also for extracting information from transcription factor concentrations: given the typical statistics of the transcription factor signal, the hope would be to infer how many sensors are necessary to provide a certain amount of information and how they should optimally be distributed to extract this.

Before briefly introducing the information-theoretic optimization in the sensing problem introduced by Laughlin as an example of a signal processing optimization, I want to emphasize one difference between electronic signal transfer and biological systems: In electronic signal transfer, one can consider how to best represent (source coding: optimizing entropy), and how to best transmit a message (channel coding: optimizing error correction). In biological systems, it can be difficult to differentiate the signal from what the message should be (for example, for signal processing of photoreceptors in the eye, the intuition could be that the set of messages is the set of maximally distinct images; however, some animals may care less about specific features of the images). In addition, in the processing of gene regulatory signals, the source (chemical concentration) and the channel characteristics (noise profiles) can be modified biologically (i.e. have evolved evolutionarily). Thus, it is not a priori helpful to think of coding categories as having been separately optimized in a joint source-channel coding sense [59], but better to use one's biological intuition to investigate a plausible optimization goal, and see what one learns. In the spirit of not distinguishing signal-processing categories, I will, in the following, introduce Laughlin's sensing problem phrased in terms of an optimization of information (c.f. [51]), rather than entropy.

Laughlin was wondering how a class of insect eye neurons, the so-called large monopolar cells (LMCs) can sense light intensity from different natural landscapes. We can denote the light intensity or signal by J, which can increase from 0 to a particular value J_{max}; the distribution of different intensities is $P($. The information provided in this signal needs to be transferred by the LMCs in terms of a graded potential, meaning that the LMC integrates the signal intensity from a series of photoreceptors [60] and uses the value of this potential as a proxy for the value of the intensity. We call this interpretation of the signal C (we will discuss this in more detail later). We now want to optimize the mutual information

$$I(J;C) = \iint \mathrm{d}J\,\mathrm{d}C P(J,C)\log_2\frac{P(J,C)}{P(J)P(C)}, \tag{6}$$

where $P($ is the joint probability distribution and $P(C)$ the marginal distribution of the graded potential. Optimizing the mutual information corresponds to essentially maximizing the correlations between J and C. We note that the mutual expression can also be expressed as

$$I(J;C) = H(C) - H(C\,|\,J), \tag{7}$$

where

$$H(C) = -\int \mathrm{d}C P(C)\log P(C) \tag{8}$$

Biochemical Society Transactions (2022)
https://doi.org/10.1042/BST20220333

and $H(C\,|\,J)$ are the entropy and conditional entropy, respectively. The mutual information is at its maximal value, the entropy of C, when J and C are maximally correlated. We want to find an encoding $J \rightarrow C$ that maximizes this mutual information; this encoding can be written as a function that assigns one or many values of potential C (depending on the noise in the encoding) to a value of light intensity J. It is important to note that we will not be able to calculate what numerical value C should have for any particular J: from a decoding perspective, it is irrelevant if high light intensities should be encoded by a low potential or by a high potential, as long as the mapping is clear.

To make the problem simple, we assume that the noise in the encoding $J \rightarrow C$ is very low. Often, we can assume that the encoding $J \rightarrow C$ has a probabilistic encoding in which C has a Gaussian distribution around a mean, $\bar{C} = f($; this means that

$$P(C\,|\,J) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(C-\bar{C})^2}{2\sigma^2}\right]. \tag{9}$$

Then, we can calculate the probability distribution $P(C)$ analytically and relate it to the inverse of the derivative of $\bar{C}$ with respect to J, $P(C) = |\mathrm{d}f/\mathrm{d}J|^{-1}_{J_0} P($ [51, 61]. Thus, optimizing the mutual information over all possible encodings $\bar{C} = f($ for Gaussian distribution corresponds to, modulo normalization factors, optimizing over all possible probability distributions $P(C)$.

If the noise σ is lower than a reasonable discretization of C (i.e. if we think the graded voltages can only be resolved to a certain value), we can assign a single value C to each value J. Then, the conditional entropy is zero. Thus, in Laughlin's case, maximizing the mutual information corresponds essentially to maximizing the entropy in the encoding variable C. Maximizing the entropy is a simple information-theoretic problem, which can be solved using the method of Lagrangian multipliers [59]: the distribution $P(C)$ that maximizes the entropy is the uniform distribution. Since we now know that $P(C) = const = |\mathrm{d}f/\mathrm{d}J|^{-1}_{J_0} P($, we can see that $\mathrm{d}f/\mathrm{d}J = P($. This means that the best possible encoding for light intensities is one where the slope of the encoding matches the distribution over typical light intensities that typical insects see. This information-theoretic result is exactly what Laughlin found in his data [58].

In the next section, we will apply this sensing optimization to transcription factors.

Sensing applied to transcription factors

We note that similarly to the light intensity, we assume that the 'intensity' or concentration of the transcription factor provides relevant information to cells. Especially in early fly development, the concentrations of certain transcription factors, such as the maternal morphogen Bicoid, provide information about a particular cell's fate [62]: cells close to the head of the embryo at high bicoid concentration differentiate differently from cells close to the tail end of the embryo. In fact, neighboring cells along the embryonal axis distinguish almost uniquely into different cell fates (e.g. different body segments, such as thorax and abdomen, regulated by the hox genes, and even within a segment, different pair-rule genes are expressed at different concentrations). Cells will need to read the transcription factor concentration signal in a way that maximizes the information between the signal and the future cell fate. For simplicity, we can label the cell fate by its position along the embryonal axis, X. This idea goes back to Wolpert's idea about 'positional information', as in the French flag model introduced above, but was developed and made more precise in a series of papers by Bialek, Gregor, Wieschaus and colleagues [45, 61, 63, 64].

In the case for fly development, there is thus a clear variable we care about (cell fates along the embryonal axis X). The signals that provide information about this cell fate are the four so-called gap genes (see Figure 4): they are expressed just downstream of Bicoid and two other maternally supplied signals. They form a complete set of inputs, because their expression profiles provide enough information about the downstream cell fates [63], and this information can be used to predict the expression pattern of the downstream pair-rule genes [64]. Thus, for the question of how to extract information from transcription factor signals, we have, in the fly embryo, a clear set of candidate signals where we can investigate whether and how information can be extracted, based on data.

The complication compared with Laughlin's example in the previous section is that the signals and the variable, we care about, are different. In Laughlin's example, the intuition was that the different intensities J were the signals that needed to be maximally distinguished. Here, the gap gene expression concentrations are signals

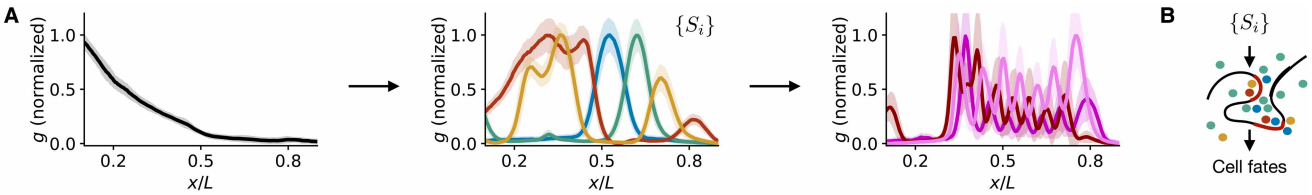

Figure 4. Sketch for signal processing in the fly embryo. (**A**) Expression patterns of the maternal Bicoid gradient, which regulates the four gap genes, which, in turn, regulate the seven pair-rule genes (three shown). The pair-rule genes, together with the hox genes, determine the fly's segmentation along the embryonal axis, X, or its cell fates. Data from [45, 64]. (**B**) Gap gene expression patterns present signals that need to be interpreted by the gene regulatory apparatus for the correct cell fates differentiation.

$i \in \{Hb, Gt, Kr, Kni\}$ for $i \in \{Hb, Gt, Kr, Kni\}$ to represent the four genes *Hunchback*, *Giant*, *Krupel* and *Knirps*, and the cell fates of the set of possible positions along the axis, X. We are *not* interested in optimizing $I(S;X)$ (which would correspond to designing a set of signals, $\{S\}$, which maximize the different responses or cell fate decisions along X). Instead, we take for granted the shape of the signals or gap expression profiles, and we are interested in how this biological signal can be interpreted by the cell in order to learn about the future cell's fate. Thus, what we need to optimize is the cell's reading of the signal (see Figure 4B); we denote this measurement or interpretation by C, which stands for compression, as a compression can be seen as an efficient measurement of the signal. In other words, we want to maximize the information that cells have, after their measurement of the signal, about their future cell fate decisions, i.e. $I(C;X)$. We know that this reading or measurement is noisy, because of the stochastic noise with transcription factor arrival and binding discussed above. If we knew the mechanism for binding and arrival, we could calculate the probability distribution $P(C\,|\,S)$, i.e. the cell's internal measurement for every value of the signal, and then we could calculate $I(C;S)$. However, we do not want to make an assumption about this mechanism; instead, we want to capture the essence but not the details of the limitations in any mechanism. Thus, we maximize $I(C;X)$ for various values of $I(C;S)$, or for various values of noisy measurements. For each value of $I(C;S)$, we want to infer the encoding $P(C\,|\,S)$ that extracts the most information about X; we can then compare this to calculations of $I(C;S)$ and $I(C;X)$ from various mechanisms. This inference will allow us to see how the cell would optimally set up the measurement if it had to be noisy.

This optimization procedure can be phrased as the optimization goal

$$\max_{P(C\,|\,S)} I(C;X) - \lambda I(C;S). \tag{10}$$

This optimization goal corresponds to the information bottleneck optimization goal [65]; for this optimization goal, an analytic expression for $P(C\,|\,S)$ exists that allows one to calculate $P(C\,|\,S)$ self-consistently for each value of λ, and, due to the numerical discretization, for each value of discrete levels of C. According to this self-consistent equation, $P(C\,|\,S)$ reads.

This information bottleneck algorithm is a compression algorithm that has recently enjoyed an increase in popularity, due to interest from machine learning [66, 67]: in image recognition, one is also interested in compressing away aspects of an image that do not contribute to our recognition of it. Similarly, the question when extracting transcription factor signals is which aspects of the signal are most informative, so that the organism can concentrate on sensing them more precisely. To go towards continuous C, we can simply ensure that the number of discrete levels of C is large.

We performed this calculation in [68], and I briefly summarize the key results here. For example, we can look at how much $I(C;X)$ we can obtain at best for each value of $I(C;S)$. We show this optimal trace in the information plane (where we plot $I(C;S)$ on the x and $I(C;X)$ on the y-axis) in Figure 5. We note that all possible values on the information plane are below the diagonal (where $I(C;X)I(C;S)$) and below the top dashed line (where $I(C;X)I(S;X)$). This top line is at $I(S;X) \approx 4.1$bit, which is the amount of information that the gap genes provide about cell fates [63]. This upper bound is due to the data processing inequality: effectively, it means that the cell can never obtain more information by its measurement than the signal provides. The optimal sensing bound calculated by the bottleneck algorithm is close to the best possible bounds, in that it

Biochemical Society Transactions (2022)
https://doi.org/10.1042/BST20220333

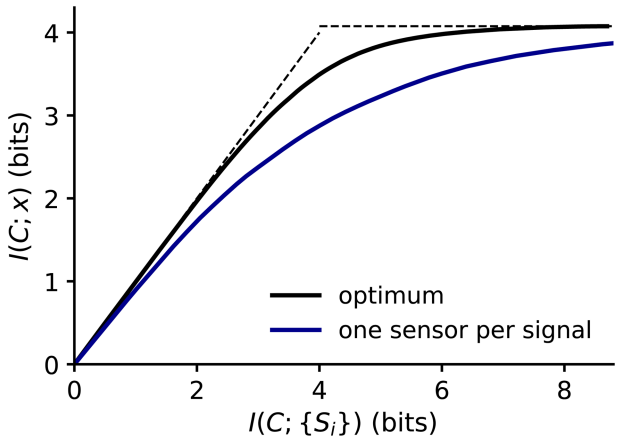

Figure 5. The optimal bottleneck curve for a single sensor for all genes (black) and with four sensors optimized for all genes separately (blue). The *x*-axis shows the information capacity of the sensor and the *y*-axis the information that the sensor has cell fates, which we want to maximize. Data replotted from [68].

increases quite steeply along the diagonal initially. This is not necessarily the case when one tries to find the best possible signal processing from a set of neurons [69], and suggests that the gap transcription factors here really provide a complete signal that can be sensed well.

What do enhancers need to do if they extract signals optimally?

What can one infer mechanistically about how enhancers need to sense these gap transcription factors, if they sensed optimally? To do this, one can either compare the optimal information bottleneck curve to calculations from various mechanisms, or calculate where on the optimal curve various sensors would be. To simplify the question about mechanism, we can use a single gap transcription factor Hb: we imagine that cells need to infer theire fate (or position) from the concentration of Hb, and ask how much information an optimal sensor C can infer given a limit on its capacity $I(C;S)$. Figure 6 shows the optimal bottleneck curve for Hb (with a lower maximum, at $I(Hb,X) = 2.1$ bits, as it is this time only a single transcription factor that is analyzed for cell fate decision making). We compare this optimal curve to the threshold model by optimizing the position of one, two and several thresholds to maximize the information $I(\theta;X)$, where θ is a thresholded variable. These thresholds lie exactly on the bottleneck curve. This is important because it means that thresholded measurements of transcription factors, which only trigger when transcription factors concentrations are above or below values, while not mechanistically feasible, are information-theoretically optimal. Thus, the biological intuition that led to the suggestion that the important features of the gap genes were their boundaries is information-theoretic intuition; we can thus make mathematically precise intuition that biologists had for several decades, and expand on it.

Further analysis of these thresholded measurements showed that the threshold positions did not need to be fine tuned: specifically, thresholds at higher transcription factor concentrations could be placed more loosely. Intuitively, this means that in concentration regimes where Hb is expressed noisily, the precise levels are not as important. Biologically, transcription factor concentrations at high concentrations are often measured with weak binding sites. We deduced that this suggests that how many weak binding sites an enhancer has does not matter as much; this, again, is known in biology.

Second, we see that about 10 thresholds are required to sense hunchback correctly. When we use realistic estimates for how well a single enhancer can sense in a Hill-function model with the Berg–Purcell noise (equation 4), we obtained 1–3 bits. While this may just be enough information to sense Hunchback correctly, it is not enough when we want to obtain information about all four gap transcription factors together: there, we needed about 3.8 bits or ca 50 thresholds to get to an accuracy that gets to about 10% of the information provided. This shows that many enhancers are required to read the gap transcription factor signals.

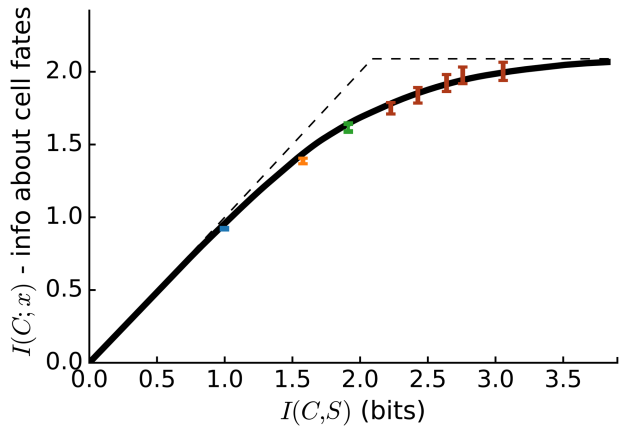

Figure 6. The bottleneck curve for an optimal sensor for the single protein Hb. The dashed lines correspond to the data processing inequality, and separate inaccessible from accessible regions of the information plane. An abstract sensor that measures with one (blue), two (orange), three (green) and more (red) thresholds is also on this curve and thus also information-theoretically optimal. Data replotted from [68].

Finally, in order to determine what enhancer architectures should look like if they sensed optimally, one can perform a comparative calculation. We optimized four separate sensors with the constraint that each sensor should only sense one gap transcription factor each. We found that this was always worse than having a single sensor that sensed them together (see blue line in Figure 5). This means that having four enhancers, one of which would sense a single transcription factor, would not be information-theoretically optimal. Indeed, we know that the enhancers that sense the gap transcription factors do have binding sites for many of them at the same time; for example, the Eve stripe 2 enhancer has binding sites for the gap proteins Hb, Kr and Gt [70].

Perspectives

- We were able to apply a sensing approach to transcription data and found that this captured several aspects of the transcriptional architecture for this network: multiple enhancers which measure gap proteins together (in combinations of expression levels that cannot easily be separated) and with degeneracies for weak binding sites allow the fly to extract most of the protein signal that is provided in the gap transcription factors, and this, in turn, allows the fly to make the correct cell fate decisions.

- The hope is that sensing or inference approaches can, together with mechanistic approaches, help us understand faster why certain regulatory features are there; this could be important not only for a better *in vivo* applications, but also for an appreciation of the regulatory complexity.

- Future directions: Especially for synthetic gene regulation, where one hopes to engineer gene regulatory systems [71–73], often unforeseen bottlenecks arise (see e.g. [74]). A conceptual framework that can identify how important various transcription factor signals are and how they might be sensed in natural systems could help to transfer ideas to the synthetic systems, or help identify what is different.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

Funding

This work was supported by TU Delft, by the National Science Foundation through the Center for the Physics of Biological Function (PHY-1734030), and was finished at Aspen Center for Physics, which is supported by the National Science Foundation grant PHY-1607611.

Acknowledgments

I am very grateful to William Bialek for many inspiring discussions, and for comments on a draft of this manuscript. I also acknowledge helpful conversations with colleagues in the Princeton biophysics, fly, and neuroscience groups in the process of this work, in particular with Eric Wieschaus on fly development.

Abbreviations

LLPS, liquid–liquid phase separation; LMCs, large monopolar cells.

References

1 Jacob, F. and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 https://doi.org/10.1016/S0022-2836(61)80072-7
2 Watson, J., Baker, T., Bell, S., Gann, A., Levine, M. and Losick, R. (2008) *Molecular Biology of the Gene*, Pearson/Benjamin Cummings, San Francisco, CA, USA
3 Goodwin, B.C. (1965) Oscillatory behavior in enzymatic control processes. *Adv. Enzyme Regul.* **3**, 425–437 https://doi.org/10.1016/0065-2571(65)90067-1
4 Griffith, J.S. (1968) Mathematics of cellular control processes I. Negative feedback to one gene. *J. Theor. Biol.* **20**, 202–208 https://doi.org/10.1016/0022-5193(68)90189-6
5 Hill, A.V. (1910) A new mathematical treatment of changes of ionic concentration in muscle and nerve under the action of electric currents, with a theory as to their mode of excitation. *J. Physiol.* **40**, 190–224 https://doi.org/10.1113/jphysiol.1910.sp001366
6 de Jong, H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **9**, 67–103 https://doi.org/10.1089/10665270252833208
7 Wolper, L. (1969) Positional information and the spatial pattern of cellular differentiation. *J. Theor. Biol.* **25**, 1–47 https://doi.org/10.1016/S0022-5193(69)80016-0
8 Sharpe, J. (2019) Wolpert's French Flag: what's the problem? *Development* **146**, dev185967 https://doi.org/10.1242/dev.185967
9 Jaeger, J. (2009) Modelling the *Drosophila* embryo. *Mol. Biosyst.* **5**, 1549 https://doi.org/10.1039/b904722k
10 Estrada, J., Wong, F., DePace, A. and Gunawardena, J. (2016) Information integration and energy expenditure in gene regulation. *Cell* **166**, 234–244 https://doi.org/10.1016/j.cell.2016.06.012
11 Wong, F. and Gunawardena, J. (2020) Gene regulation in and out of equilibrium. *Annu. Rev. Biophys.* **49**, 199–226 https://doi.org/10.1146/biophys.2020.49.issue-1
12 Hornos, J.E., Schultz, D., Innocentini, G.C., Wang, J., Walczak, A.M., Onuchic, J.N. et al. (2005) Self-regulating gene: an exact solution. *Phys. Rev. E* **72**, 051907 https://doi.org/10.1103/PhysRevE.72.051907
13 Tkacik, G. and Walczak, A. (2011) Information transmission in genetic regulatory networks: a review. *J. Phys.: Condens. Matter.* **23**, 15310 https://doi.org/10.1088/0953-8984/23/15/153102
14 Tkačik, G., Walczak, A.M. and Bialek, W. (2009) Optimizing information flow in small genetic network. *Phys. Rev. E* **80**, 031920 https://doi.org/10.1103/PhysRevE.80.031920
15 Elowitz, M.B., Levine, A.J., Siggia, E.D. and Swain, P.S. (2002) Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 https://doi.org/10.1126/science.1070919
16 Berg, H.C. and Purcell, E.M. (1977) Physics of chemoreception. *Biophys. J.* **20**, 193–219 https://doi.org/10.1016/S0006-3495(77)85544-6
17 Bialek, W. and Setayeshgar, S. (2005) Physical limits to biochemical signaling. *Proc. Natl Acad. Sci. U.S.A.* **102**, 10040–10045 https://doi.org/10.1073/pnas.0504321102
18 Kaizu, K., de Ronde, W.H., Paijmans, J., Takahashi, K., Tostevin, F. and ten Wolde, P.R. (2014) The Berg–Purcell limit revisited. *Biophys. J.* **106**, 976–985 https://doi.org/10.1016/j.bpj.2013.12.030
19 Bialek, W. and Setayeshgar, S. (2008) Cooperativity, sensitivity, and noise in biochemical signaling. *Phys. Rev. Lett.* **100**, 258101 https://doi.org/10.1103/PhysRevLett.100.258101
20 Endres, R.G. and Wingreen, N.S. (2009) Maximum likelihood and the single receptor. *Phys. Rev. Lett.* **103**, 158101 https://doi.org/10.1103/PhysRevLett.103.158101
21 ten Wolde, P.R., Becker, N.B., Ouldridge, T.E. and Mugler, A. (2016) Fundamental limits to cellular sensing. *J. Stat. Mech.* **162**, 1395–1424 https://doi.org/10.48550/arXiv.1505.06577
22 Skoge, M., Meir, Y. and Wingreen, N.S. (2011) Dynamics of cooperativity in chemical sensing among cell-surface receptors. *Phys. Rev. Lett.* **107**, 178101 https://doi.org/10.1103/PhysRevLett.107.178101
23 Furlong, E.E. and Levine, M. (2018) Developmental enhancers and chromosome topology. *Science* **361**, 1341–1345 https://doi.org/10.1126/science.aau0320

24 Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J. et al. (2005) Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* **15**, 116–124 https://doi.org/10.1016/j.gde.2005.02.007
25 Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J. et al. (2005) Transcriptional regulation by the numbers: applications. *Curr. Opin. Genet. Dev.* **15**, 125–135 https://doi.org/10.1016/j.gde.2005.02.006
26 Scholes, C., DePace, A.H. and Sánchez, Á. (2017) Combinatorial gene regulation through kinetic control of the transcription cycle. *Cell Syst.* **4**, 97–108 https://doi.org/10.1016/j.cels.2016.11.012
27 Harden, T.T., Vincent, B.J. and DePace, A.H. (2021) biorxiv.
28 Martinez-Corral, R., Park, M., Biette, K., Friedrich, D., Scholes, C., Khalil, A. et al. (2020) biorxiv.
29 Dufourt, J., Trullo, A., Hunter, J., Fernandez, C., Lazaro, J., Dejean, M. et al. (2018) Temporal control of gene expression by the pioneer factor Zelda through transient interactions in hubs. *Nat. Commun.* **9**, 5194 https://doi.org/10.1038/s41467-018-07613-z
30 Hannon, C.E., Blythe, S.A. and Wieschaus, E.F. (2017) Concentration dependent chromatin states induced by the bicoid morphogen gradient. *eLife* **6**, e28275 https://doi.org/10.7554/eLife.28275
31 Gaskill, M.M., Gibson, T.J., Larson, E.D. and Harrison, M.M. (2021) GAF is essential for zygotic genome activation and chromatin accessibility in the early *Drosophila* embryo. *eLife* **10**, e66668 https://doi.org/10.7554/eLife.66668
32 McDaniel, S.L., Gibson, T.J., Schulz, K.N., Garcia, M.F., Nevil, M., Jain, S.U. et al. (2019) Continued activity of the pioneer factor Zelda is required to drive zygotic genome activation. *Mol. Cell* **74**, 185–195.e4 https://doi.org/10.1016/j.molcel.2019.01.014
33 Eck, E., Liu, J., Kazemzadeh-Atoufi, M., Ghoreishi, S., Blythe, S.A. and Garcia, H.G. (2020) Quantitative dissection of transcription in development yields evidence for transcription-factor-driven chromatin accessibility. *eLife* **9**, e56429 https://doi.org/10.7554/eLife.56429
34 Levo, M., Raimundo, J., Bing, X.Y., Sisco, Z., Batut, P.J., Ryabichko, S. et al. (2022) Transcriptional coupling of distant regulatory genes in living embryos. *Nature* **605**, 754–760 https://doi.org/10.1038/s41586-022-04680-7
35 Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B. and Gregor, T. (2018) Dynamic interplay between enhancer-promoter topology and gene activity. *Nat. Genet.* **50**, 1296–1303 https://doi.org/10.1038/s41588-018-0175-z
36 Barinov, L., Ryabichko, S., Bialek, W. and Gregor, T. (2020) Preprint arXiv:2012.15819.
37 Batut, P.J., Bing, X.Y., Sisco, Z., Raimundo, J., Levo, M. and Levine, M.S. (2022) Genome organization controls transcriptional dynamics during development. *Science* **375**, 566–570 https://doi.org/10.1126/science.abi7178
38 Shin, Y. and Brangwynne, C.P. (2017) Liquid phase condensation in cell physiology and disease. *Science* **357**, 4382 https://doi.org/10.1126/science.aaf4382
39 Hyman, A.A., Weber, C.A. and Jülicher, F. (2014) Liquid–liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* **30**, 39–58 https://doi.org/10.1146/cellbio.2014.30.issue-1
40 Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K. and Sharp, A.P.A. (2017) A phase separation model for transcriptional control. *Cell* **169**, 13–23 https://doi.org/10.1016/j.cell.2017.02.007
41 Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V. et al. (2018) Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* **175**, 1842–1855.e16 https://doi.org/10.1016/j.cell.2018.10.042
42 Zamudio, A.V., Dall'Agnese, A., Henninger, J.E., Manteiga, J.C., Afeyan, L.K. and Hannett, N.M. et al. (2019) Mediator condensates localize signaling factors to key cell identity genes *Mol. Cell* **76**, 753–766.e6 https://doi.org/10.1016/j.molcel.2019.08.016
43 Cho, W.K., Spille, J.H., Hecht, M., Lee, C., Li, C., Grube, V. et al. (2018) Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412–415 https://doi.org/10.1126/science.aar4199
44 McSwiggen, D.T., Mir, M., Darzacq, X. and Tjian, R. (2019) Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. *Genes Dev.* **33**, 1619–1634 https://doi.org/10.1101/gad.331520.119
45 Gregor, T., Tank, D.W., Wieschaus, E.F. and Bialek, W. (2007) Probing the limits to positional information. *Cell* **130**, 153–164 https://doi.org/10.1016/j.cell.2007.05.025
46 Abu-Arish, A., Porcher, A., Czerwonka, A., Dostatni, N. and Fradin, C. (2010) High mobility of bicoid captured by fluorescence correlation spectroscopy: implication for the rapid establishment of its gradient *Biophys. J.* **99**, L33–L35 https://doi.org/10.1016/j.bpj.2010.05.031
47 Keenan, S., Blythe, S., Marmion, R., Djabrayan, N.-V., Wieschaus, E. and Shvartsman, S. (2020) Rapid dynamics of signal-dependent transcriptional repression by Capicua. *Dev. Cell* **52**, 794–801.e4 https://doi.org/10.1016/j.devcel.2020.02.004
48 Mir, M., Reimer, A., Haines, J.E., Li, X.-Y., Stadler, M., Garcia, H. et al. (2017) Dense Bicoid hubs accentuate binding along the morphogen gradient. *Genes Dev.* **31**, 1784–1794 https://doi.org/10.1101/gad.305078.117
49 Huang, S.-K., Whitney, P.H., Dutta, S., Shvartsman, S.Y. and Rushlow, C.A. (2021) Spatial organization of transcribing loci during early genome activation in *Drosophila*. *Curr. Biol.* **31**, 5102–5110.e5 https://doi.org/10.1016/j.cub.2021.09.027
50 Hodgkin, A.L. and Huxley, A.F. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**, 500–544 https://doi.org/10.1113/jphysiol.1952.sp004764
51 Bialek, W. (2012) *Biophysics: Searching for Principles*, Princeton University Press, Princeton, NJ, USA
52 Bialek, W., Steveninck, R.D.R.V. and Tishby, N. (2006) in *Proceedings of the IEEE International Symposium on Information Theory*, pp. 659–663, IEEE, Piscataway, NJ
53 Marzen, S. and DeDeo, S. (2016) Weak universality in sensory tradeoffs. *Phys. Rev. E* **94**, 060101 https://doi.org/10.1103/PhysRevE.94.060101
54 Meshulam, L., Gauthier, J.L., Brody, C.D., Tank, D.W. and Bialek, W. (2017) Collective behavior of place and non-place neurons in the hippocampal network. *Neuron* **96**, 1178–1191.e4 https://doi.org/10.1016/j.neuron.2017.10.027
55 Ramirez, L. and Bialek, W. (2021) Preprint arXiv:2112.14334.
56 Tesileanu, T., Conte, M.M., Briguglio, J.J., Hermundstad, A.M., Victor, J.D. and Balasubramanian, V. (2020) Efficient coding of natural scene statistics predicts discrimination thresholds for grayscale textures. *eLife* **9**, e54347 https://doi.org/10.7554/eLife.54347
57 Fairhall, A.L., Lewen, G.D., Bialek, W. and de Ruyter van Steveninck, R.R. (2001) Efficiency and ambiguity in an adaptive neural code. *Nature* **412**, 787–792 https://doi.org/10.1038/35090500
58 Laughlin, S. (1981) A simple coding procedure enhances a Neuron's information capacity *Z. Naturforsch. C* **36**, 910–912 https://doi.org/10.1515/znc-1981-9-1040

59 Cover, T.M. and Thomas, J.A. (2012) *Elements of Information Theory*, John Wiley and Sons, New York, NY, USA
60 Van Hateren, J. and Laughlin, S. (1990) Membrane parameters, signal transmission, and the design of a graded potential neuron. *J. Comp. Physiol. A* **166**, 437–448 https://doi.org/10.1007/BF00192015
61 Tkačik, G., Callan, Jr, C.G. and Bialek, W. (2008) Information flow and optimization in transcriptional regulation. *Proc. Natl Acad. Sci. USA* **105**, 12265–12270 https://doi.org/10.1073/pnas.0806077105
62 Nüsslein-Vollhard, C. and Wieschaus, E. (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature* **287**, 795–801 https://doi.org/10.1038/287795a0
63 Dubuis, J.O., Tkačik, G., Wieschaus, E.F., Gregor, T. and Bialek, W. (2013) Positional information, in bits. *Proc. Natl Acad. Sci. U.S.A.* **110**, 16301–16308 https://doi.org/10.1073/pnas.1315642110
64 Petkova, M.D., Tkacik, G., Bialek, W., Wieschaus, E.F. and Gregor, T. (2019) Optimal decoding of cellular identities in a genetic network. *Cell* **176**, 844–855.e15 https://doi.org/10.1016/j.cell.2019.01.007
65 Tishby, N., Pereira, F.C. and Bialek, W. (1999) in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing* (Hajek, B., Sreenivas, R.S., eds), pp. 368–377, University of Illinois, Champaign, IL
66 Alemi, A.A., Fischer, I., Dillon, J.V. and Murphy, K. (2016) Preprint arXiv:1612.00410.
67 Chalk, M., Marre, O. and Tkačik, G. (2018) Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl Acad. Sci. U.S.A.* **115**, 186–191 https://doi.org/10.1073/pnas.1711114115
68 Bauer, M., Petkova, M., Gregor, T., Wieschaus, E.F. and Bialek, W. (2021) Trading bits in the readout from a genetic network. *Proc. Natl Acad. Sci. U.S.A.* **118**, e2109011118 https://doi.org/10.1073/pnas.2109011118
69 Palmer, S., Marre, O., Berry, I.I.M.J. and Bialek, W. (2015) Predictive information in a sensory population. *Proc. Natl Acad. Sci. U.S.A.* **112**, 6908–6913 https://doi.org/10.1073/pnas.1506855112
70 Arnosti, D.N., Barolo, S., Levine, M. and Small, S. (1996) The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**, 205–214 https://doi.org/10.1242/dev.122.1.205
71 Kung, H.-F., Redfield, B., Treadwell, B., Eskin, B., Spears, C. and Weissbach, H. (1977) DNA-directed in vitro synthesis of beta-galactosidase. Studies with purified factors. *J. Biol. Chem.* **252**, 6889–6894 https://doi.org/10.1016/S0021-9258(17)39933-7
72 Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K. et al. (2001) Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 https://doi.org/10.1038/90802
73 Birnie, A. and Dekker, C. (2021) Genome-in-a-box: building a chromosome from the bottom up. *ACS Nano* **15**, 111–124 https://doi.org/10.1021/acsnano.0c07397
74 Doerr, A., Foschepoth, D., Forster, A.C. and Danelon, C. (2021) In vitro synthesis of 32 translation-factor proteins from a single template reveals impaired ribosomal processivity. *Sci. Rep.* **11**, 1898 https://doi.org/10.1038/s41598-020-80827-8