Towards a probabilistic theory of life[1]

Thomas Heams[2]

Comments by Franck Varenne[3]

Prisme N°12
October 2008

[1] This text is a translation of the transcription by Léo Coutellec of the seminar "The Probabilism Sessions", held on 21 March and 5 July 2008 at the Cournot Centre for Economic Studies, Paris.
[2] Associate professor of genomics at AgroParisTech.
[3] Associate professor of philosophy of knowledge at the University of Rouen.

Summary

Biology has long been dominated by a deterministic approach. The existence of a genetic code, even a "genetic programme", has often led to descriptions of biological processes resembling finely-regulated, precise events written in advance in our DNA. This approach has been very helpful in understanding the broad outlines of the processes at work within each cell. However, a large number of experimental arguments are challenging the deterministic approach in biology. One of the surprises of recent years has been the discovery that gene expression is fundamentally random: the problem now is to describe and understand that. Here I present the molecular and topological causes that at least partly explain it. I shall show that it is a wide-spread, controllable phenomenon that can be transmitted from one gene to another and even from one cell generation to the next. It remains to be determined whether this random gene expression is a "background noise" or a biological parameter. I shall argue for the second hypothesis by seeking to explain how this elementary disorder can give rise to order. In doing so, I hope to play a part in bringing probability theory to the heart of the study of life. Lastly, I shall discuss the possibility of moving beyond the apparent antagonism between determinism and probabilism in biology.

Introduction[4]

The object of this talk is to discuss the relevance of the probabilistic approach to understanding gene expression, a fundamental subject in biology. I would like to show you that such an approach, which can also be described as random or stochastic, is relatively new in our disciplines, both in terms of research programme and in terms of perspective for biology.

To start, I wish to explain the concept of random gene expression and why it is counter-intuitive or even contradictory to what we have been reading for years in the scientific literature in biology. Quotations from recent scientific articles suggest that the idea of a genetic programme is still very strong. The classic idea that there is a genetic programme that governs all cellular phenomena can be clearly seen in this affirmation: the large number of cell states that can be observed over the course of an organism's life and their reproducibility indicate not only the existence of programmes, but also of mechanisms that ensure their reliable reproduction (Martinez-Arias and Hayward, 2006). This idea is often developed in prestigious publications. It reaches its peak when it represents the regulation of embryonic development in terms of a network of very precise relations between the genes and the proteins, a sort of printed circuit. This representation gives the idea of a programme, similar to what one can find in a computer, which enables a given function to be performed (Davidson *et al.*, 2002). These works all have their strengths and relevance. Such developments all lie within a deterministic approach where each gene has a precisely-defined role attributed to it by a programme.

If I start here, it is also because the past few years have seen steady growth in the number of expressions and articles that go against these ideas. A first theoretical article published in 1983 (Kupiec, 1983) set out a probabilistic scenario for cell differentiation. An experiment conducted six years later was presented in an article about the transcription of any gene in any cell type, described as "illegitimate transcription" (Chelly *et al.*, 1989). This was a pioneering expression, and we must recognize it as such, but it still implied the existence of a "legitimate

[4] Authors' note: we have sought to preserve the oral nature of the presentation and discussion. Nevertheless, certain expressions inappropriate to written text have been modified in the transcription. The reader is therefore requested to consider the written version to be authentic.

transcription", from which we can sometimes escape. So this was still a deterministic view, but with some added qualifications. From 1994, articles started to appear affirming that "the transcription of individual genes in eukaryotic cells occurs randomly and infrequently" (Ross *et al.*, 1994). At the beginning of the 2000s, this view started to spread; we learned that "gene expression is a stochastic or 'noisy' process" (Swain *et al.*, 2002) and that "cells are intrinsically noisy biochemical reactors" (Thattai and Oudenaarden, 2001). Subsequently, stochastic mechanisms have been presented as being ubiquitous in biological systems (Ozbudak *et al.*, 2002), in contradiction with everything we had read and learned during at least 40 years.

What happened to cause this change in perspective? Before answering that question, I shall present two hypotheses about genetic expression. According to a first, classic hypothesis, under given conditions and in a homogeneous environment, for the cells of one same organ, which therefore possess exactly the same genes, the use and expression of those genes is homogeneous. According to a second hypothesis, these cells can have an unpredictable and random use of their genes, despite possessing the same genes and being in the same environment. It is these two hypotheses that I wish to balance. The initial question therefore becomes: "what happened in the space of a few years to cause this change in paradigm?". The change in perspective occurred with the adoption of new techniques. During the 1990s, technological advances made it possible to examine what happens in individual cells, and so to decide between the two hypotheses. If we look at cells separately in a given environment, and we adopt the first hypothesis, then we should expect to observe each cell synthesizing roughly the same thing (the same genes expressed, the same proteins). What we observe at the level of the population as a whole should therefore reflect fairly faithfully what happens in each cell. The random view, that is to say the second hypothesis, predicts that the average synthesis observed at the level of the cell population is only an average, hiding a certain variability. For many years this could not be tested, because the experiment was technically impossible. We had to wait for improvements in microscopy, enabling the study of individual cells, to discover this variability. In the deterministic frame of reference, we might never have observed it. The inter-individual variability of cells possessing the same genes has become a research subject of the highest importance. A brief bibliometric survey, such as searching *PubMed* for joint occurrences of the terms “stochastic + gene + expression” reveals the explosion in the number of publications on this theme, half of which have been written in the last five years.

A first demonstration of random gene expression

First I will give a concrete demonstration of random gene expression, before moving on to assess the consequences it may have. An experiment conducted by the Elowitz team, published in *Science* in 2002, explored the following question: is the gene expression of bacteria homogeneous or random? The protocol consisted in introducing two foreign genes into the deoxyribonucleic acid (DNA) of a bacterium. These genes enabled the synthesis of fluorescent red or green proteins in order to observe and follow the expression of these genes *in vivo*. By construction, there was no reason to believe that the level of expression of these two genes, in the same environment, would differ. This was the authors’ first hypothesis: whatever the time, roughly the same quantities of red and green proteins appear. The resulting fluorescence is on average yellow; it is always the same, and thus demonstrates this first hypothesis. According to the second hypothesis, the two genes have a random expression in relation to each other. Their functioning is unpredictable, and there is therefore no correlation between the levels of expression of these two genes – no correlation between the colours. There is therefore coexistence of bacteria, some greener, some redder and some yellow. Elowitz and his team tested these two hypotheses. Their publication described the first results, revealing heterogeneity in the fluorescence of the bacteria, even though, I repeat, they all possessed the same genome. This shows an unpredictable expression of some of these proteins from one cell to another, whereas it had hitherto been reasonable to believe that they functioned homogeneously. This experiment did not remain unnoticed for long: it made the cover of *Science*. That is not in itself a criterion of scientificity, but it shows how far we have come from the long-established viewpoint, challenging it and giving much food for thought.

If the phenomenon was limited to bacteria, it would already be very interesting, since bacteria constitute 99.9 per cent of living organisms on Earth, which is more than a simple detail. What is more, these organisms can be found in all environments, even the most hostile. Movements of random gene expression have also been observed in mushrooms, animals and plants, although there are less data available for this last group. At this stage, random gene expression has therefore been demonstrated experimentally and it is a widespread phenomenon of life.

The causality of random gene expression: biochemical and topological origins

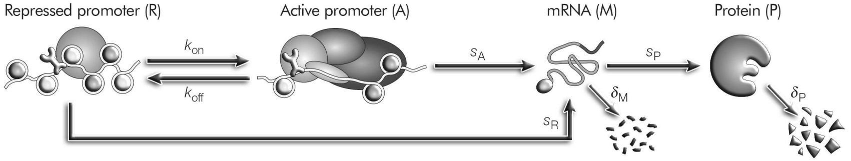

Fig.1 Multistep process of gene expression.

Why is gene expression random? What is it that makes it not strictly programmed as we might have thought? Here we need to do a bit of molecular biology and specify that gene expression is a process that occurs in several steps (see Fig. 1). The genes are located on the DNA; they are transcribed into RNA (a sort of intermediary, a small molecule which resembles DNA, but which is much shorter and which can move out of the nucleus where it is synthesized); this RNA serves to make proteins. The transition from DNA to proteins is not direct: the multistep process intervenes. At every step, it is possible that the process does not work perfectly; there is a certain room for manoeuvre. Differences from step to step can be observed between two cells although they are supposed to possess exactly the same genes. In a way that is counter-intuitive with respect to the classic literature, the association of nuclear structures clustered around the DNA is not as stable as usually thought. This association is very dynamic, which explains why it can create differences in the activity of the chromatin, which can have an influence on the first steps in the expression of genes. Moreover, the transcription activators, necessary for the DNA to be transcribed into RNA, are sometimes very low in number, sometimes with an average of far less than one per cell. In concrete terms, in some cells, *at a given moment*, there are none, while in some others there are more than one, which can create a difference between two cells. Likewise, the transcription initiation complex is also extremely dynamic. This results in constructing “transcription factories”, regions within the nucleus where the DNA is preferentially transcribed. This enables an increase, locally and at a given moment, in the probability of the DNA to be transcribed, but as a result it also decreases the probability of expression of all the other DNA regions. All these transcription factories – about 2,500 sites per nucleus at any given moment – do not amount to much when you consider that there are about 30,000 genes. Therefore, we begin to understand how, topologically, there could be an absence of homogeneity between two cells in the expression of certain genes. For this to happen, for example, it is sufficient that the transcription factories are not situated at exactly the same place; there is, moreover, no reason to believe that they have been programmed for that.

This first step enables us to refute the postulate of the constant availability of regulatory molecules. This postulate implies that there is always a sufficient number of available molecules to keep the machinery running smoothly. It would then be enough to send a signal to cells for them to express the genes corresponding to that signal. In fact, it is more complicated than that. Fundamentally, the proteins, especially those that serve to regulate gene expression or the functioning of the DNA, are sometimes very low in number. Note that 80 per cent of the proteins in a cell are present in numbers of less than 100 copies per cell, and this notably includes the proteins involved in very important phenomena such as cell division, the initiation of DNA replication or DNA reparation. These low numbers can lead to a sampling effect and variations between cells. Some cells may possess more copies than others, and therefore different gene expression activity, even though they have exactly the same genes.

The topological origins of the stochastic component

I must stress the topological origins of random gene expression. The cell is often represented as a bag containing water with a few molecules circulating freely within it. We must forget that representation. You only have to look at the picture of bacterium drawn by the biochemist David Goodsell (see Fig. 2) to observe molecular crowding. The cell is not a sort of swimming pool with a few molecules spread through

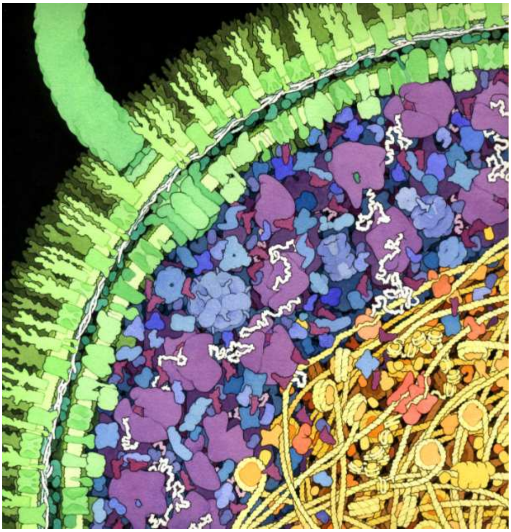

Fig. 2 Representation of molecular crowding in a cell (Goodsell. 1993).

it, but rather a high concentration of very different proteins and molecules of a large size. The light area in the bottom right-hand corner is where the genes are expressed in the DNA of the bacterium. Molecular crowding prevents the molecules from moving freely, and this can lead to a difference in accessibility depending on the time and the place at which the protein happens to be. I can complete this example by evoking the theme of what are called "chromosome territories", which have been studied for the last 10 years. In the next figure (see Fig. 3), you are looking at the nucleus of a chicken cell (*Gallus gallus*). You are probably used to seeing chromosomes represented in the form of clearly-defined little sticks, but here you can see that they have been decondensed to allow access to the genes ("allow" is perhaps rather teleological, but this is for the purposes of the demonstration). Although they are decondensed, each chromosome occupies a specific place within the nucleus, a territory that can be represented as a sort of skein of chromosomes. It has been shown that the level of expression of a gene depends on its distance from the centre of the nucleus and its position within a chromosome territory. Within these more or less fluctuating territories, a gene can, under the influence of cell divisions and the state of the cell, end up being more or less accessible to transcription factors, according to its position within the chromosome territory. It is therefore not enough to know the gene composition of a given cell, the precise sequence of its genome, to understand where the gene will be, physically and topologically, within the nucleus. And yet these data appear to be necessary for understanding the level of expression. Although it is possible to determine the broad trends, these examples all show that it is not enough

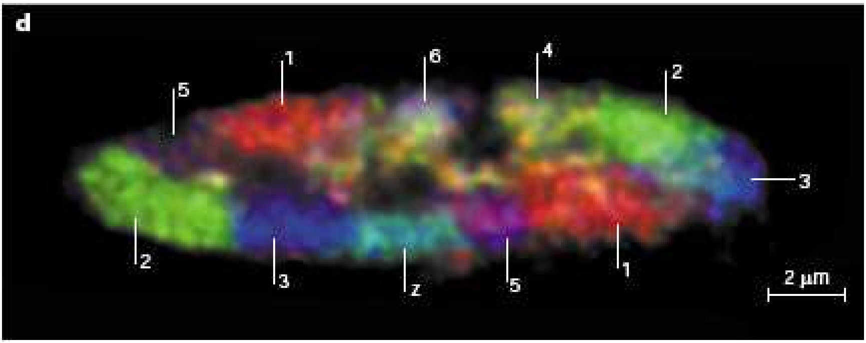

Fig. 3 The existence of functional chromosome territories in *Gallus gallus* (Cremer *et al.* 2001).

to know the genome of an individual, whether bacterial or animal, to predict what use will be made of its genes by the cells of that individual.

The modalities of random gene expression

Up to this point, I have shown that the phenomenon of random gene expression exists and is widespread, and what its biochemical and topological causes are. Now I will try to dissect this random expression. In this section, we shall look at the question of "noise" in gene expression. The concept of noise can imply a non-random way of functioning, with the noise being a sort of perturbation. That is not the idea that I shall develop later. But in this section, the term noise allows me to present an explanation. Let us return to our bacterial chromosomes with their two fluorescent proteins. Now, the postulate of homogeneous gene expression entails homogeneity, whatever the cells. In other words, by taking the level of expression of protein 1 and that of protein 2, at the intersection of these two levels, we find the place where the expected genetic expression will be located. We have seen that it does not happen like that. The dispersion of points is observed around the expected value, and this dispersion happens in two directions, allowing us to distinguish two types of noise. The first type, called extrinsic, depends on the concentration, state and location of the molecules that play a role in transcription, therefore in the first step of the gene expression process. This noise is responsible for the difference in configurations between two cells, or within one cell over time. The second type of noise, intrinsic, is due to the fundamentally random nature of microscopic elements, which creates differences between two "reporter" genes (used to identify a signal) in the same cell. In the case that interests us, it is rather this type of noise that is involved, although it is possible for a combination of the two to exist. The modelling of noise in gene expression has helped to distinguish between them and shown that they have functional consequences.

These noises are associated with an autocorrelation time, that is to say, a characteristic time for returning to the mean. If you imagine the mean expression of the genes, and you see a fluctuation in that expression, which you call "the noise", then the characteristic time it takes to return to the mean is what we call the autocorrelation time. It is independent from the magnitude of the noise. In other words, if you have two noises of pretty much the same magnitude, you can have a very short autocorrelation time, in which case there will be a lot of oscillations. If the autocorrelation time is longer, then you will observe undulations. The characteristic time of the extrinsic noise is longer than that of the intrinsic one: of the order of about 40 minutes for the former, compared with nine minutes for the latter. When one is in the time scale of several tens of minutes, it is interesting to note that there is possible overlap with the length of the cell cycle of certain bacteria. This is the same order of magnitude, meaning that the bacteria can end up being modified by the noise. This parameter must therefore be taken into account.

Control of random gene expression

We shall now try to understand how this random expression can be controlled. It appears to be necessary, at least as a first approximation, for the cells not to express any gene at any time, but that they should be able, as liver cells or skin cells, to follow the behaviour of relatively homogeneous liver cells or skin cells, so that the organ can function. Although this point can be called into question, let us start with the idea that the random expression needs to be controlled.

It can be controlled in several ways. The first is related to a simple effect of the mean. If, for example, a liver cell serves only to synthesize certain molecules for metabolizing or controlling the arrival of nutrients in the blood, then we can say that it is not too serious if some liver cells do this better than others. The important thing is for there to be a given concentration of these molecules in the total secretion. The functional effect does not take place at the level of individual cells, but at the level of the secretion of the organ; it is therefore a mean effect. The large number of cells makes it possible to rely on statistics. Yeasts provide another example: 75 per cent of the genes are at a level of expression lower than or equal to one transcript per cell. In other words, there is necessarily a wide variability in the contribution of these transcripts (and then of the proteins) within a population that is nevertheless homogeneous and genetically identical.

There is a second way of considering the control of stochastic gene expression. This involves taking into account the importance of the upstream steps in the random expression. If we go back a bit, we said that gene expression is a multistep process, that genes are "machineries" that do not function at full capacity and that if they do not function at full capacity, then some steps can be more efficient than others. If the upstream steps are more efficient than the downstream ones, then this tends to minimize the noise. All else being equal, when the transcription is more efficient than the translation, which is a downstream step, there is a lower intercellular variation in expression, and *vice versa*. It has also been shown, in yeasts, that when the activation of transcription (upstream step) is more efficient than transcription itself (downstream step), the noise is also limited. It should be understood that the more efficiently the upstream steps function, the more latitude there is for the downstream steps to function less efficiently, and there is still a reduction in noise. This helps to answer a question couched in evolutionary terms: if the downstream steps are not very efficient, why have the upstream steps, which are efficient (that is, they produce a lot, possibly a surplus, of molecules), been selected if these surpluses are not used by the downstream machinery? We can consider it as an "energy sink", evolutionarily selected to ensure the accuracy of the expression. That briefly covers the importance of the upstream steps.

Random gene expression can also be controlled by the number of copies of a gene. It has been established that the size of the noise is a decreasing function of the number of copies of the gene (Raser and O'Shea, 2004). In other words, the more genes there are (for example, 50 times in a genome), the more probability they have of being expressed, although there are channelling mechanisms to ensure that single-copy genes are found and used. Still, it is reasonable to believe that having several copies of one gene can guarantee an acceptable level of expression if there is topological randomness in gene expression (remember my earlier description of crowding inside the cell). This is therefore a possible evolutionary hypothesis for the keeping of several copies of certain genes after duplication. This phenomenon, which can be observed in different genomes, may also partly explain what is called polyploidy, that is, the fact of possessing several copies of given chromosomes instead of just one pair. The control can also be carried out by controlling the location of the gene copies, because, as we have seen, their position in relation to the centre of the nucleus and in relation to the periphery of a chromosome territory can also influence the level of expression. If we assume that these territories, the structure of which is currently being discovered, are not themselves formed completely randomly, then an evolutionary advantage could thus situate certain genes in certain regions of these territories. Once again, if we do not know the position of the genes in relation to each other within the volume of the nucleus, then simply possessing the sequence of this genome is not enough for us to determine the potential evolutionary advantage that certain topological positions might represent compared with others. Next, and this is more classic in molecular biology, we can talk about what is called the feedback loop. We can imagine that random gene expression is also the object of various feedback loops, especially "negative" ones. What is a negative feedback loop? If we consider a gene *a* that is going to be activated and therefore expressible, it is classically observed in biology that the downstream product of the chain of expression has a negative effect, in other words it tends to reduce the expression of this gene *a*. Logically, when the same gene *a* is expressed less, the downstream product is also less present, and therefore has less of an inhibitory effect on *a*: you can then observe an increase in the expression of *a*. In this way, the expression of the gene is more or less constantly maintained.

It can be observed, however, that this feedback is parameterized by a number of constants, notably the constant of association with the gene and the rate of destruction of this product into several by-products. These details are important, because one can observe that there is much less noise when the constant of association with the gene is substantially higher than the rate of destruction of the product of that gene. This shows that even the fairly classic mechanisms in biology can help to explain how random expression can be controlled.

The transmissibility of stochastic expression within a network of genes

To summarize the above: the stochastic or random expression of genes is a widespread phenomenon in life. I have presented the causes and shown that it is controlled. Now, with the help of recent results, I would like to explain how it is a transmissible phenomenon. First, it is transmissible throughout the cascade of gene expression. What is this cascade? When we say that genes function in a cascade, that means that often, a gene expresses a protein, which in turn influences the expression of another gene, and so on. So there are several intermediary genes between the first one activated and the final product of this cascade, a protein. Using fluorescence-based methods, it can be observed that the intercellular variability of expression and the response time for the expression of this protein both increase with the length of the cascade. This makes it possible to demonstrate something simple, namely that random gene expression can be transmitted from one gene to another. Noise can be transmitted all the way down the cascade (Hooshangi *et al.*, 2005), however long the cascade. We therefore conclude that noise is transmissible. This phenomenon, for which we have as yet a limited number of results, is nevertheless functionally useful.

Noise is transmissible within a cascade; now let us show that it is transmissible from one cell generation to the next. Noise is thus described as "heritable", in itself a surprising adjective. Let us take, in the case of yeast, a mother cell and a daughter cell. We would expect what happens in the mother cell to be independent from what happens in the daughter cell, which became an autonomous yeast following the asymmetric division of the mother. Being autonomous organisms, they are therefore independent. Recent studies, however, have produced counter-intuitive results. To understand them, we have to bring into play two pieces of information: the first is that yeasts can switch between "on" and "off" states. In other words, they can be in two different overall states of expression. Different groups of genes are expressed in each of these two states. They are, in a way, two different developments. We can reveal these states through the use of colouring techniques: the cells switch from one state to the other randomly. The second item of information is that we are capable, under specific conditions, of following the yeast cell genealogies: by setting up a video camera, we can observe what happens to the cells, making it possible to identify which cell came from which.

When we combine these two pieces of information, when we see cells switching from "on" state to "off" state and when we can tell whether or not these cells are mother and daughter, we can verify that the probability for a daughter cell to switch from on to off depends on whether the mother has switched from on to off. In other words, it is more likely that a daughter cell will switch from one state to the other when the mother cell has switched (which it does unpredictably). There is a kind of "heritability" of random behaviour. This curious and not yet clearly understood characteristic shows, in any case, that random behaviour can be inherited from a mother cell to a daughter cell. The fact that this random behaviour is transmissible and the object of inheritance leads us quietly towards the idea that these phenomena serve a functional purpose and can be used by the organism.

The biological significance of stochastic gene expression

What is the biological significance of random gene expression? How can we prove that the "background noise" version, which is in keeping with the classical view, is not the right one? According to this classical view, it is possible to envisage that cells are not robots, and that all the conditions are not necessarily strictly identical. It is therefore legitimate to imagine that on the margins, cells may behave slightly differently from each other. There is a genetic programme that imposes on cells precisely what they are supposed to do, and then from time to time they do not do what was expected. This is the first hypothesis that allows us to retain the classical paradigm. As we have seen that the phenomenon of randomness is widespread, controlled (that is to say that evolution appears to have led to mechanisms to control it), and more or less transmissible (with all due reserves on this last point), we challenge this status of simple "noise". Why should it not, on the contrary, be a biological parameter? Why should cells, or even organisms, not have been able to integrate this completely random functioning as a biological parameter providing them with flexibility in the face of environmental variations?

Let us take another example from the world of bacteria. Imagine two networks of genes, r1 and r2, functioning independently of each other in the bacterium. Each network of genes is controlled by its own specific promoter. At the centre of the r1 network, we have an essential gene (if this gene is not expressed the cell cannot function). The experiment consists in shifting this essential gene from the r1 network to the r2 network by modifying its promoter, replacing it by the promoter of the r2 network. Now we block the functioning of the r2 network (which is technically possible). This network, which now includes the essential gene that was originally in r1, ceases to function: none of the genes are expressed any longer. The bacterium is thus deprived of this essential gene, leaving it in a very bad state.

The experiment shows a drastic fall in the number of cells and then, although the mechanisms are not yet clearly understood, a number of solutions is found by a sub-population of the initial population. A solution emerges for the cells, maybe thanks to a gene that has not been used until now, resembling the one that has been moved and making it possible to start functioning again and reach an acceptable level of expression. Of course, this new population is checked to ensure that the essential gene from r1 is still controlled by the r2 network into which it was moved. To verify that there has not been an inversion, which could explain the recovery of the population, a second experiment consists in weakening even further the structure of the essential gene. Here we observe the cells displaying exploratory behaviour, which leads them to find a solution. This appears to show that the cells are capable of "digging through" their genes as if they were a toolbox, even though some of the genes have been prevented from functioning. A sub-fraction of the cell population (obviously not all of them, otherwise we would still be in the context of a programme) thus find a way of functioning in the face of an environmental problem. This is one way of explaining how the exploratory behaviour of cells, and the way they find possibly unexpected genes by "digging through" the gene pool contained in the genome, can be a means of adapting to unforeseen environmental variations.

Self-amplification and bistable equilibrium: the effect of random bifurcation

Another means of using random gene expression appears in the phenomenon of bistable equilibrium. Let us take a scenario that can be completely incorporated into the classic deterministic paradigm of gene expression (see Fig. 4). Take a molecule A capable of favouring the expression of two different genes, gene *b* or gene *c*. Gene *c*, if it is expressed, tends to inhibit *b*, and if *b* is expressed it tends to inhibit *c*. So it is either one or the other. The cell that possesses this network of bistability moves towards either the expression of gene *b* or the expression of gene *c* in a completely unpredictable way. Two cells in very close proximity possessing this exact network and in which the molecule A is activated can each go in opposite directions with regard to the expression of either *b* or *c*. It is therefore unpredictable and random, even if the network is completely deterministic. While the molecules have precise influence on the expression of a gene, we can see that they do not all go in the same direction. We shall see that biologically this can be very useful. Ultimately, this bistable equilibrium makes it possible to make ingenious use of a random and unpredictable feature of gene expression.

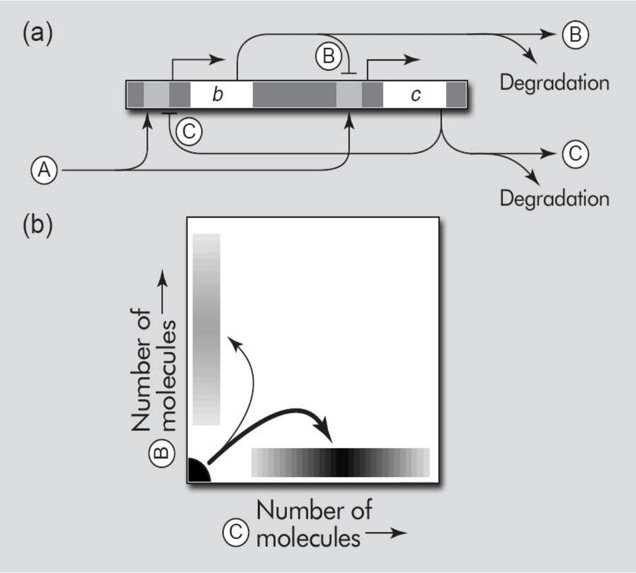

Fig. 4 Bistable equilibrium

Let us take the example of the eye of the *drosophila* (fruit fly), composed of facets for which the technical term is ommatidia (see Fig 5). Each of these ommatidia is composed of eight cells, and two of these eight cells are responsible for capturing light. For the fruit fly eye to function correctly, 30 per cent of the cells must be capable of capturing ultraviolet light and 70 per cent the visible spectrum. Let us imagine a genetic programme to achieve that: it would have to tell each cell what sort of light to capture and it would have to know that it indicated one thing to one cell and not to another in order to achieve the right proportions. With bistable equilibrium, there is no need for a precise programme imposing a precise future on each cell. The cells will have a 70 per cent chance of going in one direction of expression, enabling them to capture visible rays, and a 30 per cent chance of going in the other direction of expression, enabling them to capture wavelengths in the invisible spectrum. In the end, the probabilities of future state, for each cell, are transformed into a proportion of the population of cells. This is due simply to a phenomenon of bistable equilibrium, the proportions of which have been evolutionarily selected. Here we have an economy of effort: the cells can be seen to behave at random, and the desired result is achieved by means of evolutionarily selected proportions, rather than having an omniscient genetic programme. This means that a random dimension can be put to functional use. When we think in terms of individuals through a population, this functioning is very useful.

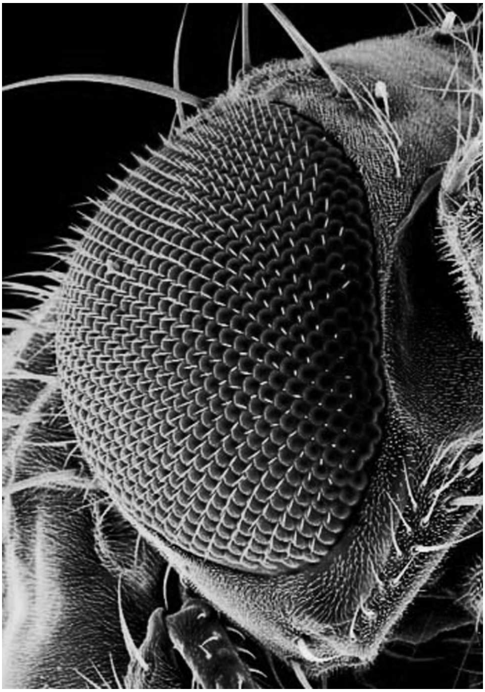

Fig. 5. Drosophila eye.

Other observed cases of stochastic differentiation

I now wish to show, without going into too much detail, that there are cases of random functioning of cell differentiation (what we have seen is indeed a kind of cell differentiation, when the cells are changed from one state to another) throughout the animal kingdom (using this rather dated term as a first approximation). It can be observed, notably, in a little nematode much used in the laboratory (C. *elegans*), which has the particularity of possessing a fixed number of cells, about 950. In developmental biology, they are typically used within a deterministic perspective, meaning that we can predict the future of each of these cells through the acquisition of a programme of precise instructions for each of them at each stage of development. What is interesting in this example, paradigmatic of deterministic development, is that some of these cells display random functioning, stochastic differentiation. This is the case in particular for the cells Z1 and Z4, which can differentiate into one of two different lineages, and when one cell differentiates into one lineage, the other cell differentiates into the other lineage. This is a good example, because even within an organism that is well-known for functioning, at least seemingly, in a deterministic fashion, there appear to be places where we can see a little bit of probabilism pop up.

In the clawed toad (*Xenopus laevis*), we can observe, even during the phase of embryonic development (when there are not a lot of cells), more or less random differentiation of the mesoderm. We observe that at each stage in the differentiation, only a small proportion of the cells become active, and that everything gradually becomes homogeneous. In other words, not all the cells react in the same way during the differentiation of the mesoderm, even though they are in one precise place and subjected to the same micro-environment. Likewise, they do not all react at the same time. Their behaviour therefore includes random responses.

Stochastic differentiation and stem cells

Now it is time to look at the modelling of cell differentiation. It was Jean-Jacques Kupiec, in 1983, who proposed the general idea that random gene expression was not necessarily a problem to be channelled, but a biological parameter that could be used by organisms. According to Kupiec, the appearance of “order” (cells that all resemble each other, for example, in the same tissue) can be explained in terms of a preliminary disorder. Cells initially possessing the capacity to switch unpredictably into different states, resulting in several different cell “types”, could have been evolutionarily selected to reach a stable state through interactions, for example, membranous interactions between surface molecules specific to each of these “types”. Thus, the cell populations of one “type” cannot reach a stable state until the second “type” appears. Each type appears, with a certain probability, at the cell level; it therefore appears in a given proportion at the level of the population. In this way, Kupiec formulated the idea that the stochastic behaviour of cells could be selected (in this case by stabilization): a sort of “random variation followed by selection”. The term “endo-Darwinism” was coined (Heams, 2004) to describe Darwinism within an organism. This is original insofar as it is a form of Darwinism involving cells that possess exactly the same genes, whereas Darwinian processes are generally processes of random variation followed by selection between individuals or entities that have different gene pools and that are selected on account of those differences. Kupiec postulated that it was not so much the gene composition of cells as the way those genes were used that enabled some cells to reach a stable state. Recently, through modelling research, he has modified these propositions slightly by explaining the phenomenon not in terms of potential membrane receptors, but by trophic interactions between cells of complementary “types” that appear stochastically. That is to say, the cells do not necessarily emit surface receptors, so much as secrete molecules enabling them to feed each other reciprocally, to be dependent on each other. For his pioneering theoretical propositions, which he sought to verify through experiments and modelling, Jean-Jacques Kupiec must be considered the father of the modern probabilistic theory of life.

Conclusion

I would like to conclude this presentation with a reflection on the possible compatibility between deterministic and probabilistic phenomena. Has life selected certain niches in which random phenomena can play a role within modes of functioning that are broadly deterministic? No, I simply defend the idea that on the contrary, all these phenomena take place within a general probabilistic context, with determinism being no more than an extreme case of probabilism. What I mean is that phenomena whose probability of occurring is close to 1 appear to be deterministic. A probability of 99.9 per cent is still a probability, and the functionally deterministic aspects of some biological phenomena are perhaps the evolutionary manifestations of a more general mode of functioning that is fundamentally probabilistic. It is observed differently depending on the constraints placed on these biological phenomena. We could represent this idea in the following way. Imagine that a given phenomenon has a probability of 1 in 1000 of occurring. Let us take a population of 100 cells where there is 1 chance in 1000 of the cells reproducing themselves. If you observe that the phenomenon is reproduced in all the cell populations – even those with less than 1000 cells – then you can say that there are channelling mechanisms at work to enable the phenomenon to take place. It thus appears to be deterministic. Now, let us take a population of 1 million cells with again a probability of 1 in 1000 that the cells will reproduce themselves. Here, you can be certain that the phenomenon will occur. You need only let chance take its course: about every 1 in 1000 cells (so, roughly 1000) will conform to this probability. These 1000 new cells will then multiply and in turn invade the population of the preceding generation. There is thus no need for any precise control, which is costly in energy and information storage.

Ultimately, we can say that there is a balance to be found between the sizes of the populations under consideration and the probabilities, per cell, of a phenomenon occurring. If the population is of a large enough size in relation to the probability of the elementary event, then the actual occurrence of the event is no miracle. Of course, these assertions need to be qualified in a number of ways. The sub-population that corresponds to the stochastic event must be able to spread; the mechanisms must be found to enable this population to be homogeneous at that precise moment in time; the observed duration of the transition between the two states must be modulated. Given the appropriate means, all these little qualifications could enable us to calculate a sort of synthetic index to determine the relevance of invoking either stochasticity or more deterministic mechanisms of channelling. In the coming years, this would allow us to define a refutable domain of validity for stochasticity in gene expression and cellular biological phenomena.

Comments and questions from **Franck Varenne**
Answers from **Thomas Heams**

Franck Varenne

First, I would like to return to two passages in your presentation. You started by recalling that a number of classical postulates have been refuted by many recent publications. You described recent developments: the idea of stochasticity in gene expression is really starting to spread. What I found striking was the idea that the postulate of homogeneity was linked to an instrument. In the past, techniques did not allow us to access the components of individual cells, but only populations. How do you view this link between the history of instruments and the shift in paradigm that appears to be happening now? Second, what seems to me to be just as decisive is the postulate of constant availability that takes us to the level of the individual to inform us that the cell is not a swimming pool with the components floating around in it, but that there is a sort of scarcity and a kind of management of that scarcity, which is also related to the appearance of visualization tools. Has the individual cell approach supplanted the cell population approach?

Thomas Heams

On the postulate of homogeneity, I went over that rather fast and it's true that I attributed that first to developments in techniques. However, some of my colleagues have qualified the importance of the role of instruments: it has been technically possible to detect heterogeneity since at least as early as the 1990s, although not, admittedly, in such detail as I have been able to do today. In particular, there were flow cytometry techniques, enabling cells to be classified by size, nucleus size or density. This technique has been used for many years to draw scatter plots, which show that cells all placed in the same environment do not all respond in the same way, that they have, on the contrary, room for manoeuvre. This did not particularly awaken the interest of biologists, in the sense that they stuck to the idea that admittedly, cells are not robots; they do not all do exactly the same thing, but overall they do, and so one should focus on the centre of the scatter plot, where the true value lies. In the end, this missed opportunity leads us to wonder whether it was not a problem of technique. For my part, I don’t think so. Molecular biology exploded at the end of the Second World War, especially during the 1950s when James Watson and Francis Crick discovered the molecular structure of DNA, in a context of strong growth in computing, information theory and the idea of programmes, of cybernetic determination. All the early models, including the simplest ones, functioned by the transposition of deterministic concepts to biology.

Franck Varenne

In a very conceptual context, therefore, focused on the mean, with the concept of noise coming from signal theory, for example.

Thomas Heams

Exactly. I think there was a kind of percolation, or at least infusion, of this concept that was very strong. Information theory and computing produced very concrete, powerful results. There were technical limitations, but there were also real presuppositions that went unnoticed. In particular the presupposition that all the cells studied were equal to the average obtained from a mass recovery of proteins from a tissue. This was a real presupposition, *a priori* we could have said from the start that there was a Gaussian distribution and real variability between cells, but we preferred to consider them as homogeneous.

Franck Varenne

At the same time, in biological systems, as Erwin Schrödinger observed, when one starts to deal with the microscopic level, one cannot apply the law of large numbers. This is perhaps another problem of the statistical approach?

Thomas Heams

Schrödinger’s book (1944) is fascinating. The author was one of the great thinkers on probabilism in physics; he tackled biology with intelligence and with a sort of anticipation, as he foresaw, in writing about the aperiodic crystal, what the molecule of DNA would be like eight years before it was demonstrated by Watson and Crick. In his approach to biology he made calculations that were formally correct, but he failed to calculate the right things. He calculated the number of atoms there could be in a molecule and asked whether those atoms could function according to the law of large numbers and thus allow a probabilistic mode of functioning. He did not imagine – he couldn't have done at the time – how all the genes can interact with each other. With 30,000 genes in an organism, which can all interact with each other, we have probabilities of interaction that exceed the number of molecules in the universe. So we can say that Schrödinger came very close to a marriage between probabilism and biology; he opened the door, but then closed it too quickly.

Franck Varenne
Because he didn't incorporate a combinatorial dimension?

Thomas Heams
Not on the right scale, in any case. I feel this to have been an incredible missed opportunity, and we kept a largely deterministic functioning with, however, arguments that are understandable. Before discovering how 1000 genes function, it is worth trying to understand how genes in a small network function in relation to each other with a protein that may or may not express itself. So we started with very simple things; this was all the work of the school of François Jacob and Jacques Monod, who cleared an immense field. I have no intention of questioning the scale of this work. I just want to show that it was approached from a deterministic conceptual perspective, which didn't cause any problems during the early years, but which did cause problems later. For example, when the question arose of the genes responsible for cancer (a programme in which billions of dollars were invested in the 1970s) the answers could not be found, despite all the power of North American molecular biology. All the genes identified over the last 30 years are involved in other things besides cancer. We haven't found the specificity we expected. So much for homogeneity. As for availability, we have observed that the cells are heterogeneous and that there are problems of availability with certain molecules. Some molecules are only present with an average of one or two copies per cell, so with a Gaussian distribution some cells have none and others have three or four. This can trigger threshold effects and differences in the functioning of cells. I should add, to qualify things a little, that although cells are not the swimming

pools through which molecules can spread without restraint, recent research has nevertheless shown that there are very fast molecular movements within cells. Despite the topological crowding that has been observed, the molecules move much faster than one might have thought possible. This is an element that counterbalances part of my argument.

Franck Varenne

One thing that particularly struck me was this return of space, this return to topology. Taking into account the asymmetries of space and the variety of topologies allows us to observe that the level of expression cannot be known only through the sequence. This is a very important point, and what you told us must be emphasized: it is not enough to know the genome sequence to understand what is going on.

Thomas Heams

I'm not saying that knowing the sequence is of no interest; I am myself involved in sequencing research, and for good reason: the work of sequenced animal genomes is precious. However, that is just as true for all the work on chromosome territories, nuclear topology, and I'm thinking in particular of the Cremer team in Germany, pioneers in the field. Knowledge of the topology greatly influences our understanding of the genome. And the topology is typically outside the genome. This ties up with a general theme called epigenetics, which is very fashionable at the moment, for good or bad reasons. The term epigenetics itself is worth studying. We could go back to the time of Ptolemy, who had a system for explaining celestial mechanisms, and when nothing worked out right he added epicycles as *ad hoc* explanations of everything that didn't work right in his system. With the term epigenetics you get the feeling that it lumps together everything that genetics cannot explain, whereas a few years ago genetics was supposed to explain everything: it covers chromosome territories, the methylation of genes, modifications of DNA that are not visible in the sequence, but which can influence the function of each gene (the same genes do not necessarily function in exactly the same way). In any case, genomics on its own is not a sufficient tool for understanding life.

Franck Varenne

This is precisely the point I wanted to come back to. It appears to me that there is a sort of theoretical ambition when you say, first, that we can explain how forms of determinism are due to highly-channelled stochastic behaviour, and second, that this type of modelling, with reference to the work of Jean-Jacques Kupiec, is economical. That strikes a chord with an epistemologist, because there is a tradition of theoretical research under conditions of the economy of explanatory postulates. Third, that raises the question of causal explanation in biology. Lastly, you spoke at one time of the exploratory behaviour of cells as if there was a form of intelligence at work: we might think of artificial intelligence programmes with trial and error or genetic algorithms: this is the question of teleology.

Thomas Heams

Yes, I need to be clearer about this last question. When I spoke about the exploratory behaviour of genes I was not attaching any particular intelligence to any of these cells; the question of intelligence is beyond me. To put it in concrete terms, I am just speaking of a cell population in which the cells do just about anything, under conditions of stress, for example, and some of them find the right combination by chance. There is random behaviour during a given period. Moreover, one can well imagine that if, after a given time, none of the cells has found the right combination of expressible genes, because the probability of finding the right solution is too low, then the population will die. I do not, therefore, attribute any particular intention to them. For the principle of economy, that is something I am attached to. Moreover, it ought to be demonstrated, it is not as simple as that, but I reason in evolutionary terms. I work on the assumption that making DNA, controlling the reliability of that sequence of DNA, has an energy cost, and I consider to be economical anything that enables cells and organisms to function acceptably in a given environment with a certain amount of flexibility without necessarily having to encode everything in the DNA; I take that to be economical. But I am aware that this is open to debate.

Franck Varenne

I was thinking of the epistemology of economical theories, where a small number of economical principles are sought for reasons of epistemic convenience. Here, your economical approach appears to enjoy further grounding, as we see that life is in itself economical. For many physicists, as well as epistemologists, science must seek to be economical in its principles: they must be kept to a minimum. To explain, in a sense, always means to condense, to make a coherent unit or single identity out of varied parts, or, at the least, a small number of enunciations. For someone such as Ernst Mach, for example, science must be considered above all as a strategy of "the economy of thoughts" (1893). He held that scientific laws are nothing more than convenient summaries of experimental events. Your approach also has this economical concern. The difference is that you offer theoretical proof that is grounded in the object of study itself and not limited to a general strategy of knowledge (a strategy that would only be epistemological in this context). Such a stochastic conception makes it possible for life to have complex and adaptive behaviours without everything being programmed or without scientific laws taking into account every detail.

Thomas Heams

The idea of *programme* (from the Greek "written in advance") is an anthropomorphic projection; in fact nothing is written in advance. DNA is fascinating, but it is not central. There is no centre in a cell; there are proteins and there is the DNA that produces those proteins, but without the proteins there can be no DNA. That is all thoroughly distributed, decentralized. So I feel that using the concept of programme means making a hypothesis in epistemological terms that needs to be verified. We must not reverse the burden of proof. I find that asserting the existence of a programme is very heavy in scientific terms. To me, it appears more intellectually satisfying to work on the assumption that there may not be one, and that there are channelling phenomena that sometimes make it look as if there is one. I would just like to add that I am aware – and this is a problem that faces biology in general – that after a presentation like this one, I could be accused of choosing the 10 examples in the literature that illustrate randomness, and of reasoning through selected hypotheses and well-chosen examples. This is true for all demonstrations in biology; we do not have the theoretical power of physics, and it is a challenge for biology to produce demonstrations. At this stage, the proposal that appears honest to me is to understand, not how to weigh determinism and probabilism against each other, but how we can set up a dialogue between them and bring them together in one overall view. I would not be happy simply setting some examples against others.

Franck Varenne

In that case, determinism would be a variant, even a degeneration of probabilism. This leads me to another question. There was something of a Popperian flavour to some of the views you expressed. One could turn that back against you by saying: "Are we not going to end up with an irrefutable theory if we say that determinism is simply when the probability tends to one?" My second question is: does this not also raise the issue of changes in level? Thus, the causal explanation should not be sought at the level of the molecule, but perhaps at another level, which remains rather unclear for me, that of the population?

Thomas Heams

This is what I tried to show at the end of my talk. Obviously, when a phenomenon has a very low probability of occurring and yet it occurs in every generation, in all the "4-cell stage" embryos, you cannot speak solely of probability, there are mechanisms in place that allow some kind of channelling to take place. When the ovum is fertilized, some RNA are present that enable the cell to divide and produce proteins during the first divisions before it uses its own genome. There, we can say that there is a kind of determinism at work. I have no problem with that, I am not trying to refute determinism. On the other hand, there is one thing that appears fairly coherent to me, and that is to believe that the general mode of functioning is probabilistic and that there can exist phenomena of channelling.

Franck Varenne

It's true that mathematical techniques exist for refuting a probabilistic model. Now, I would like to move on to a more general question to wrap up our discussion, a question that falls within a perspective of the history of science. There is perhaps a tradition, in biology, of repeatedly returning to probabilism. Biometrics, for example, began fighting against the physiological approach at the beginning of the twentieth century. Biometrics – born out of the systematic questioning of the spread of genetic characteristics – developed new mathematical instruments of observation and discernment that showed that there was stochastic behaviour at almost every level of life, including, for example, the metabolic or biomechanical level. Is what you are experiencing in molecular biology today something of this nature? Do you see similarities with the biometricians of the 1920s and 1930s who warned biologists against naïve causal approaches claiming that causality can only be observed on average behaviours? There is a persistent tension between the traditional physiological approach – first illustrated by the notion of internal environment and later closely linked to the notion of average behavior – and the biometrical approach – here, biology studies something that is changing and subject to unknown factors and fluctuations, in line with the mathematical theory of probability of Francis Galton and later Ronald Aymler Fisher. This tension is still perceptible in works such as the manifesto by Eugène Schreider (1967). Schreider wrote that there is no functional link in biology; an organism is nothing more than a large number of random variables, a "multivaried stochastic process". In the same vein, Fisher, in 1934, put forward an indeterminist and probabilistic vision of the causal system as a whole, and in particular of biological systems.

Thomas Heams

Although some of my propositions may be slightly provocative, my intention is not to discredit past approaches. I include them in a historical whole, and I am aware of how much I owe them. There can be no ambiguity about that; biology would be nothing without the deterministic work of molecular biology carried out after the Second World War. Nevertheless, I have the impression that things are evolving radically. The probabilistic movement in genetics has this ambition to stimulate a reaction against the excesses of determinism.

References

Chelly, J., J.-P. Concordet, J.-C. Kaplan, A. Kahn (1989), "Illegitimate Transcription: Transcription of Any Gene in Any Cell Type", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, n°8, pp. 2617–2621.

Cremer, M., J. von Hase, T. Volm, *et al.* (2001), "Non-Random Radial Higher-Order Chromatin Arrangements in Nuclei of Diploid Human Cells", *Chromosome Research*, vol. 9, n°7, p. 541–567.

Davidson, E.H., J.P. Rast, P. Oliveri, *et al.*, (2002), "A Genomic Regulatory Network for Development", *Science*, vol. 295, n°5560, pp. 1669–1678.

Elowitz M., A.J. Levine, E.D. Siggia, P.S. Swain (2002), "Stochastic Gene Expression in a Single Cell", *Science*, vol. 297, n°5584, pp. 1183–1186.

Felix, M.A. and P.W. Sternberg, "Symmetry Breakage in the Development of One-Armed Gonads in Nematodes", *Development*, vol. 122, n°7, p. 2129–2142.

Fisher R.A. (1934), "Two New Properties of Mathematical Likelihood", Proceedings of the Royal Society, A, 144: 285–307.

Goodsell, D. (1993), *The Machinery of Life*, New York: Springer-Verlag.

Heams, T. (2004), *Approche endodarwinienne de la variabilité de l'expression génétique*, doctoral thesis, INA P-G.

Hooshangi, S., S. Thiberge and R. Weiss (2005), "Ultrasensitivity and Noise Propagation in a Synthetic Transcriptional Cascade", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, n° 10, pp. 3581–3586.

Kupiec J.-J. (1983), "A Probabilistic Theory for Cell Differentiation, Embryonic Mortality and DNA C-Value Paradox", *Speculations in Science and Technology*, vol. 6, n°5, pp. 471–478.

Kupiec J.-J. (2008), *L'origine des individus*, Paris: Arthème Fayard.

Mach, E. (1893) [1883], *The Science of Mechanics: A Critical and Historical Account of Its Development*, [*Die Mechanik in ihrer Entwickelung Historisch-kritisch dargestellt*], translated from German by Thomas J. McCormack, Chicago, U.S.A.: The Open Court Publishing Company.

Martinez-Arias, A. and P. Hayward (2006), "Filtering Transcriptional Noise during Development: Concepts and Mechanisms", *Nature Reviews Genetics*, vol. 7, n°1, pp. 34–44.

Ozbudak, E.M., M. Thattai, I. Kurtser *et al.* (2002), "Regulation of Noise in the Expression of a Single Gene", *Nature Genetics*, vol. 31, n°1, pp. 69–73.

Raser, J.M. and O'Shea, E.K., (2004), "Control of Stochasticity in Eukaryotic Gene Expression", *Science*, vol. 304, n° 5678, pp. 1811–1814.

Raser, J.M and O'Shea, E.K. (2005), "Noise in Gene Expression: Origins, Consequences, and Control", *Science*, vol. 309, n°5743, p. 2010–2013.

Ross, I.L., C.M. Browne, D.A. Hume, (1994), "Transcription of Individual Genes in Eukaryotic Cells Occurs Randomly and Infrequently", *Immunology & Cell Biology*, vol. 72, n°2, pp. 177–185.

Schreider, E. (1967), *La biométrie*, Paris, PUF.

Schrödinger E. (1944), *What Is Life?*, New York: McMillan.

Swain, P.S., M. Elowitz, E.D. Siggia (2002), "Intrinsic and Extrinsic Contributions to Stochasticity in Gene Expression", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, n°20, pp. 1279–12800.

Thattai, M. and A. van Oudenaarden (2001), "Intrinsic Noise in Gene Regulatory Networks", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, n°15, pp. 8614–8619.