1

Status: it's complicated!

In 1635, the English scientist Robert Hooke made a fantastic discovery. Studying a slice of cork through a microscope, he discovered cavities that, as he said, used to contain the "noble juices" that once had nurtured the living tree from which the cork had been cut. He called the cavities *cellulae*, a Latin word for storage rooms and the root of the term that we still use today: cells. Because the cork was dead, Hooke was only able to see cell walls forming a honeycomb structure. As exciting as this discovery must have been at the time, we now know that Hooke merely saw the tip of the iceberg, actually missing most of what makes a *cellula* such an impressive object. Peeking into a cell today with an optical or electron microscope, we see how a whole new world of structures and molecules opens up. Most cells have a nucleus, mitochondria, and ribosomes, and there are all kinds of small organelles, vesicles, and membrane enclosures. Going even further, modern visualization and tagging techniques of molecular biology allow us to see more and finer structures, all the way down to the level of large individual molecules. A whole world, invisible to the human eye, is emerging, leaving no doubt: life is complicated!

We systems biologists love to work on real puzzles. For many of us, living systems are huge Sudokus, where some information is available, but lots of gaps in-between are to be filled in through experiments and with advanced logic that evaluates the experimental results in a systemic context. Trying to figure out how the multitudinous parts in cells work together to create something as incredible as a brain is very attractive to us. We are fully aware that we will not solve the whole puzzle in our

lifetimes, but nature is modular, and every systems biologist hopes to solve a large sub-puzzle, or at least a few smaller puzzles. The intellectual challenge is the enormous complexity of every cell and organism, which requires us to invent new tools and methods, and that's what systems biology is all about.

The complexity of living systems is due to different features. First and foremost, there are just very many parts. The lowly bacterium *E. coli* contains between four and five thousand genes. Nobody is sure how many different proteins are in a single plant or animal cell. Suffice it to say, there are easily tens of thousands. Our brains allegedly contain several hundred trillion connections between neurons, and the human body altogether supposedly consists of roughly five octillion atoms, that is:

5,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000.

Now, it is unlikely that we will ever have the need to follow each atom explicitly, but it does seem rather evident that a deep understanding of the inner workings of life will require at least a good account of our genes and proteins. And therein lies our first grand challenge. The cellular Sudoku consists of a huge number of grid boxes. These sheer numbers already tell us that we need to generate our own support strategies, starting with tools that can handle the enormous bookkeeping tasks we have at our hands. Fortunately, computers are very good at dealing with gazillions of data points and not forgetting any of them.

While storing information is a considerable challenge, data points by themselves are usually not all that stimulating. Much more intriguing are the interwoven processes that lead to the data and, in particular, features of biological systems that look quite harmless at first, but can really play games with our minds. One such feature that we often take for granted, but that biological systems often violate, is linear scaling. We like to expect that there is a strict correlation between an input and the corresponding output. If we invest $1,000 in the stock market and after a couple of years receive $1,100 in return, we would have received $110,000 had we invested $100,000. Not so in biology. If we fertilize our roses with 2 tablespoons of fertilizer, they might produce 50 blossoms, but fertilizing them with 200 tablespoons will most assuredly not result in 5,000 blossoms. Rather, the roses will probably die. Thus, more is not necessarily better in biology, and there is often a strongly reduced return on investment. Many biological phenomena are even more complicated

in that both very large and very small inputs are disadvantageous, if not lethal, while normal functionality requires an input of just the right intermediate magnitude. Other systems, particularly in the microbial world, are so robust that they can tolerate quite large variations in their environments without changing their functionality in response. So, the assumption of linearity is more often than not a problem in biology, and it is clear that we need nonlinear thinking. But that is a real challenge as nonlinear input–output relationships make it difficult for us to predict responses to perturbations to which a cell or organism is exposed.

Related to the issue of linear scaling is the principle of superposition, which is a cornerstone of many applications in engineering and in daily life. It addresses the relationship between inputs to a machine or system and the outputs with which it responds. The principle says that, if some Stimulus-1 leads to Response-1 and if Stimulus-2 leads to Response-2, then the Response to Stimulus-1 and Stimulus-2 together is equal to the sum of Response-1 and Response-2. As an example, consider a car on a suspension bridge. Due to the car's weight the bridge bends down very slightly. For a truck, the bending is a little stronger. If the car and the truck are both on the bridge, the total bending is the sum of the two. A typical example from physics is the force (vector) that results from two independent forces (vectors).

Biological systems often operate quite differently. Two inputs may lead to a much stronger response than expected or, in other cases, the response might be much weaker. More than 2,000 years ago, Aristotle already pondered what has almost become a cliché; namely, that a system can be more than the sum of its parts and that there may be synergism between actions or processes. He did not make up the fancy word *synergism* for this observation, as it is simply Greek for *collaboration*. Instead of collaborating and enhancing each other, it is also possible that processes work against each other and, in an antagonistic manner, diminish each other's impact. As an example, suppose that a cell can generate a metabolite through two pathways. If the activity of one is increased, the activity of the other is usually slowed down. An important advantage provided by the superposition principle, if it applies, is that it is legitimate to analyze parts of a system with respect to their input–output relationships one at a time. Recording all results then allows valid predictions of how the system will respond to combined inputs. Many engineered systems are designed such that the superposition principle holds and this type of

analysis can be performed. The situation is dramatically different in a synergistic biological system, which often shows responses that none of the components by themselves can explain and where parts or subsystems, taken out of their contexts, frequently cease to function altogether. A lot of research has been devoted to creating artificial environments in which subsystems work properly, but the fact that isolated parts act differently if they are inside or outside their normal milieu remains a formidable challenge.

Western tradition trains us to think in chains of causes and effects. Our brain is good at following these chains, even if they are as convoluted as a Rube Goldberg contraption. However, as soon as the situation is a little bit more complicated, we scratch our heads. Let's do a thought experiment with the ubiquitous example of a negative feedback loop. Imagine a chain of events involving a gene G, which, when expressed, leads to the formation of the matching mRNA R, which is translated into a protein P. Let's assume that P is an enzyme that catalyzes the production of metabolite M. If that is all there is, then we easily predict what happens if G is expressed. Namely, in strict order we will see R, P, and finally M appear or rise in amount. If G is turned off, R, P, and M will eventually disappear. Now suppose a seemingly small addition to the system in the form of M directly affecting the expression of G. This situation is quite frequent and we find it, for instance, in the famous lactose genes that led to an enormous spectrum of insights into gene regulation and for which François Jacob, André Lwoff, and Jacques Monod obtained a Nobel Prize. Let's suppose that M represses the expression of G. Our natural thought process then is probably the following. If G is expressed, we find more R, more P, and more M. Now, the increase in M feeds back and represses G. As a consequence, we'll have less G, less R, less P, and less M. Less M means less repression of the expression of G, which should lead to more G, R, P, M, and so on. What should we conclude? The system appears to oscillate, but does it really and, if so, for how long? Hmm. We might believe there must be a way of solving the puzzle by thinking harder, but the truth is that we cannot figure it out with the information given. The reason is that the response of such a feedback system depends critically on the numerical features of the system, such as the number of events in the chain, the time delay between the expression of G and the production of M, and the strength of the feedback. As a

daily life example, imagine a picky and impatient copilot in your car who finds the temperature much too cold. So he turns on the heater. It takes a while for the car to warm up, and all of a sudden it is too hot. So he turns the heating off. If the temperature adjustments are done in a well-measured manner, the temperature will eventually be just right, but if the copilot keeps overreacting with his feedback, it may oscillate between too cold and too hot for a long time. Nature is full of feedbacks, and they are often nested in complicated patterns. Biological systems also use feedback activation, as well as feedforward regulation. Furthermore, signals may compete with each other. If a process is simultaneously activated by one factor and inhibited by another, which one will win? Actual systems, even of moderate size, contain many such controls, and these often form complex regulatory webs. They constitute intriguing puzzles, which we can only solve effectively by setting up mathematical models.

The regulatory webs in cells and organisms often work simultaneously on different organizational scales and different time scales. For instance, cells exposed to physical or chemical stresses may alter their gene expression and protein profile, as well as their metabolite concentrations, in response. Multicellular organisms furthermore show different fast and slow physiological responses, such as shivering for immediate warming and creating a fat layer to ward against cold temperatures that occur on a regular basis. Over short periods, regulatory control systems ensure that the cell or organism remains close to its normal state of homeostasis. However, if the system is perturbed for a long time, the regulatory web mounts an adaptive, long-lasting response. In many cases this adaptation is successful, and the cell or organism lives essentially a normal life, even though it is exposed to inferior conditions. In other cases, the system may settle in a different, suboptimal state, such as a disease.

A good example of slow and fast adaptation is the reward system in the human brain. This system responds to pulses of neurotransmitters, such as dopamine and glutamate. Eating a piece of chocolate or looking at a beautiful sight triggers dopamine production and makes us feel good. However, if we repeatedly trigger or even overwhelm the reward system, for instance with the recreational drug methamphetamine, the number of dopamine receptors eventually begins to change, and we require ever-more input to achieve the same reward of feeling good. This regulatory response occurs at several organizational levels in the

brain and involves electrical signals, metabolites, proteins, genes, and the restructuring of cell membranes.

A different, deceivingly simple-looking puzzle comes from threshold effects: a slight increase in some concentration or amount leads to a correspondingly slight change in some output, whereas a stronger increase has a totally different effect. An intuitive illustration is a visit to the beach. If we stay for a short period of time, our skin does not respond much; if we stay longer, we may develop a tan; if we stay too long, we go home with a bad sunburn. Tanning and sunburn are very different biological responses, but unless we have a lot of experience with our own body, it is difficult to predict when exactly one is replaced by the other. We experience the same type of input (sunlight) in both cases, but somewhere there is a threshold in its amount, which distinguishes one response from the other. Thresholds are very common in nature, and one might speculate that many diseases are caused by processes exceeding their "normal" thresholds. The challenge is that we do not really know what these thresholds are in most cases, and if several thresholds are in play, predictions regarding the responses of the system become very difficult. Moreover, thresholds frequently change over time. Systems get used to stresses, they adapt, and this adaptation may happen within hours, months, or on an evolutionary time scale.

Our innate way of linear thinking in terms of causes and effects has dominated biological research for a long time. In particular, it is directly in line with the paradigm of reductionism. The core idea of reductionism is that knowledge of all parts of a biological machine will tell us how the machine works. Therefore, understanding how an organism functions requires that we understand what its organs do. To understand organs, we investigate tissues and cells. In order to understand cells and their function, we study the details of intracellular structures, processes, and molecules. The implicit expectation is the following: if we work hard enough and characterize every constituent within a cell, we will grasp the secrets of biology. There is no doubt that we need to know the parts of biological systems and their features, but is this knowledge sufficient? The answer is no; we need additional techniques for putting the system together again.

A prominent example is the Central Dogma, which Nobel Laureate Francis Crick proposed about half a century ago and which somewhat simplistically states the following. Genes consist of DNA. DNA is

transcribed into a matching RNA, and this RNA is translated into a string of amino acids, that is, a peptide or protein. Proteins are responsible for the processes of daily life; for instance, by providing physical structure and serving as signaling molecules or enzymes that control metabolism. The Central Dogma proposes a neat linear chain of causes and effects, and the entire blueprint of life therefore appears to be coded within the genes. Following through with this argument, it seems that we need to know the genes, and all else will follow. The reductionist paradigm therefore emphasizes the identification of genes and genomes very strongly.

The basic tenets of the Central Dogma are still undisputed, and genes do contain an enormous amount of information. However, the more we learn, the more we fathom how complicated the details are. We now know that the process of expressing a gene is controlled by transcription factors, which are proteins, and also by the three-dimensional structure of DNA and the way it is stored in cells. Additionally, expression may be affected by repressors or inducers, for instance in the form of the metabolite that is the ultimate product of this very gene. The seemingly simple inclusion of these modulators of gene expression introduces an enormous complication, as we discussed earlier, because instead of a linear chain of causes and effects, the Central Dogma has become a feedback system with different control loops. To make things even more interesting, rather than just one transcription factor per gene, there are often many, and the transcription factors themselves form hierarchical networks, where a high-level transcription factor controls the expression of numerous genes that in turn produce transcription factors controlling the expression of a whole set of other genes.

In addition to this feedback system, scientists more recently stumbled upon another fascinating control mechanism, which relies on hundreds, if not thousands, of small regulatory RNAs. Some of these have the ability to silence the expression of target genes, while others control how other RNAs are spliced together. As a consequence, it has become evident that small RNAs can be involved in a number of diseases. Thus, in addition to the feedback loops from genes to proteins and back, and from genes to metabolites and back, small RNAs form yet another loop from genes to RNAs and back.

Reductionism and the Central Dogma used to make biology look deceivingly simple, at least conceptually. Alas, if we really want to understand biology, health, and disease, the time has come to accept

nature's complexity with all its beauty and challenges, while "making things as simple as possible, but not simpler," as Albert Einstein famously said. Life is complicated and it has become undisputedly evident that we need tools beyond our intuition to decipher its secrets. Such tools are at the heart of the emerging field of systems biology. Among them are experimental approaches that probe large and small biological systems and collect data and contextual information. Complementing these experimental approaches are mathematical and computational strategies. Some of these help us keep track of the many heterogeneous components of cells and organisms. Others have the goal of integrating biological information and of constructing models that permit exploration, explanation, and the formulation of novel hypotheses. Many of these models are initially conceptual, but as soon as they become more detailed and specific, they rely on the crisp language of mathematics, which alone is able to capture and evaluate complex numerical relationships among the components of biological systems. As we move into the future and learn more about living systems, the methods and approaches will certainly change, but the intricate features and characteristics of nature's complexity will always be with us.

2

I'd rather be fishin'

Biological research has had a long and esteemed history. So it is not surprising that its concepts, approaches, and methods have been subjected to dramatic changes time and again. Early trial and error in agriculture and animal domestication matured into simple plant manipulations and animal husbandry. Observations of birth and death, growth and decay, led to methods for preserving food for times of dearth. Exploratory dissections of corpses turned into primitive forms of surgery. The worldview of biology exploded with the invention of the microscope, which opened a window into an entirely new world of cells and microorganisms and pathogens. The exploration of medicinal herbs and poisons, as well as the procedures of alchemy and chemistry, motivated the invention of ever-more accurate methods and refined measurement tools.

The search for scientific truth reached a high point in the seventeenth century with the acceptance of the so-called scientific method, which is still considered fundamental today. According to this method, scientific inquiry advances through well-structured, iterative cycles of posing a hypothesis, testing it with experiments, analyzing results, making predictions, testing them, and formulating new hypotheses. In all fairness, one should mention that the roots of this structured type of scientific thinking and experimentation can actually be traced back two millennia to the third century BC Greek physician and anatomist Herophilus, who cofounded the most famous medical school of the time in the Egyptian city of Alexandria. Herophilus performed systematic dissections, which he documented in great detail, and maintained that trustworthy scientific

knowledge can only be found on an empirical basis. Nevertheless, the scientific method became the gold standard only in the seventeenth century.

Then the twentieth century rolled along and modern biomedical research exploded. Powerful experimental tools and custom-tailored machines rendered it possible to characterize biological phenomena with a resolution never seen before, down to the level of individual molecules. A prominent highlight was the identification of the structure of DNA, but many other classes of molecule were identified and characterized, and uncounted small and large discoveries occurred during the second half of the century. Most of these breakthroughs resulted directly from the application of the scientific method, which brought forth incredible amounts of precise data and unprecedented insights into the inner workings of life.

In the shadows of this hugely successful modus operandi, an alternative approach began to take hold around the turn of the millennium. This approach was driven by novel combinations of molecular biology and ingenious engineering advances in miniaturization and robotics, which suddenly permitted the execution of very many experiments at once. Whereas it once had taken an entire thesis project to determine the sequence of a short gene, sequencing became a quick routine task. Quasi overnight, almost every molecular biology lab became enabled to characterize the expression levels of thousands of genes simultaneously, with no need for specifically targeting a gene of interest. It became feasible to identify hundreds of proteins or metabolites with techniques of mass spectrometry. Biology witnessed the birth of a new era of large-scale, high-throughput data generation.

While very exciting to many researchers, this type of investigation was seen by many others as the antithesis of the scientific method, a despicable distraction from real research. The idea of "let's see what happens if we check all genes" was derided as a "fishing expedition." But alas for the critics, fishing has been successful and often enjoyable throughout human history, and it quickly started receiving appreciation and acceptance within the biological science community. Fast-forward, and fishing for molecular targets is now largely considered an equal partner to traditional, hypothesis-driven research.

Of course, the new methods needed their own names, but unfortunately for everyone who likes words or is a linguist in disguise, they became collectively known as terms with the suffixes –ome or –omics. Thus, high-throughput data generation on proteins was named