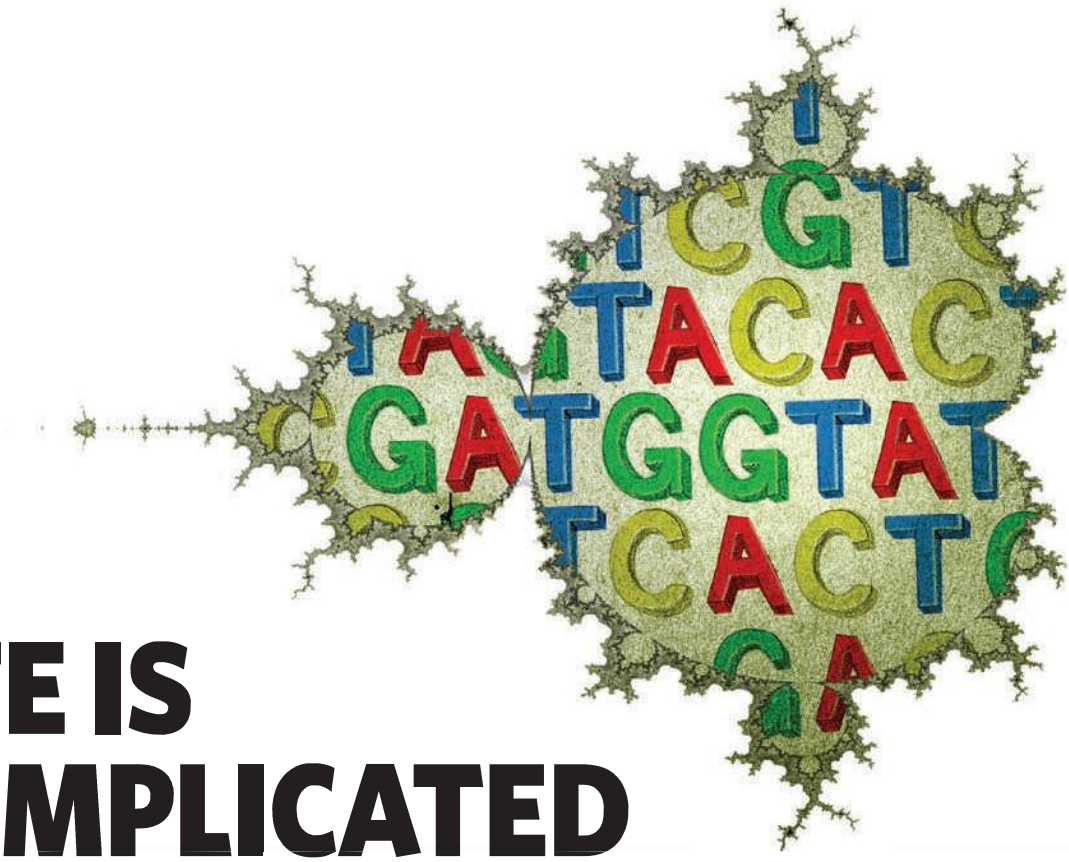


LIFE IS COMPLICATED

The more biologists look, the more complexity there seems to be. **Erika Check Hayden** asks if there's a way to make life simpler.

ILLUSTRATIONS BY JONATHAN BURTON

Not that long ago, biology was considered by many to be a simple science, a pursuit of expedition, observation and experimentation. At the dawn of the twentieth century, while Albert Einstein and Max Planck were writing mathematical equations that distilled the fundamental physics of the Universe, a biologist was winning the Nobel prize for describing how to make dogs drool on command.

The molecular revolution that dawned with the discovery of the structure of DNA in 1953 changed all that, making biology more quantitative and respectable, and promising to unravel the mysteries behind everything from evolution to disease origins. The human genome sequence, drafted ten years ago, promised to go even further, helping scientists trace ancestry, decipher the marks of evolution and find the molecular underpinnings of disease, guiding the way to more accurate diagnosis and targeted, personalized treatments. The genome promised to lay bare the blueprint of human biology.

That hasn't happened, of course, at least not yet. In some respects, sequencing has provided clarification. Before the Human Genome Project began, biologists guessed that the genome could contain as many as 100,000 genes that code for proteins. The true number, it turns out, is closer to 21,000, and biologists now know what many of those genes are. But at the same time, the genome sequence did what biological discoveries have done for decades. It opened the door to a vast labyrinth of new questions.

Few predicted, for example, that sequencing the genome would undermine the primacy of genes by unveiling whole new classes of elements — sequences that make RNA or have a regulatory role without coding for proteins. Non-coding DNA is crucial to biology, yet knowing that it is there hasn't made it any easier to understand what it does. "We fooled ourselves into thinking the genome was going to be a transparent blueprint, but it's not," says Mel Greaves, a cell biologist at the Institute of Cancer Research in Sutton, UK.

Instead, as sequencing and other new technologies spew forth data, the complexity of biology has seemed to grow by orders of magnitude. Delving into it has been like zooming into a Mandelbrot set — a space that is determined by a simple equation, but that reveals ever more intricate patterns as one peers closer at its boundary.

With the ability to access or assay almost any bit of information, biologists are now struggling with a very big question: can one ever truly know an organism — or even a cell, an organelle or a molecular pathway — down to the finest level of detail?

Imagine a perfect knowledge of inputs, outputs and the myriad interacting variables, enabling a predictive model. How tantalizing this notion is depends somewhat on the scientist; some say it is enough to understand the basic principles that govern life, whereas others are compelled to reach for an answer to the next question, unfazed by the ever increasing intricacies. "It seems like we're climbing a mountain that keeps getting higher and higher," says Jennifer Doudna, a biochemist at the University of California, Berkeley. "The more we know, the more we realize there is to know."

Web-like networks

Biologists have seen promises of simplicity before. The regulation of gene expression, for example, seemed more or less solved 50 years ago. In 1961, French biologists François Jacob and Jacques Monod proposed the idea that 'regulator' proteins bind to DNA to control the expression of genes. Five years later, American biochemist Walter Gilbert confirmed this model by discovering the lac repressor protein, which binds to DNA to control lactose metabolism in *Escherichia coli* bacteria[1]. For the rest of the twentieth century, scientists expanded on the details of the model, but they were confident that they understood the basics. "The crux of regulation," says the 1997 genetics textbook *Genes VI* (Oxford Univ. Press), "is that

a regulator gene codes for a regulator protein that controls transcription by binding to particular site(s) on DNA."

Just one decade of post-genome biology has exploded that view. Biology's new glimpse at a universe of non-coding DNA — what used to be called 'junk' DNA — has been fascinating and befuddling. Researchers from an international collaborative project called the Encyclopedia of DNA Elements (ENCODE) showed that in a selected portion of the genome containing just a few per cent of protein-coding sequence, between 74% and 93% of DNA was transcribed into RNA[2]. Much non-coding DNA has a regulatory role; small RNAs of different varieties seem to control gene expression at the level of both DNA and RNA transcripts in ways that are still only beginning to become clear. "Just the sheer existence of these exotic regulators suggests that our understanding about the most basic things — such as how a cell turns on and off — is incredibly naive," says Joshua Plotkin, a mathematical biologist at the University of Pennsylvania in Philadelphia.

Even for a single molecule, vast swathes of messy complexity arise. The protein p53, for example, was first discovered in 1979, and despite initially being misjudged as a cancer promoter, it soon gained notoriety as a tumour suppressor — a 'guardian of the genome' that stifles cancer growth by condemning genetically damaged cells to death. Few proteins have been studied more than p53, and it even commands its own meetings. Yet the p53 story has turned out to be immensely more complex than it seemed at first.

In 1990, several labs found that p53 binds directly to DNA to control transcription, supporting the traditional Jacob–Monod model of gene regulation. But as researchers broadened their understanding of gene regulation, they found more facets to p53. Just last year, Japanese researchers reported[3] that p53 helps to process several varieties of small RNA that keep cell growth in check, revealing a mechanism by which the protein exerts its tumour-suppressing power.

Even before that, it was clear that p53 sat at the centre of a dynamic network of protein, chemical and genetic interactions. Researchers now know that p53 binds to thousands of sites in DNA, and some of these sites are thousands of base pairs away from any genes. It influences cell growth, death and structure and DNA repair. It also binds to numerous other proteins, which can modify its activity, and these protein–protein interactions can be tuned by the addition of chemical modifiers, such as phosphates and methyl groups. Through a process known as alternative splicing, p53 can take nine different forms, each of which has its own activities and chemical modifiers. Biologists are now realizing that p53 is also involved in processes beyond cancer, such as fertility and very early embryonic development. In fact, it seems wilfully ignorant to try to understand p53 on its own. Instead, biologists have shifted to studying the p53 network, as depicted in cartoons containing boxes, circles and arrows meant to symbolize its maze of interactions.

"The more we know, the more we realize there is to know."

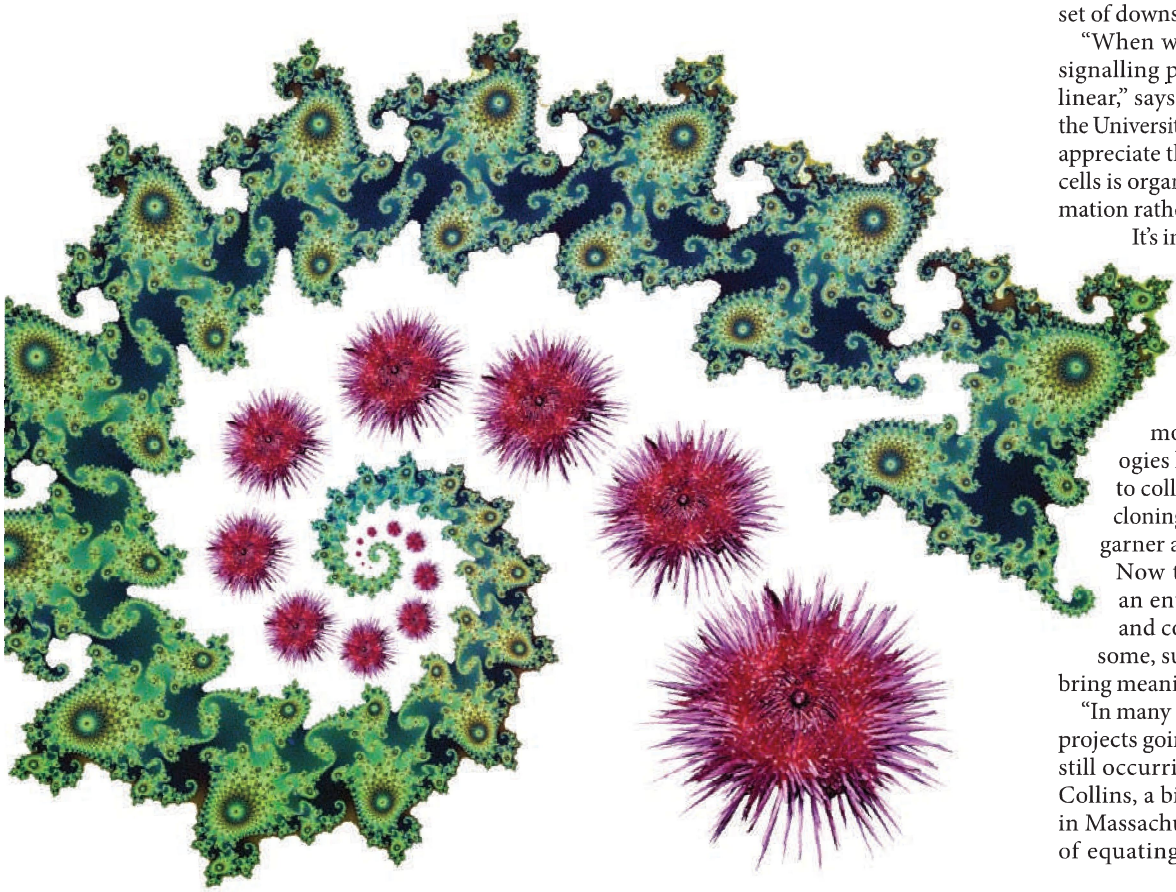
Data deluge

The p53 story is just one example of how biologists' understanding has been reshaped, thanks to genomic-era technologies. Knowing the sequence of p53 allows computational biologists to search the genome for sequences where the protein might bind, or to predict positions where other proteins or chemical modifications might attach to the protein. That has expanded the universe of known protein interactions — and has dismantled old ideas about signalling 'pathways', in which proteins such as p53 would trigger a defined set of downstream consequences.

"When we started out, the idea was that signalling pathways were fairly simple and linear," says Tony Pawson, a cell biologist at the University of Toronto in Ontario. "Now, we appreciate that the signalling information in cells is organized through networks of information rather than simple discrete pathways. It's infinitely more complex."

The data deluge following the Human Genome Project is undoubtedly part of the problem. Knowing what any biological part is doing has become much more difficult, because modern, high-throughput technologies have granted tremendous power to collect data. Gone are the days when cloning and characterizing a gene would garner a paper in a high-impact journal. Now teams would have to sequence an entire human genome, or several, and compare them. Unfortunately, say some, such impressive feats don't always bring meaningful biological insights.

"In many cases you've got high-throughput projects going on, but much of the biology is still occurring on a small scale," says James Collins, a bioengineer at Boston University in Massachusetts. "We've made the mistake of equating the gathering of information



SEA URCHINS: J. ROTMAN/NATUREPL.COM

with a corresponding increase in insight and understanding."

A new discipline — systems biology — was supposed to help scientists make sense of the complexity. The hope was that by cataloguing all the interactions in the p53 network, or in a cell, or between a group of cells, then plugging them into a computational model, biologists would glean insights about how biological systems behaved.

In the heady post-genome years, systems biologists started a long list of projects built on this strategy, attempting to model pieces of biology such as the yeast cell, *E. coli*, the liver and even the 'virtual human'. So far, all these attempts have run up against the same roadblock: there is no way to gather all the relevant data about each interaction included in the model.

A bug in the system

In many cases, the models themselves quickly become so complex that they are unlikely to reveal insights about the system, degenerating instead into mazes of interactions that are simply exercises in cataloguing.

In retrospect, it was probably unrealistic to expect that charting out the biological interactions at a systems level would reveal systems-level properties, when many of the mechanisms and principles governing inter- and intracellular behaviour are still a mystery, says Leonid Kruglyak, a geneticist at Princeton University in New Jersey. He draws a comparison to physics: imagine building a particle accelerator such as the Large Hadron Collider without knowing anything about the underlying theories of quantum mechanics, quantum chromodynamics or relativity. "You would have all this stuff in your detector, and you would have no idea how to think about it, because it would involve processes that you didn't understand at all," says Kruglyak. "There is a certain amount of naivety to the idea that for any process — be it biology or weather prediction or anything else — you can simply take very large amounts of data and run a data-mining program and understand what is going on in a generic way."

"It's people who complicate things. Some people are simplifiers and others are dividers."

This doesn't mean that biologists are stuck peering ever deeper into a Mandelbrot set without any way of making sense of it. Some biologists say that taking smarter systems approaches has empowered their fields, revealing overarching biological rules. "Biology is entering a period where the science can be underlaid by explanatory and predictive principles, rather than little bits of causality swimming in a sea of phenomenology," says Eric Davidson, a developmental biologist at the California Institute of Technology in Pasadena.

Such progress has not come from top–down analyses — the sort that try to arrive at insights by dumping a list of parts into a model and hoping that clarity will emerge from chaos. Rather, insights have come when scientists systematically analyse the components of processes that are easily manipulated in the laboratory — largely in model organisms. They're still using a systems approach, but focusing it through a more traditional, bottom–up lens.

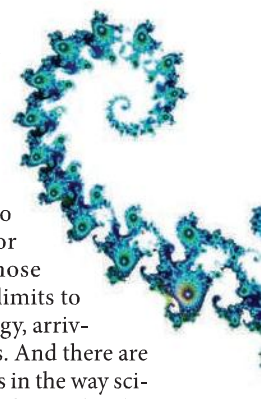
Davidson points to the example of how gene regulation works during development to specify the construction of the body. His group has spent almost a decade dissecting sea-urchin development by systematically knocking out the expression of each of the transcription factors — regulatory proteins that control the expression of genes — in the cells that develop into skeleton. By observing how the loss of each gene affects development, and measuring how each 'knockout' affects the expression of every other transcription factor, Davidson's group has constructed a map of how these transcription factors work together to build the animal's skeleton[4]. The map builds on the Jacob–Monod principle that regulation depends on interactions between regulatory proteins and DNA. Yet it includes all of these regulatory interactions and then attempts to draw from them common guiding principles that can be applied to other developing organisms.

For example, transcription factors encoded in the urchin embryo's genome are first activated by maternal proteins. These embryonic factors, which are active for only a short time, trigger downstream transcription factors that interact in a positive feedback circuit to switch each other on permanently. Like the sea urchin, other organisms from fruitflies to humans organize development into 'modules' of genes, the interactions of which are largely isolated from one another, allowing evolution to tweak each module without compromising the integrity of the whole process. Development, in other words, follows similar rules in different species.

"The fundamental idea that the genomic regulatory system underlies all the events of development of the body plan, and that changes in it probably underlie the evolution of body plans, is a basic principle of biology that we didn't have before," says Davidson. That's a big step forwards from 1963, when Davidson started his first lab. Back then, he says, most theories of development were "manifestly useless".

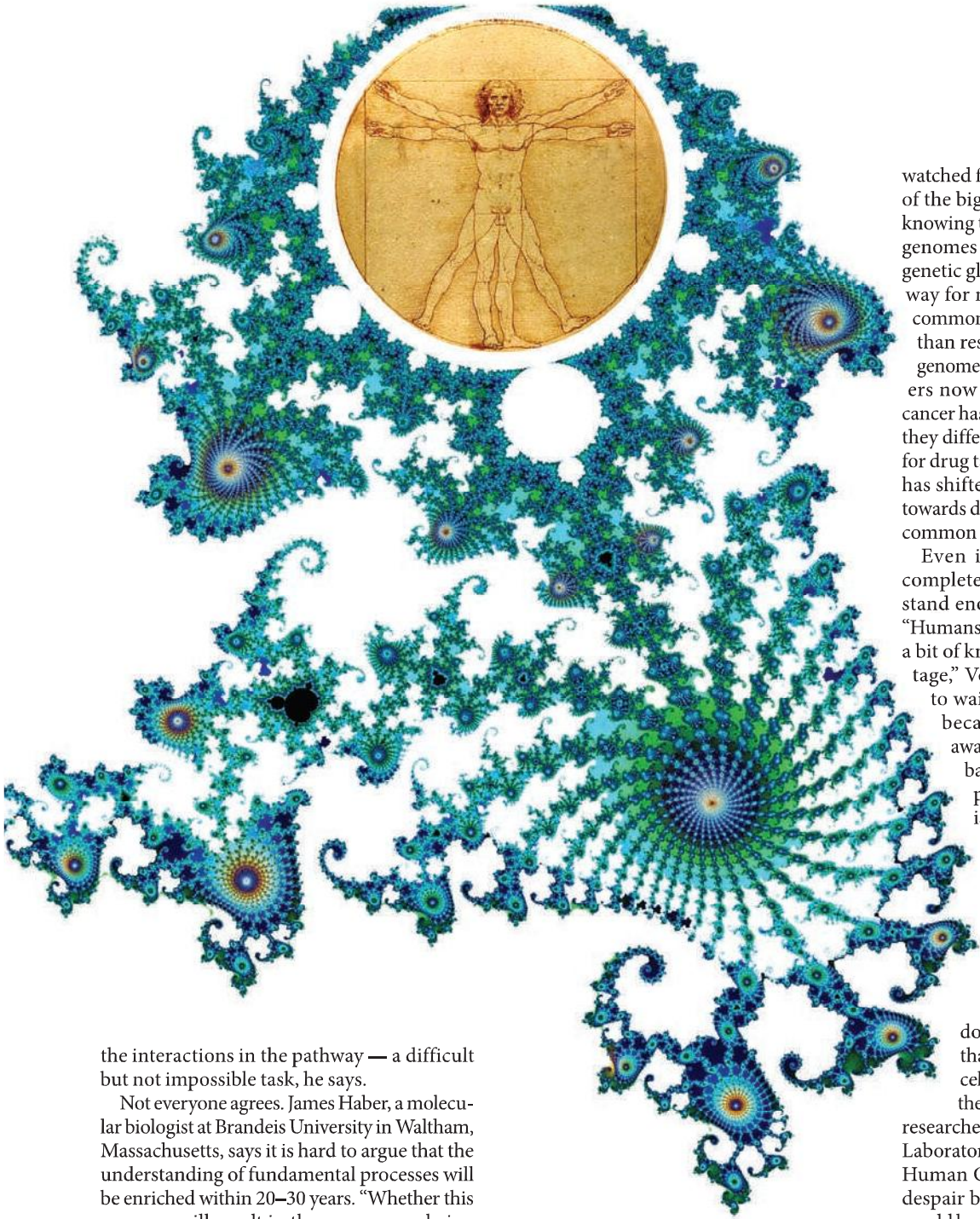
Davidson calls his work "a proof of principle that you can understand everything about the system that you want to understand if you get hold of its moving parts". He credits the Human Genome Project with pushing individual biologists more in the direction of understanding systems, rather than staying stuck in the details, focused on a single gene, protein or other player in those systems. First, it enabled the sequencing of model-organism genomes, such as that of the sea urchin, and the identification of all the transcription factors active in development. And second, it brought new types of biologists, such as computational biologists, into science, he says.

The eye of the beholder



So how is it that Davidson sees simplicity and order emerging where many other biologists see increasing disarray? Often, complexity seems to lie in the eye of the beholder. Researchers who work on model systems, for instance, can manipulate those systems in ways that are off-limits to those who study human biology, arriving at more definitive answers. And there are basic philosophical differences in the way scientists think about biology. "It's people who complicate things," says Randy Schekman, a cell and molecular biologist at the University of California, Berkeley. "I've seen enough scientists to know that some people are simplifiers and others are dividers." Although the former will glean big-picture principles from select examples, the latter will invariably get bogged down in the details of the examples themselves.

Mark Johnston, a yeast geneticist at the University of Colorado School of Medicine in Denver, admits to being a generalizer. He used to make the tongue-in-cheek prediction that the budding yeast *Saccharomyces cerevisiae* would be "solved" by 2007 when every gene and every interaction has been characterized. He has since written more seriously that this feat will be accomplished within the next few decades[5]. Like Davidson, he points out that the many aspects of yeast life, such as the basics of DNA synthesis and repair, are essentially understood. Scientists already know what about two-thirds of the organism's 5,800 genes do, and the remaining genes will be characterized soon enough, Johnston says. He works on the glucose-sensing pathway, and says he will be satisfied that he understands it when he can quantitatively describe



the interactions in the pathway — a difficult but not impossible task, he says.

Not everyone agrees. James Haber, a molecular biologist at Brandeis University in Waltham, Massachusetts, says it is hard to argue that the understanding of fundamental processes will be enriched within 20–30 years. "Whether this progress will result in these processes being 'solved' may be a matter of semantics," he says, "but some questions — such as how chromosomes are arranged in the nucleus — are just beginning to be explored." Johnston argues that it is neither possible not necessary to arrive at the quantitative understanding that he hopes to achieve for the glucose-sensing pathway for every other system in yeast. "You have to decide what level of understanding you're satisfied with, and some people respond that they're not satisfied at any level — that we have to keep going," he says. This gulf between simplifiers and dividers isn't just a matter of curiosity for armchair philosophers. It plays out every day as study sections and peer reviewers decide which approach to science is worth funding and publishing. And it bears on the ultimate question in biology: will we ever understand it all?

The edge of the universe

Some, such as Hiroaki Kitano, a systems biologist at the Systems Biology Institute in Tokyo, point out that systems seem to grow more complex only because we continue to learn about them. "Biology is a defined system," he says, "and in time, we will have a fairly good understanding of what the system is about."

Others demur, arguing that biologists will never know everything. And it may not matter terribly that they don't. Bert Vogelstein, a cancer-genomics researcher at Johns Hopkins University in Baltimore, Maryland, has watched first-hand as complexity dashed one of the biggest hopes of the genome era: that knowing the sequence of healthy and diseased genomes would allow researchers to find the genetic glitches that cause disease, paving the way for new treatments. Cancer, like other common diseases, is much more complicated than researchers hoped. By sequencing the genomes of cancer cells, for example, researchers now know that an individual patient's cancer has about 50 genetic mutations, but that they differ between individuals. So the search for drug targets that might help many patients has shifted away from individual genes and towards drugs that might interfere in networks common to many cancers.

Even if we never understand biology completely, Vogelstein says, we can understand enough to interfere with the disease. "Humans are really good at being able to take a bit of knowledge and use it to great advantage," Vogelstein adds. "It's important not to wait until we understand everything, because that's going to be a long time away." Indeed, drugs that influence those bafflingly complex signal-transduction pathways are among the most promising classes of new medicines being used to treat cancer. And medicines targeting the still-mysterious small RNAs are already in clinical trials to treat viral infections, cancer and macular degeneration, the leading cause of untreatable blindness in wealthy nations.

The complexity explosion, therefore, does not spell an end to progress. And that is a relief to many researchers who celebrate complexity rather than wring their hands over it. Mina Bissell, a cancer researcher at the Lawrence Berkeley National Laboratory in California, says that during the Human Genome Project, she was driven to despair by predictions that all the mysteries would be solved. "Famous people would get up and say, 'We will understand everything after this,'" she says. "Biology is complex, and that is part of its beauty." She need not worry, however; the beautiful patterns of biology's Mandelbrot-like intricacy show few signs of resolving. ■

Erika Check Hayden is a senior reporter for *Nature* based in San Francisco.

1. Gilbert, W. & Muller-Hill, B. *Proc. Natl Acad. Sci. USA* **56,** 1891–1898 (1966).
2. The ENCODE Project Consortium *Nature* **447,** 799–816 (2007).
3. Suzuki, H. I. *et al. Nature* **460,** 529–533 (2009).
4. Oliveri, P., Tu, Q. & Davidson, E. H. *Proc. Natl Acad. Sci. USA* **105,** 5955–5962 (2008).
5. Fields, S. & Johnston, M. *Science* **307,** 5717 (2005).

See Editorial, page 649, and human genome special at www.nature.com/humangenome.