CHAPTER 1

The Code of Life

The digital era of life commenced with the most famous understatement in the history of science:

> *We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.*

Thus began a paper that appeared in the journal *Nature* on April 25, 1953, in which its authors, James Watson and Francis Crick, suggested the now-famous double helix form of DNA. The paper was extraordinary in several ways: first, because Watson and Crick, both relatively young and unknown researchers, had succeeded in beating many more famous rivals in the race to explain the structure of DNA. Second, their proposal managed to meld supreme elegance with great explanatory power—a combination that scientists prize highly. Most of all, the paper was remarkable because it ended once and for all decades of debate and uncertainty about the mechanism of inheritance. In doing so, it marked the starting point for a new era in genetics, biology, and medicine—an era whose first phase would close exactly 50 years after Watson and Crick's paper with the announcement of the complete elucidation of human DNA. The contrast of that half-century's dizzying rate of progress with the preceding centuries' slow groping towards an understanding of inheritance could hardly be greater.

One hundred and fifty years ago, Gregor Mendel, an Augustinian monk working in what is now the city of Brno in Moravia, carried out the first scientific investigations of heredity. Prior to his meticulous work on crossbreeding sweet peas, knowledge about heredity had existed only as a kind of folk wisdom among those rearing animals or propagating plants.

Mendel crossed sweet peas with pairs of traits—two different seed shapes or flower colors—in an attempt to find laws that governed the inheritance of these characteristics in subsequent generations. After thousands of such experiments, painstakingly recorded and compared, he deduced that these traits were passed from parent to offspring in what he called *factors*. Mendel realized that these factors came in pairs, one from each parent, and that when the two factors clashed, they did not mix to produce an intermediate result. Rather, one factor would dominate the other in the offspring. The subjugated factor would still persist in a latent form, however, and might reappear in subsequent generations in a remarkably predictable way.

Although it offered key insights into the mechanism of inheritance, Mendel's work was ignored for nearly half a century. This may have been partly due to the fact that his work was not widely read. But even if it had been, his factors may have been too abstract to excite much attention, even though they turned out to be completely correct when recast as the modern idea of genes, the basic units of heredity. In any case, work on heredity shifted to an alternative approach, one based on studying something much more tangible: cells, the basic units of life.

Hermann Muller used just such an approach in 1927 when he showed that bombarding the fruit fly with X-rays could produce *mutations*—variant forms of the organism. This was important because it indicated that genes were something physical that could be damaged like any other molecule. A chance discovery by Fred Griffith in 1928 that an extract from disease-causing bacteria could pass on virulence to a strain that was normally harmless finally gave researchers the first opportunity to seek out something chemical: the molecule responsible for transmitting the virulence. It was not until 1944, however, that Oswald Avery and his coworkers demonstrated that this substance was deoxyribonucleic acid—DNA.

In many ways, this contrasted sharply with the accepted views on the biochemical basis for heredity. Although DNA had been known for three quarters of a century—Johann Friedrich Miescher discovered it in pus-filled bandages discarded by a hospital—it was regarded as a rather dull chemical consisting of a long, repetitive chain made up of four ingredients called nucleotides. These nucleotides consist of a base—adenine, cytosine, guanine or thymine—each linked to the sugar deoxyribose at one end and a phosphate

group at the other. Chemical bonds between the sugar and phosphate group allow very long strings of nucleotides to be built up.

The conventional wisdom of the time was that genetics needed a suitably complex molecule to hold the amazing richness of heredity. The most complex molecules then known were proteins. They not only form the basic building blocks of all cells, but also take on all the other key roles there such as chemical signaling or the breakdown of food. It was this supposition about protein as the chosen carrier for heredity that made Watson and Crick's alternative proposal so daring. They not only provided a structure for DNA, they offered a framework for how "boring" DNA could store inherited traits.

This framework could not have been more different from the kind most researchers were using at the time. The key properties of a protein are its physical and chemical properties; to use a modern concept, its essence is analogue. Watson and Crick's proposal was that DNA stored heredity not physically (through its shape or chemical properties), but through the information encoded by the sequence of four nucleotides. In other words, the secret of DNA—and of life itself—was digital.

Because it is the information they represent rather than the chemical or physical properties they possess that matters, the four nucleotides can, for the purposes of inheritance and genetics, be collapsed from the four bases (adenine, cytosine, guanine, and thymine) to four letters. The bases are traditionally represented as A, C, G, and T. This makes explicit the fact that the digital code employed by Nature is not binary—0 and 1—as in today's computers, but quaternary, with four symbols. But the two codes are completely equivalent. To see this, simply replace the quaternary digit A with the binary digits 00, C with 01, G with 10 and T with 11. Then any DNA sequence—for example AGGTCTGAT—can be converted into an equivalent binary sequence—in this case, 00 10 10 11 01 11 10 00 11. Even though the representation is different, the information content is identical.

With the benefit of hindsight, it is easy to see why a digital mechanism for heredity was not just possible but almost necessary. As anyone knows who has made an analogue copy of an audio or video cassette from another copy, the quality of the signal degrades each time. By contrast, a digital copy of a digital music file is always perfect, which is why the music and film industries have switched from a semi-official tolerance of analogue copying to a rabid

hatred of the digital kind. Had Nature adopted an analogue storage method for inheritance, it would have been impossible to make the huge number of copies required for the construction of a typical organism. For example, from the fertilized human egg roughly a hundred thousand billion cells are created, each one of which contains a copy of the original DNA. Digital copying ensures that errors are few and can be corrected; analogue copying, however, would have led to a kind of genetic "fuzziness" that would have ruled out all but the simplest organisms.

In 1953, computers were so new that the idea of DNA as not just a huge digital store but a fully-fledged digital program of instructions was not immediately obvious. But this was one of the many profound implications of Watson and Crick's work. For if DNA was a digital store of genetic information that guided the construction of an entire organism from the fertilized egg, then it followed that it did indeed contain a preprogrammed sequence of events that created that organism—a program that ran in the fertilized cell, albeit one that might be affected by external signals. Moreover, since a copy of DNA existed within practically every cell in the body, this meant that the program was not only running in the original cell but in all cells, determining their unique characteristics.

Watson and Crick's paper had identified DNA as the digital code at the heart of life, but there remained the problem of how this was converted into the analogue stuff of organisms. In fact, the problem was more specific: because the analogue aspect of life was manifest in the proteins, what was needed was a way of translating the digital DNA code into analogue protein code. This endeavor came to be known as "cracking the DNA code." The metaphor was wrong, though—perhaps it was a side effect of the Cold War mentality that prevailed at that time. DNA is not a cryptic code that needs to be broken, because this implies that it has an underlying message that is revealed once its code is "cracked." There is no secret message, however.

DNA is another type of code—computer code. DNA is the message itself—the lines of programming that need to be run for the operations they encode to be carried out. What was conventionally viewed as cracking the code of life was in fact a matter of understanding *how* the cell ran the DNA digital code.

One step along the way to this understanding came with the idea of *messenger RNA* (mRNA). As its name suggests, ribonucleic acid (RNA) is closely related to DNA, but comes as a single strand rather than the double helix. It, too, employs a digital code, with four nucleotides. Thymine is replaced by uracil and the deoxyribose sugar by ribose, but for information purposes, they are the same.

It was discovered that mRNA is transcribed (copied) from sections of the DNA sequence. In fact, it is copied from sections that correspond to Mendel's classical *factors*—the genes. Surrounding these genes are sections of DNA text that are not transcribed, just as a computer program may contain comments that are ignored when the program is run. And just as a computer copies parts of a program held on a disc and sends them down wires to other components of the system, so the cell, it seemed, could copy selected portions of DNA and send them down virtual wires as mRNA.

These virtual wires end up at special parts of the cell known as *ribosomes*. Here the mRNA is used to direct the synthesis of proteins by joining together chemical units called amino acids into chains, which are often of great length. There are twenty of these amino acids, and the particular sequence in the chain determines a protein's specific properties, notably its shape. The complicated ensemble of attractions and repulsions among the constituent atoms of the amino acids causes the chain of them to fold up in a unique form that gives the protein its properties. The exact details of this protein are determined by the sequence of amino acids, which are in turn specified by the mRNA, transcribed from the DNA. Here, then, was the device for converting the digital data into an analogue output. But this still left the question of how different mRNA messages were converted to varying amino acids.

A clever series of experiments by Marshall Nirenberg in the early 1960s answered this question. He employed a technique still used to this day by computer hackers (where *hacker* means someone who is interested in understanding computers and their software, as opposed to malevolent *crackers*, who try to break into computer systems). In order to learn more about how an unknown computer system or program is working, it is often helpful not only to measure the signals passing through the circuits naturally, but also to send carefully crafted signals and observe the response.

This is precisely what Nirenberg did with the cell. By constructing artificial mRNA he was able to observe which amino acids were output by the cell's machinery for a given input. In this way he discovered, for example, that the three DNA letters AAA, when passed to a ribosome by the mRNA, always resulted in the synthesis of the amino acid lysine, while CAG led to the production of glutamine. By working through all the three-letter combinations, he established a table of correspondences between three-letter sequences—known as codons—and amino acids.

This whole process of converting one kind of code into another is very similar to the process of running a computer program: the program lines are sent to the central processing unit (CPU) where each group of symbols causes certain actions that result in a particular output. For example, this might be

a representation on a monitor. In the same way, the ribosome acts as a kind of processing unit, with the important difference being that its output consists of proteins, which are "displayed" not on a screen but in real, three-dimensional space within the cell.

Viewed in this way, it is easy to understand how practically every cell in the body can contain the same DNA code and yet be radically different in its form and properties—brain, liver, or muscle cells, for example. The DNA can be thought of as a kind of software suite containing the code for every kind of program that the body will ever need. Among this is operating system software, basic housekeeping routines which keep cells ticking over by providing energy or repairing damaged tissue. There are also more specialized programs that are only run in a particular tissue—brain code in brain cells or liver code in liver cells, for example. These correspond to more specialized kinds of programs like word processors or spreadsheets: very often they are present on a computer system, but they are only used for particular applications. The operating system, however, is running constantly, ensuring that input is received from the keyboard and output is displayed on the screen. The details of the analogy are not important; what is crucial is that DNA's information is digital. From this has flowed a series of dramatic developments that are revolutionizing not just biology but medicine, too. All of these developments have come about from using powerful computers to search the digital code of life for the structures hidden within.

It may not be immediately apparent why computing power is important or even necessary. After all, on one level, the totality of information contained within an organism's DNA—termed its genome—is not complex. It can be represented as a series of letters, turning chemicals into text. As such, it can be read directly. This is true, but even leaving aside the problem of interpretation (what these letters in a particular order mean), there is another fundamental issue that genome researchers must address first: the sheer quantity of the data they are dealing with.

So far, the digital content of the genome has been discussed in the abstract. To understand why computers are indispensable, though, it is helpful to consider some specific facts. For example, the DNA within a typical human cell is twisted into a double helix; this helix is wound up again into an even more convoluted structure called a chromosome. Chromosomes were first noted within the nucleus of certain cells over one hundred years ago, but decades were to pass before it was shown that they contained DNA. Normal human cells have 46 chromosomes—22 similar pairs, called autosomes, and the two sex chromosomes. Women have two X chromosomes, while men possess one

X chromosome and one Y chromosome. The number is not significant; chromosomes are simply a form of packaging, the biological equivalent of CD-ROMs.

Even though these 46 chromosomes (23 from each parent) fit within the nucleus, which itself is only a small fraction of the microscopic cell's total volume, the amount of DNA they contain collectively is astonishing. If the DNA content of the 23 chromosomes from just one cell were unwound, it would measure around 1 meter in length, or 2 meters for all 46 chromosomes. Since there are approximately one hundred thousand billion cells in the human body, this means that laid end-to-end, all the DNA in a single person would stretch from the earth to the sun 1,200 times.

Things are just as dramatic when viewed from an informational rather than physical point of view. Each of the two sets of 23 chromosomes—found in practically every human cell—makes up a genome that contains some 3 billion chemical digits (the A*s*, C*s*, G*s* and T*s*). Printed as ordinary letters in an average-sized typeface, a bare listing representing these letters would require roughly 3,000 books each of 330 pages—a pile about 60 meters high. And for any pair of human beings (except twins deriving from the same fertilized egg), every one of the million pages in these books would have several letters that are different, which is why some people have brown eyes and others blue.

Now imagine trying to find among these 3,000 volumes the subprograms (the genes) that create the particular proteins which determine the color of the iris, say, and the letter changes in them that lead to brown rather than blue eyes. Because genes have about 12,000 chemical letters on average—ranging from a few hundred to a couple of million—they spread over several pages, and thus might seem easy enough to spot. But the task of locating these pages is made more difficult by the fact that protein-producing code represents only a few percent of the human genome. Between the genes—and inside them, too, shattering them into many smaller fragments—are stretches of what has been traditionally and rather dismissively termed "junk DNA." It is now clear, however, that there are many other important structures there (control sequences, for example, that regulate when and how proteins are produced). Unfortunately, when looking at DNA letters, no simple set of rules can be applied for distinguishing between pages that code for proteins and those that represent the so-called junk. In any case, even speed-reading through the pile of books at one page a second would require around 300 hours, or nearly two days, of nonstop page flicking. There would be little time left for noting any subtle signs that might be present.

The statistics may be simplistic, but they indicate why computers have become the single most important tool in *genomics*, a word coined only in 1986 to describe the study of genomes. Even though the data are simple almost to the point of triviality—just four letters—the incomprehensible scale makes manipulating these data beyond the reach of humans. Only com-

puters (and fast ones at that) are able to perform the conceptually straightforward but genuinely challenging operations of searching and comparing that lie at the heart of genomics.

The results of marrying computers with molecular biology have been stunning. Just fifty years after Watson and Crick's general idea for DNA's structure, we now have a complete listing of the human genome's digital code—all 3 billion chemical letters of it. Contained within them are the programs for constructing every protein in our bodies. There are instructions that tell the fertilized egg how to grow; there are specialized programs that create muscles, skin, and bone. As we begin to understand how this happens, we can also appreciate how things go wrong. Like all software, the DNA code has bugs, or errors, in it. Most of these are of no consequence, occurring in noncritical places of the program. They are the equivalent of misspelled words in the comments section of programming code. However, some errors can be devasting. Consider the following two listings:

```
AGTAATTTCTCACTTCTTGGTACTCCTGTCCTGAAAGATAT
TAATTTCAAGATAGAAAGAGGACAGTTGTTGGCGGTTGCTG
GATCCACTGGAGCAGGCAAGACTTCACTTCTAATGATGATTA
TGGGAGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGT
GGAAGAATTTCATTCTGTTCTCAGTTTTCCTGGATTATGC
CTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTA
TGATGAATATAGATACAGAAGCGTCATCAAAGCATGCCAA
```

```
AGTAATTTCTCACTTCTTGGTACTCCTGTCCTGAAAGATAT
TAATTTCAAGATAGAAAGAGGACAGTTGTTGGCGGTTGCTG
GATCCACTGGAGCAGGCAAGACTTCACTTCTAATGATGAT
TATGGGAGAACTGGAGCCTTCAGAGGGTAAAATTAAG
CACAGTGGAAGAATTTCATTCTGTTCTCAGTTTTCCTGGAT
TATGCCTGGCACCATTAAAGAAAATATCATTGGTGTTTCCTA
TGATGAATATAGATACAGAAGCGTCATCAAAGCATGCCAA
```

The two listings show only a tiny fraction of the 250,000 DNA letters that code for an important human protein. The difference between the two portions of code is just three chemical letters—CTT is missing in the second listing. The absence of these three letters, however, is enough to result in cystic fibrosis for many people who have this apparently trivial software glitch. Similarly, just one wrong letter in another region can lead to sickle cell anemia, while the addition of a few extra letters in the wrong place elsewhere

causes Huntington's disease. Even more serious errors can mean embryos fail to develop at all—a fatal flaw in the operating system that causes the human system to crash as it boots up.

With the cell's digital code in hand, scientists can begin to understand these problems and even treat them. Often a DNA software bug causes the wrong protein to be produced by the ribosomes. Drugs may be able to block its production or operation in some way. Similarly, knowledge about the genomes of viruses and bacteria can aid pharmaceutical companies in their search for effective drugs and vaccines to combat them.

Driving these developments is bioinformatics: the use of computers to store, search through, and analyze billions of DNA letters. It was bioinformatics that turned the dream of sequencing the human genome into reality. It is bioinformatics that will allow humanity to decode its deepest secrets and to reveal the extraordinary scientific riches contained in the digital core of life.

NOTES

1. p. 7 *it would measure around 1 meter in length* 20 facts about the human genome. Online at *http://www.sanger.ac.uk/HGP/draft2000/facts.shtml.*
2. p. 7 *one hundred thousand billion cells in the human body* 20 facts about the human genome. Online at *http://www.sanger.ac.uk/HGP/draft2000/facts.shtml.*
3. p. 7 *genes have about 12,000 chemical letters* Tom Strachan and Andrew P. Read, *Human Molecular Genetics* 2 (1999): 150.
4. p. 7 *a word coined only in 1986* P. Hieter and M. Boguskis, "Functional genomics: it's all how you read it," *Science* 278 (1997): 601–602.