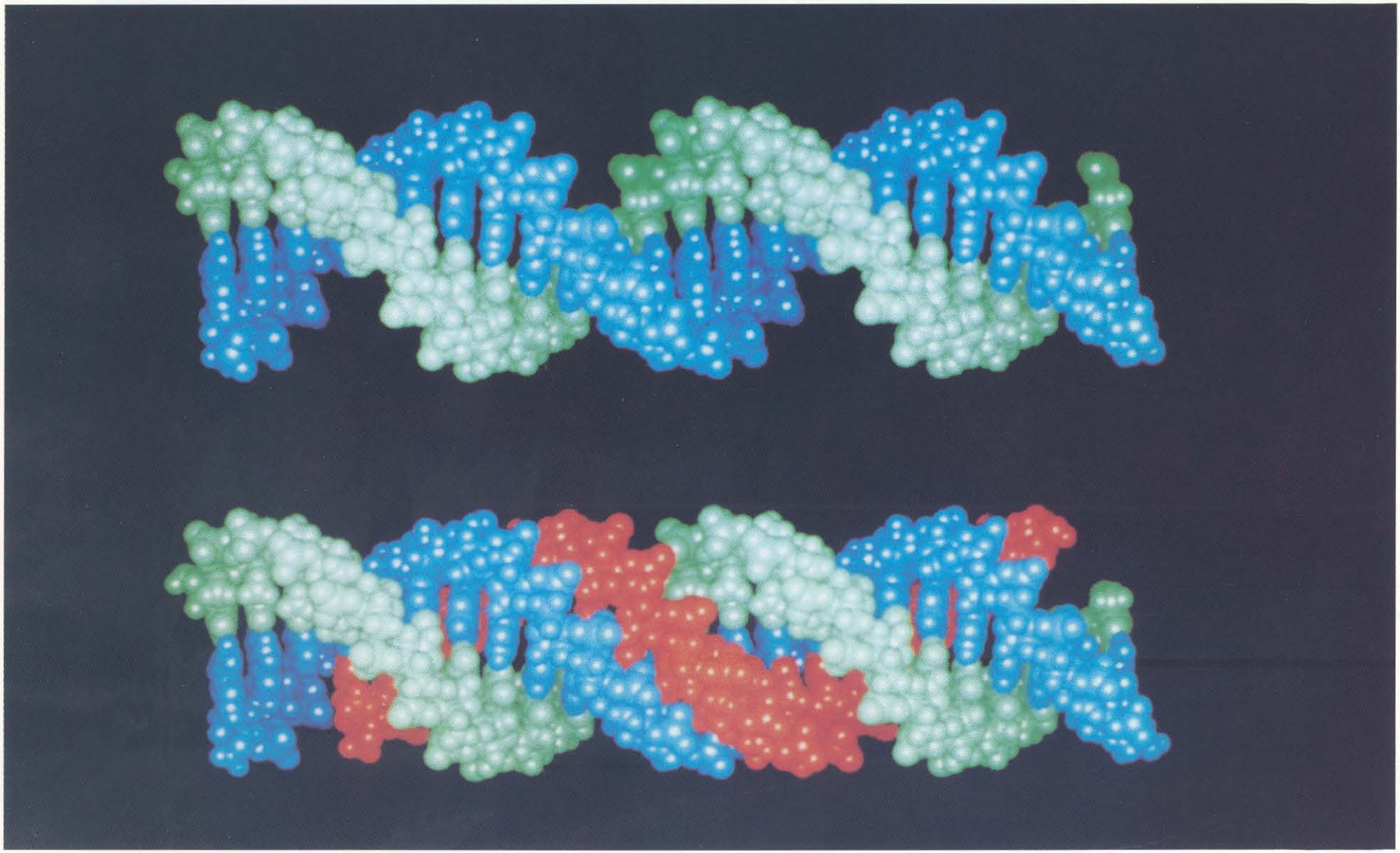


Designing Molecular Machines to Read the Genetic Blueprint

What is the genome project, and what does chemistry at Caltech have to do with it?

by Peter B. Dervan

Opposite, left: Two strands of DNA twist around each other into Watson and Crick's famous double helix. In this computer-generated image, one strand is colored blue, the other green. Each horizontal link between the strands is a letter in the genetic code—there are 24 letters in this image, and about 3 billion in a human cell. Right: A third strand of DNA, colored Caltech orange, can bind to the Watson and Crick strands without disrupting them. This chemical approach may be a general method for locating single sites in the human genome. This Caltech strand is 18 letters long.

The human genome project is an ambitious effort to map all of the 100,000 or so genes that make up the blueprint of man. I'm not going to talk about how much money we should spend on this, or how fast we should do it. Suffice it to say that it will happen sooner or later, and that it will affect everybody's life when it does. But what is the genome project, and what does chemistry at Caltech have to do with it?

Physicians have been mapping the human body for hundreds of years—charting where the bones are, and the muscles, and the blood vessels, and so on. Mapping the genome means finding the genes that make us what we are—the coded instructions that govern how we develop and grow, and determine what makes one person different from another—and pinpointing their specific locations in the genetic material. So in fact, this is the highest-resolution map of man.

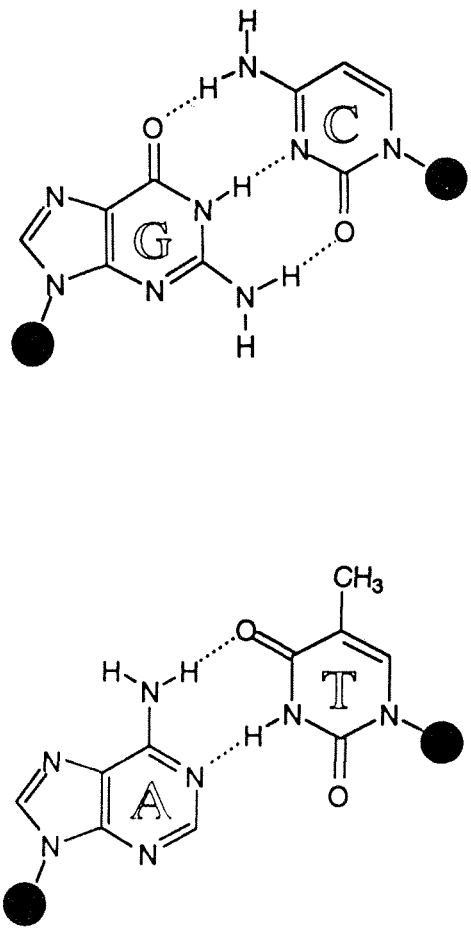
You can think of this genetic blueprint as an encyclopedia containing 2,000 volumes, each having 500 pages, and with 3,000 letters on each page. Say you want to know what makes your eyes blue, or predisposes you to cardiovascular disease. You need to be able to find out that the pertinent information is in Volume III, say, on page 357, and then you can turn to that page and look at that gene, or set of genes. So by mapping the genome, we are really writing the encyclopedia's index.

Our cells actually store this information in coded form in a molecule called deoxyribonucleic acid (DNA). The code is written in an alphabet much simpler than that of English, having only four letters instead of 26. The letters are chemical entities that we designate A (adenine), C (cytosine), G (guanine), and T (thymine). These letters are strung together in long sequences, like beads on a string, to make DNA. The DNA is such a valuable set of reference books that the library—a cell's nucleus—keeps it on reserve. When the cell needs to use the information, it doesn't let the DNA circulate out into the cell, but copies the information onto another molecule called RNA (ribonucleic acid), which is chemically very similar to DNA but not quite as stable. The RNA carries the blueprint's instructions to the cell's manufacturing centers, which make all the protein machines that give us hair, or make our muscles work, or digest our food. And when the cell has finished making the protein, it breaks down and recycles the RNA.

DNA is pretty sturdy stuff. It will last for millions of years in water at room temperature. So the chemical bonds—called covalent bonds—that hold the letters together in their correct sequence are very strong. This makes good sense—after all, if you are a cell, you don't want your master blueprint to fall apart on you. A human analogy to these strong bonds would be the bond between my elbow and my wrist. Chemists know a lot about these strong bonds—we synthesize small bits of genes in the laboratory routinely, on a machine. This machine is basically a lot of fancy plumbing and computer-controlled valves that mix the chemical ingredients in the right order.

But there's another set of weaker bonds that are very important to this story, and our understanding of these bonds is quite poor. This is the

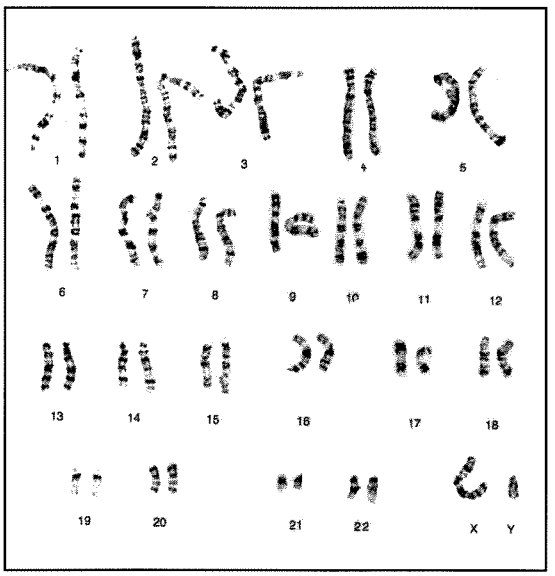
My research group is trying to build a molecular machine that can scan this whole meter of DNA and find one single location on it, reading its bumps and edges, its nooks and crannies, like Braille.



The four letters—C, G, A, and T—of the genetic code, and how they recognize each other. The hydrogen bonds that, taken together, constitute each "secret handshake" are shown as dotted lines. These bonds result from the attraction between a hydrogen atom (H) in one code letter and an atom of oxygen (O) or nitrogen (N) in the other code letter. The solid black circles represent the DNA molecule's backbone.

set of bonds that allows the stored information to be communicated so that the RNA copies can be made. You see, an A only talks to a T, and a G only talks to a C. Each pair of letters interacts with each other in a very specific way—a secret handshake, if you will, that allows each letter to recognize its partner. To carry the anatomical analogy further, this handshake is a very specific interaction—we don't shake shoulders—and it's strong enough that, if I have you by the hand, I could pull you from a river and save your life. But the interaction is weak enough that we can break it in an instant at a very specific place. If I shake your hand and then we turn and walk away from each other, you wouldn't tear my hand off and take it with you. You could also think of these weak bonds as being made of Velcro. It's these weak bonds—some of them are called hydrogen bonds—that give proteins and other biopolymers their specific three-dimensional shape, and it's a molecule's shape that allows it to perform its function. Chemists today are struggling to understand these weak bonds to the point where we can predict their behavior, so that we can design our own proteins from scratch.

Cellular DNA is actually two strands of letters laid head to toe, with each letter in one strand paired up with its partner in the other strand by these secret handshakes. The whole arrangement resembles a ladder, with the pairs of letters—base pairs—being the rungs. In fact, the molecule is twisted, so it really looks like a spiral staircase—Watson and Crick's famous double helix. When the cell wants to copy a particular piece of genetic information, it unwinds the relevant stretch of



A set of human chromosomes.

You might say that my assignment as a chemist is to develop a general method for finding needles in haystacks.

DNA and then separates the two strands from each other, like a zipper unzipping. Then pieces of RNA come in and make their handshakes with the exposed letters, so that the assembled RNA molecule transcribes the DNA's sequence and the information it contains.

The DNA ladder in each one of your cells contains about three billion rungs. Each rung is 3.4 Ångstroms tall—an Ångstrom is a ten-billionth of a meter—so at three billion rungs, that's roughly one meter of DNA per cell. The DNA obviously has got to be very tightly packed to fit in the cell. The DNA is tightly coiled like a telephone cord, except that the DNA coil is held together by proteins, and the coil twists around itself the way that the cord does when you hang up the phone. This tightly wound tangle of DNA is called a chromosome, and it's big enough to be visible under the microscope. A human being's meter of DNA is divided up into 23 chromosomes.

My research group is trying to build a molecular machine that can scan this whole meter of DNA and find one single location in it, reading its bumps and edges, its nooks and crannies, like Braille. DNA looks ribbon-smooth from a distance, but it's really quite lumpy when you look at it up close. (I should mention here that we understand the details of DNA's contours imperfectly, even today. It's only in the last few years that we've begun to get our first high-resolution glimpses of the double helix's stair-steps.) If we could learn a set of general rules for reading those contours, then we could design a set of molecules that would behave like a child's Lego set. We could assemble a bunch of pieces and the assembly would automatically snap onto the stretch of DNA that fits its shape The analogy is an apt one—each block has knobs, almost like teeth, that fit precisely into the holes in another block. If there's an extra knob sticking out, or the spacing between the holes is a bit off, the two blocks won't bind. Each and every knob–hole pair has to make the right handshake. We need such exact matching in the handshakes between our molecule and the DNA to guarantee letter-perfect sequence recognition. The problem is very difficult, because we need to be able to read DNA that's sitting on the library shelf, as it were—DNA in its compact, twisted-up form with the two strands zipped together. The zipped-up form only leaves a little bit of the edge of each base pair exposed, so we don't have much to work with.

But let's say we *can* find the rules to make the right set of Lego blocks to get that one-to-one recognition. Then, since we know the shape of each of the possible base pairs—and there are only four of them: AT, TA, CG, and GC—a biologist could give us a sequence of letters and say, "Here's part of the gene for cystic fibrosis," or "This belongs to a cancer gene," and we could assemble a molecule that would bind precisely, exclusively to the one spot in all that DNA where the gene actually is. You might say that my assignment as a chemist is to develop a general method for finding needles in haystacks. The biologists and the medical researchers will tell me what needles to look for—what sequences are important. In many cases, biologists know part

1.

2.

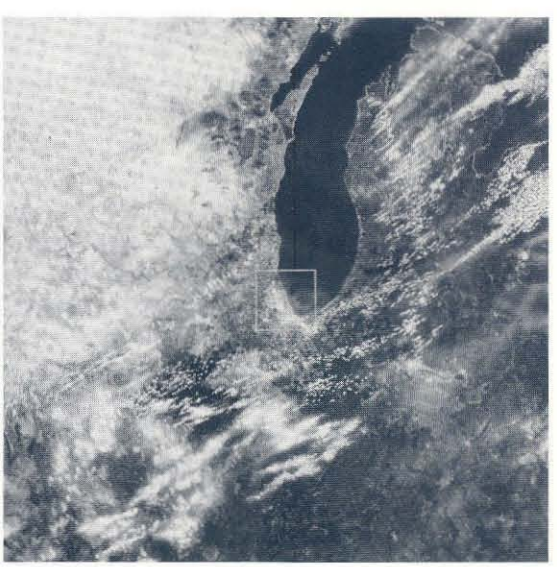
of the sequence of a disease-causing gene without knowing where the gene is. This is because genes are the blueprints for proteins, and if an aberrant or malfunctioning protein can be tied to a disease, then biologists can work backward from the protein to deduce what the gene looked like that gave rise to the protein.

We want to do more than just find genes, which can contain as many as 100,000 letters. We now know that a change of one—or a few—letters out of the whole three billion is sufficient to cause some diseases—not all, but some. There's no need to get too nervous about this news, though, because there can be lots and lots of mistakes all through your DNA, and they won't affect your health at all. And there are bits of machinery in every cell that go around all the time, fixing mistakes and repairing the DNA. But some errors, in some specific locations, can be very bad. We want to be able to find these errors, too.

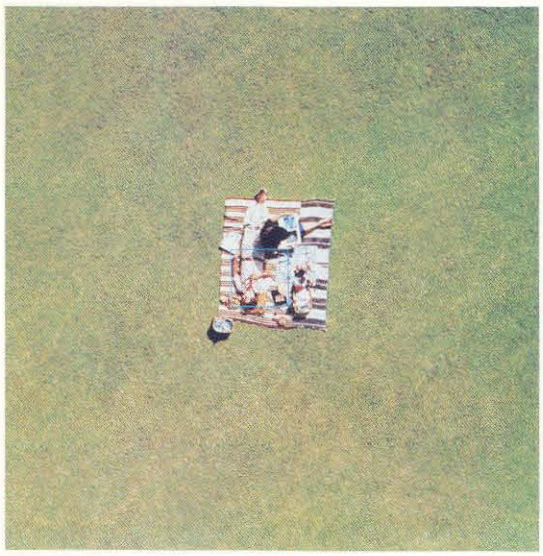
What does it really mean, finding one letter in three billion? According to the 1980 census, there are roughly 100 million residences in the United States. Let's assume that each one has 30 electrical outlets. (That may sound like a lot, but try counting the ones where you live sometime. You'd be surprised.) If there's a single dead wall socket anywhere in the U.S., I want to be able to find it rapidly. And I want to develop a general method, so that if I identify a bad socket in Wisconsin, and another one shorts out in Vermont, I can find it instantaneously.

I can explain our strategy by returning to the encyclopedia analogy. If we pull a volume off the

5.



8.



3.

4.

What does 3 billion really mean? The back of an adult human hand (opposite, top left) is about 3 inches from wrist to knuckle. Each successive picture shows an area ten times wider than the previous one, but centered on the same spot. The small square in each picture outlines the previous picture. Thus the hand belongs to a man having a picnic in Grant Park, in downtown Chicago, on the shore of Lake Michigan, on the continent of North America, on planet Earth. The final picture spans a view roughly 3 billion inches wide.

6.

7.

9.

10.

shelf and open it at random, a three-letter word like THE would occur often, but a 16-letter word like PREDETERMINATION would probably appear rarely. The larger the word, the rarer it's going to be. It's a simple mathematical exercise, really. There are 64—4 x 4 x 4, or 4^3—possible three-letter words we can make with a four-letter alphabet. In the three billion letters of the genome, each one of those 64 words should appear about 16 million times, assuming that all four letters, on average, are equally distributed throughout the genome. But there are roughly four and a quarter billion—4^{16}—ways to write a 16-letter word. Statistically, in the three billion letters of the genome, each one of those 16-letter sequences should appear rarely, or only once. In reality, some sequences occur over and over again in many different genes, but the point is that if we know a 16-letter sequence that's unique to the gene we're looking for, we can find it. And if we're looking for a single-letter error, what we need to do is look for a 16-letter sequence that includes our errant quarry. (Think what this would be like if we were working with English words—26^{16} is roughly 40,000,000,000,000,000,000,000!)

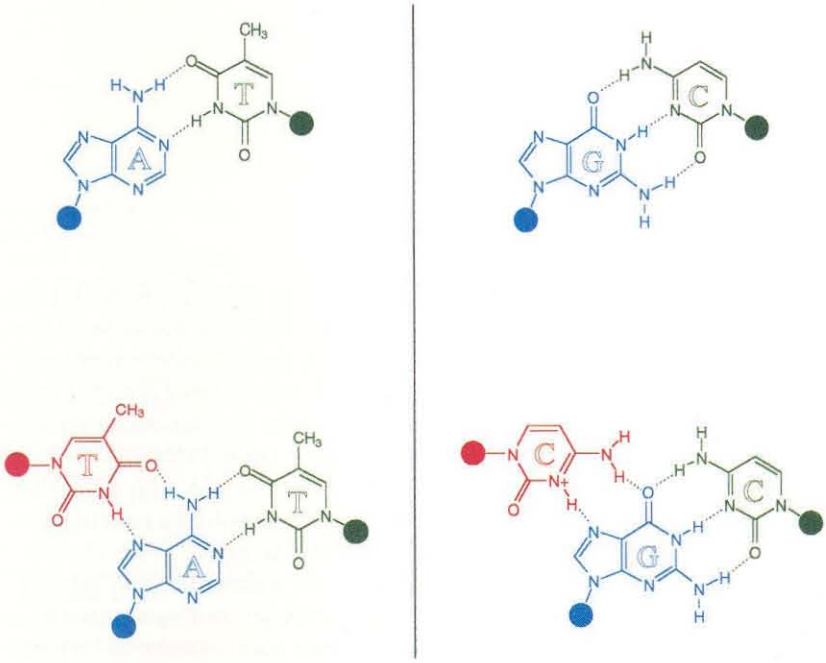
Biology—if not biologists—solved this problem a long time ago. Our cells are turning genes on and off at will every moment of our lives. Nature uses proteins—another class of polymer, another set of beads on a string—as molecular on/off switches. The protein alphabet is more complicated, having 20 letters. These letters also make handshakes with each other that cause the protein to fold up into a complex three-dimensional shape, and one portion of this shape's exterior surface reads the texture of the steps of the DNA spiral staircase. When the protein finds a location on the staircase that matches its reading surface exactly, it snaps onto that spot in yet another handshake. This DNA–protein handshake is an extraordinary one that scientists would dearly love to reproduce. The whole problem of how proteins fold to create such precisely engineered surfaces is a very complicated one that will probably take 10 years, and many researchers, to crack.

The ancients watched birds fly, and built themselves bird's wings, feathers and all, and jumped off of cliffs. That would be like me trying to duplicate how proteins recognize DNA.

But I'm impatient. I don't want to wait another decade (or two) until we figure out how proteins fold. Chemists are inventors—we're always creating new materials or rearranging old ones. Is there a way for chemists to make something that mimics nature's function—something whose behavior we could predict in advance, so that we could custom-design it to read the right shape? The key to the problem is the relationship between structure and function. Biology has evolved a structure that performs this function, but there might be other, less complex structures that are easier for humans to work with. The ancients watched birds fly, and built themselves bird's wings, feathers and all, and jumped off of cliffs. That would be like me trying to duplicate how proteins recognize DNA. But then people realized that a wing could be built out of wood and cloth—and later out of aluminum. It didn't look like a bird's wing anymore, but it had the same function. And now we can fly from Los Angeles to London in comfort, with air-conditioning and a movie, without having to ride on a bird's back, or, worse, flap our arms the whole distance.

Scientists are always building on other scientists' work. Every once in a while a Watson or Crick do something stunning that changes a whole field, but most science is built brick by brick. Sometimes a paper sits in the literature for a long time before someone sees an application for that work. Such a paper was written back in 1957, just after Watson and Crick proposed their double helix. Davies, Felsenfeld, and Rich—three physical chemists—reported that if you took double-helical RNA and simply added magnesium salts, the two-stranded polymer wound itself into a three-stranded polymer—a triple helix! This was an interesting anomaly but no one knew if it was really important. It was just a laboratory curiosity—an amusing chemical oddity—so it was duly written up. Nobody knew how the three strands bound together—they had no high-tech instruments back then to determine its detailed chemical structure.

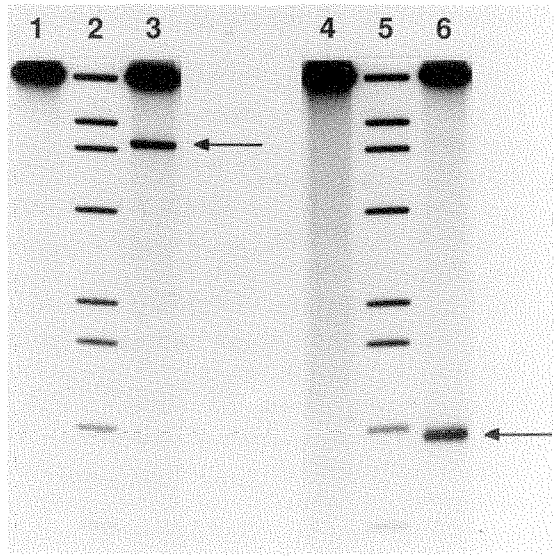
Could this three-stranded structure—the details of which are still imperfectly understood, and whose biological use, if any, remains unknown—be used for a new function: sequence recognition?



How the Caltech strand makes three-way handshakes with the Watson and Crick strands. The upper drawings on either side of the vertical line show the normal base-pair recognition seen previously on page 4. In the lower drawings, the Caltech strand is binding to the normal base pairs. The color scheme is the same as in the three-dimensional view on page 2.

Thirty years later, we read this paper and realized that if the third strand was lying on the steps of a normal, two-stranded piece of DNA like a carpet runner on a staircase, then we might be able to read a single site within a large piece of double-helical DNA by creating a short piece of DNA that would form a local third strand at that one site. In other words, could this three-stranded structure—the details of which are still imperfectly understood, and whose biological use, if any, remains unknown—be used for a new function: sequence recognition? And, in fact, the third strand can make very special three-way handshakes with the letters in the two normal strands. If we have an A-T pair as a step in the Watson-Crick spiral staircase, a T on the third strand can make a new handshake with the A without disturbing the rest of the staircase. Similarly, if we modify a C a little bit by putting an extra hydrogen ion on it, it will read the G in a G-C pair on the staircase. So if we have a DNA sequence on one Watson–Crick strand that consists only of As and Gs, we can make a third strand of Ts and Cs that will read that sequence and bind only to it. This is a very simple idea, because we can string together short, i.e., 16 letters long, sequences of Ts and Cs—called oligodeoxyribonucleotides—in our machines, and we don't have to worry about how the molecule we've made will fold up.

It's all very well to say that we're binding to one 16-letter sequence in three billion base pairs of DNA and no other, but how do we prove it? Well, we just add some new chemistry—we're inventors, after all. We give our molecule an



The result of the first site-specific recognition by triple-helix formation experiment. The DNA sample started at the top of each of the numbered lanes, and the fragments were drawn down the page by the electric field. Lanes 1 and 4 are the uncut DNA; lane 3 is the 3,000-letter fragment; lane 6 is the 1,000-letter fragment; lanes 2 and 5 contain sets of standard DNA fragments of known lengths that allow biologists to estimate the lengths of fragments in the other lanes.

Wherever our molecule binds, it will cut the DNA right next to that spot, leaving a permanent record of where it's been.

attachment that cuts DNA. We put it on the end of the molecule, like a stinger on a scorpion's tail. Wherever our molecule binds, it will cut the DNA right next to that spot, leaving a permanent record of where it's been. So if we've built a smart scalpel that finds a single site and cuts there, the DNA will be broken into two pieces. And if we use a piece of DNA whose sequence is already known, then we'll know exactly where that binding occurred, and how long each of the two broken pieces was. If we start with a DNA sequence 4,000 letters long, for example, and intend to cleave it after the 3,000th letter, we should get one fragment 3,000 letters long and one 1,000 letters long. But if we've built a molecule that doesn't recognize its target, then it will bind anywhere, and the DNA will be sliced into a million bits of different sizes.

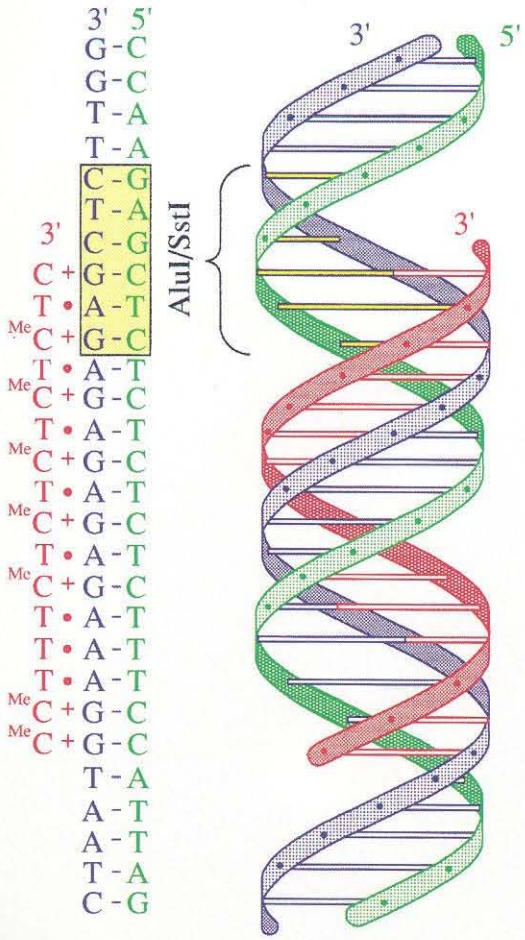
Biologists have a powerful separation technique to measure the size of DNA fragments. It's called gel electrophoresis. They put the DNA sample on one end of a slab of polymer, called a gel, that the DNA wants to stick to. Then they apply an electrical field across the gel. Now DNA is a polyanion, meaning that it has lots of negative charges scattered along its length, so the electrical field attracts it and starts to drag the fragments along the gel. The little fragments are easier to move than the bigger ones. After a time, the little fragments have moved a long way down the gel, with the smallest fragments moving farthest, while the big fragments are still lying near where they started. So if we've really cut our DNA sample in just one spot to make two pieces of unequal length, we will see two nice, sharp bands on the gel—one for each piece. But if we've cut the DNA at random, we'll get one long smear down the gel, made up of fragments of all sizes. Postdoc Heinz Moser, a member of our group, did the first experimental site-specific recognition by triple-helix formation in 1987. He used a piece of DNA a bit more than 4,000 letters long, and behold! he got the two fragments he expected to get.

Then we raised the stakes. We still weren't ready for human DNA yet, but two years ago we moved up to brewer's yeast—*saccharomyces cerevisiae*—which has 14 million base pairs of DNA in its genome. We picked one of its 14 chromosomes, which happens to be 340,000 base pairs long, and found that we could break it at a single site of our choosing. We did the cleavage with 95 percent yield, so we then knew that the method works on large DNA from a real organism.

Now we're ready to take on the real challenge—a human chromosome. We want to take this basic research, which started as a purely academic study of the chemical principles behind weak bonds, and perhaps do something useful with it, while at the same time we explore its scope and limitations. It turns out that the gene for Huntington's disease, an inherited neurological disorder, is on the tip of human chromosome 4. This has been known for several years, ever since Nancy Wexler did a pioneering study on a group of Venezuelan villagers. Everyone in the village came from just a few ancestors, and a large fraction of the village's population had Huntington's disease. So she was able to draw up a genealogy for every villager, and trace the

That doesn't mean that we've isolated Huntington's disease, but we've got a piece of DNA in our test tube that's 3.6 million base pairs in size, and it comes from the right place.

Gene hunters at work. Left: Designing a Caltech strand to bind to human DNA. The color scheme is the same as on pages 2 and 9. The set of six base pairs in the yellow box is the enzyme's cleavage site, which overlaps the 16-letter sequence that the Caltech strand can recognize. The normal base-pair handshakes are shown as dashes, while the three-way handshakes are symbolized as plus signs or bullets. The 3' and 5' indicate in which direction the DNA strand is running. Right: The Caltech strand nestles into the larger one of the DNA's two grooves—called the major groove—like a carpet runner on a spiral staircase.



gene's inheritance. Then she took DNA samples from the villagers, and, with the help of some brilliant biologists in Boston, was able to prove that the gene was on the tip of chromosome 4. The gene has been localized to the final three to four million base pairs of the chromosome—only a few percent of the overall length—so it's truly right on the tip. Now if we could isolate that piece in a test tube, perhaps the biologists could go in and actually pull out the specific gene.

We found a 16-base-pair sequence that we could recognize, because each of its letters was capable of making one of our three-way handshakes. It was in a piece of DNA, called a genetic marker, that is known to lie very close to the gene and is inherited along with it. Scott Strobel, a graduate student in our laboratory, has built the "Caltech strand" that should bind to that sequence and direct the cutting at that site by a DNA-cleaving enzyme. He's doing an experiment, in collaboration with David Housman's group at MIT, to see if that strand will find the Huntington's marker out of a background of 10 billion base pairs—10 billion, because he's working with mouse cells that have been altered to include human chromosome 4—and will snip the tip off of that chromosome. I believe he's done it. That doesn't mean that we've isolated Huntington's disease, but we've got a piece of DNA in our test tube that's 3.6 million base pairs in size, and it comes from the right place.

Whether isolating this particular fragment has placed a key piece of the puzzle in the hands of the geneticists who're searching for the Huntington's gene isn't yet known. It's fair to say, however, that our method will in all likelihood be sufficiently general to isolate almost any fragment of the human genome that's a few million base pairs in size. This may speed up the physical mapping of the thousands of genes important in human disease, and accelerate progress in the rational design of novel therapeutics that will alleviate human suffering. □

Peter B. Dervan, Bren Professor of Chemistry, got his BS in chemistry from Boston College in 1967, and came to Caltech in 1973 after earning his PhD in physical organic chemistry at Yale. This article is adapted from his Centennial Seminar Day talk.