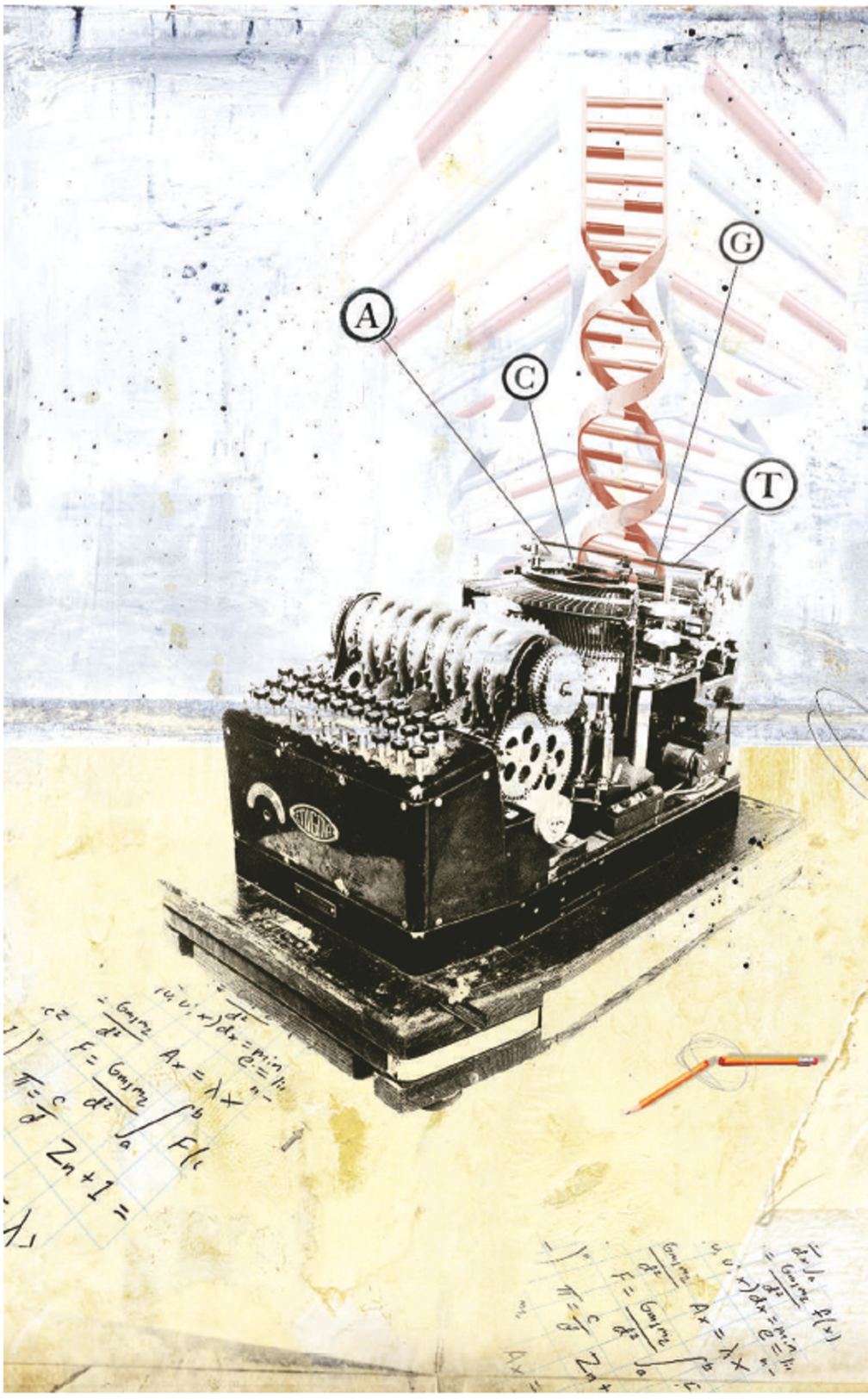


US NATL CRYPTOL. MUS./D. ALLISON

CODES AND ENIGMAS

There's more than one way to read a stretch of DNA, finds **Helen Pearson** — and we need to understand them all.

"It was known that they were a little acquainted but not as syllable of real information could emma procure as to what the truly was…" Reduce it to just a sequence of letters, and even a delicate phrase from Jane Austen's *Emma* becomes virtually impenetrable gobbledygook. So it was something of a triumph for Simon Shepherd when, in 2001, an algorithm he had written reconstructed all of *Emma*, word for separated word, from just such an uninterrupted string, despite being unacquainted with English vocabulary or syntax. The software worked out which groupings of letters were most likely to appear together, and thus have distinct meanings.

Shepherd, a researcher at the University of Bradford, UK, picked up much of his expertise during ten years cracking Russian codes in British Naval Intelligence. But he was not really interested in *Emma* — that was just a demonstration. His real goal was the far longer sequences of As, Gs, Cs and Ts that make up the world's genomes. Within those strings there is information that no one knows how to extract — codes that regulate, control or describe all sorts of cellular processes. And if the information is there, Shepherd thinks that number crunching should be able to pry it loose. "We are treating DNA as we used to treat problems in intelligence," he says. "We want to break the code at the most fundamental level."

That DNA contained at least one code was realized as soon as the molecule's structure was discovered. That code, cracked in the 1950s and 1960s, parses passages of DNA into three-letter combinations that correspond to particular amino acids. This is a code in the strictest sense; input determines output.

But researchers now know that there are numerous other layers of biological information in DNA, interspersed between, or superimposed on, the passages written in the triplet code. Human DNA contains tissue-specific information that instructs brain or muscle cells to produce the suite of proteins that make them brain or muscle cells. Other signals in the sequence help decide at what points DNA should coil around its scaffolds of structural proteins. These are the codes that computer buffs such as Shepherd want to crack with raw processing power — and that mainstream biologists are attacking, too, although using a rather more lab-based approach. "We need all these codes together to understand the dynamics of the cell," says computational biologist Manolis Kellis at the Massachusetts Institute of Technology in Cambridge.

The DNA sequence contains information

not just about the make-up of proteins but also about the interactions of DNA with some of those proteins, and the diverse antics of RNA. The analysis of DNA sequences is revealing patterns that have meanings at all of these levels. "Biology has probably figured out a way to squeeze every bit of information from that molecule it can," says Jason Lieb, who studies DNA–protein interactions at the University of North Carolina at Chapel Hill.

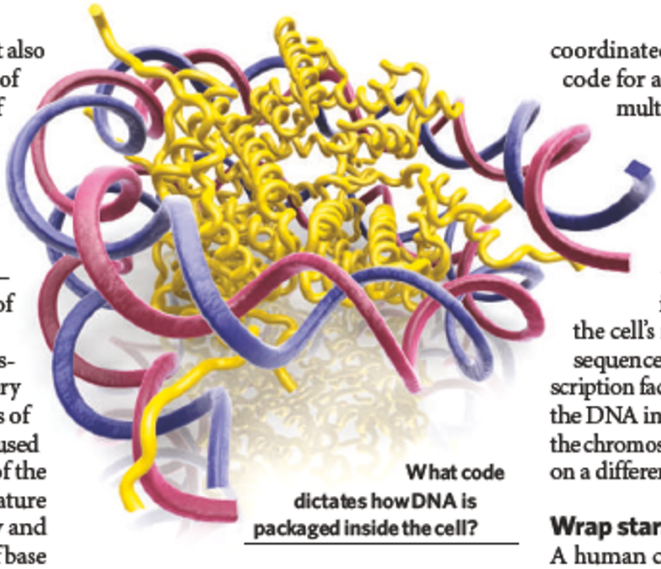
The code that is currently most exercising the minds of geneticists is the 'regulatory code' that directs the production of suites of proteins tailored to specific cell types and used at specific times. The idea is that many of the genes switched on in DNA contain signature sequences in 'promoter' regions nearby and 'enhancer' regions that may be millions of base pairs away. In a blood cell, say, these signature sequences might be bound by proteins A, B, C and D, whereas genes switched on in skin may be regulated by signature sequences that bind proteins B, C, Y and Z.

"The biggest obstacle after the sequencing of the genome has been to understand how genes are regulated and how we can see that from the sequence," says Jussi Taipale, who studies gene regulation at the University of Helsinki, Finland. "It's a more complex code than the genetic code." The first difficulty is the sheer scale of the problem. Human cells contain more than 20,000 protein-coding genes, roughly 1,500–2,000 transcription factors, which switch genes on and off, and numerous other regulatory proteins and RNAs that direct their production. The possible permutations and combinations are bewildering.

Lost in translation

One way to start solving the regulatory code en masse would be to find all the positions where each of the regulatory proteins binds within the genome. Many transcription factors show a penchant for binding specific short motifs in DNA, such as a six-letter sequence. In theory, a computer could scan for any such motifs that occur more often than might be expected by chance.

But there are drawbacks. For one thing, a given six-base-pair sequence will sometimes be a binding site and sometimes not, probably depending, in part, on whether the DNA is folded up in a way that prevents transcription factors from gaining access. For another, the way that these sites are recognized is not as specific as the binding between the bases that translate the triplet code into protein. Transcription factors recognize DNA sequences from the effects of the sequence on the outside of the helix, and although this recognition is still sequence dependent, it is not quite so precise. Some of these proteins will bind to a range of related sequences — sometimes more tightly, sometimes less so — and those subtleties of affinity, like the nuances of a social embrace, may themselves have biological meaning.



PHANTATOMIX/SPL

What code dictates how DNA is packaged inside the cell?

Len Pennacchio at the Lawrence Berkeley National Laboratory in California and his colleagues have begun to fathom some of these subtleties by identifying a rudimentary tissue-specific code for the human brain[1]. They teased out the relevant enhancers from the human genome by comparing the human sequence to those of distant relatives such as the pufferfish (*Takifugu rubripes*), pulling out regions that didn't describe proteins but that evolution had nevertheless deemed important enough to keep intact. They then systematically inserted 167 such regions into mouse embryos and found that 45% of them provided tissue-specific ways to switch on genes.

The team identified four enhancers that boost gene activity in the developing forebrain and share several short-sequence motifs that are presumably binding sites for control proteins. By searching for similar signature sequences in the human genome, they located other forebrain enhancers, suggesting that they have found some of the sequence information that 'means' brain-specific in the regulatory code.

Taking a slightly different tack, Richard Young at the Whitehead Institute for Biomedical Studies in Cambridge, Massachusetts, and his colleagues have come up with a preliminary code that distinguishes human embryonic stem cells[2]. They extracted human DNA bound by three key transcription factors and determined all the sequences to which those proteins chose to bind. The proteins recognize sequences near genes that need to remain active for stem cells to stay stem cells; they also recognize other sites where they seem to help shut down the genes needed for the stem cells to differentiate into other cell types. So these proteins, in combination with others, seem to stop stem cells from becoming other cell types.

Many researchers are now talking about a coordinated effort to identify the regulatory code for all human transcription factors in multiple tissues. But they are unlikely to resolve this code without simultaneously extracting other layers of overlapping information in DNA.

A section of DNA can contain two or more layers of information that are used at different times or in different ways depending on the cell's requirements. So whether a given sequence is read as a binding site for a transcription factor to some extent depends on how the DNA involved is packaged at that point in the chromosome — and that packaging depends on a different code stored in the DNA.

Wrap stars

A human cell has to fit about two metres of DNA into a nucleus a few micrometres in diameter; that requires packing it together with proteins in a complex hierarchy of folding back and wrapping round. The fundamental element underlying all this packaging is the nucleosome — 147 base pairs of DNA wrapped around a globule of eight proteins called histones. Up to 90% of DNA is bundled up into nucleosomes, and their position influences the DNA's activity. Sequences wrapped up in nucleosomes are often less accessible to transcription factors and so less likely to be transcribed. It has been known for more than two decades that in the test-tube certain sequences are more likely to be packaged up in nucleosomes. But in the real hustle and bustle of the cell, it was unclear to what extent such preferences get honoured.

"We are treating DNA as we used to treat problems in intelligence." — Simon Shepherd

Earlier this year, Eran Segal at the Weizmann Institute of Science in Rehovot, Israel, Jonathan Widom at Northwestern University in Evanston, Illinois, and their colleagues came the closest yet to defining a code for the position of nucleosomes[3]. They took DNA wrapped up in nearly 200 yeast nucleosomes, and 177 from chickens, and exposed it to enzymes that would eat up all sequences in between the nucleosomes. They then sequenced the DNA left intact in the nucleosomes, and used computational methods to align the sequences and search for common patterns.

The team came up with a set of rules that could predict where more than 50% of nucleosomes lie in yeast and chicken DNA. "It's much less than perfect but way better than random," Widom says. The main rule is that the sequences AA, TT or TA are more likely to be found where the spiralling DNA backbone grazes the histone — they seem to help the DNA bend around the protein core.

But Segal and Widom's rules can't predict the position of a significant fraction of the

nucleosomes. DNA's overlapping codes mean that an individual nucleosome might be usurped if regulatory proteins are already tightly bound there. The nucleosome code depends on the regulatory code, just as the regulatory code depends on the nucleosome code. In addition, the position of a nucleosome might be influenced by the way in which the nucleosome-wrapped sequence is folded and condensed yet further. "The code specifies the initial state and the cell can mess with what happens afterwards," says Oliver Rando, who studies nucleosome positioning at Harvard University.

"It's much less than perfect but way better than random."
— Jonathan Widom

The goal now is to find codes that govern those larger-scale features of DNA packaging, such as how the nucleosomes are twisted up into a cable of chromatin and eventually coiled into the tightly interwoven ropes of the chromosome. As yet, though, researchers have not found landmarks equivalent to nucleosomes that can guide the search for meaning — nor is it clear that they will. "There could be diffuse information spaced at hundreds of kilobases that helps package even larger pieces of the genome together," says Lieb. "Or it could be that the exact position of those structures is not important."

Room for manoeuvre

DNA seems well adapted to supporting a number of codes. For a start, only 1–2% of the human genome is occupied with protein-coding sequences, which leaves plenty of intervening DNA to hold other information. But many stretches of DNA in humans and other organisms manage to multitask: a sequence can code for a protein and still manage to guide the position of a nucleosome. This is possible because the triplet code is 'degenerate'. Several slightly different triplets can code for the same amino acid, and many positions in a protein can be filled by different amino acids — so different sequences can effectively mean the same thing. This allows other signals to be imprinted on top of the first — especially when those other signals are themselves encoded with some slack.

This elegance is surely the handiwork of evolution — and if the way in which that hand had worked to solve these problems were clearer, the simultaneous decoding of all the messages involved might become easier. Perhaps ancestral organisms had simpler sequence patterns that evolution has optimized, taking advantage of its degeneracy to layer in additional information that helped organisms acquire extra complexity. Hanspeter Herzel, who specializes in statistical analyses of DNA at Humboldt University, Berlin, speculates that the space constraints of the cell may have favoured the development of nucleosomes that wound up unruly DNA — and that their existence then encouraged the evolution of a nucleosome code in the sequence because this lowered the energetic cost of coiling up DNA. But as yet such ideas, and any help they might offer, remain tentative. "We don't really have a phylogeny of these signals," he says.

And in some cases, it seems that evolution may have generated patterns that have no clear biological function. In 1992, Gene Stanley at Boston University, Massachusetts, and his co-workers created waves when they suggested that there were patterns in DNA that spanned hundreds and thousands of base pairs[4]. Stanley used the types of statistical techniques that identify correlations in climate and financial data and applied them to all the DNA sequences available in databases at the time.

Essentially, the study showed that a region with a particular chemical composition, such as one loaded with the bases A and G, is likely to be followed by a similar region hundreds or thousands of base pairs away, and that the probability of this pattern declines in a predictable way with distance. It also found that this correlation existed predominantly in DNA that did not code for protein, leading Stanley to propose that DNA previously written off as junk actually carries biological information.

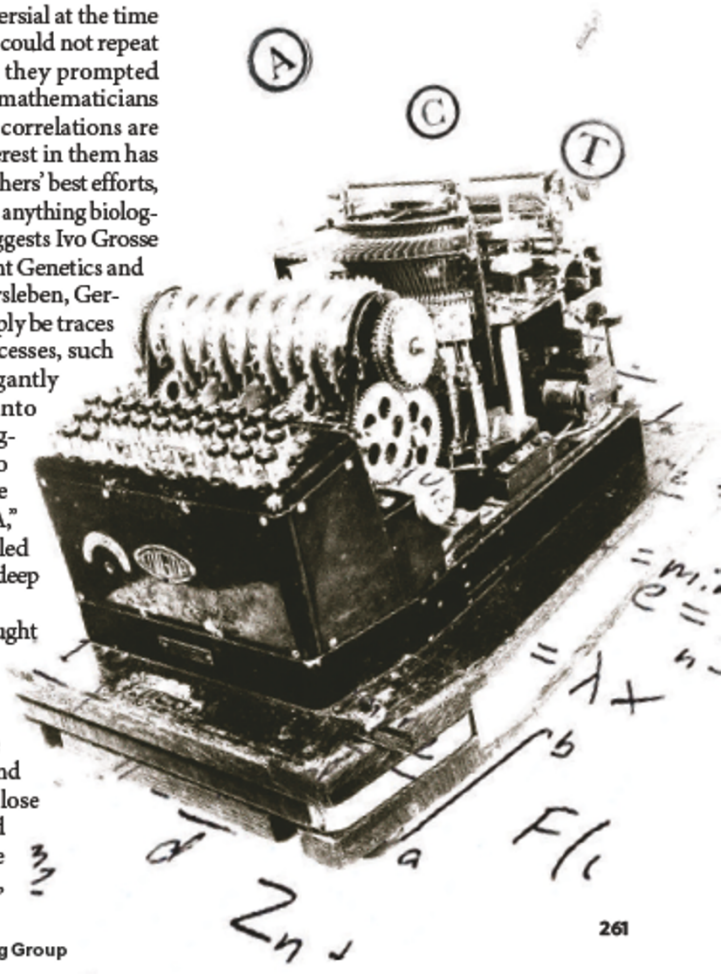
The findings were controversial at the time because several other groups could not repeat aspects of the analysis, and they prompted huge interest in DNA from mathematicians and physicists. Today, these correlations are thought to be real — but interest in them has faded because, despite researchers' best efforts, the patterns have not revealed anything biologically important. Perhaps, suggests Ivo Grosse of the Leibniz Institute of Plant Genetics and Crop Plant Research in Gatersleben, Germany, the patterns could simply be traces of random evolutionary processes, such as the erosion patterns elegantly but accidentally carved into sandstone by the wind. "Long-range correlations definitely do exist, but I don't think it's some supercode imprinted in DNA," Grosse says. "We just stumbled on a feature with probably no deep biological meaning."

But to some people the thought of order with no meaning is an affront. To such minds, the idea of teasing out nature's secrets with little more than mathematical cunning and processing power will never lose its allure. When Shepherd and his graduate student Natalie Kay, in unpublished work, ran the software that they had tried out on *Emma* over the (admittedly small) genome of Ebola virus, it identified as meaningful some sequences that, at the time, bore no annotations in genetic databases. Only later, Shepherd says, were these motifs recognized by biologists as passages that control the activity of genes or mark their ends. He thinks that approaches based on almost pure number crunching will go on to rock the field: "I firmly believe that major advances in this over the next 20, 30, 50 years will be made by the theorists, not the medics."

But researchers versed in the complexities of how DNA and proteins actually work remain convinced that their type of knowledge will remain vital to sorting the meaningful from the circumstantial. When the triplet code was first being studied, there were any number of fanciful mathematical and logical approaches to it — but the approaches that paid off were the ones informed by the greatest degree of biological insight. "Computer scientists think they can just walk in the door and solve things," says bioinformatics expert Wyeth Wasserman at the University of British Columbia in Vancouver, Canada. "But they come to realize you need biology too." ■

Helen Pearson is a reporter for *Nature* based in New York.

1. Pennacchio, L. A. *et al. Nature* doi:10.1038/nature05295 (2006).
2. Boyer, L. A. *Cell* **122**, 947–956 (2005).
3. Segal, E. *et al. Nature* **442**, 772–778 (2006).
4. Peng, C.-K. *et al. Nature* **356**, 168–170 (1992).



U.S. NATL. CRYPTOL. MUS./D. ALLISON