BIOLOGY

AI Helps Researchers Find Meaning in Molecules

At the Imperial Cancer Research Fund in London, a group of researchers led by Dominic Clark has puzzled out the rough structure of a protein that plays a role in the spread of certain cancers. The structural information, which the group intends to publish later this year, should help guide efforts to disarm the protein and reduce the cancer's threat. But just as significant as the clinical promise is how the discovery was made. Part of the credit goes to a nonhuman collaborator: a computer running an artificial intelligence (AI) program.

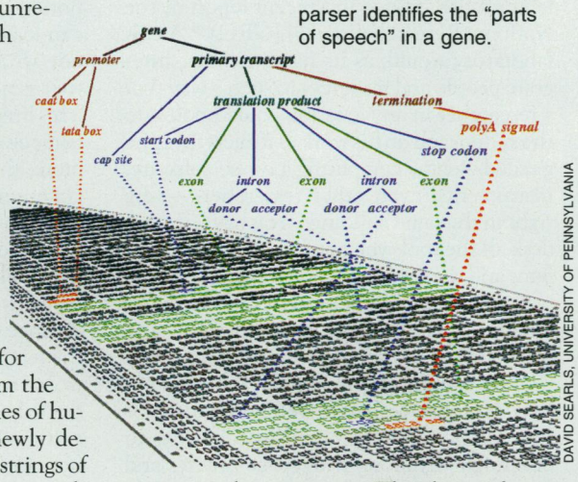
Clark and company may be ahead of their time—but not by much. They are part of a small but growing group of molecular biologists who have turned to artificial intelligence for help in making sense of sequence data on DNA and proteins. Besides zeroing in on protein structures, AI is helping researchers find genes in large stretches of DNA and evaluate the effects of mutations on certain genes. Unlike conventional computer programs, which can only carry out steps that are specified ahead of time, AI programs can reason on their own and make connections between seemingly unrelated pieces of information—which makes AI especially valuable to anyone looking for patterns in otherwise overwhelming amounts of data. "These kinds of tools are extremely useful," says Chris Fields, director of a fully automated sequencing lab that churns out 400,000 bases worth of DNA sequence data each week. "The payoff is that we can do better biology."

A major source of employment for these software assistants comes from the effort to sequence the entire genomes of humans and other creatures. With newly decoded genetic sequences—the long strings of nucleic acid base pairs that make up strands of DNA—rolling out of laboratories at a rate of millions of base pairs per month, "we've got sequences up the wazoo," groans Temple Smith, a Boston University biologist. "We need a way to keep from being overwhelmed."

Enter Grail, one of AI's first triumphs in molecular biology. First made available to researchers over the Internet computer network in 1991, Grail employs a "neural network"—a technique modelled after the human brain's approach to pattern recognition, in which an input signal (such as an image on the back of the eye) cascades through a forest of interconnected neurons until an output pattern (a recognition of the image as, say, a car) emerges. Grail's assignment is to pore over raw sequence data—long, unbroken strings of G's, T's, A's and C's, the four nucleic acid bases—and pick out the genes hiding in those random seeming sequences from the long stretches of DNA regarded as "junk."

Actually, explains one of Grail's creators, computational biologist Ed Uberbacher of Oak Ridge National Laboratory, the original goal was simply to identify the protein-coding segments of genes. These are the parts of a gene that contain the recipe for stringing amino acids together to make a protein, and they may constitute only 2% or so of a gene's sequence. The rest of the gene consists of such things as regulatory regions, which determine when the gene will become active, and introns, which separate the protein-coding segments of DNA and tend to be of less interest to molecular biologists.

Over time, biologists have learned certain rules for picking out genes and their protein coding segments. They know, for instance, that the three-base sequences TAA, TAG, and TGA are "stop codons" that mark the end of a protein-specifying section. But simply building such rules into a computer program wouldn't give it sufficiently keen discrimination, since many of DNA's important codes are still unknown.



Parlez-vous DNA? An AI parser identifies the "parts of speech" in a gene.

© 1992 DAVID SEARLS, UNIVERSITY OF PENNSYLVANIA

Because Grail is a neural network, it could bypass these limitations by learning to recognize protein-coding segments on its own. Uberbacher "trained" Grail by showing it DNA sequences in which the genes were already known and having it guess where the protein-coding sections were, then rewiring Grail's software "neurons" so as to strengthen connections that gave accurate guesses and weaken those that produced poor results. Eventually, the program came to recognize protein-coding sequence segments with about 80% accuracy.

Not content with that figure, Uberbacher revised Grail to take non-protein-coding segments into account. Just as trying to identify an unlabeled state on a U.S. map is easier when the neighboring states are identified, so picking out protein-coding segments becomes easier when neighboring segments are known. Uberbacher says the program's accuracy is now 90%.

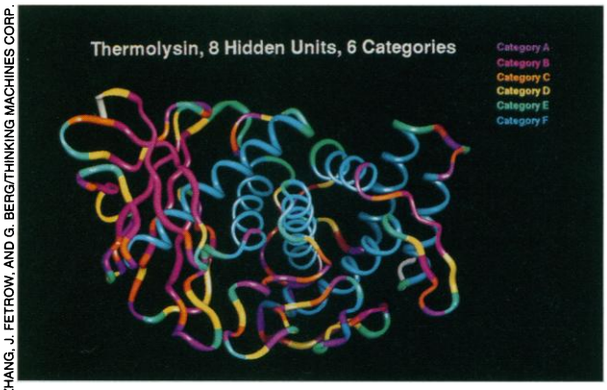
The grammar of DNA. At the University of Pennsylvania, David Searls, a biologist and—like many in his field—also a computer scientist, is taking a different approach to AI-based gene mapping: He treats DNA as a language in which genes are "sentences" and tries to teach a computer to pick out which DNA sequences are grammatical sentences (i.e., genes) and which are nongrammatical nonsense. The roots of this linguistic approach go back to the pioneering work of Massachusetts Institute of Technology linguist Noam Chomsky, who in the 1950s defined a language as the set of "strings" that can be formed by chaining together a group of symbols—whether the symbols are the letters of the English alphabet, the 0s and 1s of computer instructions, or the four nucleic acid bases.

Viewed in this light, DNA is a bit like a series of English sentences whose words are run together without capital letters or punctuation. Making sense of the sentences requires figuring out where they start and end, and how they are broken into phrases by commas. We can do that in English by enlisting our knowledge of grammar—the set of rules defining how parts of speech can be related—and Searls hopes to unlock DNA's "grammar" in order to recognize the beginning and end of genes and their component parts. Just as English grammar dictates the way nouns, verbs, and other parts of speech can be combined into phrases which in turn make up sentences, so DNA's grammar should explain how "words" (such as codons) can be arranged to form "phrases" (such as the parts of the gene that regulate how and where enzymes bind to the DNA in order to read the gene sequence), and how such phrases are structured into a complete genetic sentence (a gene). Searls notes that genes can even be grouped into "paragraphs"—strings of genes that address a single "theme," such as the group of genes that code for cell surface receptor proteins.

To discover DNA's grammar, Searls has developed a computer program called a "parser," which attempts to split a genetic sequence into increasingly lower-level parts—

first into genes, then into protein-coding versus non-protein-coding segments, then into codons—just as sixth graders break down English sentences into phrases and then parts of speech. The hope is that, by practicing on known sequences, the parser can learn the genetic grammar. Later, it could apply the grammar to unknown sequences to see whether they can be parsed into genes and components of genes. Right now, Searls' system is able to identify parts of a gene with about 75% accuracy, and Searls expects that number to improve.

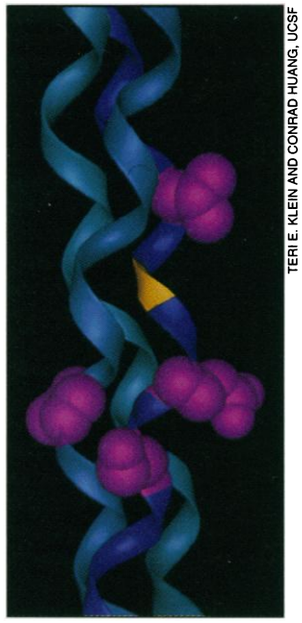
Puzzling out protein structure. Besides deciphering DNA sequences, biologists are faced with the job of making sense of another kind of raw data: the strings of amino acids that make up proteins. Many of a protein's properties, including how it interacts with other molecules in the body, depend on the specific three-dimensional structure into which the string of amino acids folds and curls, but the protein's sequence, at least at first glance, reveals little about this structure. Biologists can sometimes determine a protein's structure with x-ray crystallography, but this is a slow, tedious process. Researchers would like to be able to predict the structure merely by knowing the one-dimensional sequence of amino acids that constitute the protein.



X. ZHANG, J. FETROW, AND G. BERG/THINKING MACHINES CORP.

Body by AI. A neural network determined the local structures of this protein by examining the amino acid sequence.

This has proved an exceptionally difficult problem. Some conventional computer programs have had limited success in predicting "secondary structure"—the handful of more-or-less generic shapes such as the corkscrew-shaped alpha-helixes and multi-stranded beta-sheets that crop up along the protein and can provide important clues to a protein's functions. Essentially, these programs work by tracking the relative abundance of various amino acids in sections of those few proteins whose structures have been determined through crystallography; given a new, unknown protein, the programs assume that similar ratios of amino acids imply similar secondary structures. Unfortunately, their predictions are wrong nearly half the time.



TERI E. KLEIN AND CONRAD HUANG, UCSF

Bad influences. In the protein collagen, AI picked out regions *(magenta)* that affect the lethality of a point mutation *(yellow)*.

Now AI is improving the odds. The current champ in secondary structure prediction is a system developed by Xiru Zhang of Thinking Machines Corp. in Cambridge, Massachusetts, and his colleagues. Zhang's system comprises three "expert" modules that follow different predictive strategies and hone their skills on a database of known structures. The first expert is a neural network that learns to associate the patterns in long sequences of amino acids with particular secondary structures. The second simply looks for the known sequence that most closely resembles the unknown sequence and offers the known sequence's structure as its prediction.

The third expert uses statistical techniques to calculate which short sequences are most often associated with which type of structure. A "combiner" module melds the predictions from these three molcules into a single prediction. The system predicts the correct secondary structure 67% of the time, Zhang says, which he and many other researchers suspect may be about as good as it gets without taking into account amino acids beyond the immediate area under consideration. "It may be that interactions between amino acids far apart on the protein are affecting the way it folds," Zhang explains.

If predicting secondary structure is tough, predicting "tertiary structure"—the complete three-dimensional form—is next to impossible, but Clark at the Imperial Cancer Research Fund has got a good start on it. His AI program tackles a protein's "topological structure"—the rough spatial relationship between its helixes, sheets, and plain strands. Clark's system is equipped with a handful of rules about protein folding gleaned from biochemistry and statistical observation. One rule, for example, states that when a helix is connected to a plain strand on either end, the whole assembly will tend to fold up like an accordion, with the helix parallel to and sandwiched between the two strands. In operation, the system examines secondary structures and tries to apply the rules wherever possible so it can at least eliminate implausible topological structures. In some cases, that's enough to make an experimentally verifiable prediction of the structure, as was the case with the sought after cancer-related protein. Clark won't identify the protein pending further verification, but he claims that early results suggest the computer got it right. "We're all very excited about this," he says.

Some AI work has gone beyond identifying sequences and structures to making predictions about function. National Institutes of Health computer scientist Lawrence Hunter and University of California, San Francisco, chemist Teri Klein are studying osteogenesis imperfecta, a bone disease caused by mutations in the gene that codes for the protein collagen. Depending on which part of the protein it affects, the mutation can be fatal to fetuses. As a first step toward developing tests for carriers of the mutation and perhaps even treatments, the two scientists would like to predict which mutations will be lethal.

AI is aiding the prediction effort by helping them devise rules about mutation lethality. Hunter and Klein's system begins by studying a small database of lethal and non-lethal mutated collagen genes and picking out features of the gene sequence that appear to be correlated to lethality; having narrowed the search, it analyzes these features to produce sets of rules that predict the lethality of a mutation. Finally the system performs statistical tests of its hypotheses to decide if any of them are promising, extracting and combining those rules that best survive the tests.

One discovery so far: It's not just the site of the mutation that determines its lethality, but also which amino acids show up an even number of positions down the protein from the site of the mutation—a finding that promises a clue not only to how the mutation works its effects, but also to the structure of the protein. "It was a totally unexpected result," says Hunter. There are certain to be many more unexpected patterns lurking in the blizzard of genetic code and protein sequences facing molecular biologists. And Hunter, like many of his colleagues, sees no alternative to unleashing AI's computer code on this storm of code from biology. As he puts it: "If AI can't do it, I don't know what can."

–David Freedman

David Freedman is a free-lance science writer in Brookline, Massachusetts.