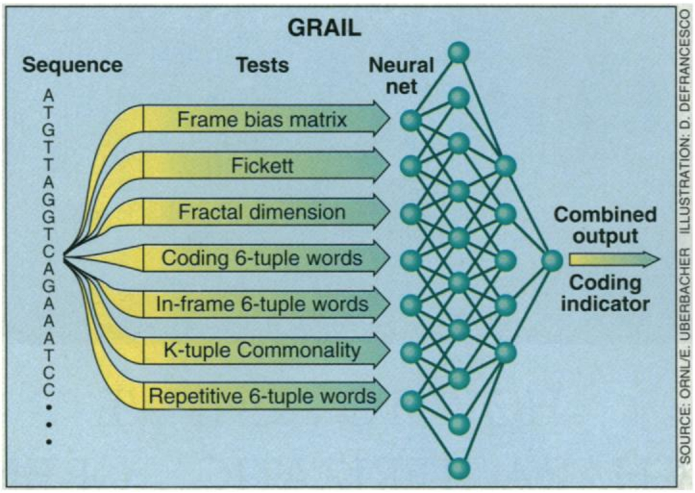
GRAIL Seeks Out Genes Buried in DNA Sequence

When the Human Genome Project achieves its ultimate goal, supposedly around 2005, biologists will have in hand the exact sequence of all 3 billion nucleotides arrayed along the human chromosomes. But they have never been entirely sure how they will read the language of that long string of As, Gs, Ts, and Cs. How will they even be able to pick out the genes, which account for a mere 5% of the genome, from the mass of letters in between?

Now Edward Uberbacher, a biophysicist-turned-computational-biologist at Oak Ridge National Laboratory, has come one step toward providing an answer: a new artificial intelligence program, called GRAIL, that can pick out the coding regions of genes in a long stretch of sequence data. Uberbacher and his colleagues—molecular biologist Richard Mural and computer scientists Ralph Einstein, Xiaojun Guan, and Reinhold Mann—made the program available to other labs by e-mail last July, and he described it at *Science*'s Human Genome III meeting in San Diego in mid-October.

So far, the Oak Ridge team has analyzed 5 million bases of DNA, "and I doubt we have missed more than a few genes," asserts Uberbacher. "One year ago, even 6 months ago, it was virtually impossible to go into human genomic sequence and find genes by computer with any reliability. Now we can go in and find 90% of the genes very quickly." What's more, he says, GRAIL can be used on a PC, not a supercomputer, and it provides an answer almost instantly.

"It is the niftiest thing I have encountered this year," says Francis Collins, a gene hunter at the University of Michigan, of the new program. "It starts to make me a believer in the assumption that having the sequence will enable us to find the genes." Collins has good reason to be enthusiastic: He is using GRAIL to zero in on the region of chromosome 4 where the Huntington's gene is thought to reside. So far, he says, the program's predictions have been "superb," though the long-sought gene still remains elusive.



SOURCE: ORNL/E. UBERBACHER ILLUSTRATION: D. DEFRANCESCO

GRAIL asks seven questions of each 100-base section of DNA sequence, then feeds the answers into a neural network, which learns, by trial and error, to pick out the coding regions.

The basic idea in developing GRAIL, Uberbacher explained in San Diego, was to feed as much biological information as possible to a neural network and then let it learn to distinguish the exons, the coding regions of genes, from the introns, or noncoding regions. The team developed several new algorithms, or statistical tests, for identifying exons. They also adopted some existing ones, such as that developed by computational biologist Jim Fickett at Los Alamos National Laboratory in the early 1980s. The various tests recognize certain patterns in the sequence that are characteristic of exons. "Each of them is imperfect and works only part of the time," explains Uberbacher, but the Oak Ridge technique combines the tests—a novel feature that makes the approach more powerful.

Specifically, the program analyzes sequence 100 bases at a time, asking seven statistical questions of each "window" of DNA. For instance, Uberbacher's group and two others recently found that there are certain 6-base "words" that occur much more frequently in coding regions. "An analog is, if you looked at a page of an engineering text and a page of a romance novel, you could tell by looking at a few dozen random words which was which," says Uberbacher.

Based on these statistical tests, the program accumulates circumstantial evidence, which it then presents to the neural net. "We train the neural net," explains Uberbacher. "We show it a bunch of examples of coding and noncoding sequences that are known, and it learns how to use that information. It decides how important each of the tests is to the final answer. As you show it more and more examples and then test it, just as you would a student, it is eventually able to distinguish coding from noncoding regions with very high success," though it cannot pinpoint their exact location.

When the Oak Ridge team tested GRAIL on 19 human genes that had not been used in training, "it located 90% of the coding exons of 100 bases or more in length," says Uberbacher, but it found only about half of the smaller exons. The false negative rate is 10% and the false positive rate is 20%—a significant improvement over existing exon-recognition programs. "Virtually every other method produces more false positives than real exons," says Uberbacher, who adds that it is still too early to say how GRAIL compares with another neural net program being developed by computer scientist Alan Lapedes at Los Alamos National Lab.

Although GRAIL is a major step forward, it is still not the final answer for finding human genes. Like other exon-recognition programs, it says nothing about gene structure—for instance, precisely where the exons start and stop or the location of the dividing lines between exons and introns. And gene structure is what biologists will ultimately need to know to be able to discern, by merely looking at the sequence, what protein the gene encodes. Uberbacher, Mural, and colleagues are already working on just such a "gene assembly" program—one that will predict not just the exons but all the functional sites. Several other computational biologists are pursuing the same tack, including Chris Fields, now at the National Institutes of Health, who introduced the first such program in 1990; Phil Green at Washington University in St. Louis; and Temple Smith at Harvard. It will be an awesome task; indeed, Fields says it is "very likely impossible to solve with complete reliability." Still, these groups have made a promising start. "For the first time, this gives a rationale for just sequencing," said David Smith of DOE's Office of Health and Environmental Research after hearing Uberbacher's talk. "We are beginning to be able to read the language." ■ LESLIE ROBERTS