Gene Counters Struggle to Get the Right Answer

Until researchers determine what constitutes a gene, they can't tally up how many humans have. In the meantime, gene-hunting programs are becoming more sophisticated

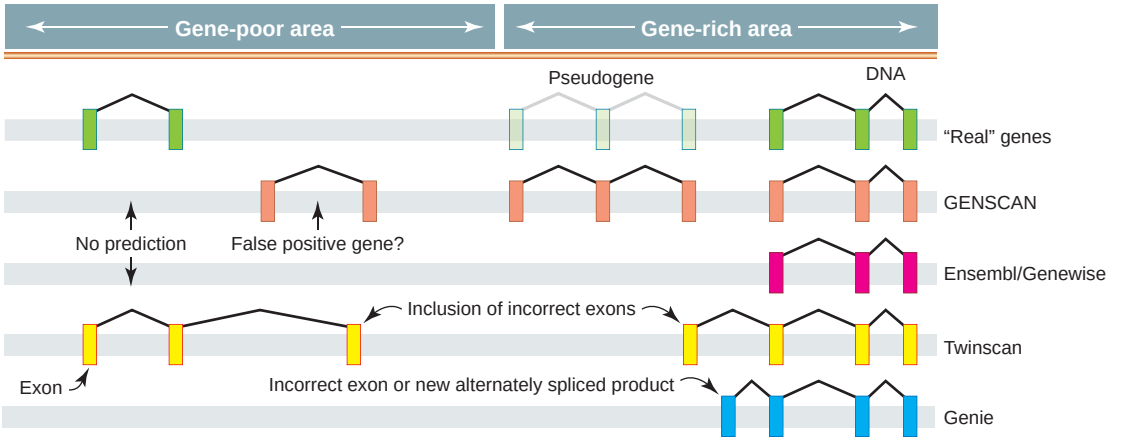
Researchers have been counting human genes for decades, but the numbers just don't add up. The best estimate soared to 100,000 a few years ago, dropped to about 30,000 when the human genome sequence was published, and recently sank as low as 20,000. To take full advantage of the sequencing of the human and other genomes, researchers say, they need a better accounting.

In more optimistic times—a mere 3 years ago—the genome-sequencing community started a betting pool called GeneSweep on what the number of human genes would turn out to be once the sequence was finished. This summer, researchers admitted they were nowhere near establishing a final count. They decided to end the suspense, saying that the books balance out at 24,500 protein-coding genes for now and naming a winner (*Science*, 6 June, p. 1484).

The problem, says David Kulp, a computer scientist at Affymetrix in Santa Clara, California, is that when it comes to defining a gene, "it's very difficult to say definitely what's right or wrong." Molecular biologists are finding that some genes are shorter than anybody expected a gene to be. Sometimes it's hard to tell whether a piece of code is a single gene or two that overlap. And the community is not quite sure how to classify genes that code for multiple proteins or genelike sequences that code only for RNA.

This complexity has taxed bioinformaticists to the limits of their software-writing abilities. People in this decade-old field design computer programs to analyze DNA sequence data, which includes detecting genes. Their mathematics are increasingly sophisticated, with algorithms that take into account the geneticist's best knowledge of genes and proteins, as well as the molecular biologist's insights into how genes are hidden in DNA. Some of these computer buffs have even started doing their own experiments to characterize genes better.

They have a lot of work to do. Often they can tell that a stretch of DNA codes for an amino acid sequence, but the size, number, and exact distribution of protein-coding and noncoding regions in that gene remain elusive. Most worrisome to some is so-called dark matter, seemingly geneless regions in a genome that might contain hidden coding sequences.



Never perfect. No program calls all genes correctly. Some see genes (shown here as coding regions, or exons, connected by bent lines) where there are none; some miss a gene altogether; and some don't put all the gene's parts in the right places.

Elusive prey

In the 1930s, George Beadle and Edward Tatum suggested that each gene codes for just one protein, an assumption that remained the conventional wisdom for decades. Now it's known to be oversimplified. One gene can yield multiple proteins or even be transcribed into RNA rather than a protein.

The protein-coding regions of human genes, called exons, take up only 2% of the DNA and can get lost in the other 98%. Genetic oddballs complicate gene counts as well. Some very simple genes consist of just one exon. They are so small that they are easily overlooked by both human and computer gene counters. In contrast, genes that no longer function because of some aberration in their DNA—so-called pseudogenes—artificially inflate gene numbers.

Among the 24,500 genes in the current assessment, "3000 could be pseudogenes," points out Ewan Birney, one of the chief gene counters at the European Bioinformatics Institute in Cambridge, U.K. And he's not the only one who is stuck trying to decide which genes are real. "I believe all gene-prediction programs suffer from this," says Michael Brent, a computer scientist at Washington University in St. Louis. "Everyone will do better [at their predictions] once we get the pseudogenes taken care of."

Even worse, parts of the genome have proven completely impenetrable to the best gene-prediction programs. They include dark matter, regions named because they "are apparently devoid of genes," says Roderic Guigó, a computational biologist at Pompeu Fabra University in Barcelona, Spain. Gene-prediction pros know nothing about this dark matter. Many worry that this void contains genes that researchers just can't see; dark matter genes "may have characteristics other than the ones we recognize," Guigó points out.

One gene ... two genes ...

Any gene-prediction program worth its computer time must do a decent job of finding genes outside the dark matter zone. Typically they do this in one of two ways. The "ab initio" approach recognizes genes by detecting distinctive patterns in DNA sequences, such as those that characterize the beginnings and ends of genes. The other approach is comparative: It uncovers new genes based on their similarity to known proteins and genes. The two create a Goldilocks dilemma. Ab initio programs classify anything that looks vaguely like a gene as a gene, so their totals are too high. Comparative approaches don't recognize unfamiliar genes, so their estimates are too low. And nothing seems to estimate gene numbers just right. "But the programs perform a lot better than they used to," says Gary Stormo, a computer scientist at Washington University in St. Louis.

Many people trace the field's beginnings to an ab initio program called Gene Modeler,

SOURCE: EWAN BIRNEY/SANGER INSTITUTE/EBI

produced in 1990 by Chris Fields and Cari Soderlund, who were then at New Mexico State University in Las Cruces, to find genes in the nematode *Caenorhabditis elegans*. Other software in existence at the time was much clunkier and took less direct approaches. For example, BLAST and FASTA translated DNA sequence into protein sequence that could be compared to existing protein data.

The field grew quickly. Other early predictors included Guigó, who adopted Gene Modeler's approach to build GeneID for finding human genes instead of worm ones. In 1991, one of Stormo's graduate students, Eric Snyder, wrote software called GeneParser that incorporated a technique called dynamic programming to separate exons from introns, gene regions that don't code for proteins. It worked more efficiently than other approaches by allowing the computer to consider just subsets of the data as it evaluated sequences.

Snyder, now at Pennington Biomedical Research Center in Baton Rouge, Louisiana, let the project lapse. "If I were to do it over, I would have kept working on GeneParser and gene prediction," he says. But at the time, granting agencies didn't think the problem of counting genes was particularly important, and he was not alone in leaving his software behind for other projects.

A few have been lucky enough to be in the right place for keeping up their gene prediction work. Steven Salzberg and his crew at The Institute for Genomic Research (TIGR) in Rockville, Maryland, have been improving their programs for finding human genes since their first one, an ab initio approach, came out in 1994. They have recently come up with several new programs, one of which incorporates more background information to generate predictions, such as clearer rules about sizes of exons and introns. Another program works with two whole genomes at once, computationally laying one on top of the other for comparison.

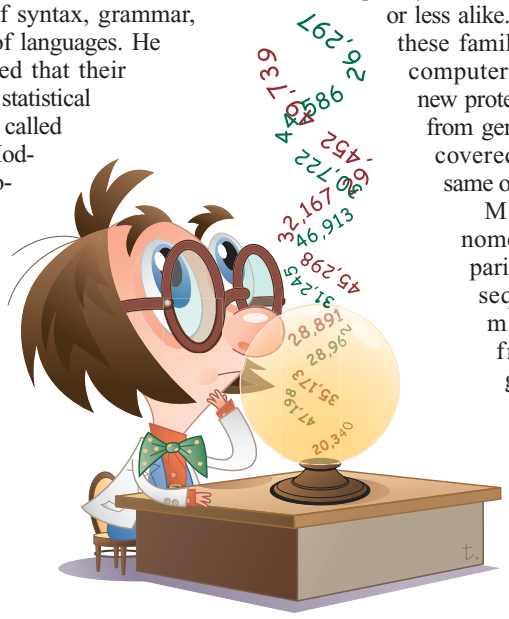
Many researchers are taking this latter approach because similar species tend to have genes with very similar sequences. Protein-coding regions are likely to match and thus stick out among the unmatched nonsense sequence surrounding them. Not only genes match, says Eric Green, a genomicist at the National Human Genome Research Institute in Bethesda, Maryland. He and his colleagues compared DNA from 13 species, including the dog, cow, chicken, and puffer fish. In addition to genes, regulatory regions match, they report in the 14 August issue of *Nature*. And those regions, too, can confound gene counts.

One program, GENSCAN, stands out among the others as having set the standard for the field. When Chris Burge, now at the Massachusetts Institute of Technology in Cambridge, began writing the program in 1996, many of his colleagues were advocating a comparative approach. They picked out genes in a newly sequenced genome by matching its DNA against known genes in existing databases. But Burge disagreed. "We had human sequence, but there was really nothing to compare it to," he recalls. No other vertebrate genomes were very far along, and the matches in sequenced genomes of the fruit fly, nematode, and microbes were fairly limited.

Gene Counts

Program	Prediction
Ensembl/Genewise	24,500
Twinscan	25,600
GeneID	32,400
GENSCAN	45,000

Instead, Burge took a lesson from David Haussler, a computer scientist at the University of California, Santa Cruz. Three years earlier, Haussler had realized that the gene-prediction problem was similar to the challenge faced by linguists who were trying to pick out patterns of syntax, grammar, and other features of languages. He and others suggested that their colleagues borrow a statistical tool from linguistics called a Hidden Markov Model. It calls for the program to make predictions based on a set of benchmarks it acquires from existing information.



"There are a whole bunch of patterns and rules that distinguish parts of genes," Burge points out. For example, all —or at least almost all—genes begin and end with a particular sequence. The ends of exons also have a characteristic sequence that tells enzymes to slice out the intron that follows. Burge "taught" the model these rules by having it analyze the sequences of several hundred genes with known intron and exon positions. The patterns it learned became the grist of a Hidden Markov Model for predicting whether a stretch of DNA includes a gene.

The approach proved a great success. Today, Hidden Markov Models are standard in most gene-prediction algorithms. As for GENSCAN itself, "it was significantly better than what was out there," says Salzberg. Adds Brent, "It was just the best."

But even the best have their flaws. GENSCAN's is overenthusiastic gene identification: It predicts 45,000 genes for the human genome, almost double the currently accepted total. Burge admits that GENSCAN has this problem but thinks that too many genes are better than too few; one can always eliminate the false positives.

GENSCAN will probably never predict the correct number of genes, Burge says. And much has changed in the genome world since the program was introduced. Taking account of new sequence data from humans and other species may be key to getting the gene tally just right. "If I were working on gene finding today, then the comparative approach would be the way to go," he says.

Several programs, such as the official GeneSweep counter, Ensembl/Genewise, pick out genes based on their resemblance to what's already known. But they are much more sophisticated than earlier comparative efforts. Genewise, developed by Birney and his colleagues, works backward from known proteins. Proteins come in families whose members' amino acid sequences, and consequently DNA sequences, look more or less alike. Taking advantage of these family resemblances, the computer program compares new protein sequences derived from genes to previously discovered proteins from the same or different organisms. Matching entire genomes rather than comparing short stretches of sequence is becoming more feasible and fruitful as more genomes are sequenced. Together, these comparative approaches are the most promising route right now, Affymetrix's Kulp says. And several programmers are melding multiple gene-prediction strategies.

Despite these advances, "we have probably reached a plateau," Guigó complains. Few of the next-generation programs come up with similar gene totals. And dark matter still looms as a big unknown, one that no current program can touch. To decipher it, what's needed is knowledge about why genes have the characteristics they do and what dark matter genes might look like. In short, Guigó says, "to get better we will need to better understand the biology." Nobody is betting on when that will happen.

–ELIZABETH PENNISI

ILLUSTRATION: TERRY F. SMITH