New Sequencers to Take on the Genome

A major push is under way in DNA sequencing technologies. New instruments can sequence thousands of bases a day; DOE wants to boost the rate to thousands of bases per second

In the early 1970s it took more than a year for a skilled biologist to work out the nucleotide sequence of a single gene. Sequencing, and much of molecular biology, was given a major boost by the development of rapid sequencing techniques in the late 1970s, which raised the rate to 15,000 finished bases per year. In 1986 Leroy Hood, Lloyd Smith, and their colleagues at the California Institute of Technology developed the first automated DNA sequencer, which has the theoretical capability of sequencing 10,000 to 15,000 bases a day.

Such advances make sequencing the entire human genome—all 3 billion base pairs—technically feasible, if not necessarily politically and economically desirable. The Department of Energy (DOE) has launched a concerted program to look into doing just that. The National Institutes of Health (NIH), although not exactly leading the charge, is being swept along by the enthusiasm among the sequencing fraternity. Other biologists, however, remain to be convinced of the wisdom or necessity of a focused endeavor. Congress, too, is weighing the pros and cons (*Science,* 31 July, p. 486).

Few question the benefits to medicine and basic knowledge, which are likely to be substantial. Rather, the stumbling block is the massive scale of the project, which might consume one to several billion dollars and take 15 to 20 years to complete. Not surprisingly, the proposal has given considerable impetus to the development of new automated DNA sequencing technologies to shave time and money off the task.

Applied Biosystems Inc. (ABI) of Foster City, California, was the first on the market, earlier this year, with an automated DNA sequencer based on Hood's technology. Now several other companies are scrambling to stake out their share of what promises to be a lucrative market. In an article in this issue of *Science* (p. 336) James M. Prober and his colleagues at E. I. DuPont de Nemours & Co. describe their new automated DNA sequencer, expected on the market in early 1988. EG&G Biomolecular of Watertown, Massachusetts, is also venturing into DNA sequencing, but with a smaller instrument based on conventional radioisotopic technology. Since 1981 the Japanese government has been spending $1 million a year on the development of automated DNA sequencing technologies. Its goal is to develop equipment capable of sequencing 1 million bases a day, for 17 cents a base.

DOE is exploring admittedly "far-out" technologies that may allow sequencing thousands of bases a second.

None of the current approaches looks fast enough to DOE, which is exploring admittedly "far-out" technologies that may allow sequencing thousands of bases a second, for less than a penny a base. There is considerable skepticism about whether that rate can actually be attained anytime soon. At this stage, most efforts are focusing on working out the bugs of the first generation of automated DNA sequencers, which Hood has likened to a Model T.

Sequencing by conventional techniques, whether the Sanger method or the Maxam-Gilbert method, involves generating a series of DNA fragments that vary in length by one nucleotide base—that is, they start at the same point and end at different bases, an A, G, T, or C. In the Sanger method, on which the new automated machines are based, an oligonucleotide primer is used to stimulate synthesis of a chain of DNA. Included in the normal reaction mixture is a dideoxy form of one of each of the bases. When the dideoxy base, say a dideoxy adenine, is incorporated into the growing DNA chain instead of the usual deoxy form, the chain stops elongating.

Four of these reactions are performed, each with a dideoxy form of a different base. Each reaction produces a series of different length fragments ending in the same base, say an A, and tagged with a radioisotope. The fragments are then forced through an electrophoretic gel, which separates the fragments by length—the shorter fragments run faster; the longer, slower. A film image is then taken, and from the relative positions of the fragments on the gel, the sequence can be inferred.

Hood and his Caltech colleagues modified that procedure in several ways. Instead of using a single radioisotope to label the fragments, their machine uses four fluorescent dyes, a different color for each base. Because they can be distinguished by their color, the fragments can be combined in one lane for electrophoresis, rather than run in separate lanes, as is the case with manual sequencing. An argon ion laser is used to excite the dyes, and when the fluorescent fragments near the bottom of a gel, a detector identifies the emission fluorescence and a computer reads the sequence in real time. The ABI machine has 16 lanes for electrophoresis and can thus accommodate 16 different DNA samples per run.

The DuPont team has taken the same overall approach—the Sanger technique, using four fluorescent dyes and a laser detector—but has modified it to simplify the chemistry. Their approach differs primarily in how the fluorescent labels are attached. Whereas ABI uses a fluorescent primer to label the DNA fragments, DuPont instead labels the four dideoxy nucleotides that terminate the chain. This means that the four sets of fragments can be generated in one reaction rather than four. But the chief advantage, DuPont scientists say, is that fluorescent dideoxy nucleotides can be used with all sequencing strategies and vectors. Using the ABI instrument, the sequencing strategy is currently constrained by the need to use special primers.

Another advantage, according to Mark L. Pearson, director of molecular biology at DuPont, is that this approach avoids the problem of polymerase-pausing artifacts, known as "false stops." With the primer-labeled approach, whether conventional or automated, the DNA chain sometimes stops elongating, for reasons largely unknown, before the terminating dideoxy nucleotide has been incorporated. As a result, false bands appear on the gel that must be sorted out. With the DuPont approach, Pearson says, while false stops may still occur, they are not detected because the products do not contain the fluorescent dye.

One of the problems Hood and his colleagues encountered is that attaching the fluorescent dyes affects the speed at which the DNA fragments move through the gel, and does so differently for each dye. The ABI instrument compensates for this vari-

able mobility in the software. To avoid this problem, DuPont selected a family of similar fluorescent dyes. Thus, although these dyes still affect the mobility of the fragments, they do so consistently. In other words, although the labeled fragments move at a different rate than do unlabeled ones, the overall pattern they create is the same.

The similarity of the fluorescent dyes, however, makes them potentially more difficult to discriminate. (DuPont's dyes are four shades of green, whereas ABI's are four different colors.) DuPont addressed this problem by using two detectors, with complementary optical filters, that are extremely sensitive to slight wavelength shifts of the emission bands.

This is still clearly a shake-out period for these new technologies, and their performance—their speed, error rate, and cost per base—is changing almost weekly. But both Hood and Pearson of DuPont predict roughly comparable performance for the two instruments. Working flat out, both instruments can theoretically churn out up to 10,000 or 15,000 raw bases of sequence data a day. This assumes multiple runs, with fully loaded machines—an unlikely scenario in a typical lab. And this estimate is for *raw* sequence data—each base determined once. Before these data could be published, the strand would have to be resequenced several times to ensure accuracy.

Both instruments can sequence about 300 bases per lane per run. Thus, with 16 lanes, the ABI machine can determine about 4800 bases in an 12-hour run. For the DuPont instrument, which has 12 lanes, the total is about 3600 per run. The DuPont sequencer is faster, however, with each run taking 6 hours as opposed to 12.

The limiting factor, in terms of output, is the length of fragments the instruments can handle. With current DNA separation technology, resolution breaks down after about 300 bases. However, the two groups report success in their recent attempts to maintain resolution out to 600 bases, which could double the output. Output can also be boosted by adding additional lanes, which is also being actively investigated. The current error rate is about 1% for both instruments, which puts them in the ballpark of human sequencers, and both groups are striving for a rate of less than 0.1%.

In terms of large-scale sequencing efforts, such as the genome project, the key question is the cost of sequencing per base. Hood and Pearson expect the figure for the two instruments to be about the same, but getting agreement on what it actually is can be tricky, as everyone seems to calculate it differently. For the ABI sequencer, a reasonable figure for the cost per raw base—that is, one base determined once—is 10 cents, and 6 to 8 cents under optimal conditions, Hood says. This figure assumes starting with a cloned piece of DNA ready for sequencing and covers just the cost of materials and labor to run the instrument. If overhead were factored in, the cost would roughly double.

DuPont, on the other hand, includes overhead in its estimates, and thus cites a figure of 20 cents a raw base, including materials and labor—about the same as the ABI instrument. But, Pearson cautions, this estimate is for predicted performance, under optimal conditions. At this stage, he says, a more realistic figure is 30 cents. The equivalent figure for conventional sequencing, with materials, labor, and overhead, is about $1 a base.

Calculated this way, the cost for a *finished* base, assuming three runs to check accuracy, ranges from about 60 cents with the automated technology to $3 manually. Costs of DNA preparation would then have to be added on. All of these estimates are rather hazy, however, since the figures used for overhead and salaries vary by as much as 50%. Another caveat is that the performance of these machines, and thus the cost, varies significantly according to the skill of the operator. As Pearson noted at a recent meeting, "I think we have to be very careful in feeding numbers to other people who are not familiar with the technology so that they appreciate the real cost and not what it takes your very best graduate student in his final year who is really cranking it through."

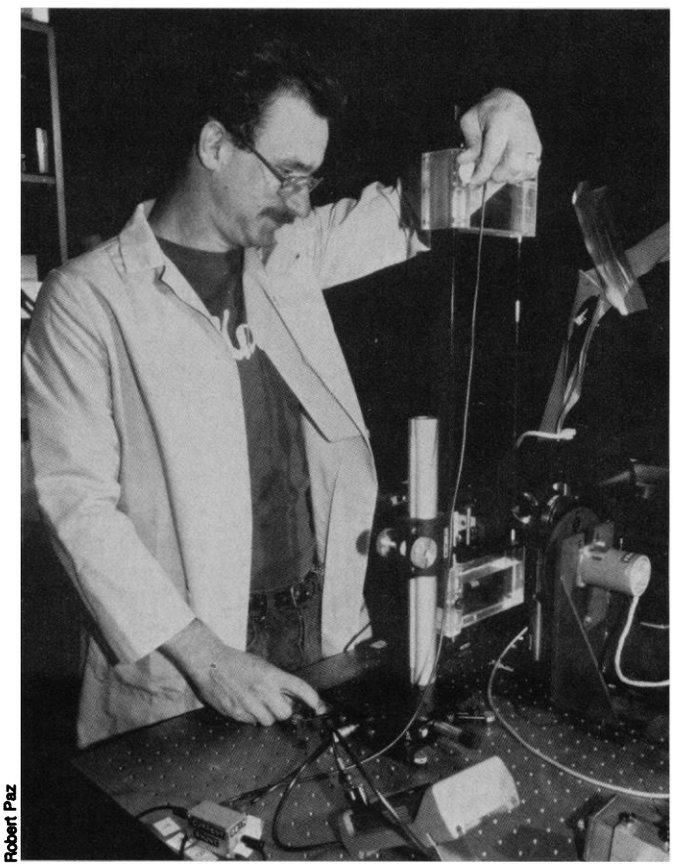
However the current cost per base is figured, most of those involved agree that it will have to drop substantially before it makes sense to begin sequencing the entire genome, or even a major chunk of it. Hood anticipates "a series of incremental changes that will improve the performance by a factor of 10 in the next few years, probably at reduced expense." At ABI, for instance, they are experimenting with stronger dyes, different gels, and new filters, as well as modifying the software so that it will tolerate a broader range of conditions, thus increasing speed and reducing the error rate.

"I'm quite confident that we'll have instruments that work effectively in 1 to 2 years," says Hood. "Now our machine works effectively in experienced labs, with experienced sequencers. But it is complicated technology."

Perhaps the biggest gain will come from automating the "front end," as the laborious task of cloning, mapping, and otherwise preparing the DNA for sequencing is called. "We've moved the burden of effort from running gels and reading sequence to the front end, preparing subclones," Pearson

Lloyd M. Smith

Works on the Caltech DNA sequencer.



Robert Paz

says. "The costs of cloning will make the base pair costs look like chicken feed." Hood, whose group is one of several now tackling the "front end," expects to see automated cloning devices within a year or so. The eventual goal is to automate all the sequencing steps and tie them together.

At DOE, however, "we are not interested in small incremental improvements in existing technology, but in new methods that offer great possibilities," according to Gerald Goldstein, who is overseeing the technology development for the genome project. "We must have vastly improved technologies," he says. And vast improvement, according to Charles DeLisi, who headed the genome effort at DOE before moving to Mount Sinai School of Medicine this fall, means thousands of bases per second.

For fiscal year 1988, DOE has set aside several million dollars for such technology development. Proposals are due 2 November, but some work is already under way. For the past few years a group at Los Alamos National Laboratory has been trying to apply flow cytometry to DNA sequencing. The idea is to tag the bases with a fluorescent label, then cleave the bases off one at a time and flow them by a detector. Currently, flow cytometers detect single cells, but the Los Alamos group is working on an "extra-sensitive" optical system capable of detecting a single molecule, Goldstein says.

DOE is also funding work at Brookhaven National Laboratory on scanning transmission electron microscopy. So far the effort has been focused mainly on mapping techniques, but it could conceivably be used for sequencing, Goldstein says. This might involve labeling the bases with clusters of gold or tungsten atoms. "It's not exactly a crazy idea, but it is a long way from being proved."

Other possibilities include mass spectrometry—DOE is evaluating several proposals—and scanning tunneling microscopy, which creates a sort of contour map with atomic resolution. Both are "highly speculative," Goldstein admits. DOE may fund a feasibility study of another approach that involves immobilizing DNA in a solid matrix and then knocking off the bases one by one, perhaps with an ion beam, and then detecting them in some as yet unidentified way.

Few expect these innovative approaches to figure in the genome project any time soon. Hood, for one, does not anticipate any "earth-shattering new approaches that will change things fundamentally. Maybe in 5 years someone will have a new idea and there will be a big jump." But for now, most of the gains will come from tinkering with the current generation of DNA sequencers. ■ LESLIE ROBERTS

My Close Cousin the Chimpanzee

Recent evidence of molecular biology indicates that humans and chimpanzees are each others' closest relative, a conclusion that remains at odds with most anatomical inferences

"If Morris Goodman is correct in his conclusion, we will just have to go back to the anatomical evidence and find out what we've been missing," says Lawrence Martin, an anthropologist at the State University of New York at Stony Brook. The conclusion to which Martin refers is that, contrary to most expectations, humans are genetically closer to chimpanzees than either is to the gorilla. On the basis of both superficial physical similarity and more formal anatomical analysis, chimpanzees and gorillas certainly appear to be each others' closest relative. "It would be remarkable if this proved not to be the case," says Martin.

And yet, if a score had been kept during the past few years of the various lines of molecular evidence that have emerged on the human-chimp-gorilla relationship, the unexpected would be seen to be gaining majority support, by more than two to one. The latest offering, by Goodman and his colleagues at Wayne State University and the University of Florida, is published on page 369 of this issue, and is described by Martin as "by far the best molecular dataset to date." Goodman's data, which he collected with Michael Miyamoto and Jerry Slightom, are in the form of a 7100–base pair sequence of a locus in the beta-globin region in humans, chimpanzees, gorillas, and the orangutan.

"If we had only our dataset, the question of a human-chimpanzee association wouldn't be decisive," acknowledges Goodman. "And maybe putting all the datasets together still would leave room for some doubt. But I think it is getting pretty close to being decisive." Backing up this conclusion are two additional papers about to be published.

The first is by Nobuyo Maeda and her colleagues at the University of Wisconsin and presents a further 3000–base pair sequence from the same genetic region that Goodman analyzed. The second is a new batch of DNA-DNA hybridization data by Charles Sibley of San Francisco State University and Jon Ahlquist of Ohio University. Although there had been earlier indications scattered in the literature, it was Sibley and Ahlquist's publication 3 years ago of their first set of DNA-DNA hybridization data that forced molecular biologists and anthropologists alike to take seriously the possibility that the chimpanzee's closest genetic relative might be *Homo sapiens*, not the gorilla.

Impressive though the recent accumulation of genetic results in favor of the human-chimp association is, resolution of the issue will not be settled by majority vote, not least because the data are not equivalent. In addition to the basic divide between anatomical and molecular information, there are different types of genetic data: some are more direct than others.

For instance, DNA sequence data offer direct information about the species being compared, and the sequences themselves can be thought of as being analagous to series of anatomical characters, such as the shape of a bone or the pattern of muscle attachment. By contrast, DNA-DNA hybridization data—which match the overall fit of two separate genomes—are an indirect reflection of two species' relatedness, and simply offer a measure of the genetic distance between them: the poorer the fit, the greater the distance.

In general, biologists with an interest in reconstructing phylogenies—or family trees—prefer to use characters rather than distance data because, in principle, characters allow unique links between species to be identified. For this reason Goodman's latest DNA sequence results are seen as being of special importance in resolving what has clearly become a hot issue in anthropology.

The issue is hot for several reasons. First, if Goodman and others are correct, ideas about the beginnings of the human lineage would be significantly altered. Specifically, because both chimpanzees and gorillas move about by means of a mode of locomotion known as knuckle-walking, it becomes more likely than not that the common ancestor of these two African apes and humans was also a knuckle-walker. Sherwood Washburn, of the University of California, Berkeley, has advocated just this scenario for many years, but with virtually no support from any