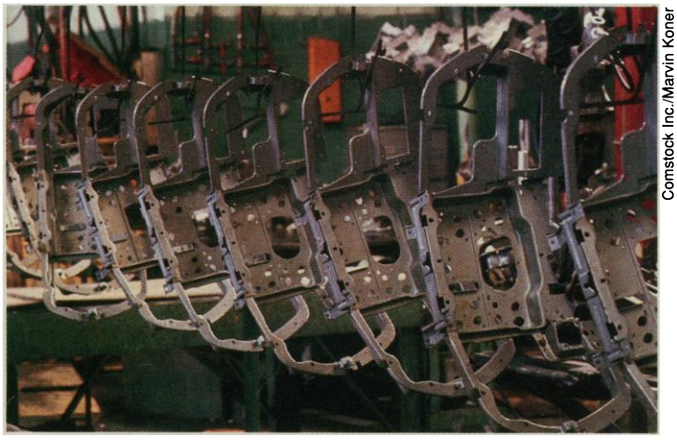
Large-Scale Sequencing Trials Begin

As genome sequencing gets under way, investigators are grappling not just with new techniques but also with questions about what is acceptable accuracy and when data should be released

FOUR GROUPS ARE EMBARKING ON PROJECTS that could make or break the human genome project. They are setting out to sequence the longest stretches of DNA ever tackled—several million bases each—and to do it faster and cheaper than anyone has before. If these groups can't pull it off, then prospects for knocking off the entire human genome, all 3 billion bases, in 15 years and for $3 billion will look increasingly unlikely.

Harvard's Walter Gilbert, who shared the Nobel prize for his pioneering work on DNA sequencing in the mid-1970s, is first tackling the genome of a tiny organism called *Mycoplasma capricolum*. At Stanford, David Botstein and Ron Davis are tackling the budding yeast *Saccharomyces cerevisiae*. In a collaborative effort, Robert Waterston at Washington University and John Sulston at the Medical Research Council lab in Cambridge, England, have already started on the nematode *Caenorhabditis elegans*. And in the only longstanding project of the bunch, University of Wisconsin geneticist Fred Blattner is already several hundred kilobases into the *Escherichia coli* genome.

The future of the genome project? *Sequencing is a production job, argues Walter Gilbert, and should be tackled the same way cars are built.*



Comstock Inc./Marvin Koner

Each has vowed to sequence about 3 million bases at 75 cents a base within 3 years, and to drop the cost to 50 cents a base within 5 years—not a trivial task, considering that the going rate is now roughly $2 to $5 a base and the largest genome ever sequenced is that of cytomegalovirus, a mere 240 kilobases. And that took 12 person years to complete. But until these four groups, and several others that are likely to be funded in early 1991, show they can get the cost of sequencing down and the speed up, genome project director James Watson says he won't even contemplate an all-out assault on the human chromosomes, the ultimate goal of the genome project. "We need a factor of 10 improvement," he said at the recent Human Genome II meeting in San Diego where NIH's new pilot projects were described. "We would like a factor of 100, but that is not essential. If we can do it for 50 cents a base, we should."

As the four groups gear up, each experimenting with different approaches, they are finding that the questions are far broader than which techniques to use. How, for example, do you motivate people for what will ultimately be a boring and monotonous task? Do you lure in top notch young scientists by promising them a first crack at the interesting biology you uncover? Or do you turn it into an assembly line, cranking out DNA sequences the way other factories crank out computer chips? And what error rate is achievable—and what is acceptable, bearing in mind that high accuracy means higher costs? Finally, the researchers are asking, when should they release their sequence data, much of which will never be published in a conventional sense?

The sleeper issue, from the outset of the human genome project, has been who will actually do the work once sequencing becomes routine. The bane of sequencing, even proponents of the project agree, is its drudgery—even with automated sequencing machines and DNA preparation robots to take over some of the task. Sydney Brenner, who oversees the genome effort at the Molecular Research Council lab in Cambridge, has joked that sequencing should be meted out as punishment to convicts, say, 12 million bases for a typical crime, and more for a really atrocious one. Barring an available penal colony, the four groups are tackling the issue head on, and coming to quite different conclusions.

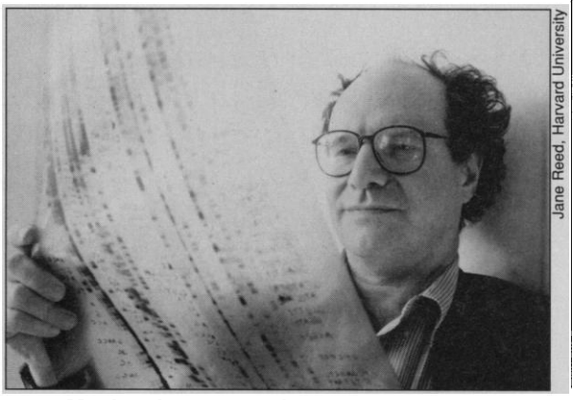
Stanford geneticists David Botstein and Ron Davis have opted for bright young scientists who will work on the project not because they love sequencing but because of what the sequence will eventually enable them to do. As Botstein explains, "The yeast genome is really small, just 12.5 megabases, and it is almost all information." He means that, unlike the human genome, which is interrupted by long stretches of noncoding, or "junk," DNA, the yeast genome is practically solid genes—many of which have close human counterparts. The Stanford workers plan to sequence the genome in a way that will make it easier to elucidate the functions of the yeast genes when they are done.

By setting up the project this way, Botstein and Davis have been able to recruit three "genome fellows," who will be appointed as research assistant professors at Stanford and who will do the bulk of the work. "We have arranged the work to hold their interest and greatly enhance their careers," says Botstein. "I think the fellowships will be seen like a stint at Cold Spring Harbor or the Whitehead. The fellows will be working on a project they think is of great importance and is beyond the means of someone starting their own lab."

Others are not so sure. "Botstein is following the wrong line," asserts Gilbert, who believes that if large-scale sequencing is to work, it must be treated not as science but as a production job. "A sequencing group needs to be interested in the problem of producing a large amount of sequence, which is a pure technological problem quite apart from interpreting the sequence. The idea is to attack the problem like building an automobile," the same way computer chips or biotech drugs are manufactured, he says. "And that work is ultimately done by production workers interested in doing that job very well. It is not done by research scientists."

In Gilbert's shop, which will have a staff of 15, research scientists will invent the methods, troubleshoot, and interpret the sequence as it is completed. But the sequencing itself will be done by six "production technicians." Once they knock off the mycoplasma, he plans to move on to another, more typical bacterium or perhaps to a yeast chromosome. With a streamlined production facility and an innovative new sequencing strategy, Gilbert predicts that he will sequence faster, more accurately, and cheaper—at 35 cents a finished, or checked base—than his colleagues.

Because of the sheer size of the nematode genome—about 100 million bases—Robert Waterston and John Sulston also think a production approach will be essential. "While Botstein's genome fellowships are appealing, I don't think they'll be practical" on this scale, says Waterston. During the first 3 years, however, while the worm groups are devising new techniques, it won't be the smooth assembly line that Gilbert has in mind either, Waterston concedes. Rather, "it will be somewhere between a production job and fiddling."



Motivation question. *Walter Gilbert believes sequencing should be done by production technicians who take pride in their craft.*

Fred Blattner, meanwhile, has chosen a middle ground. In his *E. coli* project he is using undergraduates for the steps he can't automate—an approach closer to Gilbert's than to that of the Stanford team, but still off the mark in Gilbert's view. "Blattner's students are research scientist types who will ultimately get bored."

Although emotions run high on how best to organize the project, it is accuracy that is emerging as the most divisive issue in sequencing. At first blush, there would seem to be little to argue about. Indeed, a few years ago, accuracy goals were barely discussed; the assumption was that you got it right. But now the sequencing community is deeply divided over what degree of accuracy can be achieved, what level is necessary—and, most important, what the costs will be.

Gilbert threw down the challenge at the recent genome meetings in San Diego and Hilton Head, South Carolina, when he said that he was shooting for "zero defects." He admitted, however, that he would probably be hard pressed to get beyond an error rate of 1 base in every 100,000—a tall order nonetheless, since the rate for the sequences currently in databases, while somewhat murky, seems to be several errors per thousand bases.

Gilbert's goal was instantly dismissed as unnecessary and probably unattainable by pragmatists such as Blattner, Botstein, Tom Caskey of Baylor College of Medicine, and Craig Venter of NIH, who all argue that pushing for perfection will drive the cost of sequencing up too far. Achieving ultra high accuracy, they point out, would involve sequencing the same stretch many times, not just the standard three or four. And besides, they say, slightly sloppy sequence may be fine for most of the uses people envision. The pragmatists think an error rate of about 1 in 1000, or perhaps 1 in 4000, would be just fine.

"I think the biologic community would find sequence with [1 error in 1000] tremendously useful," asserts Caskey. Indeed, Botstein and David States of the National Center for Biotechnology Information recently reported on a simulation they did on yeast sequences, in which they found that for certain purposes, like searching for similarities to other proteins, they could tolerate a surprisingly large amount of error, between 1% and 5%. Venter has obtained similar results in his early sequencing of the complementary DNAs expressed in the human brain.

Scientific lure. *For his group's yeast project, David Botstein favors scientist-sequencers who will be motivated by insights they will uncover.*

But Gilbert is not persuaded and dismisses the "illusion" that accuracy is expensive. "The Japanese auto industry has shown the fallacy of that argument," he says. "It is cheaper in the long run to set up a production job and strive for zero defects than to say we can't possibly do that and then put up with errors." He agrees with his Harvard colleague George Church, who points out that quick and dirty sequencing has a hidden cost: "If you produce inaccurate sequence, someone else has to redo it."

Taking his cues from the Japanese, Gilbert is setting up a separate team, distinct from the sequencing technicians, to monitor the quality of his operations and to devise new tests for accuracy. Gilbert readily admits that, "I am standing on a limb and saying I can do it this accurately for this cost. And if you want to do it better, I applaud you. If you strive for something lower, then shame on you."

What's missing from all this debate, which takes on religious overtones, are some solid estimates of what different degrees of accuracy will cost, says Waterston. At this stage, he says, no one really knows how much it will cost per base to sequence at the 95% accuracy rate versus 99%, or 99% versus 99.9%, much less 99.999%. Having comparative figures in hand, he says, would go a long way toward making this "less of a religious discussion and more of a practical one." In addition, says Waterston, data are sorely needed on just what the current accuracy rate is—a topic of considerable debate—and how it can be measured in the first place. "Many people haven't spent a lot of time thinking about how to quantify accuracy, other than to do more sequencing," he says. "But if there are systematic errors in sequencing, they will never be discovered."

On one point, however, nearly everyone agrees: There ought to be a way for investigators to attach confidence statements, or disclaimers, to the sequence they submit to a database. "As Botstein has shown, 95% [accuracy in] sequence can be very useful," says Waterston. But if it is going to be entered in a database, then it has to be flagged somehow to say, "This is rough, use it at your own peril."

And finally, there is the contentious matter of when sequence data should be deposited in a database. Much of the argument ostensibly concerns accuracy, says Waterston, but that is really a surrogate for the question of how long someone gets to hang on to the data and study it in private before sharing it with his peers.

"That gets complicated," Watson said in San Diego. "The sequence has to be correct before it is released. But we don't believe that someone should hang on to it for 10 or

20 years with their lawyers patenting everything of interest. There will have to be a compromise. We are all aware of the problem, but we have to get big sequencing under way before we get the answer."

The genome center has had a committee looking into the data release issue for about a year now and expects to come out with a policy soon. The bottom line, says Joseph Sambrook of the University of Texas Southwestern Medical Center in Dallas, who chairs the committee, is that once a reasonable chunk of sequence is finished and checked, it should be released as quickly as possible, within some outside deadline. Sentiment is strong, says Sambrook, for making that a condition of funding. But so far, committee members haven't yet decided what that outside date should be. Meanwhile, in the absence of guidelines—and after much wrangling and some peer pressure—a consensus is starting to emerge that sequence data should be deposited within 3 to 6 months of completion.

Gilbert believes the problem will essentially disappear over the next few years anyway. At the moment, the handful of researchers doing large-scale sequencing are passionately interested in the information the DNA encodes. But eventually, he predicts, "There will be a split between those who actually get the sequence and those who analyze it." In the interim, the community will come to some truce that will give investigators one first—and probably brief—look at their long-sought data. ■ LESLIE ROBERTS

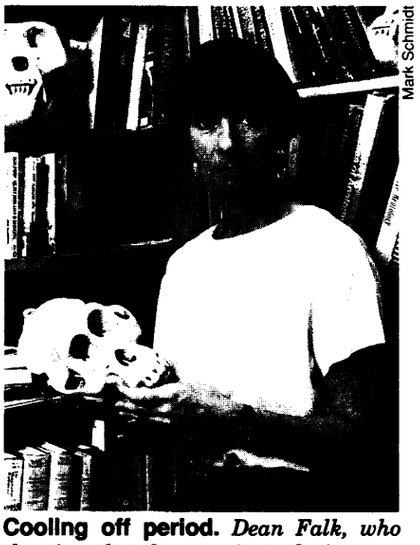
Did Cooler Heads Prevail?

A tenacious paleoanthropologist has set the pot boiling with her theory that hominid brains needed a special cooling system before they could expand to their present size

ALTHOUGH SCIENTIFIC INSPIRATION IS known to come from diverse sources, few scientific theories have been inspired by auto mechanics. Yet that was just the source of anthropologist Dean Falk's recent controversial notion of brain evolution. As her mechanic was explaining her engine's cooling system, Falk realized that a car engine and the human brain might have something crucial in common: a radiator. Falk theorized that the brain's size, like the engine's, is restricted by the capacity of the cooling system to keep it running within a certain temperature range. That insight led Falk to propose a "radiator theory" of brain evolution, an idea that has been generating a lot of heat lately among anthropologists and physiologists.

In a report published last summer in *Behavioral and Brain Science*, Falk proposed that the development of the "radiator"—in the form of a vast network of hair-thin veins that drain heat-carrying blood—was a crucial preadaptation that enabled the human brain to swell from slightly larger than the size of a chimp's to its current weight in just 2 million years. Falk also claims she can trace the evolution of this radiator from its origins in some species of early hominids to its complex form in modern humans. Indeed, she thinks certain hominid species may have hit evolutionary deadends largely because they lacked the brain radiator.

Few anthropologists or physiologists have been able to keep a cool head over Falk's theory since she published it last summer. Her backers express wild enthusiasm, claiming the radiator theory offers an entirely new perspective on brain evolution. "Falk's evidence is convincing. Her paper will be a classic in paleoanthropology," says Harry Jerison, a professor at the University of California at Los Angeles Medical School and author of a classic text on brain evolution.



Mark Schmidt

Cooling off period. *Dean Falk, who theorizes that the capacity to drain away heat was crucial to brain evolution.*

The critics are equally extreme. They argue that Falk hasn't mastered brain physiology and therefore doesn't appreciate that the network of veins she describes has no significant role in brain cooling. "I think she's dead wrong," says Ralph Holloway, an anthropologist at Columbia University and a well-known specialist in brain evolution. "The radiator theory has too many leaks to be taken seriously."

Falk, a respected anthropologist who is known for tenacity, isn't fazed by the roller coaster of praise and criticism, much of which emerged in 26 commentaries by other scientists that were published along with her paper. "Paleoanthropologists are a contentious lot," she says. "It was what I expected."

One reason for all the overheating is that Falk has jumped into an area that has a long history—but few hard results. Many attempts have been made to explain how and why the human brain expanded so rapidly, with most theories attributing the explosive growth to changes in behavior: walking erect, hunting, using language and tools. But those theories are speculative, because there is no way to prove that coincident changes in behavior caused brain expansion. Which is one reason paleoanthropologists have long looked for direct evidence of anatomical changes in the brain over time.

Falk begins her case with physiology. A growing brain would need a radiator, she argues, because its tissues would be exquisitely sensitive to heat. A change of only 4 degrees Celsius, for example, begins to disturb the functions of the contemporary human brain; in children, high fevers can cause convulsions—a sign that the orderly coordination of neurons has been disrupted. But, in general, the body is cooled by the skin and the sweat glands. Does the brain have a special vascular radiator? Falk argues that it does by citing recent studies of brains of human corpses by Michel Cabanac of Laval University in Quebec.

Cabanac's work shows that the network of so-called emissary veins is spread throughout the skull. He thinks this system cools the brain during hyperthermia because its veins bring blood from the brain to the face and the surface of the skull where it is cooled by evaporation of sweat before being returned to the deeper recesses of the cranium.

Falk draws on Cabanac to postulate the existence of the "radiator" as a significant brain-cooling mechanism. The evidence hasn't convinced many physiologists, how-