Proposal to Sequence the Human Genome Stirs Debate

Some molecular biologists fear that the proposal to sequence the entire human genome is developing unstoppable momentum, despite its uncertain merits

DURING the past 12 months there have been half a dozen separately organized small gatherings scattered across the country, each one discussing the prospect of obtaining a complete nucleotide sequence of the human genome. Such a project would be gargantuan in scale, even when measured against the Big Science ventures of high energy physics, but particularly so against the really rather modest standards of biological science. With current technology, sequencing the 3 billion nucleotides in the human genome could consume 30,000 person-years of effort and upward of $2 billion.

For comparison, the proposed Superconducting Super Collider is billed at $3 billion and the space station project $8 billion.

Proponents of the genome sequencing venture foresee immense potential benefits, both in basic biological research and in health care. "The total human sequence is the grail of human genetics," said Walter Gilbert at a 1-day workshop organized by the Department of Energy (DOE) in Santa Fe earlier this year. "It would be an incomparable tool for the investigation of every aspect of human function."

Given the magnitude of the project—both in practical terms and in the flood of data that would be generated—it is perhaps surprising that the various deliberations of the past year have hardly surfaced in the molecular biology community as a whole. But the first that many practitioners heard about it, including some of the most prominent names in the field, was at this summer's Cold Spring Harbor Symposium.*

The symposium was entitled "Molecular Biology of *Homo sapiens*," which has a certain grandeur in itself. In effect, the symposium was a celebration of the impact of molecular biology on the understanding of the human condition, including genetic disease, cancer, and evolution. Appropriate, then, that this grandeur should potentially be enhanced to its ultimate level: to wit, the prospect of knowing everything there is to know about the human genetic blueprint.

*"Molecular Biology of *Homo sapiens*," Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 28 May to 4 June.

The response, however, was extremely mixed. In contrast with the DOE's Santa Fe workshop, which one participant described as being "distinguished by a rare and impassioned esprit," the Cold Spring Harbor gathering revealed a significant level of doubt as to the wisdom of the venture. For instance, although Walter Bodmer, of the Imperial Cancer Research Fund, London, described the prospect as "the most exciting human endeavor," Maxine Singer, of the National Cancer Institute, Bethesda, argued that there were more productive ways of finding out the things we need to know about the human genome. "An approach that included mapping, genetics, and biochemistry makes a lot more sense."

> ***"The total human sequence is the grail of human genetics."***
> **—Walter Gilbert**

An overriding concern, and one that was shared by all sides, was that a project of this magnitude might divert funds from existing biological research. David Botstein, of the Massachusetts Institute of Technology (MIT), reminded participants that they were amateur politicians at best. Against the professionals in Washington, the chances were good that, in such a high stakes game as this would be, existing research funding would suffer. "It endangers all of us, especially the young researchers," he said.

It quickly emerged during discussions that the idea of pushing biology into the Big Science league has built up a substantial momentum in a relatively short period of time. Not yet a fait accompli, the proposal nevertheless appears to have a sense of inexorability about it.

With current methods a single person can sequence 100 kilobases (kb) a year, at a cost of about $1 a base. These numbers are certain to change, however, as new technology is developed: witness the automatic DNA sequencer just reported by Leroy Hood and co-workers at the California Institute of Technology, which speeds sequencing tenfold and cuts the cost in half. Nevertheless, even with several orders of magnitude improvement in techniques, the task of sequencing the entire human genome, or even one chromosome, remains nontrivial. Clearly, the drive to embark on such a task must be motivated by something other than the lure that it is now technically feasible.

That drive comes in part from human genetics, both classical and molecular. The combination of the two has served to generate a genetic map, albeit sketchy in parts, of more than 4000 loci over the 25 human chromosomes (22 autosomes, two sex chromosomes, and one mitochondrial chromosome). These loci include well-characterized genes, such as those of the globin family, and the as yet elusive genes associated with genetic diseases, such as Huntington's disease. Knowledge of where precisely the Huntington's gene is, and what its sequence is, could offer real possibilities of treatment and cure, as it could for the other 1000 or so known genetic diseases.

But there is a huge difference between knowing where a particular loci is on a large-scale genetic map and being able to specify precisely the location of the gene in question and read its sequence in its complete molecular biology context. Hence the interest in knowing the total sequence of the human genome.

One of the first occasions on which the idea of obtaining the entire human genome sequence was discussed seriously was at the molecular genetics Gordon conference last summer. About the same time Robert Sinsheimer, of the University of California at Santa Cruz, called together an informal gathering on the subject. The conclusion in both cases was that, though mammoth, the project was technically feasible. The ball was rolling.

Meanwhile, the DOE, which already has spent upward of $2 billion through the years on energy-related effects on human genetics, began to take the initiative. Charles DeLisi, the new director of the DOE's Office of Health and Environmental Research (OHER), became intrigued with the idea toward the end of last year, and the

result was the Santa Fe gathering of 50 participants, chaired by Frank Ruddle of Yale University.

The DOE had been instrumental in establishing the National Laboratory Gene Library Project, which is based jointly at the Los Alamos and Lawrence Livermore Laboratories. The library contains, chromosome by chromosome, workable fragments of human DNA, which are freely available to researchers. "We see the sequence project as a natural offshoot of the library," says David Smith of OHER. There is a desire, he says, "to utilize DOE laboratory resources in a beneficial way." Sequencing the entire human genome is seen as achieving that goal.

Discussion at DOE's Santa Fe meeting was, therefore, how the sequence could be achieved, not whether it should be done at all. Organization of a project of this scale was recognized as being as important as the technology that would be involved. At one end of the scale Walter Gilbert proposed the establishment of a Human Genome Institute, which in the first instance at least would be devoted entirely to churning out sequences. The strategy, he suggested, could be to go for the most interesting regions first, that is, known genes and regions of known importance.

Although participants felt that some kind of central control was necessary, most considered that the work itself would best be distributed in laboratories throughout the country, and perhaps throughout the world. Sherman Weissman of Yale University said that such an arrangement would engender more creative approaches to the problems that each laboratory would separately face. Improving sequencing technology would obviously be crucial.

The workshop also recognized the importance of effective handling of the vast amount of data that would pour from the project. A Cray-class computer would be required at the heart of the operation, with a great deal of effort also being devoted to developing analytical manipulation of the data. The problems of data-handling currently experienced by the U.S. national DNA database, known as GenBank, provide a salutory lesson for what might be required for close to 1000 times as much information. GenBank has spent 30 cents a base in its storage operations for six million bases, and has been overwhelmed by the task (see box on this page). At an estimated cost of 3 cents a base, expenditure on data-handling for the genome project is projected to be in the order of $100 million over 10 years.

Smith told the Cold Spring Harbor meeting that the "near unamimous enthusiasm" of the Santa Fe workshop had been very well

DNA Databases Are Swamped

In the 4 years since they were established, the major DNA databases, located at the Los Alamos National Laboratory in the United States and the European Molecular Biology Laboratory (EMBL), Heidelberg, have grown in size some 25 times. And the rate of increase in production of DNA sequence information could almost double the size of the databases by June 1987, at which time the initial $3.5 million, 5-year contract for the U.S. facility comes to an end. The National Institutes of Health recognizes that it will have to find significantly increased resources when it awards a new contract for the continuation of the U.S. database because the Los Alamos operation, known as GenBank, has simply been unable to cope with the volume of sequence information coming through the journals.

"The amount of sequence data being produced is far greater than was anticipated," explains Walter Goad of the Los Alamos Laboratory. "In addition, annotation of the sequences and entering them into GenBank is more time-consuming than had been calculated." As a result the facility faces a large backlog of sequences that are yet to get into the database, some of which goes back as far as 2 years. For instance, only 19% of the sequences published in 1985 are yet in GenBank.

"We have launched a crash effort to catch up," says Goad. This effort includes hiring more people to enter the information. The principal tactic, however, is to cut back on annotation of the sequences, which to some extent diminishes the value of the database. Annotation involves the indication of start and stop positions in a gene, intron and exon boundaries, the location of enhancers, and the addition of other relevant biological information. The assembly of these data has so far been done by technically qualified database personnel, who comb through the source paper and other relevant publications. "Since the beginning of this year we have mainly limited ourselves to getting the raw sequence into GenBank and citing the source," explains Goad. In some "hot" areas, such as AIDS research, Goad and his colleagues have made a selective effort to keep up-to-date, but inevitably this has meant that other less glamorous topics have fallen even further behind.

GenBank shares the job of collecting sequence information with the EMBL, and the two databases then pool their information. The EMBL facility, like GenBank, found itself flooded with data, but for a number of reasons has already been able to clear up much of its backlog. Not only did the EMBL database have a 6-month jump on GenBank in beginning the data-collection job, starting in April 1982 as against GenBank's October, but it also initiated its crash catch-up program a year earlier than GenBank. "Unlike GenBank, we did not concentrate on any particular subject area," says Greg Hamm, who is responsible for the EMBL database.

GenBank and EMBL are collaborating closely to find ways of getting sequence information into their databases more rapidly and efficiently. For more than a year GenBank has been writing directly to authors of sequences, asking for annotation details in standard format. The response has been poor, about 30%. Even fewer authors take the opportunity of submitting their information on a floppy disc, which method greatly facilitates input to the database. Goad, Hamm, and their colleagues are currently exploring various schemes with journal editors, by which submission of sequences by authors to GenBank and EMBL would become an integral part of the publication process. "We need to develop a feeling among researchers that the job is not completed until the annotated sequence is in the database," says Hamm. In fact, authors might soon find that having their sequence on the databases will be the only way of making their work public, as journals become reluctant to occupy page upon page of their publications with virtually unreadable sequences.

Dieter Söll, who recently chaired a meeting on the status of GenBank, is already looking to the future and sees the need for what he describes as "a second generation" of databases. "The ever-increasing volume of information is the key to this," he notes. Also important is the basic handling of data, especially the melding of related information, which currently is enormously time-consuming. Lennart Philipson, director of the European Molecular Biology Organization, is expected to initiate discussion of organizational and funding issues surrounding second-generation databases later this year. "These discussions will be at too early a stage to affect the new GenBank contract," says Söll. "But I would expect sufficient flexibility in the new contract to allow for the evolution of old structures into new ones." ■ **R.L.**

received by high level DOE officials. "We were encouraged to proceed," he said. The next steps include further review of the idea during this summer at higher levels within the department, the establishment of a scientific advisory committee that would help steer future decisions, and the funding of a set of research proposals to the tune of "a few million dollars over the next 3 years." Although nothing is approved as yet, this could be in the order of $20 million.

Smith identified three areas in which his department would expect to be supporting research over the near term. First is improvement of sequencing technology, which would include automation and increased speed. Second is work toward the development of a physical map of the genome. And third is in the area of data-handling.

There was a palpable unease among the Cold Spring Harbor audience about the prospect of the DOE being so deeply involved in what essentially is a project in the biological sciences, an unease that James Watson expressed directly. By way of response, Smith reminded the audience of the department's involvement with the gene library project and GenBank. He also pointed out that DOE was well used to handling research projects of this magnitude. "And there's no doubt that sequencing the human genome is more of an organizational challenge than it is a technical challenge," he added. "We are organizationally set up to do Big Science." Gilbert's opinion on DOE involvement was that "we don't want NIH or NSF running this, because if they did there would be a greater likelihood that their funds for other research would be cut back."

Smith also told the meeting that the DOE was conferring with the Howard Hughes Medical Institute, whose already considerable investment in research on genetic diseases makes an interest in whole genome sequencing a natural progression. The institute is to host a discussion meeting on the idea in Bethesda at the end of July.

Paul Berg of Stanford University cochaired the Cold Spring Harbor discussion with Gilbert, and he tried to get participants to put questions of funding aside and to focus only on the potential value of a genome sequencing project. Berg, who describes his position as one of "qualified strong support," asked, "Is it worth the cost, not in terms of dollars but in terms of its impact on the rest of biological science?"

No vote was taken, no resolutions passed, so it was not possible accurately to determine the overall response. Nevertheless, several levels of reservation were clearly expressed.

One level was that voiced by Maxine Singer, mentioned earlier—simply, that more information is gained if sequencing goes hand in hand with other lines of investigation, which might include genetics and biochemistry. "We wouldn't know if it was worth doing as a project until we've done it," she said. "But we do know that when we do sequencing and biochemistry together we really learn a lot." Singer told *Science* that "Of course we are interested in having the sequence, but the important question is the route we take in getting it."

"The idea is gathering momentum. I shiver at the thought."
—David Baltimore

Eric Lander of the Whitehead Institute, Cambridge, concurs with this level of criticism. "The ability to sequence the human genome has just arrived," he says. "We could embark on a space-lab scale of project. But what we've been good at is devising new methods and techniques in small-scale projects. Most of the best developments of techniques in biological science have been adventitious, not goal directed." Lander spoke for many when he said that one way or another the human genome will be sequenced by the year 2000, but a mega-scale project directed toward that goal would change the nature of biological research. "The structures necessary to cope with the expenditure of $2 billion could be inimical for biology. It could create immovable structures."

A second level of criticism concerned the type of tactics most researchers might wish to employ. "In one sense, the sequence is trivial," said Botstein. "What we really want is a physical map of the genome." The existing genetic map of the human genome represents a very low resolution sketch of the whereabouts of certain identified loci on each chromosome. It does not allow a researcher to pluck out manageable sections of DNA of interest. A physical map, which would give a much higher resolution picture, would permit this. Such a map might be constituted of overlapping segments some 40 kb in length.

With a map of this sort it would be a relatively straightforward matter to pin down the location of, say, the cystic fibrosis gene to one particular 40 kb fragment, which could then be sequenced relatively rapidly.

David Baltimore of the Whitehead Institute supports the idea of a physical map as the more profitable goal to aim for. "I don't see the lack of the sequence of the human genome as a limiting factor in anyone's research," he told *Science*. "If the sequence existed, of course you could look it up. But you can easily get the sequence once you've identified the piece of DNA you're interested in."

Sydney Brenner of the Medical Research Council's Laboratory of Molecular Biology, Cambridge, England, has been involved in some of the discussion of the genome project during the past year. Nevertheless, he considers an all-out sequencing effort to be "premature." Like Baltimore, he supports a drive to build up a physical map, and intends to begin work on such a project at the end of this year. "One just has to get started," he says.

Producing a physical map would be a nontrivial exercise in itself. For instance, using the method developed by Brenner and his colleague John Sulston, a map of the complete human genome might occupy 20 scientists for 3 to 5 years. Several approaches to mapping are possible, each of which would cost in the order of $10 million.

It may be that a decision of what scale of project to embark upon, and how it should best be approached, will be forced upon the molecular biological community sooner than most researchers would like. The reason is that the Japanese have apparently already set their sights on a sequencing project. Eiichi Soeda, who was at the Cold Spring Harbor meeting but did not openly discuss his country's plans, is leading a genome sequencing project at the Riken Institute, Tokyo. Some of the best known company names—such as Fuji and Hitachi—are involved in automating a sequencing effort that, within 2 years, will be charting up a million bases a day.

Soeda made it clear that although the sequence is regarded as a highly valued end product, at least as important is the expected improvement of biotechnology that would flow from the effort. Indeed, it was apparent at the Cold Spring Harbor meeting how very important new technologies had been in ushering the molecular biology of *Homo sapiens* into previously unreachable territories. The economic and basic research potential of such advances was abundantly clear to everyone at the meeting. So, might the perceived prospect of Japanese superiority in biotechnology, driven by a human genome sequencing project, be sufficient to spur a decision to go ahead with a similar project in the United States, no matter what the real merits are?

"The idea is gathering momentum," notes Baltimore. "I shiver at the thought." ■

ROGER LEWIN