February 13, 2001, Tuesday

READING THE BOOK OF LIFE; Grad Student Becomes Gene Effort's Unlikely Hero

By NICHOLAS WADE

A surprising hero helped the consortium of academic scientists decoding the human genome to avoid a drubbing by its rival, the Celera Genomics company. Scientists throughout the world are now beating an electronic path to his Web site, where they can analyze and download the human genome sequence. He is a graduate student at the University of California at Santa Cruz, and his name is not Clark but James Kent.

In four hectic weeks last spring, Mr. Kent wrote a computer program that the consortium's leaders hadn't realized how much they needed, one that assembles the 400,000 fragments of DNA they had decoded into a coherent sequence. Using 100 computers that his senior colleague, Dr. David Haussler, had persuaded the university to buy for the purpose, Mr. Kent performed his first assembly on the human genome on June 22, just four days before Dr. Francis S. Collins, the consortium's informal leader, and Dr. J. Craig Venter of Celera, announced at the White House on June 26 that each had assembled the human genome.

Since Dr. Venter has now stated that Celera finished its computer assembly just the day before, on the night of June 25, it turns out that Mr. Kent's brilliant improvisation was the first assembly of the human genome, even though one that had and still contains many gaps.

"Without Jim Kent, the assembly of the genome into the golden path wouldn't have happened," said Dr. Collins, referring to the nickname for the GigAssembler, as the program is known.

Dr. Haussler, a professor of computer science, described his student as a superstar. "He's unbelievable," Dr. Haussler said. "This program represents an amount of work that would have taken a team of 5 or 10 programmers at least six months or a year.

"Jim in four weeks created the GigAssembler by working night and day," Dr. Haussler said. "He had to ice his wrists at night because of the fury with which he created this extraordinarily complex piece of code."

Having finished the GigAssembler, Mr. Kent then wrote another program known as a browser that enables many other kinds of genomic information to be aligned in tracks above the raw sequence of DNA bases, the chemical letters in which the human hereditary information is encoded. This extra information is essential for making sense of the DNA sequence and in particular for discovering where the genes are.

How did a mere private see the weakness in the order of battle that the consortium's generals had missed? Their competitor, Dr. Venter, had always relied on heavy-duty computation to hedge the enormous risks he took in his sequencing strategy. In setting up Celera he ordered a machine with vast processing power and four terabytes of memory, creating the largest computer in civilian use.

The consortium's DNA sequencing centers had nothing to match it because their strategy did not seem to require large-scale computation. The consortium breaks the human genome into large pieces known as BAC's and works from a simple map that shows how one BAC overlaps with the next in a tiling path that extends across the genome.

In December 1999, Dr. Eric S. Lander of the Whitehead Institute, who heads the consortium's genome analysis group, approached Dr. Haussler for help finding the genes in the human genome sequence, the first step in annotating it. Dr. Haussler decided the genome could not be properly analyzed until the pieces being generated by the consortium were in the right order.

The pieces of DNA were so small that their average size was less than that of the average gene.

Most of the consortium's BAC's were unfinished, each one consisting of up to 80 different pieces of unknown orientation and of unknown order within each BAC -- hardly worthy of being called a sequence, as Dr. Venter liked to observe of his competitors' efforts.

Dr. Haussler believed there was enough information, some generated inadvertently by the consortium and some available from other

sources, to orient and align the BAC fragments correctly.

He immediately started writing an assembly program. He persuaded the university's chancellor, Dr. Marcia Greenwood, that there was a chance to make a historic contribution and that she needed to advance him $250,000 to buy 100 computers with Pentium III processors. It was a previous chancellor of U.C. Santa Cruz, Dr. Robert L. Sinsheimer, who in 1985 held the first meeting to explore the idea of sequencing the human genome.

But progress was slow. In May 2000, when Mr. Kent sent an e-mail message to his colleague to ask how the assembly program was going, Dr. Haussler said he had to reply, "Jim, it's looking grim."

Mr. Kent sent an e-mail message back that said he felt he could write an assembly program using a simpler strategy. "I sent back a one word e-mail, 'Godspeed,' " Dr. Haussler said.

Mr. Kent, who turned 41 last Saturday, is a graduate student in his second career. In his first, which lasted more than 10 years, he ran a computer animation programming business.

Then he decided to go back to school and become a biologist. He started analyzing the DNA of C. elegans, the laboratory roundworm. When Dr. Haussler accepted the human genome assignment, Mr. Kent started adapting his worm programs to human DNA.

Mr. Kent said he offered to write a human genome assembly program for the consortium because of his concern that the genome would be locked up by commercial patents if an assembled sequence was not made publicly available for all scientists to work on. "The U.S. Patent Office is, in my mind, very irresponsible in letting people patent a discovery rather than an invention," he said. "It's very upsetting. So we wanted to get a public set of genes out as soon as possible."

After receiving his supervisor's "Godspeed" message, Mr. Kent started work on May 22. By June 22 he had written the GigAssembler together with subsidiary programs, in 10,000 lines of code, and had completed his first assembly of the human genome. "The two hard things, one is the repeat structure," he said, referring to the numerous very similar sequences of DNA letters in the human genome, "and then for me, something that makes the engineering very challenging is that you read DNA in two directions, so you are merging a lot of code that has eight directions."

The program must decide which of DNA's two strands each of the fragments belongs to because the sequences of letters in each strand run in opposite directions.

Mr. Kent then started work on another program, called the U.C.S.C. browser, which allows the assembled DNA sequence to be viewed in terms of its structural features, like genes and chromosomes. The browser allows one to search 21 tracks of information, each aligned with the underlying DNA sequence. The extra information helps to identify genes.

Although the consortium had deposited its BAC data in GenBank, a public DNA database, the U.C.S.C. browser gave biologists their first opportunity to view the assembled human genome unless they were subscribers to Celera's database. On July 7 the browser was posted on the World Wide Web. There was an overwhelming response. On that day the U.C.S.C. servers put out half a terabyte of information, Dr. Haussler said, and the browser now gets 20,000 hits a day from all over the world. "Nothing crashed, we just kept putting it out," he said. "People wrote back it was fantastic we had put it out with no intellectual property requirements, a gift with no strings attached."

Mr. Kent has designed the browser so that other biologists can add tracks of data to it, increasing the wealth of annotation. "It's been a wonderful stone soup, where other people have contributed bits," he said.

The consortium now lists the U.C.S.C. browser as the principal source for viewing its human genome sequence.

The consortium's three major centers, with their large staffs of computational biologists, did not write an assembly program of their own because they had not needed one in their pilot project, on the worm genome, and perhaps did not realize how useful such a program would be. The National Center for Biotechnology Information completed an assembly program only recently.

"It's easy for Venter to say of course you should have had this all planned out," Dr. Haussler said. "But he had the ability to organize an industrial-scale effort. Given the culture of the public project, it would have been very difficult to graft on an autocratic new software project, and if you had assigned a team to build this whole thing over several months it would have been a very difficult proposition. So what Jim has done is miraculous in many ways. No one expected anyone could come in and put this together in four weeks."

An even more telling accolade comes from Dr. Venter, who was astonished that the consortium managed to put the human genome sequence together, given what he regarded as the poor quality of its data.

"They used every piece of information available," he said. "It was really quite clever, given the quality of their data. So honestly, we are impressed. We were truly amazed because we predicted, based on their raw data, that it would be nonassemblable. So what Haussler did was he came in and saved them. Haussler put it all together."

Organizations mentioned in this article:
University of California (Santa Cruz); Celera Genomics

Related Terms:
Genetics and Heredity; Deoxyribonucleic Acid; Computers and the Internet; Computer Software; Human Genome Project