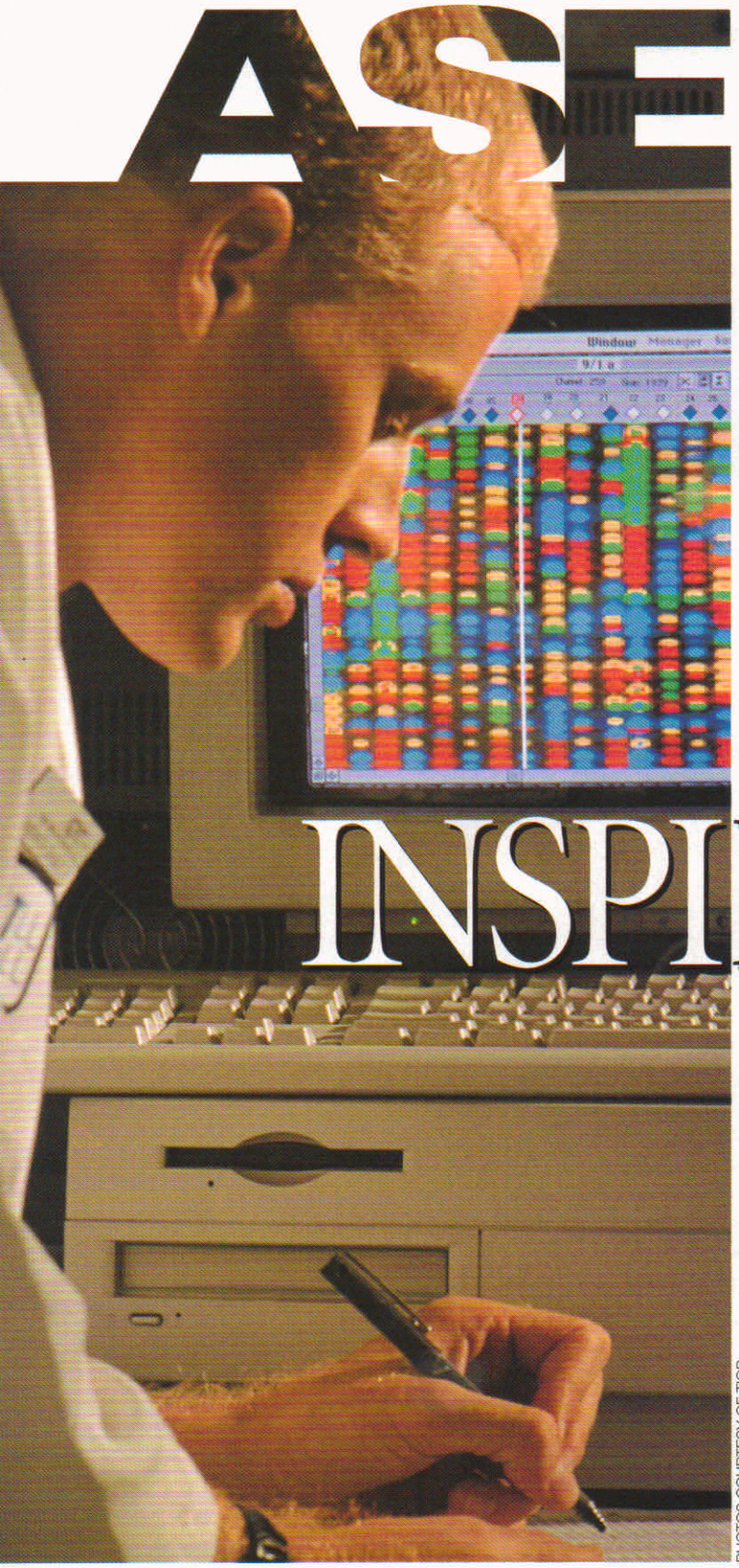
ASE

BY ALAN KARLS

The Institute for Genomic Research relies on Sybase to read the genome "book"

At the Speed of INSPIRATION



PHOTOS COURTESY OF TIGR

JOHN HEIDELBERG AND NAJIB EL-SAYED DON'T APPEAR AT FIRST glance to be doing earth-shattering work. Heidelberg references and cross-references computer-generated data. El-Sayed's group labors in a lab, running chemical experiments on containers of clear fluid.

In spite of outward appearances, however, they and their colleagues at The Institute for Genomic Research (TIGR) will provide huge benefits to mankind. That's because they're studying the genetic structure of organisms.

Their science is called *genomics*—learning to transcribe, read, and understand the genetic code of living organisms. Genetic material, with the code containing the instructions for all cellular structures and activities of the organism, is called the *genome*.

TIGR, a leader in the genomic revolution, performs structural, functional, and comparative analysis of the genomes of plants, animals, and pathogenic bacteria.

Although still in its infancy, genomics is predicted to lead in the next ten years to the end of some bacterial and viral diseases, remedies for inherited diseases, the development of new energy sources and efficient toxic waste cleanup, infallible criminal forensics, definitive histories of human population diversity, and other results as yet undreamed of.

TIGR, a leader in this genomic revolution, performs structural, functional, and comparative analysis of the genomes of plants, animals, and pathogenic bacteria. It was the first organization to develop the technique for analyzing the genetic blueprint and the first to map the genome of a free-living organism. To date, TIGR has mapped more complete genomes than any other entity.

TIGR couldn't have done it without Sybase's Adaptive Server Enterprise. Michael Heaney, the institute's database manager, says, "I cannot imagine anyone at the institute who does not make use of Sybase in his day-to-day job—it is fundamental to the vital work we do at TIGR." And the work TIGR does is incredibly complex, requiring the agility, durability, and speed that Adaptive Server Enterprise offers.

Immensely Difficult Task

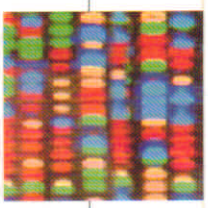
Say you have a book of three billion letters of a language you don't know—a book in several million disorganized fragments of about 500 letters each. A single contiguous letter sequence may appear on several duplicate fragments that start and end at different points in the letter sequence. You must match the last letters of one fragment to the first letters of another fragment and keep doing it until you put the entire sequence of three billion letters together beginning to end. Once it is assembled, you look at it, try to determine where the words begin and end, decide what the words mean, and then read it.

Complicating the task, about 40 percent of the words in the complete book don't mean anything; they are "junk" material left over from previous versions of the book, edited out, but remaining in the manuscript. You must find which of the words are relevant amid those that are not.

In this example, the "book" is the human genome. The "letters" are nucleotides, four different chemical units arranged in pairs—called base pairs—and linked in chains. Specific sequences of base pairs are called genes, the "words" in the analogy. Genes produce proteins, which define the organism's structure and life cycle. The particular order of the base pairs determines the complete instructions for an organism to build and maintain life.

Although an organism's code is found in a complete and contiguous "book," we can't read that book with the present state of technology. Because we can't read it, we have to re-create pieces of the book that we can read.

"Just one of the chromosomes of a malaria parasite took about two weeks to assemble."



Genomic scientists use gene-produced proteins to clone string fragments of 500 to 700 base pairs. A process called *sequencing* identifies the base pairs on the fragment. This information, the fragment name, and the base pairs on that fragment—called a *sequence read*—are the raw data entered into Adaptive Server Enterprise.

TIGR Assembler Does the Job

A powerful computer program called the TIGR Assembler, developed by TIGR, pieces together sequence reads in correct order by matching identical sequences of base pairs on the ends of separate sequence reads. All the millions of fragments must be reassembled in the right order to re-create the entire genome of a species.

"This is where it gets computationally intense," says Heaney. "The Assembler may run for weeks" to put the pieces together. "For instance, just one of the chromosomes of a malaria parasite took about two weeks to assemble. If the database goes down in

the middle of this, you have to start the process all over again," says Heaney. It requires a database with power and reliability. "Adaptive Server Enterprise doesn't go down."

"Here's an example of ASE's rock-solid stability," Heaney adds. "Last month, we had some electricians come in to do some testing of the fire-suppression system in our computer room. In the course of their work, they managed to accidentally shut off the power in the room, bringing every computer to an immediate and complete halt. Doing this to production systems usually has dire consequences and, in fact, the UNIX guys spent many hours trying to get their systems up and running, with one machine so damaged that it required new disks.

"The querying speed of ASE enables TIGR scientists to follow their inspiration."

"All five of my Sybase servers also went down, but all five servers, with a total of 435 data-bases on them, came up without a single problem. It was a beautiful thing to watch as Sybase went through each database in turn, applying any changes that were present within the transaction log, and then bringing the database online. Within 90 minutes, all of my servers were up and ready for business."

Once the Assembler has utilized Adaptive Server Enterprise to rebuild the genome, scientists have to make sense of it. In the book analogy, where do the words start and stop? What does each word mean? In the genome, where do the genes start and stop? What does each gene do?

The process, according to Heaney, involves a scientist saying, "If we read from here, then this is a gene and this amino acid must make this protein. Therefore, these proteins come from this gene. This protein makes this happen; therefore, this gene makes this happen."

This process is called "annotating" the genome—identifying the gene list, pointing out where a gene appears on the genome, and identifying its biological function. To do all of this, the scientists run searches of TIGR's own database and the databases of all other research institutes to find matches of previously annotated genes—genes with known functions. If they can't find matches for a gene, they progress to searches for protein matches, then amino acid matches. If they find no matches at all, they must run chemical analyses to find out firsthand the nature of the gene sequence.

"Genomes produce far more questions than they do answers," says Dr. Heidelberg, a scientist at TIGR who analyzes the computer-generated genomic sequence data to find the function of genes. Why is this gene duplicated? Where has this gene or pattern appeared elsewhere in this or another organism? Which of these genes is active? When does this gene turn on? When the answers come, they will have a significant impact.

Dr. El-Sayed, another scientist at TIGR, runs chemical experiments on gene molecules of unknown function. He is working on a parasite that has two life-cycle stages, living inside an insect that transmits it to humans, then in the bloodstream of humans, where it causes disease.

"We experiment to identify genes that are turned on in one stage of the parasite life cycle versus the other. For example, a gene of unknown function that is expressed only when the parasite is in its human host may be an important factor in molecular events that define its pathogenicity. We're also looking for unique molecules in the parasite that can be targeted without affecting the human host in any way."

When El-Sayed and his colleagues identify such targets, they will be able to develop new routes for chemotherapy and vaccine development against the deadly parasite.

continued on page 47

continued from page 16

Heaney adds, "When our scientists hook onto a train of thought, they have to be able to quickly explore the various avenues that are presented to them. Otherwise the inspiration can elude them. The querying speed of Adaptive Server Enterprise enables them to follow their inspiration, and its flexibility allows them to interrogate and drill down into the data in whichever ways they choose. You can say that Adaptive Server Enterprise runs at the speed of inspiration."

TIGR places the sequenced and annotated genomes into Adaptive Server Enterprise databases for free and open access by the organization's own scientists and researchers from any other genomic research institute. TIGR's gene banks are unique in the number of species sequenced and in the manner in which they can be used. Through its Adaptive Server Enterprise databases, TIGR offers the public the most comprehensive information on genomic sequences, providing great value to all researchers making cross-species comparisons.

Says Heidelberg, "No single individual can ask all the questions—to make those leaps of inspiration that advance genomic research. We put the results of our research, held in Adaptive Server Enterprise, out on the Web so that scientists with diverse specialties can ask those questions and begin to answer them." □

TIGR by the Numbers

TIGR now has five Sybase Adaptive Server Enterprise systems, including one dedicated to system development. Two production servers deployed on Sun Solaris and one on Linux run more than 220 research project databases, totaling a massive 340 gigabytes of data. A fifth Sybase Adaptive Server running on Linux forms the database engine behind the TIGR Web site.

TIGR also develops research software programs, many of which it makes freely available to the nonprofit scientific community via the TIGR Web site. The institute has developed numerous programs over the years using Sybase Open Client and has recently begun to employ other Sybase development tools such as PowerDesignerjConnect, and PowerJ.

www.sybase.com/corporate/events

WEDNESDAY, 10
Gartner ITEXPO, Orlando, FL. Information: www.sybase.com/corporate/events

THURSDAY, 11
Gartner ITEXPO, Orlando, FL. Information: www.sybase.com/corporate/events

MONDAY, 15
Government in Technology, Ottawa, Canada. Contact: www.sybase.com/corporate/events

TUESDAY, 16
Government in Technology, Ottawa, Canada. Contact: www.sybase.com/corporate/events

WEDNESDAY, 17
Government in Technology, Ottawa, Canada. Contact: www.sybase.com/corporate/events

Sybase Worldwide Marketing Conference, Emeryville, CA. Contact: www.sybase.com/corporate/events

THURSDAY, 18
Government in Technology, Ottawa, Canada. Contact: www.sybase.com/corporate/events

Sybase Worldwide Marketing Conference, Emeryville, CA. Contact: www.sybase.com/corporate/events

FRIDAY, 19
Sybase Worldwide Marketing Conference, Emeryville, CA. Contact: www.sybase.com/corporate/events

TUESDAY, 23
Financial Technology Expo, New York, NY. iAnywhere Solutions showcase. Information: www.sybase.com/solutions

WEDNESDAY, 24
Financial Technology Expo, New York, NY. iAnywhere Solutions showcase. Information: www.sybase.com/solutions

EP Technical Briefing, Spain. Call: 1-800-8SYBASE

THURSDAY, 25
Financial Technology Expo, New York, NY. iAnywhere Solutions showcase. Information: www.sybase.com/solutions

KM Studio Seminar, Denmark. Information: amortens@sybase.com

MONDAY, 29
TMA 2001, United Kingdom. Call: 1-800-8SYBASE

TUESDAY, 30
TMA 2001, United Kingdom. Call: 1-800-8SYBASE

Windows on Healthcare, San Diego, CA. iAnywhere Solutions. Information: www.sybase.com/corporate/events

WEDNESDAY, 31
TMA 2001, United Kingdom. Call: 1-800-8SYBASE

Windows on Healthcare, San Diego, CA. iAnywhere Solutions. Information: www.sybase.com/corporate/events