

Stanford University

Advancing Computational Biology Databases with Sun™ Technology



Highlights

Industry

Education/Research

Applications/Solutions

Computational Biology Databases

Products/Services

- Sun Fire™ V480 dual-processor server
- Sun Fire V120 server
- Sun Enterprise™ 220R, 420R, and 4500 servers
- Sun StorEdge™ A1000 and D1000 disk arrays
- Java™ technology
- JavaServer Pages™
- JDBC™
- Solaris™ 8 Operating Environment
- SunSpectrum™ Support

Key Challenges

- Implement state-of-the-art knowledge base for storage and analysis of diverse pharmacogenetic research data
- Implement state-of-the-art facility to store, edit, analyze, and disseminate large volumes of experimental microarray data
- Provide computational power and storage capacity to accommodate exponential growth

Key Results

- Data repository and research tool for multi-million dollar government studies
- Position of leadership in pharmacogenetic knowledge base technology
- Largest academic microarray database in the world

“Windows, or Windows NT never even entered into the discussion. We didn’t consider it to be sufficiently robust, and you can’t get large enough Windows machines to do what we needed to do.”

– Dr. Gavin Sherlock, Director of the Stanford Microarray Database

Without the aid of massive databases, and networked computational systems, it would be virtually impossible to store, disseminate, and analyze the avalanche of genetic data that is being generated on a daily basis by research facilities around the world. For more than 25 years, Stanford University has pioneered advanced research to develop new forms of informatics technology focused on the biomedical sciences. Stanford’s Microarray Database (SMD) and Pharmacogenetics Knowledge Base (PharmGKB) are two systems on the vanguard of the bioinformatics revolution.

The Stanford Microarray Database

The Stanford Microarray Database (SMD) is the largest academic microarray database in the world, storing experimental data from approximately 200 labs around the world—and making this data publicly accessible over the Web. The system runs on an eight-processor Sun Enterprise 4500 server, allowing researchers to load, edit, search, filter, and analyze large volumes of experimental microarray data. Such microarray experiments focus on gene discovery, disease diagnosis, drug discovery, toxicological research, and more.

The Sun™ E4500, running the Sun Solaris Operating Environment, was selected for the SMD because of its stability, scalability, processing power, compatibility with third-party software products, and support program. “Windows, or Windows NT never even entered into the discussion,” said Dr. Gavin Sherlock, Director of the Stanford Microarray Database. “We didn’t consider it to be sufficiently robust, and you can’t get large enough Windows machines to do what we needed to do.”

World’s Largest Academic Microarray Database

The first version of the SMD system was intended simply to support a set of internal Stanford laboratories. This early implementation entailed using a Perl/CGI wrapper to store experimental data as a collection of flat files.

Microarrays contain thousands of microscopic DNA “probes” affixed to a small slide. These probes are used to establish the presence of a particular gene, or to verify the gene’s expression in a particular cell. DNA contained in a solution applied to the microarray will bind, or “hybridize,” to a particular complementary probe site on the slide. Those DNA fragments that have hybridized

“From my perspective, as Principal Investigator, the goal is to have the hardware work, and shrink to the background, so that we can focus on the hard information modeling and analysis tasks. So far, Sun hardware has done this for us.”

– Dr. Russ Altman, Principal Investigator for Stanford University’s PharmGKB Project

can then be detected (and quantified) via a laser that excites a fluorescent dye used to chemically label the experimental sample strands.

But such microarray experiments can generate extremely large volumes of data. A single microarray can contain 20,000 DNA spots. And up to 40 different types of data are typically generated for each spot. That translates to approximately 1 million data points for a single array. “And an actual experiment may use dozens of microarrays,” said Sherlock, “so an experiment using 50 arrays might easily generate 50 million data points.”

With these data needs in mind, it soon became apparent that the flat file implementation of the SMD would not be sufficiently scalable—particularly with an increasing number of both internal and external labs wanting to store their data on the system. “There are about 80 labs on campus doing microarray experiments,” said Sherlock, “and they have collaborations with probably another 120 labs around the world.”

Approximately a year and a half ago, the SMD project purchased a Sun E4500 server to house their system, using the Solaris 8 Operating Environment, and the Oracle8i Enterprise Edition DBMS. This current implementation employs the E4500, with 8 processors at 400 MHz, 8 GB of RAM, and 4 I/O boards controlling 2 Sun StorEdge D1000 disk arrays and 7 Sun StorEdge A1000 disk arrays. The total usable disk space allocated for the SMD is 1.2 TB. Additionally, a Sun Enterprise 420R server, with 4 processors at 450 MHz, and 4 GB of RAM, is used for remote processing and analysis. Both systems are supported under the SunSpectrum Silver contract.

Hardware to Spare

A key aspect of the SMD is its facility of connecting microarray data with the underlying biological data that pertains to the DNA on the slide. For each experiment, extensive metadata is stored—including the name of the researcher, a category and subcategory that describe details of the experiment, and the organism that served as the source of the DNA spotted on the microarray. Researchers can apply complex logical filters to the data, view clickable image maps of the array results, as well as explore associated biological information.

The data are currently loaded into the database via a Web form. The system accepts data produced by both GenePix and Scanalyze (microarray image analysis software). In the future, the SMD hopes to store information sufficient to actually reproduce a given experiment. And a second goal is to implement a means whereby the results of analyses conducted on the system can be stored. In essence, the database could store the results of computer-based experiments, with information as to how to recreate those experiments.

The SMD currently holds approximately 200 GB of data, but this is growing all the time. As a result, the scalability of the Sun hardware is of paramount importance. “The majority of the data are held within a single table,” said Sherlock. “We actually have a table with approximately 600 million rows in it, which is quite substantial compared to most databases. So the ability to add additional storage to the system is extremely important.”

Stanford's PharmGKB

Stanford's Pharmacogenetics Knowledge Base (PharmGKB) sits at the forefront of storing, organizing, and logically interconnecting genetic, clinical, and phenotypic research data—all within a publicly accessible Web-based facility. The purpose of the knowledge base is to aid researchers in better understanding how genetic variations between individuals contribute to their differing responses to pharmaceutical drugs. Funded by the National Institutes of Health, PharmGKB houses experimental data from ongoing pharmacogenetic studies, providing tools to submit, edit, browse, query, and process the information.

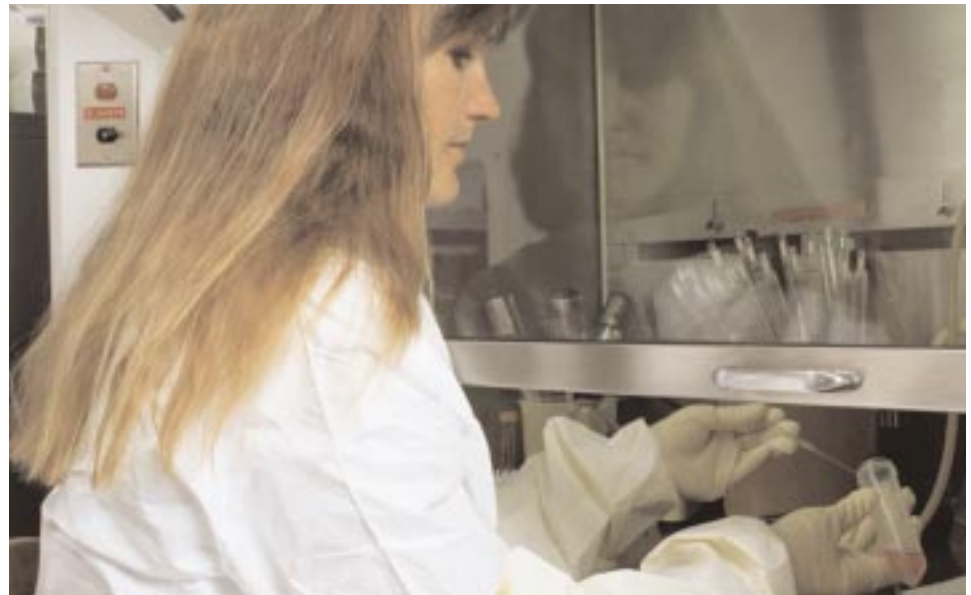
The system will soon undergo a major hardware overhaul, comprised entirely of Sun equipment. This new three-tiered implementation of PharmGKB will consist of:

- **Database tier:** Sun Fire V480 server (using the Solaris 8 Operating Environment), with 2 processors at 900 MHz, 4 GB of RAM, and Sun StorEdge A1000 RAID disk arrays (12x36 GB).
- **Application tier:** Dual-processor Sun Enterprise 220R server.
- **Web tier:** Sun Fire V120 server (using the Solaris 8 Operating Environment), with 1 processor at 650 MHz, and 1 GB of RAM.

All three systems will be supported by the SunSpectrum Silver contract. The system will be replicated in three complete sets—for development, beta, and production environments.

"The reasons for choosing Sun are:

1) scalability—as our usage goes up to the levels we are anticipating, 2) reliability—this is a federally funded service activity that requires 24 x 7 uptime, and 3) a mature software environment for delivering both web services and Java software, and integrating a number of other resources," said Dr. Russ Altman, Principal Investigator for PharmGKB. "From my perspective, as Principal Investigator, the goal is to have the hardware work, and



shrink to the background, so that we can focus on the hard information modeling and analysis tasks. So far, Sun hardware has done this for us."

The Quest for Individualized Medicine

It has long been known that patient responses to pharmaceutical drugs are variable and sometime unpredictable. Adverse drug reactions led to more than 2 million hospitalizations and 100,000 deaths in one year studied. And it also known that such variability in drug response is in good part genetically determined. Pharmacogenetics is the study of how genetic variations contribute to differences in drug responses—with an ultimate goal of establishing individualized drug therapies based upon a better understanding of the relationship between genotype and phenotype (physical manifestation of genotype).

In the year 2000, the National Institutes of Health sponsored the Pharmacogenetics Research Network, a group of multi-million dollar research projects, with a mandate to accumulate, process, and store pharmacogenetic experimental results. PharmGKB serves as a central knowledge base of data collected by the research network, and provides tools for submitting, viewing, editing, and interpreting the information.

A key challenge of such a knowledge base, however, is that of establishing a logical mapping between genetic, clinical, and phenotypic data, interconnected by relations, and organized by levels of abstraction. "We've had to design our own infrastructures, in terms of modeling the data, and storing the data," said Dr. Teri Klein, Project Director of PharmGKB. There are currently over 600 relationships used in PharmGKB, and researchers can submit 38 different types of data, containing information pertaining to genomics, drugs, diseases, populations, and more. Submissions are made through Web forms, or are uploaded via PharmGKB-defined XML elements.

Room To Grow

The first version of PharmGKB went online in 2001. But the system consisted of multiple tiers residing on the same machine, which ultimately resulted in scalability problems. Also, this early incarnation of PharmGKB used Stanford's frame-based Protégé tool as both its knowledge representation infrastructure, and storage facility. Protégé is a suite of data modeling and acquisition tools. But without a formal relational database beneath Protégé, the system soon ran into processing bottlenecks in terms of searching and storage.

“We needed this equipment to scale to our needs for the next three to four years, so we were pretty picky. We chose our hardware based upon memory, processor speed, and storage capacity. That’s one of the advantages of Sun architecture—it’s easier to grow in capacity and processing power in a more structured fashion.”

– Dr. Teri Klein, Project Director of Stanford University’s PharmGKB Project

In March 2002, PharmGKB employed a hybrid system consisting of Protégé 2000, and Oracle 8i as its underlying relational database.

With the Sun hardware, the PharmGKB system will be well positioned for its expected exponential growth phase. “If you look back at GenBank, the NIH’s genetic sequence database, it started off with a very gradual growth curve for the first five years or so,” said Klein, “then all of a sudden, it just took off. And I think we’ll see the same kind of growth here. We will eventually be storing terabytes of data.”

In the government medical research sector, funding is often at a premium. So selection of the proper hardware is always of the utmost importance. “We needed this equipment to scale to our needs for the next three to four years, so we were pretty picky,” said Klein. “We chose our hardware based upon memory, processor speed, and storage capacity. We bought machines with two processors, but we can easily go to four processors if we need them. That’s one of the advantages of Sun architecture—it’s easier to grow in capacity and processing power in a more structured fashion.”

Future Functionality

Mechanisms for surveillance of, and interaction with, external data will provide even more powerful query capabilities to PharmGKB. The system monitors the NIH’s dbSNP database, looking for new information about genes of interest to the various associated research groups. And PharmGKB’s dbSNP tool also formulates relevant submissions to dbSNP. This automatic submission and surveillance allows PharmGKB to both contribute to an external database and to correlate its information with theirs.

And the hardware and software infrastructure of PharmGKB will serve it well as interconnections and associations between related genetic database become ever more common. The system uses JavaServer Pages and Java servlets for generating dynamic content, JDBC for cross-DBMS connectivity, and XML for platform-neutral data exchange.

“We’re very happy with the Sun architecture that we’re moving to,” said Klein. “We think it will really help our developers for the future, and remove some of the bottlenecks we’ve seen in the past. And it’s important that what we’ve bought will be able to grow with us, because what we will need a year from now is not what we will need today. We expect to be humming!”

Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA Phone 1-650-960-1300 or 1-800-555-9SUN Web sun.com



Sun Worldwide Sales Offices: Argentina: +5411-4317-5600, Australia: +61-2-9844-5000, Austria: +43-1-60563-0, Belgium: +32-2-704-8000, Brazil: +55-11-5187-2100, Canada: +905-477-6745, Chile: +56-2-3724500., Colombia: +571-629-2323, Commonwealth of Independent States: +7-502-935-8411, Czech Republic: +420-2-3300-9311, Denmark: +45 4556 5000, Egypt: +202-570-9442, Estonia: +372-6-308-900, Finland: +358-9-525-561, France: +33-134-03-00-00, Germany: +49-89-46008-0, Greece: +30-1-618-8111, Hungary: +36-1-489-8900, Iceland: +354-563-3010, India-Bangalore: +91-80-2298989/2295454; New Delhi: +91-11-6106000; Mumbai: +91-22-697-8111, Ireland: +353-1-8055-666, Israel: +972-9-9710500, Italy: +39-02-641511, Japan: +81-3-5717-5000, Kazakhstan: +7-3272-466774, Korea: +82-2-193-5114, Latvia: +371-750-3700, Lithuania: +370-729-8468, Luxembourg: +352-49 11 33 1, Malaysia: +603-21161888, Mexico: +52-5-258 6100, The Netherlands: +00-31-33-45-15-000, New Zealand-Auckland: +64-9-976-6800; Wellington: +64-4-462-0780, Norway: +47 23 36 96 00, People’s Republic of China-Beijing: +86-10-6803-5588; Chengdu: +86-28-619-9333; Guangzhou: +86-20-8755-5900; Shanghai: +86-21-6466-1228; Hong Kong: +852-2202-6688, Poland: +48-22-8747800, Portugal: +351-21-4134000, Russia: +7-502-935-8411, Saudi Arabia: +9661 273 4567, Singapore: +65 6438-1888, Slovak Republic: +421-2-4342-94-85, South Africa: +27 11 256-6300, Spain: +34-91-596-9900, Sweden: +46-8-6311-0-00, Switzerland-German: 41-1-908-90-00; French: 41-22-999-0444, Taiwan: +886-2-8732-9933, Thailand: +662-344-6888, Turkey: +90-212-335-22-00, United Arab Emirates: +9714-3366333, United Kingdom: +44-1-276-20444, United States: +1-800-555-9SUN OR +1-650-960-1300, Venezuela: +58-2-905-3800, Or Online at sun.com/store

SUN™ THE NETWORK IS THE COMPUTER © 2002 Sun Microsystems, Inc. All rights reserved. Sun, Sun Microsystems, the Sun logo, Sun Fire, Sun Enterprise, Sun StorEdge, Java, JavaServer Pages, JDBC, Solaris, SunSpectrum, and The Network Is The Computer are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and other countries.

LF3.2 Printed in USA 09/02 FE1890-0/2K