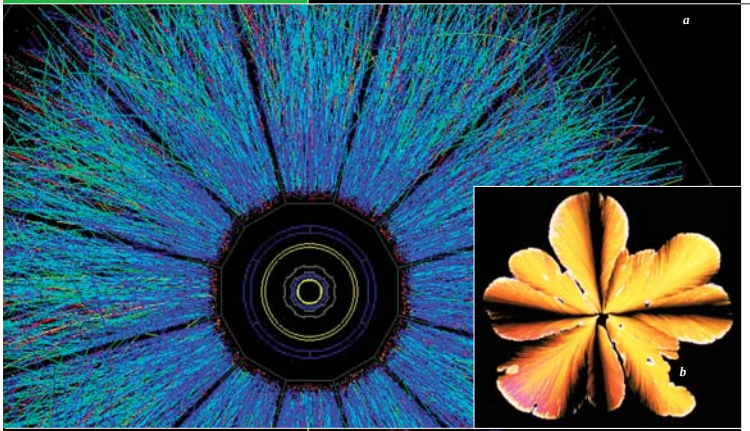
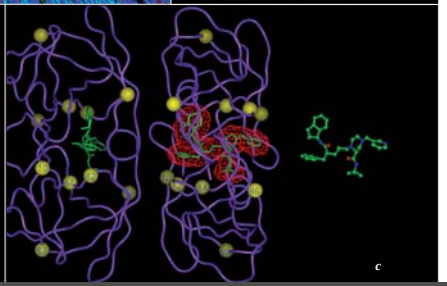
Foundations for SCIENTIFIC LEADERSHIP

LARGE-SCALE SCIENTIFIC COLLABORATION





Never would I have dreamed in 1953 that my scientific life would encompass the path from DNA's double helix to the 3 billion steps of the human genome. The completion of the Human Genome Project is a truly momentous occasion for every human being around the globe.

James D. Watson, co-discoverer of the structure of DNA, during DOE/NIH double celebration in April 2003 marking the discovery's 50th anniversary and the successful conclusion of the decoding effort it made possible.

a) Spray of subatomic particles produced by high-speed collision of gold ion beams in DOE/SC's Relativistic Heavy Ion Collider. A 1,200-ton instrument called the Solenoid Tracker at RHIC (STAR) records data from collision. High-end computation and data-sharing are critical in such unique experiments, enabling many scientists to reconstruct and visualize the sequence of events from the moment of impact and to analyze the data for evidence of new forms of matter – a central goal of high-energy physics research.

The first great scientific triumph of this century – the decoding of the human genome announced in February 2001 – dramatically introduced the new era of large-scale collaboration enabled by computing and networking. When DOE/SC and NIH launched the Human Genome Project in 1990, the most powerful computers were 100,000 times slower than today's high-end machines, the private citizens using networks could send data at only 9600 baud (bits per second), and many geneticists did calculations by hand. The challenge – figuring out how the genetic instructions for life are organized in the four chemical compounds that make up the biomolecule deoxyribonucleic acid (DNA), called life's blueprint – was understood to be critical to the future of medical science. But the work was expected to take decades.

Ultimately, the international decoding effort, in which more than 1,000 scientists participated, became a showcase for the central role of IT in advanced research. The distributed teams of Human Genome Project researchers each computed pieces of possible chemical sequences and transmitted them over high-speed networks to the project's data repositories for other scientists to examine and use. Researchers devised new software tools that automated sequence computations and analyses. In a sign of the enormous IT-driven acceleration of the work, a June 2000 announcement of a "rough draft" of the genome noted that more than 60 percent of the code had been produced in the prior six months alone, at a rate of 1,000 bases (sequences of DNA's four nucleotides) per second around the clock. The total raw sequence calculations submitted numbered more than 22 billion.

Connecting people and resources

In every scientific domain today, investigating complex phenomena requires a great many minds, multiple skills, and state-of-the-art tools. Such research can be staggeringly data-intensive, often involving unique measurements generated by one-of-a-kind instruments. Just one collision experiment in an advanced particle accelerator, for example, produces millions of physics "events" per second, amounting to 40 or more terabytes of data all told, filling 8,000 high-capacity magnetic tapes. If the tapes were stacked, the pile would be 500 feet high. Such massive data sets exceed the capacity of all but the largest supercomputing systems to store or process.

The NITRD agencies are leaders in envisioning and developing revolutionary IT capabilities to connect people in any location to such massive data sets, as well as to colleagues and computing resources, in secure high-bandwidth infrastructures for real-time scientific collaboration. In FY 2003, efforts funded by DOE/NNSA, DOE/SC, NASA, NIH, and NSF continue to build out the high-speed networks, terascale computing and petabyte storage resources, and advanced software for grid computing that enable large groups from all points of the compass to work together in a joint enterprise. In these leading-edge research environments, the NITRD community is exploring and prototyping technologies and methods for large-scale collaborative human activity that will become the norm for many American workers in the 21st century. Citing the cost-effectiveness of collaboration and resource sharing, businesses are already adopting NITRD-developed grid technologies to connect distributed teams and computing resources.

In a significant milestone for large-scale collaboration, NITRD-engineered high-speed networks (including DOE/SC's Energy Sciences Network, ESnet, and the NSF-sponsored TeraGrid and StarLight links) in February 2003 enabled an international team of particle physicists and corporate IT partners to set a new world network-speed record. The team sent 6.7 gigabytes of uncompressed data – the equivalent of four hours of DVD-quality movies – from Sunnyvale, California to Amsterdam, Netherlands in 58 seconds.

The High Performance Wireless Research and Education Network (HPWREN) sponsored by NSF at the University of California-San Diego uses advanced wireless capabilities to provide Internet links at educational institutions in remote sites, such as Native American communities. HPWREN is also prototyping a variety of distributed data-collection applications for high-speed wireless networking, including real-time measurement and multicasting of seismic events in southern California along the San Jacinto fault, wildfire tracking, and sensor networks for ecological monitoring in remote habitats.

b) DNA molecule in liquid crystalline form.
c) Perspectives of HIV-1 protease, showing protease inhibitor indinavir in green. Red mesh represents computed solvent accessible surface, a key indicator of protein structure. Such computational modeling is greatly speeding the work of identifying the structures of the some 100,000 proteins in the human genome, which will enable new treatments for diseases like HIV.