North Carolina Bioinformatics Grid

Information Infrastructure for the Biology of the 21st Century

The sequencing of entire genomes from microbes to plants, mice, and humans is the departure point for a new biology—the biology of the 21st century. The goal of this new biology is to achieve, *for the first time*, a fundamental, comprehensive, and systematic understanding of life and, from this, new powers of prediction, prevention, and remedy that will impact the health and well-being of all of the peoples of the world. Information science—computational modeling and data mining and analysis—will play a major role in realizing the new biology. The *North Carolina Bioinformatics Grid* will provide access to, and coordination of, the computing, data storage, and networking infrastructure required for researchers and educators throughout North Carolina to take full advantage of the genomics revolution.

Genomes
Foundations for a New Biology

In June 2000, the President of the United States and the Prime Minister of Great Britain announced, with great fanfare, the completion of the draft sequence of the human genome. This event, when coupled with publications of the genomes of a number of bacteria, plants, and animals, heralds a new era in biology. With this information, and that to be produced in the coming years, it will be possible to achieve an understanding of life at its most fundamental level. Technologies based on this knowledge promise a new generation of unique therapeutics: disease-, and even patient-, specific drugs that are far more effective than the drugs of today. Equally as important, genomics can lead to the development of hardier crop plants to benefit farmers, more nutritious foods to fight hunger and malnutrition, and enhanced microbial systems and plants for cleaning up toxic waste, oil spills, or polluted air.

Genomic Data
An Opportunity and a Challenge

The availability of massive compilations of genomic and related data is merging biology with information science. Over the next several years, genomic data on a broad range of organisms will be generated by publicly and privately financed sequencing programs. These data sets will be augmented by data on RNA and the proteome, the ensemble of proteins encoded in the genome, and on the bio-networks that describe the metabolic, regulatory, and other processes of life (these are the actors and scenes in the play of life). Data mining and analysis, along with computational modeling, will play a critical role in synthesizing this data into a holistic description of life—the goal of the new biology.

Storage and management of these data sets will require data systems capable of managing petabytes (millions of billions of bytes) of data. Analysis of these data sets and modeling biological processes will require high performance computing systems capable of tens to hundreds of trillions of arithmetic operations per second. This is 100,000 times more data than can be stored on the hard disks found in today's high-end personal computers, and 100,000 times more computing power than these advanced PCs possess. The new biology is both compute- and data-intensive and will require innovations in mathematics and information technology as well as biology to succeed.

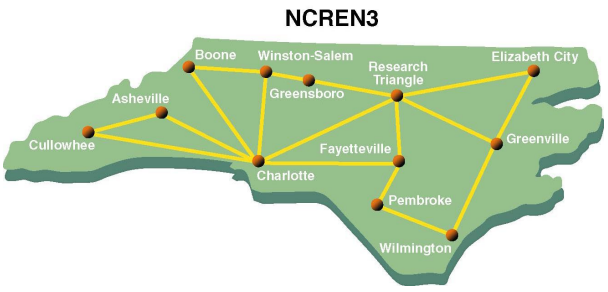
North Carolina
Bioinformatics Grid

To provide the computing, data storage, and networking capabilities to support the genomics revolution, members of the North Carolina Genomics and Bioinformatics Consortium are working with computer and networking companies to create the *North Carolina Bioinformatics Grid*. The NCGBC was established in December 2000 by the North Carolina Biotechnology Center to promote genomics, proteomics, and bioinformatics education, research, and development in North Carolina. The Consortium consists of over 70 organizations—universities and colleges; biomedical, biotechnology and information technology companies; nonprofit institutions; and foundations.

The *NC BioGrid* will accumulate the vast library of genomics, proteomics, and related data being created throughout the world, combine it with non-proprietary data from Consortium members, and make this invaluable collection of data available to researchers and educators throughout North Carolina. The BioGrid will

also make available high-performance computing hardware and software to mine, analyze, and model this data and a high speed network to move the data sets among Consortium sites. The base infrastructure for the *NC BioGrid*—terascale computer, petascale data store, and high-speed network—will be provided by the North Carolina Supercomputing Center and the North Carolina Research and Education Network. These resources will be multiplied by the shareable resources available at other institutions in the Consortium. Although these institutions will, by and large, tailor their computing and data storage systems to meet their own needs, they may make resources available to other Consortium members when they are not being otherwise used. Any institution may also elect to be responsible for collecting and curating selected data sets.

Grid middleware—the software that gives intelligence to the network—will be used to bind the computing and data resources scattered across the network into a unified environment for genomics research and education. The *NC BioGrid* will allow Consortium members to share computing and data resources as well as software for mining, analyzing, and modeling the data. The BioGrid will also facilitate collaboration among North Carolina's researchers and educators in computational biology, bioinformatics, genomics, and proteomics.



The *NC BioGrid* will draw on other grid research projects underway in the U.S. and Europe, including *GriPhyN* (the *Grid Physics Network*), the European *DataGrid*, and the United Kingdom's *eScience Grid*. These efforts are largely focused on applications in physics, especially high-energy physics, but many of the fundamental capabilities needed for a physics grid and a bioinformatics grid, *e.g.*, managing large distributed data sets, are identical.

Benefits of NC BioGrid to Researchers and Educators in North Carolina

For researchers and educators, the *NC BioGrid* will streamline access to the genomic, proteomic and related data being compiled by various national and international efforts. To the scientist, it will appear as if these data reside on his or her personal computer and that the analysis is being performed there. In reality, the data may be stored at one or more remote Consortium sites and the calculations performed at yet another Consortium site, with the middleware that underlies the *NC BioGrid* co-scheduling computing, data, and networking resources as needed. The BioGrid will allow biologists to concentrate on biology, not the arcana of computing, networking, and data storage, thereby increasing their productivity and creativity as well as their opportunities for collaboration.

Benefits of the NC BioGrid to Citizens of North Carolina

For the citizens of North Carolina, the *NC BioGrid* will help the State take advantage of the opportunities offered by the revolution in genomics. The BioGrid will create a collaborative computing, data management, and networking infrastructure for life sciences research and education that will establish North Carolina as a leader in this highly competitive field. This capability will attract additional investments to North Carolina—new federal and private grants for the State's research and educational institutions and new businesses to stimulate and sustain the State's economy. But, most important, these investments will provide North Carolina with an unparalleled opportunity to contribute to the world's store of knowledge and to the welfare of the peoples everywhere.

For more information on the North Carolina Bioinformatics Grid, see the BioGrid web site at:

http://www.ncbiogrid.org/
(available after December 21, 2001)

For more information on the North Carolina Genomics and Bioinformatics Consortium, see the Consortium web site at:

http://www.ncgbc.org/
