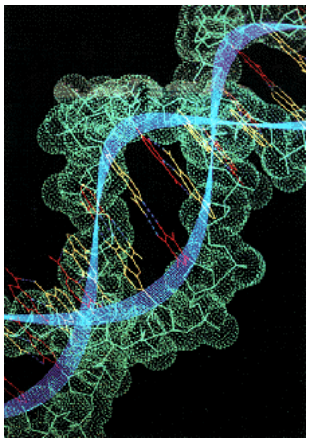
Genomes to Life Requires New Life from Networks

Building on the successes of the Human Genome Project, DOE offices of Advanced Scientific Computing Research and Biological and Environmental Research has initiated an ambitious program to achieve the most far-reaching of all biological goals: a fundamental, comprehensive, and systematic understanding of life. DOE's Genomes to Life program will make important contributions in the quest to venture beyond characterizing such individual life components as genes and other DNA sequences toward a more comprehensive, integrated view of biology at a whole-systems level. As we enter the age of system-level biology, it is becomes necessary to examine different types of biological data from many different sources, reflecting different aspects of a biological system. Such task requires the collective expertise from different disciplines to analyze, understand and draw conclusions form the data. Experts are located at various national laboratories, universities and other institutions, and currently can interact over the Internet, but are severely constrained by its limitations. In particular, the next generation of these system-level biological tasks require networks and networking technologies with capabilities significantly beyond the current state-of-the-art.



Genomes to Life: The Big Picture

The major goal is to use DNA sequences from microbes and higher organisms, including humans, as starting points for systematically tackling questions about the essential processes of living systems. Advanced technological and computational resources will help to identify and understand the underlying mechanisms that enable organisms to develop, survive, carry out their normal functions, and reproduce under myriad environmental conditions. This approach ultimately will foster an integrated and predictive understanding of biological systems and offer insights into how both microbial and human cells respond to environmental changes. In a related program, the Microbial Cell Project (MCP) provides the ultimate test of GTL findings. MCP takes a whole-genome approach to understanding the function and regulation of genes for a single living system and the pathways in which the protein products interact. The applications of this level of knowledge will be extraordinary and will help DOE fulfill its broad missions in energy, environmental remediation, and the protection of human health.

The Network Challenge

To support various system-level biology tasks, it is becoming necessary to form interdisciplinary centers consisting of experts located at different geographical locations. The data related to different aspects of a biological system is collected, processed, analyzed and stored at different locations. This data must be cooperatively accessed and analyzed by teams of experts. Many bio-informatics prediction/modeling software modules have been developed to help biologists to make sense of such data. It is most cost-effective to support such efforts over wide-area networks with high-performance capabilities. The software typically, at different stages of the computation, needs to make repeated accesses to various bio-databases that are scattered all over the country. These databases are typically several Gigabytes in size and increasing rapidly. The research groups install local copies of these large databases to save the data transfer time. But as the amount of bio-data increases exponentially due to the advanced capabilities in analytical technologies for biology, it will soon become unrealistic to keep local copies of all bio-databases in every single biology lab. Then the bottleneck in using the databases is the network communication time. Another important network-related limitation manifests in the prediction or modeling tasks for biological sciences, such as the simulation of dynamics of bio-molecules over large time-scales. Supercomputers at different physical sites will be harvested to provide the computational power needed for these long time-scale simulations, which must be tightly coordinated to ensure the effective utilization of computers. It is important that the end-to-end message delays be minimized over the networks to ensure that the supercomputers do not idle waiting for control messages. Note that a single second of idle time represents the loss of several teraflops of compute power. While the first problem of data rates can be addressed by building a faster network infrastructure (e.g. with multiple OC 192 channels) together with high-performance data transport protocols, the end-to-end delay minimization represents a significant challenge to networking technologies. Such problems are currently addressed in a limited way in overlay networks and daemons, but highly focused research and development efforts are needed for an effective solution to this class of applications.

Elements of an Office of Science Program

The networking technologies required in this class of applications represent a quantum step forward in the area of high-performance networks and provide a unique opportunity to Office of Science to be a national leader in enhancing the productivity of a wide class of Genomes to Life projects.