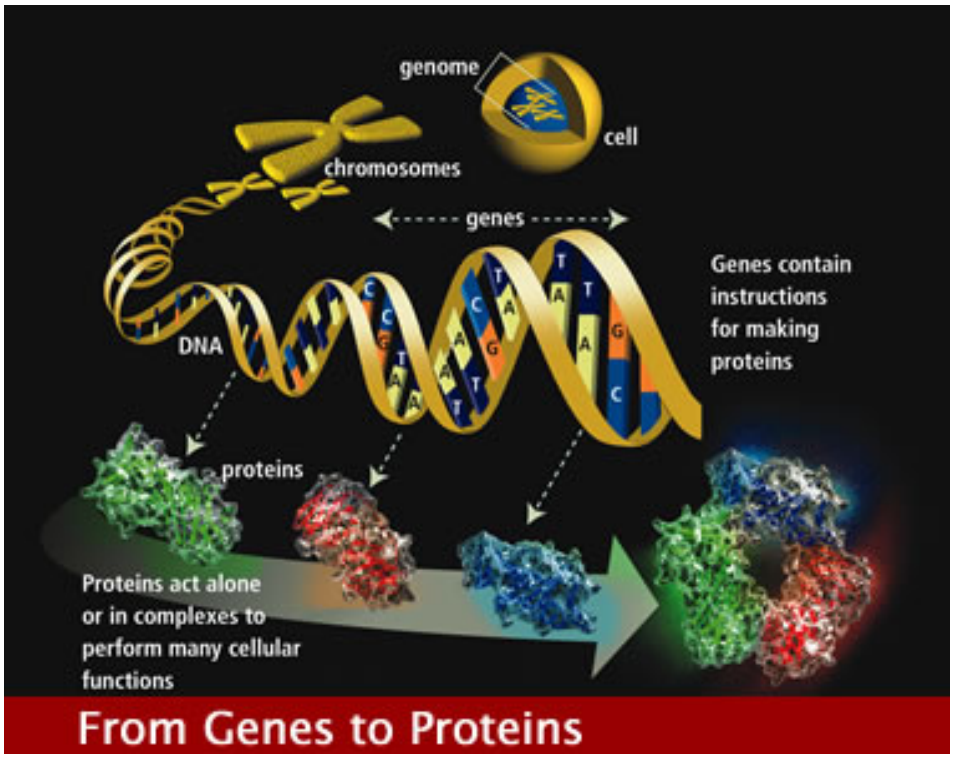
Beyond the Human Genome Project : Genomes to Life

The DNA sequences generated in hundreds of genome projects now provide scientists with the "parts lists" containing instructions for how an organism builds, operates, maintains, and reproduces itself while responding to various environmental conditions. But we still have very little knowledge of how cells use this information to "come alive." The functions of most genes remain unknown. Nor do we understand how genes and the proteins they encode interact with each other and with the environment. If we are to realize the potential of the genome projects, with far-ranging applications to such diverse fields as medicine, energy, and the environment, we must obtain this new level of knowledge.

One of the greatest impacts of having whole-genome sequences and powerful new genomic technologies may be an entirely new approach to conducting biological research. In the past, researchers studied one or a few genes or proteins at a time. Because life doesn't operate in such isolation, this inherently provided incomplete—and often inaccurate—views. Researchers now can approach questions systematically and on a much grander scale. They can study all the genes expressed in a particular environment or all the gene products in a specific tissue, organ, or tumor. Other analyses will focus on how tens of thousands of genes and proteins work together in interconnected networks to orchestrate the chemistry of life—a new field called "systems biology".

The Genomes to Life (GTL) program of the U.S. Department of Energy (DOE) builds on these successes by combining DNA sequence data with advanced technologies to explore the amazingly diverse natural capabilities of microbes—the invisible organisms that thrive in every known environment on earth. The ultimate goal is to understand and use their diverse functions to meet critical DOE mission challenges in energy security, global climate change, and toxic waste cleanup.

Although we now have the entire genome sequences for hundreds of microbes, we still have very little understanding of how the information in DNA creates, sustains, and reproduces living systems. Obtaining this knowledge, a critical first step in harnessing microbial functions, requires a comprehensive approach extending from individual cells to many cells functioning in communities. Such studies must encompass proteins, multimolecular assemblies (sometimes called "molecular machines"; see figure) of components that work together, the intricate labyrinth of pathways and networks in which they interact, and cells. The wealth of data to be collected must be assimilated, understood, and modeled on the scale and complexity of real living systems and processes.



Just as DNA sequencing capability was completely inadequate at the beginning of the HGP, the quantity and complexity of data that must be collected and analyzed for systems biology research far exceed current capabilities and capacities. Dozens of advanced large-scale technologies and approaches must be developed, with mathematics and computing guiding the research questions and interpretation at every step. Computational tools must manage and integrate the data into mechanistic models that describe how cells work. These studies eventually will enable an integrated and predictive understanding of how living cells function and respond to environmental changes, opening the door to using microbial capabilities. Several examples follow:

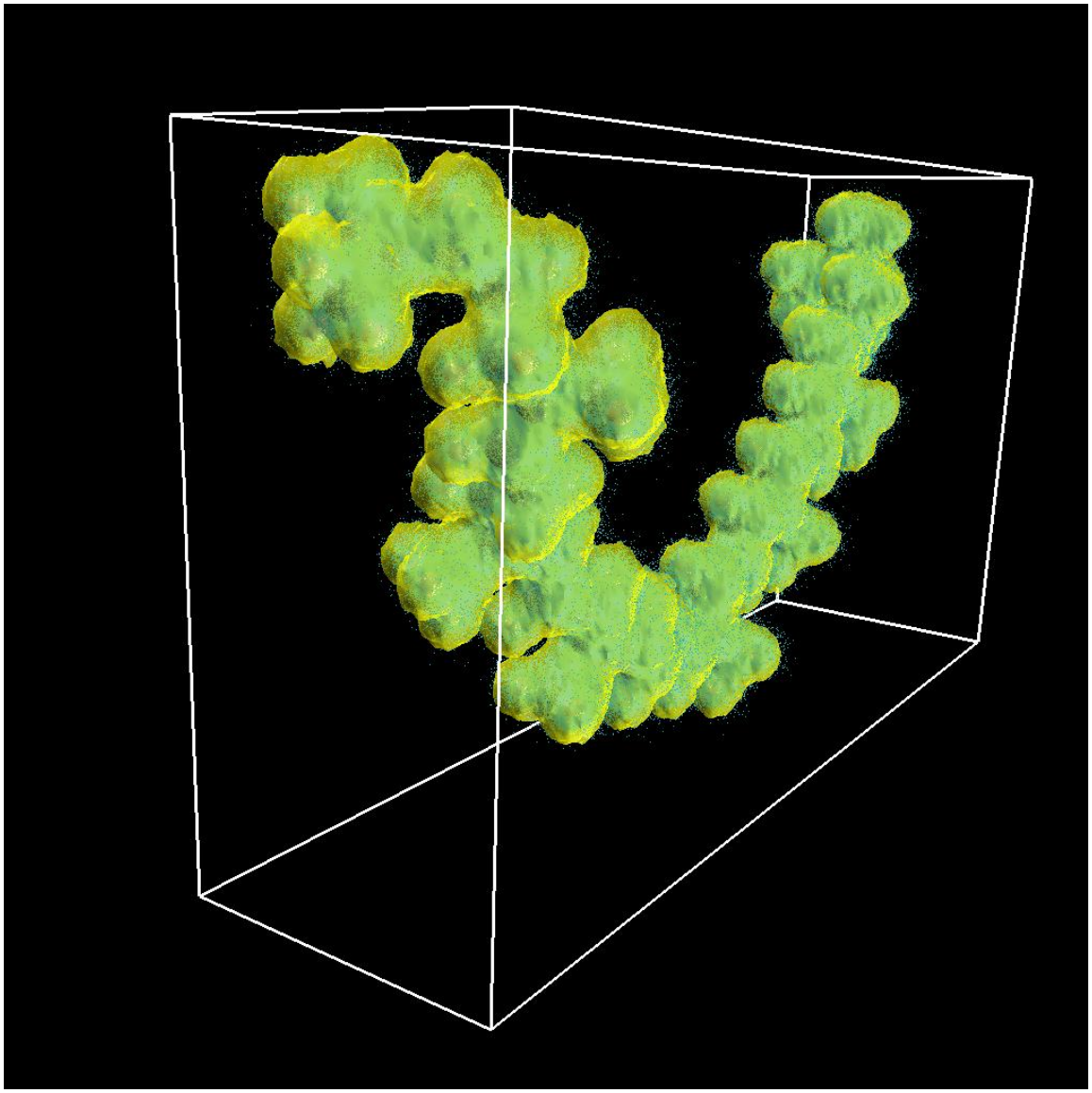
Quantum Monte Carlo Study of Photoprotection via Carotenoids in Photosynthetic Centers

This project aims to increase understanding of the complex processes which occur during photosynthesis, the process by which plants and bacteria convert the sun's light into energy, taking in carbon dioxide and producing oxygen in the process. This project is important on several levels. First, plants and bacteria are the world's foremost means of "carbon sequestration," or storing carbon from the atmosphere — a process which has enormous implications for climate change and global warming. Additionally, photosynthesis is an example of fundamental electron chemistry and is an efficient energy transfer system — processes which are fundamental in many areas of scientific research. The "Monte Carlo" in the title refers to simulations in which data are obtained by simulating a statistical model in which all parameters are numerically specified.

After the project was under way, it was determined that the calculation could be made much more efficient by expressing the wave functions in a less dense representation and by using a Slater basis. In order to accomplish the latter, NERSC has acquired and installed the ADF software package for first-principles electronic structure calculations on NERSC's IBM SP. As a result, the calculation is more capable of scaling up to large systems involving hundreds of electrons. A sparse representation of the wave function was accomplished via a grid acceleration technique. The improvements yield a 16-fold reduction in wallclock time for one of the systems of interest (the spheroidene molecule shown in the figure).

The goal of the project was to compute the excitation energies of the spheroidene and bacteriochlorophyll molecules and to use the obtained electron densities to estimate a rate of energy transfer between the two molecules. The rate of energy transfer as well as further analysis of the results using analysis tools such as the electron-pair localization function was possible due to the storage capabilities at NERSC: configurations from the Monte Carlo random walk were stored for analsys.

This project led to at least an order of magnitude improvement on the efficiency of the Zori quantum Monte Carlo code developed at UC Berkeley.



Quantum Monte Carlo electron density of Spheroidene. This molecule is responsible for photoprotection in photosynthetic reaction centers. This is the largest biological molecule ever treated using the accurate quantum Monte Carlo method.

Molecular Dynameomics

This project combines molecular dynamics and proteomics to create an extensive repository of the molecular dynamics structures for protein folds, including the unfolding pathways. There are approximately 1,130 known, non-redundant protein folds, of which about 30 have been simulated. They plan to use the information from these simulations to improve algorithms for predicting protein structure. Structure prediction remains one of the elusive goals of protein chemistry. It is necessary to successfully predict native states of proteins, in order to translate the current deluge of genomic information into a form appropriate for better functional identification of proteins and drug design.

In summary, the future needs of GTL will require ever greater facilities for the analysis of complex biological systems and the integration of these technologies with advanced computational resources.