

Grand Challenges in Computational Structural and Systems Biology (DA Dixon, TP Straatsma, T. Head-Gordon – PNNL & LBNL)

The biological sciences have experienced tremendous growth over the past five decades. For the first time there is the potential to understand living organisms as complex dynamic systems and to use large-scale computation to simulate their behavior in a variety of new ways thought impossible just a few years ago. Modeling multiple levels of biological complexity is well beyond even the next generation of supercomputers, but each increment in the computing infrastructure makes it possible to move up the biological complexity ladder and tackle problems that could not be solved previously. Biological processes occur on a wide range of spatial and temporal scales. The time scales of biological function range from very fast femtosecond molecular motions, to multi second protein folding pathways, to cell cycle and development processes that take place over the order of minutes, hours and days. Similarly, the dimensions of biological interest range from small organic molecules to multi-protein complexes to cellular processes to tissues to the interaction of human populations with the environment. The linking of biological phenomena at all levels of temporal and spatial scale is leading to the transformation of the separate areas of biology to that of a “systems biology”. The new science of systems biology is based on the ability to “read” and understand the complexity of biology beginning with genome sequences and other sources of high throughput data including global experimental strategies coupled with a detailed understanding of the behavior of proteins individually and in complexes. We are making important progress in developing new scientific methods and technologies that integrate physical, chemical and biological approaches with information science, mathematics, and computational science. However, these developments can only lead to scientific breakthroughs if accompanied by dramatic changes in our access to high performance computing resources.

Societal Needs for High Performance Computing Advances in Computational Biology

All biological processes are eventually determined by molecular interactions and conformational changes. However, these processes typically involve a complex of a large number of individual molecules interacting with assemblies of very large molecules. At the smallest spatial scale, we can evaluate how small to medium-size molecules can interact with proteins to lead to cell signaling events and how signals are amplified in cells. Examples include the role of neurotoxins in protein coagulation leading to swelling in neurons, protein phosphorylation reactions important to signal amplification, and the way that protein switches work. However, a detailed understanding of cell signaling pathways, such as the epidermal growth factor (EGF) or tumor necrosis factor (TNF α) pathways that control cell differentiation, growth, and death and inflammatory response requires that complex protein-protein interactions be studied. How complex is such a signaling process? Activation of the EGF signaling pathway is under control of growth factors that bind to a receptor site on the exterior of a cell. Binding of the receptor initiates a cascade of protein conformational changes through the cell membrane, involving a complex rearrangement of many different proteins, including Ras. Ras is a molecular switch that, upon binding of other proteins such as Raf, initiates a cascade of protein kinases that transfer the external signal to the cell nucleus where it controls cell proliferation and differentiation. Disruption of this signaling pathway can have dire consequences as illustrated by the finding that mutations of the Ras enzyme have been found in 30% of human tumors. A thorough understanding of this pathway is of crucial importance for the development of cancer treatments as well as understanding how cancer is initiated. Because computer simulations can provide atomic level detail that is difficult or impossible to obtain from experimental studies, computational studies are essential. However, this requires the modeling of an extremely large complex of biomolecules, including bilayer lipid membranes, transmembrane proteins, and a complex of many intercellular kinases, and, of course, all of thousands of waters of solvation. Another example of the need to simulate very large molecular systems is in biogeochemistry. The understanding of the role of gram negative bacteria in moderating subsurface reduction/oxidation chemistry and the role of such systems in bioremediation technologies requires that we study how cell walls, including many trans-membrane protein substituents, interact with extracellular mineral surfaces and solvated atomic and molecular species in the environment. These are two of many examples that can be given of systems that require many millions of atoms to be included in our simulations.

In addition to the general areas of health and environment mentioned above, computational biology techniques can play an important role in countering bioterrorism. Since these methods are used to determine structure-function relationships in proteins in order to understand the biological pathways, it also provides the tools to study the factors that lead to toxin formation and the interruption of such pathways, either for detection, prevention, remediation or health impacts. Computational approaches can be used for the rational redesign of enzymes to degrade chemical agents. An example is the enzyme phosphotriesterase (PTE), which could be used to degrade nerve gases. Combined experimental and computational efforts can be used to develop a series of highly specific PTE analogues, redesigned for optimum activity at specific temperatures or for optimum stability and activity in non-aqueous and low humidity environments or in foams, for improved degradation of warfare neurotoxins. It is also possible to use advanced computations to design better reactivators of the enzyme acetylcholinesterase (AChE) that can be used as more efficient therapeutic agents against the highly toxic phosphoester compounds such as the nerve warfare agents DFP, sarin, and soman and insecticides like paraoxon. AChE is a key protein in the hydrolysis of acetylcholine and inhibition of AChE through a phosphorylation reaction with such phosphoesters can rapidly lead to severe intoxication and death.

The Goal of Quantitative Simulations of Large Molecular Systems

An enormous range of computational methods is required for adequate biological simulations. The enormous difficulty of using information from small model systems to address complex, collective phenomena at large scales, requires significant advances in theoretical methods, algorithms and software, and computing hardware. At one extreme are quantum chemical and molecular dynamics approaches, which predict the energetics and motions of biological systems at the atomic scale. These methods are currently limited by their need for extremely high speed arithmetic calculations; hence, the enormous improvements in computer performance provided by the next generation of massively parallel supercomputers will greatly extend the applicability of the such atomic-level biological simulations. At a higher scale of system complexity and size are methods that combine experimental gene and protein sequence and structure data, heuristic methods and physical simulations to predict the structure and function of proteins. Such methods offer tremendous promise for increasing the value of structural biology by predicting protein structures based on their similarity to known folds and identifying novel proteins for experimental characterization. Managing the huge databases of experimental biological data will also require tremendous advances in computer memory and speeds. Beyond the level of individual genes and proteins, simulations have a critical role in understanding the complex biological processes such as passage of ions through channels across cellular membranes, multi-spectral analysis of cellular signaling processes, and metabolic networks. Such simulations will involve "mesoscale" models that include three dimensional continuum transport and chemical processing of ions and signal molecules. These simulations are computationally complex, but will be critical because quantitative descriptions of these processes will be required both to complete our understanding of the functions of the living world and for many promising future applications. In many of these cases, appropriate statistical accuracy requires that many extended time simulations need to be carried out on a single, large, complex system. With appropriate implementation such simulations can be efficiently carried out on very high performance computers.

The enormous complexity of biological systems and the difficulty of using information from small model systems to address complex, collective phenomena at large scales, requires significant advances in theories, algorithms, software, and hardware. Currently available computing resources, allow computer simulations of biomolecular systems to be routinely carried out for about 100 thousand atoms for tens of nanoseconds. Computer resources will need to increase in power by at least three orders of magnitude in order to be able to allow microsecond simulations of systems with several million atoms. This is an ambitious goal, but absolutely necessary if we want to make a significant impact on the nation's key scientific problems in health, environmental protection, and national security.