Biosciences Area LDRD FY18 Foci

1. Biological Dark Matter

Labwide with computing sciences for mathematical modeling and simulation, algorithm design, data storage, management and analysis, computer system architecture and high-performance software implementation.

Metagenomics and single-cell sequencing have enabled, for the first time, glimpses into the vast metabolic potential of Earth's collective biological systems. Yet, for the most part we can't accurately predict nor identify the products of most biosynthetic pathways. Most of what we know of microbial biochemistry is based on characterization of a few model microorganisms, and these findings have been extended through sequence correlations to the rest of sequence space. Unfortunately, these extrapolations have questionable validity for the vast majority of environmental microbes and therefore requires fundamentally different approaches for directly linking novel sequences to their biochemical functions.

Because of the sensitivity of MS, tandem mass spectra are often the first (or only) data obtained on unknown compounds in complex samples. Tandem mass spectrometry involves fragmentation of analyte-derived ions via a sudden infusion of energy into the molecule. Mass spectrometry (MS) can detect hundreds of compounds at high-sensitivity from complex mixtures, and mass spectra are often the first data available on unknown samples in microbial ecology, metabolomics, as well as analysis of metabolism for synthetic biology. Although the metabolome has been an extensive target of study since the dawn of biochemistry, the number of natural products on Earth is unknown; likewise, methods for definitively identifying them are lacking. Ideal approaches for the exhaustive mapping of relationships within and between molecules and linking these back to sequence space are likely not readily achievable with today's computing technology.

2. Modeling the Biology: Environment interface

We seek proposals that pair advanced computing with environmental and biological data for advances in mechanistically informative predictive modeling (mechanistic-predictive modeling). Examples include (1) biological data science that merges and models information across scales and modalities for predictive modeling of molecular, cellular, and population behavior; (2) interpretable learning and adaptive database architectures: knowledge-engines for discovery science from multi-modal environmental data; (3) molecular modeling of the soil-water interface. An area of particular interest is the integration of hyperspectral and videographic imaging data with omics and molecular modalities.

Computational infrastructure has been vital to the biological and environmental sciences in the 21st century, particularly in genomics and climate science. However, next-generation ecological and biological experiments leverage more complex, multi-modal and time-course measurement strategies, often spanning molecular level "omics" data and large-scale environmental information. Examples include large-scale data on plant, metazoan and soil systems from high-throughput phenotyping, hyperspectral data, and ecological surveys, through to metabolomic and transcriptomic studies of individual cells and organisms. The fusion of diverse, spatiotemporally resolved data streams yields an unprecedented opportunity to derive high-precision predictive-mechanistic models of biological and ecological systems. However, many roadblocks exist to achieving this goal - we need new, flexible and adaptable data structures, storage paradigms and analysis methods, as current approaches typically assume data is coming from a single instrument or assay type, and/or rely on unrealistic regression or network modeling strategies. The capacity to accurately quantify uncertainty and misspecification in non-ergodic settings will be transformative for climate modeling – where currently the most trusted models fail to agree on the sign of terrestrial C02 compositions over the next fifty years. Such techniques are equally applicable to molecular information collected for individual cells, and by constructing flexible and general frameworks for data organization and analysis, we have an opportunity to make radically better use of emerging exascale architectures than will otherwise be possible. Finally, we will need collaboration to make this work, to connect researchers across different silos, and to train them in methods that make their data more reusable, and to make maximal use of community data.

3. The Molecules-to-Minds Program

The BRAIN Initiative aims to accelerate neuroscience through interdisciplinary technology development. LBL has made significant investment into the BRAIN Initiative through development of next generation devices for electrical and optical measurement/manipulation and advanced data analysis tools. Most recently, these activities have been focused on a single biological platform for accelerating the development of these advanced neurotechnologies and to experimentally begin to determine the brain's functional connectome.

1. The **Molecules-to-Minds (**M2M) program will strengthen the LBNL contributions to the national BRAIN initiative, the DoE role in that initiative, and the BER mission in environment, with important spin-offs for neuroscience and mental health. The scientific goals of M2M are to:
2. Understand brain function in model animal systems across multiple temporal and spatial scales, from "molecules to minds".
3. To determine mechanisms to enhance brain and behavioral bio-resilience in complex ecosystems, and
4. Uncover sources of inter-individual variations in neurological function, behaviors, and risks for disruption due to exposure to environmental stressors (such as toxicants found in synthetic biological processes and energy byproducts), and neuro-modulating host microbiomes.

M2M builds on long-standing strengths of BSA and ongoing collaborations in genetically defined animal model systems, molecular analyses, cellular imaging, behavioral analyses, microbiomes in soil communities and animal hosts, synthetic biology, environmental exposure biology, and DOE facilities such as the Joint Genome Institute, NERSC, and the Molecular Foundry.

4. **Molecular to Mesoscale Analysis:**

To identify the principles driving biological systems of microbes, plants and multi-species communities relevant to bioenergy and the environment new approaches are needed that integrate measurements across length and time scales. With these principles in hand it will be possible to predict the consequences of systems biodesign. The integration of information from disparate sources, including genomics, functional genomics, structural biology, and imaging will require new computational methods that both integrate information and provide the means for automated knowledge generation, and ultimately accurate prediction of biological systems.