Exploration of Collaborative Opportunities across SciDAC Institutes and Genomic Science Program

Paul Adams (LBNL), Leonid Oliker (LBNL)

Rommie Amaro (UCSD), William Cannon (PNNL), Gloria Coruzzi (NYU), Lori Diachin (LLNL), Roger Ghanem (USC), Jeff Hollingsworth (UMD), Paul Hovland (ANL), Costas Maranas (PSU),Todd Munson (ANL), Habib Najm (SNL), Esmond Ng (LBNL), Robert Ross (ANL), Oliver Ruebel (LBNL), Arie Shoshani (LBNL), Jeremy Smith (ORNL), Kranthi Varala (NYU)

1. Introduction

As part of the DOE/BER Genomic Sciences Program Annual Principal Investigator meeting, a joint session was held to bring together biosciences researchers and representatives from the ASCR SciDAC Institutes on February 7, 2017. The session began with presentations from each of the four SciDAC3 Institutes, describing their goals and examples of impacts on research projects across numerous computational domains. These talks were followed by five talks given by bioscientists, highlighting areas of interest to BER/BSSD: genomics, metabolic modeling and biosystems design, multiscale modeling, and biological simulation. The session ended with a panel discussion about potential collaborations between the bioscientists and computational scientists. Short summaries of the biological presentations are given below along with potential connections to the SciDAC program that were identified in the panel discussion. Overall we believe there are numerous opportunities to form effective collaborations between the BSSD and SciDAC communities.

2. Metabolic Modelling and Biosystems Design

Costas Maranas (Penn State University) gave a talk titled “Computational Bottlenecks in Metabolic Networks and Protein Design”. Several types of computational challenges were discussed. There are limits on current approaches to metabolic modeling that result from limited interoperability between metabolic models, databases, and ability to trace atoms through reactions. As researchers attempt to create more realistic models, tractability challenges are encountered in computations arising from spatially and temporally distributed metabolism in microbial communities and plants in the environment. It remains challenging to integrate heterogeneous datasets of metabolite concentrations, metabolic fluxes and omics data within mechanistically accurate models of metabolism. Prospecting both existing and hypothetical pathways from feedstocks towards biofuels and biorenewable products rapidly encounters combinatorial explosion. At the atomic level, de novo enzyme design for improved/altered substrate specificity is hampered by current search algorithms and energy functions.

2.1 Potential SciDAC Connections

The discovery of network models based on data is generally a difficult discrete optimization problem along with classical inverse problem challenges of non-uniqueness and ill-posedness.

Some of these challenges have been dealt with in SciDAC in a continuous setting by introducing regularization, including the use of Bayesian priors in the statistical inverse problem formulation, and the use of L0 and L1 regularization to identify sparse solutions. The Bayesian formulation provides also well-grounded handling of multiple heterogeneous noisy data sources, as well as information-theoretic means for control of graph complexity and avoidance of overfitting. Extending these methods to a discrete setting is an opportunity. Analysis and methods for systems of ordinary and partial differential equations with nonsmooth or discontinuous right-hand sides are also relevant, where the connections among the equations are represented by a graph. DOE-funded applied mathematics research in solving discrete numerical optimization problems (e.g. advances in mixed-integer linear and nonlinear optimization methods in conjunction with ordinary or partial differential equation constraints) and heuristics are relevant to this area along with work in robust estimation and optimization that accounts for certain types of uncertainty. Bilevel optimization problems and non-cooperative games, possibly with discrete variables, appears in strain design and when modeling communities. The bilevel problems resemble models of sensor placement in adversarial networks. Multiobjective methods where the Pareto surface needs to be explored are also relevant to this area. SciDAC Institutes can provide expertise in differential equations and numerical optimization (FASTMath); multiobjective methods and performance optimization (SUPER); uncertainty quantification, model selection, statistical inversion, and stochastic optimization (QUEST); and storage of large amounts of data including provenance information from computational experiments and visualization of the large graphs and their evolution in time (SDAV).

3. Genomics

Kranthi Varala (NYU) gave a talk titled “EvoNet: A Phylogenomic and Systems Biology Approach to Identify Genes Underlying Plant Survival in Marginal, Low-Nitrogen Soils”. The presentation described the application of phylogenomics (reconstruction of species trees using all possible gene sequences) with the Phylogeneious pipeline to identify genes that underlie phenotypes. This pipeline makes use of a combination of gene clustering, sequence alignment, gene family tree construction and orthology assignment to enrich for genes that define the identified phenotype. Current calculations with plant genomes encompassing 70 taxa take approximately 1.5 calendar months. Future calculations will increase the number of species to 90, with the ultimate goal of analyzing 1,400 species to determine the genes that have been critical to the emergence of the major land plant lineages. Some strategies to improving computational throughput were discussed, including migration of the code to a massively parallel computing environment.

3.1 Potential SciDAC Connections

The same above SciDAC connections pertaining to discrete optimization methods, multiobjective methods, inference/estimation of graphs based on data, including L0 and L1 regularization, compressive sensing, statistical inversion, control of complexity, and handling of heterogeneous data, can also relevant here. Further, the inference of graphs is a major field in machine learning, with the availability of statistically-based methods for graph model selection. Analysis pipelines and workflows and data analysis methods, including gene context analysis, are also relevant to this area. SciDAC Institutes can provide expertise in differential equations and numerical optimization (FASTMath); multiobjective methods and performance optimization

(SUPER); uncertainty quantification, statistical learning, and Bayesian inference (QUEST); and analysis pipelines and data analysis methods (SDAV).

4. Multiscale Modelling for Microbiomes

Bill Cannon (PNNL) gave a talk titled “Some Challenges in Multiscale Modeling: Molecules to Microbiomes”. BER/BSSD emphasizes the development and application of multiscale approaches. Examples were presented of work in progress or planned that spans time and length scales to link macroscopic observations to the underlying molecular or cellular components. Going from the molecular to the cellular, the Multiscale Model of Circadian Rhythms project takes enzymatic reactions and propagates them using statistical thermodynamics methods to replicate oscillatory dynamics. In practice, challenges are encountered because the dynamic models need to expanded or reduced in complexity depending on the questions being asked. The use of relative time in the thermodynamic state functions leads to nonlinear effects and the ODEs can become very stiff, and reliable nonlinear solvers are needed. In the Modeling Microbiomes Across Scales project the goal is to simulate from cells to the microbiome. This requires multi-scale, multi-physics models that are currently of insufficient flexibility. Multiple challenges are encountered including the lack of parallel ODE/PDE solvers and uncertainty quantification for the model of the interaction network. It was argued that inference from data and simulation of models should each use the same mathematical frameworks.

4.1 Potential SciDAC Connections

There are recent developments on the analysis and reduction of stiff reaction network models, using computational singular perturbation theory, including accounting for uncertainty, that can be of some utility here. Further, the intimate coupling of data and computations, including both top-down and bottom-up modeling, is strongly connected to both inverse and forward uncertainty quantification. Moreover, the exploration of model space can benefit from recent developments on the Bayesian modeling and estimation of model structural error. Similarly, there are opportunities for using Bayesian and information-theoretic methods for informing model selection, targeting both fitness skill and optimal complexity, in the overall context of physical model growth for representing observable data. The DOE-funded applied math community has a great deal of expertise in solving (deterministic and stochastic) ordinary and partial differential equations and there are opportunities for close collaborations in dealing with some of the issues in multi-scale, multi-physics modeling. In particular, if the models involve electronic and atomic scales, the experience and work from the recent SciDAC Partnership Projects in BES will be relevant. This area also includes large graph network layouts describing the interactions between the differential equations, which can be modeled as differential equations with nonsmooth or discontinuous right-hand sides. The development of reliable nonlinear numerical optimization solvers is another DOE-funded activity relevant to this area and the inference of models from data include the previously mentioned topics in discrete optimization, multilevel methods, and estimation. Additionally, the institutes have broad experience in identifying the performance bottlenecks for these classes of computations and developing high-performance scalable solutions on DOE HPC systems.

5. Multiscale Modelling at the Cellular Level

Rommie Amaro (UCSD) gave a talk title "Multi-scale Dynamics: Molecules to Cells". The extraordinary increase in computing power available makes it possible to simulate biological systems at ever increasing size and time. It is now possible to perform atomic simulations of whole viruses (with more than 200 million atoms) for up to 100 microseconds. At the same time experimental techniques are able to probe cells and molecules at increasing resolution leading to datasets with trillions of pixels containing hundreds of thousands of molecules. It is conceivable that the experimental data will be available to locate individual molecules in cells and place those cells in the context of tissues. The application of high performance computing may ultimately make it possible to understand the molecular and chemical mechanisms underlying disease through simulation. Examples were provided of how simulation of single proteins at the atomic level can lead to insights into new drug targets. Some of the emerging tools were described for building higher order biological objects, such as viruses and cells, from the underlying molecules, and how simulation can be used to link these to experimental observations. Finally there was a discussion of a broad set of improvements that will be needed to fully exploit simulation in biology, including better data integration, data storage and accessibility, and multiscale algorithms.

5.1 Potential SciDAC Connections

The previous discussion of multiscale and multiphysics methods is relevant to this area and are needed to investigate dynamics across scales. Eigenvalue problems are also encountered in this area. Tomographic reconstruction problems from the experimental data sets can be posed as numerical optimization problems that need to be solved. Large scale data sets need to be stored and retrieved efficiently (on the order of 1.2 trillion pixels) and visualization methods for simulation with 200 million atoms are important. Load balancing and performance optimization becomes a significant issue in these simulations. SciDAC Institutes can provide expertise in multiscale methods and eigenvalue problems (FASTMath), performance optimization (SUPER), model fitting/selection (QUEST), and data management and visualization (SDAV).

6. Biological Simulation

Jeremy Smith (ORNL/UT) gave a talk titled "Exascale Concepts In Biological Simulation". The final presentation began with a discussion of the opportunities to be had from combining genomic, and structural information to predict macromolecular function. An example was given of combining experimental data and simulation to understand the cosolvent pretreatment of lignocellulosic biomass. Different imaging and measurement technologies probe different lengths and time scales leading to multiscale experimental and computational solutions. One particular area of promise is the combination of simulation methods with experimental data collection facilities so that real time feedback can be provided from computation to help guide the experiment. This would require extensive improvements in the scale of codes on supercomputers, real time analysis, better uncertainty quantification, improved load balancing, and methods for analysis of very large eigenvalue problems.

6.1 Potential SciDAC Connections

There are opportunities for using uncertainty quantification and numerical optimization methods in the context of model-experiment comparisons, including associated model fitting, selection, and validation challenges, and the use of optimal experimental design strategies for guiding experimental campaigns. Areas of interest to the applied math community are fast partial differential equation solvers and fast integral evaluations. There are also opportunities in eigenvalue calculations. One example is the normal mode analysis. Another example is to understanding the dynamics of state transitions, which will to be very large eigenvalue problems. In the area of imaging (crystallography, solution diffraction), recent work by members of the CAMERA project may be relevant. Management, coordination, and movement of data from the facilities that generate data to the supercomputers that analyze it are also important, as are mechanisms for coordination of complex distributed workflows and performance optimization and load balancing, which will be critical to achieve the scale of computation targeted in this work.

7. Summary

Overall it was found that there is a large intersection between the capabilities of the needs of the computational biology community within BER and the SciDAC Institutes and DOE-funded applied mathematics and computer science research. Building upon these connections would accelerate the pace of the biological research and lead SciDAC Institutes in new directions and capabilities. Further discussion among these communities is warranted to expand upon these connections in finer detail and to identify additional collaborative opportunities that would advance the BER and ASCR research goals in both the short- and long-term.