ASCAC-BERAC Joint Report on Modeling and Simulation for GTL

Rick Stevens and John Wooley

(co-Chairs)

August 6th, 2008

Review of the Charge

The overall charge was to address the issue of computational models for GTL and how progress could be accelerated through targeted investments in applied mathematics, computer science, and computational biology. Specifically, we were asked to address the following questions:

1. Is the current ASCR long-term goal too ambitious given the status and buy-in from the community?

2. What intermediate goals might be more relevant to the two programs?

3. What are the key computational obstacles to developing computer models necessary to characterize and engineer microbes for DOE missions such as biofuels and bioremediation?

Charge Question 1. (PART Goal)

Recommendation 1. *The ten-year OMB PART goal for ASCR and BER for the joint modeling and simulation activity of ASCR and BER should be modified to read as follows:*

(ASCR/BER) By 2018, ***validate capability to predict phenotype from an organism's genome and to predict genotype from an organism's physiology***

This PART goal should be accompanied by a specific set of metrics of progress. Such metrics could include for a given organism *the number of correct metabolic phenotype measurements predicted, the fraction of an organism's genes and gene products included in a model, the number of transcription regulatory elements in a model, the number of correct gene expression experiments predicted, the fraction of correct predictions of essential genes, and the number of organisms for which predictive models can be generated*.

In general we imagine the effort to be joint with the computational infrastructure and tools being supported primarily by ASCR, the experimental work supported primarily by BER and the modeling, methods and simulation work jointly supported.

Charge Question 1. (PART Goal)

Recommendation 2. *DOE should develop an explicit research program aimed at achieving significant progress on the overarching goal of predictive modeling and simulation in DOE relevant biological systems (e.g. organisms and communities). This program should be a joint effort between ASCR and BER and should include a diversity of modeling approaches (e.g. FBA, Boolean Networks, Stochastic ODEs).*

The program should leverage existing experimental activities as well as support the development of new experimental activities that are directly tied to the needs of developing predictive models. This new research program should be aimed at advancing the state of the art of cell modeling directly, should include equal participation from biologists and mathematicians, computer scientists, and engineers; and should be indirectly coupled to the more applied goals of bioenergy, carbon cycle research, or bioremediation.

This program will need to be supported at a large-enough scale that a multiple-target approach can be pursued that will enable progress on many intermediate goals simultaneously by different research groups.

Since the subcommittee has been working DOE has been moving ahead with two relevant workshops; GTL Knowledgebase and the planned Exascale workshop in Biology.

Charge Question 1. (PART Goal)

Recommendation 3. *DOE should establish an annual conference that focuses on highlighting the progress in predictive modeling in biological systems. This should be an open meeting and separate from any programmatic PI meeting.*

One goal of the meeting would be to establish a series of scientific indicators of progress in predictive modeling, similar to successful indicators associated with the competitive assessment of structure prediction (CASP). These types of measures will enable the community to benchmark progress on methods and will be critical to assessing the impact of the research program on fundamentally advancing the state of the art.

Example metrics could include *predicting essentiality in microbial genomes, predicting gene expression patterns in novel environments, predicting flux values through key reactions in microorganisms, or predicting yields in metabolic engineering scenarios.*

The proposed conference would be a cross between CASP and the SciDAC conference, providing an annual opportunity to qualitatively and quantitatively assess scientific and technical progress

Charge Question 2. (Relevant Goals)

Intermediate goals that could be considered more relevant for the two programs fall into two general areas.

TOOLS and INFRASTRUCTURE:

The first area is building needed tools, curated databases, and computational and collaborative infrastructure that directly accelerate the communities' ability to develop models and simulations. Examples of these are tools for curation of genomes and reconstruction of metabolic networks, integrated databases enabling the community to share data needed to build and test models and validation datasets, and mathematical libraries and core model components that would enable many groups to leverage the work of others.

PROBLEMS:

The second area is focusing on a targeted set of biological modeling and simulations problems that build on each other and that over time would expand the modeling capabilities in the appropriate directions. Examples of these are models of cellular metabolism, motility, global transcription regulation and differentiation, and life-cycle development. Each of these models could play a role in advancing toward the overarching goal of a complete cell model that can be used to predict phenotypic traits or behaviors of a cell from genomic and other "omic" data sources.

Charge Question 2. (Relevant Goals)

Recommendation 4. *The GTL modeling and simulation research program should be supported by an explicit series of investments in modeling technology, databases, algorithms, and software infrastructure needed to address the computational challenges.*

The appropriate early targets for a comprehensive attack on predictive biological modeling are specific functions of microbial organisms (e.g., *cellular metabolism, motility, global transcription regulation, differentiation, and life-cycle development*).

The focus should include advancing the predictive skill on well-studied models (e.g., *E. coli, B. subtilis)* but begin to extend to those organisms that stretch the capability beyond the existing well-studied model systems (e.g., Clostridium, Shewanella, Synechocystis) and small consortia (communities) of microorganisms relevant to DOE missions, such as those associated with bioremediation, carbon sequestration, and nitrogen fixation and fermentation and degradation.

We also recommend that the lower eukaryotes (e.g., diatoms, coccolithopores, single-cell fungi) and plants be included as targets in longer-term modeling and simulation goals. Such inclusions will advance the goals of Genomics:GTL by strengthening efforts to integrate the modeling and advancing systems-level and synthetic knowledge for microbes and plants.

An approach to the problem of intermediate goals is to build up from submodels starting with those that are reasonably well developed (e.g. metabolism)to those that are less well developed (e.g. signaling). Each is pursued in a specific model organism and then extended.

Charge Question 3. (Obstacles)

Finding #6. A number of obstacles remain to reaching the visionary goal of a predictive model useful for engineering of an organism derived largely from its genome and related data. Here we describe five of the most relevant ones.

➔ First, we lack integrated genomics databases and the associated computational methods for supporting curation, extension, and visualization of comparative data explicitly focused on supporting the development of modeling and simulations for DOE-relevant organisms.

Second, for current systems-level computational analyses (e.g., flux balance analysis) work is needed to further integrate additional cellular physio-chemical constraints into modeling frameworks in order to generate computational predictions with greater accuracy towards defining the actual physiological state of the cell.

Third, we lack robust mathematical frameworks and software implementing those frameworks for integrating models of metabolism with those of gene regulation that are two of most highly developed areas of modeling and simulation at the whole cell level, but whose mathematical representations are quite different.

Fourth, we lack the multiscale mathematics and associated software libraries and tools for integrating processes in cellular models of disparate scales (e.g., molecular scale to that of the whole cell and microbial community) that would enable the modeling community to begin development of integrated whole-cell-scale models with atomistic simulations of specific mechanisms.

Fifth, we lack a computational and analytical theory for framing all of computational biology. Such a theory should incorporate evolution as the basis for understanding and interpreting the results from comparative analysis. For example, we have not yet developed the algorithms needed to make rapid progress on questions such as understanding the major forces governing the evolution of metabolism and regulatory networks. Understanding these forces will be critical to creating the stable engineered strains needed for large-scale bioproduction of materials.

Charge Question 3. (Obstacles)

Recommendation 5. *DOE should establish a mechanism to support the long-term curation and integration of genomics and related datasets (annotations, metabolic reconstructions, expression data, whole genome screens, etc.) to support biological research in general and specifically the needs of modeling and simulation in particular in areas of energy and the environment that are not well supported by NSF and NIH.*

This mechanism should target the creation of a state-of-the-art community resource for data of all forms that are relevant to organisms of interest to DOE. This should be a joint activity of ASCR and BER, with ASCR responsible for the database and computational infrastructure to enable community annotation and data sharing. It should also leverage the work of established groups.

N.B. The GTL Knowledgebase Initiative is addressing this issue

Charge Question 3. (Obstacles)

Recommendation 6. *DOE should work with the community to identify novel scientific opportunities for connecting modeling and simulation at the organism level to modeling and simulation at other space and temporal scales. (atomistic to the global scale)*

Examples that could be investigated include integration of microbial models into ocean and terrestrial ecology models which in turn are coupled to global climate models, and models of bioremediation environments that can couple organism metabolic capabilities to external biogeochemistry. This multiscale coupling in beginning to be explored, but much more can be done, and it is likely to yield significant scientific insight.

It is expected that planned Exascale Advanced Modeling and Simulation Workshop will address the issues of mathematical frameworks and multiscale modeling and simulation

Program Scope and Funding Levels

July Revised Draft Report ASCR/BERAC Subcommittee on Modeling and GTL

Level ($M)	Recommended Action (Assume 50% split among BER, ASCR)	Impact
5	Fund a GTL KB Coordination Center; w a cell-centric GTL DB & Rich Portal Environment; Coordinate bioIT at Centers" = Focus on Common Goals /Needs among Centers Choose max 3 pilots; Coordinate with other databases. KBCC Distribute Mature Modules to 3 BioEnergy Centers Ontology; key Interactions with other KBs, major resources Also $0.6M to each from CC to provide BioIT support, interface	Note: BER funding only; just initiates knowledgebase, for minimum GTL effort. NO Implementation of the Joint AC Study Findings, ONLY GTL KB.
10	Expand CC; advance Core Knowledge Environment, include all GTL projects; flexible approach in case additional GTL Centers exist Add more pilot projects as part of CC and begin some modeling KB better linked to other major knowledge resources; accelerate Community Collaborations on tool development (improve use of KB) Ensure all GTL projects have strong internal bioIT. Expand Viz & Remote Collaboratory Tools; Run broad Annual Meeting	Equal funding BER, ASCAC; Allows more complete GTL KB; Begin to build community tools. Ensure robust GTK KB with ASCR research contributions.
30	Grow Algorithm and Software support across community rapidly. Provide high level incentives, maximize funding for collaborations among experimentalists and quantitative scientists. Seek joint solicitations w other agencies to extend additional Expand annual meeting to validation/CASP-like process	Implement initial modeling goals; accelerate GTL through extensive agency partnerships. Validation will also increase the buy-in by wet lab science.
50	Full implement of the opportunities in this the joint report. Robust individual and group efforts in algorithm and software Validation of Packages and deep engagement with experimentalists and modeling groups; Interplay to accelerate progress, outcomes of GTL KB science to serve society	Community able to integrate experimental work, computing / modeling and simulation, feed back to web lab and back to modeling for full impact

Impact and Funding Levels

- 30M/yr would provide a significant implementation of the goals outlined in this report, and enable partnerships with other agencies in order to broaden support for software for systems and synthetic biology; contributions in basic biology by other agencies, facilitated by DOE, would help GTL underpin DOE applied mission needs in biology.

- 50M/yr would allow full delivery for DOE of the opportunities for GTL from modeling; this would include intensive collaborative modeling and simulation efforts, and create a major opportunity for interactions to accelerate discovery, minimize false or inferior directions, by interactions among experimentalists and quantitative scientists for all DOE organisms and for potentially any project within GTL. At this level of funding, DOE Office of Science can play a leadership role and significantly catalyze the growth of GTL and its impact on the societal goals of bioremediation, carbon sequestration and implementation of bioenergy options.