World Technology Evaluation Center, Inc.

NSF Workshop on

QUANTITATIVE SYSTEMS BIOTECHNOLOGY

September 13-14, 2000

Report & Proceedings

http://www.wtec.org/qsb

World Technology Evaluation Center, Inc.
R.D. Shelton, President

2601 N. Howard St.
Baltimore, MD 21218655

Table of Contents

This document was sponsored by the National Science Foundation (NSF) of the U.S. government under NSF Award No. BES-0004066, awarded to the International Technology Research Institute (ITRI, Inc.). The government has certain rights to this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States government, the authors' parent institutions, or ITRI, Inc.

I. EXECUTIVE SUMMARY

INTRODUCTION

On September 13 and 14, 2000, the Engineering Directorate of the National Science Foundation convened a workshop on "quantitative systems biotechnology" (QSB). For the purposes of the workshop, QSB was defined as engineering research to augment the process of predicting the phenotypic behavior of a living organism from the genomic information being generated for that organism and the environmental conditions that influence the expression of that genome. The objectives of the workshop were to provide detailed suggestions for the scope of a new NSF/ENG Program Solicitation (http://www.nsf.gov/cgi-bin/getpub?nsf0137) on QSB, and to examine the possible case for conducting a World Technology Evaluation Center (WTEC) international benchmarking study in this area.

Leading U.S. experts in the fields of biotechnology, chemical engineering, biochemical engineering, experimental biology, computational biology, molecular biology, systems analysis, bioinformatics, metabolic pathway analysis, theoretical biology, and genetics were invited to make presentations on recent research and development activities related to the forthcoming NSF QSB initiative, and to participate in extensive open discussion on the possible foci and boundaries of the initiative.

In addition to the invited researchers active in the field, participants included many NSF staff members, including representatives from the Bioengineering and Environmental Systems (BES), Electrical and Communications Systems (ECS), Chemical and Thermal Systems (CTS), and Engineering Education and Centers (EEC) divisions of NSF's Directorate for Engineering, as well as from the Computer and Information Science and Engineering (CISE) Directorate and the Biological Sciences Directorate at NSF. The National Institute of Standards and Technology was also represented. WTEC provided logistical and reporting support. A complete list of participants in the workshop is included as Appendix B to this report.

Presentations in the first day of the workshop were divided into sessions on (1) opportunities and unmet needs, (2) needs for test beds and screening systems, and (3) analysis needs. The morning of the second day featured breakout sessions on each of the above topics, during which participants were asked to identify specific topics that might be included in (or excluded from) the NSF/ENG initiative.

CONCLUSIONS

The primary recommendation of the workshop is that this initiative should focus on developing a systems framework to more efficiently relate genome-level data to phenotype. Phenotype in the broadest sense results from the interaction of thousands of genes as well as environmental factors, and so is best understood from a systems approach. The availability of new genome-level data (array expression technology, proteomics, etc.) provides new opportunities for systems-based studies of biological problems. The construction of this framework will require the collaborative input of bioengineers, systems engineers, and biologists. Development of a framework will require both development of "*in silico*" (or modeling) methods as well as the engineering basis for new experimental methods to obtain genome-level data.

The workshop demonstrated the need for novel mathematical modeling methods that facilitate the integration of diverse pieces of genome-scale and metabolic information into functioning cellular networks. These methods should assist in the interpretation of genome-scale data and aid the conception of biotechnological experiments that elucidate the design of cellular systems. The modeling methods may focus on one or more levels of cellular organization. They may address the interactions among genetic and metabolic components and, in particular, emphasize processes involving DNA, RNA, proteins, and metabolites. They may also integrate cellular components into larger, multi-scale systems within cells, such as gene regulation networks, assemblages of metabolic pathways, or the coordination of genetic and metabolic processes. The envisioned models should be amenable to efficient mathematical analysis, simulation, and optimization and allow the quantitative investigation of system responses to external stimuli.

With regard to the possibility of conducting a WTEC international benchmarking study in this area, there was a belief among workshop participants that there are centers abroad that are ahead of the U.S. in such areas as proteomics (Switzerland, Australia and Japan) and modeling (U.K. and Netherlands). A benchmarking study might thus be useful.

II. BREAKOUT SESSION SUMMARIES

INTRODUCTION

As was the case with the detailed presentations on September 13, the breakout sessions on the morning of September 14 were divided up into three topical sub-areas: (1) opportunities and unmet needs (overview session); (2) needs for testbeds and screening systems; and (3) analysis needs. Each group was given a set of questions to address and discuss. Summaries of the questions and the discussion from each session are presented below.

OPPORTUNITIES AND UNMET NEEDS

The following participants met to discuss opportunities and unmet needs for the initiative in general: Michael Shuler (chair), John Yin, Sangtae Kim, Rosemarie Hunziker, Bruce Hamilton, and George Komatsoulis.

Discussion in this session was guided by four questions:

1. *What are the opportunities that will emerge in industry as a result of an ability to quantitatively relate genomic information to phenotype? What are the other potential societal benefits?*
2. *Are there intermediate targets that are technologically important in themselves?*
3. *What are the primary obstacles to using genomic information for phenotype prediction?*
4. *Considering the above, where would an NSF program have the most impact?*

Discussion:

The overall challenge is immense; it is to efficiently relate genome-level data on genetic structure, RNA expression, protein content, metabolite content, and so forth, to the phenotype of a whole organism, while including both genetic and environmental factors. Complete model frameworks for a variety of organisms, including humans, is an extremely long-term goal. Such frameworks, when they exist, will greatly assist in the processes of biological discovery, identification of drug leads, improved bioprocesses, and a better understanding of the environment. Even partial models will provide some of the benefits. Complete models for complex, multi-cellular organisms are not attainable in the foreseeable future, but complete models of unicellular organisms and models of subcellular systems in the context of multicellular organisms are feasible goals.

This endeavor is at an early stage, and there are many unmet needs. Examples include a need for improvements in existing devices to obtain genome-level data, development of databases that can be accessed readily by the public, better approaches to predicting structure-function relationships in proteins, and knowledge-based data mining tools. However, many of these needs are or can be best addressed outside the mechanisms available within the purview of this NSF/ENG initiative.

The group members believe that NSF/ENG has a unique opportunity to support the development of a systems framework for the integration of genome-level information into a conceptual structure that also allows inclusion of environmental factors. This structure will likely be an "*in silico*" representation of a living organism. This systems framework will facilitate biological discovery through generation of testable hypotheses and identification of properties that emerge through the interaction of the non-linear components. Further, the framework will encourage the testing of the proposition that sufficient generality exists among organisms that the general "design principles of life" can be discovered. The practical significance of the systems framework is the potential for the rational manipulation of organisms for the production of specific

metabolites, for the identification of drug targets and potential side effects in humans, and targets in pathogens.

Specific points of agreement among the participants in this session were as follows:

- There is a need to integrate information from genome-level data sources into a systems framework, if the maximum value of such data is to be captured.
- Current methods for obtaining genome data are not complete for the purposes of constructing full models. Development of the framework and new methods to obtain quantitative data need to proceed in an interactive manner. In addition to sequence, RNA expression, and proteomic data, methods to obtain data for non-macromolecular metabolite concentrations, protein or enzyme functions and DNA/protein interactions need to be developed.
- Development and application of new systems engineering tools that can be applied to understanding genome-level interaction in living organisms are needed.
- Often these challenges will be best addressed by collaborations of systems engineers, bioengineers and biologists.
- A rigorous evaluation of the quality of genome-level data (e.g., RNA expression arrays) is necessary, as is assessment of model dependence (e.g., sensitivity) on the quality of this data.

TEST BEDS & SCREENING SYSTEMS

The following participants met to discuss data acquisition and integration challenges and opportunities: Kelvin Lee, Rob Fleischmann, Doug Selinger, Scott Peterson, and Mike Domach (chair).

Discussion in this session was guided by four questions:

1. *What are the limits of the current experimental tools for tracking gene expression?*
2. *Are new experimental analyses (e.g. non-RNA analysis) and screening systems needed?*
3. *On what cellular systems should effort be focused to enable progress in the near term, yet have long term scientific and technological relevance?*
4. *Other issues?*

Discussion:

Methods and strategies that offer increased reliability, reproducibility, and resolution for gene expression and protein profile determination are required. The data provided must have the attribute of being more directly useable for quantitative model development and/or testing. In particular, moving beyond ratio analysis is needed where, for example, the information provided is in the molecules per cell, concentration, etc.

New biomolecular screening strategies need to be developed that are complementary to RNA expression data (e.g. DNA microarrays). Examples include the large scale screening of: protein and metabolite levels, second messenger abundance, protein-protein interactions, and protein-DNA interactions.

Unicellular organisms such as bacteria are ideal candidate systems. If the project entails using an organism as a test-bed for proving the concept behind a new screening method, then the genome need not be totally sequenced. Unicellular systems such as *E. coli,* or even simpler systems, may make the best candidates for developing generic genome-to-phenotype rules or relationships although there are trade-offs between modeling tractability and biological importance associated

with any level of complexity. Funds for limited infrastructure grants or a reference to other NSF programs should be provided.

ANALYSIS NEEDS

The following participants met to discuss analysis needs, including modeling: Eberhard Voit (chair), Christodoulos Floudas, Vassily Hatzimanikatis, and James Liao.

Discussion in this session was guided by three questions:

1. *For using genomic data, are there existing mathematical tools that should be investigated more for their applicability for enhancing understanding of what phenotypes are displayed and what underlying biological design principles are?*
2. *What new combinations of mathematical methods need to be developed further to elucidate design principles and/or to enable outcome prediction?*
3. *What are some key issues that modern models must address?*

Discussion:

The ultimate goals of the QSB initiative are (1) the prediction of physiological function from knowledge about genomes; (2) a comprehensive interpretation of post-genomic data; and (3) a true understanding of the design principles of genomic, post-genomic, and metabolic systems.

It was suggested that these goals can only be reached with integrative modeling approaches. It was furthermore advised not to delay the development of such modeling approaches until all necessary data are complete and perfect.

An intermediate target in the pursuit of the goals of this QSB initiative is the development of "good" models of small-scale systems. These models should yield explanations and predictions concerning these small systems and allow the scaling up to larger phenomena. These models should produce testable results and guide the development of new experiments. A different intermediate target, which should be pursued in parallel, is the development of (possibly less detailed) large-scale models. It appears that models at the desired level of complexity will be approximate in nature. They may address the questions of interest from the top down or from the bottom up. The models should be tested and validated in collaboration with bench scientists.

The participants in this session agreed that the systems aspect of research funded under the QSB initiative should be the overriding criterion. There is no doubt that the identification and characterization of individual genomic or metabolic components is of extreme importance. Nevertheless, the QSB initiative should address interactions between such components and only support research on the components themselves, if such research would enhance understanding of their role within a functional network of interactions.

It was agreed that the modeling effort should be limited to "post-genomic" data, i.e., to data at genome-wide and metabolic scales. The QSB initiative should not support methodologies of sequencing, molecular modeling, or genomic data acquisition, unless these are demonstrated to be crucial to our understanding of the interconnected nature of processes governing the emergence of phenotypes from genotypes. In a typical case, a model supported by this initiative would use as components a set of known sequences, whether coding or non-coding, a set of regulators, and/or a number of metabolites and pathways, and connect them in a fashion that would elucidate aspects of their integrated function, coordination, and regulation. As far as it is beneficial for this

type of integrated assessment and understanding, analyses of molecular mechanisms, such as ligand binding, protein-protein interactions, and molecular recognition should be encouraged.

Ideally, different modeling frameworks should be developed and compared. These may include differential equation models, large-scale discrete models and small-scale models that account for the small numbers of substrates, enzymes, and effector molecules in individual cells. The models should be tested with respect to robustness and sensitivities in response to changes in parameters and inputs.

A goal of this initiative should be the investigation of fundamental rules and design principles that govern post-genomic phenomena within cells. Questions that would ideally be answered might include the following:

1. *Why are cellular systems organized the way they are?*
2. *Is there a limited number of design principles or motifs that govern gene regulation?*
3. *Is there a limited number of design principles or motifs that govern metabolic regulation?*
4. *How much commonality in regulation is there among different species?*
5. *Is it possible to establish a catalog of regulatory "units" or "modules" that can be assembled for the prediction and understanding of larger regulatory circuits? If so, what types of methods are necessary to construct large circuits from units or modules?*
6. *What does it take to rewire regulatory networks successfully?*
7. *What does it take to design functioning pathways or cells from scratch?*

Although it was not felt to be mandatory, the collaboration of investigators from different disciplines should be encouraged. Relevant disciplines might include, but are not limited to bioengineering, systems engineering, biology, mathematics, and computer science.

III. APPENDICES

A. WORKSHOP AGENDA

B. LIST OF PARTICIPANTS

C. ABSTRACTS AND VIEWGRAPHS PRESENTED ON SEPTEMBER 13, 2000

Appendix A: Workshop Agenda

National Science Foundation

QUANTITATIVE SYSTEMS BIOTECHNOLOGY WORKSHOP

Chair, Michael L. Shuler

AGENDA
SEPTEMBER 13 & 14, 2000

September 13, 2000
Room 110

7:30-8:00	Coffee and refreshments.	
8:00–8:15	Remarks from NSF: **F. HEINEKEN** and **L. MARTIN-VEGA**	
8:15-8:30	Introduction to session on ***Opportunities & Unmet Needs*** **M. SHULER**	
	8:30-9:00	**S. KIM**
	9:00-9:30	**G. KOMATSOULIS**
	9:30-10:30	Discussion initiated by **J. YIN**
10:30-10:45	Break	
10:45	Introduction to session on ***Needs for Test Beds and Screening Systems*** by **M. DOMACH**	
	10:50-11:20	**R. FLEISCHMANN**
	11:20–11:50	**D. SELINGER**
11:50-1:00	Lunch Break	
	1:00-1:30	**K. LEE**
	1:30-2:30	Discussion initiated by **S. PETERSON**
2:30-2:45	Break	
2:45	Introduction to session on ***Analysis Needs*** **E. VOIT**	
	2:50-3:20	**V. HATZIMANIKATIS**
	3:20-3:50	**E. VOIT**
	3:50-4:50	Discussion initiated by **J. LIAO** & **C. FLOUDAS**
4:50-5:00	Summary & wrap up **M. DOMACH**	
5:00	Adjourn for dinner & rumination	

National Science Foundation

QUANTITATIVE SYSTEMS BIOTECHNOLOGY WORKSHOP

Chair, Michael L. Shuler

AGENDA
SEPTEMBER 13 & 14, 2000

September 14, 2000
Room 580

9:00-10:00	Open Discussion
10:00-11:00	Generation of 3-5 key issues from 3 areas by self-selecting subgroups for the purpose of outlining and phrasing the summary document.
11:00-12:00	Read, explain, and discuss the language that the three co-chairs can synthesize after lunch.
12:00	Adjourn for lunch and departure for those with early afternoon flights

Appendix B: List of Participants

Experts

Robert Fleischmann
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850

Christodoulos A. Floudas
Department of Chemical Engineering
Room A-325
A-217 Engineering Quadrangle
Princeton University, NJ 08544

Vassily Hatzimanikatis
Robert R. McCormick School of Engineering and Applied Science
Department of Chemical Engineering
Northwestern University
Evanston, IL 60208

Sangtae Kim
Vice President Information Technology
Eli Lilly and Company
Lilly Corporate Center, Drop Code 0725
Indianapolis, IN 46285

George Komatsoulis, Ph.D.
Senior Bioinformatics Scientist
Human Genome Sciences, Inc.
9410 Key West Avenue
Rockville, MD 20850

Kelvin Lee
School of Chemical Engineering
Olin Hall, Cornell University
Ithaca, New York 14853-5201

James Liao
Department of Chemical Engineering
5531-E Boelter Hall
University of California, Los Angeles
Los Angeles, CA 90095-1592

Scott Peterson
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850

Michael Savageau
Department of Microbiology & Immunology
5641 Medical Science Building II
University of Michigan Medical School
Ann Arbor, MI 48109

Doug Selinger
Department of Genetics

Harvard Medical School
Alpert Building 5th Floor
200 Longwood Av
Boston MA, 02115

Michael Shuler
Director, School of Chemical Engineering
Director, Bioengineering Program
270 Olin Hall, Cornell University
Ithaca, New York 14853-5201

Eberhard O. Voit
Professor, Department of Biometry and Epidemiology
Medical University of South Carolina
Carolina Medical Center
171 Ashley Avenue
Charleston, SC 29425

John Yin
Department of Chemical Engineering, University of Wisconsin
3633 Engineering Hall
1415 Engineering Drive
Madison, WI 53706-1691

NSF Hosts

Michael Domach
Fred Heineken

NSF Staff

W. Richards Adrion, CISE/EIA
Eugene Bruce, IBN/BIO
Maria Burka, ENG/CTS
Frederica Darema, CISE/EIA
Hector Flores, BIO/MCB
Bruce Hamilton, ENG/BES
James A. Hickman, CISE/EIA
Tapan Mukherjee,ENG/EEC
Sharman O'Neill, IBN/BIO
Dewey Ryu, ENG/BES
Frederick Thompson, ENG
Louis Martin-Vega, ENG
Paul Werbos, END/ECS

NIST

Rosemary Hunziker, ATP

WTEC, Inc.

Stephen Gould
Geoff Holdridge
Roan Horning
Laura Pearson
Robert Shelton