

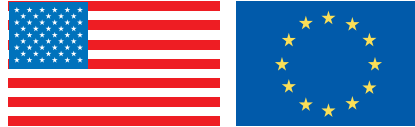
Report from the
US-EC Workshop on

INFRASTRUCTURE NEEDS OF SYSTEMS BIOLOGY

3-4 May 2007
Boston, Massachusetts, USA



Workshop facilitation, logistics, & report editing and publishing
provided by the World Technology Evaluation Center, Inc.



**Report from the
US-EC Workshop on**

**INFRASTRUCTURE
NEEDS OF
SYSTEMS BIOLOGY**

**3-4 May 2007
Boston, Massachusetts, USA**

Jean Mayer USDA Human Nutrition Research Center on Aging (HNRCA)

Tufts University

711 Washington Street

Boston, MA 02111-1524

This workshop was supported by the US National Science Foundation through grant ENG-0423742, by the US Department of Agriculture, and by the European Commission.



Any opinions, findings, conclusions, or recommendations contained in this material are those of the workshop participants and do not necessarily reflect the views of the United States Government, the European Commission, or the contributors' parent institutions.

Copyright 2007 by WTEC, Inc., except as elsewhere noted. The U.S. Government retains a nonexclusive and nontransferable license to exercise all exclusive rights provided by copyright. Copyrights to graphics or other individual contributions included in this report are retained by the original copyright holders and are used here under the Government's license and by permission. Requests to use any images must be made to the provider identified in the image credits.

First Printing: February 2008.

Preface

This report summarizes the presentations and discussions of the US-EC Workshop on Infrastructure Needs of Systems Biology, held on May 3–4, 2007, at the Jean Mayer USDA Human Nutrition Research Center on Aging, Tufts University, in Boston, Massachusetts under the auspices of the US-EC Task Force on Biotechnology Research. It brought together 24 scientists from member states of the European Union, the United States, and Canada.

Current scientific infrastructure cannot keep pace with the progress being made in systems biology research. The purpose of this workshop was to bring together international experts in systems biology and scientific infrastructure to discuss and identify needed improvements in experimental and informatics tools to better support continued research in systems biology and related fields.

The workshop was organized into four sessions: (1) experimental tools; (2) databases; (3) modeling applications; and (4) software infrastructure. The participants at the workshop discussed a wide range of topics including the development of new measurement tools, the establishment of standards for databases with specific capabilities for systems biology, the issues and challenges in modeling biological systems, and the characteristics and standardization of bioinformatics.

The workshop was co-chaired by Marvin Cassman, formerly of the National Institute of General Medical Sciences (NIGMS) of the U.S. National Institutes of Health (NIH), and Søren Brunak of the Technical University of Denmark. They also compiled and edited this report, which is also available on the US-EC Task Force web site, http://ec.europa.eu/research/biotechnology/ec-us/index_en.html. We would like to thank them for their outstanding efforts.

The views expressed in this document are those of the workshop participants, and do not necessarily reflect the views of the sponsors or governments.



Christian Patermann, EC Chairperson



Kathie L. Olsen, US Chairperson

EC-US Task Force on Biotechnology Research



Infrastructure Needs of Systems Biology

Table of Contents

Preface	i
Table of Contents	iii
Workshop Report	
Introduction.....	1
Experimental Tools.....	1
Databases	1
Models, Modeling, and Software	3
Organization and Education	4
Presentation Summaries	
A Human Protein Atlas (M. Uhlén)	7
Measurements for Dynamic Modeling (S. Hohmann).....	8
Ontologies for Data Integration (M. Ashburner).....	9
Enabling Systems Biology: Development and Implementation of Proteomics Standards and Services (R. Apweiler)	11
Towards a European Bioinformatics Infrastructure—In the International Context (J. Thornton)	12
The BioSapiens NoE Experience in High-Throughput GenomeAnnotation and The ENCODE Pilot Project Case Story (A. Valencia).....	13
Can We Design Biological Systems in a Predictable Manner? (P. Silver).....	15
Modeling and Analysis Challenges in Biology: From Genes to Cells to Systems (F. Doyle).....	16
Model-Driven Discovery (M. Covert)	17
Standardization Efforts for Computational Modeling in Biology (M. Hucka)	18
Pathway Commons: A Public Library of Biological Pathways (G. Bader).....	19
Loosely Coupled Bioinformatics Workflows for Systems Biology (D. Kell).....	20
Appendices	
A. Workshop Agenda	25
B. Workshop Participants.....	27
C. Glossary.....	29



Infrastructure Needs of Systems Biology

Workshop Report

INTRODUCTION

Systems biology is in a state of rapid development, characterized by an inability of the infrastructure to keep up with the demands of the science. The purpose of the workshop is to identify the unique aspects of science that define systems biology, and to identify the tools needed—especially in experimental and informatics infrastructure—to support these research areas. The primary topics were experimental tools, databases, models and modeling applications, and software. These correspond to the characteristic aspects of systems biology: time-dependent, quantitative measurements; model-driven experiments; and analytic software yielding predictive outcomes.

EXPERIMENTAL TOOLS

The discussion ranged from the development of new tools such as the creation of a more complete assessment of the human proteome to the characteristics of dynamic data.

Depending how one defines the human proteome, the numbers of proteins and their variants can range from tens of thousands to hundreds of thousands to millions. Since many systems biology models involve the dynamics of components of the proteome, it is clear that the analysis of dynamic cell function is limited by our incomplete understanding of the proteome. One approach is to develop a set of reagents, including antibodies that can be used to generate quantitative, time-dependent, spatially-resolved data. Several such approaches are currently underway, including the creation of a Human Protein Atlas (HPA) www.proteinatlas.org.

A complete understanding of proteomics is necessary to create a comprehensive representation of cellular networks. The tie to systems biology is that the distinguishing characteristic of systems biology is its focus on the behavior of networks rather than on individual molecules. A prerequisite is identification and validation of the networks. This is largely a data-driven approach using high-throughput

datasets, and proteomics is a key element. Also a characteristic of systems biology is model-driven data collection that often focuses on the analysis of dynamic processes. Unlike most currently available high-throughput data, this requires quantitative, time-dependent measurements. Further, such analysis is often context dependent and consequently must be accompanied by significant annotation of experimental conditions (the metadata) to be broadly usable. These measurements for dynamic modeling are commonly small-scale and, because they are model-driven, need to be reported together with the model. Although there are model repositories currently available, there are no databases that provide easily accessible and transportable data (discussed in the Databases section). Finally, multiple data types are frequently required for modeling and the measurements are technically challenging. New tools need to be developed that will allow rapid and easy measurements of quantitative and time-resolved data. These often require single-cell measurements because ensemble-averaging of nonlinear processes obscures their dynamics; processes in individual cells may be out of phase with those in other cells.

Recommendations

- Support the joint creation of common experimental protocols, selection of truly validated common cell types, tools for single cell analysis, globally useful reagents, reporter constructs, etc., thus making experimental data more valuable for modeling.
- Make established experimental techniques broadly available to the community.
- Create a large-scale proteomics effort which would include alternative modifications, localization, structure, etc.

DATABASES

The discussion primarily revolved around several issues: the development of standards, the need for specific capabil-



Infrastructure Needs of Systems Biology

ities for systems biology, and the difficulties in supporting databases. There was general agreement that there is a need in systems biology for a controlled vocabulary, a uniform exchange format, and a set of minimum elements that are required in reporting a systems biology experiment. The effective absence of all of these constitutes a significant impediment in communication of information. The result is an inability to transfer data between laboratories and an inability to communicate information across cell types and local experimental conditions.

This is hardly a new problem. There was extensive discussion of two approaches for dealing with these issues, the GO (Gene Ontology) model and, more recently, the Human Proteome Organization Proteomic Standards Initiative (HUPO PSI). GO, begun in 1998 through a collaboration between three databases, is an excellent example of the way a community organized its efforts to provide a consistent identification of biological elements (in this case, genes and gene products) across a number of databases. The use of GO terms by collaborating databases facilitates uniform queries across them. The controlled

vocabularies are structured so that they can be queried at different levels. The HUPO PSI is a much more recent development. It was initiated, in 2002, with the intent of developing data exchange formats, controlled vocabularies, and reporting guidelines for Minimum Information About A Proteomics Experiment (MIAPE) in three broad fields—mass spectrometry, gel chromatography, and molecular interactions. Significant progress has been reported.

The key issue is an understanding of what data are needed for systems biology and how they should be represented. As noted above, most model-driven data are small-scale and dependent on specific cell types and experimental conditions. It is also unclear what “completeness” would mean for systems biology data. Since dynamic processes almost certainly involve both constant and transient interactions, the representation of such processes in a database would be very different from the “core” databases such as GenBank, UniProt, and the Protein Structure Database. The characteristics of “core” databases are defined by Janet Thornton in Table 1.

Table 1
Core Database Characteristics

- They are universally relevant to biomolecular science.
- They have a huge user community.
- They aim to be complete collections.
- Their completeness is assured by exchange agreements with other data centres worldwide (typically the USA, Japan, and Europe, at present).
- The science they represent is stable enough to allow standardization of the data structure.
- They follow standards, where available.
- They are actively involved in relevant standard development.
- Journals insist that data be deposited in these databases.

At the moment, systems biology data meet none of these criteria. This is particularly critical since it is precisely the criteria of usage, stability, standards, etc. that are the factors in determining whether a database will be supported. And yet easily transportable and usable data are critical for progress in systems biology. Precisely because it is model-driven, access to data is needed for documentation

of model predictions and to test other models against the same data. Further, it will be necessary to integrate multiple experimental data into larger cell/organismal databases. However, funding of database infrastructure is neither easily accessible nor adequate when it is provided. It is further complicated by the question of how to get national entities to pay for an international effort.

Workshop Report

The most common approach for the development of nascent databases is through small groups piggybacking on existing collaborations and scavenging from existing funding. This begins with the establishment of standards, ontologies, etc. It has been done in a number of cases, e.g., Systems Biology Markup Language (SBML) (originally supported by Japanese Science and Technology Corporation) and the GO (out of a consortium of existing funded databases). Basically, local initiatives emerge as de facto standards. Is this the best way to do it? The question is irrelevant. It is a normal progression for consumer technology. Nevertheless, it would certainly help if there were funding mechanisms that would ease the process. There are programs underway in the US (Continued Development and Maintenance of Software) and the EC (European Life-Science Infrastructure for Biological Information, ELIXIR), to provide a transnational infrastructure for biological information that may be of help in providing support for such efforts. To begin the process of developing a mechanism for annotating and storing systems biology data a number of recommendations were made by the Workshop participants.

Recommendations

- Establish criteria for long-term support for systems biology relevant databases.
- Support the development of standard representations enabling interoperability between databases and tools.
- Support data capture incorporating minimal information, using standard formats and semantics.
- Support and broaden BioMart-like data integration schemes going beyond sequence centric approaches.
- Promote access to full-length paper text and repositories and promote semantic enrichment efforts.
- Support ‘workflow’ schemes in the context of systems biology.

MODELS, MODELING, AND SOFTWARE

Models are part of science at every level and in every discipline. Every hypothesis is in essence a model to be tested. In systems biology models mean something more specific. They are representations and analytic tools for examining networks. These networks can be anything from small motifs

with a limited number of elements to cellular, metabolic, and regulatory structures to networks on the physiological level, such as organ systems. Certainly the application of predictive models in biology has a long history, going back at least to the Lotka-Volterra systems and the Hogkin-Huxley equation. However, modern biology has largely neglected the application of quantitative predictive models. What defines systems biology in particular is the iterative application of predictive models followed by experimental testing of the predictions. Although such approaches are still at a relatively early stage, there have been numerous examples of their value. Some examples are the application of modeling to circadian rhythms, *E. coli* metabolism, signal transduction, and the synthesis of switches and other biological processes based on predictive models. The diversity of systems addressed and their levels of organization indicate the generality of the approach and its potential utility. In every case a specific experimental system is queried using a quantitative predictive model. The value of the model can be to test competing hypotheses, to (in)validate data sets, and to suggest new experiments. The result is frequently not only the ability to address a specific biological question, but to uncover new biology. There are also benefits in the applied sciences and medicine. In the broadest sense this can include biological design concepts such as modularity and robustness as well as the ability to engineer, or re-engineer, biological processes.

An important issue is the extent to which current infrastructure supports modeling efforts. In general, development of tools to facilitate access to and communication of models is considerably advanced compared to data. There are a variety of databases that reflect network and pathway representations and annotation, including Kyoto Encyclopedia of Genes and Genomes (KEGG), BioModels, Reactome, etc. Additionally, there is an attempt through Pathway Commons to integrate all this information through a single point of access.

A comparable effort involves standardizing the representation of computational models. There has been significant progress in this area, through the development of SBML and Cell Markup Language (CellML) as model exchange formats; MIRIAM (Minimum Information Requested in the Annotation of Biochemical Models), which provides guidelines for annotation and curation on quantitative models; and the BioModels Database, which stores quantitative models. All of these are under continuous

Infrastructure Needs of Systems Biology

development, as is the creation of standards for graphical representation of models.

Still missing is a mechanism to access the modeling software currently being developed in laboratories across the world. At the moment there are two classes of software. Probably the most common are off-the-shelf commercial packages such as Matlab. These are fine (perhaps) for skilled and knowledgeable software developers. They are not so good for the naïve, or modestly involved, user. In particular, if a goal of systems biology is to make its concepts and tools as widely used by the working biologists as the concepts and tools of molecular genetics are now, then something more is needed—specifically, software packages that can easily be used by someone other than the developers. This kind of prepackaged software is generally developed to address specific problems such as simulations of regulatory pathways or gene networks. Here there are real problems with access. Software development in systems biology is essentially a cottage industry. A piece of software is developed for local consumption and, with a few exceptions, is rarely available with any ease outside the local environment. This leads to much duplication, wasted effort, and a total inability to link to other pieces of software, not to mention an inability to test the software capability by others. This is not because people are reluctant to distribute their software, but because there is no easy way to find out what exists and what it does. The very common absence of decent documentation generally makes it impossible to use without great effort. A reasonable set of expectations is that such software should be interoperable, transparent to the user, and sufficiently well documented so that it can be modified and adjusted to circumstances. No resource currently provides these capabilities.

Finally, there is the general problem of associating all the data and modeling capabilities in a framework that allows (relatively) seamless integration. There have been a number of attempts at this, including the Systems Biology Workbench (SBW) and Biological Simulation Program for Intra- and Inter- Cellular Evaluation (Bio-SPIICE), none of which have been widely used. More recently, “workflow environments” such as Taverna have been created to loosely integrate the elements involved in systems biology studies. Such approaches require further development and testing.

Recommendations

- Support initiatives in multi-scale modeling spanning molecular to multi-tissue organism levels.
- Support the use of standards and environments that permit interoperability and integration.
- Initiate an infrastructure-related software support mechanism in the EC (like in the US).
- Support systems biology software repositories which incorporate software curation .
- Support education in the use of software within systems biology.

ORGANIZATION AND EDUCATION

Of all the infrastructure issues, education is arguably the most important. Biology has made two great leaps in the 20th century and both have come from merging with other disciplines. First, at the turn of the century, chemistry turned the cell, from an opaque and unknowable structure with rules of behavior that were unique to itself, into an understandable construct made of chemical components that followed the rules of chemistry. Then, almost exactly in mid-century, genetics and biochemistry partnered to create a second revolution to convert the gene into a physically knowable object. Now, at the beginning of the 21st century, engineering, mathematics, computer science, and to some degree physics are about to transform biology again, into a predictive science. This will be a harder adjustment, since the fields are so disparate and the backgrounds of the participants will have to accommodate very different bodies of knowledge. Mathematics is at the core, and this is a major shortcoming in the training of biologists. Over the past 30 years biology has become a discipline for people who want to do science without learning mathematics. Consequently, it is by no means clear how best to accomplish this synthesis of disciplines, but there are a lot of attempts in progress. At this stage it is important to try all the reasonable approaches that people can come up with, but it is equally important to convey the results to help others to see what works and come up with best practices.

A major organizational issue in the provision of infrastructure is the prosaic question of how to provide funding. Fundamentally, the issue is how to support

Workshop Report

international issues such as databases and other central resources through national funding mechanisms, given that international bodies are not (yet) available for this purpose. There are several attempts in progress. In the US, the NIH has a program called “Continued Development and Maintenance of Software.” In the EC, the “European Strategy Forum on Research Infrastructures (ESFRI)” evaluated and recommended a number of biologically important infrastructure initiatives to the EC. The list was adopted by the EC but without funding. (A common phenomenon generally referred to as “approved but not funded.”) Currently underway is a proposal to the EC known as “ELIXIR” which will establish a transnational infrastructure for biological information.

Recommendations

- Support education in the use of software within systems biology.
- Initiate US-EC collaboration on establishing curricula in systems biology.
- Support activities similar to competitions like Internet Geotechnical Engineering Machine Competition (iGEM) (see www.igem2007.com).

- Support community building around concrete projects, e.g., ontologies and databases, funded jointly by the US-EC. In effect international glue-grant funding.
- Establish joint US-EC panels for assessment of research projects.
- Joint US-EC systems biology benchmark studies such as A European Network of Excellence (ENFIN-DREAM) www.enfin.org/dokuwiki/doku.php?id=wiki:wp7 and <http://magnet.c2b2.columbia.edu/news/DREAMInitiative.pdf>. This could possibly include funding for pre- and post-prediction experimental data generation, evaluation, and creation of standards.
- Generate procedure for US involvement in ESFRI.

Specific recommendations for prompt action

- Create a mechanism to support ongoing joint US-EC benchmark efforts with special emphasis on data generation.
- Start effort on standards and interoperability for databases, software, and experimental systems.
- Exchange information on training programs.

Infrastructure Needs of Systems Biology

Presentation Summaries

Summaries of the speakers' presentations follow. Adobe Acrobat (PDF) versions of their individual slide presentations are available at www.wtec.org/ec-us_sysbio_workshop/.

A HUMAN PROTEIN ATLAS

Matthias Uhlén (based on a transcript of his presentation)

Mathias Uhlén began his presentation by noting that we are in the era of “omics,” in which genomics, transcriptomics, proteomics, and interactomics all contribute to systems biology which in turn leads to discoveries related to pathways, biomarkers, and drug targets. Currently, the number of DNA sequences in public databases doubles every ten months, while more than one genome is finished every week. The status of genome projects at the U.S. NIH illustrates the progress being made.

The human proteome consists of around 22,000 non-redundant proteins, over 200,000 protein variants, over 10 million combinatorial variants, over 100,000 protein species, and over 57,000 protein alleles. Analysis of the proteome can lead to improvements in *in vivo* imaging and diagnostics, increase our knowledge about cells, tissues, and diseases, and improve pharmaceuticals and biotechnology products.

The Human Antibody Initiative, co-chaired by Dr. Uhlén, Michael Snyder, and Peter Hudson, has two subprograms: the Protein Atlas (Sweden) and the Antibody Resource (US/EC). The Swedish Human Proteome Resource (HPR), established in 2003, is funded by the Kurt and Alice Wallenberg Foundation, with additional funding from the EC, MolPAGE, and ProteomeBinder. HPR is located in eight sites Royal Institute of Technology in Stockholm, Uppsala, Karolinska Institute, Lund, Umeå, Seoul, Beijing, and Mumbai. Its current throughput is 10 new validated antibodies per day and 10,000 images, which are added to the publicly available HPA.

The process of annotating the immunohistochemical images begins with a review by an experienced pathol-

ogist who annotates the pattern, distribution, intensity, and localization. Web-based annotation software is being developed to assist this process. Tissues are annotated automatically, while cells require manual annotation. A staff of six full-time curators annotates over 10,000 images per day; five software engineers are required to support the workflow. Antibody quality assurance is provided by the HPR center at Stockholm/Uppsala. In the HPR pipeline, high-throughput cloning and antigen expression results in 200 genes initiated every week; antibody generation results in 150 antibodies purified every week; and antibody-based annotation results in over 10,000 images annotated every week. As of April 2007, 14,185 genes have been initiated (representing more than 50% of all human genes); 23,623 protein expressed sequence tags (PrESTs) have been designed (over 45,000 primers synthesized); 14,883 PrESTs have been cloned (and their sequences verified); 11,798 antigens have been sent for immunizations; 3,143 antibodies have been annotated (with over 2 million annotated images); and 1,514 antibodies have been released in version 2.0 of the HPA, which was released in October 2006.

Version 3.0 of the Atlas is scheduled for release in the third quarter of 2007 at the 6th HUPO conference in Korea. It will include approximately double the number of antibodies (3,000). Version 4.0, scheduled to be released in the second quarter of 2008 at the 7th HUPO conference in Amsterdam, will again contain approximately double the number of antibodies as the previous version (6,000).

A new affinity reagents portal based on the existing HPA database structure will include a list of all submitted antibodies with a short description, chromosome location, and links. The affinity reagent and antigen types will also be included along with validation data including reliability scores.

Looking at the future, the goal is to have 8,000 human proteins in the HPA by 2009, with a first draft of the complete human proteome in 2015, extended content



Infrastructure Needs of Systems Biology

that will include subcellular localization, serum analysis, transcriptomics, additional diseases, and other species.

MEASUREMENTS FOR DYNAMIC MODELING

Stefan Hohmann

Systems biology is a rapidly developing scientific approach that aims at describing the properties of networks of biomolecules and the rules with which they operate in cells and organisms. Inherent to systems biology is the integration of experimentation and computational modeling. Presently, systems biology comes in two mainstreams:

- Top-down or data-driven systems biology that employs high-throughput datasets to discover molecular networks and study their properties.
- Bottom-up or model-driven systems biology that commonly tries to understand the dynamic operation of already known molecular networks and processes. This approach is generally in need of data, especially quantitative, time-resolved data.

Data relevant for dynamic modeling include, among others and depending on the system under study, physico-chemical properties of the molecules, their concentration, rates of changes of such concentrations, velocity of movements or diffusion rates, rates of changes of interactions, rates of changes of protein modifications, etc.

The European Commission supports several Specific Targeted Research Projects (STREPs) on dynamic modeling and has published a call for larger projects with a deadline on April 2007. The speaker coordinates the projects Quantifying Signal Transduction (QUASI) (2004–2007) and systems biology of the Activated Protein Kinase (AMPKIN) (2006–2009).

QUASI is a consortium of five partners and seven research groups (all in either the fields of biology, chemistry, or computational biology) studying yeast mitogen-activated protein kinase (MAPK) signaling using experimentation and modeling. Types of measurements performed include rates of changes of MAPK phosphorylation, certain protein interactions, messenger ribonucleic acid (mRNA) and protein levels as well as certain relevant metabolites. In addition, nuclear-cytoplasmic shuttling of MAPK is monitored as a real-time measure for signaling. Perturbations to test and optimize mathematical models include genetic changes, environmental treatments as well as

specific kinase inhibitors (mutant kinases and ATP analogues). Questions successfully addressed by QUASI include feedback control mechanisms in pheromone and high-osmolarity signaling MAPK pathways, control of cell cycle by MAPK pathways, control of a eukaryotic osmolyte system, regulation of gene expression by high osmolarity glycerol (Hog1) MAPK, integration of converging branches of signaling pathway (HOG branches), and pathway crosstalk. Issues and challenges raised by QUASI include modeling processes operating according to different biochemical rules and in different time frames (signaling, gene expression, cell cycle, metabolism), monitoring intermediates of signaling pathways (phospho-proteins), interpreting the use of genetic perturbation versus specific inhibitors, and distinguishing response profiles according to cell-to-cell variations.

AMPKIN studies the control of the adenosine-5'-monophosphate-activated protein kinase (AMPK) signaling system in both yeast and mammalian cells. The pathway controls energy homeostasis in cells and whole organisms and hence is critically important for treatment of type II diabetes and the metabolic syndrome. The consortium has four partners and five research groups (all in the fields of biology, physics or computational biology). Types of measurements performed include glycolytic flux and rates of changes of metabolite levels, rates of changes of phospho-AMPK, rates of changes of phosphorylated forms of certain target proteins, activity of target enzymes, absolute levels and rates of changes for many pathway components, and rates of changes of mRNA levels for reporter genes. AMPKIN also performs single cell analysis to assess by fluorescence-activated cell sorting (FACS) analysis population profiles using fluorescent protein reporter genes (reporter-XFP) and by microscopy nuclear shuttling of the Mig1 transcription factor as read-out for real-time signaling. Perturbations employed include genetic changes in the signaling and metabolic pathways, specific kinase inhibitors as well as changes in experimental conditions. Significant questions addressed by AMPKIN encompass among others comparative modeling of yeast and mammalian pathways, integration of metabolism and signaling, mechanisms controlling pathway activity and cross talk to other pathways, signaling via protein kinases or protein phosphatases as well as the quantitative contributions of parallel pathways. Issues raised by AMPKIN are defining treatments for activating/deactivating the

Presentation Summaries

pathway, rapid sample preparation for measurements, using genetic perturbation versus specific inhibitors, as well as distinguishing response profiles according to cell-to-cell variations.

At this stage, quantitative time course data for dynamic modeling are scarce, not available in databases (there are no databases for that purpose), and commonly not produced (yet) in high-throughput and at global scale. Rather, those data are mainly produced in dedicated, small-scale experiments. Another challenge concerns the generation of connected datasets, i.e., sets of different types of data (transcriptome, proteome, metabolome, etc.) from one and the same experiment. All these aspects call for guidelines and standards for reporting, depositing, and disseminating such datasets produced for dynamic modeling. The EC-funded coordination action YSBN (Yeast Systems Biology Network addresses some of these issues and a project proposal has been submitted to significantly extend on those.

QUASI and AMPKIN, as well as many other systems biology activities, highlight the need for single-cell analyses to collect data for dynamic modeling. Two types of data are critically important:

- Population profiles that provide information on the fraction of cells that respond to a certain stimulus at different times. Such data are needed to properly interpret response profiles obtained from cell extract from billions of cells for instance by western blotting or reverse transcriptase polymerase chain reaction. For instance, observing amplitude diminished by 50% in a mutant as compared to wild type could mean that all mutant cells respond to 50% or that only 50% of the mutant cells respond but with maximal amplitude. Typically, population profiles can be obtained by flow cytometry, requiring suitable fluorescent reporters. Those reporters presently constitute a bottleneck.
- Single cell analyses allow monitoring signal transduction in real time. Again, suitable reporters are a bottleneck. Typically, cells are exposed to altered environmental conditions (stress, hormone, etc.) in a microfluidic device linked to a fluorescent microscope. Reporters could be protein movements due to signaling (e.g., nuclear-cytosolic shuttling) or transient protein-protein interaction monitored by fluorescence resonance energy transfer (FRET).

In conclusion:

- Measurements for dynamic modeling are commonly small-scale and highly dedicated.
- Data collection is model-driven, but high-throughput approaches (such as protein phosphorylations via mass spectrometry) are becoming feasible.
- Data should be reported in conjunction with the model and suitable reporting schemes and repositories are needed.
- Measurements are often technically challenging and different types of technological advances, especially in proteomics, metabolomics, and bioimaging are needed.
- Live-cell imaging and single cell analyses methods are important and need to be further developed, suitable reporters especially present a bottleneck.

Eventually the two present mainstreams of systems biology, top-down and bottom-up, will need to come together to approach the vision of dynamic models of whole cells and organism at molecular resolution.

ONTOLOGIES FOR DATA INTEGRATION

Michael Ashburner

That biology is an “information-rich science” is a truism.

Nevertheless, it is one that must be appreciated by funders, since without an information infrastructure (and the funding that requires) much, or even most, of today’s biological and biomedical research will be a major opportunity missed, and a major misuse of funding. While the highlights of today’s research will continue to be published in the scientific journals, the data upon which these are based can only be published in well-structured and open databases.

There are well over 1,000 different databases in the general domain of biomedicine. Galperin’s last annual list (M. Galperin, 2007, The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Research* 35:D3-D4, http://nar.oxfordjournals.org/cgi/content/full/35/suppl_1/D3/DC1) lists 968 databases, a growth of 110 in 2006. A distinction should be made between “community” and “hobby” databases. The former are core databases, generally, though not always, funded for the long-term, and they are funded with the objective of

Infrastructure Needs of Systems Biology

providing an open community resource; they are generally, but not always, institutionalized, with the consequence that their survival does not depend upon the vagaries of individual employment. (Good examples would be European Molecular Biology Laboratory (EMBL)-Bank, GenBank, and UniProt.) The latter are generally built (at least initially) for the purpose of helping the research of an individual lab or a small group of labs. They depend, if lucky, on short-term funding, and if this funding fails, or the individuals responsible for the database fall under a bus, retire, or get fed-up, then the database folds or disappears into the commercial domain. That is not to say that “hobby” databases cannot be valuable resources, simply that they cannot be relied upon as a component of infrastructure.

The number of community (core) databases that are absolutely required for biomedical research is probably not very large: less than 50 would be a reasonable guess. This number will include databases of the literature, primary sequences (nucleic acid and protein), structures (proteins, small molecules, etc.), and genomes and their polymorphisms, transcriptomic and proteomic data, model organism databases, etc.

The funding climate for community (core) databases in the US is actually not too bad. For example, the NIH has a funding modality (P41) specifically designed for the support of infrastructure. It is this mechanism, for example, that supports UniProt, a host of model organism databases (FlyBase, *Saccharomyces* Genome Database, Mouse Genome Database, etc.) as well as such valuable resources as Reactome and the University of California Santa Cruz Genome Browser.

By contrast, Europe has completely and consistently failed to provide this sort of support, not only at the level of the Commission but also through Member States. We all recall the funding crisis for SwissProt, which drove it into commercial licensing and, later, into being funded by the NIH. We can all name other valuable European resources that were driven into the commercial domain. A Data Base of Eukaryotic Transcription Factors (TRANSFAC) is an example, the “up-to-date” version of which is only available by commercial license: www.gene-regulation.com/pub/databases.html#transfac. We all further recall the funding crisis that hit the European Bioinformatics Institute (EBI), Europe’s premier provider of community

databases, in the summer of 1999. This is a problem that must be solved at the European level. The Commission spends billions of euros on supporting biomedical research. Without a community database infrastructure much of this money will, if not be wasted, then not gear the appropriate return. As a guideline the FY2008 budget for the National Human Genome Research Institute (NHGRI) is US\$484 million. Of this, just under 10% (a little over US\$30 million) supports database infrastructure (Peter Good, NHGRI, pers. comm.). This funding supports core data resources in the community, in addition to the major US funding stream for the National Center for Biotechnology Information (NCBI), which probably totals over US\$50 million per year.

However, simply supporting databases is by far not sufficient. The funders have a duty to ensure that the databases they support are not only effective and efficient, but also work together. Biomedical data cannot be simply partitioned into different sealed boxes. We need to be able to work with data from a wide range of sub-domains. There are many actions that can facilitate this. One is to ensure that, at a technical level, the databases share standards and tools. This the NHGRI does through its GMOD (Generic Model Organism Database) project (www.gmod.org). The other is by encouraging databases to use a common set of semantic standards. This is being done by the NIH Directors’ Roadmap initiative project, the National Center for Biomedical Ontology (NCBO) (www.bioontology.org). The NCBO grew out of the great success of the Gene Ontology Consortium (www.geneontology.org/) in enabling all of the major model organism (genomic) databases to share a structure and semantics for the description of gene function. This in itself lead to the Open Biomedical Ontology infrastructure (<http://obofoundry.org/>), a clearinghouse for interoperable ontologies in the biomedical domain.

Therefore, long-term funding support in this area is needed for:

- Support of community (core) databases.
- Development of semantic and technical standards for databases.
- Support for the development and integration of the most valuable “hobby” databases.

Presentation Summaries

- Support that will enable a network of databases to work together for the greatest benefit to the community.

It would not be too great a task to draw up, at high-level, a blueprint for this infrastructure. There are signs in the US that the funders will be encouraging more interaction than now occurs between database providers. Ideally, this should be done at the international, rather than national, level. ELIXIR, an EU initiative to coordinate infrastructure development and funding within Europe, is one step in this direction, but this initiative should be broadened to include international scientific and funding cooperation and planning.

What can the Commission do?

- Establish criteria for long-term support for community (i.e., core) databases needed for systems biology, and give long-term support for these databases.
- Support the development of standard representations, including ontologies, enabling interoperability between databases and tools.
- Support data capture incorporating minimal information, using standard formats and semantics.
- Support and broaden BioMart-like data integration schemes going beyond sequence-centric approaches.
- Promote access to full-length paper text and repositories and promote semantic enrichment efforts.

ENABLING SYSTEMS BIOLOGY: DEVELOPMENT AND IMPLEMENTATION OF PROTEOMICS STANDARDS AND SERVICES

Rolf Apweiler

Rolf Apweiler spoke about the importance of standardized data reporting in systems biology and illustrated that by explaining the work of the HUPO PSI. The HUPO was formed to consolidate national and regional proteome organizations into a single worldwide body. In April 2002, the PSI committee was formed by HUPO and tasked with standardizing data formats within the field of proteomics. Databases and repositories could then use to exchange their existing content and encourage pre-publication of experimental data. In parallel, reporting standards were to be written, comparable to those already published in the field of microarray data, to improve the standard and consistency of data reporting in published literature.

The HUPO-PSI has concentrated its efforts in 3 broad fields:

1. Mass spectrometry
2. Gel chromatography
3. Molecular interactions

Achievements in Mass Spectrometry

- mzData eXtensible Markup Language (XML) data interchange format released 2004, implemented by wide range of instrumentation manufacturers.
- Minimum reporting requirements document MIAPE Mass Spectrometry (MS) in second round of review by *Nature Biotechnology* (NBT).
- Work in hand to merge mzData with Institute for Systems Biology (ISB) mzXML to create a single, universal format, mzXL. Planned release 2007.
- Protein/peptide identification format in final stages of completion. Planned release 2007.
- Workshop planned June 2007 to finalize both formats ready for public consultation and review.
- Proteomic Data Collection (ProDaC) collaboration focused on implementation and journal acceptance.

Achievements in Gel Chromatography

- Gel Markup Language (GelML) interchange format completing public review and in preparation for 1.0 release.
- MIAPE-Gel in second round of review by NBT.
- Informatics format GelInfoML in preparation.

Achievements in Molecular Interactions

- Proteomics Standards Initiative-Molecular Interactions (PSI-MI) XML1.0 published (NBT) in 2004—implemented by all major public domain databases and many large-scale data producers.
- PSI-MI XML2.5 made public 2006, schema now more flexible, can exchange more data types. Most databases have now upgraded to new version. Publication under second round of review by *Biomed Central Bioinformatics* (BMC).
- Minimum reporting requirements document accepted by NBT.

Infrastructure Needs of Systems Biology

- International Molecular Exchange consortium formed to exchange data/share curation effort.
- Work in hand to produce HUPO PSI interaction data scoring system and gold standard test sets for benchmarking large interaction data sets.

Achievements in over-arching activities

- Working with FuGE (Functional Genomics group) to produce data model to describe experimental meta data (sample preparation, etc.) common to all “omics” experiments.
- Further minimum standards documents in preparation to cover other areas, such as column chromatography.
- Developing Minimum Information for Biological and Biomedical Investigations (MIBBI) to ensure this work remains non-redundant, current and accessible to the user community.
- Working with journals to expedite publication process including data deposition.

TOWARDS A EUROPEAN BIOINFORMATICS INFRASTRUCTURE—IN THE INTERNATIONAL CONTEXT

Janet Thornton (based on a transcript of her presentation)

Janet Thornton began by providing a review of the current status of biomolecular information resources in Europe,

particularly the molecular databases at European Bioinformatics Infrastructure (EBI) used for genome and nucleotide sequencing; gene expression; protein sequencing, families, structure, and interaction; chemical entities; pathways; and systems. Data are freely deposited, exchanged daily, and made available to the large and growing user community.

EBI’s core data resources are growing rapidly, and it is aiming to develop complete collections through exchanges with other data centers around the world (typically the USA, Japan, and Europe at present). The science represented by the collections is stable enough to allow standardization of the data structure. The databases follow existing standards, and EBI is actively involved in developing relevant standards. Journals insist on data deposition in these databases.

EBI has more than 700 specialized molecular data resources, of which 30% are in Europe (see Fig. 1). Non-core data resources at EBI are more specialized (e.g., one species or family) and do not aim to be comprehensive. They are investigator-led products of research groups with specialized content that may require expert users. Many are derivative of a range of other databases. Some may be candidates for the core collection. Criteria for assessing data resources include usage cost and value, stability, standards, size, and international status. Current international resources for DNA sequences, genomes, protein sequences, structures, and interactions, and model organism data resources were then reviewed.

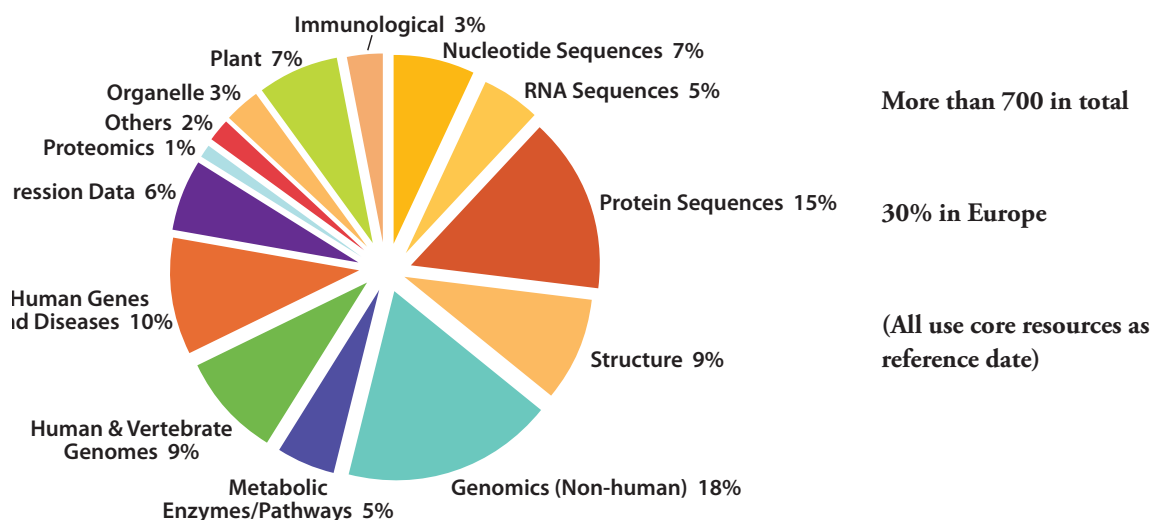


Figure 1. Specialized Molecular Data Resources (Galperin: 2005 NAR).

Presentation Summaries

The biomolecular community requires exchange agreements and protocols to be established for data related to expression, human variation, images, proteomics, and metabolomics, among others. The European Model for Bioinformatics Research and Community Education (EMBRACE) EU Network of Excellence makes tools and web services available to researchers, while the BioSapiens EU Network of Excellence provides annotations.

Integration with other data is increasingly important, however funding for these infrastructures in Europe is neither sensibly organized nor adequate. Given that bioinformatics is international and its resources are used by everyone, who should pay? Most nations do not want to pay for an international effort, although the US has been very proactive in its support for public data resources and making data freely available through the Web. Commercial solutions are not available. The EMBL provides core funding but has a limited budget; the EU has provided less than 20% of EBI funding. Apart from Switzerland and the recent UK Biotechnology and Biological Sciences Research Council (BBSRC) initiative, there appears to be little national funding for core bioinformatics data resources. Private foundations, e.g., Wellcome Trust, support some funding.

Dr. Thornton reviewed the ESFRI, established in 2002 by the Competitiveness Council. Its goal is to describe Europe's scientific infrastructure needs for the next 10–20 years, including the establishment of six biomedical and life sciences centers. Although adopted by the EU, little funding is available to support it. Therefore a 'Preparatory Phase' has been established to create consortia of European Funding Bodies willing to support individual infrastructures.

The proposed ELIXIR project will establish a transnational infrastructure for biological information and service providers, including existing national infrastructures and networks. It will also upgrade EBI and construct a European Biomolecular Data Center, and promote the use of state-of-the-art information technology for data integration and database interoperability and distributed annotation technologies for large-scale European collaborations. Other outcomes will include biological information infrastructures in new accession states, appropriate legal and financial frameworks, and training and outreach. ELIXIR will consist of an interlinked collection of 'core'

and specialized biological data resources and literature that will facilitate rapid search and access.

Funding for ELIXIR will be provided by EMBL member states (see www.embl.org/aboutus/generalinfo/membersmap.html), other EU states, non-member states in the EC, and others. Stakeholders include the funders, EMBL, the EU, national governments, private foundations, industry, data resource providers, and others. The total budget will be €567 million over 7 years (2007–2013). New funding is needed for data resources in chemicals in biology and medicine, imaging, human variation, literature resources, distributed specialist resources as well as hardware, software, and personnel. These are included in the ELIXIR proposal.

The presentation concluded with a review of items for further discussion, including how best to link core and specialist resources, which new resources are needed for systems biology, how best to develop the ELIXIR model in an international context, what are the necessary links between funders, how to integrate national funding of specialized resources, and what to do on the issue of centralized vs. distributed resources.

THE BIOSAPIENS NOE EXPERIENCE IN HIGH-THROUGHPUT GENOME ANNOTATION AND THE ENCODE (ENCYCLOPEDIA OF DNA ELEMENTS PROJECT) PILOT PROJECT CASE STORY

Alfonso Valencia

Alfonso Valencia presented four projects with a clear focus in systems/synthetic biology in which the Bioinformatics resources are essential components.

ENFIN is a collaborative project dedicated to foster the interrelation between bioinformatics predictions/modeling and experimental approaches, with the aim of making progress in areas of relevant biological significance.

In the working model of ENFIN explicit predictions are derived from the computational models and tested by experimental groups. The results will be used in additional cycles of prediction/experimentation. The project is now (first quarter 2007) at the end of the first cycle of experimental validation.

Infrastructure Needs of Systems Biology

ENFIN is collaborating with the U.S.-based DREAM initiative that offers a complementary view of the problem of assessing computationally derived system models.

Two basic infrastructures are essential for the work carried out in ENFIN. First repositories of text (full articles) accessible in the appropriate format (XML) and the tools to identify entities (gene and protein names, but also others), their sequence database correspondences, and their relations (protein interactions, gene control, and others) are used for the prediction of protein associated to biological processes (i.e., formation of mitotic spindle). Various text mining approaches were used in combination with other bioinformatics approaches provided by members of ENFIN to predict new spindle-associated proteins.

Beyond the current application in ENFIN, text repositories and text mining tools that facilitate the access experimental biologist to the information are proposed as essential pieces for the development of systems biology approaches.

Infrastructure Needs in Text Mining

- Text repositories in biomedicine
- Open Access/Editorial policies/Abstracts versus full text
- Formats (HTML, PDF, text)
- Pictures, figures, tables, other non-textual information
- New sources of information: Web, Nature scramble text, others
- Lack of Annotated text sources for training Neuro-Linguistic Programming (NLP) systems (annotation efforts)

Methods for Information Extraction and Text Mining

- Comparative assessment (Biocreative and others)
- Access to methods, common formats, and distributed annotations

End Users' Needs

- Database, curation, and annotation
- Biologist and communities
- Others?

In the context of protein interactions and protein interactions network, ENFIN is interested the interaction of proteins with basic protein interaction modules (e.g., SH3, SH2, and others). A number of computational approaches based on the combined use of sequence family information and structural information are used to predict interactions and hypotheses about how to control those interactions with point mutations.

At the level of infrastructures, the availability of sequences (in particular, complete genomes of similar species) and structures of protein complexes are well-recognized needs. It is perhaps less obvious, but equally important, to availability of curated databases and sufficiently precise computational methods able to lead with the complexity of this information.

Infrastructure Needs in Protein Interaction Networks

Insufficient quantitative experimental data

- Protein-protein interaction specificity data
- Structures of protein complexes
- Genomes of closely related species

Sharing data of prediction methods

- Quality standards, definitions of methods, updates, dependencies
- Relation of genome-and structure-based prediction methods
- Protein engineering of surfaces and interactions
- Manipulation of specificity (small molecules)

End users needs

- Data representation, linking to databases

For the work carried out by the BioSapiens Network of Excellence (NoE) on the annotation of the protein complement of the human genome, it is particularly relevant the cooperative effort carried out to analyze the potential splice forms in the 1% of the human genome mapped by the ENCODE consortium (Tress et al., PNAS 07). The development of annotation pipelines, in particular the adaptation of new ones for the annotation of splice variants, are the essential scaffold in which systems biology is built.

Presentation Summaries

Infrastructure Needs in ENCODE Protein Annotation

Insufficient experimental information about protein structure and function.

- Expression and stability of proteins (antibodies, mass spec databases)
- 3D Structures of proteins and domains

Sharing predictions from different methods

- Quality standards, definitions of methods (ontologies)
- Stability of servers, updates, dependencies
- Reliability of assigned predictions

Several national infrastructures, including the Spanish National Bioinformatics Institute (INB) and a large European network (EMBRACE NoE) are dedicated to the construction of Web services and workflows and to the development of the underlying Web technology. These efforts, together with the work carried out in ontologies and classifications of bioinformatics methods, are important to increase the inter-relation of the resources (methods and databases) and to offer to bioinformaticians a better/more accessible range of tools.

With the aim of increasing the accessibility of an experimental biologist, without formal training in bioinformatics, a new interesting collection of clients based on Web technology (widgets) is being developed (e.g., see the Comparison of Approaches to Risk Governance—CARGO—project).

Infrastructure Needs Web Services

Technical difficulties

- Management, ontologies of services, and definition of methods
- Reliability, reliance, response time
- Computationally demanding services

Sociological difficulties

- Owner interfaces and additional services offered by local Web servers

- Who is the user? Profiles of user and user-machine interfaces
- Interfaces and user feedback

All these initiatives are paving the way to the new ESFRI, in particular to the proposal in Bioinformatics (ELIXIR) coordinated by Janet Thornton. One of the specific work packages in the preparatory phase of the project is dedicated to the organization of text repositories and text mining.

Finally, a new set of projects related with the integrated analysis of biological systems (SYS-BIO) and the development of infrastructure for synthetic biology (EMERGENCE CA) have recently started. In both cases the emphasis is very much in the creation of common standards and reusable infrastructures, including the connectivity with existing efforts. For example, the collaboration with the Massachusetts Institute of Technology repository of parts is one of the main objectives of EMERGENCE.

CAN WE DESIGN BIOLOGICAL SYSTEMS IN A PREDICTABLE MANNER?

Pam Silver (based on a transcript of her presentation)

Pam Silver's presentation focused on the value of models and design, with an illustrative example of building a system with predictable properties. She also provided an overview of the infrastructure needs for training scientists who are well versed in collaboration and mathematical modeling of biological systems.

The question of why it is necessary to make a predictable biology can be answered by saying that the redesign of a system can test our understanding of its components. Biology presents an array of engineering possibilities that have thus far been unexplored. Design concepts to be explored include sensation, signal processing and communication, modularity, and easy duplication. Biologists need to understand whether self-repair and evolvability are “bugs” or “features” of the system.

Dr. Silver then discussed biological modularity and pointed to examples of modularity including genes (e.g., promoters, open reading frames (ORFs), introns, and enhancers), RNA (e.g., translation, stability, export, and localization), proteins (e.g., targeting, DNA binding, dimerization, and degradation), and pathways (e.g.,

Infrastructure Needs of Systems Biology

signaling and metabolism). Biological design can test the limits of modularity.

Scientific challenges that biologists face today include whether functional components can be made and whether that component function can be measured quantitatively. The creation of higher order networks and the ability to predict their behavior is also a significant challenge.

Dr. Silver reviewed a recent case study of success in building cellular memory in eukaryotes as an example of the value of models and design study, broadening the example to discuss applicable lessons pertaining to the construction of networks using standardized parts.

With regard to training the next generation of scientists, Dr. Silver began by reviewing Harvard University's Systems Biology Ph.D. program, noting that students come from a wide range of disciplines including biology, computer science, mathematics, chemistry, physics, and engineering. Interdisciplinary approaches are used to address important biological and medical questions. Rather than instructing students in the state-of-the-art in systems biology, the Harvard program expects students to actually participate in its creation. Challenges and goals include:

- Can we enable collaboration and synergy amongst our students?
- Can we teach the biologists mathematical modeling?
- Can we teach the modelers to answer biological problems?

The challenge of defining a standard systems biology curriculum is driven largely by the fact that the field does not yet have unified principles, which inhibits the development of a coherent textbook. However, it is certain that “numbers matter,” that is, human intuitions about the behavior of complex biological systems are frequently wrong and mathematics helps us to draw accurate lessons. Quantitative observation leads to discovery.

Dr. Silver reviewed the strengths and weaknesses of two essential computational tools, Matlab 7 and Mathematica, noting that neither of these programs is free for academic use. Possible free alternatives include Octave and R.

The presentation concluded with a discussion of the learning curve for systems biologists. For biologists, the curve is learning how and when to use the available mathematical tools for modeling a problem when intuition is lacking,

e.g., in the cases of mRNA and protein. For modelers, the curve comes from understanding the physical principles of biology to help determine if the model is sound. She noted that all models are wrong to some degree, and must be based on sound principles and shown to be robust. Models offer the opportunity for a better understanding of the physical principles of biology.

MODELING AND ANALYSIS CHALLENGES IN BIOLOGY: FROM GENES TO CELLS TO SYSTEMS

Frank Doyle

The field of systems biology is receiving increasing attention as a paradigm shift in the life sciences, and also as a unifying discipline that links biology, chemistry, physics, mathematics, and engineering. The emphasis is on understanding the principles of regulation in biology from a “systems perspective,” that is the coordinated interactions of many components, as opposed to focusing on individual components (genes, proteins, cells, etc.). A recent benchmark study commissioned by NSF/NIH/DOE/NIST/DOD set forth the following definition for this discipline: “The understanding of network behavior through the application of modeling and simulation, tightly linked to experiment” (<http://wtcc.org/sysbio/report/SystemsBiology.pdf>). Despite the emphasis on networked behavior, and the promised goals of drug discovery, the early results in systems biology have largely focused on understanding behavior at a single horizontal level in these hierarchical networks. There are many “systems” studies of gene regulatory networks, and there are also many “systems” studies of protein networks (signal transduction, etc.). The vast majority of these studies have been intracellular and there are virtually no studies that link gene to protein to cell, and ultimately, to organism phenotype and behavior. Transcending the vertical directions in such networks, crossing orders of magnitude in component concentrations, system size, and time scale, is the ultimate challenge in multiscale systems biology. This is essential to realize the promise of improved medical treatments, and more important for the present context, is critical to advance the applications of methods from systems biology to the military arena, with profound impact for the performance of the soldier. One of the clear hurdles is the scaling of methods of simulation and analysis from the single level of gene regulation to the coordinated layers that underlie an entire organism.

MODEL-DRIVEN DISCOVERY

Markus Covert

Model-driven discovery is the use of computational modeling to describe a biological process and direct an experimental program. Briefly, a computational model of a particular biological network or process is reconstructed. This model is used to run simulations and quickly predict which experiments will be of the greatest value and which will yield comparatively trivial or redundant information. Focusing on the former class of experiments leads to relatively rapid identification of new network components and/or interactions. Two examples will be covered here, metabolism and transcriptional regulation in *Escherichia coli* (*E. Coli*) and the dynamics of NF- κ B signal transduction in mammalian cells.

First, the group compiled the first integrated genome-scale computational model of a transcriptional regulatory and metabolic network for *E. Coli*. The model accounts for 1,010 genes, including 104 regulatory genes whose products together with other stimuli regulate the expression of 479 of the 906 genes in the reconstructed metabolic network. The *E. Coli* model was used to drive an experimental program intended to elucidate the aerobic/anaerobic response in this organism, and found that this model was able to predict the outcomes of high-throughput growth phenotyping and gene expression experiments. Fey knockout strains were then constructed, Affymetrix gene chip technology was used to monitor gene expression. Then these results were compared to model predictions to determine many previously unknown interactions in the regulatory network. Based on these results, the team found that a model-driven discovery approach that combines genome-scale experimentation and computation can systematically generate hypotheses on the basis of disparate data sources.

More recently, model-driven discovery has been used to study the dynamics of the nuclear factor-kappa B (NF- κ B) signaling network under lipopolysaccharide (LPS) stimulation. NF- κ B is an important transcription factor family involved in a variety of diseases, notably many types of cancer. Understanding the specificity and temporal mechanisms that govern NF- κ B activation, as well as how NF- κ B evokes a transcriptional response, are therefore important in understanding cancer progression, and model-driven discovery is particularly well suited to address this issue.

Using a combined computational modeling/experiment approach, the group was able to characterize a previously unknown part of the Toll-IL-1 receptor (TIR) domain-containing adapter inducing interferon- β -dependent pathway, which led from LPS stimulation to NF- κ B activation. Surprisingly, the pathway depended on autocrine stimulation by tumor necrosis factor alpha (TNF α) and therefore identified unexpected crosstalk between the TNF α and LPS signaling pathways. This approach was also used to explain a complex behavior—the observed stable activation of NF- κ B under LPS stimulation—as the interaction between two signaling pathways.

More recently, a system has been built that enables quantitative characterization of the response of near-endogenous levels of NF- κ B in primary cells. This system was used to look at the contributions of single cells to a general behavior, with the finding that in the case of LPS stimulation, cells exhibit qualitatively different behaviors at the single-cell level. Current efforts center on identifying the cause of these different behaviors.

Efforts with NF- κ B and *E. coli* have led to certain observations about the practice of systems biology, which are grouped in terms of the four focus areas of the workshop. First, the entire process has been driven by new experimental tools and technologies. The group couldn't build a genome-scale model of *E. coli* metabolism until the genome had been sequenced, and likewise the transcriptional regulatory model was infeasible until the development of gene expression microarray technology. At the same time, it is interesting that in many cases, high-throughput data is being generated with little thought as to what knowledge will be extracted from it.

Databases are extremely useful on the whole, but the experience thus far is that they have not been used extensively to generate these models. This is largely because of concerns about the reliability of the data, and the finding that often the data required for these models is not easy to extract from a database. As a result, in building the models, the group generally searched the primary literature in detail (about 1.5 man-years of effort for regulation in *E. coli*) rather than relying primarily on databases (although RegulonDB was extremely helpful in pointing to the key literature).

Concerning the actual modeling process and applications, it has been found that many of the biggest “success stories”

Infrastructure Needs of Systems Biology

in model-driven discovery have used relatively simple math. One important result of this is that the math most likely could be explained to a typical life scientist. It will be important in the coming years to demonstrate to biologists that they truly can get involved in systems biology, and showing them some of the interesting results with simple mathematics will be a part of that effort. Unfortunately, very often in talks on systems biology the math is glossed over, even when it would be relatively simple to explain.

In summary, model-driven discovery is a powerful method that has been demonstrated in a few systems, and the integration of modeling with efforts in database development and high-throughput biological assaying will be of key importance in the coming years.

STANDARDIZATION EFFORTS FOR COMPUTATIONAL MODELING IN BIOLOG

Michael Hucka

Systems biology by its nature requires collaborations between scientists with expertise in biology, chemistry, computer sciences, engineering, mathematics, and physics. Successful integration of these disciplines depends on bringing to bear both social and technological tools: namely, consortia that help forge collaborations and common understanding, software tools that permit analysis of vast and complex data, and agreed-upon standards that enable researchers to communicate and reuse each other's results in practical and unambiguous ways.

An important prerequisite for effective sharing of computational models is reaching agreement on how to communicate them, both between software and between humans. The SBML project is an effort to create a machine-readable format for representing computational models at the biochemical reaction level. By supporting SBML as an input and output format, different software tools can operate on the same representation of a model, removing chance for errors in translation and assuring a common starting point for analyses and simulations. SBML has become the de facto standard for this purpose, with over 100 software tools supporting it today.

A recently created sister project is the Systems Biology Graphical Notation (SBGN) project. It addresses the issue of consistent human communication, by attempting to add more rigor and consistency to the graphical network diagrams that often accompany published research on

models of biological reaction systems. The real payoff will come when more people and software adopt such a common visual notation and it becomes as familiar to them as circuit schematics are to electronics engineers.

Finally, when developing and publishing computational models, it is only natural to want to put them into a database. The BioModels Database is a project to provide a free, centralized, publicly accessible database of human-curated computational models in SBML and other structured formats. The goal of providing a resource like BioModels Database led to other standardization efforts along the way, in particular MIRIAM and the Systems Biology Ontology (SBO).

The goal of MIRIAM is to provide guidelines for minimal information in a model for reference correspondence (the association of a machine-readable model with a human-readable description of what the model does) and for annotating a model with links to data resources (such as the identifiers of chemical substances listed in databases such as UniProt, ChEBI, etc.). It is hoped that in time, model authors will routinely provide this minimal level of annotation information, thereby making it easier to curate models in databases such as BioModels Database. SBO, on the other hand, is an ontology of terms suitable for explaining the mathematical semantics of each part of a model: its rate expressions in human terms such as “enzyme kinetics with competitive inhibition,” the meaning of parameters that enter into the rate expressions, the roles of participants in the expressions, and so on.

While computational systems biology is beginning to standardize and reach agreement on these many aspects of representations for predictive computational models, the acceptance of standard means of achieving interoperability between software has been lagging. Complex, general-purpose computer software standards such as Common Object Request Broker Architecture (CORBA) have fallen out of favor, and simplified frameworks such as the SBW and Bio-SPIICE, both of which were designed for systems biology applications, have in truth only seen limited levels of adoption from software developers. This is unfortunate, because greater use of application communication frameworks would benefit users in being able to work more easily with software tools. A coalition of open-source developers standardizing around a system such as SBW would be a welcome development for everyone.

Presentation Summaries

Development of software in academic environments is in many ways antithetical to the purposes of academic institutions. Robust, industrial-grade software requires testing, documentation, portability to different operating systems, packaging, distribution, managing user feedback, fixing bugs, and performing other maintenance operations—all of which are activities which rarely produce significant publications for the persons involved. There is therefore a strong disincentive for individual students, postdocs, and faculty members to go beyond doing the typical “research-grade” software implementations, and consequently, the effort and time and money that is put into these implementations is wasted because no one will ever be able to reuse the results without the documentation, packaging, distribution, etc., required for software to be usable by people other than its creators. One can argue that this is merely Darwinian survival of the fittest software, but the truth is that most software systems in these cases falls into disuse not because they are not useful or smartly developed, but because its original developers cannot afford to harden and maintain the software after its initial production.

This leads to a final question that merits further exploration: should there be a service organization whose job it is to take over software hardening and maintenance after the initial production by academic developers? Or if this is unsuitable, can we find other means of stopping the current waste of effort?

PATHWAY COMMONS: A PUBLIC LIBRARY OF BIOLOGICAL PATHWAYS

Gary Bader

Biological pathways represent knowledge about molecules, processes, and their interactions, the cellular wiring diagram. Researchers use pathway maps to design and analyze experiments and to make predictions about the behavior of biological systems. Pathway information is, unfortunately, extremely difficult for biologists to use in its current fragmented and incomplete state. The plan is to overcome this roadblock to biological research by developing the Pathway Commons Research Resource.

Pathway Commons benefits researchers as a convenient single point of access to diverse biological pathway information translated to a common data language (www.pathwaycommons.org). To provide this service, datasets will be aggregated from many existing pathway

databases; translate, store, validate, index, integrate, hyperlink, and maintain the information for maximum quality access; freely distribute pathway information to the scientific public, both academic and commercial; and, provide open-source end user software for pathway browsing and analysis. Pathway Commons promotes and supports convergence, by the community, to a truly integrated representation of cellular biological processes.

Using Pathway Commons, biology research groups will be able to quickly find and analyze information about cellular biological processes. Primary database groups will be able to share their pathway data, thus increasing access, reducing software development and curation costs, and avoiding duplication of effort. Bioinformatics software developers will be able to increase efficiency by sharing open-source software components.

The completion of the human genome sequence and advances in molecular technologies are driving biology towards increased use of computational tools. Pathway Commons will make biological knowledge available for computational processing and help create predictive computational models of biological processes, which will revolutionize many areas of biology, including health research.

Pathway Data Exchange Standards

BioPAX (Biological Pathway Exchange, www.biopax.org/) is a collaborative effort to create an OWL–XML data exchange format for biological pathway data. BioPAX covers metabolic pathways, molecular interactions, and protein post-translational modifications. Future versions will expand support for signaling pathways, gene regulatory networks, and genetic interactions. This is a large international collaborative effort involving many different academic and commercial groups. The format is open under the Lesser General Public License (LGPL).

Database Software

cPath is an open source database and Web application for collecting, storing, browsing, and querying biological pathway data. cPath makes it easy to aggregate custom pathway data sets available in standard exchange formats from multiple databases, present pathway data to biologists via a customizable Web interface, and easily export pathway data via a web service to third-party software for

Infrastructure Needs of Systems Biology

visualization and analysis. cPath is software only, and does not include new pathway information. The cPath software is freely available under the LGPL open source license for academic and commercial use.

http://cbio.mskcc.org/dev_site/cpath/

Demo: <http://cbio.mskcc.org/cpath/> (Protein-protein interactions)

Demo: <http://cancer.cellmap.org> (Pathways)

Analysis and Visualization

Cytoscape (www.cytoscape.org) is a bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data. Additional features are available as plugins. Plugins are available for network and molecular profiling analyses, new layouts, additional file format support and connection with databases. Plugins may be developed using the Cytoscape open Java software architecture by anyone and plugin community development is encouraged. Cytoscape was originally developed at the Institute of Systems Biology and is now a collaborative effort involving many different academic and commercial

groups. Cytoscape is freely available under the LGPL open source license for academic and commercial use.

Pathguide

Pathguide, the Pathway Resource List (www.pathguide.org/), contains information about hundreds of online biological pathway resources. Databases that are free and those supporting BioPAX, CellML, PSI-MI or SBML standards are highlighted.

LOOSELY COUPLED BIOINFORMATICS WORKFLOWS FOR SYSTEMS BIOLOGY

Douglas Kell

Progress in systems biology—or in “understanding complex systems”—depends on new technology, computational assistance, and new philosophy, but probably not in that order. Some developments include all three. The focus here is on the informatics needs, and these are illustrated with respect to the strategy for the Manchester Centre for Integrated Systems Biology (MCISB). The overall structure of the MCISB bottom-up systems biology pipeline is given in Fig. 2.

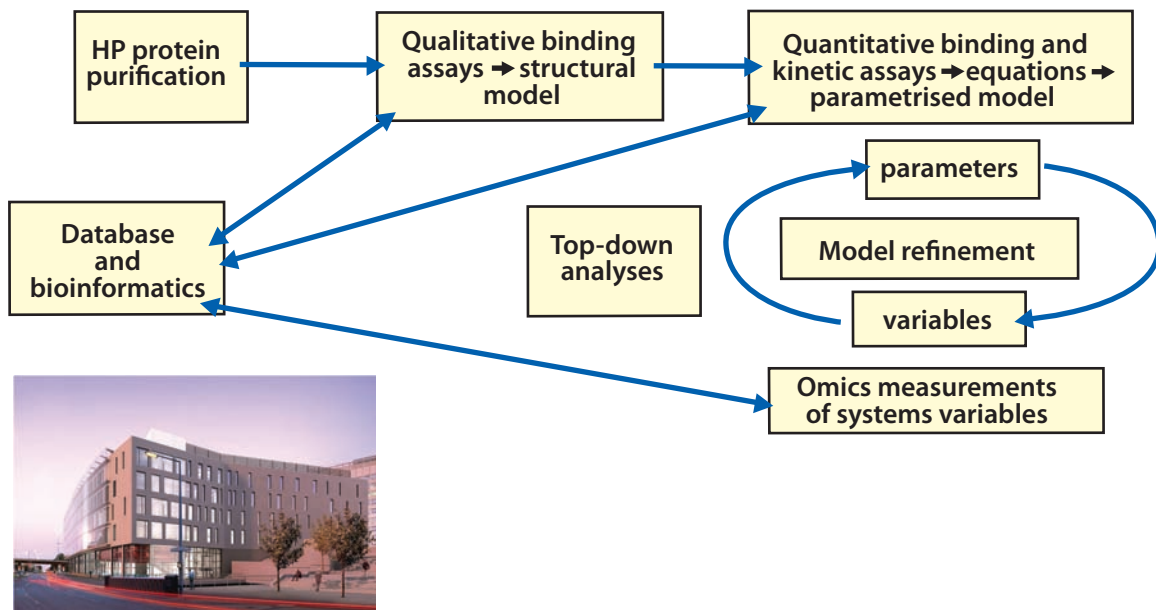


Figure 2. The overall structure of the systems biology agenda within MCISB.

Presentation Summaries

The basic problems are easily stated. There is a distributed set of data sources and software, much of it 'legacy' (i.e., funded by project grants that have now expired), and these need to be joined up to make suitable pipelines or workflows. Thus, in Manchester there are databases for genome, transcriptome, proteome, and metabolome data. An important part of the informatics strategy has involved the more or less automated building of consistent interfaces for both data entry and database interrogation, exploiting the fact that the XML schema of the database in question contains many of the facts necessary to do this.

For systems biology models, the focus at MCISB is their representation in SBML (www.sbml.org), and this is used as the starting point. There are then a great many things one might wish to do with this model, including creating it via text mining, storing it in a suitable database, running it in an environment such as Gepasi or Complex Systems Simulator (COPASI), performing sensitivity analyses of various kinds, and so on (Fig. 3).

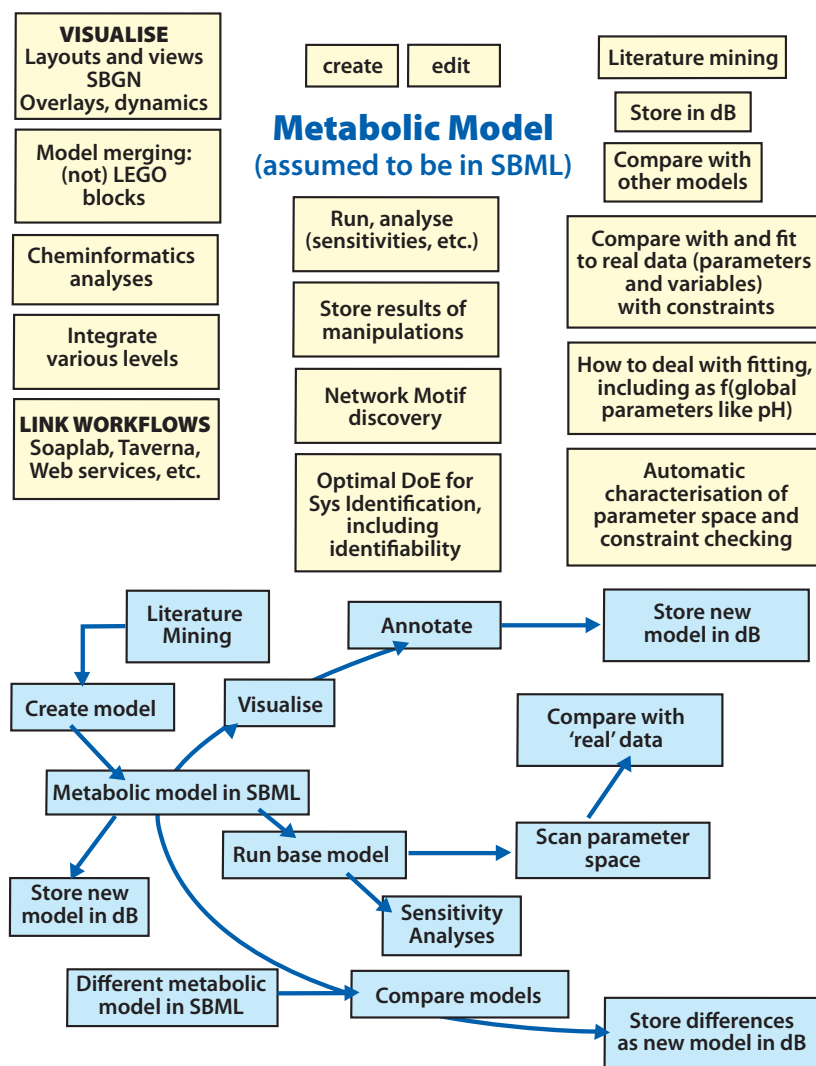


Figure 3. A variety of possible actions can be performed on, with and to systems biology models encoded in SBML (upper part), and strung together as workflows (lower part).

Infrastructure Needs of Systems Biology

The issues here, then, are that the producers and consumers of the elements or modules of these kinds of bioinformatic workflows are distributed and disparate, and somehow these need to be joined together. One might try and write a huge piece of software that would do all this in one place, with the data integrated in a data warehouse of some kind. However, such monolithic infrastructures follow an *integrate-in-anticipation* approach that is best suited to stable applications, in which the relevant data sets, data models and analyses are well established. As systems biology is rapidly evolving, it would be preferable to follow an *integrate-on-demand* approach, built around a loosely coupled collection of data and analysis resources, in which everyone continues to make available the

kinds of data and software for which they have the best expertise. This translates the problem into one of creating an environment with which to perform this coupling *loosely*. The key requirement for the producers of data and software modules then is that they make their material available programmatically via Web Services, preferably with a suitably rich description in Web Service Definition Language (WSDL).

The environment used for this kind of loose coupling is Taverna (<http://taverna.sourceforge.net/>) and this has the considerable attraction of a body of existing knowledge and, most importantly, the ability to reuse workflows. The overall architecture is given in Fig. 4.

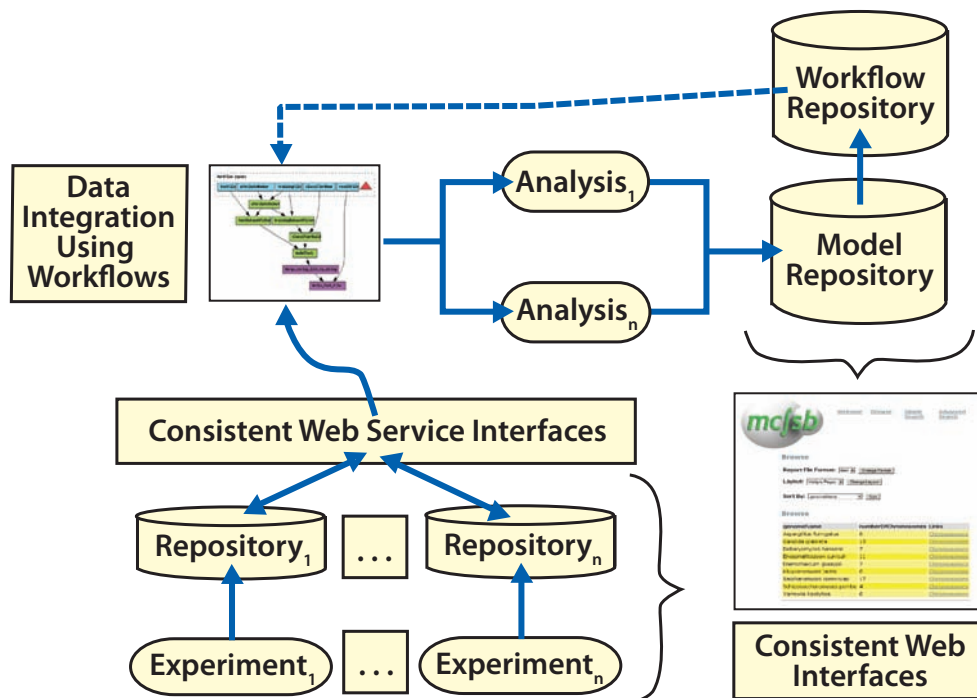


Figure 4. Overall architecture of the present state of the informatics infrastructure in MCISB.

Presentation Summaries

This approach has already been used successfully for the development of systems biology workflows, and the tasks now are to build up (and make available—possibly via www.myexperiment.org) a body of useful workflows, to make them as easy as possible for systems biologists to use, to exploit the rich semantic content and ontologies that can now be encoded in SBML, and to provide much improved means of network visualization.

This latter is especially significant as we enter the era of whole-genome network models (exemplified for instance by that for the human metabolic network recently published by Palsson and colleagues. Thus, in unpublished work to assess the available environments for analyzing and comparing genome-scale models of yeast the NCISB group has noted that Cell Designer (www.celldesigner.org) cannot deal at all with models of such scale, COPASI

(www.copasi.org) will load them but has no network visualisation tools, Cytoscape (www.cytoscape.org) can visualize them but not in any useful way since its more readable automatic layouts do not work for such large models, the Systems Biology Workbench (<http://sbw.sourceforge.net/>) simply but politely informs the user that the network is too large to be dealt with, and the Edinburgh Pathway Editor (www.bioinformatics.ed.ac.uk/epe/) does not *import* SBML. Other network visualization tools, such as Osprey do not speak SBML either.

Overall, the group wishes to develop and promote the idea that the data and software necessary for performing systems biology are best brought together in a loosely coupled fashion, and that Taverna represents a very suitable environment for performing this.

Infrastructure Needs of Systems Biology

Appendix A. Workshop Agenda

US-EC WORKSHOP ON INFRASTRUCTURE NEEDS OF SYSTEMS BIOLOGY

MAY 3-4, 2007

Tufts University Medical Center Campus
Boston, MA

Thursday

- 8:30 AM Welcome and Introductions
- Fred Heineken, NSF
 - Martha Steinbock, USDA
 - Søren Brunak, for the European Commission
- 9:00 AM Experimental Tools (Chair: Michael Ashburner)
- Matthias Uhlén: “A Human Protein Atlas”
 - Stefan Hohmann: “Measurements for Dynamic Modeling”
 - Michael Ashburner: “Ontologies for Data Integration”
- 10:30 AM Break
- 11:00 AM Databases (Chair: Rolf Apweiler)
- Rolf Apweiler: “Enabling Systems Biology: Development and Implementation of Proteomics Standards and Services”
 - Janet Thornton: “Towards a European Bioinformatics Infrastructure in the International Context”
 - Alfonso Valencia: “The Biosapiens NoE Experience in High-Throughput Genome Annotation and The ENCODE Pilot Project Case Story”
- 12:30 PM Lunch
- 2:00 PM Modeling Applications (Chair: Pamela Silver)
- Pamela Silver: “Can We Design Biological Systems in a Predictable Manner?”
 - Frank Doyle: “Modeling and Analysis Challenges in Biology: From Genes to Cells to Systems”
 - Markus Cover: “Model-Driven Discovery”



Infrastructure Needs of Systems Biology

3:30 PM	Break
4:00 PM	Software Infrastructure (Chair: Michael Hucka) <ul style="list-style-type: none">• Michael Hucka: “Standardization Efforts for Computational Modeling in Biology”• Gary Bader: “Pathway Commons: A Public Library of Biological Pathways”• Douglas Kell: “Loosely Coupled Bioinformatics Workflows for Systems Biology”
7:30 PM	Dinner at Les Zygomates Wine Bar and Bistro
Friday	
8:30 AM	Discussions and Conclusions
12:00/Noon	Lunch
12:30 PM	Discussions and Conclusions

Note: Adobe Acrobat (PDF) versions of slide presentations are available at www.wtec.org/ec-us_sysbio_workshop/. Underlined names in the agenda above are direct hyperlinks to the respective presentations of those participants.

Appendix B. Workshop Participants

Rolf Apweiler
European Bioinformatics Institute

Michael Ashburner
University of Cambridge

Gary Bader
University of Toronto

Judith Blake
The Jackson Laboratory

Søren Brunak
Technical University of Denmark

Andrea Califano
Columbia University

Marvin Cassman

Markus Covert
Stanford University

Frank Doyle
University of California, Santa Barbara

Fred Heineken
United States National Science Foundation

Maryanna Henkart
United States National Science Foundation

Michael Hucka
California Institute of Technology

Stefan Hohmann
Gothenburg University, Sweden

Douglas Kell
University of Manchester, United Kingdom

Yves Moreau
Katholieke Universiteit Leuven, Belgium

Chris Sander
Memorial Sloan-Kettering Cancer Center

Pamela Silver
Harvard Medical School

Martha Steinbock
United States Department of Agriculture

Janet Thornton
European Bioinformatics Institute

Matthias Uhlén
Royal Institute of Technology (KTH), Sweden

Alfonso Valencia
National Center for Biotechnology (CNB-CSIC), Spain

Steven Wiley
Pacific Northwest National Laboratory

Via conference call

Claire Tomlin
Stanford University



Infrastructure Needs of Systems Biology

Appendix C. Glossary

INFRASTRUCTURE NEEDS OF SYSTEMS BIOLOGY

AMPK	adenosine-5'-monophosphate-activated protein kinase
AMPKIN	Activated Protein Kinase
Bio-SPICE	Biological Simulation Program for Intra- and Inter-Cellular Evaluation
BioPAX	Biological Pathway Exchange
CARGO	Comparison of Approaches to Risk Governance
CellML	Cell Markup Language
COPASI	Complex Systems Simulator
CORBA	Common Object Request Broker Architecture
DOD	U.S. Department of Defense
DOE	U.S. Department of Energy
EBI	European Bioinformatics Institute
EC	European Commission
ELIXIR	European Life-Science Infrastructure for Biological Information
EMBL	European Molecular Biology Laboratory
EMBRACE	European Model for Bioinformatics Research and Community Education
ENCODE	ENCyclopedia Of DNA Elements project (U.S. National Human Genome Research Institute)
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
GelML	Gel markup language
GMOD	Generic Model Organism Database
GO	Gene ontology
HNRCA	Jean Mayer USDA Human Nutrition Center on Aging
HOG	High osmolarity glycerol
HPA	Human Protein Atlas
HPR	Human Proteome Resource
HUPO	Human Proteome Organization
ISB	Institute for Systems Biology



Infrastructure Needs of Systems Biology

KTH	Royal Institute of Technology in Stockholm, Uppsala
LGPL	Lesser General Public License
LPS	Lipopolysaccharide
MAPK	mitogen-activated protein kinase
MCISB	Manchester Centre for Integrated Systems Biology
MIAPE-MS	Minimum Information about a Proteomics Experiment in Mass Spectrometry
MIAPE	Minimum Information about a Proteomics Experiment
MIRIAM	Minimum Information Requested in the Annotation of Biochemical Models
mRNA	Messenger RNA
mzXML	mzData extensible markup language
NBT	Nature Biotechnology
NCBO	National Center for Biomedical Ontology
NF- κ B	nuclear factor-kappa B signaling network
NHGRI	National Human Genome Research Institute
NIGMS	National Institute of General Medical Sciences, NIH
NIH	U.S. National Institutes of Health
NIST	U.S. National Institute of Standards and Technology
NoE	Network of Excellence
NSF	U.S. National Science Foundation
PNAS	Proceedings of the National Academy of Sciences
PrEST	protein expressed sequence tag [note: PrEST typically denotes “protein epitope signature tag”]
PSI-MI	Proteomics Standards Initiative-Molecular Interactions
PSI	Proteomic Standards Initiative, HPO
QUASI	Quantifying Signal Transduction
SBML	Systems Biology Markup Language
SBO	Systems Biology Ontology
SBW	Systems Biology Workbench
TNF α	tumor necrosis factor alpha
TRANSFAC	Data Base of Eukaryotic Transcription Factors
Trif	Toll-IL-1 receptor domain-containing adapter inducing interferon- β
UCSC	University of California Santa Cruz
XML	Extensible markup language

