Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

CHALLENGES FOR MODELING AND SIMULATION METHODS IN SYSTEMS BIOLOGY

Herbert M. Sauro (Moderator)

Keck Graduate Institute
535 Watson Drive
Claremont, CA 91711, U.S.A.

Adelinde M. Uhrmacher (Moderator)

Department of Computer Science
University of Rostock
D-18051 Rostock, GERMANY

David Harel

Department of Computer Science and Applied Mathematics
The Weizmann Institute of Science
Rehovot 76100, ISRAEL

Michael Hucka

Biological Network Modeling Center (BNMC)
Beckman Institute M/C 139-74
California Institute of Technology
Pasadena, CA 91125, U.S.A.

Marta Kwiatkowska

School of Computer Science
University of Birmingham
Edgbaston, B15 2TT, U.K.

Pedro Mendes

Virginia Bioinformatics Institute
Virginia Tech
Blacksburg, VA 24061, U.S.A.

Clifford A. Shaffer

Department of Computer Science
Virginia Tech
Blacksburg, VA 24061, U.S.A.

Lena Strömback

Department of Computer and Information Science
Linköpings Universitet
S-581 83, Linköping, SWEDEN

John J. Tyson

Department of Biological Sciences
Virginia Tech
Blacksburg, VA 24061, U.S.A.

ABSTRACT

Systems Biology is aimed at analyzing the behavior and interrelationships of biological systems and is characterized by combining experimentation, theory, and computation. Dedicated to exploring current challenges, the panel brings together people from a variety of disciplines whose perspectives illuminate diverse facets of Systems Biology and the challenges for modeling and simulation methods.

1 INTRODUCTION

The goal of Systems Biology is to analyze the behavior and interrelationships of functional biological systems (Kitano 2002). Systems Biology is also characterized by the synergistic combination of experimentation, theory and computation. Challenges for the future lie in each of these areas. What follows is a series of commentaries from some of the leading researches in the field on some of the current and future challenges in systems biology.

1.1 The Critical Importance of Experimentalists

The application of reductionism in biology has proved to be a highly successful strategy and has enabled us to uncover the molecular details of biological systems in unprecedented detail. So successful has been this approach that there has been considerable skepticism as to the need for an alternative approach such as systems biology. The real test for systems biology is whether its application can generate novel biological insight that cannot be uncovered by pure reductionism.

Part of the problem lies in the wide gap that exists between the computationalists and the experimentalists. On the one hand, the computationalists are unaware of the difficulties in experimental work while the experimentalists in turn are unaware of the kinds of questions that computationalists could help them answer. Moreover, until recently generating the appropriate data to fuel the computationalists' appetite has proved extremely challenging for the experimentalists. However, most of the technology for making the required measurements is now in place. In particular, light microscopy of single cell dynamics is a reality and enables a researcher to track the concentration of a small number of proteins in real-time. Using light microscopy and cell counting techniques, large amounts of *high* resolution data on a *small* number of observables can be collected. This is in contrast to contemporary high-throughput approaches which collect *low* resolution data on *many* observables which is of limited use in building quantitative models.

Some of the most interesting and probably far reaching experiments done in recent years is the work being carried out by the growing single cell community, the work by Alon and colleagues (Lahav et al. 2004, Geva-Zatorsky et al. 2006) on p53 dynamics is extremely noteworthy in this respect. This work illustrates how a combination of experimentation and theory can lead to not only new insight but also opens up a whole new set of questions, one of the hallmarks of good science.

Although systems biology in its current reincarnation is now over six years old, the success of the discipline still remains patchy. There are noteworthy and significant success stories such as the work on p53 (Lahav et al. 2004, Geva-Zatorsky et al. 2006), the cell cycle models of Tyson (Sha et al. 2003) or the growing field of synthetic biology (Kaern and Weiss 2006), to name only a few. Another success has been the slow but steady diffusion of the importance of dynamics into the mainstream molecular biology community. For example, there is now much more emphasis placed on teaching dynamics to molecular biology students than ever before. All these developments are very welcome and are opening up entire new areas of discourse and research in the biology community. The reductionist approach will remain an essential part of biological research, but along side this the application of systems approaches should continue to be encouraged with particular emphasis on drawing in more experimentalists. The continued success of systems biology largely rests with engaging the experimentalists with the computationalists and therein lies the challenge.

1.2 Challenges in Theory Development

In the development of new theory, there are many interesting challenges to be met. Some of these include the following.

We will never be able to comprehend cellular networks in their entirety. In fact, viewing cellular networks in such a manner leads to the common remark that cellular networks are complex. Often, the complexity arises because we choose to see the entire network at once. In engineering disciplines, especially electrical engineering, large systems are modularized into distinct functional subsystems. Such subsystems carry out a well defined, and relatively easily understood function. By building a heterarchy of modules it is possible to rationalize a seemingly complex device.

The difficult question arises, what is a functional module in a biological network? There have been numerous discussions of this issue in the literature (Hartwell et al. 1999, Tyson et al. 2003, Wolf and Arkin 2003) and a number of common themes have emerged. A key idea is replacement, where a module can be replaced without disturbing the rest of the system behavior. With replacement comes the notion of an interface, where a module has a defined interface which is the point of contact between the module and the rest of the system. Finally, the number of contact points at a module interface will often be smaller than the number of interactions internal to the module. This latter aspect is of interest because it has been used to uncover modules in complex networks. In particular, a common metric (Newman and Girvan 2004) used to uncover topological modularity in networks is based on this very idea.

Many current approaches to modularization rely on topological modularity, whereas in fact networks should be functionally modularized. The problem of course is that it is not known how to functionally modularize a network, and therein lies a great challenge.

1.3 Computational Challenges

We will find the theoretical challenges reflected in development and requests for computational methods of modeling, simulation, and analysis.

If we assume that we can build biological systems in a modular manner, modeling formalisms are required that support this modularity, i.e., to compose models out of other models. Thus, we need to develop standard representations for building models from submodels and extending such approaches to multicellular systems. The issues surrounding standardization is largely sociological, technically many of the problems can be and have been resolved, e.g., (Cuellar et al. 2003, Hucka et al. 2003), the real issue is community acceptance. Many of the problems, for example, have been solved by the group who developed CellML and the associated MLs. However, the solutions are very complex and one wonders if a simpler approach is not possible. Also, the question of semantically correct re-use of models still looms large (Novre et al. 2005). The re-use of models requires understanding what information is needed to support reuse and how it should be presented, developing mechanisms to collect and record this information, understanding how to design for reuse, developing search

tools to locate model components, and developing criteria to decide when model reuse is desirable (Overstreet, Nance, and Balci 2002). Thereby, an extensive use of ontologies will only provide part of the answer.

To reduce a model's complexity, the level of detail at which subsystems are described might be chosen differently, leading to multi-level modeling (Uhrmacher, Degenring, and Zeigler 2005). If part of the model is described as a deterministic, continuous model and other parts as stochastic, discrete event models, reliable hybrid methods for combining continuous with stochastic models become necessary. Stochastic simulations of biological systems are known to be computationally intensive. Approaches towards addressing this problem include partitioning the model and using hybrid simulation methods, introducing improved scheduling algorithms, applying parallel and distributed simulation methods, or approximating future events (Burrage et al. 2005, Takahashi et al. 2004). The latter also leads to a more abstract view on the scheduled events.

In addition to simulation, the analysis of models is also important as it helps us explore the dynamics that are inherent in a model and to compare it with our knowledge. However, one of the most obvious gaps in the systems biologist tool box are reliable and user friendly analysis packages. Most prominent among the analysis techniques currently applied is bifurcation analysis (Doedel, Keller, and Kernevez 1991). However, analysis techniques based on verification methods are increasingly attracting attention. Questions of whether certain states can be reached and what the preconditions for certain behavior patterns are questions that can be answered in logic-based approaches (Kwiatkowska et al. 2006, Talcott 2006). However, these tools are currently only available for a small community of experts, to make those, in addition to simulation and animation tools, available in a user-friendly manner will be important in establishing modeling and simulation in wet-labs. It is to be expected that visualization techniques will play a central role in this endeavor.

2 A GRAND CHALLENGE: FULL REACTIVE MODELING OF A MULTI-CELLULAR ORGANISM BY DAVID HAREL

Biological systems can be modeled beneficially as reactive systems, using languages and tools developed for the construction of man-made systems. The Grand Challenge I proposed in 2003 (see Harel 2003, Harel 2005) is to model a full multi-cellular organism as a (hybrid) reactive system. I suggest the *C. elegans* nematode worm as a possible example of a fitting animal, which is complex, but very well-defined in terms of anatomy and genetics. The challenge is to construct a full, true-to-all-known-facts, 4-dimensional, fully animated model of the development and behavior of this worm (or of a comparable multi-cellular animal), which is multi-level and interactive, and is easily extendable - even by biologists - as new biological facts are discovered. The proposal has three premises:

- That satisfactory frameworks now exist for reactive system modeling and design.
- That biological research is ready for an extremely significant transition from analysis (reducing experimental observations to elementary building blocks) to synthesis (integrating the parts into a comprehensive whole).
- That the true complexity of the dynamics of biological systems - specifically multi-cellular living organisms - stems from their reactivity.

In the last seven or eight years I have been working with students and colleagues on exhibiting the feasibility of modeling biological systems as reactive systems, and the results are very encouraging. We have done work on T-cell development in the thymus, vulval cell fate determination in *C. elegans*, embryonic development of the pancreas, development of the lymph node, and generic cell behavior (sample publications Efroni et al. 2003, Kam et al. 2002, Fisher et al. 2005, Swerdlin et al. 2006).

Achieving this Grand Challenge could result in an unprecedented tool for the research community, both in biology and in computer science. We feel that much of the research in systems biology will be going this way in the future: grand efforts at using computerized system modeling and analysis techniques for understanding complex biology. And I truly believe that computer science, and especially ideas from systems and software engineering, will play a role in life sciences research of the 21st century similar to the role played by mathematics in the physics of the 20th century.

3 SCALING UP THE ANALYSIS – CAN WE USE COMPOSITIONAL REASONING? BY MARTA KWIATKOWSKA

In the context of biological processes, the term 'complexity' refers not only to non-linearity and emergent behavior, but also the sheer size of the systems as measured by the number of components and the complex pattern of interactions between them. Though computational modeling in biology has made tremendous progress in recent years, the problem of scalability of the techniques to models of realistic size remains a major challenge. The growth in complexity is often exponential, and arises independently of the representation: for example, in signaling pathways with parallel state changes the number of differential equations grows exponentially with the number of molecules, as does the number of global system states in discrete stochastic models. This pattern of increase in complexity cannot

be simply addressed by enlarging the capacity of existing machinery, but instead calls for a paradigm shift and more sophisticated techniques.

Process calculi such as the stochastic pi-calculus (Priami et al. 2001, Regev and Shapiro 2004) have recently been proposed as a convenient modeling framework for biological processes, since they support a natural decomposition into concurrently interacting modules, for example, proteins reacting with other molecules, in a manner that enables population- and individual-based models. An important advantage of process calculi is that, in addition to conventional analysis by simulation, they admit automated verification and falsification of models using techniques such as probabilistic model checking (Rutten et al. 2004). Process calculi are inherently compositional, though compositionality is usually only exploited for model description and construction, not analysis. Thus, for example, a replacement of a module with a smaller but provably equivalent one is possible, but there is limited support for compositional quantitative analysis that enables the derivation of properties of the composed system based on the analysis of individual components.

Compositional verification frameworks have been proposed in the mid-80s (Pnueli 1985), but the results have remained largely theoretical until recently. A well-known paradigm for compositional reasoning is assume-guarantee: starting from a decomposition of a system into components, we verify each component separately by making assumptions about its environment, and then discharge the assumptions for the parallel composition, thus avoiding the need to build a representation of the full model. This reasoning is frequently circular: A is verified under the assumption that the environment B behaves as expected, and, symmetrically, B is verified under the assumption that A behaves as expected, but non-circular rules are also available. Much progress has been made in automation of compositional verification for the qualitative analysis of large-scale systems (Cobleigh, Giannakopoulou, and Pasareanu 2003). Unfortunately, these techniques are less well developed in the case of stochastic, and more generally quantitative, modeling frameworks where only partial solutions exist, for example product-form (Hillston 2005); see also (Chatterjee et al. 2006). Future progress in modeling and rigorous analysis of real-world biological processes can only be made through advances in quantitative compositional analysis and automation of the techniques.

4 THE CURRENT STATE AND FUTURE PROSPECTS FOR STANDARDIZATION BY MICHAEL HUCKA

With the increasing interest in computational modeling in biology today, there has come an increasing awareness that continued intellectual progress demands better sharing of data, models, software and knowledge. In recent years, this has lead to a surge in efforts to establish common standards. The fact that many of these efforts have become prominent lately is not surprising for the fact that it is happening—after all, standards for common information exchange are a facet of virtually every other human endeavor—but perhaps more surprising for its apparent suddenness. However, the timing is probably easily explained simply by the combination of critical mass (enough people now feel the pain of not having standards) and the proliferation of information technology in life sciences in the last decade.

Efforts to standardize model representation languages — e.g., SBML (Hucka et al. 2003, Finney et al. 2006), CellML (Hedley et al. 2001, Cuellar et al. 2003), BioPAX (Bader et al. 2005, Strömbäck and Lambrix 2005) — have been among the most successful. The Systems Biology Markup Language (SBML) has had widespread acceptance in its domain of mathematical models of biochemical networks, and BioPAX is now emerging as the leading standard for representing pathways for exchange between databases. Both show signs of continued increasing adoption in the near future. Both are software-level standards, not intended for direct human consumption. For the latter, the Systems Biology Graphical Notation (SBGN) (Kitano et al. 2005, SBGN Team 2006) is a very recent effort to begin working on standardizing the icons and other visual notations used in biological network diagrams. Given the natural predisposition humans have towards using visual diagrams, SBGN stands to garner more enthusiasm (and controversy) than any of the other efforts at standardizing model representations.

If one has agreement on model representations (SBML, CellML, or other), a natural next step is to want a centralized database where models can be stored and found. This is also a crucial enabler for scholarly publications, whose editors can recommend that authors deposit their models in the database much as is done for sequences in sequence databases. BioModels Database (Le Novère et al. 2006, BioModels Database Team 2006) is the now leading frontrunner in this area, having gained acceptance from Nature and PloS so far, with more sure to come. Among the standout features of BioModels Database are the employment of human curators who verify and annotate every model, and the use of a relational database allowing much more sophisticated searches than would be possible in simpler repositories of models.

In the software interoperability domain, efforts have been somewhat less successful. Complex, general-purpose computer software standards such as CORBA have fallen out of favor, and simplified frameworks such as the Systems Biology Workbench (SBW) (Sauro et al. 2003) and Bio-SPICE (Kumar and Feidler 2003), 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 com-

munication 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.

It is worth noting that the successful efforts in these areas have all been bottom-up, with communities forming around common needs. This is likely to continue for some time, as the rate of innovation and development of new areas continues rapidly. As Quackenbush (2006) and others noted recently, top-down standardization requirements are almost certainly doomed to failure. Multiple special-purpose standards developed by the communities who need them, seem to be the direction in which we are headed for the near future.

5 THE ROLE OF TOP-DOWN STUDIES IN BUILDING COMPUTATIONAL MODELS BY PEDRO MENDES

The construction of models of biochemical and cellular behavior has been traditionally carried out through a bottom-up approach, which is essentially a process of synthesis that combines *in vitro* enzyme kinetic data and knowledge of a reaction network to produce a dynamic model of the same network. This process requires that the reaction network be known and that it is possible to carry out the various enzymatic reactions *in vitro*. This process of modeling carries the implicit assumption that the reactions *in vitro* proceed in a manner similar to *in vivo*, an assumption that is not always correct. Such bottom-up modeling is thus a process of synthesis that collects a great deal of detailed enzyme kinetic or protein-protein binding data to form a dynamic model of a biochemical network.

While bottom-up modeling has been a very successful methodology, it is unfortunately not possible to apply it in many circumstances. Conditions in which this method is not appropriate or feasible are: a) when the reaction network is not well known, such as in some signaling pathways or in secondary metabolism; b) when the purified proteins lost significant interaction partners (e.g., other proteins that are bound to them in the cell and alter their function); c) when substrates are not available in purified form, which is unfortunately a common situation. On the other hand, modern genomic, proteomic and metabolomic advances mean that the most abundant data sets are composed of snapshots of cellular states at the molecular level. When these data are obtained as time series they are trajectories that reflect the cellular process of interest. Since these data are now easier to obtain than traditional purified enzymological assays, at least in the conditions listed above, there is a great need to use these data for the construction of dynamic models. Indeed, it could be argued that such a *top-down* modeling strategy is closer to the spirit of systems biology exactly because it makes use of systems-level data, rather than having originated from a more reductionist approach of molecular purification. Both these approaches are useful, though, and I prefer to think of them as complementary approaches, since each of them has their own advantages as well as disadvantages.

The top-down modeling approach is essentially a systems identification problem, also known as an inverse problem. One is presented with the behavior of a system and from that one desires to infer which molecules are involved in interactions (network structure), how these interactions proceed (kinetic laws), and by how much (kinetic parameter values). Top-down modeling is an active area of research and it would be fair to say that it is currently without a general working solution. Despite the fact that there is an abundance of publications on network reverse engineering (e.g., Arkin and Ross 1995, de la Fuente et al. 2002, Gardner et al. 2003, Laubenbacher and Stigler 2004), it is fair to say that none of them have yet been applied to a real set of "omic" data and unraveled a new biochemical pathway and its dynamics. This reflects the difficulty of the process, indeed a common feature of all inverse problems. The biochemical network inverse problem is possibly one of the hardest of all, due to the inherent nonlinearity of biological systems. This is complicated by the nature of current "omic" experiments, where the measured variables largely outnumber the number of samples (states) collected. This, in turn, leads to a severe under-determination that implies that only a few of the variables can be used for model construction.

It is appropriate to enumerate here the specific issues involved in top-down model construction. These are technical problems that are in need of solutions if we are to be able to use "omic" data directly for model construction. Given a set of trajectories of the biological system in response to environmental or genetic perturbations:

- Select a set of variables from the trajectories that will serve as the basis for the modeling process; these have to reflect the process of interest and one hopes that they are indeed directly involved in it. (Note that this is also an inverse problem in itself!)
- Given the trajectories and a selection of variables, find out how these variables are related to each other, ultimately representing the network of interactions and/or reactions in which they are involved.
- Given the trajectories, a set of variables, and the network formed by them, identify the rate laws (or other transfer functions) that characterize the interactions that compose the network. The result of this process will often be in the form of a set of equations (e.g., ordinary differential equations).
- Given a set of equations that characterizes a network of some variables from measured trajectories of the system, identify the numerical values of the

equation parameters that best describe the system (goodness of fit). This is often referred to as the parameter estimation problem. Often, this step may have to be solved simultaneously with the previous one.

These four steps in the top-down modeling process are described here in generic terms, but will take specific forms depending on the mathematical modeling formalism used. Often, this is in the form of ordinary differential equations, but the top-down process is general and is equally applicable with other formalisms. Their solution is likely to involve optimization algorithms, since the solution to inverse problems can generally be stated as an optimum of an objective function. The parameter estimation problem is the one that is better studied (Mendes and Kell 1998) and is commonly addressed as a maximization of a likelihood function, which usually translates into the minimization of the difference between model and observations. It is an open question whether the discrete modeling steps enumerated above are amenable to be solved independently or if they need to be solved simultaneously. While parameter estimation can be carried out independently when the correct rate laws of a model are known, it is probably not possible to infer the rate laws themselves without simultaneously estimating their parameter values.

It is clear that the development of a robust methodology for top-down modeling is one of the grand challenges of systems biology. The solution of the four problems enumerated above will bring closer the realization of that goal.

6 FROM WET LAB DATA TO COMPUTER SIMULATION: PROBLEMS IN CELL CYCLE MODELING
BY CLIFF SHAFFER AND JOHN TYSON

In the field of molecular systems biology, computational models should be relevant to a defined set of experimental observations that provide information on the molecular machinery underlying some aspect of cell physiology. This experimental data set usually includes some combination of biochemical measurements and physiological observations on wild-type and mutant cells under normal and "perturbed" conditions. The relevant data are usually quite diverse, including accurate quantitative measurements (e.g., the half-life of protein X is 13 ± 2 minutes), reliable qualitative observations (e.g., mutants *a* and *b* are viable, but the double mutant *ab* is inviable), and imprecise observations (e.g., enzyme *Y* is much less active under conditions *P* compared to conditions *Q*). From this collection of information, it is the modeler's job to devise an appropriate mathematical model that is reasonably consistent with the available data, that provides some new insights into the underlying molecular mechanism, and that makes useful predictions about novel experimental studies of the system.

In building and testing such models, three issues must be kept in mind. First, the model must be "bounded." That is, the modeler must decide what parts of the cell's molecular machinery are to be included in the model, and then how the rest of the cell's physiology serves as boundary conditions of the model (inputs, demands, etc.). In this sense, the model is a "module" with a well-defined interface to the remainder of the cell's internal machinery. The model itself might be composed of sub-modules, and it might later become a sub-module of a larger model that covers a larger collection of data.

Second, the model must be "appropriate" to the available data. It must contain variables that connect to all the available observations on genes, proteins, metabolites, etc. If the model is too simple, it will not be able to account for the available data. If it is too complex, there will be insufficient experimental observations to constrain the model. If the model is "appropriate," then it should be possible to estimate the parameters of the model from the available data, and to have some data "left over" to test the model. Also, an appropriate model should be able to successfully predict the outcome of novel experiments within the confines of that part of the cell's physiology being modeled.

Third, when building models, brute-force simulations are usually insufficient to make progress, because the parameter space that must be searched, even for models of moderate complexity, is enormous. The modeler needs some analytical tools to explore the mechanisms first in qualitative terms: steady state analysis, bifurcation analysis, sensitivity analysis, network analysis, etc. These tools help to define the general capabilities of a model and to delineate regions in parameter space where the model is likely to be successful in explaining the experimental data.

Our experience modeling eukaryotic cell cycle regulation illustrates these issues. In the early 1990's, only a few parts of the molecular regulatory system were known (CDK1, cyclin B, APC, Wee1, Cdc25), and the first models were primitive but effective (Novak and Tyson 1993). Later, as more genes and protein interactions were discovered in the wet lab, the models became increasingly more complex, sophisticated and successful, building incrementally on the limited successes of earlier models. The most complete model to date, for the basic cell cycle engine in budding yeast (Chen et al. 2004) is composed of over 30 ODEs, involving about 140 rate constants, and constrained by the observed phenotypes of 130 mutant strains with aberrations in different genes of the control system.

This level of complexity stretches the ability of experienced and dedicated modelers to build, analyze, simulate, verify, and test their models by hand. For example, when a model fails to account for all the observations in the

experimental data set (as is almost always the case!), the modelers must determine where the problem lies:

1. With the parameter set?
2. With the model itself? Maybe the molecular wiring diagram is incorrect. Maybe some crucial molecular interactions have been left out of the model.
3. With the experimental data? Maybe the unfitted experimental observations are mistaken in some way.

To make this decision, modelers need efficient software tools for exploring parameter space automatically, for modifying wiring diagrams quickly and accurately, and for analyzing and simulating equations easily and reliably.

To these ends, we and other research groups (see for example Sauro et al. 2003, Copasi 2006) have been developing software for building models, analyzing them, running suites of simulations, comparing simulations to available data, and automated parameter estimation. Our tools include the JigCell Model Builder (Vass et al. 2006) for creating and editing models, the JigCell Run Manager (Allen et al. 2003) for organizing the various mutants, and the JigCell Comparator (Allen et al. 2003) for analyzing the goodness of fit between the experimental data and the simulation outputs. Such tools allow for the automation of model validation procedures. Automated model validation allows the modeler to institute validation checking early into the model lifecycle, and to cheaply validate the model at each step in its development. Our Parameter Estimation Tool (Zwolak 2006) supports automatic exploration of parameter space by local gradient-based optimization and by global deterministic search algorithms (Panning et al. 2006). Oscill8 (Conrad 2006) is a user-friendly environment for exploring the bifurcation structures of a model.

As has been stated earlier, ultimately models will be too complex to understand without some form of structuring into units. Model composition, where models are decomposed into structural parts, will be required. Such parts must be understood in terms of their interfaces to other parts. We examine techniques for such modeling in greater detail in (Shaffer, Randhawa, and Tyson 2006) in these proceedings.

7 DATABASES, SCHEME-MATCHING, AND ONTOLOGIES – FROM BIOINFORMATICS TO COMPUTATIONAL SYSTEMS BIOLOGY BY LENA STRÖMBÄCK

One important goal for systems biology is a complete understanding and description of the protein interaction network underlying cell physiology, and how these processes contribute to the function of the cells and organisms (Collins et al. 2003; Hermjakob et al. 2004). To achieve this goal, it is important that researchers can access and reuse each other data from results from single experiments to models for analysis and simulations in a transparent way. The tradition within the field is to publish results in databases available on the Internet which makes the field unique by making large quantities of data available. However, to strive towards more automatic processing, there is a large need for development of standardized descriptions, methods for integration of data, and software components capable to work on several standards.

Descriptions, or formats, for exchange of data have developed from formats aimed at export of information from one particular tool or database towards standardized descriptions of how to represent information within a particular area. SBML (Hucka, Finney, and Sauro 2003), PSI MI (Hermjakob, Montecchi-Palazzi, and Bader 2004), and BioPAX (BioPax 2006) are good examples of this. In parallel to this there has been a development of biomedical ontologies to allow standardization of concepts, e.g., GO (Ashburner, Ball, and Blake 2000), and OBO (OBO 2006). Currently, there is a merge of efforts where many of the standards make use of ontologies. This can either be done by making references to existing ontologies or by specifying controlled vocabularies as part of the standard.

There is a large difference in scope between available standards for systems biology (Strömbäck and Lambrix 2005). This is visible in terms of which concepts they cover but also in terms of purpose of the formalisms, i.e., whether the standard is intended for the recording of results, models for simulation, or something else. This purpose determines which terminology and sets of attributes are provided for every concept within a standard. Standards that have been created for a particular and well-defined purpose have often been more popular than general standards, meaning that also in the future it is probable that there will be parallel standards with different purposes. This means that for a complete understanding of the area, technology for schema matching and alignments of ontologies will be of importance (Lambrix and Tan 2005, Strömbäck 2006). Here, the integration of ontology concepts within standards is an important aid for matching and alignment of datasets.

There are currently many tools for analysis and simulation of systems biology data. For data management and storage, there is a limited number of specialized tools and traditional database technology is a good option. For standards implemented in XML, there are in principle two options, either a translation of data to a traditional relational database, or to use the newer XML-database approach. The latter has the benefit of allowing direct access of data on the XML representation via the query language XQuery (Strömbäck 2005, Strömbäck and Hall 2006). This technology does, however, require a detailed knowledge about the standard from the user, which in many cases can be a drawback if the user needs to work on data available in different standards.

As a summary, we can see that the development of new standards and ontologies within systems biology is very important for reaching the goal of complete understanding of interactions networks. For this, there is a need of transparent flow of data from experiments to models which is supported by recent development within the semantic web and database community. This is though only a start; to reach the final goal, further development within all fields discussed here is required.

REFERENCES

Allen, N., C. Shaffer, M. Vass, N. Ramakrishnan, and L. Watson. 2003. Improving the development process for eukaryotic cell cycle models with a modeling support environment. *Simulation* 79:674–688.

Arkin, A., and J. Ross. 1995. Statistical construction of chemical-reaction mechanisms from measured time-series. *Journal of Physical Chemistry* 99 (3): 970–979.

Ashburner, M., C. A. Ball, and J. A. Blake. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25:25–29.

Bader, G., E. Brauner, M. Cary, R. Goldberg, C. Hogue, P. Karp, T. Klein, J. Luciano, D. Marks, N. Maltsev, E. Marland, E. Neumann, S. Paley, J. Rick, A. Regev, A. Rzhetsky, C. Sander, V. Schachter, I. Shah, and J. Zucker. 2005. BioPAX–Biological Pathways Exchange language. Level 1, Version 1.4 documentation. Available on the Internet at <http://www.biopax.org>.

BioModels Database Team. 2006. BioModels Database. Available on the Internet at <http://www.ebi.ac.uk/biomodels/>.

BioPax 2006. <http://www.biopax.org>.

Burrage, K., P. Burrage, N. Hamilton, and T. Tian. 2005. Compute intensive simulations for cellular models. In *Parallel Computation for Bioinformatics and Computational Biology*, ed. A. Y. Zomaya. Wiley.

Chatterjee, K., L. de Alfaro, M. Faella, T. Henzinger, R. Majumdar, and M. Stoelinga. 2006. Quantitative compositional reasoning. In *Proceedings of the 3rd International Conference on Quantitative Evaluation of Systems (QEST'06)*: IEEE CS. To appear.

Chen, K., L. Calzone, A. Csikasz-Nagy, F. Cross, B. Novak, and J. Tyson. 2004. Integrative analysis of cell cycle control in budding yeast. *Molecular Biology of the Cell* 15 (8): 3841–3862.

Cobleigh, J., D. Giannakopoulou, and C. Pasareanu. 2003. Learning assumptions for compositional verification. In *Proceedings of the 9th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'03)*, 331–346.

Collins, F. S., E. D. Green, and A. E. Guttmacher. 2003. A vision for the future of genomics research: a blueprint for the genomic era. *Nature* 422:835–847.

Conrad, E. 2006. Oscill8. <http://oscill8.sourceforge.net/>.

Copasi. 2006. <http://www.copasi.org/tiki-index.php>.

Cuellar, A., C. Lloyd, P. Nielsen, D. Bullivant, D. Nickerson, and P. Hunter. 2003. An overview of CellML 1.1, a biological model description language. *Simulation: Transactions of the Society for Modeling and Simulation International* 79 (12): 740–747.

de la Fuente, A., P. Brazhnik, and P. Mendes. 2002. Linking the genes: inferring quantitative gene networks from microarray data. *Trends in Genetics* 18:395–398.

Doedel, E., H. Keller, and J. Kernevez. 1991. Numerical analysis and control of bifurcation problems in finite dimensions. *International Journal of Bifurcation and Chaos* 1:493–520.

Efroni, S., D. Harel, and I. Cohen. 2003. Towards rigorous comprehension of biological complexity: modeling, execution and visualization of thymic T cell maturation. *Genome Research* 13:2485–2497.

Finney, A., M. Hucka, B. Bornstein, S. Keating, B. Shapiro, J. Matthews, B. Kovitz, M. Schilstra, A. Funahashi, J. Doyle, and H. Kitano. 2006. Software infrastructure for effective communication and reuse of computational models. In *System Modeling in Cellular Biology*, ed. Z. Szallasi, J. Stelling, and V. Periwal, Chapter 17. MIT Press.

Fisher, J., N. Piterman, E. Hubbard, M. Stern, and D. Harel. 2005. Computational insights into C. elegans vulval development. *Proceedings of the National Academy of Sciences* 6 (102): 1951–1956.

Gardner, T. S., D. di Bernardo, D. Lorenz, and J. J. Collins. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301 (5629): 102–105.

Geva-Zatorsky, N., N. Rosenfeld, S. Itzkovitz, R. Milo, A. Sigal, E. Dekel, T. Yarnitzky, Y. Liron, P. Polak, G. Lahav, and U. Alon. 2006. Oscillations and variability in the p53 system. *Molecular Systems Biology* 2:2006–2006.

Harel, D. 2003. A grand challenge for computing: towards full reactive modeling of a multi-cellular animal. *Bulletin of the European Association for Theoretical Computer Science (EATCS)* (81): 226–235. (Reprinted in *Current Trends in Theoretical Computer Science: The Challenge of the New Century*, Algorithms and Complexity, Vol I (Paun, Rozenberg and Salomaa, eds.), World Scientific, pp. 559-568, 2004).

Harel, D. 2005. A turing-like test for biological modeling. *Nature Biotechnology* 23:495–496.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray. 1999. From molecular to modular cell biology. *Nature* 402 (6761 Supplement): C47–C52.

Hedley, W., M. Nelson, D. P. Bullivant, and P. Nielson. 2001. A short introduction to CellML. *Philosophical Transactions of the Royal Society of London A* 359:1073–1089.

Hermjakob, H., L. Montecchi-Palazzi, and G. Bader. 2004. The HUPO PSI's molecular interaction format - a community standard for the representation of protein interaction data. *Nature Biotechnology* 22:177–183.

Hillston, J. 2005. Process algebras for quantitative analysis. In *Proceedings of the Twentieth Annual IEEE Symp. on Logic in Computer Science, LICS 2005*, ed. P. Panangaden, 239–248: IEEE Computer Society Press.

Hucka, M., A. Finney, and H. M. Sauro. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19:524–531.

Hucka, M., A. Finney, H. M. Sauro, H. Bolouri, J. Doyle, H. Kitano, A. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19 (4): 524–531.

Kaern, M., and R. Weiss. 2006. Synthetic gene regulatory systems. In *System Modeling in Cellular BiologyFrom Concepts to Nuts and Bolts*, ed. Z. Szallasi, J. Stelling, and V. Periwal, Chapter 13, 269–298. MIT Press.

Kam, N., D. Harel, H. Kugler, R. Marelly, A. Pnueli, E. Hubbard, and M. Stern. 2002. Formal modeling of C. elegans development: a scenario-based approach. In *Proceedings of the 1st International Workshop on Computational Methods in Systems Biology (ICMSB 2003)*, Number 2602 in LNCS, 4–20: Springer. (Revised version in *Modeling in Molecular Biology* (G. Ciobanu and G. Rozenberg, eds.), Springer, Berlin, 2004, pp. 151-173.).

Kitano, H. 2002. Systems biology: a brief overview. *Science* 295 (5560): 1662–1664.

Kitano, H., A. Funahashi, Y. Matsuoka, and K. Oda. 2005. Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology* 23 (8): 961–966.

Kumar, S., and J. Feidler. 2003. BioSPICE: A computational infrastructure for integrative biology. *OMICS* 7 (3): 225.

Kwiatkowska, M., O. Tymchishyn, G. Norman, E. Gaffney, and J. Heath. 2006. Computational modelling of signalling pathways: comparing simulation, verification and differential equation approaches. In *Proceedings of the 2006 Winter Simulation Conference*, ed. L. Perrone, F. Wieland, J. Liu, B. Lawson, D. Nicol, and R. Fujimoto. Monterey, California, USA.

Lahav, G., N. Rosenfeld, A. Sigal, N. Geva-Zatorsky, A. Levine, M. Elowitz, and U. Alon. 2004. Dynamics of the p53-mdm2 feedback loop in individual cells. *Nature Genetics* 36 (2): 147–150.

Lambrix, P., and H. Tan. 2005. A framework for aligning ontologies. *Current Trends in Theoretical Computer Science: The Challenge of the New Century, Algorithms and Complexity* 3703:17–31.

Laubenbacher, R., and B. Stigler. 2004. A computational algebra approach to the reverse engineering of gene regulatory networks. *Journal of Theoretical Biology* 229 (4): 523–537.

Le Novère, N., B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J. Snoep, and M. Hucka. 2006. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research* 34 (suppl-1): D689–691.

Mendes, P., and D. Kell. 1998. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 14:869–883.

Newman, M., and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E Statistcal, Nonlinear, and Soft Matter Physics* 69 (2 Pt 2): 026113–026113.

Novak, B., and J. J. Tyson. 1993. Numerical analysis of a comprehensive model of m-phase control in xenopus oocyte extracts and intact embryos. *Journal of Cell Science* 106 (Pt 4):1153–1168.

Novre, N., A. Finney, M. Hucka, U. Bhalla, F. Campagne, J. Collado-Vides, E. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J. Snoep, H. Spence, and B. Wanner. 2005. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology* 23 (12): 1509–1515.

OBO. 2006. Open biomedical ontologies. <`http://obo.sourceforge.net`>.

Overstreet, C. M., R. E. Nance, and O. Balci. 2002. Issues in enhancing model reuse. In *International Conference on Grand Challenges for Modeling and Simulation*. San Antonio, Texas, USA.

Panning, T., L. Watson, N. Allen, C. Shaffer, and J. Tyson. 2006. Deterministic parallel global parameter estimation for a model of the budding yeast cell cycle. In

Proceedings of the 2006 High Performance Computing Symposium (HPC'06), 195–201.

Pnueli, A. 1985. *In transition from global to modular temporal reasoning about programs*. 123–144. New York: Springer-Verlag.

Priami, C., A. Regev, E. Shapiro, and W. Silverman. 2001. Application of a stochastic name-passing calculus to representation and simulation of molecular processes. *Information Processing Letters* 80:25–31.

Quackenbush, J. 2006. Top-down standards will not serve systems biology. *Nature* 440:24.

Regev, A., and E. Shapiro. 2004. The pi-calculus as an abstraction for biomolecular systems. In *Modelling in Molecular Biology*, 219–266. Springer.

Rutten, J., M. Kwiatkowska, G. Norman, and D. Parker. 2004. *Mathematical techniques for analyzing concurrent and probabilistic systems*, ed. P. Panangaden and F. van Breugel, Volume 23 of *CRM Monograph Series*. American Mathematical Society.

Sauro, H. M., M. Hucka, A. Finney, C. Wellock, H. Bolouri, J. Doyle, and H. Kitano. 2003. Next generation simulation tools: the systems biology workbench and BioSPICE integration. *OMICS* 7(4):355–372.

SBGN Team. 2006. The Systems Biology Graphical Notation (SBGN). Available on the Internet at <http://www.sbgn.org>.

Sha, W., J. Moore, K. Chen, A. Lassaletta, C. Yi, J. Tyson, and J. Sible. 2003. Hysteresis drives cell-cycle transitions in xenopus laevis egg extracts. *Proceedings of the National Academy of Sciences* 100 (3): 975–980.

Shaffer, C., R. Randhawa, and J. Tyson. 2006. The role of composition and aggregation in modeling macromolecular regulatory networks. In *Proceedings of the 2006 Winter Simulation Conference*, ed. L. Perrone, F. Wieland, J. Liu, B. Lawson, D. Nicol, and R. Fujimoto. Monterey, California, USA.

Strömbäck, L. 2005. Possibilities and challenges using XML technology for storage and integration of molecular interactions. In *Proceedings of the Sixteenth International Workshop on Database and Expert Systems Applications*, 575–579. Denmark, Copenhagen.

Strömbäck, L. 2006. A method for alignment of standardised XML information within systems biology. In *Proceedings of the 2006 Winter Simulation Conference*, ed. L. Perrone, F. Wieland, J. Liu, B. Lawson, D. Nicol, and R. Fujimoto. Monterey, California, USA.

Strömbäck, L., and D. Hall. 2006. An evaluation of the use of XML for representation, querying, and analysis of molecular interactions. In *Proceedings of the 2006 Winter Simulation Conference*, ed. L. Perrone, F. Wieland, J. Liu, B. Lawson, D. Nicol, and R. Fujimoto.

Strömbäck, L., and P. Lambrix. 2005. Representations of molecular pathways: an evaluation of SBML, PSI MI, and BioPAX. *Bioinformatics* 21:4401–4407.

Swerdlin, N., I. Cohen, and D. Harel. 2006. Towards an *in-silico* lymph node: a realistic approach to modeling dynamic behavior of lymphocytes. Submitted.

Takahashi, K., K. Kaizu, B. Hu, and M. Tomita. 2004. A multi-algorithm, multi-timescale method for cell simulation. *Bioinformatics* 20 (4): 538–546.

Talcott, C. 2006. Pathway logic: a logical approach to modeling cellular processes. In *Proceedings of the 2006 Winter Simulation Conference*, ed. L. Perrone, F. Wieland, J. Liu, B. Lawson, D. Nicol, and R. Fujimoto. Monterey, California, USA.

Tyson, J., K. C. Chen, and B. Novak. 2003. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current Opinion in Cell Biology* 15:221–231.

Uhrmacher, A., D. Degenring, and B. Zeigler. 2005. Discrete event multi-level models for systems biology. *Transactions on Computational Systems Biology* 1:66–89.

Vass, M., C. Shaffer, N. Ramakrishnan, L. Watson, and J. Tyson. 2006. The JigCell model builder: a spreadsheet interface for creating biochemical reaction network models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 3:155–164.

Wolf, D. M., and A. P. Arkin. 2003. Motifs, modules and games in bacteria. *Current Opinion in Microbiology* 6:125–34.

Zwolak, J. 2006. Parameter estimation tool (PET). Available on the Internet at <http://jigcell.biol.vt.edu/parameter.html>.

AUTHOR BIOGRAPHIES

HERBERT SAURO was originally educated as a biochemist/microbiologist but became interested in simulation and theory to understand cellular networks after accidentally coming across a paper by David Garfinkel on the simulation of glycolysis. He wrote one of the first bio-chemical simulators for the PC (SCAMP) in the 1980s to assist work on extending metabolic control analysis (a theory closely related to biochemical systems theory). However, with the lack of community interest in systems biology during the late 80s and early 90s, he left science to start a software company and offer consultancy work to finance firms in the UK. With the surge in interest in systems biology in the US in the late 90s, he secured a position at Caltech to assist in the development of the Systems Biology Markup Language. Since then he moved to a faculty position at the Keck Graduate Institute where he continues to do research on network motifs, theory and software. His email address is <hsauro@kgi.edu>.

ADELINDE M. UHRMACHER is head of the Modeling and Simulation Group being part of the Institute of Computer Science at the University of Rostock, Germany. Her research interests are in modeling and simu-

lation methodologies, including multi-level modeling and simulation and biological applications. From 2000-2006 she has been editor-in-chief of the SIMULATION - Transactions of the SCS and is on the Editorial Board of Transactions on Computational Systems Biology. Email: <lin@informatik.uni-rostock.de>, webpage: <wwwmosi.informatik.uni-rostock.de>.

DAVID HAREL has been at the Weizmann Institute of Science since 1980. He was Department Head from 1989 to 1995, and was Dean of the Faculty of Mathematics and Computer Science between 1998 and 2004. He is also co-founder of I-Logix, Inc. He received his PhD from MIT in 1978, and has spent time at IBM Yorktown Heights, and at Carnegie-Mellon and Cornell Universities. In the past he worked mainly in theoretical computer science, and now he works in software and systems engineering, modeling biological systems, and the synthesis and communication of smell. He is the inventor of statecharts and co-inventor of live sequence charts, and co-designed Statemate, Rhapsody and the Play-Engine. He received the ACM Outstanding Educator Award (1992), the Israel Prize (2004), the ACM SIGSOFT Outstanding Research Award (2006), and two honorary doctorates. He is a Fellow of the ACM and of the IEEE.

MICHAEL HUCKA received his PhD from the University of Michigan studying computational neuroscience. After a postdoc in the same area at the California Institute of Technology, he switched to working in systems biology with the Kitano ERATO Systems Biology Project in the group of Professor John Doyle. As part of a team with Herbert Sauro, Andrew Finney and Hamid Bolouri, he helped develop the Systems Biology Workbench and Systems Biology Markup Language. Today he is a Senior Research Fellow at Caltech and co-director of the Biological Network Modeling Center at the Beckman Institute, continuing work with SBML and new efforts such as BioModels Database.

MARTA KWIATKOWSKA is Professor of Computer Science in the University of Birmingham, UK. Her research is mainly concerned with developing modelling frameworks and novel methods for analysing large complex systems, especially automatic verification techniques. She led development of the state-of-the-art probabilistic model checker PRISM and is on the Editorial Board of Transactions on Computational Systems Biology and Logical Methods in Computer Science. Email: <mzk@cs.bham.ac.uk>, webpage: <www.cs.bham.ac.uk/~mzk>.

PEDRO MENDES is an Associate Professor at the Virginia Bioinformatics Institute. His research is centered broadly around computer simulation and analysis of biochemical networks. This is comprised of three components: development of simulation software (Gepasi and now COPASI), modeling of gene expression in the context of metabolic networks, and bioinformatic support for metabolic profiling. He is the author of the widely used Gepasi software application.

CLIFFORD A. SHAFFER is an associate professor in the Department of Computer Science at Virginia Tech since 1987. He received his PhD from University of Maryland in 1986. His current research interests include problem solving environments, bioinformatics, component architectures, visualization, algorithm design and analysis, and data structures. His Web address is <www.cs.vt.edu/shaffer>.

LENA STRÖMBÄCK is an Assistant professor at Linköpings Universitet. She has a solid background in working with databases and XML. She holds a PhD degree in computer science within natural language processing (1997). After her PhD, she worked at Nokia with research and development of products for the information society. This work included responsibility for European projects and work with future standards in XML. Her current research focuses on standards and tools for management of standards, mainly within the area of bioinformatics. Her e-mail address is <lestr@ida.liu.se> and her Web address is <http://www.ida.liu.se/~lestr>.

JOHN J. TYSON is University Distinguished Professor of Biological Sciences at Virginia Tech. He received his PhD in chemical physics from the University of Chicago in 1973 and has been specializing in theoretical cell biology since that time. His current interests revolve around the geneprotein interaction networks that regulate features of cell physiology such as cell division, circadian rhythms, intracellular signaling networks, and programmed cell death.