Ten Top Reasons for Systems Biology to Get Into Model-Driven Engineering

Magali Roux-Rouquié
LIP6 CNRS – UPMC
8, rue du Capitaine Scott
75015 Paris, France
(+33) 144 277 094
magali.roux@lip6.fr

Debora Schuch da Rosa
LIP6 CNRS – UPMC and DIT – UNITN
via Sommarive 14
38050 Trento, Italy
(+39) 0461 883 942
debora.schuch.rosa@acm.org

ABSTRACT
Recent progresses in post-genomic biology allow addressing issues at the system level through the coupling of experimental and modeling approaches that constitute the core of Systems Biology (SB). The challenge for researchers is to integrate disparate data into models to infer useful biological knowledge. Problems which are questioning Model-Driven Engineering (MDE) deal with complexity, systems modeling (including modularity and structure/behavior relationships), models networking and integration, domain-specific language, etc. Among these issues, ten arguments are presented to consider the conceptual convergence and cross-fertilization of SB with MDE. Current models in SB are reviewed and the main requirements to conduct models integration are identified. Some perspectives are provided towards the translation of high-level SB descriptions into formal languages for performing static and dynamic analyses.

Categories and Subject Descriptors
I.6.5 [**Simulation and Modeling**]: Model Development – *Modeling methodologies*.

General Terms
Design, Standardization, Languages, Theory, Verification.

Keywords
Complex systems, systems biology, Bioambients, UML.

1. INTRODUCTION
The 20th century will be reminded as the century of gene: starting with the rediscovery of Mendel's laws on inheritance, it ended with the sequencing of the human genome. Among major advances, the genetic code was broken in 1961, enabling the decoding of genetics messages in a large variety of species. These efforts were crucial in setting the stage for the development of biotechnologies. Over the past few years a variety of high-throughput methodologies (0mics sciences) was developed that enable large-scale studies of the molecular landscape and contribute more and more data, making biology amenable to comparison with large-scale systems in computer science [1].

GaMMa'06, May 22, 2006, Shanghai, China.

The 21st century is claimed to be the century of information society [2] and the invention of novel computing paradigms for improved processing of human and biological data is identified as one of major challenges [3]. It was suggested that such issues would greatly benefit of convergence in sciences that cross-fertilize each other concepts; using the analogy between biological systems and large-scale computing systems could contribute to develop such innovative design paradigms. To approach these tasks, we provided some insights to open a debate on convergence of systems biology (SB) and model-driven engineering (MDE) through concepts and methods exchanges. In section 2, we examined related work; in section 3, we reviewed some of the problems addressed by systems biology that could bring new clues in computer sciences and conversely; in section 4, we examined the main requirements to design a minimal framework draws on the experience of metamodeling; in section 5, we presented some perspectives towards the translation of high-level descriptions in systems biology into formal languages**.**

2. RELATED WORK
In computer sciences, model-driven engineering is a new paradigm that allows the integration of different technologies spaces based on their modeling in conformance to a metamodel [4]. This produces a design that satisfies a set of requirements that can be used to check for design'correctness [5]. Formal methods that have mathematical precise foundations are used for model-based testing; they include, notably, statecharts [6] and process algebra [7].

Paralleling these considerations, systems biology offers different technological spaces named "Omics" (see subsections 3.1 and 4.1) and there is a need to integrate these different views into a more global representation; currently, this is achieved using clustering methods [1]. Otherwise, model-checking of biological observations is developing very fast by adapting different formal methods [8, 9]. In this position paper, we suggest to adopt a MDE approach to deal with biological heterogeneity at data (syntactic) level and/or functional (semantic) level. Coupling these approaches to model-based testing will make it possible to address several problems, from data integration to the analysis and the prediction of systems behaviour.

3. CONVERGING PROBLEMS ADDRESSED IN SYSTEMS BIOLOGY
In a MDE perspective, we present below our personal view of main issues in systems biology.

3.1 Toward a system view

The scientific approach of reducing complex systems to their component parts has produced remarkable results over the last fifty years in molecular biology and is likely to produce more. Among reductionnist approaches, Omics sciences focused on components through the selected study of genes (genomics), proteins (proteomics), sugars (glycomics) and other biologically crucial molecules. To develop a systems perspective, the knowledge on the components, theirs relations, etc. has to be aggregated into a system level understanding; this is the central goal of the new called systems biology discipline.

3.2 Complexity of design

Examples of the data explosion include data from systematic DNA sequencing of a growing variety of organisms, simultaneous measurement of the expression levels of tens of thousands of genes; 3D models of small molecular structures and large macromolecular structures (proteins, RNA, DNA); models of biological pathways, etc. The challenge consists in assembling these massive amounts of information into the right framework.

3.3 Designing complexity

Living systems are made of incredibly complex sets of interconnected systems. As reported previously [10], one approach to deal with complexity is to differentiate systems in as many subsystems or levels (the so-called nearly decomposable systems), each level being modeled by its network and being interpreted in a relatively autonomous way, provided the coupling relations between sublevels are carefully identified. This approach sounds particularly well-suited for the design of living systems that show high compositionality.

3.4 Context awareness

One of the best biological examples of context awareness is achieved by the immune system, which reacts specifically to one foreign molecule among billions of possible substances in new and unforeseen contexts. Recent developments in systems biology have made it possible to seek a global understanding of embedded cellular and molecular networks involved in immunity allowing new insights on mechanisms involved in adaptation and robustness [11].

3.5 Star-abilities (*-abilities)

Biological systems show reliability, adaptability, composability, reusability, etc. One challenging issue aims to understand how biological systems achieve high reliability based on unreliable components (for example, some components like proteins have a limited life time). Otherwise, their adaptability is fundamental to the ability to persist despite external perturbation and internal noise, it is due to their capacity to variability, notably, through mutations.

3.6 Modelling at the heart

In biology, the art of mathematical modelling relies on four items: (1) a deep understanding of the biological problem, (2) a realistic abstraction of the phenomena, (3) the finding of (preferably) quantitative solutions, (4) the interpretation of the results in predictive terms. Actually mathematical modeling has a long history in biology, but only recently has there been a strong increase of interest due to explosion of data, placing modelling at the heart of systems biology.

3.7 Computational evaluation

Two kinds of models are developed: (1) using bottom-up approaches, models are designed for the detailed analysis of well characterized processes (continuous or discrete models), (2) using top-down approaches, models integrate 'omics' information of hundreds or thousands of interacting components (mostly clustering methods). In both case, models and simulations are becoming more and more important for computational evaluation of biological processes, contributing to limit the explosive cost of experiments.

3.8 Models integration

Complex composite models could be achieved through the integration of submodels. Although central to systems biology, this issue has only be partially addressed in SBML [12] and CellML [13] which are repositories for biochemical models. Inspired from software engineering and notably from model-driven engineering, we and others [14] propose to adopt a metamodel approach to delineate a general framework that could be used for model design and models integration.

3.9 Domain specific modeling language

No modeling language exists at the moment in systems biology, i. e. allowing describing physiological processes at different spatial, temporal, hierarchical and functional scales. Actually, the design of such a framework would have to identify categories of systems building blocks (concepts and rules), each of which describes a particular facet of the system; based on these building blocks, complex systems could be further specified. Such efforts would allow delineating a modelling language specific to the domain to be used for designing analysable and transplantable domain models.

3.10 Biological systems as engineering systems

Using a sequential software engineering process, a complex system (SARS viral infection) was reverse-engineered and represented as an object-oriented software system. The resulting artifact was shown not only to capture information about individual components but also to display the system-level architecture and relations among system components [15]. We suggest adapting formal verification approaches for the checking of models in systems biology.

4. MAIN REQUIREMENTS FOR THE DESIGN OF A MINIMAL FRAMEWORK IN SYSTEMS BIOLOGY

Our first goal is to enable heterogeneous models integration based on the experience in software engineering, in particular metamodeling; this would guide the design of a domain specific modeling language. As a first step, we identified two main requirements as follows:

4.1 Integrating the current frameworks and standards

Frameworks have been designed for every Omics technologies and UML class diagrams were used to represent physiological

processes and experimental protocols. Some of them were approved as community standards including MIAME [16] (Minimum Information About a Microarray Experiment), PSI-MIF [17] (Proteomics Standard Initiative Molecular Interaction format) or MIRIAM [18] (Minimum Information Requested in the Annotation of biochemical Models). In any case, all data in schemas can be rearranged into Biological Information (the biological phenomena under investigation) on the one hand and Experimental Data (that characterize the phenomena) on the other hand (Fig. 1, dark grey box). Similarly, models in SBML or CellML repositories as well as the recent BioModels Database [19] show variables and parameters related to Biological Information (Fig. 1, light grey box), emphasizing the pivotal role of the Biological Information (Fig. 1, white box). Accordingly, the three major building blocks for a minimal framework in systems biology might concern Experimental Information, Biological Information and Models. Other packages in figure 1 refer to complementary data provided by specific frameworks.

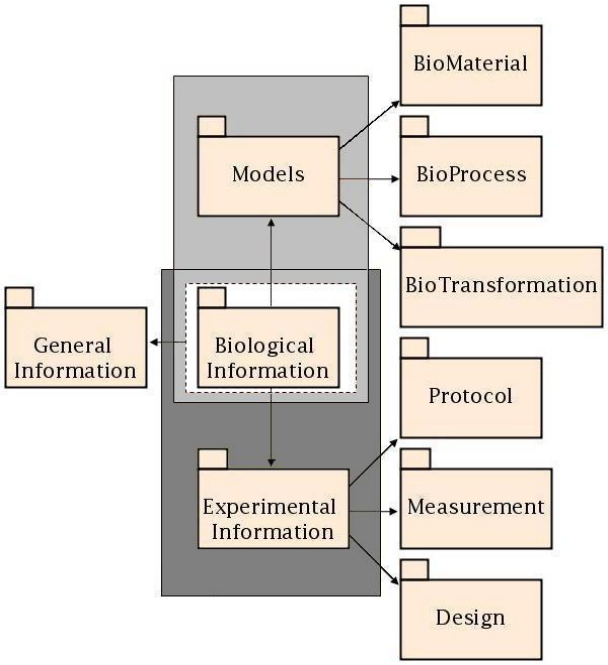

Figure 1. Main building blocks for a minimal framework for systems biology.

4.2 Capturing the system dynamics (linking structure to behavior)

Most of mathematical models in SBML, CellML and BioModels Database are ODE-based models. These models have some well known difficulties, including non-linearity (no analytical solutions) and non-modularity (when new components are added to the system, the complete mathematical model has to be changed). In the future, other mathematical formalisms could be of help to analyze the behavior of biological systems, and especially strategies based on model verification to represent or to predict the difference in performance between two scenarios or two model configurations. To approach these problems, careful descriptions of systems behaviour have to be achieved. Software engineering uses to deal with these issues putting the ideas in the UML world (or in domain specific modeling language) to perform model transformation, for example from UML statecharts to Pi-calculus [20]. In this perspective, we [21] and others [8] have shown that the dynamics of biological systems can be captured using statecharts.

Taking advantage of the UML language [21], work is in progress to delineate a profile for systems biology that will combine these two major requirements to achieve a process view of biological entities [22].

5. THE TRANSLATION OF HIGH-LEVEL DESCRIPTIONS IN SYSTEMS BIOLOGY REQUIRES IMPROVEMENTS IN FORMAL LANGUAGES

The next availability of a domain specific modeling language will make it convenient to specify biological systems in a semi-formal ways and to share easily knowledge between scientific communities. The extraction of process algebra from high-level descriptions is a step towards the use of formal methods for performing further analyses (although hiding formal details to the designers). The first attempts to model biological systems using process algebra were done in lambda-calculus [23] (RNA modeling) and then in stochastic pi-calculus [24]. After some new primitives were proposed to deal the specificity of biological processes, a novel calculus named BioAmbients was created [25]. Actually, our research is going to bring some innovative changes to the existing formal languages and a new view of the operational semantics is being proposed. To take into account the concurrency and mobility of biological processes, we suggest a refined methodology to code biological systems into BioAmbients. This new language offers novel nested processes, the ambients, which are very appropriate to represent system evolution, for example, shape transformation of system components over time. This formalism, which will incorporate data provided by our framework, will circumvent problems found in the use of ODE, as mentioned above. These improvements in process algebra will delineate a new language with expressiveness similar to Petri nets (with respect to data, time, etc.), in addition to its strong mathematical foundation.

The ambient codes generated from biological specifications will be the input source for tools that were developed recently for static [26] and dynamic [27] analyses; in particular, static analysis techniques developed in the context of optimizing compilers were applied to biological specifications in formal language and were found to be a promising direction.

6. DISCUSSION AND PERSPECTIVES

Interdisciplinarity is the way to share concepts and methods from different disciplines to synthesize new cognitive structures (novel concepts, approaches, frameworks), this is a deeper involvement than having just several disciplines concerned by adapting and modifying existing concepts and methods within their discipline and occasionally borrowing ideas from others [28]. With this respect, our ultimate perspective is to design innovative paradigms for systems biology through a MDE way of thinking. In this paper, we presented some of the arguments to account for our approach and the main steps to afford, including metamodeling and model translation. To achieve these goals,

many of the problems mentioned above would have to be solved. By doing this, we should learn from each other's communities, hoping that sharing our efforts in the context of systems biology will inspire new ideas in computer sciences.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge Corrado Priami for many useful discussions. Work was supported by grants from CNRS (STIC-SANTE) and UNITN.

8. REFERENCES

[1] Ge H., Walhout A. J., Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*. 19 ((2003)), 551-60.

[2] http://www.itu.int/wsis/

[3] http://www.nsf.gov/pubs/2005/nsf05626/nsf05626.htm, Emerging Models and Technologies for Computation.

[4] Kurtev, I., Bézivin, J., and Aksit, M. Technological spaces: an initial appraisal. CoopIS, DOA 2002 Federated Conferences, Industrial track, Irvine, 2002.

[5] Brilliant, S., Knight, J., and Leveson, N. Analysis of Faults in an N-Version Software Experiment. *IEEE Trans. on Software Engineering*, Vol. SE-16, No. 2, February 1990 (http://sunnyday.mit.edu/papers.html#reqs)

[6] Harel, D. Statecharts: A visual formalism for complex systems. *Sci. Comput. Programming*, 8 (1987) 231-274.

[7] Milner (R. Milner (1989): *Communication and Concurrency*, Prentice Hall)

[8] Fisher, J., Piterman, N. Hubbard, E., et al. Computational insights into Caenorrhabditis elegans vulval development. *Proc. Natl. Acad. Sci. USA* 102 (2005) 1951-1956.

[9] Priami, C. *Design Environments for Complex Systems.* (2003) Technical Report DIT-03-048, Informatica e Telecomunicazioni, University of Trento.

[10] Simon, H.A. The Sciences of the Artificial (3rd ed.). Cambridge, MA: *The MIT Press*, 1997.

[11] Aderem, A., Hood, L. Immunology in the post-genomic area. *Nature Immunology*. 2, (2001) 373-375.

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

[13] Lloyd, CM; Halstead, MDB; Nielsen, PF. CellML: its future, present and past. *Prog Biophys Mol Biol*. 85(2-3) (2004) 433-450

[14] Finkelstein, A., Hetherington, J., Li, L., Margoninski, O.et al. *Computational Challenges of Systems Biology*. IEEE Computer, 37(5) (2004) 26-33.

[15] Shegogue D, Zheng WJ. Object-oriented biological system integration: a SARS coronavirus example. *Bioinformatics*. (2005) 21:2502-9.

[16] Brazma A, Hingamp P, Quackenbush J, Sherlock G,. et al. Minimum Information About a Microarray Experiment. Towards a standard for microarray data. *Nature Genetics* 29, (2001), 365-371.

[17] Hermjakob H, Montecchi-Palazzi L, Bader G, et al. The HUPO PSI's molecular interaction format-A community standard for the representation of protein interaction data. *Nature Biotechnologie* 22, (2004), 177-183.

[18] Le Novère N, Finney A, Hucka M, Bhalla US, et al. Minimum Information Requested in the Annotation of biochemical Models (MIRIAM) *Nature Biotechnology* 23, (2005), 1509-1515.

[19] Le Novère N, Bornstein B, Broicher A, Courtot M, et al. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* 34, (2006), 689-691.

[20] Korenblat, K., Priami, C. *Extraction of Pi-calculus specifications from a UML sequence and state diagrams*. DEGAS IST-2001-32072, Technical Report No.(2003) DIT-03-07.

[21] Roux-Rouquié, M. , Caritey, N., Gaubert, L.et al. Using the Unified Modelling Language (UML) to guide the systemic description of biological processes and systems. *BioSystems* 75 (2004) 3–14.

[22] Roux-Rouquié M., Caritey, N., Gaubert, L., Le Grand, B., and Soto M. Metamodel and modeling Language: Towards an Unified Modeling Language (UML) Profile for Systems Biology (*SCI05*, Orlando, Florida, USA, July 10-13, 2005).

[23] Fontana, W., Buss, L., W. The Arrival of the Fittest: Toward a Theory of Biological Organization, *Bull. Math. Biol.*, 56 (1994), 1-64.

[24] Priami, C., Regev A., Shapiro E. and Silverman W.., Application of a stochastic passing-name calculus to representation and simulation of molecular processes, Information Processing Letters, 80, (2001), 25-31.

[25] Regev. A. Computational system biology: A calculus for biomolecular knowledge. PhD Thesis, Tel Aviv Univ., 2003.

[26] Nielson, F., Nielson, H. Riis, Priami, C. and Rosa, D. S. Control Flow Analysis for BioAmbients. *Elsevier's Electronic Notes in Theoretical Computer Science serie BioConcur*, 2003.

[27] The BioSpi project: http://www.wisdom.weizmann.ac.il/~biopsi/

[28] EURAB, Interdisciplinarity in research, April 2004.