Connecting the dots: role of standardization and technology sharing in biological simulation

Samik Ghosh[1], Yukiko Matsuoka[1,2] and Hiroaki Kitano[1,3,4]

[1] The Systems Biology Institute, Tokyo 108-0071 Japan
[2] JST ERATO Kawaoka Infection-induced Host-response Network Project, Tokyo 108-8639 Japan
[3] Okinawa Institute of Science and Technology, Okinawa, Japan
[4] Sony Computer Science Laboratories, Tokyo 141-0022 Japan

The role of biological modeling and simulation in enhancing productivity across the drug discovery pipeline has been increasingly appreciated over the past decade by the pharmaceutical industry. However, adoption of *in silico* modeling and simulation techniques has been sparse due to skepticism in the associated pay-offs and knowledge gap in research. While biological simulations have been successfully applied in specific projects, a standardized, community-wide platform is imperative for making the final leap of faith across the domain. This review outlines the issues and challenges involved in fostering a private-public collaborative effort for the development of standard modeling and biosimulation platforms and concludes with insights into possible mechanisms for integrating an *in silico* pipeline into the drug discovery and development process.

In the spring of 2008, a group of scientists and thought leaders from across academia and industry met in Tokyo [1] to brainstorm future challenges in systems biology and its application to the pharmaceutical industry. While unanimously acknowledging the burgeoning role of systems biology in tackling complex diseases, the researchers laid out a bold objective – 'to create over the next 30 years a comprehensive, molecule-based, multi-scale, computational model of the human ('the virtual human'), capable of simulating and predicting, with a reasonable degree of accuracy, the consequences of most of the perturbations that are relevant to healthcare' [2].

The scale and timeline of the project outlined in the Tokyo Declaration [2] underscore the complexity of the problem facing researchers in life sciences and pharmaceutical companies alike. With the current economic scenario, value of impending major blockbuster drug patent expirations and thinning pipelines for anticipated phase III drug approvals [3,4], the pharma industry is undergoing a paradigm shift in its efforts to develop safe and more efficacious drugs for the complex and life-threatening diseases facing humankind. Recent industry reports on future visions for the industry [3,4] have highlighted the need to shift trajectories from target-based drug discovery to more system-oriented, holistic approaches that embrace knowledge from basic academic research and computational and systems engineering techniques. In particular, predictive biosimulation software – based on interconnected sets of mathematical equations with calibrated parameters to represent biological and physiological behaviors – has been successfully applied across the different stages of drug discovery and development, from target identification and validation, lead optimization and candidate selection to clinical trial design and development [5–7].

Simulation is an indispensable tool in all engineering designs and has been successfully applied in the automobile, aerospace and telecommunication industries for decades. Computational fluid dynamics (CFD), for example, is an essential design process in aircraft design, ship design and automobile design. Any high-rise building has to carry out a series of structural integrity simulations even to be approved for construction; chipmakers model, modify and simulate their designs on computers before sending them to the fabrication plants; 'virtual cars' are driven and 'virtual aircrafts' flown under simulated conditions before hitting the manufacturing floor [8]. Although the application of advanced

Corresponding author:. Kitano, H. (kitano@sbi.jp)

1359-6446/06/$ - see front matter © 2010 Published by Elsevier Ltd. doi:10.1016/j.drudis.2010.10.001

modeling and simulation techniques has resulted in immense cost savings and standardized procedures for such R&D-intensive industries, the pharmaceutical industry has historically lacked these approaches, leading to astronomical costs in drug development (~25% of its revenue, almost twice that of other knowledge-driven industries [8]).

Although appreciation and awareness for the potential benefits of computational approaches in biological sciences and drug design have been on an increasing trajectory in both industry and academic circles, it is important to keep in perspective the unique hurdles and cosiderable challenges of applying *in silico* techniques in the life sciences. Identification of the specific features desired from computational tools in the pharmaceutical industry, together with an open, collaborative mindset between all players, would form the key stepping stones in the development of safer, efficacious and cost-efficient drugs for complex diseases such as cancer, metabolic and cardiac disorders.

Issues and challenges

The adoption of simulation techniques in the life sciences requires careful and detailed consideration of the unique challenges of multi-scale modeling– from cells, to tissues and organs, to whole human body and host–pathogen interactions. A series of issues have to be addressed before simulation can be accepted as normal practice in the industry.

First, a set of fundamental technical issues must be solved to further improve accuracy of simulation. Different flavors of simulation technologies exist, from deterministic, differential-equation-based systems to non-deterministic, stochastic techniques, agent-based and discrete-event based simulations; each presents a unique set of assumptions and system conditions that need to be considered before successful application to specific biological problems, as elucidated schematically in Fig. 1 [9]. Cellular modeling or physiological modeling with molecular details will require the integration of heterogeneous computational models that are on different spatial and temporal scales, and the basic equations still need to be defined [10].

The purpose and goal of a simulation system applied in drug design should be clearly defined: for example, in Formula 1 aerodynamics design, the goal is to design an aerodynamically optimal body with maximum down force with minimum drag. This forms a key step in defining and determining the eventual success of a biological simulation system. Merrimack Pharmaceuticals [11], for example, used computer simulation to identify a novel drug target for specific cancer sub-types that resulted in the development of a monoclonal antibody for ErbB3, now in clinical trial. Simulation models need to be designed to sufficiently capture essential features to accomplish the task defined, but features that are unlikely to affect prediction accuracy of the given task can be ignored.

Sophisticated models with molecular details that can predict cellular behaviors in various conditions are crucial for elucidating system-level properties of cellular systems. Such models should be able to provide predictions on how cells and organs respond when certain perturbations, such as drug administration, are given. Although there are some successful cases of computational modeling of limited-scale biological networks, there is no established method for developing high-precision models.

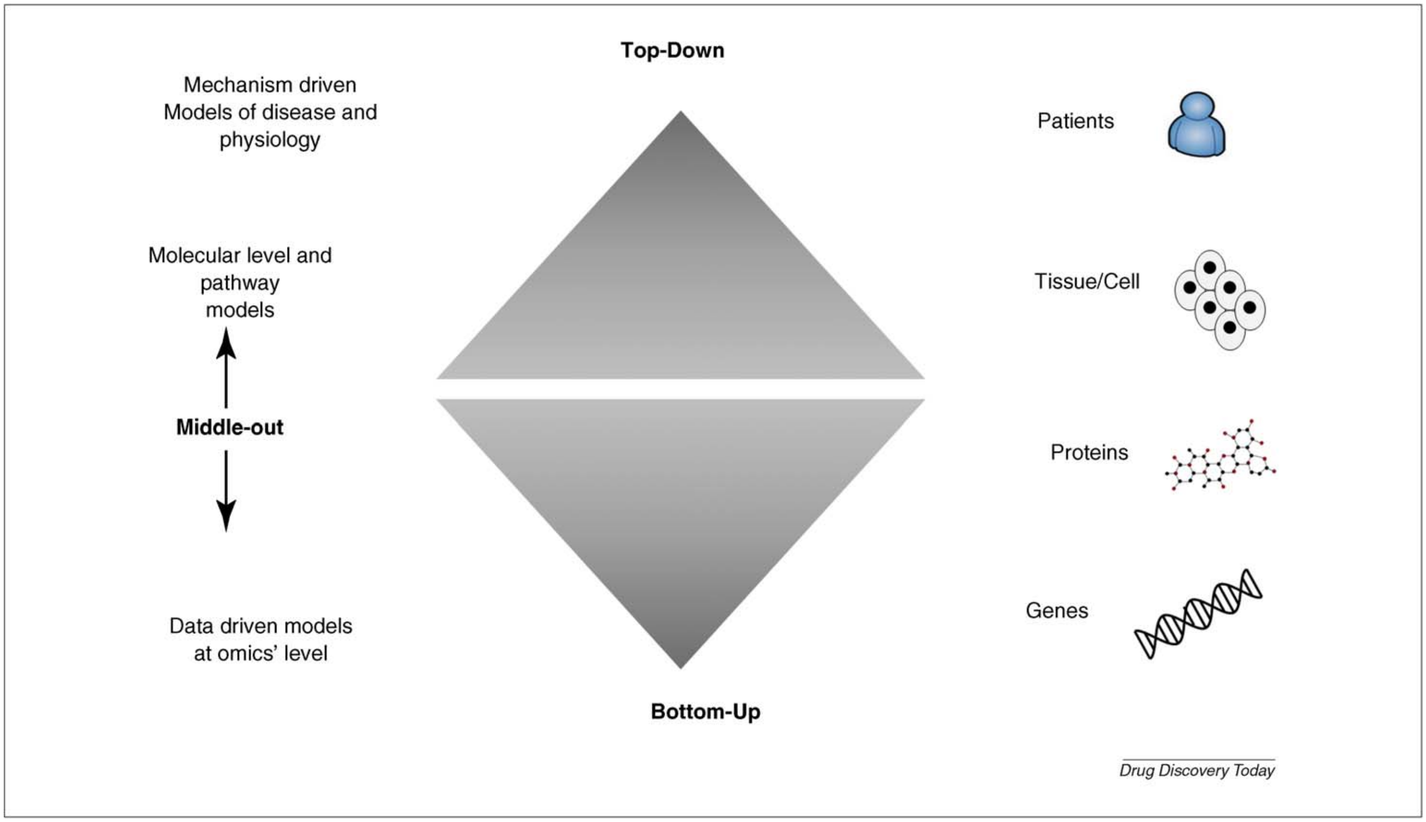

Top-Down
Mechanism driven Models of disease and physiology
Molecular level and pathway models
Middle-out
Data driven models at omics' level
Bottom-Up
Patients
Tissue/Cell
Proteins
Genes
Drug Discovery Today

FIGURE 1

Schematic showing different simulation and modeling approaches.

In establishing such practices, it might be prudent to borrow ideas from other fields with more mature and well-established practices of simulation modeling. In particular, the integration of computational modeling and experimental data acquisition has to be promoted. Delving into the design process of Formula 1 racing cars provides a picture of an iterative design cycle – several designs are tested with CFD and some of these designs are then tested using a wind tunnel, leading to the selection of one or two designs that are actually implemented and tested in a test course before one design is selected for final production. In this process, CFD models are calibrated against wind tunnel data for further improvement of accuracy, instead of data from the test course or from actual racing telemetry data.

This comparison delivers two messages. First, we need to develop highly controllable experimental systems comparable to wind tunnels in aerodynamics. This means that we need to be able to precisely control exposure to chemical substances and other environmental conditions. Second, efforts need to be made to create high-precision models against well-controlled experimental systems, instead of uncontrollable systems. The identification and integration of structural, spatial and temporal dynamics of both interaction networks and cellular structures is an essential prerequisite for defining such high-precision models. The dynamics of cellular structure and interaction networks need to be quantified by taking comprehensive, high-resolution, measurements of intracellular status – such as the concentrations, interactions, modifications and localizations of molecules – and of cellular structures in different dimensions and environmental conditions. In addition, the problem of how to identify unknown interactions from such data sets still remains.

These problems are fundamental and require collaborative efforts on a community-wide scale. Although various ongoing efforts exist for tackling the problems of building molecular network maps, simulation tools, data resources and web services for sharing information, an open, integrative platform for sharing and exchange of computational approaches is of fundamental importance.

Standards and technologies

The sharing of knowledge accelerates progress in science. This is perhaps more true for the biological sciences, in which the complexity and size of the problem domain make it imperative for different research groups to focus on different sections of it. Computational tools have already played an important part in the collection, storage and intelligent retrieval of vast amounts of information in the life sciences. With the adoption of biosimulation tools, the ability to store and share information in a seamless, unambiguous fashion became imperative, leading to the definition of The Systems Biology Markup Language (SBML; http://www.sbml.org), a set of standards developed to facilitate effective and efficient sharing of models defined as a set of biochemical reactions. The effort was initiated by the ERATO Kitano Symbiotic Systems Project, funded by the Japanese government but soon grew to be a global community-wide initiative. Importantly, the SBML community has evolved a proper procedure to elect editors, implemented voting procedures and formalized discussion forums. Thus, it is no longer a project belonging to any one institution but is truly a community effort. Currently, more than 180 pieces of bio-modeling and simulation software comply with SBML [12], enabling the sharing of models across them.

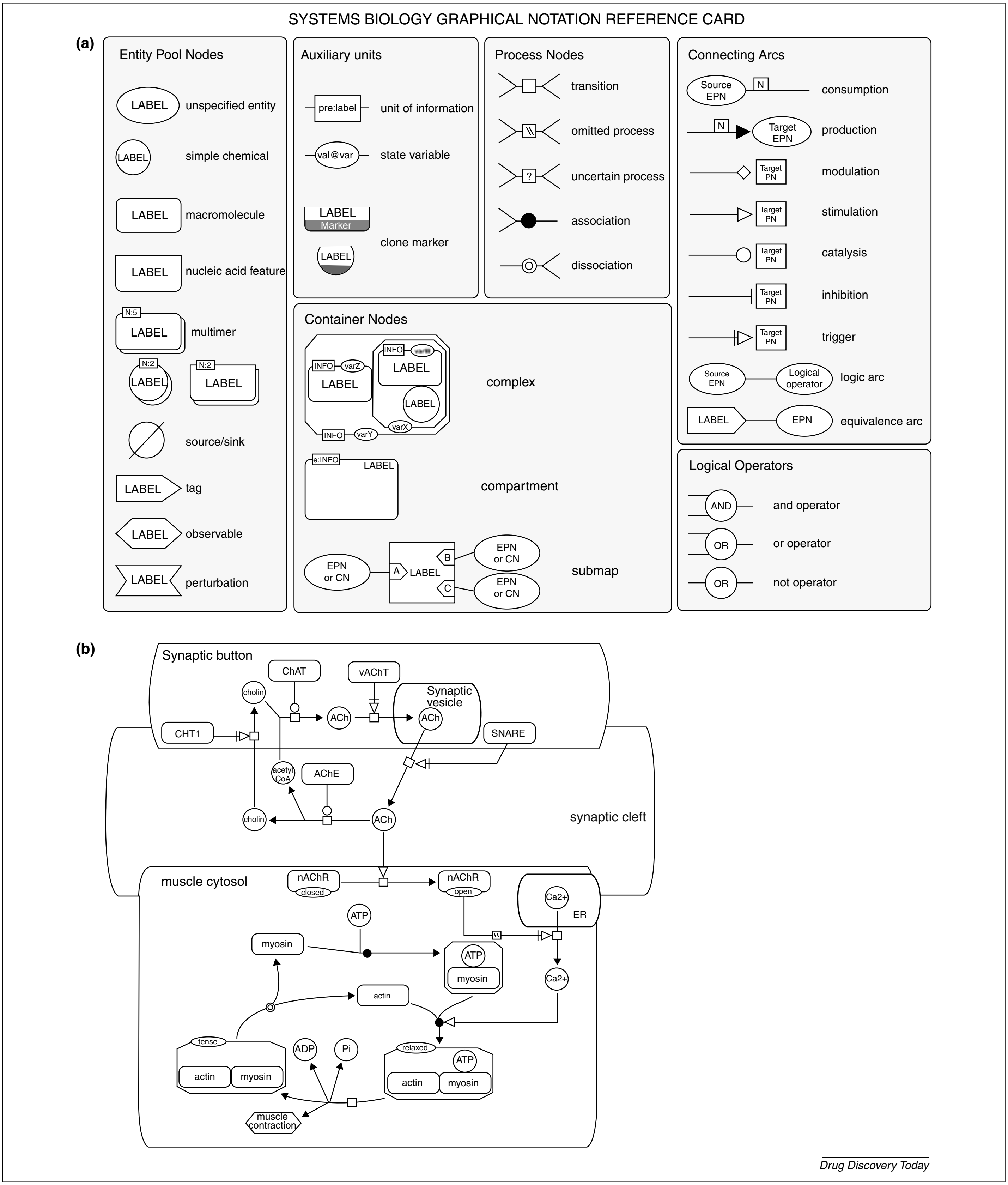
Although the definition of a common *lingua franca* is an important step in the standardization of *in silico* technologies, biology has traditionally been a descriptive science, in which the role of pictures and diagrams cannot be overstated. Keeping in mind the need for a common graphical representation standard in the life sciences, a community-wide effort has been undertaken to define The Systems Biology Graphical Notation (SBGN; http://www.sbgn.org/) [13]. The SBGN community is working to formulate a set of rules for human-readable visual representations of biological networks. The goal of SBGN is to define a set of visual glyphs and syntax so that anyone can understand exactly what each diagram means (Fig. 2), in the same vein as the electrical circuit diagrams used by chip designers.

With SBML and SBGN, standards for computational modeling and representation languages are being defined that will have a notable impact on standardization and model sharing. For models to be informative, they must be properly annotated; sufficient information must be attached with the models to enable third parties to use them. With the objective of proper model annotation, the MIRIAM project (http://www.ebi.ac.uk/miriam/) defines the minimum information that has to be attached to a model so that model can be informative by itself. A series of standard formation efforts are now underway to cover the whole process of modeling development and analysis.

Although standardization is an integral part of the process of computational systems biology, the development of a suite of software tools for model building, distribution and running simulations is another important dimension. In this direction, a plethora of modeling building and simulation tools such as Cellarator [14], Copasi [15] and Dizzy [16] in the academic community and SimBiology® from Mathworks Inc. (http://www.mathworks.com/products/simbiology) and PhysioLab® (Entelos Inc; http://www.entelos.com) exist, each catering to different modeling techniques (see Ref. [12] for comprehensive coverage of systems biology tools).

One of the most popular and widely used tools in this space [12] is CellDesigner™ [17] – a modeling and simulation tool to visualize, model, and simulate gene regulatory and biochemical networks. Two major characteristics embedded in CellDesigner boost its usability to create, import and export models: the solidly defined and comprehensive graphical representation (SBGN) of network models and SBML as a model-describing basis, which function as inter-tool media to import and export SBML-based models. Moreover, CellDesigner provides the ability to embed – or smoothly connect via Systems Biology Workbench (http://www.sys-bio.org/research/sbwIntro.htm) – different simulation and analysis packages, which enable the simulation of the pathways using various simulation techniques (Copasi, SBML ODE solver, and so on), as shown in Fig. 3a.

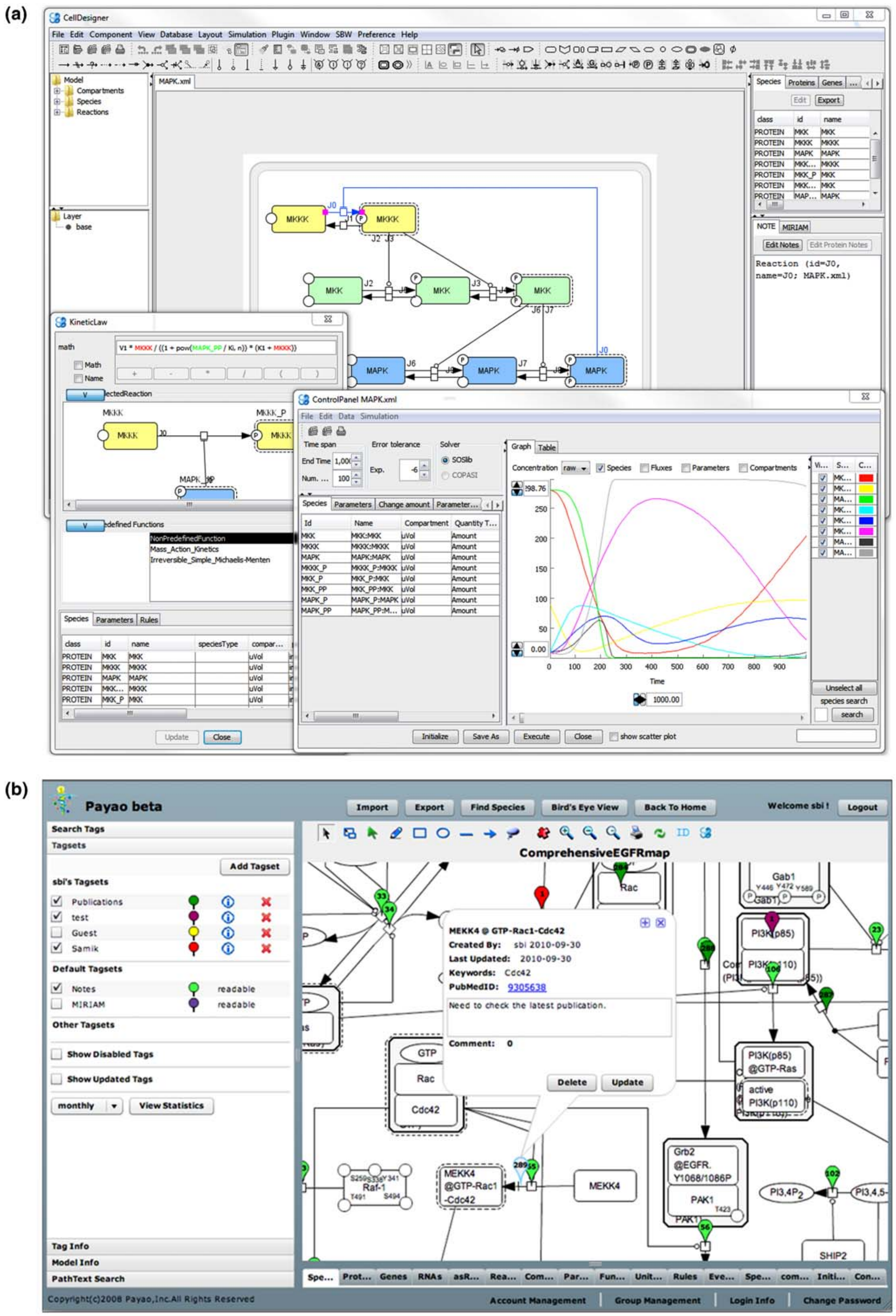
With the explosive growth in proliferation and adoption of the Internet and web services, knowledge in the life sciences has shifted to the World Wide Web with online access to scientific literature, biological databases, and knowledge-sharing and/or discussion forums. As more biological information comes online, it is important to develop an infrastructure that enables the community to leverage the vast web resources. This online, com-



Drug Discovery Today

FIGURE 2

(a) The Systems Biology Graphical Notation (SBGN) glyphs and **(b)** sample model representation.



class	id	name
PROTEIN	MKK	MKK
PROTEIN	MKKK	MKKK
PROTEIN	MAPK	MAPK
PROTEIN	MKK...	MKKK
PROTEIN	MKK_P	MKK
PROTEIN	MKK...	MKK
PROTEIN	MAP...	MAPK

Reaction (id=J0, name=J0; MAPK.xml)

math: V1 * MKKK / ((1 + pow(MAPK_PP / Ki, n)) * (K1 + MKKK))

NonPredefinedFunction
Mass_Action_Kinetics
Irreversible_Simple_Michaelis-Menten

class	id	name	speciesType	compar...
PROTEIN	MKK	MKK		uVol
PROTEIN	MKKK	MKKK		uVol
PROTEIN	MAPK	MAPK		uVol
PROTEIN	MKK...	MKKK		uVol
PROTEIN	MKK_P	MKK		uVol

End Time: 1,00(
Num. ...: 100
Exp.: -6
● SOSlib
○ COPASI

Id	Name	Compartment	Quantity T...
MKK	MKK:MKK	uVol	Amount
MKKK	MKKK:MKKK	uVol	Amount
MAPK	MAPK:MAPK	uVol	Amount
MKKK_P	MKKK_P:MKKK	uVol	Amount
MKK_P	MKK_P:MKK	uVol	Amount
MKK_PP	MKK_PP:MKK	uVol	Amount
MAPK_P	MAPK_P:MAPK	uVol	Amount
MAPK_PP	MAPK_PP:M...	uVol	Amount

☑ Publications			
☑ test			
□ Guest			
☑ Samik			
☑ Notes		readable	
□ MIRIAM		readable	

MEKK4 @ GTP-Rac1-Cdc42
Created By: sbi 2010-09-30
Last Updated: 2010-09-30
Keywords: Cdc42
PubMedID: 9305638
Need to check the latest publication.
Comment: 0
Delete Update

Drug Discovery Today

FIGURE 3

(a) CellDesigner snapshot and **(b)** Payao web interface snapshot.

munity collaboration paradigm motivated the development of Payao [18], a web-based biological pathway sharing and tagging service (http://www.payaologue.org). A snapshot of the interface is shown in Fig. 3b. The goal of Payao is to provide a Google Maps equivalent for biological pathways, wherein researchers can share large-scale, curated and annotated (MIRIAM-compliant) network maps (SBGN and SBML compliant) using software such as CellDesigner and publish it to the community online. With the built-in tagging and collaborative system, the community can participate in enhancing the biological entities in the map or navigate their specific areas of interest.

Other important assets for online community collaboration are the databases that provide pathway models fully curated and compliant with standards. Two major online data resources providing access to biological pathway models are the BioModels.net Initiative (http://www.biomodels.net), which provides a one-stop shop for SBML-compliant simulation models of biological pathways published in various literature, and The Panther Pathways database [19], which provides a database of annotated pathways created in CellDesigner (SBML and SBGN compliant).

We have provided a slice of the spectrum of activities initiated by different research groups in developing biosimulation tools and technologies. Whereas each of these tools provides a niche solution to a particular biological problem in isolation, an integrative framework leveraging the advantages of the diverse techniques holds the promise of providing a comprehensive computational pipeline.

Community collaboration and open flow model

The standards and technologies in systems biology, as elucidated in the previous section, represent various pieces of a comprehensive computational pipeline for the pharmaceutical industry. Crucial to the success of such a pipeline, however, is the clarification of the promises and pitfalls of the different components and the identification of a cohesive set of strategies to add value to the drug discovery process.

In June 2008, a representative group of systems biology scientists from academia, biotechnology and the pharmaceutical sector gathered in Portofino, Italy [20], with the goal of brainstorming a set of recommendations for such a strategic path to systems biology. While identifying various action items like setting data standards, modeling drug actions and toxicity, the leaders proclaimed that a 'network solution' [20] (i.e. community-wide collaboration, communication and outreach) was the key to solving complex biological network problems.

As outlined in the 'Issues and challenges' section, the size and complexity of living systems present several challenges to the systems biology community. Overcoming the obstacles of simulation speed, accuracy, high-precision experiments and knowledge sharing would require the expertise of scientists and engineers from diverse backgrounds and disciplines. Thus, developing an ecosystem of communication and information sharing empowered by powerful computational simulation tools and web services can provide the right plan for advancing biosimulation approaches in the pharmaceutical domain.

Concomitant with the development of a community ecosystem is the need for an open flow model of research, in which researchers from industry and academia share knowledge in a collective commons of data, analytical tools, information technology, biospecimens and disease models [21]. The past decade has seen several efforts in open-source biomedical science – the Cancer Genome Atlas (http://www.cancergenome.nih.gov/), Pathway Commons (http://www.pathwaycommons.org) and World Community Grid (http://www.worldcommunitygrid.org). Efforts for establishing collaboration and outreach through joint projects have also been launched in North America and Europe, including the Alliance for Cellular Signaling by the NIH (http://www.afcs.org) and the HepatoSys project (http://www.hepatosys.de) in Germany. The Open Source Drug Discovery project (http://www.osdd.net), initiated by the Council of Scientific and Industrial Research, India, is a uniquely novel effort dedicated to developing drugs for neglected tropical diseases (*Mycobacterium tuberculosis*) that plague developing countries and have very thin pipelines (in terms of the number of candidate drug compounds in development) in the pharmaceutical industries.

An open innovation strategy has also gained traction recently in the pharmaceutical industry, fueled by a need to reconfigure and streamline the drug discovery pipeline, particularly in the early phases of target biology and biomarker development [22]. An industry–academic consortium involving researchers from The University of California at San Diego, the California Institute of Technology, the Massachusetts Institute of Technology, the University of Massachusetts, Entelos, and several research groups from Pfizer has initiated a collaborative project on insulin-resistive pathways [23]. More recently, in May 2010, London-based GlaxoSmithKline and Novartis deposited over 300,000 structures of chemical compounds active against the malaria parasite *Plasmodium falciparum* in an open database archive, ChEMBL Neglected Tropical Disease from the European Bioinformatics Institute (http://www.ebi.ac.uk/chemblntd) [22].

Several socioeconomic obstacles to leveraging the power of an open-source approach exist, however – participant willingness to share data, incompatibility of formats, quality control of data, intellectual property conflicts and sustainable funding avenues. To develop a generic model for collaboration, it is necessary to develop a framework that is open, providing standardized, license-free access to biological data; integrative, providing common sets of pathway curation, annotation, modeling, analysis and simulation tools; and 'share-and-care', an incentive system for providing pre-competitive, early access to the results and data to participants.

An open, integrative, share-and-care model

The goal of an open, integrative, share-and-care framework is to provide a common set of tools, principles and practices for the application of biosimulation techniques in a systematic and cohesive manner. It would involve a standardization platform that enables the incorporation of standards and interfaces for the systems biology community; a computational bio-networking platform that provides a suite of network building tools, databases and web services for sharing information in a standards-compliant, interoperable manner; and an advanced simulation platform that incorporates the different simulation tools and technologies and is capable of encompassing information from the standard compliant biological network resources and databases. A schematic representation of such a scheme is shown in Fig. 4.

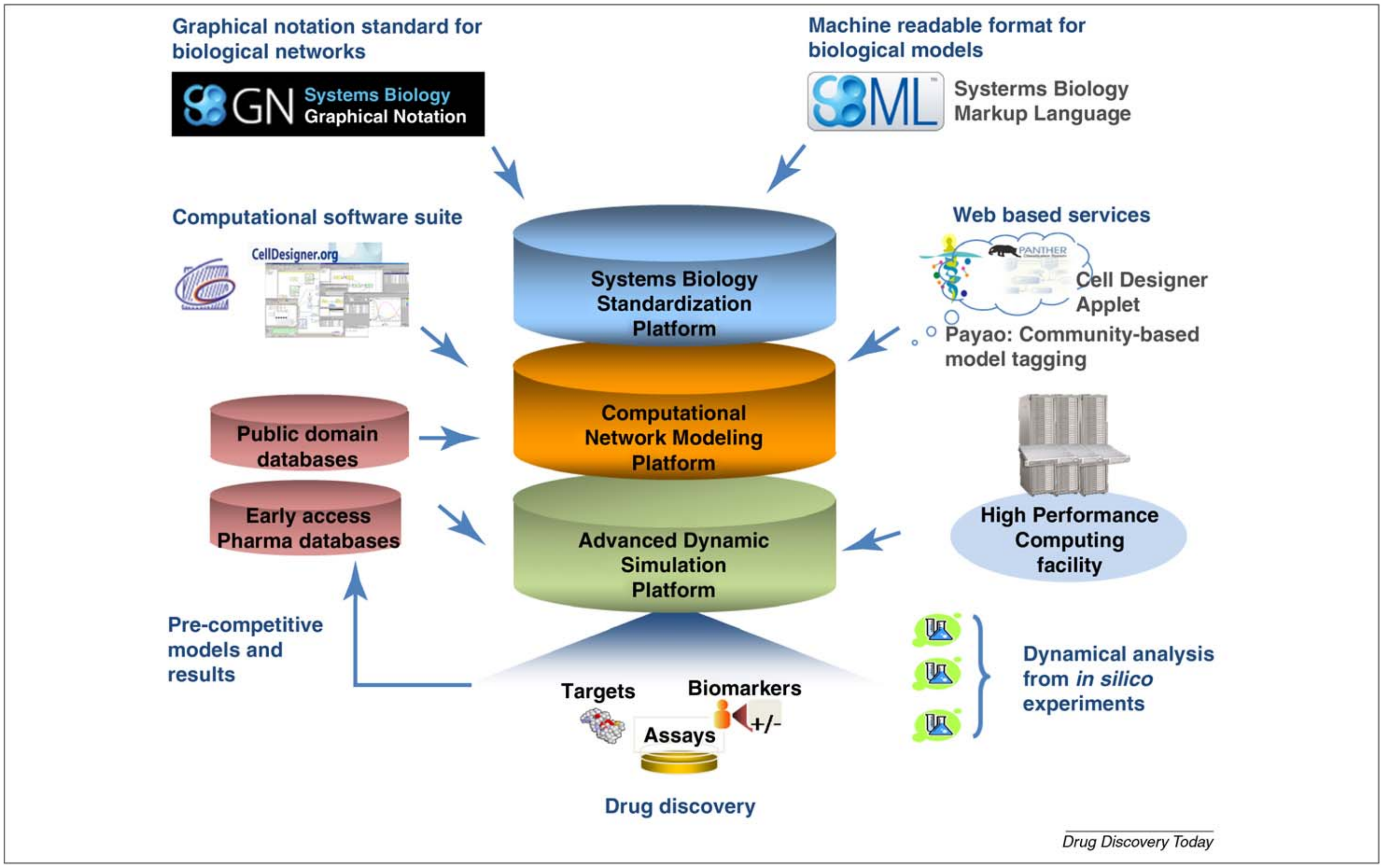

Graphical notation standard for biological networks
SGN Systems Biology Graphical Notation
Machine readable format for biological models
SBML Systems Biology Markup Language
Computational software suite
CellDesigner.org
Web based services
PANTHER
Cell Designer Applet
Payao: Community-based model tagging
Systems Biology Standardization Platform
Computational Network Modeling Platform
Advanced Dynamic Simulation Platform
Public domain databases
Early access Pharma databases
High Performance Computing facility
Pre-competitive models and results
Targets
Biomarkers
+/-
Assays
Drug discovery
Dynamical analysis from *in silico* experiments
Drug Discovery Today

FIGURE 4

An open, integrative, share-and-care model for systems biology.

Although the technical components of such a framework are all or mostly in place, the success of the model hinges on being able to foster a bi-directional flow of knowledge between academia and industry. An open access paradigm for information and data sharing would enable the fast and effective dissemination of knowledge and the identification and filling in of information gaps in crucial biological processes and pathways.

A standard-driven, integrative framework would enable a plug-and-play model for diverse simulation and modeling tools, providing researchers with the freedom to test various tools and techniques.

The most important part, perhaps, is ensuring the flow of information between diverse research groups. In this respect, a share-and-care model would provide incentives to all participants: academic researchers would gain access to specific data on drugs and clinical trials (e.g. access to failed drugs from phase II trial databases [24]), enabling them to develop models with higher predictive powers, whereas the business imperatives of pharmaceutical companies would be respected through early, pre-competitive access to model predictions and insights.

Such a system would work in a positive feedback loop, because high-precision data from the drug industry would drive the precision of biosimulation models, which, in turn, would be able to provide better decision-making conditions for the drug companies. In a recent development, a Boston, MA based startup called EnlightBio (http://www.enlightbio.com), working on similar paradigms, started operations focused on developing breakthrough innovations through partnership with multiple drug companies. Another effort initiated in early 2010 is Sage Bionetworks (http://www.sagebase.org), a non-profit, open-source research organization that aims to develop and share large-scale network models of diseases. Such initiatives would foster collaborative research while paving the way for systems biology to empower the life sciences industry.

Future perspectives

The value of applying systems biology techniques to the different stages of the drug design and development process has been demonstrated through various academic–industry collaborations spanning projects of different scales and sizes. Although such isolated success stories further motivate the large-scale adoption of *in silico* biosimulation practices, the development of a structured and coherent route to facilitate and accelerate the process is imperative. As outlined in this review, collaboration is the key to achieving the goal – not only in terms of joint projects but also – in developing a suite of standards-compliant platforms for sharing domain-specific knowledge and expertise. It might not be a distant dream that the community will develop a set of recommendations for good simulation practices (GSP), as already exists for good manufacturing practices (GMP) and good engineering practices (GEP) – standards and protocols accepted and complied with in the application of computational simulation tools for drug design and development.

Acknowledgements

This research is, in part, supported by the ERATO-SORST Program of the Japan Science and Technology Agency (JST); the Genome Network Project of the Ministry of Education, Culture, Sports, Science, and Technology; the NEDO Fund for International Standard Formation from the New Energy Development Organization; and the Okinawa Institute of Science and Technology.

References

1 Editorial, (2008) All systems go. *Nat. Rev. Drug Discov.* 7, 278–279

2 FCSB, (2008) *The Tokyo Declaration, International Workshop on Future Challenges for Systems Biology*. http://www.systems-biology.org/~myukiko/FCSB2008/doku.php?id=workshop:statement

3 PriceWaterhouseCoopers, (2008) *Pharma 2020: Virtual R7D – which path will you take?* http://www.pwc.com/extweb/pwcpublications.nsf/docid/91BF330647FFA402852572F2005ECC22

4 Futurology, P. (2007) *Joined-up healthcare, 2016 and beyond.* British Telecommunications http://www2.bt.com/static/i/media/pdf/BT_Pharma_Lowres.pdf

5 Kansal, A.R. and Trimmer, J. (2005) Application of predictive biosimulation within pharmaceutical clinical development: examples of significance for translational medicine and clinical trial design. *Syst. Biol. (Stevenage)* 152, 214–220

6 Rullmann, J.A.C. *et al.* (2005) Systems biology for battling rheumatoid arthritis: application of the Entelos PhysioLab platform. *Syst. Biol. (Stevenage)* 152, 256–262

7 Sams-Dodd, F. (2006) Drug discovery: selecting the optimal approach. *Drug. Discov. Today* 11, 465–472

8 No authors listed. (2005) Models that take drugs. *The Economist* June, S23–S24.

9 Michelson, S. and Scherrer, D. (2003) Predictive biosimulation for lead optimization. *Curr. Drug Discov.* 1, 18–22

10 Kitano, H. (2010) Grand challenges in systems physiology. *Frontiers in Physiology*. doi:10.3389/fphys.2010.00003.

11 No author listed. (2008) Merrimack Pharmaceuticals initiates enrollment in a phase 1 study of MM-121, an ErbB3 antagonist. *Medical News Today* August.

12 Klipp, E. *et al.* (2007) Systems biology standards – the community speaks. *Nat. Biotechnol.* 25, 390–391

13 Le Novère, N. *et al.* (2009) The systems biology graphical notation. *Nat. Biotechnol.* 27, 735–741

14 Shapiro, B.E. *et al.* (2003) Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations. *Bioinformatics* 19, 677–678

15 Hoops, S. *et al.* (2006) COPASI – a complex pathway simulator. *Bioinformatics* 22, 3067–3074

16 Ramsey, S. (2005) Dizzy: stochastic simulation of large-scale genetic regulatory networks. *J. Bioinform. Comput. Biol.* 3, 415–436

17 Funahashi, A. *et al.* (2008) CellDesigner 3.5: a versatile modeling tool for biochemical networks. In *Proceedings of the IEEE 96* pp. 1254–1265

18 Matsuoka, Y. *et al.* (2010) Payao: a community platform for SBML pathway model curation. *Bioinformatics* 26, 1381–1383

19 Thomas, P.D. *et al.* (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141

20 Henney, A. and Superti-Furga, G. (2008) A network solution. *Nature* 455, 730–731

21 Patlak, M. (2010) Open-source science makes headway. *J. Natl. Cancer Inst.* 102, 1221–1223

22 Strauss, S. (2010) Pharma embraces open source models. *Nat. Biotechnol.* 28, 631–634

23 No authors listed. (2008) The Insulin Resistance Pathways Project. http://www.pfizercambridge.com/home.php?id=groups/tgu/focus_insulin

24 Petsko, G.A. (2010) When failure should be the option. *BMC Biol.* 8, 61