ELSEVIER

Biochemical modeling with Systems Biology Graphical Notation

Andreas Jansson[1,2] and Mats Jirstrand[1]

[1] Fraunhofer-Chalmers Centre, Chalmers Science Park, 412 88 Göteborg, Sweden
[2] Infofusion, Systems Biology Research Centre, University of Skövde, Box 408, Skövde, Sweden

The Systems Biology Graphical Notation (SBGN) is an emerging standard for graphical notation developed by an international systems biology community. Standardized graphical notation is crucial for efficient and accurate communication of biological knowledge between researchers with various backgrounds in the expanding field of systems biology. Here, we highlight SBGN from a practical point of view and describe how the user can build and simulate SBGN models from a simple drag-and-drop graphical user interface in PathwayLab.

Introduction

Biochemical reaction networks or pathways have been used for decades in biology and medicine to graphically describe cellular processes. The aim of these maps ranges from sketches of qualitative findings to very accurate descriptions of molecular mechanisms; hence, the amount of effort necessary to transform these graphical formalisms to mathematical models that can be validated by experimental data varies a lot. Recent advances in experimental techniques and the increasing amount of data have triggered a shift of focus in many biological fields from an entity-oriented view to a systems approach (i.e. from the study of parts in isolation to the study of how they interact and their functional roles). Mathematical models are indispensable tools in this work, and there is a need for software tools that support mathematical modeling, computational analysis, and documentation. The scientific area known as 'systems biology' has drawn increasing attention from the systems and control community, which is natural because the subject applies systems and control methodologies and techniques to biological problems. Systems biology makes complex systems understandable by combining experimental data with a theoretical approach that enables computer simulations [1–3]. Such computer simulations can be done in many different types of software packages, and the Systems Biology Markup Language (SBML) [4] has been a crucial step in systems biology to enable exchanging models between different software tools. SBML is an open standard for encoding biological models in XML-based format, which facilitates the establishment of public model libraries such as BioModels Database and JWS Online [5,6]. A published model from such a library can easily be simulated by others using a software tool that supports SBML, even though the original model was implemented in different software. CellML is another XML-based markup language [7] that aims to address similar goals as SBML. The Biological Pathway Exchange standard, BioPAX, focus on the representation of biochemical networks and information about their constituents. However, it does not support kinetic modeling (i.e. mathematical reaction rate laws etc) so BioPAX models cannot be simulated. SBML and CellML on the other hand have their main focus on kinetic modeling and have a less detailed way of representing information about the networks and their elements.

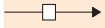
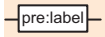
One of the greatest challenges in systems biology is to encourage scientists to use similar standards for graphical representation. This is a vital step for systems biologists that come from different backgrounds. If, for example, a biologist and a mathematician use the same graphical notation for building models, this would increase their ability to communicate. A first attempt by Kurt Kohn to create a unified visual notation for biological pathways was his Molecular Interaction Map, which is a notation for defining symbols and syntax to describe molecular interactions [8]. Kohn's work has been followed by alternative notations and extensions [9–12]. From an initiative by Hiroaki Kitano, an international meeting was held in Tokyo in 2006 to investigate the feasibility of a unified graphical notation for systems biology. A result of the meeting was an initial proposal constituting the basis

Corresponding author:. Jirstrand, M. (mats.jirstrand@fcc.chalmers.se)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2010.02.012

BOX 1

A selection of Systems Biology Graphical Notations

Symbol	Name	Comment
Entity pool nodes		
	Unspecified entity	An unspecified entity is used when the entity type is unknown or simply not relevant to the purposes of the model.
	Simple chemical	A simple chemical is a chemical compound that is not formed by the covalent linking of pseudo-identical residues (e.g. ATP).
	Macro-molecule	A macromolecule is a biochemical substance built up from the covalent linking of pseudo-identical units (e.g. proteins).
	Nucleic acid feature	A nucleic acid feature is a representation of a molecule carrying genetic information (e.g. RNA or DNA).
	Multimer	A multimer is an association of multiple identical or pseudo-identical entities (macromolecule, in this example) held together by non-covalent bonds. A unit of information can be attached to the multimer, indicating the number of aggregated entities.
	Source/sink	Source/sink is an entity for something that one does not need or wish to make precise. For example, this symbol is useful for representing the production or decay of entities without bothering to represent the details of the process.
Transitions		
	Transition	A transition is the flow from one compartment or entity to another.
	Association	The association between one or more entities represents the non-covalent binding of the biological objects into larger complexes. Note that if the association involves more than two entities, another consumption arc can be connected to this process.
	Dissociation	The dissociation of an entity into one or more entities represents the breaking of non-covalent bindings between the entities.
Controls		
	Modulation	A modulation is a general representation in which the modulator can affect the process both positively and negatively, or if one does not know the exact direction of the effect.
	Stimulation	Stimulation affects the flux positively of a process represented by the target transition.
	Catalysis	A catalysis is a specific case of stimulation, in which the effector affects the flux positively by lowering the activation energy of the reaction.
	Inhibition	An inhibition affects the flux negatively of a process represented by the target transition (e.g. competitive or uncompetitive inhibition).
	Trigger	A trigger is an absolute stimulation in which a process modulated by a trigger can only occur when this trigger is active.
Container nodes		
	Compartment	A compartment is a logical or physical structure (e.g. a cell membrane) that contains entities. Entities located in two different compartments are considered as two different variables. Note that in SBGN, a compartment can take any geometry and that two adjacent compartments should be separated by two lines.
	Complex	A complex represents an association of biochemical entities (macromolecules, simple chemicals, etc.). The resulting entity can have its own identity, properties and function.
Auxiliary units		
	Unit of information	The SBGN unit of information can be attached to the border on entities or container nodes. Entities must in most cases be prefixed with a controlled vocabulary term indicating the type of information being expressed (www.sbgn.org).
	State variable	A state variable can be attached to the border of an entity, indicating different states of the entity. For example, indicating whether a macromolecule has been phosphorylated or not.
	Clone marker	A clone of an entity (simple chemical in this example) is indicated by a filled area and is used to indicate that at least one other occurrence of this entity can be found in the map.

Please note that this is just a selection of SBGN notations (for more information, see www.sbgn.org) and that transitions are predefined, including a consumption arc, a production arc and a process node.

of the current Systems Biology Graphical Notation (SBGN). The idea behind SBGN is to have several views of the same model by enabling three different graphical representations of a model: process description maps, entity relationship maps and activity flow maps [13,14]. A process description map illustrates the changes of entity pools in the system and the transitions between connected species. A given entity might appear several times in the same map, once all copies are properly marked. For example, ATP can be involved in several processes and should be defined as a clone marker (Box 1). A drawback of process description maps is that they quickly become very complex, which is typically seen in – for example – signaling cascades. To reduce all the underlying biochemical details of one entity affecting others, activity flow maps can be used instead. Such maps focus on the flux of information between the entities in the system. Entity relationship maps, by contrast, only describe the relationship between the entities and do not show the sequence of events. Entity relationship maps are particularly useful when one wants to illustrate the current understanding of molecular interactions of a specific system [14]. In this review, we focus on process description maps illustrated by the software PathwayLab, which is a commercial product with a 30 day trial version (http://www.innetics.com).

Modeling objects

SBGN process description maps consist of several different symbols, which are divided into different classes (Box 1). Note that Box 1 contains only the most fundamental symbols to represent a process description map, whereas a full and up-to-date view of all

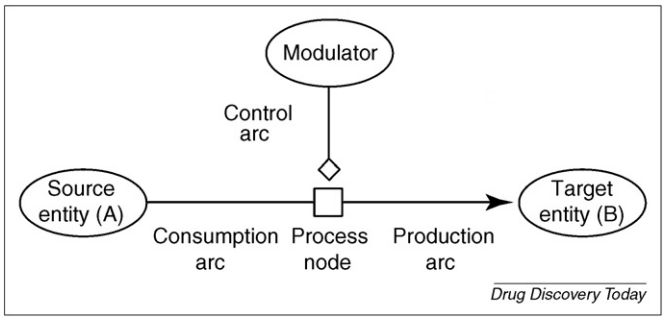


FIGURE 1

A Systems Biology Graphical Notation illustration of how a source entity is connected to a target entity through a process node. The graph also illustrates how a modulator should be connected to such process.

symbols can be found at www.sbgn.org. Building an SBGN process description map usually starts by drawing the types of entities that are involved in the system (simple chemical, macromolecule, nucleic acid feature, and so on), followed by attaching transitions between the entities. The transitions usually consist of a consumption arc, a process node and a production arc (Figure 1). Software tools such as PathwayLab and CellDesigner [15] have a convenient way of representing predefined transitions consisting of the connection arc, the process node and the production arc. Process nodes are generally represented as a square box or as a filled or open circle if the transition involves an association or dissociation, respectively, between the entities. If a transition is modulated by another entity, the process node should be connected to the

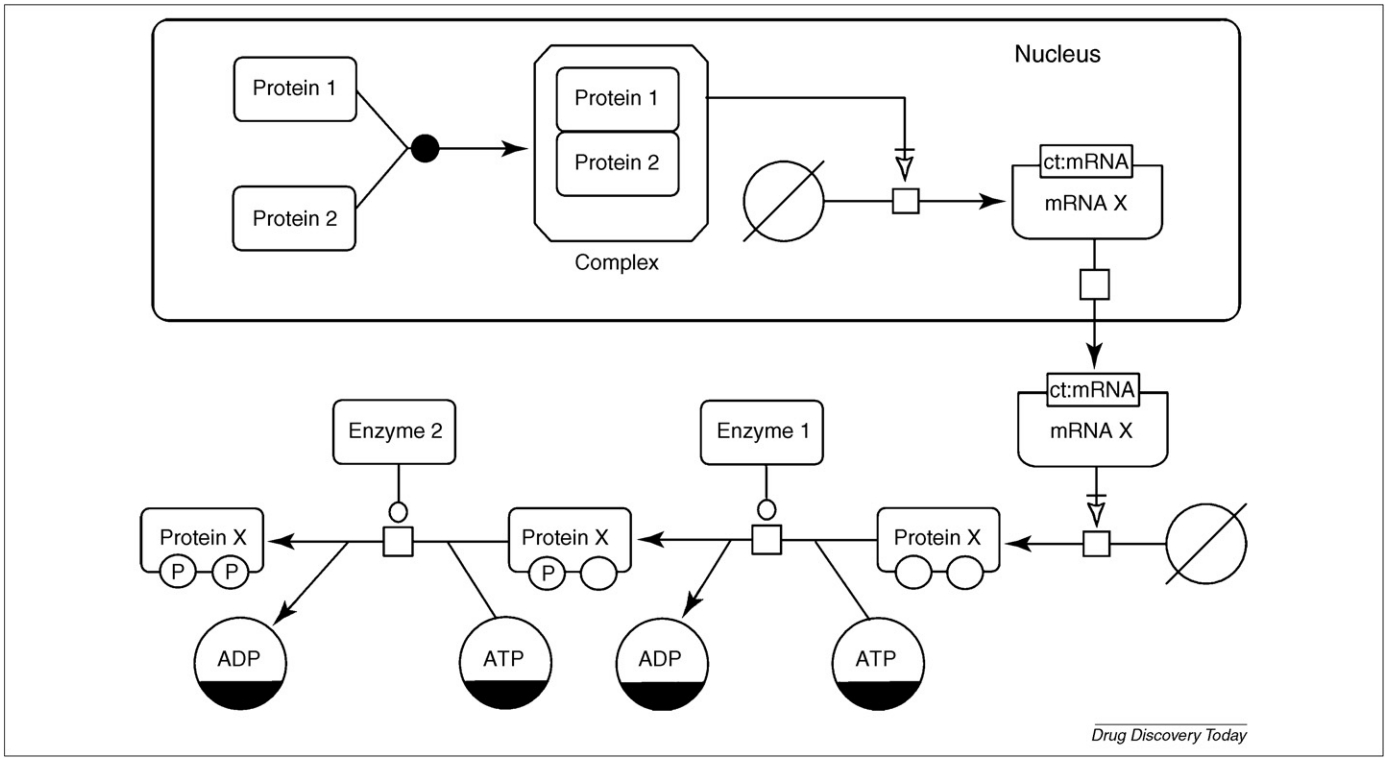


FIGURE 2

An illustrative example of a Systems Biology Graphical Notation process description map. Within the nucleus compartment, there is a complex formation of a transcription factor that triggers mRNA expression. The expressed mRNA is transported out of the nucleus compartment and triggers the translation of Protein X, which is phosphorylated in two steps, catalyzed by different enzymes.

modulator entity by a specific control arc, illustrating the type of modulation (e.g. catalysis or inhibition). There are complex nodes also, representing complexes formed by different entities and compartments that can represent whether entities are within a certain cellular compartment or not. Entity pool nodes and container nodes can hold auxiliary units on their borders (e.g. unit of information indicating the number of molecules within a multimer or state variable representing whether the entity is phosphorylated or not). To represent that an entity is duplicated in a map, a clone marker auxiliary unit should be used, which is illustrated by a filled area (less than 30% of the entity). A brief overview of some of the concepts in building an SBGN process description map is exemplified in Figure 2. This map is an illustrative example, showing how two macromolecules associate into a complex. This complex triggers the production of a nucleic acid feature (mRNA) within the nucleus compartment. The source symbol is used to ignore all the underlying details behind this production process. The mRNA also holds a unit of information, indicating its conceptual type. The mRNA is exported from the compartment and is triggering the production of a protein. The protein is then phosphorylated in two steps, where the state variables indicate the phosphorylation state. ATP is consumed in this process, and ADP is produced. These simple chemicals contain clone markers (partly filled), which indicate that there exist multiple instances of these entities. The control arcs attached to the nodes of these phosphorylation steps indicate that the enzymes catalyze these processes.

Mathematical modeling

A drawback of SBGN when translating graphics into mathematical equations is the lack of information on the underlying mathematical framework that should be used. In addition, the symbols do not map directly to mathematical functions; for example, the control arc 'inhibition' can be translated to many different types of mathematical functions of the underlying kinetics (competitive, uncompetitive, and so on). However, using individual symbols for each type of kinetic expression would cause an explosion of symbols and SBGN would no longer be an understandable standard for a broad scientific community. Kitano *et al.* [12] have, in addition, proposed the use of indexing to each arrow that could enable such information to complement the symbols.

A common type of theoretical approach in modeling intracellular reaction networks is ordinary differential equations (ODE). They are particularly useful for identifying drug targets because they can be analyzed with powerful simulation tools, such as TIde [16], to find key control points. In addition, they permit the direct use of kinetic constants that can be measured experimentally [17,18]. The disadvantages with ODE models, however, are that they assume high concentrations of the components in the system and that the components are uniformly mixed. If such assumptions are not fulfilled for the system under study, other methods have to be considered, such as spatial or stochastic models.

Here, we illustrate how an SBGN model is converted into ODEs using PathwayLab as a representative software tool. PathwayLab is a software tool, built on top of Microsoft Visio, for modeling, analysis, and information management of biochemical pathways. The tool streamlines the pathway building process by using its rich and flexible set of graphical building blocks for specifying biochemical entities, reaction and control mechanisms. The pathway models are built by dragging and dropping from SBGN stencils into a drawing page. This makes it very easy to rapidly build SBGN models of biochemical reaction networks such as metabolic or signaling pathways and gene regulatory networks. Each modeling object described in the previous section has a mathematical meaning, which will be described below. Hence, a biochemical reaction network built using these objects directly maps to a mathematical model, which can be used for simulations such as transient analysis, steady-state analysis, metabolic control analysis, or other sensitivity techniques.

The mapping between a biochemical reaction network described by the modeling objects of PathwayLab and its mathematical description results in the so-called 'reaction rate equations' for the network. This is a set of ODEs describing the time evolution of the concentrations (or amounts) of the biochemical entities involved in the network. Bold letters represent vectors, and the set of ODEs take the form

$$\dot{\boldsymbol{x}} = \boldsymbol{MR}$$

where the state vector $x \in \mathbb{R}^n$ is a vector of concentrations, $M \in \mathbb{R}^{n \times m}$ is the stoichiometric matrix and $r \in \mathbb{R}^m$ is the vector of reaction rates. The concentration of each entity can be represented as a dynamic variable or as an exogenous signal. Choosing 'dynamic variable' means that the concentration of the entity will be a component in the state vector $\mathbf{x}$. Initial conditions for entities represented by dynamic variables can also be set. An entity represented by an exogenous signal will only appear as an argument of the reaction rates. Exogenous signals can either be specified as symbolic expressions or be computed by interpolation of data from external files. The stoichiometric matrix, $\mathbf{M}$, is a constant matrix, which is usually very sparse. Positions of the non-zero elements follow directly from the reaction network structure and the values are given by the stoichiometric constants of the involved reactions. Components of the reaction rate vector, $\mathbf{r}$, are functions of the concentrations of the entity. For each transition, there is a corresponding reaction rate. The mathematical expression for a reaction rate is a function of the concentrations of entities connected to the transition either directly or via control arcs. Entities directly connected to a transition are called 'substrate' or 'product', depending on whether the corresponding entity is consumed or produced in the reaction. The qualification as substrate or product also has an impact on the signs of elements in the stoichiometric matrix. An entity represented by a dynamic variable, which is also the substrate of a particular reaction, will get a negative entry (typically -1) in the corresponding position in $\mathbf{M}$. Similarly, a product will result in a positive entry. An entity connected via a control arc is called a modifier because it is only affecting the reaction rate without being affected by itself. The rate function is formulated using local variable names for one or several substrates, products and modifiers $\mathbf{s}_i$, $i = 1, \ldots, n_s$, $\mathbf{p}_j$, $j = 1, \ldots, n_p$ and $\mathbf{m}_k$, $k = 1, \ldots, n_m$. A set of local parameters θ_l, $l = 1, \ldots, n_\theta$ is also used and the sometimes large rate expressions can be broken down into smaller pieces using auxiliary expressions. Hence, each component of the reaction rate vector has the form

$$\boldsymbol{r}_i = \boldsymbol{r}_i(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{m}, \theta), i = 1, \ldots, n.$$

Consider the reaction network depicted in Figure 2. In this model, the entities are represented using dynamic variables. The source symbols and the enzymes are specified as signals with constant

values. There are six transitions in this model: *Nucleus*.r_1 (association), *Nucleus*.r_2 (production of mRNA), *Nucleus*.r_3 (export of mRNA), r_4 (production of protein X), and r_5 and r_6 (phosphorylation). For simplicity, we drop the full names below because the short names of the rates are all unique. The reaction network structure is automatically translated into a set of ten ODEs, one for each dynamic variable. As an example, we will consider the variable for the intermediate phosphorylation step, where the equation becomes

$$\frac{dx_7}{dt} = r_5 - r_6$$

corresponding to the seventh row (0,0,0,0,1,−1) of the stoichiometric matrix. The user can manually enter an appropriate kinetic formula for each transition. In this case, we have chosen a two-substrate Michaelis–Menten kinetic formula specifying the rate r_5

$$r_5 = \frac{k_0 m_1 (S_1/K_{m1})(S_2/K_{m2})}{(1 + (S_1/K_{m1}))(1 + (S_2/K_{m2}))}$$

where S_1 and S_2 are the variable names internal to transition r_5 for the ATP and unphosphorylated protein X, m_1 is the modifier enzyme 1 connected to it and the rest are parameter names.

Simulation (computer analysis)

Simultaneous to the drag-and-drop and connection of modeling objects in a drawing page, an internal symbolic structure of the corresponding modeling objects is produced. In the process of simulation, the symbolic structure is parsed and an efficient byte code for the right-hand side of the reaction rate equations is generated. This byte code is linked to an ODE solver, which solves the set of equations. Once a simulation has been carried out, the user can select entities and transitions in the drawing area and study their simulated responses in a plot window. There is also means of utilizing the symbolic structure to compute steady states of the model. In the steady-state computation, a nonlinear algebraic solver first attempts to solve the static problem. If this fails, the reaction rate equations are simulated until a norm of the derivatives becomes small.

PathwayLab has built-in abilities to perform metabolic control analysis. This analysis is based on perturbations of rates and multiple simulations to obtain so-called flux and concentration control coefficients [19]. Metabolic control analysis corresponds to the computation of several sensitivity measures for a reaction network operating in steady state. For each transition (entity), there are as many flux (concentration) control coefficients as the total number of transitions in the model. The control coefficients for a particular transition or entity give a measure of how sensitive its steady-state level is to changes in other transition rates. PathwayLab also allows export of the model to SBML, Mathematica, and Matlab code, which enables more advanced analysis of the model.

Existing software tools that use SBGN

Several existing non-commercial software tools use SBGN in their graphical visualization. A tool that enables simulation of SBGN models, similar to PathwayLab, is CellDesigner [15]. CellDesigner not only features plug-ins such as SBML ODE Solver Library and Copasi [20] but can also utilize other modules to analyze the model in other software tools. An active area of investigation is to develop tools that automatically generate SBGN maps from SBML code. The visualization tool Arcadia allows import of SBML to generate SBGN graphs in a simple semi-automated style (http://arcadiapathways.sourceforge.net/). Tools such as PANTHER Pathway (http://www.pantherdb.org/pathway/), BioUML (http://www.biouml.org) and Edinburgh Pathway Editor (http://www.bioinformatics.ed.ac.uk/epe/) can also connect to databases to retrieve specific details of the modeling objects in the network, such as kinetics and gene expression data in the context of the maps. BioModels Database enables users to store published mathematical models in SBML or CellML formats, which facilitates exchange and reuse of models [5]. This database is also connected to simulating tools, such as JWS Online, where users can simulate the models stored in the database [6]. Another database that stores biological pathways is Reactome [21], which enables experts in the field to load and edit SBGN graphs that can be browsed to retrieve up-to-date information about a topic of interest. An up-to-date list of software tools that support SBGN can be found at www.sbgn.org.

Concluding remarks

The SBGN is an emerging standard for graphical notation, which is crucial for efficient and accurate communication of biological knowledge between researchers with various backgrounds. We have introduced how to build SBGN models using PathwayLab as one of several available SBGN editors. It has also been shown how such a tool converts a graphical model into a system of ordinary differential equations, allowing the user to simulate and further analyze the model. The software infrastructure in systems biology is advancing with standards such as SBML and CellML enabling different software tools to communicate and to use special functions for analysis. The progress of SBGN is promising, and the ultimate goal is to use such a standard to be able to visualize graphical models and enhance knowledge transfer in the systems biology community. This would increase the interpretation of the models, as well as facilitating the direct link to simulate the models in software tools, supporting such standards.

Acknowledgement

We are grateful to Nicolas Le Novére (EMBL-EBI) for helpful comments on the manuscript. This work has been funded in part by the European Commission through the BioSim network of excellence and the UNICELLSYS project.

References

1 Kitano, H. (2002) Systems biology: a brief overview. *Science* 295, 1662–1664
2 Liu, E.T. (2005) Systems biology, integrative biology, predictive biology. *Cell* 121, 505–506
3 Aderem, A. (2005) Systems biology: its practice and challenges. *Cell* 121, 511–513
4 Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531
5 Le Novère, N. *et al.* (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* 34, D689–D691

6 van Gend, C. *et al.* (2007) Data and model integration using JWS Online. *In Silico Biol.* 7, S27–S35
7 Lloyd, C.M. *et al.* (2004) CellML: its future, present and past. *Prog. Biophys. Mol. Biol.* 85, 433–450
8 Kohn, K.W. (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* 10, 2703–2734
9 Pirson, I. *et al.* (2000) The visual display of regulatory information and networks. *Trends Cell Biol.* 10, 404–408
10 Cook, D.L. *et al.* (2001) A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. *Genome Biol.* 2 RESEARCH0012
11 Demir, E. *et al.* (2002) PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* 18, 996–1003
12 Kitano, H. *et al.* (2005) Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.* 23, 961–966
13 Kitano, H. (2003) A graphical notation for biochemical networks. *BIOSILICO* 1, 169–176
14 Le Novère, N. *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.* 27, 735–741
15 Funahashi, A. *et al.* (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO* 1, 159–162
16 Schulz, M. *et al.* (2009) TIde: a software for the systematic scanning of drug targets in kinetic network models. *BMC Bioinformatics* 10, 344
17 Holmén, J. *et al.* (2009) A kinetic overview of the receptors involved in 1,25 dihydroxyvitamin D_3 and 24,25-dihydroxyvitamin D_3 signaling: a system biology approach. *Crit. Rev. Eukaryot. Gene Expr.* 19, 181–196
18 Schmidt, H. and Jirstrand, M. (2006) Systems Biology Toolbox for MATLAB: a computational platform for research in Systems Biology. *Bioinformatics* 22, 514–515
19 Fell, D.A. (1992) Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem. J.* 286, 313–330
20 Mendes, P. *et al.* (2009) Computational modeling of biochemical networks using COPASI. *Methods Mol. Biol.* 500, 17–59
21 Matthews, L. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 37, D619–D622