Towards a Semantic Description of Bio-Models: Meaning Facets – A Case Study

Christian Knüpfer
Artificial Intelligence Group
FSU Jena, Germany
tral@cs.uni-jena.de

Clemens Beckstein
Artificial Intelligence Group
FSU Jena, Germany

Peter Dittrich
Bio Systems Analysis Group
FSU & JCB Jena, Germany

Abstract

Systems Biology *reconstructs* biological phenomena in order to develop explanatory models of living systems. These models are represented precisely in terms of mathematical expressions. However, the meaning of a model usually is not formally specified but only described in natural language. Here, we discuss a framework for specifying the meaning of bio-models. We show that semantics appears on different levels: the meaning of the model as a whole, the meaning of the model's components, and the meaning of the model's behaviour. Each level has an intrinsic and extrinsic facet. We illustrate our framework by sketching what must be considered for a formal semantics of two simple numerical models of the cell cycle.

1 Introduction

Systems Biology *reconstructs* biological phenomena in order to develop explanatory models of living systems. These models are represented formally in terms of mathematical expressions or algorithmic statements. But often the meaning of a model is not formally specified and only described in natural language. A formal semantics describing the meaning of bio-models would not only be useful in application scenarios, such as model search and model integration, but also would support biologists in understanding the mathematical formalism.

In order to account for such a formal semantics of bio-models we have to investigate and characterise the meaning of the models in detail (Section 2). We show that semantics appears in different facets. A formalisation of the semantics has to account for these facets and essential for the relations between them. In this paper we will introduce the meaning facets and then briefly sketch how they can be used in order to systematically reconstruct the meaning of mathematical models of the cell cycle (Section 3).

2 Meaning Facets of Explanatory Models

In biological systems complex dynamical behaviour usually arises from tangled interactions between countless entities. In order to understand the mechanisms that drive such a complicated system biologists typically substitute the real system by a simplified model which is supposed to imitate just those aspects of the system that are needed for simulation. The justification for this is *explanation*: If the results obtained under plausible boundary conditions from a simulation model are consistent with the observable behaviour of the modelled system and if in addition the components of the model possess a meaning wrt. the modelled system then this explanation plays an important role in the understanding of the real system. Following the traditional performance vs. competence distinction of (Chomsky, 1965), such a model is a *competence model*.

We will call such a mathematical competence model of a biological phenomenon a *bio-model*. A bio-model comprises a binary relation between an expression in some formalism, the model in a narrower sense, and a biological phenomenon, i.e. the *intended meaning* of the model. While the model itself almost always is given in a formal way the intended meaning typically is only described informally in natural language. The same holds for the specification of the meaning of the components of the model and for a large part of the boundary conditions. Thus, systems biologists give semantics to the bio-models they deduce from their experiments, but this semantics typically is amenable only for human biologists and not for computers.

	intrinsic	extrinsic
intentional	In which formalism is the model formulated? How are expressions in the formalism interpreted or executed? How is the formalism used to simulate the behaviour?	What do the model stands for? Which biological phenomena does the model describe? What are the described biological systems and processes?
structural	What is the formal structure of the model? Which mathematical formalism is used? What are the mathematical objects of the model (equations, terms, variables)?	What are the biological meanings of the components of the model? Which model entity maps onto which biological object or process?
behavioural	Which characteristic types of dynamical behaviour can be observed? What are typical runs of the simulation model and which parameter settings are used therefor?	Which biological phenomena correlate with characteristic types of dynamical behaviour? Which experimental data are reproduced by which run of the simulation model?

Table 1: Meaning Facets of a Bio-Model

There is a main distinction in the meaning of a bio-model given in some mathematical formalisation: (1) The mathematical expression bears meaning by itself without referring to any biological reality it stands for. We can interpret and analyse (maybe through numerical simulation) a formal expression without knowing what it represents. This is what we call the *intrinsic meaning* of a model. (2) A pure intrinsic model would be a vain effort. What a model distinguishes from a pure syntactical formal expression is its surrogateness: The model describes some piece of reality and thereby bears an *extrinsic meaning*.

Orthogonal to this intrinsic/extrinsic dimension of the meaning of a bio-model three pragmatic levels can be identified: The model for itself has an *intention*, it has some given *structure* and when it is used it shows a *behaviour*. These meaning facets of a bio-model are summarised in Table 1.

The pragmatic levels can be seen as levels of concreteness: In the intrinsic case the intensional level defines the set of valid models in the given formalism. One of the valid models is chosen on the structural level and it constrains the possible patterns of behaviour. With concrete boundary conditions one can select some of the possible behaviours on the behavioural level.

In the extrinsic case the intensional level focuses on specific biological phenomena. In order to explain these phenomena one has to choose the relevant biological objects and processes that connect them. For this purpose potential mechanisms must be identified that can explain the phenomena. On the behavioural level one looks whether the supposed mechanisms can explain and predict the observed behaviours of the biological system.

3 A Case Study: Models of the Cell Cycle

In the following we will illustrate our meaning facets with the help of a case study – two models of the cell cycle from (Tyson, 1991):

Model 1:

$$
\begin{aligned}
\mathrm{d}[\mathrm{C2}]/dt &= k_6[\mathrm{M}] - k_8[\sim\mathrm{P}][\mathrm{C2}] + k_4[\mathrm{CP}] \\
\mathrm{d}[\mathrm{CP}]/dt &= -k_3[\mathrm{CP}][\mathrm{Y}] + k_8[\sim\mathrm{P}][\mathrm{C2}] - k_4[\mathrm{CP}] \\
\mathrm{d}[\mathrm{pM}]/dt &= k_3[\mathrm{CP}][\mathrm{Y}] - [\mathrm{pM}]F([\mathrm{M}]) + k_5[\sim\mathrm{P}][\mathrm{M}] \\
\mathrm{d}[\mathrm{M}]/dt &= [\mathrm{pM}]F([\mathrm{M}]) - k_5[\sim\mathrm{P}][\mathrm{M}] - k_6[\mathrm{M}] \\
\mathrm{d}[\mathrm{Y}]/dt &= k_1[\mathrm{aa}] - k_2[\mathrm{Y}] - k_3[\mathrm{CP}][\mathrm{Y}] \\
\mathrm{d}[\mathrm{YP}]/dt &= k_6[\mathrm{M}] - k_7[\mathrm{YP}]
\end{aligned}
$$

$$
\begin{aligned}
F([\mathrm{M}]) &= k'_4 + k_4([\mathrm{M}]/[\mathrm{CT}])^2 \\
[\mathrm{CT}] &= [\mathrm{C2}] + [\mathrm{CP}] + [\mathrm{pM}] + [\mathrm{M}]
\end{aligned}
$$

Model 2:

$$
\begin{aligned}
\mathrm{d}u/dt &= k_4(\nu - u)(\alpha + u^2) - k_6 u \\
\mathrm{d}\nu/dt &= (k_1[\mathrm{aa}]/[\mathrm{CT}]) - k_6 u
\end{aligned}
$$

$$
\begin{aligned}
u &= [\mathrm{M}]/[\mathrm{CT}] \\
\nu &= ([\mathrm{Y}] + [\mathrm{pM}] + [\mathrm{M}])/[\mathrm{CT}] \\
[\mathrm{CT}] &= [\mathrm{C2}] + [\mathrm{CP}] + [\mathrm{pM}] + [\mathrm{M}] \\
\alpha &= k'_4/k_4
\end{aligned}
$$

Both models describe the building and activation of the maturation promoting factor (MPF), a hetero dimer made of the two proteins cyclin and cdc2. *Model 1* does this by means of a set of ordinary differential equations (ODEs) where each equation models the temporal evolution of the concentrations of one of the involved substances wrt. the concentrations of the other substances. Involved substances are: cdc2 ($\mathrm{C2}$), phosphorylated cdc2 (CP), inactive MPF (pM), active MPF (M), cyclin (Y), phosphorylated cyclin (YP), total cdc2 (CT), adenosine triphosphate ($\sim\mathrm{P}$) and amino acids (aa). *Model 2* is a mathematical abstraction of *Model 1* under certain additional biological assumptions. The k_i are kinetic rate constants. For details see (Tyson, 1991).

We choose these two models for the following reasons: (1) They are models of a prominent and well-known biological system. (2) They are small in the number of variables and equations. Nevertheless they exhibit surprisingly many problems in capturing and formalising their semantics. (3) There are other small models of the same biological system that we plan to use for semantics-based comparison and integration in future investigations (e.g. Goldbeter, 1991). (4) The models are contained in the *BioModels* Database (Le Novére et al., 2006), a database of bio-models that provides some basic grounding for model components (cf. Section 5). (5) One of the models (*Model 2*) is an abstraction of the other (*Model 1*).

Table 2 sketches the meaning facets of *Model 1*. Some details are suppressed here because of space limitations. There is a similar table for *Model 2* which is not shown here. The main difference in the meaning of the two models is in the mapping between the variables and the biological entities they stand for. In *Model 1* there is a simple mapping between each variable and a sort of molecules in the cell. The situation for *Model 2* is more complicated: here a variable is an arithmetic expression over variables from the first model.

Key questions that had to be answered in order to arrive at the tables were: How should a model be understood? Which facets are required to understand the model? What are the sources of the understanding? Which meaning facets are rather implicit? What is the final ground the understanding rest upon? The answers are an account of the meaning facets (cf. Table 1) from the perspective of a human biologist.

4 Semantical Application Scenarios

An important next step towards a semantic description of bio-models is to develop a formalism that allows to represent the meaning. The choice of this formalism must be guided by the intended application. In the following scenarios a formal semantic description of bio-models would be desirable:

Semantics based search: For example, both models discussed in this work should be retrievable by search queries of the following types: "Find models related to p34 protein kinase!", "Find models that describe the interaction between cdc2 and cyclin!", or "Find models that exhibit both steady state and cyclic behaviour!".

Model comparison: Given two models, do they semantically overlap? Is one model a sub-model of the other? Or is one of them an abstraction of the other? A method for model comparison in general is needed for many higher level tasks, like e.g. model matching or model integration.

Model integration: Given two models that semantically overlap, what would an integrated model look like? Again, a formal semantics of the model's components is needed in order to automate this task.

Model understanding: A mathematical model is a set of symbols. On the other hand it is also an abstract description of biological phenomena, such as the ability of a cell to change its behaviour form metaphase arrest to growth-controlled division cycle. By unifying the relation between model and its meaning, we expect models to become more accessible to biologists who frequently are not modelling experts.

Model usage: In order to simulate and predict the behaviour of a biological system the bio-model has to be implemented in a computer code. This causes further problems: Without a formal semantics a biologist has to modify the code in order to change the model. If a formal semantics is given, it would be possible to modify the model on a more abstract semantic level without the need to refer to the implementation.

Model mining: In order to semi-automatically extract the meaning of a given model (maybe together with a paper explaining it) one needs means to catch and verify this meaning: a formal semantic description.

5 Related Work

There are other projects related to our work: CMSBlib (Soliman and Fages, 2004) is a "library of computational models of biological systems", which has recently started and contains currently six models. A slightly larger database named *BioModels Database* (*BioModels* in brief) (Le Novére et al., 2006) links a model and its constituents to external resources like database entries. The links are formalised by using RDF and can be augmented by qualifiers adapted from the *Dublin Core*.

An application of our meaning facets to *BioModels* discloses certain shortcomings with this approach. Several of the semantical applications sketched above are only partially realisable

	intrinsic	extrinsic
intentional	Mathematically, the model is a system of coupled ODEs in one common independent variable t and a set of boundary conditions. It can be simulated with numerical methods.	The model describes the interaction between cdc2 and cyclin when forming MPF. MPF controls the major events of the cell cycle. Thus the model also describes the control of the cell cycle.
structural	There is a special variable t. The dependent variables $[\mathrm{C2}]$, $[\mathrm{CP}]$, $[\mathrm{pM}]$, $[\mathrm{M}]$, $[\mathrm{Y}]$, $[\mathrm{YP}]$ in the equations implicitly are functions of t. The same holds for the variables $[\sim\mathrm{P}]$, and $[\mathrm{aa}]$, which are assumed to be constants. There are two special variables $[\mathrm{CT}]$ and F, which can be seen as abbreviations of some arithmetic expressions over the other variables and constants. The other atomic symbols k_i $(i = 1, \ldots, 9)$ and k'_4 are genuine constants. The model comprises of six ODEs of the form $\mathrm{dV}/dt = E$, one for each of the dependent variables, with V being the dependent variable and E a recursively defined arithmetic term in the set of variables and constants.	For the mapping between entities of the mathematical model and objects and processes in the world the following must hold: 1. The special variable t represents time. 2. The variables $[\mathrm{S}]$ stand for concentrations of specific substances S in the cell. 3. The constants k_i and k'_4 represent reaction rates for specific reactions between the substances. 4. Some terms represent rates of change of concentration of substances through specific reactions. 5. The equations represent the rates of change of concentrations of the corresponding substances.
behavioural	An analysis of the dynamical properties (Tyson, 1991) reveals three qualitative modes of dynamical behaviour: (a) steady state with high values of $[\mathrm{M}]$, (b) spontaneous oscillation and (c) excitable switch.	One can map the modes of dynamical behaviours of the model to biological phenomena: (a) metaphase arrest, (b) rapid division cycles in early embryos, (c) growth-controlled division cycles in non embryonic cells (Tyson, 1991).

Table 2: Meaning facets of *Model 1*

with *BioModels*. *BioModels* only allows to specify the extrinsic meaning at the intentional and the structural level. Neither is there an account for the intrinsic part of the models nor for the behavioural level at all.

Because the behavioural level is not acknowledged in *BioModels* it is not possible to answer model queries such as: "Find models that exhibit both steady state and cyclic behaviour!". Another problem is that the extrinsic meaning of variables like u in *Model 2* is not specified in *BioModels*. The variable u is an abstract mathematical object which requires the integration of more than one extrinsic link for the specification of their semantics. This can only be achieved by taking into account the mathematical structure of the model and by allowing for composed concepts as extrinsic references.

6 Conclusion

In this paper we introduced important meaning facets that should be employed in order to arrive at a reasonable formal semantics of bio-models. We illustrate this using Tyson's cell cycle models. Our meaning facets can be seen as a methodological commitment that should be followed when modelling biological processes. They offer a set of criteria for systematically constructing bio-models and for reconstructing their meaning. Together they build a promising framework that can be used for the derivation of a formal semantics and that will provide a sound basis for the construction of higher level semantical applications.

We acknowledge financial support by the *Rosa Luxemburg Foundation* and by the *Federal Ministry of Education and Research* Grant 0312704A.

References

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Mass.

A. Goldbeter. 1991. A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proc Natl Acad Sci USA*, 88(20):9107–9111, Oct.

N. Le Novére, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J.L. 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(Database issue):689–691, Jan.

Sylvain Soliman and François Fages. 2004. CMBSlib: a library for comparing formalisms and models of biological systems. In Vincent Danos and Vincent Schächter, editors, *CMSB'04: Proceedings of the second Workshop on Computational Methods in Systems Biology*, Lecture Notes in BioInformatics. Springer Verlag.

JJ. Tyson. 1991. Modeling the cell division cycle: cdc2 and cyclin interactions. *Proc Natl Acad Sci USA*, 88(16):7328–7332, Aug.