




An Ontology to Structure Biological Data: The Contribution of Mathematical Models

Olivier Inizan¹(✉) , Vincent Fromion¹ , Anne Goelzer¹ , Fatiha Saïd² ,
and Danai Symeonidou³ 

¹ Université Paris Saclay, INRAE, MaIAGE, Jouy-en-Josas, France
olivier.inizan@inrae.fr

² LISN, Université Paris Saclay, CNRS UMR9015, Orsay, France

³ INRAE, SupAgro, UMR MISTEA, Université de Montpellier, Montpellier, France

Abstract. The biology is a research field well known for its huge quantity and diversity of data. Today, these data are still recognized as heterogeneous and fragmented. Despite the fact that several initiatives of biological knowledge representation have been realized, biologists and bioinformaticians do not have a formal representation that, at the level of the entire organism, can help them to organize such a diversity and quantity of data. Recently, in the context of the whole cell modeling approach, the systemic mathematical models have proven to be a powerful tool for understanding the bacterial cell behavior. We advocate that an ontology built on the principles that govern the design of such models, can help to organize the biological data. In this article, we describe the first step in the conception of an ontology dedicated to biological data organization at the level of the entire organism and for molecular scales i.e., the choice of concepts and relations compliant with principles at work in the systemic mathematical models.

Keywords: Ontology · Mathematical models · Biological data

1 Introduction

The recent advances of sequencing technologies lead to a faster and cheaper production of data in the field of biology [14]. Biologists and bioinformaticians have to deal nowadays with a huge quantity and diversity of *omics* data (such as genomics, transcriptomics, proteomics, metabolomics and metagenomics) [8]. These data are mostly obtained in a given context of an experimentation to answer a particular question. From a wider perspective they appear to be heterogeneous and fragmented [2]. Moreover, despite the fact that there are many data available for a given organism, the ability to organize and integrate these data remains a challenge [10]. Such integration can be of great importance, and we can cite, among others, the elucidation of mechanisms to understand and treat diseases [13]. It should also be noticed that, despite an active research

activity in biological knowledge representation [12], there is no formal representation dedicated to data organisation for molecular scales, at the level of the organism. The lack of such a representation prevents scientists from exploiting the full potential of these data. Since a decade, the whole cell modelling approach has showed that systemic mathematical models are a powerful tool for describing and understanding the bacterial cell behavior. More precisely, through these models, when fed with biological data, it is possible to identify organizational principles on which (unobserved) cell behavior can be predicted [3,9]. Therefore, there is a real need to develop a new formal representation that can semantically represent the links between biological data while ensuring compliance with biological principles followed in mathematical modelling of biological processes.

In this article we present the first steps in the development of a formal representation dedicated to biological data organisation and designed according concepts that hold in mathematical models. We want to underline that this work is an ongoing research, the tasks realized so far are mainly conceptual and concrete realisations have been done on examples as proof of concepts. The rest of the document is organized as follows. In Sect. 2 we present the state of the art of this work and its main motivation. The concepts and relations of the ontology are described in Sect. 3 and illustrated with an example in Sect. 4. The conclusion and perspectives are provided in Sect. 5.

2 State of the Art and Motivation

To understand the motivation of the present work we have to detail two starting points: the BiPom and BiPON [5,6] ontologies and the constraints relative to mathematical models.

2.1 BiPON and BiPom: New Potential Rules and Usage for Bio-ontologies

Biology is a rich field of knowledge where several communities can work on the same object for different purposes. Being able to avoid ambiguities when referring to the same object is then crucial. Consequently, well known bio-ontology projects (for example GO [1]) provide a hierarchy of concepts used as controlled vocabulary. Another usage can be found in the BioPax community [2] where the ontology is designed to collect and exchange data related to biological pathways. In 2017 and 2020, two OWL¹ ontologies, BiPON and BiPom, have provided new potential rules and usage for bio-ontologies: first, they introduce the systemic approach as a design principle to represent the biological knowledge. This approach originates from the field of engineering science and aims to break down a given system into linked (sub) modules [4]. In this context (Fig. 1a), the notion of systemic module is strictly defined by its inputs, outputs and the function it fulfills. Inputs, outputs and function are then tied together in a mathematical model which gives a formal description of the behavior of the module. The

¹ <https://www.w3.org/OWL/>.

authors of BiPON and BiPOm have showed that the bacteria cell can be considered as a system and be organized in linked and interlocked systemic modules. These systemic modules are OWL concepts typed as *biological processes*. Second, BiPON and BiPOm provide a high level of expressiveness in comparison with other bio-ontologies. From their initial set of concepts, relations, rules and individuals, they exploit the reasoning capacity provided by the OWL language and the Description Logic to infer new relations between individuals. As a result, the authors of BiPON have showed that a wide diversity of biological processes can be described by few concepts of mathematical models.

2.2 The Constraints of Mathematical Models

As presented in Fig. 1b, a mathematical model is associated to a *biological process*. In this section we focus on the constraints that drive the construction of such mathematical models. To understand the importance of these constraints, we first have to detail a little more the notion of biological processes defined in the ontologies BiPON/BiPOm. A biological process has one or several molecules as inputs and also one or several molecules as outputs. We consider that a process *consumes* the inputs and *produces* the outputs. Moreover, a biological process has a function which is the objective to fulfill. Finally, the process has means to transform inputs into outputs and these means are expressed through a mathematical model. In Fig. 1a the general form of a biological process is presented. In Fig. 1b, we represent a simple biochemical reaction (a molecule 'A' is converted into the molecule 'B') and the corresponding biological process.

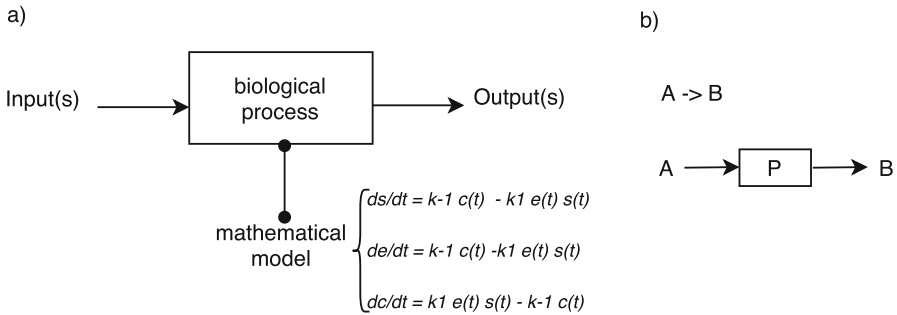


Fig. 1. a) The general form of a biological process and its associated mathematical model. b) A simple biochemical reaction and its process P.

A striking fact in the modeling community is that, whatever the mathematical model being build, three general constraints are always satisfied. Consequently, we consider that (i) these constraints are major and (ii) they drive the construction of mathematical models. These constraints, presented below, will be referred in the sequel as *model constraints*:

1. *The physical causality.* The physical causality states that if the inputs produce the outputs, then the inputs precede the outputs. Since we do not especially consider the time in the formal representation, the causality can be reformulated as follows: if the inputs are present in a sufficient quantity, then the process can consume the inputs and produce the outputs.
2. *The mass conservation.* It is an important constraint of the modelling approach that ensures the consistency of the models.
3. *The concurrency of access.* The biological processes are in concurrence to access the same type of entity. More precisely, the same type of molecule can be consumed or produced by different processes. A classic example is the ATP molecule which provides energy for the cell and that is consequently consumed by different chemical reactions.

It is important to notice that, despite the fact that the concept of the *biological process* is present in BiPON/BiPOm and that mathematical models are represented in BiPON, none of these ontologies considers these model constraints.

2.3 Motivation

Our motivation lies in the fact that the model constraints represent a powerful tool to validate the consistency of the biological knowledge and data relative to an organism. If we want to consider these constraints in a formal representation, we should first provide concepts and relations that allow us to *count* the molecules that are consumed or produced by the biological processes. Consider the simple example of the Fig. 1b: the physical causality states that at *least one* molecule A must be available for the process P. The mass conservation states that *one* molecule A must be converted into *one* molecule B. Considering the concurrence between the processes implies also counting the molecules: consider a second process P' that consumes also a molecule A. If there is *only one* molecule A in the entire cell, P and P' are in concurrence. But if there are *two* molecules A, then P and P' are not in competition. As already mentioned in Sect. 2, BiPON and BiPOm have validated the systemic approach to represent the biological knowledge. However, none of these ontologies allow to count the entities consumed and produced by the processes. This drawback prevents the representation of model constraints and leads us to build a new ontology.

3 First Components of a Bio-ontology for Data and Knowledge Organization

As mentioned in the previous section, we want to provide a representation that takes into account the model constraints that drive the construction of mathematical models. We have shown that, to achieve this goal, we have to count the entities (the molecules) that are consumed or produced by the processes. In this section, we propose a first set of concepts and relations of our bio-ontology that allow counting the entities (see Sect. 3.1). These concepts and relations can help us to give a more formal definition of the biological process. This definition is presented in Sect. 3.2.

3.1 A First Set of Concepts and Relations to Count Entities

In this work, we adopt the formal approach and the concept of *biological process* presented in BiPON/BiPOm. In order to take into account the model constraints (i.e., physical causality, mass conservation and concurrency of access), we use the concepts that are frequently manipulated by the modelers [15]. We first, create the concept *pool* that groups all the molecules of the same biochemical entity into *pools*. For example, all the molecules of water will be grouped in the H₂O *pool*. Second, since the *pool* is of a finite volume, the number of molecules is given by the *concentration* of molecules in the *pool*. Third, we state that the *processes* can communicate with each other only via the *pools*. This leads us to create three relations (i.e., *reads*, *retrieves* and *puts*) and another concept: (i) a *process reads* the *concentration* of the molecules in the *pool* and (ii) the *process retrieves* molecules from the *pool* and/or *puts* molecules into the *pool*. By this way, the *process triggers* a *flow* of molecules. The Fig. 2a illustrates how we represent the concepts of *pool*, *process*, *concentration* and *flow*. The relation *triggers* and *reads* are also represented. In the Fig. 2b we represent the simple example showed in Fig. 1b where the *process* P converts the molecule A into the molecule B.

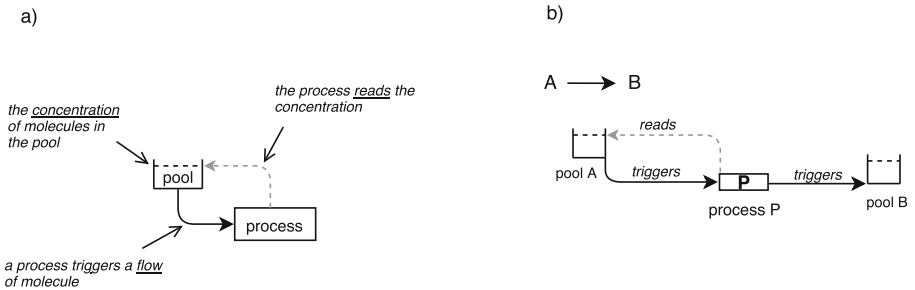


Fig. 2. a) The concepts and relations of the new ontology. b) A simple biochemical reaction represented with these concepts and relations.

The Fig. 2b can be detailed as follows: the *process* P *reads* the *concentration* of molecules in the *pool* A (dashed grey arrow). If there are enough molecules (here only one molecule is required), P *retrieves* this molecule (aka P trigger a *flow* of molecule A (first black arrow)) and *puts* a molecule B in the *pool* B (aka P *triggers* a *flow* of molecule B (second black arrow)).

3.2 A Formal Definition of a Biological Process

The set of concepts and relations designed above is a convenient way to go further in the definition of a *biological process* provided by BiPON/BiPOm. These ontologies describe a *biological process* through the relations *has_input* and *has_output* with the molecules that participate in the biochemical reaction. We propose to re-formulate the behavior of the *biological process*. We explain

this re-formulation through the example of Figs. 1 and 2. While BiPON/BiPOM state that the *process* P *has_input* the molecule A and *has_output* the molecule B, we state that the *process* *reads* the *concentration* of molecules A and (if there are enough molecules) *triggers* a *flow* of molecules A and a *flow* of molecules B. Expressing the behavior by this way is more compliant with the constraint stated by the physical causality: the fact that there is enough *concentration* of molecules is the cause of the behavior of the *process* while the *flow* of molecules is considered as its effect. With these considerations we can provide a formal definition of a *biological process*.

A *biological process* is defined as a concept characterized by its inputs and its outputs:

$$\begin{aligned} \text{BiologicalProcess} \equiv \exists \text{has_input.Input} \sqcap \forall \text{has_input.Input} \\ \sqcap \exists \text{has_output.Output} \sqcap \forall \text{has_output.Output} \end{aligned} \quad (1)$$

An input is the *concentration read* by the *process*:

$$\text{Input} \equiv \text{Concentration} \sqcap \exists \text{is_read_by.BiologicalProcess} \quad (2)$$

An output is a *flow* of molecules *triggered* by a biological process:

$$\text{Output} \equiv \text{Flow} \sqcap \exists \text{triggered_by.BiologicalProcess} \quad (3)$$

We note that the definition of *biological process* is cyclic, since it is defined by the inputs and outputs which are in their turn defined by the *biological process*. Such definitions are very common in ontology design and the cycles can be solved during the ontology population by defining the order according to which individuals are created in the ontology.

4 Illustration on an Example

We illustrate the use of the ontology with the example of a biochemical reaction catalyzed by an enzyme. This biochemical reaction is representative of the metabolic processes within the whole-cell, i.e. one of the most important set of biological processes involving almost one third of the bacterial genes. Therefore, if it can be represented by the concepts and relations introduced in Sect. 3.1, a large part of the biological processes of the cell could be described accordingly, which constitutes a first step in the ontology evaluation. The chemical model of this biochemical reaction proposed by Michaelis and Menten [11] occurs in two reactions:



First, the enzyme E binds to the substrate S to form a complex $[ES]$. This reaction is reversible i.e., the complex $[ES]$ can dissociate to release the enzyme E

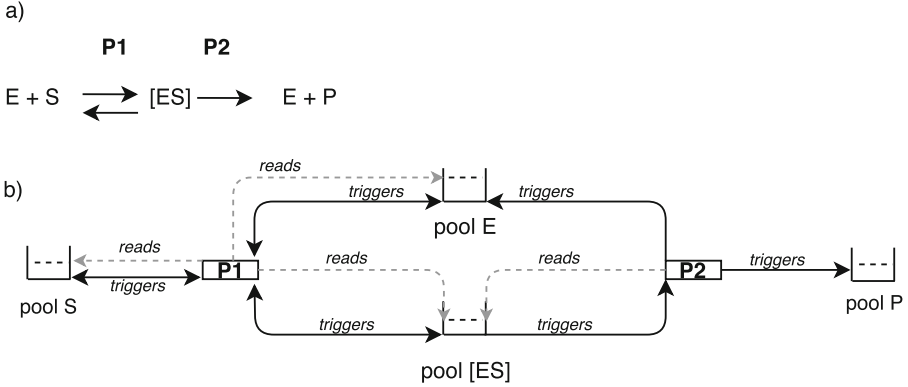


Fig. 3. a) The model provided by Michaelis and Menten b) The representation of the model with the processes, pools and relations.

and the substrate S . In contrast, the second reaction is irreversible: the complex $[ES]$ dissociates to release the enzyme E and the product of the reaction P .

To represent this chemical model with the concepts and relations proposed above, we first design two *processes*, P1 and P2, each one corresponding to the first and second reaction, respectively. We then design four *pools* named S, P, E and ES, one for each type of molecule, i.e., substrate, product, enzyme, and enzyme bound to the substrate, respectively. The *processes*, *pools* and relations are represented in Fig. 3. In Fig. 3b, P2 reads the concentration of the pool ES and (if there is enough molecules of ES) triggers a flow of molecules E, P and ES. P2 consumes E and P, and produces ES. The process P1 represents a reversible reaction. For the forward reaction ($E+S \rightarrow [ES]$) P1 reads the concentration of the pool E and S and triggers a flow of E, S and ES. For the reverse reaction ($[ES] \rightarrow E+S$), P1 reads the concentration of the pool ES and triggers a flow ES, E and S. In the ontology, the forward and reverse sub-reactions of P1 are not distinguished: P1 reads the concentrations of all pools (ES, S and E) and for each pool, P1 triggers a single flow (corresponding to the sum of the flow of each sub-reactions). By doing so, the constraint of causality is well respected.

5 Conclusion and Perspectives

In this article, we have described the first steps of the development of an ontology dedicated to the organization of biological data. This ontology has been designed according to the constraints that hold in mathematical models. The concepts and relations (i) make possible the representation of quantities, (ii) have been validated on a representative example and (iii) led us to give a new formal definition of a biological process. We plan to populate the ontology with an entire network of reactions [5], using the SBML format [7]. During this population, quantities could be associated with the concepts representing the concentrations and the

flows, and the model constraints could be expressed with SHACL² language. This work fits in the challenge of making ontologies more expressive including more quantitative knowledge. This will allow us to check the consistency and the validity of knowledge and their associated data.

References

1. Gene Ontology Consortium: The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**(D1), D330–D338 (2019)
2. Demir, E., Cary, M.P., Paley, S., et al.: The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* **28**(9), 935–942 (2010)
3. Goelzer, A., Muntel, J., Chubukov, V., et al.: Quantitative prediction of genome-wide resource allocation in bacteria. *Metab. Eng.* **32**, 232–243 (2015)
4. Hartwell, L.H., Hopfield, J.J., Leibler, S., et al.: From molecular to modular cell biology. *Nature* **402**(6761), C47–C52 (1999)
5. Henry, V., Saïs, F., Inizan, O., et al.: BiPOm: a rule-based ontology to represent and infer molecule knowledge from a biological process-centered viewpoint. *BMC Bioinform.* **21**(1), 1–18 (2020)
6. Henry, V.J., Goelzer, A., Ferré, A., et al.: The bacterial interlocked process ONtology (BiPON): a systemic multi-scale unified representation of biological processes in prokaryotes. *J. Biomed. Semant.* **8**(1), 1–16 (2017)
7. Hucka, M., Finney, A., Sauro, H.M., et al.: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**(4), 524–531 (2003)
8. Joyce, A.R., Palsson, B.Ø.: The model organism as a system: integrating ‘omics’ data sets. *Nat. Rev. Mol. Cell Biol.* **7**(3), 198–210 (2006)
9. Karr, J.R., Sanghvi, J.C., Macklin, D.N., et al.: A whole-cell computational model predicts phenotype from genotype. *Cell* **150**(2), 389–401 (2012)
10. López de Maturana, E., Alonso, L., Alarcón, P., et al.: Challenges in the integration of omics and non-omics data. *Genes* **10**(3), 238 (2019)
11. Michaelis, L., Menten, M.L., et al.: Die kinetik der invertinwirkung. *Biochem. z* **49**(333–369), 352 (1913)
12. Nicolas, J.: Artificial intelligence and bioinformatics. In: Marquis, P., Papini, O., Prade, H. (eds.) *A Guided Tour of Artificial Intelligence Research*, pp. 209–264. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-06170-8_7
13. Ramon, C., Gollub, M.G., Stelling, J.: Integrating-omics data into genome-scale metabolic network models: principles and challenges. *Essays Biochem.* **62**(4), 563–574 (2018)
14. Reuter, J.A., Spacek, D.V., Snyder, M.P.: High-throughput sequencing technologies. *Mol. Cell* **58**(4), 586–597 (2015)
15. Voit, E.O.: *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*. Cambridge University Press, Cambridge (2000)

² <https://www.w3.org/TR/shacl/>.