Theoretical Computer Science 587 (2015) 104–124

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Component identification in biochemical pathways

Giovanni Pardini *, Paolo Milazzo, Andrea Maggiolo-Schettini

Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy

ARTICLE INFO

Article history:
Received 14 April 2014
Received in revised form 3 December 2014
Accepted 3 March 2015
Available online 12 March 2015

Keywords:
Biochemical pathways
Modular modeling
Formal methods

ABSTRACT

Biochemical pathways are abstract descriptions of the interactions among the molecular species involved in a cellular process. Different molecular species mentioned in a pathway often represent different states of the same biological entity, such as the unbound and bound states of a certain molecule. Hence, a pathway can be seen as a network of interactions between entities changing state synchronously by means of reactions. We consider such biological entities as pathway components.

We define a semi-automatic algorithm to infer the components from their interactions described in the pathway. In case the interactions are not sufficient to resolve all the reactions, help from a domain expert may be needed to resolve any ambiguity that should arise. As an example of application, we apply the algorithm to a model of the EGF signaling pathway from the literature in order to identify its components. From the theoretical point of view, we formally prove the correctness of the algorithm, its termination under any input pathway and a (weak) confluence property.

1. Introduction

Biochemical pathways are networks of interactions among biological entities such as proteins, DNA, RNA and other molecules, taking place inside cells. The interactions constituting a biochemical pathway are typically bindings and unbindings of molecular species, syntheses and degradations of proteins, conformational changes and translocations of molecules (or complexes) from one compartment to another. Each biological entity involved in a pathway usually appears in several different molecular species mentioned in the pathway. For instance, the same protein may appear in its initial form (as just synthesized), but also in an activated form (e.g. phosphorylated) and as a part of some complexes. The forms that a biological entity may take are often represented as different molecular species, namely with different names. Moreover, complexes (that involve different biological entities) are often associated with a single name, that may not allow the originating biological entities to be easily identified. The reconstruction of the set of biological entities involved in a given pathway is the problem we face in this paper.

In [1] we proposed a modular verification approach based on [2] that allows properties of the pathway to be verified efficiently by applying model checking to an abstraction of the pathway semantics obtained by focusing on the behavior of a subset of the involved biological entities. To this aim we considered a notion of *molecular component* that is the counterpart in a pathway model of the notion of biological entity in the real world. In order to be able to identify molecular components, we assumed molecular species representing complexes to be replaced by as many different species as are the

* Corresponding author.
E-mail addresses: pardinig@di.unipi.it (G. Pardini), milazzo@di.unipi.it (P. Milazzo), maggiolo@di.unipi.it (A. Maggiolo-Schettini).

http://dx.doi.org/10.1016/j.tcs.2015.03.013
0304-3975/© 2015 Elsevier B.V. All rights reserved.

biological entities involved in it. For instance, if a complex C is obtained by the binding of two different proteins A and B, we assumed C to be replaced by C_A and C_B, where C_A is the part of C representing the bound form of protein A, and C_B is the part of C representing the bound form of protein B.

According to that view, and assuming the mass conservation principle for all the reactions, the reactions of a given pathway can be rewritten in a *normal form* with as many reactants as products. For example, reaction $A, B \to C$ describing the formation of complex C can be rewritten as $A, B \to C_A, C_B$. Moreover, in each reaction obtained we can establish a positional correspondence between reactants and products such that the reactant in the i-th position must be the same biological entity of the product in the same position (as it happens with A and C_A, and with B and C_B). The algorithm therefore performs a syntactical transformation of the input pathway into a normal-form pathway. Note that the mass conservation principle is needed for the successful application of our algorithm, otherwise if components could appear or disappear in between a reaction then such a reaction could not be brought into normal form. This happens, for example, in case of synthesis reactions of the form $\emptyset \to A$ and degradation reactions of the form $A \to \emptyset$.

Once all reactions in a pathway are in normal form, it is rather easy to identify molecular components. Such components essentially consist in sets of names of molecular species mentioned in the pathway. Each element of a molecular component (a name) represents a different form of the same biological entity represented by the component. As a consequence, different components do not share any element and all of the names mentioned in the pathway belong to some component. In other words, the set of components of a pathway is a partition of the set of names of biochemical species mentioned in the normal-form pathway.

In this paper we propose an algorithm for transforming pathways into their corresponding normal forms with molecular components specified. A preliminary definition of the algorithm appeared in [4]. The proposed normalization algorithm is not completely automatic, since in general there might be situations in which ambiguities on how to normalize some reactions cannot be avoided. The algorithm is designed to invoke human intervention when one of these ambiguities occurs. When a human intervention is necessary, the algorithm asks the human (a domain expert) to normalize a single reaction. There might be cases in which human intervention is invoked more than once to solve different ambiguities in the same pathway. However, we expect that in most practical cases the need of human intervention will be very limited, if not absent. This belief is supported by the extensive tests carried out on a number of SBML models from the BioModels database [5,6].

Once the molecular components are identified it is rather easy to perform syntactic transformations aimed at simplifying the visualization of the pathway itself by focusing on a subset of components (namely, on a subset of the biological entities involved). The identification of molecular components also allows formal descriptions of the pathway by means of automata or process algebra (such as those in [7–13]) to be automatically generated. This enables the application of formal methods such as model checking [14,15] and bisimulation [16] to analyze and compare behaviors of pathways.

As an application, we consider the well-known EGF signaling pathway. In particular, we consider the computational model developed by Schoeberl et al. in [17]. On this model we show that our algorithm (implemented in a Java prototype tool) can be successfully applied to infer molecular components from the reactions constituting the pathway. In this case no human intervention is needed, and the molecular components identified by the algorithm correspond to the actual biological entities involved in the pathway. We also show how to consider component-based subpathways and how to automatically generate a set of finite state automata, one for each molecular component.

This paper is an extended and revised version of [4]. With respect to the previous version, the definition of the algorithm has been updated to deal with a corner case which was not handled in the previous version. Moreover, a section investigating the theoretical properties of the algorithm has been added, in which we formally prove the correctness of the algorithm, its termination under any input pathway and a (weak) confluence property.

Structure of the paper The paper is organized as follows. In Section 2, we present the notation used for describing reactions and pathways, we formally define the concept of components, and present the problem of component identification in pathways. In Section 3 we present the algorithm for inferring components from the interactions among species described by reactions, then in Section 4 we discuss the results of application of the algorithm to a version of the well-known EGF signalling pathway from the literature. In Section 5, the algorithm is studied from the theoretical point of view, in order to prove its correctness. Finally, in Section 6 we compare our approach to related works from the literature, and in Section 7 we draw the conclusions of the paper and discuss possible future work.

2. Modelling notation for biochemical pathways

Pathways are networks of biochemical reactions occurring within a cell. Reactions can be influenced by catalysts and inhibitors, which are molecules (proteins) which can stimulate and block the occurrence of reactions, respectively. We follow the modelling notation for biochemical pathways introduced in [1]. Let us consider a set of species Σ, with metavariables $a, b, c, \ldots \in \Sigma$ and $x, y \in \Sigma$, possibly endowed with subscripts. A *reaction* R has the form

$$R : a_1, \ldots, a_n \to b_1, \ldots, b_m \quad \{c_1, \ldots, c_{w_1}\}, \{d_1, \ldots, d_{w_2}\}$$

where all a_i, b_i, c_i, d_i are in Σ. Intuitively, $a_1, \ldots, a_n$ denote the reactants, $b_1, \ldots, b_m$ denote the products, $c_1, \ldots, c_{w_1}$ denote the catalysts, and $d_1, \ldots, d_{w_2}$ denote the inhibitors.