Application of a Semi-automatic Algorithm for Identification of Molecular Components in SBML Models

Andrea Maggiolo-Schettini
Dipartimento di Informatica
Università di Pisa, Italy
maggiolo@di.unipi.it

Paolo Milazzo
Dipartimento di Informatica
Università di Pisa, Italy
milazzo@di.unipi.it

Giovanni Pardini
Dipartimento di Informatica
Università di Pisa, Italy
pardinig@di.unipi.it

Reactions forming a pathway can be rewritten by making explicit the different molecular components involved in them. A molecular component represents a biological entity (e.g. a protein) in all its states (free, bound, degraded, etc.). In this paper we show the application of a component identification algorithm to a number of real-world models to experimentally validate the approach. Components identification allows subpathways to be computed to better understand the pathway functioning.

1 Introduction

A typical aspect of a biochemical pathway is that the process described involves mainly a few chains of reactions, in which the occurrence of a reaction produces some intermediate molecule which is then transformed subsequently by other reactions. Just a few basic biological entities are often involved in a pathway, for example a simple protein undergoes a series of transformations, starting from its initial synthesized form, which can then be activated and also become part of different complexes. Therefore, intermediate molecular species can actually be seen as different states of the same initial biological entities. In accordance with this view, in the modelling of biochemical pathways we can consider a notion of *molecular component* [9, 8] that is the formal counterpart of the notion of biological entity. A molecular components will hence be associated with the set of species mentioned in the model and corresponding to different states of the same biological entity.

This paper discusses the application of a semi-automatic algorithm for the identification of molecular components in pathways which is presented in [18]. The algorithm infers components from the interactions of molecular species. In order to make components involved in each reaction explicit, the algorithm replaces each compound species (which is usually identified by a distinct name) with a fresh name for each one of its components, and transforms reactions into a normal form in which each complex is represented by as many species as are the biological entities involved in it. The algorithm is actually semi-automatic, since there are cases in which the components involved in a reaction cannot be univocally determined from the context; these cases must be solved with the help of the user.

In this paper, we present the results of the application of such an algorithm to a number of real-world pathways from the BioModels repository [12], an online database of machine-readable models of biological processes formalized in the well-known Systems Biology Markup Language (SBML). In particular, we have analyzed all of the models in the curated section, which contains only models from the literature, for a total of 436 models. In this paper, some of the models are also discussed individually in order to experimentally assess the validity of our approach when applied to real-world models.

The identification of components by the transformation of the pathway reactions into normal form allows different kinds of analysis to be performed. For example, syntactic transformations, such as the projection of the pathway over a subset of components, would allow the user to obtain insights on

A. Graudenzi, G. Caravagna, G. Mauri, M. Antoniotti (Eds.):
Wivace 2013 - Italian Workshop on
Artificial Life and Evolutionary Computation
EPTCS 130, 2013, pp. 43–52, doi:10.4204/EPTCS.130.7

the functioning of the pathway. In particular, by focusing on the reactions in which they are involved, the mechanisms underlying the pathway dynamics could be more easily identified. Moreover, from a theoretical point of view, the normalization would allow the automatic translation of a pathway into a set of automata or terms of process calculi [19, 1, 2, 5, 7], thus enabling the use of any tool available for their analysis.

2 Algorithm for Component Identification

In this section we recall the algorithm for component identification presented in [18]. The algorithm is based on the idea that a species is a "state" of a more abstract biological entity, and a reaction is a synchronized state change of a set of such entities. The algorithm assumes that biological entities involved in a reaction cannot appear from nothing and/or disappear (i.e. degradations should be modelled by using a species representing the degraded state of the biological entity). As a consequence, every entity involved in a reaction should be mentioned both among reactants and among products, and hence in a reaction there should be as many reactants as products.

Let a *component* be a set of species representing all possibile states of a given biological entity. The component identification algorithm transforms a given pathway into a *normal-form* pathway, namely such that, for all its reactions, there are as many reactants as products and there is a one-to-one positional correspondence between reactants and products which are part of the same component. The fundamental operation of the algorithm is to split a species into new subspecies, denoted by newly-introduced symbols. A split of a species is performed any time the algorithm can infer that, in a reaction, the occurrence of such a species must actually denote a complex composed of multiple bound molecules. For example, this is the case of a reaction $A,B \to C$, where C can be split into subspecies C_A,C_B, obtaining the normal-form reaction $A,B \to C_A,C_B$, with components $\{A,C_A\}$ and $\{B,C_B\}$. In general, this process may not be completely automatic since some ambiguities may arise. For example, a reaction such as $A,B \to C,D,E$ is ambiguous since it is not clear which of the two reactants has to be split into two in order to obtain as many reactants as products.

Another important aspect is that, in principle, the number of components of a pathway is not univocally determined. For example, it is possible to split the reactants/products of a reaction $A \to B$ into $A_1,\ldots,A_n \to B_1,\ldots,B_n$, thus identifying any number of components, one for each pair A_i,B_i. However, the idea is to identify only those components which can be inferred from the context, namely from the reactions in which a component is involved. For this reason, the algorithm performs a split only if it is needed to match the components of reactants and products of a reaction.

Structure of the algorithm Let us denote (in an abstract manner) by $comp(S)$ the component in which a species S occurs at a given step. Initially, the algorithm assumes that each species occurring in the input pathway is part of a different component. It then iteratively performs two alternating phases, until a normal-form pathway is obtained. In the first phase, it tries to transform the set of reactions into reactions having the same number of reactants and products. This entails inspecting each reaction in turn, in order to both (i) split species into subspecies as needed, and (ii) refine the information about components by collapsing different components into single ones, according to what can be inferred from reactions.

In case of ambiguities, the algorithm may fail to generate a normal-form pathway since either the number of reactants and products differ, or the correspondence between reactants and products is not completely specified for all the reactions. In such a case, the algorithm performs a second phase, which demands user intervention to resolve a single ambiguous reaction. Then the two phases are repeated to propagate the new information derivable from the resolved reaction to the others, as it may be useful to

resolve other ambiguous reactions, and possibly all of them.

As regards the first phase, the algorithm examines each reaction singularly. Let us consider a reaction having reactants $R = \{A_1, \ldots, A_\gamma\}$ and products $P = \{B_1, \ldots, B_\delta\}$. Moreover, assume that the components of each of the first $k < \gamma, \delta$ pairs of reactant/product match, namely $\forall i \in \{1, \ldots, k\}$. $comp(A_i) = comp(B_i)$. Let $n = \gamma - k$ and $m = \delta - k$ denote the number of remaining unmatched reactants/products of the reaction, respectively. The algorithm then distinguishes five kinds of reactions, as follows:

- Case $n = 0$, $m = 0$: the reaction is resolved, with complete correspondence between each reactant and product. This case is not necessarily definitive, since the algorithm, due to some split, may subsequently modify the reaction by replacing a species occurring in it with new symbols.
- Case $n = 1$, $m = 1$: since there is just one species remaining in each side, this means that both of them are part of the same component. Thus, the components of the reactant and product are joint into a single component by taking their union.
- Case $(n = 0 \wedge m > 0)$ or $(n > 0 \wedge m = 0)$: if either of reactants or products are empty, there is at least a component which appears in either side of the reaction which does not occur in the other side. In other words, such a component either appears or disappears in between the reaction, which is not allowed by our approach. Thus this case is regarded as an *error*, which cannot be resolved.
- Case $n = 1 \wedge m > 1$ (or $n > 1 \wedge m = 1$): in this case there is a single reactant which is actually formed from m subspecies. Thus, m new symbols A_γ-$B_k, \ldots, A_\gamma$-B_δ are introduced, by combining the name of A_γ with the B_i's. Then A_γ is replaced by the new symbols in all the reactions, and the components of each pair of species $(A_\gamma\text{-}B_k, B_k), \ldots, (A_\gamma\text{-}B_\delta, B_\delta)$ are matched, i.e. $\forall k \leq j \leq \delta$. $comp(A_\gamma\text{-}B_j) = comp(B_j)$. (The converse case, with one product and n reactants is analogous.)
- Case $n > 1$, $m > 1$: if there are multiple unmatched reactants and products, the reaction is currently *ambiguous*. This means that the current information regarding components does not allow the algorithm to decide which species must be split, thus this reaction is skipped.

The first phase continues until there are only resolved, erroneous and ambiguous reactions. In case of ambiguous reactions remaining, the algorithm asks the user to resolve one reaction by specifying the species to split and rewriting it in normal-form. The procedure is repeated from the first phase in order to propagate the updated information about components to the other reactions.

As an example, let us consider pathway $P = \{\mathbf{r_1} : Lig, rcpt \to C_1;\ \mathbf{r_2} : GDP, G_\alpha \to C_2;\ \mathbf{r_3} : GTP, G_\alpha \to C_3;\ \mathbf{r_4} : C_3 \to C_2;\ \mathbf{r_5} : C_2, G_{\beta\gamma} \to C_4;\ \mathbf{r_6} : C_4, C_1 \to C_5\}$, which models a small fragment of the well-known G protein signalling pathway, in which a ligand *Lig* (representing the signal) binds the receptor *rcpt* on cellular membrane, triggering the internal process. The algorithm identifies 5 components (representative species of which are *Lig*, *rcpt*, G_α, $G_{\beta\gamma}$ and *GDP*) and produces the follwing normal-form pathway (where species names are abbreviated):

$$\begin{aligned} P' = \{&\mathbf{r'_1} : Lig, rcpt \to C_1\text{-}Lig, C_1\text{-}rcpt; \quad \mathbf{r'_2} : GDP, G_\alpha \to C_2\text{-}GDP, C_2\text{-}G_\alpha; \quad \mathbf{r'_3} : GTP, G_\alpha \to C_3\text{-}GTP, C_3\text{-}G_\alpha; \\ &\mathbf{r'_4} : C_3\text{-}GTP, C_3\text{-}G_\alpha \to C_2\text{-}GDP, C_2\text{-}G_\alpha; \quad \mathbf{r'_5} : C_2\text{-}GDP, C_2\text{-}G_\alpha, G_{\beta\gamma} \to C_4\text{-}GDP, C_4\text{-}G_\alpha, C_4\text{-}G_{\beta\gamma}; \\ &\mathbf{r'_6} : C_4\text{-}GDP, C_4\text{-}G_\alpha, C_4\text{-}G_{\beta\gamma}, C_1\text{-}Lig, C_1\text{-}rcpt \to C_5\text{-}GDP, C_5\text{-}G_\alpha, C_5\text{-}G_{\beta\gamma}, C_5\text{-}Lig, C_5\text{-}rcpt\}. \end{aligned}$$

The sets of species forming each component are as follows: $\{Lig, C_1\text{-}Lig, C_5\text{-}Lig\}, \{rcpt, C_1\text{-}rcpt, C_5\text{-}rcpt\}$, $\{G_\alpha, C_2\text{-}G_\alpha, C_4\text{-}G_\alpha, C_5\text{-}G_\alpha, C_5\text{-}G_\alpha, C_3\text{-}G_\alpha\}, \{G_{\beta\gamma}, C_4\text{-}G_{\beta\gamma}, C_5\text{-}G_{\beta\gamma}\}, \{GDP, C_2\text{-}GDP, C_4\text{-}GDP, C_5\text{-}GDP, GTP,$ $C_3\text{-}GTP\}$. Now, we can compute subpathways describing the activity of a subset of the molecular components. For instance, by assuming that we are not interested in the role of *Lig*, *rcpt* and *GDP*, we can compute the subpathway dealing only with components G_α and $G_{\beta\gamma}$ as follows: $P'' = \{\mathbf{r''_2} : G_\alpha \to C_2\text{-}G_\alpha;\ \mathbf{r''_3} :$ $G_\alpha \to C_3\text{-}G_\alpha; \mathbf{r''_4} : C_3\text{-}G_\alpha \to C_2\text{-}G_\alpha;\ \mathbf{r''_5} : C_2\text{-}G_\alpha, G_{\beta\gamma} \to C_4G_\alpha, C_4\text{-}G_{\beta\gamma}; \mathbf{r''_6} : C_4\text{-}G_\alpha, C_4\text{-}G_{\beta\gamma} \to C_5\text{-}C_2\text{-}G_\alpha, C_5\text{-}G_{\beta\gamma}\}$. Moreover, the subpathways describing each component individually can be trivially translated into a finite state automaton or into a process algebra term to enable the application of formal analysis tools.

3 Applications

In order to test our component identification algorithm on a relevant number of real pathways we downloaded all of the pathway descriptions available in the BioModels database [12] under the category "curated models". So, our testbed consisted of 436 different SBML models of pathways.

SBML [10] is a well-established XML-based language for pathway description. A SBML pathway model includes (among others) the following elements:

- **Species:** Proteins, genes, ions and other molecules that can participate in reactions.
- **Compartments:** Well-stirred containers in which species can be located. A SBML model may contain multiple compartments (at least one) and each species must be located in one of them.
- **Reactions:** Statements describing transformation, transport or binding processes that can change the amount of one or more species. A reaction consists of reactants, products and modifiers. Moreover, the kinetic law of a reaction can be expressed by using arbitrary mathematical functions.
- **Rules:** Mathematical expressions describing how some variable values (e.g. species amounts) can be calculated from other variables, or used to define the rate of change of variables. Rules in a model can be used either together with (or in place of) reactions to determine the model dynamics.
- **Events:** A statement describing an instantaneous, discontinuous change in a set of variables when a triggering condition is satisfied.

We developed a simple translator of SBML models into CSV files accepted by our implementation of the algorithm. The translator considers only the reactions of a SBML model and transforms them into the format expected by the algorithm. SBML models in which the dynamics is governed only by rules, and not by reactions, are translated into empty CSV files, and hence are unusable by our tool. Moreover, we excluded from our test models containing reactions with fractional stoichiometry, that are meaningful only when the dynamics is described by means of ODEs. In the considered testbed, unusable models turn out to be 59 out of 436 and excluded models are 23 out of 436, hence we have 354 usable CSV files.

The 354 usable models consisted on average of 22.07 species and 31.05 reactions, with 108 models dealing with more than 20 species, 31 dealing with more than 50 species, and 6 dealing with more than 100 species. We executed our algorithm batch on all of the models. The execution (on a standard laptop) takes a few seconds for each model. The execution on a model can terminate either successfully with an automatically generated normal-form pathway (OK), or with an error message if an erroroneous reaction is encountered, or with a human intervention request if an ambiguous reaction is encountered. In the latter two cases, the execution is interrupted and the number of erroneous and ambiguous reactions, respectively, is printed. The results we obtained on the 354 usable models are shown in the table below, on line "Batch execution". Most of the models (244 out of 354) encountered some errors. By inspecting some of them we discovered that in most cases the error was due to reactions describing either synthesis or degradation of some species, in which either products or reactants were absent, respectively.

	Ok	Erroneous	Ambiguous
Batch execution	96	244	14
Preprocessing & Batch execution	241	93	20
Preprocessing & Batch execution with dynamic correction	318	0	36

We decided to solve the problem of synthesis/degradation reactions by preprocessing the SBML models. The preprocessing was performed on the CSV translation and consisted in inserting a dummy species in each empty set of reactants or products encountered. Note that every time an empty set is encountered by the preprocessor, a new fresh species is generated and used as dummy. Hence, each dummy species will appear only once in the pathway. Since dummy species added during preprocessing do not interact with other species, error situations are solved without affecting component identification.

The results of executions of the algorithm after preprocessing are in the table, on line "Preprocessing & Batch execution". Most of the previously erroneous situations can now be handled by the algorithm, and in the vast majority of the cases they did not encountered ambiguous reactions (151 errors are solved and only 6 of them turn out to need human intervention).

By inspecting some of the models still giving an erroneous result we discovered that in some cases the error was caused by a reaction describing the degradation of a part of a complex. For example, let us consider a pathway consisting of reactions $A, B \to C$ and $C \to A$. The first reaction is the formation of a complex composed by A and B, and the second is the degradation of B when it bound to A. In this case the algorithm transforms the first reaction into $A, B \to C_A, C_B$. Consequently, in the second reaction it replaces C by C_A, C_B obtaining $C_A, C_B \to A$. Now, in the second reaction the algorithm associates C_A with A (as this follows from the first reaction) and it has nothing to associate with C_B (error situation).

To solve this second kind of errors we choose to add fresh dummy objects also at runtime when errors are encountered. In the example, the second reaction could be corrected by adding a fresh dummy object D_B as follows: $C_A, C_B \to A, D_B$. This is not always correct since the dynamic insertion of dummy species may lead to different components being identified, depending on the order of processing of reactions. The results of executions of the algorithm after preprocessing and with insertion of dummy objects at runtime are in the table, on line "Preprocessing & Batch execution with dynamic correction". All of the error situations are now solved and only 36 of the 354 models turn out to need human intervention.

The component identification algorithm completed automatically its execution in the 89.83% of the cases. In the remaining 36 cases we had an average of 4.67 ambiguous reactions, that means that at most 4.67 questions are asked to the user in an average execution of a model with ambiguous reaction (with a maximum of 16 and only two models over 10). The average computed over all of the 354 cases is 0.53.

The algorithm, although semi-automatic, turns out to require very limited human intervention in practice. Let us now assess the quality of the computed results by comparing the molecular components identified by the algorithm with the biological entities the modelled pathways deal with according to the referenced literature. We consider eight "randomly" chosen SBML models (numbers 50, 100, 150, 200, 250, 300, 350 and 400 in the BioModels database), three randomly chosen big models (numbers 88, 235 and 293) and three randomly chosen models with ambiguities (numbers 82, 143 and 165). In the following paragraphs a brief summary of the analysis of each of these SBML models is given.

Analysis of model 50 (BIOMD0000000050.xml). This SBML model describes a kinetic model of N-(1-deoxy-D-fructos-1-yl)-glycine (DFG) thermal decomposition [14]. The SBML model includes 14 species and 16 reactions. During its execution the algorithm encountered one erroneous reaction, that was solved by dynamically inserting a dummy object in it. In the model DFG can be degraded into several different ways obtaining a number of different substances. The model includes some species representing unidentified intermediate components (E_1 and E_2) and unidentified carbohydrate fragments (C_n). Most of the species involved in the pathway can be transformed into these unidentified molecules, thus causing all of these species to be included in the same molecular component by the algorithm. As a consequence, the algorithm identifies only 3 components: two including all and only the species related with Glycine and methyulglyoxal, respectively, and one including all of the other species. The quality of the result in this case are hence only partially satisfactory, and the cause of unsatisfaction is that the model includes ambiguities (unidentified species).

Analysis of model 82 (BIOMD0000000082.xml). This SBML model describes the formation of an inhibitor of the Adenylate Cyclase enzyme [24]. The SBML model includes 10 species and 6 reactions. During its execution the algorithm encountered no erroneous reactions, but two ambiguous reactions. The two ambiguous reactions are the following: $DR, G\_GDP \to DRG\_GDP$ and $DRG\_GDP \to GDP, DRG$.

These reactions describe the passage (in two steps) of a G-protein from *GDP* to *DR*. This is a typical case of ambiguity (see [18]) that can be solved by making it explicit the involvement of the "hidden" component representing the *G* protein, namely by replacing the ambiguous reactions with the following ones: $DR, G_{G\_GDP}, GDP_{G\_GDP} \rightarrow DRG\_GDP$ and $DRG\_GDP \rightarrow GDP, G_{DRG}, DR_{DRG}$. So, in this case the algorithm needs human intervention, after which 4 components are correctly identified.

Analysis of model 88 (BIOMD0000000088.xml). This model describes the Rho-kinase pathway in order to study thrombin-dependent in vivo transient responses of Rho activation and Ca^{2+} increase [13]. It includes 104 species and 110 reactions, and it is one of the biggest models we have considered. During its execution the algorithm encountered no erroneous and ambiguous reactions. The algorithm identified 28 components. Species in the SBML file are represented by numbers, hence checking the correctness of the identified components was not trivial. We checked a few randomly chosen components and each of them turned out to be composed of species representing different states of the same molecule. Although we do not have a complete proof of the correctness of the result, we consider this case satisfactory.

Analysis of model 100 (BIOMD0000000100.xml). This SBML model is used to study the effects of cytosolic calcium oscillations on activation of glycogen phosphorylase [21]. The SBML model includes 5 species and 10 reactions. During its execution the algorithm encountered five errors, that were solved by preprocessing. The species involved in the model are: external calcium (EC), cytosolic calcium (Z), intravescicular calcium (Y), inositol 1,4,5-trisphosphate IP_3 (A), and glycogen phosphorlylase (GP). The reactions described in the model are: calcium influx, transportation of calcium between compartments, and IP_3 and glycogen phosphorylase syntheses and degradations. The algorithm identifies 3 components: one including all of the species representing calcium (EC, Z and Y) and the other two including IP_3 and glycogen, respectively. The quality of the result in this case is hence satisfactory.

Analysis of model 143 (BIOMD0000000143.xml). This SBML model describes the metabolism of activated neutrophils in which oscillatory behaviours have been observed [17]. The model includes 20 species and 20 reactions. During its execution the algorithm encountered three errors, solved by preprocessing. The algorithm encountered also 9 ambiguous reactions. After inspecting the model we discovered that thee reactions in particular were problematic, the solution of which solved also the ambiguities in the other 6 reactions. The three problematic reactions describe the *Peroxidase Cycle* in which an enzyme is activated by a H_2O_2 molecule and then transforms in two steps two Melatonin molecules into Melatonin-free-radical. During this cycle two molecules of water are released, and also some hydrogen ions are involved. Water and hydrogen are not mentioned in the model, and this creates ambiguities in the reactions. Such ambiguities can be solved by a human intervention aimed at clarifying the three problematic reactions. After this, the algorithm completes by correctly identifiying 5 components.

Analysis of model 150 (BIOMD0000000150.xml). This is a very simple SBML model describing formation and activation of Cdk/Cyclin Complexes [15]. The model includes 4 species and 4 reactions. The algorithm encountered no erroneous reactions. The modelled reactions are trivial (complex formation from two species, activation of the complex and inverse reactions). The algorithm correctly identifies 2 components, one for each molecule involved in the complex. The result in this case is hence satisfactory.

Analysis of model 165 (BIOMD0000000165.xml). This model deals with intracellular signalling through cAMP and its cAMP-dependent protein kinase (PKA) [22]. The SBML model includes 37 species and 30 reactions. During its execution the algorithm encountered three erroneous reactions, two of which solved by preprocessing and one by dynamically inserting a dummy object. The model included one ambiguous reaction similar to those encountered in the case of model 82. The reaction can be disambiguated by means of a single human intervention after which 16 components are correctly identified.

Analysis of model 200 (BIOMD0000000200.xml). This SBML model describes binding reactions leading to a transmembrane receptor-linked multiprotein complex involved in bacterial chemotaxis [4].

The SBML model includes 21 species and 34 reactions. During its execution the algorithm encountered no erroneous reactions. The biochemical entities involved in the pathway are the dimeric aspartate-binding receptor Tar (TT) and four cytoplasmic proteins CheW (W), CheA (AA), CheY (Y), CheB (B) and CheZ (Z). TT, W and AA are involved in a number of reactions leading to the formation of a complex TTWWAA that activates proteins Y, B, Z in different ways . The algorithm identifies exactly 6 components out of the 21 species, and such components correspond exactly to the 6 biochemical entities involved in the pathway. The quality of the result in this case is hence satisfactory.

Analysis of model 235 (BIOMD0000000235.xml). This SBML model describes the sea urchin endomesoderm network [11], a very big gene regulation network. The model includes 618 species and 778 reactions. During its execution the algorithm encountered only 3 erroneous reactions, solved by preprocessing. The algorithm identified only 47 components. By inspecting them we discovered that actually one of the components included the vast majority of the species. By inspecting reactions we discovered that actually they consisted mostly of syntheses and degradations in which species "none" was used in all the case of empty reactants or products. The algorithm associated all of the species involved in a synthesis or degradation reaction in the same component. This was due to the fact that all of such species are in relation with the same species "none". We modified the model by replacing the unique "none" by fresh dummy species (as in the case of preprocessing). After this change the algorithm identified 406 components. We checked some of them (randomly chosen) and they turned out to be correct representations of biochemical entities involved in the network. We consider hence the quality of the result satisfactory, although this model needed to be slightly modified in order to let the algorithm work as expected.

Analysis of model 250 (BIOMD0000000250.xml). This SBML model is used to study how epidermial growth factor (EGF) and heregulin (HRG) generate distinct responses of the transcription factor c-fos [16]. The SBML model includes 49 species and 78 reactions. During its execution the algorithm encountered 19 erroneous reactions solved by preprocessing. The algorithm identifies 18 components that, after an analysis of the description of the pathway in the paper, seem to correspond to the biochemical entities involved in the pathway. The quality of the result in this case is hence satisfactory.

Analysis of model 293 (BIOMD0000000293.xml). This SBML model describes an ubiquitin-proteasome system [20]. The SBML model includes 136 species and 316 reactions. During its execution the algorithm encountered 114 erroneous reactions, all solved by dynamic insertion of dummy objects. The algorithm identified only 12 components. By inspecting the model we discoverd that is suffered from the same problem of model 235, namely the same dummy species "source" and "AggP_Proteasome" were used in many syntheses and inhibition reactions, causing the most of the species to be assigned to the same component by the algorithm. We modified this model as we did for model 235. The number of erroneous reactions encounterd decreased to 77, and the number of components identified by the algorithm decreased to 16. Components seems to be rather correct, since they seems to represent a reasonable partition of the species set. However, the complexity of the model and the high number of erroneous reactions solved only at runtime does not allow us to be sure about the correctness of the result. We leave the assessment of the quality of this result as a future work.

Analysis of model 300 (BIOMD0000000300.xml). This SBML model is used to study how the activity of the heterodimeric transcription factor hypoxia inducible factor (HIF) if affected by the interaction of factors inhibiting HIF (FIH) with ankyrin-repeat domain (ARD) proteins [23]. The SBML model includes 9 species and 10 reactions. Reactions are all degradations and syntheses of species. All of the 10 reactions become erroneous reactions that are solved by preprocessing. The algorithm identifies 9 components. However, by inspecting the models it emerged that most of the dynamics is described by means of SBML rules, hence the models turns out to be unusable.

Analysis of model 350 (BIOMD0000000350.xml). This SBML model is used to study the circadian

network of higher plants. In particular, it is a model of the clock of the picoeukaryotic alga Ostreococcus tauri as a feedback loop between the genes TOC1 and CCA1 [25]. The SBML model includes 14 species and 30 reactions. Actually, most reactions are syntheses and degradations. Indeed, 27 reactions out of 30 become erroneous reactions that are solved by preprocessing. Only three reactions describe transformation of species, and consequently the algorithm correctly identifies 11 components. The quality of the result in this case is hence satisfactory.

Analysis of model 400 (BIOMD0000000400.xml). This SBML model belongs to the set of unusable models since it does not include any reaction.

4 Discussion

The component identification algorithm we proposed in [18] turned out to work pretty well. In the majority of the cases the execution of the algorithm has been completely automatic, and the computed molecular components correctly represented the biological entities involved in the pathway. In the cases in which human intervention was necessary, it usally consisted in answering very few questions on how resolve ambiguous reactions. After the detailed analysis of some of the models (numbers 235 and 293) a recurrent "error" in the modelling of pathways (from the viewpoint of the algorithm) is to use the same special species to represent the pre-synthesis or the degraded form of many different biological entities. This causes different molecular components to be erroneously merged into one. Moreover, it also emerged that ambiguities often are of the same kind, namely they are include some "hidden" species that is always bound to some other species. In these cases the algorithm cannot identify the component corresponding to such species and the reactions turn out to be ambiguous.

Most of the erroneous and ambiguous situations could be prevented by a more accurate construction of models. However, it could be an interesting further development of our work the definition of some rule or heuristics able to solve these situations when encountered. Otherwise, the prevention of erroneous and ambiguous situations could be approached by developing of a model repair preprocessing routine based for instance on static checking of conservation laws or P-invariants [6].

References

[1] Roberto Barbuti, Giulio Caravagna, Andrea Maggiolo-Schettini, Paolo Milazzo & Giovanni Pardini (2008): *The calculus of looping sequences*. In: *Formal Methods for Computational Systems Biology*, Springer, pp. 387–423. doi:10.1007/978-3-540-68894-5_11

[2] Roberto Barbuti, Andrea Maggiolo-Schettini, Paolo Milazzo & Angelo Troina (2006): *A calculus of looping sequences for modelling microbiological systems*. *Fundamenta Informaticae* 72(1), pp. 21–35.

[3] Michael L Blinov, James R Faeder, Byron Goldstein & William S Hlavacek (2004): *BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains*. *Bioinformatics* 20(17), pp. 3289–3291. doi:10.1093/bioinformatics/bth378

[4] D. Bray & R.B. Bourret (1995): *Computer analysis of the binding reactions leading to a transmembrane receptor-linked multiprotein complex involved in bacterial chemotaxis.* *Molecular biology of the cell* 6(10), pp. 1367–1380.

[5] Federica Ciocchetta & Jane Hillston (2009): *Bio-PEPA: A framework for the modelling and analysis of biological systems*. *Theoretical Computer Science* 410(33), pp. 3065–3084. doi:10.1016/j.tcs.2009.02.037

[6] Allan Clark, Vashti Galpin, Stephen Gilmore, Maria Luisa Guerriero & Jane Hillston (2012): *Formal methods for checking the consistency of biological models*. In: *Advances in Systems Biology*, Springer, pp. 461–475. doi:10.1007/978-1-4419-7210-1_27

[7] Vincent Danos & Cosimo Laneve (2004): *Formal molecular biology*. Theoretical Computer Science 325(1), pp. 69–110. doi:10.1016/j.tcs.2004.03.065

[8] Peter Drábik, Andrea Maggiolo-Schettini & Paolo Milazzo (2012): *On conditions for modular verification in systems of synchronising components*. Fundamenta Informaticae 120(3), pp. 259–274.

[9] Peter Drábik, Andrea Maggiolo-Schettini & Paolo Milazzo (2012): *Towards modular verification of pathways: fairness and assumptions*. In: Proceedings of MeCBIC 2012, *EPTCS* 100, pp. 63–81.

[10] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano & the rest of the SBML Forum (2003): *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models*. Bioinformatics 19(4), pp. 524–531. doi:10.1093/bioinformatics/btg015

[11] C. Kühn, C. Wierling, A. Kühn, E. Klipp, G. Panopoulou, H. Lehrach & A. Poustka (2009): *Monte carlo analysis of an ode model of the sea urchin endomesoderm network*. BMC systems biology 3(1), p. 83. doi:10.1186/1752-0509-3-83

[12] C. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M.I. Stefan, J.L. Snoep, M. Hucka, N. Le Novère & C. Laibe (2010): *BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models.* BMC Systems Biology 4, p. 92. doi:10.1186/1752-0509-4-92

[13] A. Maeda, Y. Ozaki, S. Sivakumaran, T. Akiyama, H. Urakubo, A. Usami, M. Sato, K. Kaibuchi & S. Kuroda (2006): *Ca2+-independent phospholipase A2-dependent sustained Rho-kinase activation exhibits all-or-none response*. Genes to Cells 11(9), pp. 1071–1083. doi:10.1111/j.1365-2443.2006.01001.x

[14] S.I.F.S. Martins & M.A.J.S. Van Boekel (2003): *Kinetic modelling of Amadori N-(1-deoxy-d-fructos-1-yl)-glycine degradation pathways. Part II–Kinetic analysis*. Carbohydrate research 338(16), pp. 1665–1678. doi:10.1016/S0008-6215(03)00174-5

[15] M.C. Morris, C. Gondeau, J.A. Tainer & G. Divita (2002): *Kinetic Mechanism of Activation of the Cdk2/Cyclin A Complex*. Journal of Biological Chemistry 277(26), pp. 23847–23853. doi:10.1074/jbc.M107890200

[16] T. Nakakuki, M.R. Birtwistle, Y. Saeki, N. Yumoto, K. Ide, T. Nagashima, L. Brusch, B.A. Ogunnaike, M. Okada-Hatakeyama & B.N. Kholodenko (2010): *Ligand-specific c-Fos expression emerges from the spatiotemporal control of ErbB network dynamics*. Cell 141(5), pp. 884–896. doi:10.1016/j.cell.2010.03.054

[17] L.F. Olsen, U. Kummer, A.L. Kindzelskii & H.R. Petty (2003): *A model of the oscillatory metabolism of activated neutrophils*. Biophysical journal 84(1), pp. 69–81. doi:10.1016/S0006-3495(03)74833-4

[18] G. Pardini, P. Milazzo & A. Maggiolo-Schettini: *An Algorithm for the Identification of Components in Biochemical Pathways*. In: Proceedings of CS2Bio 2013, ENTCS. In press.[1].

[19] A. Phillips & L. Cardelli (2007): *Efficient, correct simulation of biological processes in the stochastic pi-calculus*. In: Computational Methods in Systems Biology, Springer, pp. 184–199. doi:10.1007/978-3-540-75140-3_13

[20] C.J. Proctor, P.J. Tangeman & H.C. Ardley (2010): *Modelling the Role of UCH-L1 on Protein Aggregation in Age-Related Neurodegeneration*. PLoS ONE 5(10), p. e13175. doi:10.1371/journal.pone.0013175.s013

[21] A. Rozi & Y. Jia (2003): *A theoretical study of effects of cytosolic Ca^{2+} oscillations on activation of glycogen phosphorylase*. Biophysical chemistry 106(3), pp. 193–202. doi:10.1016/S0301-4622(03)00192-3

[22] J.J. Saucerman, J. Zhang, J.C. Martin, L.X. Peng, A.E. Stenbit, R.Y. Tsien & A.D. McCulloch (2006): *Systems analysis of PKA-mediated phosphorylation gradients in live cardiac myocytes*. Proceedings of the National Academy of Sciences 103(34), pp. 12923–12928. doi:10.1073/pnas.0600137103

[23] B. Schmierer, B. Novák & C. Schofield (2010): *Hypoxia-dependent sequestration of an oxygen sensor by a widespread structural motif can shape the hypoxic response-a predictive kinetic model*. BMC systems biology 4(1), p. 139. doi:10.1186/1752-0509-4-139

[1] Pre-proceedings version available online at `http://cs2bio13.di.unito.it/papers/Pardini.pdf`

[24] W.J. Thomsen, J.A. Jacquez & R.R. Neubig (1988): *Inhibition of adenylate cyclase is mediated by the high affinity conformation of the alpha 2-adrenergic receptor.* Molecular pharmacology 34(6), pp. 814–822.

[25] C. Troein, F. Corellou, L.E. Dixon, G. van Ooijen, J.S. ONeill, F.-Y. Bouget & A.J. Millar (2011): *Multiple light inputs to a simple clock circuit allow complex biological rhythms*. The Plant Journal 66(2), pp. 375–385. doi:10.1111/j.1365-313X.2011.04489.x