Proving the absence of unbounded polymers in rule-based models

Pierre Boutillier, Jérôme Feret, Aurélie Faure de Pebeyre

Proving the absence of unbounded polymers in rule-based models

Pierre Boutillier[1]

Harvard Medical School,
Department of Systems Biology, Boston, MA 02115, USA

Aurélie Faure de Pebeyre[2]

Centre de recherche interdisciplinaire, 75004 Paris, France
INRIA,
Centre de recherche INRIA de Paris, 75 012 Paris, France
Département d'informatique de l'École normale supérieure,
École normale supérieure, CNRS, PSL Research University, 75 005 Paris, France

Jérôme Feret[3]

INRIA,
Centre de recherche INRIA de Paris, 75 012 Paris, France
Département d'informatique de l'École normale supérieure,
École normale supérieure, CNRS, PSL Research University, 75 005 Paris, France

Abstract

Rule-based languages, such as Kappa and BNGL, allow for the description of very combinatorial models of interactions between proteins. A huge (when not infinite) number of different kinds of bio-molecular compounds may arise due to proteins with multiple binding and phosphorylation sites. Knowing beforehand whether a model may involve an infinite number of different kinds of bio-molecular compounds is crucial for the modeller. On the first hand, having an infinite number of kinds of bio-molecular compounds is sometimes a hint for modelling flaws: forgetting to specify the conflicts among binding rules is a common mistake. On the second hand, it impacts the choice of the semantics for the models (among stochastic, differential, hybrid).
In this paper, we introduce a data-structure to abstract the potential unbounded polymers that may be formed in a rule-based model. This data-structure is a graph, the nodes and the edges of which are labelled with patterns. By construction, every potentially unbounded polymer is associated to at least one cycle in that graph. This data-structure has two main advantages. Firstly, as opposed to site-graphs, one can reason about cycles without enumerating them (by the means of Tarjan's algorithm for detecting strongly connected components). Secondly, this data-structures may be combined easily with information coming from additional reachability analysis: the edges that are labelled with an overlap that is proved unreachable in the model may be safely discarded.

Keywords: Rule-based modelling, Polymers, Static analysis, Strongly connected components

1 Introduction

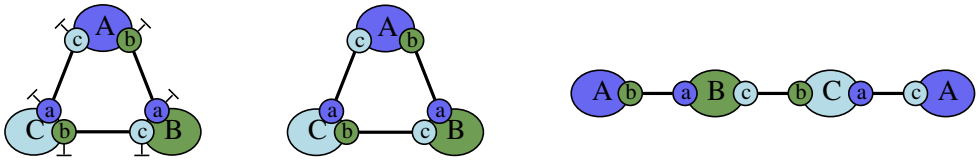
Rule-based languages, such as Kappa [8] and BNGL [2], propose a transparent way to encode models of interactions between proteins. Systems involving races for shared resources, different time- and concentration-scales, non linear feedback loops may be described by the means of rewrite rules. This allows for the description of very combinatorial models. A huge (when not infinite) number of different kinds of bio-molecular compound may arise due to the presence of scaffold and/or proteins with multiple binding and phosphorylation sites. The long term goal is then to understand how the collective behaviour of these proteins emerges from the mechanistic interactions between proteins.

Detecting whether such a model involve an infinite number of different kinds of bio-molecular compound, is important. Often, the models come from a higher level of description [14] or from automatic mining of the literature [13]. The presence of an infinite number of bio-molecular compounds is often a hint for a lack of specification. Namely, conflicts between potential bindings have not be specified enough and there is a need to refine the model. Sometimes the assembling of giant molecules is involved. In that later case, it is important to confirm that the model implements properly what the modeller has in mind. The presence of an infinite number of distinct kinds of bio-molecular compound also matters when choosing the most appropriate semantics for the models (among stochastic, differential, hybrid).

In this paper, we introduce some graph structures to abstract the potential presence of unbounded polymers in a rule-based model. These graphs either cope for the potential succession of sites along chains of proteins in the bio-molecular compounds that are reachable, or for the succession of bonds in these chains. They provide a sound and complete (with respect to the information provided by the contact map of the model) description of the potential binding between the sites of proteins. Nevertheless, the contact map encodes only non relational information: it cannot establish relationships about the different binding states of pairs of sites. To go beyond non relational information, we refine the graph of the links in bio-molecular compounds by taking into account the result of external relational static analyses [6,12,3]. Such static analyses provide a list of patterns that are known unreachable. As a result, we get a sound, but not complete approach (the detection of unreachable patterns in a rule-base language is undecidable anyway [16]) that may detect and prove that the set of non-isomorphic bio-molecular compounds of a model is finite, without executing the model.

The rest of the paper is organised as follows. Sec. 2 introduces some case studies to provide intuitions about the property that we want to infer, and to highlight the pitfalls that we will have to avoid. Sec. 3 gives some reminders about Kappa. In Sec. 4, we introduce two families of graphs and a procedure to decide whether or not the set of bio-molecular compounds that are compatible with a contact map is finite. We refine our approach to deal with black-listed patterns in Sec. 5.

[1] Email: pierre_boutillier@hms.harvard.com
[2] Email: aurelie.faure@cri-paris.org
[3] Email: jerome.feret@ens.fr



(a) Contact map. (b) Triangle ABC. (c) A repeatable pattern.

Fig. 1. The ABC example. The contact map (Fig.1(a)) provides a typing discipline. It displays every kind of protein and specifies their interfaces. The contact map also provides the potential states for each site: either free $\dashv$, or bound to another site (which is encoded as a link between pair of sites in the contact map). In Fig. 1(b) is described a bio-molecular compound that is compatible with the contact map. Every instance of proteins belongs to the contact map. Their interfaces are the same as in the contact map. Also any bond between two sites complies with one link explicitly written in the contact map. Fig. 1(c) describes a repeatable pattern. This pattern is compatible with the contact map and can be repeated in order to form arbitrarily large bio-molecular compounds.

2 Case studies

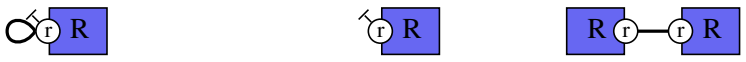
In this section, we introduce some examples to explain intuitively why there may be an unbounded number of bio-molecular compounds in a rule-based model. We also explain why naive approaches may fail in proving that the number of bio-molecular compounds is finite in a given model when it is the case, while identifying the pitfalls that shall be avoided to achieve this goal.

2.1 Elementary cycles

Let us start with a simple example. We consider a model involving three kinds of protein A, B, C. Each protein has two binding sites: the protein A has the binding sites b and c, the protein B has the binding sites a and c, and the protein C has the binding sites a and b. Each binding site may be free, or bound to another site. Only three kinds of bond are possible: the site b of an instance of the protein A may be bound to the site a of an instance of the protein B; the site c of an instance of the protein B may be bound to the site b of an instance of the protein C; and the site a of an instance of the protein C may be bound to the site c of a protein A.

These assumptions are summarised in a graph in Fig. 1(a). This graph is called the contact map of the model. It describes every kind of protein and every site in their interfaces. The potential state of each site is also indicated. In our model, every site may be free: they are all tagged with the symbol $\dashv$. Potential bonds are indicated by the means of non oriented edges between pairs of sites. The contact map provides a typing discipline. Every bio-molecular compound in our model shall satisfy the constraints that the contact map is encoding about the interface of agents, the potential states of sites, and their potential bindings. An example of bio-molecular compound that is compatible with the contact map is drawn in Fig. 1(b). This bio-molecular compound is made of three proteins A, B, and C that are bound pair-wise so as to form a triangular shape. In a bio-molecular compound, every site shall be exclusively either free, or bound to at most one other site. In general, a bio-molecular compound does not have to contain an instance of each kind of protein. Also it may contain several instances of some of them.

The contact map that is given in Fig. 1(a) is compatible with an infinite number of different (i.e. *non isomorphic*) molecular compounds. Indeed we show in Fig. 1(c), a pattern that may be repeated an unbounded number of times in order to form



(a) Contact map. (b) Exhaustive list of bio-molecular compounds.

Fig. 2. The example of a protein that may form monomers and dimers. The contact map (e.g. see Fig. 2(a)) contains a cycle, since the unique site of an instance of a protein may be linked to the unique site of another instance of another protein. However, only once instance of this cycle may occur in a given bio-molecular compound and the number of bio-molecular compound remains bounded despite this cycle (e.g. see Fig. 2(b)).

(a) Contact map. (b) A repeatable pattern.

Fig. 3. An example of a protein with two sites a and b such that the site a of a protein may be bound to the site a of another protein and the site b may be bound to the site b of another protein. The contact map (Fig.3(a)) contains two self-loops. The pattern that is made of three proteins, the first two bound via their respective sites a and the last two bound via their respective sites b (e.g. see Fig. 3(b)) is a repeatable pattern. Thus, an infinite number of bio-molecular compounds is compatible with the contact map.

arbitrary many different bio-molecular compounds. This is tempting to relate the potential presence of an arbitrary number of different bio-molecular compounds to the one of a cycle in the contact map. However we shall see in the next examples that this intuition is misleading.

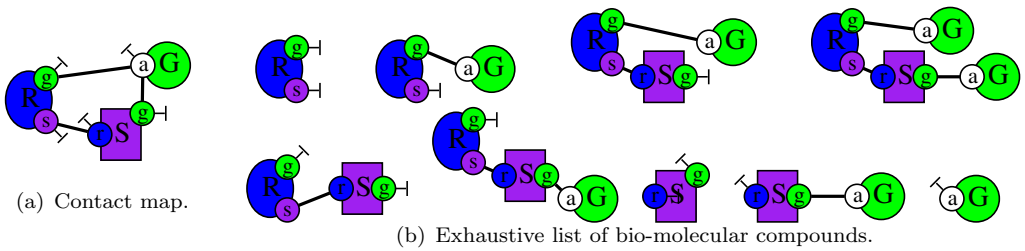
2.2 Self loops

In this example we consider a model with only one kind of protein. This protein has a single site which may be either free, or bound to the unique site of another protein of the same kind. Roughly speaking proteins may form monomers and dimers. These assumptions are encoded in the contact map that is given in Fig. 2(a). We notice a cycle in this contact map (from the unique site of the protein to itself). Yet there are exactly two kinds of bio-molecular compound that are compatible with this contact map (these bio-molecular compounds are depicted in Fig. 2(b)): there is a finite number of kinds of bio-molecular compound them despite the presence of a cycle in the contact map.

One could think that self-loops should not be considered as cycles when trying to prove the finiteness of the set of bio-molecular compounds of a model. Indeed whenever a molecular compound contains a bond that corresponds to a self-loop in the contact map, then both sites are necessarily bound together and they are no longer available to form links with other sites. Yet the contact map that is given in Fig. 3(a) shows that it is unsafe in general to discard self-loops. In this example, we consider only one kind of protein with two sites. Each site may be either free, or bound to the same site of another instance of the protein. It is then possible to form a chain of three proteins (see Fig. 3(b)) that may be repeated an arbitrary number of times in a bio-molecular compound.

2.3 Conflicting bindings

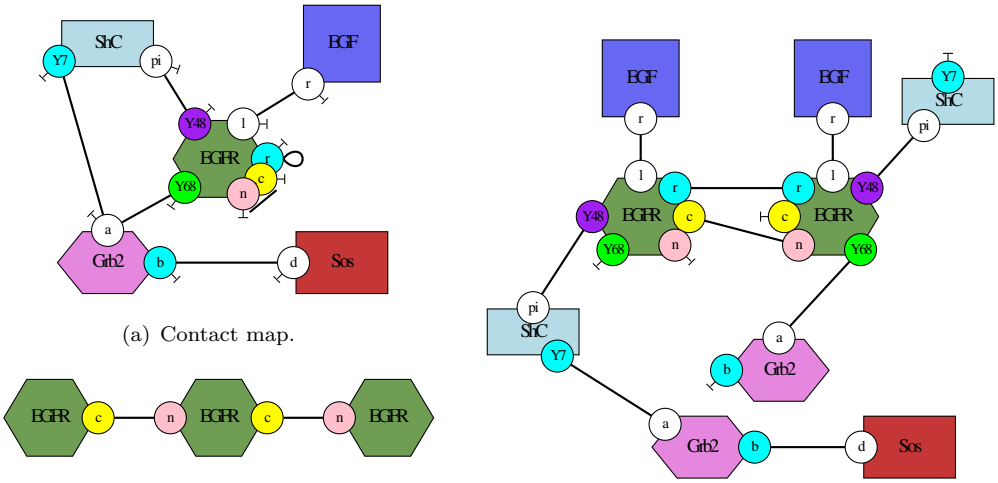
In this example, we consider three kinds of protein G, R, and S. The proteins of kind G have a single site; the proteins of kind R have two sites g and s; and the proteins of kind S have two sites g and r. Proteins R and S may bind to each-other



(a) Contact map.

(b) Exhaustive list of bio-molecular compounds.

Fig. 4. An example of a protein with a site that may be bound to two different kinds of site. As drawn in the contact map (e.g. see Fig. 4(a)), the site of the protein G may be either free, bound to the site g of the protein R, or bound to the site g of the protein S. The cycle in the contact map does not induce an infinite number of different bio-molecular compounds (e.g. see Fig. 4(b)).



(a) Contact map.

(b) A repeatable pattern.

(c) A bio-molecular compound.

Fig. 5. The example of the early events in the epidermic growth factor [1]. In Fig. 5(a) is drawn the contact map. Compared to the original model in BNGL, we have omitted phosphorylation states, since they have no impact on the binding topology. We have also added two sites in the receptor to model the asymetric bond between receptors *EGFR* in dimers. The model is constrained by the following property: whenever the site c of a receptor *EGFR* is bound, then its site r is bound as well, and both sites are bound to the same instance of protein. The contact map is compatible with the repeatable pattern that is given in Fig. 5(b). Yet this pattern does not satisfy the additional constraint. Indeed the model has only a finite set of different bio-molecular compounds. In Fig. 5(c) is given an example of a typical bio-molecular compound.

via their respective sites s and r. The unique site of proteins G may bind either to the site g of an instance of the protein R, or to the site g of an instance of the protein S. Thus, there is a competition, or a conflict, on the site of the protein G.

The contact map for this example is provided in Fig. 4(a). We notice that the competition on the site of the protein G belongs to a cycle in this contact map. Yet, in a given bio-molecular compound, the site of each instance of G is either free, or bound to at most one site. Thus the cycle of the contact map is not realisable in a concrete bio-molecular compound. In Fig. 4(b), we enumerate all the bio-molecular compounds that are compatible with the constraints encoded in the contact map. There is a finite amount of them, despite the presence of a cycle in the contact map.

2.4 Early events in the epidermic growth factor pathway

So far, we have considered only toy examples so as to try to understand which conditions on a contact map are necessary to induce only a finite number of bio-molecular compounds. In Fig. 5, we consider a model for the early events in the

integration of the epidermic growth factor (EGF) [1]. In this model, the acquisition of the protein *Sos* by the membrane of the cell is made in several steps. Firstly a pair of receptors *EGFR* on the membrane of the cell shall be activated by the ligand *EGF*. Once activated, they can form a dimer thanks to a symmetric bond via their respective sites r. Compared to the BNGL model of [1], asymmetric bonds between receptors are also considered. To stabilise dimers, pairs of receptors that are bound via their sites r form an asymmetric binding by connecting the site c of one receptor to the site n of the other receptor. The symmetric bond in a dimer cannot be released in the presence of an asymmetric one. As a consequence, whenever the site c of a receptor is bound to the site n of another receptor, these receptors are also connected by a symmetric bond. This property can be inferred by the static analysis that is described in [12,3]. Each receptor in a dimer may activate the sites *Y48* and *Y68* of the other receptor (since we focus only on the binding topology, we omit the details about these activations which are performed by the means of phosphorylation). The site *Y68* may bind to the protein *Grb2*, which may be, or not, bound to the protein *Sos*. The site *Y48* connects to the protein *Grb2* indirectly, thanks to the adapter protein *Shc*.

It is worth noticing that the contact map, that is depicted in Fig. 5(a) does not provide all the information about the model. The constraints on the sites c, n, and r emerge from some mechanisms that are described by the means of rules. Rules are omitted here so as to focus on the topology of the potential bindings between the sites of proteins. Yet some additional constraints may be provided as a list of forbidden patterns. This way, we assume that the bio-molecular compounds of our model are the ones that are compatible with the contact map and that does not contain the patterns that are black-listed.

Interestingly, the contact map of the EGF model (e.g. see 5(a)) contains both issues that we have pointed out in Sec. 2.2 and in Sect. 2.3. Indeed, the site r of a receptor may be bound to the site r of another receptor. Moreover there is a conflict on the site a of the protein *Grb2* which may be bound to the receptor directly or via an adapter protein. Another issue is raised by this model. The constraints provided by the contact map are not enough to ensure the finiteness of the set of the different bio-molecular compounds. Indeed, the pattern that is provided in Fig. 5(b) is compatible with the contact map, and could be repeated an unbounded number of times to form an infinite number of different bio-molecular compounds. Nevertheless, this pattern is not compatible with the additional constraints about symmetric and asymmetric bindings in dimers. *In fine*, there is only a finite number of different bio-molecular compounds that satisfies both the constraints from the contact map and the additional relationships among the state of the sites. In Fig. 5(c), we provide a typical example of bio-molecular compound in the EGF model. This example is made of a dimer, with one site *Y68* free, one site *Y68* connected to a *Grb2* not connected to a *Sos*, one site *Y48* connected to an adapter not connected to a *Grb2*, and a site *Y48* connected to a *Sos*. In total, a dimer may be connected to up to four instances of *Sos*.

On such a rather small model, it is possible to enumerate the different bio-molecular compounds thanks to reaction enumeration engines [2,4]. This model is made of 253 kinds of bio-molecular compound. Taking into account phosphorylation

states would lead to a model with 932 kinds of bio-molecular compound. Nevertheless, enumeration engines do not scale to large combinatorial networks such as the longer version of the EGF model (including the interactions with the proteins *Ras*, *Erk*, and *Mapk*) that is described in [5] and that involves about 10^{19} different kinds of bio-molecular compound [6] or as the model of the interactions found in the cytoplasmic portion of the Structural Interaction Network (cSIN) [9,15] that involves an infinite number of bio-molecular compounds.

We will design a well-suited data-structure to abstract the elementary repeatable patterns that are compatible with a contact map and with additional constraints.

2.5 Clique

In large combinatorial models, the set of elementary repeatable patterns may not be represented explicitly. It is important to abstract it.

Let us consider the example of a clique of n proteins. We call a clique of n proteins any n kinds of protein such that each protein has exactly $n-1$ sites and that every pair of proteins of distinct kinds may be connected by exactly one pair of sites. The number of elementary repeatable patterns in a clique of n proteins is exponential with respect to n (there are indeed $\frac{n!}{k!}$ elementary repeatable patterns with exactly $k+1$ proteins, for any k such that $2 \leq k \leq n$). Thus they cannot be all enumerated. In this paper, we will instead compute exactly the set of bonds that may occur in repeatable patterns. Our approach is based on the use of some graphs that are derived from the contact map, and for which edges correspond to the potential bonds in elementary repeatable patterns. We use Tarjan's algorithm [18] to compute the strongly connected components of these graphs. Our analysis is sound and complete with respect to the constraints that are encoded in the contact map: a bond may occur in a repeatable pattern that is compatible with a given contact map if and only if it corresponds to an edge in a non trivial strongly connected component of the graph that is associated to this contact map. Moreover, it is possible to take into account additional constraints about the patterns that are proved to be unreachable by traditional static analysis [12,3].

Outline. The rest of the paper is organised as follows. In Sec. 3, we give some reminders about Kappa. We focus only on static reasoning about graphs. We do not introduce the notion of rules. We assume that additional constraints about reachable patterns come from a black box that we do not describe in this paper. In Sec. 4, we introduce two notions of graphs: the graph of the sites and the graph of the links. Both notions can be used to reason about the finiteness of the set of bio-molecular compounds in a Kappa model. Yet we will see in Sec. 5, that the graph of the links may be refined to take into account the patterns that may be proved unreachable by an external tool.

3 Kappa

In this section, we give some reminders about Kappa. We do not introduce the full semantics of Kappa. Instead, we introduce only the notions of site-graphs and of embeddings among them. We omit the notions of rules and of rule applications. We

also omit internal states, since we focus on the topology of the potential bindings between proteins. We refer to [8,11] for a more complete description of Kappa.

3.1 Signature

Firstly we define the signature of a model.

Definition 3.1 (signature) *A signature is a triple* $\Sigma \triangleq (\Sigma_{ag}, \Sigma_{site}, \Sigma_{ag\text{-}st})$ *where:*

(i) Σ_{ag} *is a finite set of agent types,*

(ii) Σ_{site} *is a finite set of site identifiers;*

(iii) $\Sigma_{ag\text{-}st} : \Sigma_{ag} \to \wp(\Sigma_{site})$ *is a site map.*

Agent types in Σ_{ag} denote agents of interest, as kinds of protein for instance. Site identifiers in Σ_{site} represent identified loci for capabilities of interactions. Agent types $A \in \Sigma_{ag}$ are associated with sets of sites $\Sigma_{ag\text{-}st}(A)$ which may be linked.

Example 3.2 (signature (model of the triangle)) *We define the signature for the model of the triangle (e.g. see Sec. 2.1):*

$$\Sigma \triangleq (\Sigma_{ag}, \Sigma_{site}, \Sigma_{ag\text{-}st})$$

where:

(i) $\Sigma_{ag} \triangleq \{A, B, C\}$;

(ii) $\Sigma_{site} \triangleq \{a, b, c\}$;

(iii) $\Sigma_{ag\text{-}st} \triangleq [A \mapsto \{b; c\}, B \mapsto \{a; c\}, C \mapsto \{a; b\}]$.

Example 3.3 (signature) *We define the signature for the model of the early events in the epidermic growth factor (e.g. see Sec. 2.4)::*

$$\Sigma \triangleq (\Sigma_{ag}, \Sigma_{site}, \Sigma_{ag\text{-}st})$$

where:

(i) $\Sigma_{ag} \triangleq \{EGF, EGFR, Grb2, ShC, Sos\}$;

(ii) $\Sigma_{site} \triangleq \{a, b, c, d, n, l, pi, r, Y7, Y48, Y68\}$;

(iii) $\Sigma_{ag\text{-}st} \triangleq \begin{bmatrix} EGF \mapsto \{r\}, EGFR \mapsto \{c, n, l, r, Y48, Y68\}, \\ Grb2 \mapsto \{a, b\}, ShC \mapsto \{pi, Y7\}, Sos \mapsto \{d\} \end{bmatrix}$.

3.2 Σ-graphs and morphisms among Σ-graphs

Σ-graphs are graphs. Their nodes are typed agents with some sites which may bear sets of binding states. Contact maps, patterns and bio-molecular compounds are specific kinds of Σ-graph.

Definition 3.4 (Σ-graphs) *A Σ-graph is a tuple* $G \triangleq (\mathcal{A}_G, type_G, \mathcal{S}_G, \mathcal{L}_G)$ *where:*

(i) $\mathcal{A}_G$ *is a finite set of agents,*

(ii) $type_G : \mathcal{A}_G \to \Sigma_{ag}$ *is a function mapping each agent to its type,*

(iii) $\mathcal{S}_G$ *is a subset of the set* $\{(n,i) \mid n \in \mathcal{A}_G, i \in \Sigma_{ag\text{-}st}(type_G(n))\}$,

(iv) $\mathcal{L}_G$ *is a function between the set* $\mathcal{S}_G$ *and the set* $\wp(\mathcal{S}_G \cup \{\dashv\})$ *such that for any two sites* $(n,i),(n',i') \in \mathcal{S}_G$, *we have* $(n',i') \in \mathcal{L}_G(n,i)$ *if and only if* $(n,i) \in \mathcal{L}_G(n',i')$.

The set $\mathcal{S}_G$ denotes the set of binding sites. Whenever $\dashv \in \mathcal{L}_G(n,i)$, the site (n,i) may be free. Whenever $(n',i') \in \mathcal{L}_G(n,i)$ (and hence $(n,i) \in \mathcal{L}_G(n',i')$), the sites (n,i) and (n',i') may be bound together.

For a Σ-graph G, we write as $\mathcal{A}_G$ its set of agents, $type_G$ its typing function, $\mathcal{S}_G$ its set of sites, and $\mathcal{L}_G$ its set of links.

Example 3.5 (Σ-graphs (model of the triangle)) *We give two examples of* Σ*-graph for the model of the triangle (eg. Sec. 2.1).*

The graph that is depicted in Fig. 1(a) is the Σ*-graph* $\mathcal{T}_{CM}$ *defined as follows:*

(i) $\mathcal{A}_{\mathcal{T}_{CM}} \triangleq \{1,2,3\}$;

(ii) $type_{\mathcal{T}_{CM}} \triangleq [1 \mapsto A, 2 \mapsto B, 3 \mapsto C\,]$;

(iii) $\mathcal{S}_{\mathcal{T}_{CM}} \triangleq \{(1,b),(1,c),(2,a),(2,c),(3,a),(3,b)\}$;

(iv) $\mathcal{L}_{\mathcal{T}_{CM}} \triangleq \left[\begin{array}{l}(1,b) \mapsto \{\dashv,(2,a)\},(1,c) \mapsto \{\dashv,(3,a)\},(2,a) \mapsto \{\dashv,(1,b)\},\\ (2,c) \mapsto \{\dashv,(3,b)\},(3,a) \mapsto \{\dashv,(1,c)\},(3,b) \mapsto \{\dashv,(2,c)\}\end{array}\right]$.

and the bio-molecular compound that is drawn in Fig. 1(b), is the Σ*-graph* $\mathcal{T}_\Sigma$ *that is defined as follows:*

(i) $\mathcal{A}_{\mathcal{T}_\Sigma} \triangleq \{1,2,3\}$;

(ii) $type_{\mathcal{T}_\Sigma} \triangleq [1 \mapsto A, 2 \mapsto B, 3 \mapsto C\,]$;

(iii) $\mathcal{S}_{\mathcal{T}_{CM}} \triangleq \{(1,b),(1,c),(2,a),(2,c),(3,a),(3,b)\}$;

(iv) $\mathcal{L}_{\mathcal{T}_\Sigma} \triangleq \left[\begin{array}{l}(1,b) \mapsto \{(2,a)\},(1,c) \mapsto \{(3,a)\},(2,a) \mapsto \{(1,b)\},\\ (2,c) \mapsto \{(3,b)\},(3,a) \mapsto \{(1,c)\},(3,b) \mapsto \{(2,c)\}\end{array}\right]$.

Example 3.6 (Σ-graph (EGF model)) *We give two examples of* Σ*-graph for the model of the early events of the integration of the epidermic growth factor (eg. see Sec. 2.4).*

The graph that is depicted in Fig. 5(a) is the Σ*-graph* G_{CM} *defined as follows:*

(i) $\mathcal{A}_{G_{CM}} \triangleq \{1,2,3,4,5\}$;

(ii) $type_{G_{CM}} \triangleq [1 \mapsto EGF, 2 \mapsto EGFR, 3 \mapsto Grb2, 4 \mapsto ShC, 5 \mapsto Sos]$;

(iii) $\mathcal{S}_{G_{CM}} \triangleq \bigcup\{(n,i) \mid n \in \mathcal{A}_{G_{CM}}, i \in \Sigma_{ag\text{-}st}(type_{G_{CM}})\}$;

$$\text{(iv)}\ \mathcal{L}_{G_{CM}} \triangleq \left[\begin{array}{l}(1,r)\mapsto\{\dashv,(2,l)\},\\(2,l)\mapsto\{\dashv,(1,r)\},(2,r)\mapsto\{\dashv,(2,r)\},(2,c)\mapsto\{\dashv,(2,n)\},\\(2,n)\mapsto\{\dashv,(2,c)\},(2,\mathit{Y48})\mapsto\{\dashv,(4,pi)\},(2,\mathit{Y68})\mapsto\{\dashv,(3,a)\},\\(3,a)\mapsto\{\dashv,(2,\mathit{Y68}),(4,\mathit{Y7})\},(3,b)\mapsto\{\dashv,(5,d)\},\\(4,pi)\mapsto\{\dashv,(2,\mathit{Y48})\},(4,\mathit{Y7})\mapsto\{\dashv,(3,a)\},\\(5,d)\mapsto\{\dashv,(3,b)\},\end{array}\right].$$

and the Σ*-graph* G_Σ *that is defined as follows:*

(i) $\mathcal{A}_{G_\Sigma} \triangleq \{1,2,3,4,5,6,7,8,9\}$;

$$\text{(ii)}\ \mathit{type}_{G_\Sigma} \triangleq \left[\begin{array}{l}1\mapsto \mathit{EGF},2\mapsto \mathit{EGF},3\mapsto \mathit{EGFR},4\mapsto \mathit{EGFR},\\5\mapsto \mathit{Grb2},6\mapsto \mathit{Grb2},7\mapsto \mathit{ShC},8\mapsto \mathit{ShC},9\mapsto \mathit{Sos}\end{array}\right];$$

(iii) $\mathcal{S}_{G_\Sigma} \triangleq \bigcup\{(n,i) \mid n\in\mathcal{A}_{G_\Sigma}, i\in\Sigma_{ag\text{-}st}(\mathit{type}_{G_\Sigma})\}$;

$$\text{(iv)}\ \mathcal{L}_{G_\Sigma} \triangleq \left[\begin{array}{l}(1,r)\mapsto\{(3,l)\},(2,r)\mapsto\{(4,l)\},\\(3,l)\mapsto\{(1,r)\},(3,r)\mapsto\{(4,r)\},(3,c)\mapsto\{(4,n)\},\\(3,n)\mapsto\{\dashv\},(3,\mathit{Y48})\mapsto\{(7,pi)\},(3,\mathit{Y68})\mapsto\{\dashv\},\\(4,l)\mapsto\{(2,r)\},(4,r)\mapsto\{(3,r)\},(4,c)\mapsto\{\dashv)\},\\(4,n)\mapsto\{(3,c)\},(4,\mathit{Y48})\mapsto\{(8,pi)\},(4,\mathit{Y68})\mapsto\{(6,a)\},\\(5,a)\mapsto\{(7,\mathit{Y7})\},(5,b)\mapsto\{(9,d)\},\\(6,a)\mapsto\{(4,\mathit{Y68})\},(6,b)\mapsto\{\dashv\},\\(7,pi)\mapsto\{(3,\mathit{Y48})\},(7,\mathit{Y7})\mapsto\{(5,a)\},\\(8,pi)\mapsto\{(4,\mathit{Y48})\},(8,\mathit{Y7})\mapsto\{\dashv\},\\(9,d)\mapsto\{(5,b)\}\end{array}\right].$$

The Σ-graphs $\mathcal{T}_{CM}$ and G_{CM} play a specific role: we call them the contact maps of their respective models. In a contact map, each agent type occurs exactly once and each agent documents its full set of sites. Moreover every sites may be free, but may also be bound to some other sites as specified in the corresponding Σ-graph. Contact maps encode some specific typing disciplines [7]: they summarise the potential bonds between agents.

Σ-graphs may be related by structure-preserving maps of agents, called morphisms. The definition of a morphism between two Σ-graphs is given as follows:

Definition 3.7 (morphisms) *A* morphism $h : G \to H$ *from the* Σ*-graph* G *into the* Σ*-graph* H *is a function of agents* $h : \mathcal{A}_G \to \mathcal{A}_H$ *satisfying, for all agent identifiers* n, $n' \in \mathcal{A}_G$, *for all site identifiers* $i \in \Sigma_{ag\text{-}st}(\mathit{type}_G(n))$, $i' \in \Sigma_{ag\text{-}st}(\mathit{type}_G(n'))$*:*

(i) $\mathit{type}_G(n) = \mathit{type}_H(h(n))$;

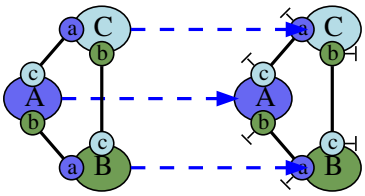


Fig. 6. The unique morphism from the Σ-graph $\mathcal{T}_\Sigma$ and the Σ-graph $\mathcal{T}_{CM}$. Each agent of the Σ-graph $\mathcal{T}_\Sigma$ is mapped to the unique agent of the Σ-graph $\mathcal{T}_{CM}$ of this type.

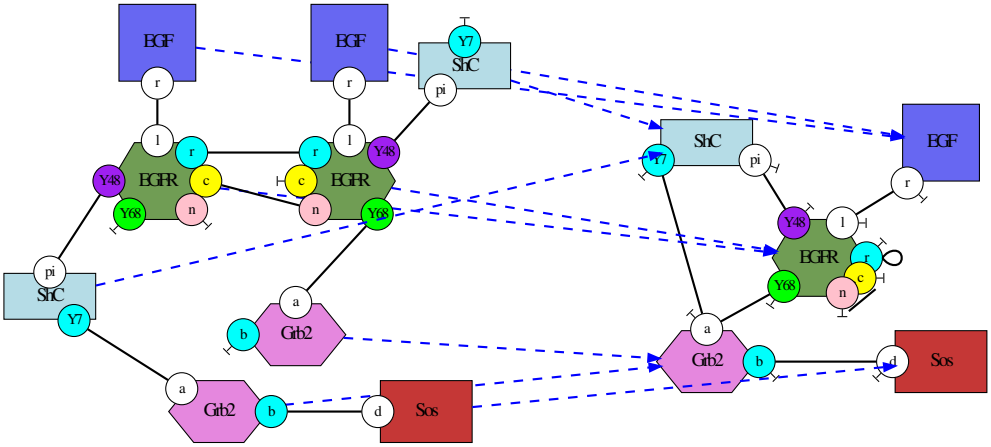


Fig. 7. The unique morphism from the Σ-graph G_Σ and the Σ-graph G_{CM}. Each agent of the Σ-graph G_Σ is mapped to the unique agent of the Σ-graph G_{CM} of this type.

(ii) *if* $(n, i) \in \mathcal{S}_G$, *then* $(h(n), i) \in \mathcal{S}_H$;

(iii) *if* $(n', i') \in \mathcal{L}_G(n, i)$, *then* $(h(n'), i') \in \mathcal{L}_H(h(n), i)$;

(iv) *if* $\dashv \in \mathcal{L}_G(n, i)$, *then* $\dashv \in \mathcal{L}_H(h(n), i)$.

Morphisms preserve the type of agents. They also preserve each agent set of sites, but more sites may be documented in the image of the morphism. A site that may be free shall be mapped to a site that may be free. Two sites that may be bound together shall be mapped to two sites that may be bound together.

Example 3.8 (morphisms (model of the triangle)) *A morphism between the Σ-graph $\mathcal{T}_\Sigma$ and the Σ-graph $\mathcal{T}_{CM}$ is depicted in Fig. 6. This morphism maps any agent of the Σ-graph $\mathcal{T}_\Sigma$ to the unique agent of the Σ-graph $\mathcal{T}_{CM}$ having the same type. This is indeed the unique morphism from the Σ-graph $\mathcal{T}_\Sigma$ to the Σ-graph $\mathcal{T}_{CM}$.*

Example 3.9 (morphisms (EGF model)) *A morphism between the Σ-graph G_Σ and the Σ-graph G_{CM} is depicted in Fig. 7. This morphism maps any agent of the Σ-graph G_Σ to the unique agent of the Σ-graph G_{CM} having the same type. This is indeed the unique morphism from the Σ-graph G_Σ to the Σ-graph G_{CM}.*

Two morphisms from a Σ-graph E to a Σ-graph F, and from the Σ-graph F to a Σ-graph G respectively, compose in the usual way (and form a morphism from the Σ-graph E into the Σ-graph G).

3.3 Patterns and embeddings

Now we restrict the definition of Σ-graphs so as to focus on the ones that may express parts of the state of the system. These Σ-graphs, that we call patterns, are defined as follows:

Definition 3.10 (patterns) *A pattern is a Σ-graph P such that, for every site $s \in \mathcal{S}_P$ both following conditions are satisfied:*

(i) *the set $\mathcal{L}_P(s)$ contains at most one element;*

(ii) *the set $\mathcal{L}_P(s)$ does not contain the element s.*

The first condition ensures that the state of every site is either unspecified, or free, or bound to a single specific site. The second condition ensures that a site is never bound to itself.

A bio-molecular compound is a connected pattern in which the state of each site is documented (no further information may be added).

Patterns may be related by embeddings. Besides preserving the structure of patterns, embeddings map agents to agents injectively.

Definition 3.11 (embeddings) *An embedding is a morphism from a pattern into another one, that is induced by an injective agent function.*

As opposed to classical notions of embeddings between graphs, embeddings between patterns preserve free sites. When there exists an embedding from a pattern E into a pattern F, we often write that the pattern E embeds in the pattern F, or that E occurs in the pattern F. The composition of two embeddings is an embedding. Two patterns E and F are isomorphic whenever there exist an embedding from the pattern E to the pattern F and an embedding from the pattern F to the pattern E, which is denoted as $E \approx F$. We also denote as $[E]_{\approx}$ the $\approx$-equivalence class of the pattern E. The $\approx$-equivalence class $[E]_{\approx}$ of the pattern E is made of all the patterns that are isomorphic to the pattern E.

4 Reasoning on repeatable patterns

In this section, we formalise the problem of deciding whether or not a contact map is compatible with an infinite set of bio-molecular compounds. Then we introduce two kinds of graph to reason about this problem.

4.1 Interpretation of a contact map

Intuitively, a contact map may be interpreted as the set of the bio-molecular compounds which may be projected into that contact map by the means of a morphism. However this notion is not relevant to reason about the finiteness of the set of the bio-molecular compounds of a given model. Indeed with such a definition, each model admitting at least one bio-molecular compound would admit an infinite number of bio-molecular compounds due to isomorphisms. Instead we consider $\approx$-equivalence classes of bio-molecular compounds.

Definition 4.1 (interpretation of a contact map) *The interpretation $[\![G_{CM}]\!]$ of a contact map G_{CM} is defined as the set of all the $\approx$-equivalence classes of bio-*

molecular compounds $[G]_{\approx}$ *such that there exists a morphism from the site graph* G *into the contact map* G_{CM}*.*

We can now state properly the problem we want to solve:

Problem 4.2 *Let* G_{CM} *be a contact map. We are looking for an automatic procedure to decide whether the set* $[\![G_{CM}]\!]$ *is finite, or not.*

4.2 Chains

In this section, we introduce a kind of pumping lemma in order to reduce Problem 4.2 to the one of detecting a repeatable pattern.

Firstly, we define properly a repeatable pattern as a chain of agents which may be iterated to form arbitrarily long patterns.

Definition 4.3 (chain) *A pattern is called a chain if and only if it satisfies the following properties:*

(i) *it is connected;*

(ii) *every agent documents at most two sites;*

(iii) *there is at least one agent which does not have two sites bound.*

A chain is formed either of a single agent with at most two sites, or of a linear chain of agents with exactly two extremities. In the former case, each site of the single agent is either free, or in an unspecified binding state. In the latter case, every agent not in the extremities has two sites and these sites are bound whereas every agent on the extremity has exactly one site that is bound and potentially at most one other site (which may be free, or with an unspecified binding state).

A chain is a repeatable pattern whenever it contains at least two agents and its extremities may be replug to each other. This is formalised as follows.

Definition 4.4 (repeatable pattern) *A chain is called a repeatable pattern if and only if the following conditions are satisfied:*

(i) *it has two distinct extremities;*

(ii) *it has no free sites;*

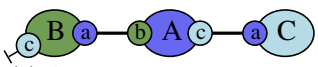
(iii) *the agents at both extremities are of the same kind;*

(iv) *the bound sites at both extremities have different names.*

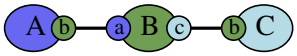
A repeatable pattern is said elementary if and only if it contains no occurrence of repeatable patterns (besides itself).

Example 4.5 *We consider four patterns in Fig. 8. All these patterns are chains. The pattern in Fig. 8(a) is not repeatable because one of its extremity has a site that is free. The pattern in Fig. 8(b) is not repeatable because its extremities are not of the same kind. The pattern in Fig. 8(c) is not repeatable because its extremities document the same site. The pattern in Fig. 8(d) is repeatable (and elementary).*

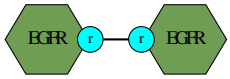
Several instances of a repeatable pattern may be combined in order to form arbitrary long chains of agents. We define formally the iterations of a repeatable pattern as follows:



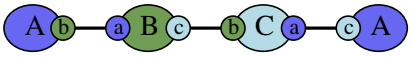
(a) A chain with one site free.



(b) A chain with two extremities of different kinds.

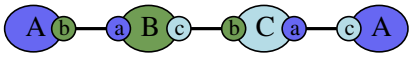


(c) A chain with two extremities of the same kind but with the same bound site.

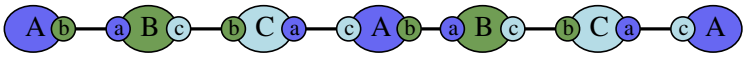


(d) An elementary repeatable pattern.

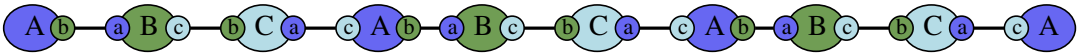
Fig. 8. Four patterns. Each of them is a chain. But only the last one is repeatable.



(a) First iteration.



(b) Second iteration.



(c) Third iteration.

Fig. 9. Three iterations of the pattern of Fig. 1(c). Each iteration is obtained by plugging this pattern at the end of the previous iteration of it.

Definition 4.6 (iterations of a repeatable pattern) *Let P be a repeatable pattern. The iterations of the pattern P are defined recursively as follows:*

(i) *the pattern P is an iteration of the pattern P;*

(ii) *for every iteration P' of the pattern P, the pattern that is obtained by fusing one extremity of P' with one extremity of P that is compatible with this extremity, is an iteration of P as well.*

In Def. 4.6, the choice of the extremity of P' does not matter, the result will be the same up to isomorphism.

Example 4.7 *We give in Fig. 9 the first three iterations of the pattern that is depicted in Fig. 1(c).*

Now we establish our pumping lemma.

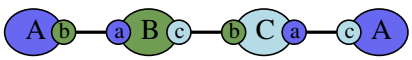
Lemma 4.8 (pumping lemma) *Let G_{CM} be a contact map. Both following assertions are equivalent:*

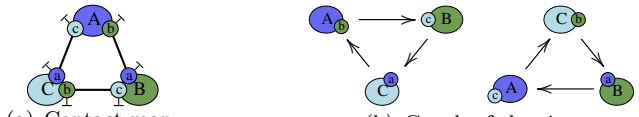
(i) *The set $[\![G_{CM}]\!]$ is infinite;*

(ii) *There exist an elementary repeatable pattern P and a morphism between the pattern P and the contact map G_{CM}.*

4.3 Graph of the sites

It is tempting to interpret the following repeatable pattern:





(a) Contact map. (b) Graph of the sites.

Fig. 10. ABC model. In 10(a), we recall the contact map. In Fig. 10(b), we give the graph of the sites that is associated with this contact map. The nodes of these graphs are the sites of the contact map. There is an oriented edge between a node s and a node t if and only if there is link in the contact map between the site s and a site of the protein that carries the site t but on a different site.

as the sequence of sites b of A, a of B, c of B, b of C, a of C, and c of A. Yet in this sequence, sites are polarised. Each site on a odd position and the next one always belong to the same kind of protein, whereas there always exists a link between each site on an even position and the next one. Due to this polarisation, it is tempting to consider the sub-sequence that is made of each other site in that sequence of sites.

Next we define a graph that stands for all the potential sequences of sites that may occur on even occurrences in the repeatable patterns that are compatible with a given contact map. This graph is called the graph of the sites of this contact map.

Definition 4.9 (graph of the sites) *Let G_{CM} be a contact map.*

The contact map G_{CM} is associated with a classical graph $(\mathcal{V}, \mathcal{E})$, called the graph of the sites of the contact map G_{CM}, which is defined as follows:

- *$\mathcal{V}$ is the set $\mathcal{S}_{G_{CM}}$ of the sites of the Σ-graph G_{CM}.*
- *$\mathcal{E}$ is the subset of $V \times V$ such that $((n,i),(n',i')) \in E$ if and only if there exists a site $i'' \in \Sigma_{ag\text{-}st}(type_{G_{CM}}(n'))$ such that: $i'' \neq i'$ and $(n',i'') \in \mathcal{L}_{G_{CM}}(n,i)$.*

In the edges of the graph of the sites, the sites via with we enter the target agent is kept implicit.

The following theorem relates the existence of cycles in the graph of the sites to the existence of repeatable patterns in the model.

Theorem 4.10 *Let G_{CM} be a contact map.*

Let A and B be two kinds of agent and i and i' be two site names.

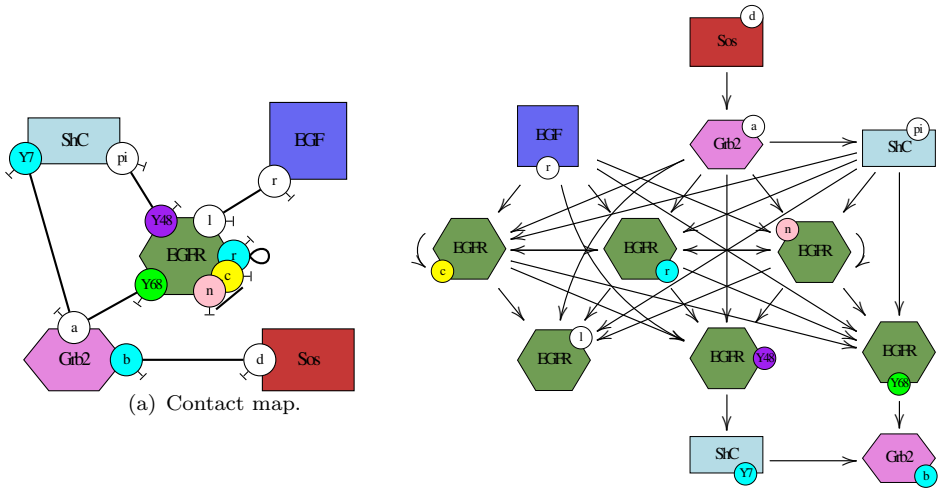
Both following properties are equivalent:

(i) *There exists a repeatable pattern with an agent of kind A connected via its site i to one site of an agent of kind B itself connected to another agent on site i'.*

(ii) *There exist two agents n and n' respectively of kinds A and B in the contact map G_{CM}, and a cycle in the graph of the sites of the contact map G_{CM} that passes by the edge $((n,i),(n',i'))$.*

Thus, Thm. 4.10 reduces the problem of deciding whether a contact map is compatible with an infinite number of non-isomorphic bio-molecular compounds to the one of computing the strongly connected components of the graph of the sites of this contact map.

Example 4.11 (graph of the sites (ABC model)) *In Fig. 10, we compute the graph of the sites for the contact map of the model with three kinds of protein that may form a triangle. It is worth noticing that this graph is made of exactly two non trivial strongly connected components. Each one corresponds to the triangle ABC depending whether it is scanned clockwise or counter-clockwise. Further constraints*



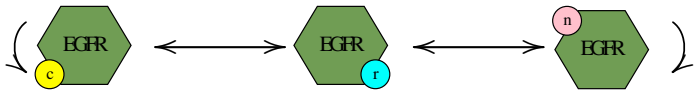
(a) Contact map.

(b) Graph of the sites.

Fig. 11. EGF model. In 11(a), we recall the contact map. In Fig. 11(b), we give the graph of the sites that is associated with this contact map.

would be required on the bio-molecular compounds of the model to prove that there is a finite amount of them (the contact map of the model is compatible with an infinite number of bio-molecular compounds).

Example 4.12 (graph of the sites (EGF model)) *In Fig. 11, we compute the graph of the sites for the contact map of the model of the early events in the integration of the epidermic growth factor. It is worth noticing that this graph has only one non trivial strongly connected component, which is depicted as follows:*

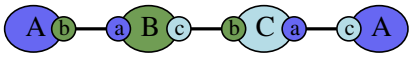


Further constraints are required on the bio-molecular compounds of the model to prove that there is a finite amount of them (the contact map of the model is compatible with an infinite number of different bio-molecular compounds).

4.4 Graph of the links

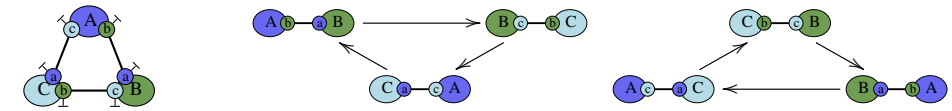
We do not know how to refine the graph of the sites of a given contact map to take into account further constraints about the bio-molecular compounds that are reachable. We consider in this section another kind of graph which focuses on the different links in the contact map and that will be easier to refine.

Now we interpret the following repeatable pattern:



as the sequence of (oriented) links from the site b of A to the site a of B, from the site c of B to the site b of C, and from the site a of C to the site c of A.

In the following, we define a graph that stands for all the potential sequences of links that may occur consecutively on the repeatable patterns that are compatible



(a) Contact map. (b) Graph of the links.

Fig. 12. ABC model. In 12(a), we recall the contact map. In Fig. 12(b), we give the graph of the links that is associated with this contact map. The nodes of these graphs are obtained by orienting the links of the contact map (hence there are two nodes per link).

with a given contact map. This graph is called the graph of the links.

Definition 4.13 (graph of the links) *Let G_{CM} be a contact map.*

The contact map G_{CM} is associated with a classical graph $(\mathcal{V}, \mathcal{E})$, called the graph of the links that is defined as follows:

- *$\mathcal{V}$ is the subset of the pairs of elements (s, s') of the set $\mathcal{S}_{G_{CM}}$ of the sites of the Σ-graph G_{CM}, such that $s' \in \mathcal{L}_{G_{CM}}(s)$.*
- *$\mathcal{E}$ is the subset of the pairs $((s, s'), (s'', s'''))$ of pairs of sites in $\mathcal{V} \times \mathcal{V}$ for which there exists an agent $n \in \mathcal{A}_{G_{CM}}$ and two different site names i and $i' \in \Sigma_{ag\text{-}st}(type_{G_{CM}}(n))$ such that $s' = (n, i)$ and $s'' = (n, i')$.*

The condition on the edges of the graph of the links ensures that edges connect bonds that may appear consecutively in a repeatable pattern.

The following theorem relates the existence of cycles in the graph of the links to the existence of repeatable patterns in the model.

Theorem 4.14 *Let G_{CM} be a contact map.*

Let A and B be two kinds of agent and i and i' be two site names.

Both following properties are equivalent:

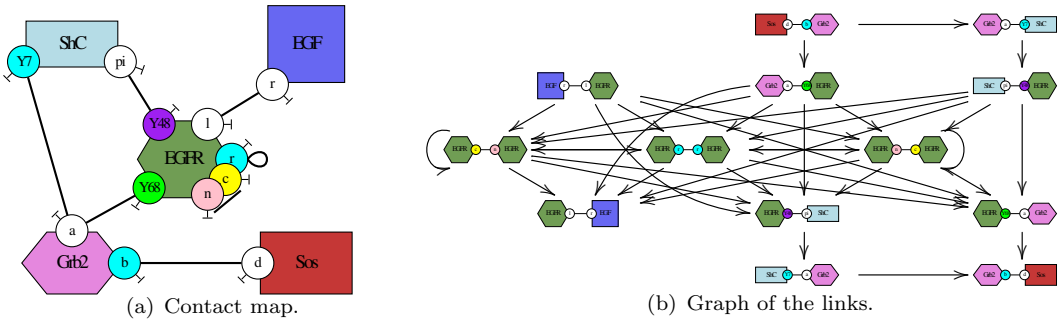
(i) *There exists a repeatable pattern with an agent of kind A connected via its site i to the site i' of an agent of kind B;*

(ii) *There exist two agents n and n' respectively of kinds A and B in the contact map G_{CM} and a cycle in the graph of the links of the contact map G_{CM} that passes by the vertex $((n, i), (n', i'))$.*

Thus, Thm. 4.14 reduces the problem of deciding whether a contact map is compatible with an infinite number of non-isomorphic bio-molecular compounds to the one of computing the strongly connected components of the graph of its links.

Example 4.15 (graph of the links (ABC model)) *In Fig. 12, we compute the graph of the links for the contact map of the model with three kinds of protein that may form a triangle. It is worth noticing that this graph is made of exactly two non trivial strongly connected components. Each one corresponds to the triangle ABC depending whether it is scanned clockwise or counter-clockwise. Further constraints would be required on the bio-molecular compounds of the model to prove that there is a finite amount of them (the contact map of the model is compatible with an infinite number of bio-molecular compounds).*

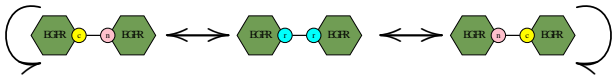
Example 4.16 (graph of the links (EGF model)) *In Fig. 13, we compute the graph of the links for the contact map of the model of the early events in the integration of the epidermic growth factor. It is worth noticing that this graph has only*



(a) Contact map. (b) Graph of the links.

Fig. 13. EGF model. In 13(a), we recall the contact map. In Fig. 13(b), we give the graph of the links that is associated with this contact map. There are two nodes per link, except for the link between the site r of $EGFR$ and itself, for which there is a unique node.

one non trivial strongly connected component:



Further constraints are required on the bio-molecular compounds of the model to prove that there is a finite amount of them (the contact map of the model is compatible with an infinite number of bio-molecular compounds).

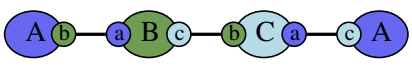
5 Taking into account the result of a static analysis

In this section, we explain how to refine the graph of the links of a given contact map, in order to take into account some additional constraints about the bio-molecular compounds that are potentially reachable. These constraints may come from a static analysis [12,3] taken as a black box and they may take the form of a set of patterns that shall occur in no reachable bio-molecular compounds. These constraints cannot be written in the contact map because the contact map describes only non relational information about the potential state of sites.

In the case of the model of the early events of the integration of the epidermic growth factor, the analysis that is described in [12] can infer automatically, from the set of rules and the initial state, that none of the following patterns:

is reachable. That is to say that a receptor cannot be bound to two different other instances of receptors.

The analysis that is described in [10] generalises this approach to arbitrary cycles of proteins. In the example of the three kinds of protein that may form a triangle, this static analysis infers that no two As may occur in a reachable bio-molecular compound, by proving that the following pattern:



is unreachable.

We refine the statement of Problem 4.2 so as take into account the constraints potentially coming from an external static analysis.

Definition 5.1 (interpretation with a set of forbidden patterns) *The interpretation $[\![G_{CM}, \mathcal{P}]\!]$ of a contact map G_{CM} with a set of forbidden patterns $\mathcal{P}$ is defined as the set of the $\approx$-equivalence classes of bio-molecular compound $[G]_{\approx}$ such that there exists a morphism from the site graph G into the contact map G_{CM} and that G contains no occurrence of patterns from the set $\mathcal{P}$.*

Problem 5.2 *Let G_{CM} be a contact map and $\mathcal{P}$ be a set of patterns.*

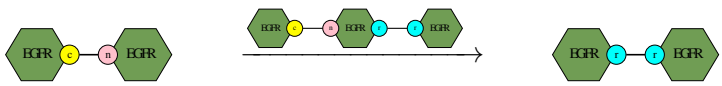
We are looking for an automatic procedure to decide whether the set $[\![G_{CM}, \mathcal{P}]\!]$ is finite, or not.

In the following, we propose a graph structure to answer to Problem 5.2. Our approach is sound but not complete. It can detect and prove that the set of bio-molecular compounds is finite. But when it warns about potential repeatable patterns, it may be a false positive. We do not look for a complete procedure because on the first hand detecting whether or not a pattern is reachable is not decidable in Kappa [16], and on the second hand detecting whether a pattern may occur in a set of bio-molecular compounds that do not contain patterns from a given set is not so easy due to potential overlaps between patterns. Thus we rely on a sound but not complete procedure.

We perform in two steps.

Firstly we label every edge of the graph of the links by a chain of agents. More precisely, the label of an edge is obtained by fusing the second agent of the source node with the first agent of the target node. We keep this orientation for the chain of agents that are now used to label the edges of the graph of the links (that is to say that two links are identified as respectively the source and the target of the chain of agents).

Example 5.3 (labelled edge) *We give as follows an example of a labelled edge in the graph of the links for the EGF model:*



The edge is labelled with a chain of three agents. The source of the chain is the link between the site c and the site n whereas the target of the chain if the link between the two sites r.

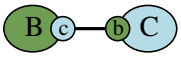
We notice that in our model this chain is indeed unreachable.

The label of an edge must be understood as an explanation about how the link of the source of this edge may be connected to the link of its target within a bio-molecular compound that is potentially reachable. Whenever an edge is labelled with a chain that contains a pattern that is unreachable, this edge may be safely discarded. The longer a chain of agents is, the more constraints it imposes. The second step consists in combining consecutive edges, in order to extend their labels into longer chains. It is worth noticing that given a node in the graph of the links, the target of the label of every incoming edge and the source of the label of every outcoming edge are indeed the same as the pattern that is labelling this node. We

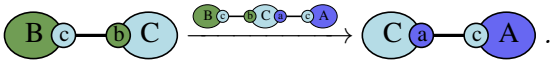
will keep this property as a structural invariant of the graph when combining the edges of the graph. Given a node in the graph of the links, an incoming edge and an outcoming edge may be composed by fusing the target of the incoming edge with the source of the outcoming edge to form an edge from the source of the incoming edge to the target of the outcoming edge.

Now we can define precisely the second step: the second step consists in selecting both a node and an incoming edge of this node so as to replace this edge with the set of all the edges that may be obtained by combining it with an outcoming edge of the node that has been selected. This transformation preserves the structural invariant and increases the length of the labels of the edges in the graph. Here again, we can safely discard every edge that is labelled with a chain that contains a pattern that is unreachable.

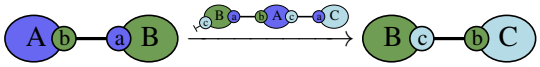
Example 5.4 (graph refinement (ABC model)) *In the model with three kinds of protein that may form a triangle (e.g. see Fig. 12(b)), the node:*



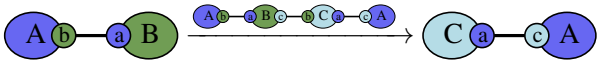
has only the following outcoming edge:



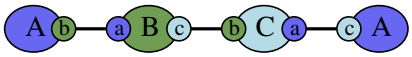
Thus, its incoming edge:



may be safely replaced with the following one:

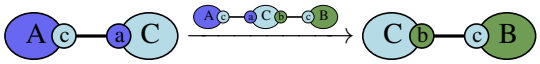


that is obtained as the composition of both edges. This new edge may then be discarded since the pattern:

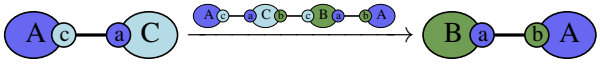


is black-listed.

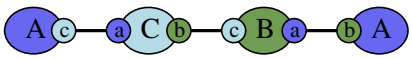
The same way, the edge:



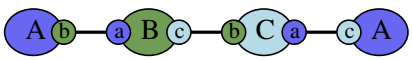
may be safely replaced with the following one:

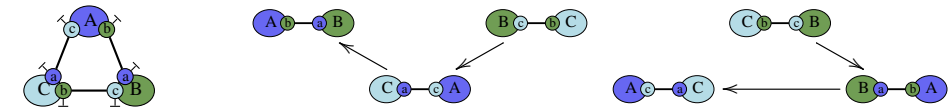


which may then be discarded, since the pattern:



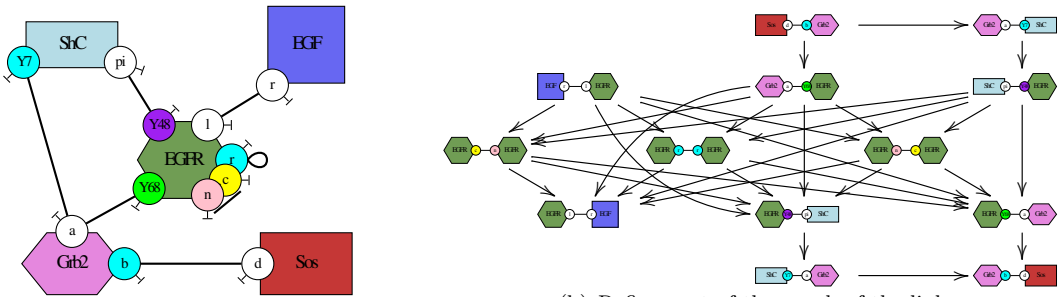
is isomorphic to the pattern:





(a) Contact map. (b) Refinement of the graph of the links.

Fig. 14. ABC model. In 14(a), we recall the contact map. In Fig. 14(b), we refine the graph of the links to take into account the constraints that two instances of A may not occur in a same connected component.



(a) Contact map. (b) Refinement of the graph of the links.

Fig. 15. EGF model. In 15(a), we recall the contact map. In Fig. 15(b), we refine the graph of the links that is associated with this contact map, by taking into account that a given receptor cannot be bound simultaneously to two different other receptors.

which is black-listed.

As a result, we obtain the refined graph of the links, that is depicted in Fig. 14(b).

The following theorem states the soundness of our approach.

Theorem 5.5 *Let G_{CM} be a contact map. Let $\mathcal{P}$ be a set of patterns. Let G be a refinement of the graph of the links of the contact map, according to the set of patterns $\mathcal{P}$. We assume that there exists a bio-molecular compound S such that $[S]_{\approx} \in [\![G_{CM}, \mathcal{P}]\!]$ that contains a repeatable pattern P such that no iteration of the pattern P contains an occurrence of a pattern in the set $\mathcal{P}$.*

Then, for every repetition Q of the pattern P, for every two agent identifiers n, n' and every two site names i, i' such that $\mathcal{L}_Q(n, i) = \{(n', i')\}$, there exists two agent identifiers n'', n''' such that $type_P(n) = type_{G_{CM}}(n'')$, $type_P(n') = type_{G_{CM}}(n''')$ and there exists a cycle in the graph G passing by the vertex $((n'', i), (n''', i'))$.

Intuitively, if an iteration of a repeatable pattern P contains an occurrence of a forbidden pattern P', then, the pattern P cannot be iterated an unbounded number of times in a reachable bio-molecular compound, otherwise eventually its iterations will contain occurrences of the pattern P', which is forbidden. The theorem states that vertices that belong to non trivial strongly connected components in a refined graph is a super-set of the bonds that may occur in a repeatable pattern all the iterations of which are compatible both with the contact map and with the black-listed patterns. If the refined graph is acyclic, then the set of the bio-molecular compounds that are reachable is necessarily finite.

Example 5.6 (refined graph of the links (model with the triangle)) *In Fig. 14, we refine the graph of the links for the contact map of the model ABC by taking into account that any pattern with several instances of the protein A is unreachable. We follow the steps that have been described in Exmp. 5.4 to prune*

two edges. The graph that is obtained this way (see Fig. 14(b)) is acyclic, which proves that the set of reachable bio-molecular compounds is finite in this model.

Example 5.7 (refined graph of the links (EGF model)) *In Fig. 15, we refine the graph of the links for the contact map of the model of the early events in the integration of the epidermic growth factor, by taking into account the fact that a given receptor cannot be bound simultaneously to several other receptors. Indeed every edge of the strongly connected component is initially labelled with a black-listed pattern, thus they can be discarded directly. The graph that is obtained (see Fig. 15(b)) is acyclic, which proves that the model involves only a finite set of reachable bio-molecular compounds.*

6 Conclusion

In this paper, we have provided some decision procedures to detect whether or not the set of bio-molecular compounds of rule-based models, such as the ones that are written in Kappa [8] or in BNGL [2], is finite or not. Our approach is mainly based on top of the contact map, a Σ-graph which summarises the potential links between the binding sites of proteins. The contact map is translated into a classical graph which encodes either the potential succession of sites, or the potential succession of links within bio-molecular compounds. Non trivial strongly connected components in this graph correspond to patterns that may be repeated an arbitrary number of times in the bio-molecular compounds that are reachable in the model. They can be detected using classical depth-first exploration without having to enumerate every elementary cycle [18]. The graph that stands for the potential succession of links in bio-molecular compounds can be refined in order to take into account some additional constraints computed by reachability analysis [6,12,3].

Our approach has been partially integrated in the static analyser KaSa [3]. More precisely, the construction of the graph of the potentially successive links has been implemented as well as the reduction with the static analysis that is described in [12]. This way, the analyser can cope accurately with the constraints involving potential cycles of two proteins. We plan to implement the generalisation that has been proposed in [10] that can handle precisely with models that can generate cyclic structures without creating arbitrary long bio-molecular compounds.

As future works, we plan to use weakly relational domains [17] to abstract more precisely the chains of proteins that may be embedded within the bio-molecular compounds that are reachable in a model. This analysis will allow to analyse accurately the rules that behave differently when applied in a uni-molecular or in a bi-molecular context.

References

[1] Blinov, M., J. Faeder, B. Goldstein and W. Hlavacek, *A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity*, Bio Systems **83** (2006), pp. 136–151.

[2] Blinov, M., J. R. Faeder, B. Goldstein and W. S. Hlavacek, *Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains.*, Bioinformatics (Oxford, England) **20** (2004).

[3] Boutillier, P., F. Camporesi, J. Coquet, J. Feret, K. Q. Ly, N. Theret and P. Vignet, *Kasa: a static analyzer for kappa*, in: *Proc. CMSB'18*, LNCS/LNBI **11095**, pp. 285–291.

[4] Camporesi, F., J. Feret and K. Q. Lý, *Kade: a tool to compile kappa rules into (reduced) odes models*, in: *Proc. CMSB 2017*, LNCS/LNBI **10545**, pp. 291–299, supplementary information available at www.di.ens.fr/~feret/CMSB2017-tool-paper.

[5] Danos, V., J. Feret, W. Fontana, R. Harmer and J. Krivine, *Rule-based modelling of cellular signalling, invited paper*, in: L. Caires and V. Vasconcelos, editors, *Proc. CONCUR'07*, LNCS **4703** (2007), pp. 17–41.

[6] Danos, V., J. Feret, W. Fontana and J. Krivine, *Abstract interpretation of cellular signalling networks*, in: *Proc. VMCAI'08*, Lecture Notes in Computer Science **4905** (2008), pp. 83–97.

[7] Danos, V., R. Harmer and G. Winskel, *Constraining rule-based dynamics with types*, MSCS **23** (2013).

[8] Danos, V. and C. Laneve, *Formal molecular biology*, TCS **325** (2004).

[9] Deeds, E. J., J. Krivine, J. Feret, V. Danos and W. Fontana, *Combinatorial complexity and compositional drift in protein interaction networks*, PLoS ONE **7** (2012).

[10] Faure de Pebeyre, A., *Static analysis of the formation of polymers in rule-based models* (2018), master 1 internship report (Master of interdisciplinary approached in life science).

[11] Feret, J., H. Koeppl and T. Petrov, *Stochastic fragments: A framework for the exact reduction of the stochastic semantics of rule-based models*, IJSI **7** (2013).

[12] Feret, J. and K. Q. Lý, *Reachability analysis via orthogonal sets of patterns.*, in: *Proc. SASB'16*, ENTCS, pp. 27–48.

[13] Gyori, B. M., J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu and P. K. Sorger, *From word models to executable models of signaling networks using automated assembly*, Molecular Systems Biology **13** (2017).

[14] Harmer, R., Y. L. Cornec, S. Légaré and I. Oshurko, *Bio-curation for cellular signalling: The KAMI project*, in: J. Feret and H. Koeppl, editors, *Proc. CMSB'17*, Lecture Notes in Computer Science **10545** (2017), pp. 3–19.

[15] Kim, P., L. Lu, Y. Xia and M. Gerstein, *Relating three-dimensional structures to protein networks provides evolutionary insights*, Science **314** (2006).

[16] Kreyßig, P., "Chemical Organisation Theory Beyond Classical Models: Discrete Dynamics and Rule-based Models," Ph.D. thesis, Friedrich-Schiller-University Jena (2014).

[17] Miné, A., *A few graph-based relational numerical abstract domains*, in: M. V. Hermenegildo and G. Puebla, editors, *Proc. SAS'02*, Lecture Notes in Computer Science **2477** (2002), pp. 117–132.

[18] Tarjan, R., *Depth first search and linear graph algorithms*, SIAM Journal On Computing **1** (1972).