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]

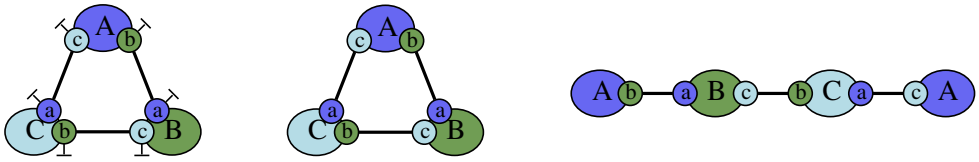
INRIA,
Centre de recherche INRIA de Paris, 75 012 Paris, France
Département d'informatique de l'École normale supérieure,
École normale supérieure, CNRS, PSL Research University, 75 005 Paris, France

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, it 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 of which are labelled with patterns while edges are labelled with overlaps between these 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 algortihm 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] 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)) specifies 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 species.

1 Introduction

2 Case studies

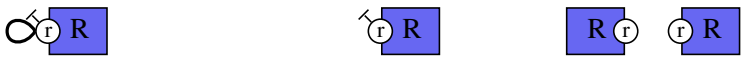
In this section, we introduce some examples to explain intuitively why there may be an unbounded number of bio-molecular compounds in a rule-based model. We also explain why naive approaches cannot be used to ensure that the number of bio-molecular compounds is finite in a given model, while identifying 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 type discipline. Every bio-molecular compound in our models shall satisfy the constraints 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 may not contain each kind of protein. Also it may contain several instances of a given one.

The contact map that is given in Fig. 1(a) is compatible with an infinite number



(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.

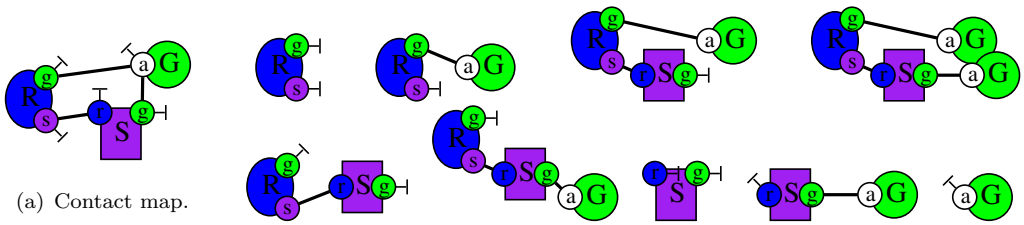
Fig. 3. An example of a protein with two sites a and b such that the site a of a protein may be bound to the site a of another protein and the site b may be bound to the site b of another protein. The contact map (Fig.3(a)) contains two self-loops. The pattern that is made of three proteins, the first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

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 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 site of another protein of the same kind. Roughly speaking a protein may form a monomer (when its site is free), or belongs to a dimer (when its site is bound). 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 only the two bio-molecular compounds that are depicted in Fig. 2(b) are compatible with this contact map: there is a finite number of kinds of bio-molecular compound them despite the cycle in the contact map.

One could think that self-loops should not be considered as cycles when trying to prove that the number of bio-molecular compounds of a model is finite. 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 the self-loops from the contact map. 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 a three proteins (see Fig. 3(b)) that may be repeated an arbitrary number of times in a bio-molecular compound.



(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 species (e.g. see Fig. 4(b)).

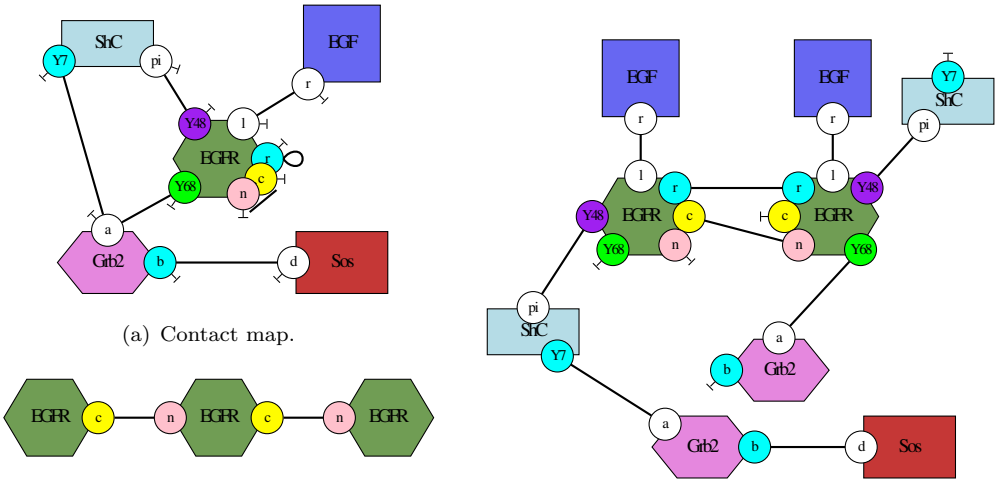
2.3 Conflicting bindings

In this example, we consider three kinds of protein G, R, and S. The proteins of kind G has 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 via their respective sires s and r. The unique site of proteins G may be bound either to the site g of an instance of the protein R, or to the site g of an instance of the protein S. We say that 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 species, 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 witht the constraints encoded in the contact map. We notice that 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, since we tried 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 ligands *EGF*. Once bound to their respective ligands, they can form a dimer by establishing a symmetric bond via their respective sites r. Compared to the BNGL model that is decribed in [1], we have also considered the asymmetric binding between receptors. To stabilize dimers, pairs of receptors that are bound via their sites r establish an asymetric 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 until the asymmetric one is. As a consequence, whenever the site c of a receptor is bound to the site n of another receptor, then both bond are also connected by a symmetric bond. This property can be computed by the static analysis that is described in [13,3]. Each receptor in a dimer may active the sites *Y48* and *Y68* of the other receptor



(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 species.

(since we focus only on the binding topology, we have omitted 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, bond to the protein *Sos*. The site *Y48* connects to the protein *Grb2* indirectly, thanks to the adapter protein *Shc*.

It is worth nothing that the contact map, that is depiceted in Fig. 5(a) does not provide all the information of the model. The constraint on the sites c, n, and r emerges from some mecanisms that are described by the means of rules. We do not describe the rule here since we focus on the topology of the potential bindings between the sites of proteins. Yet we assume that we may be provided with additional constraints of the form of some 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 any instance of the forbidden patterns.

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 and there is a conflict on the site a of the protein *Grb2* which may be bound to the receptor directly or via an adapater 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 species. Nevertheless, this pattern is not compatible with the additional constraints about symmetric and assymetric bindings in dimers: there is only a finite number of different bio-mollecular compounds that satisfies both the constraints from the contact map and

the additional constraint. In Fig. 5(c), we provide a typical example of bio-molecular compound in the EGFR model. This example is made of a dimer, with one site *Y68* free, one site *Y68* connected to a *Grb2* not connected to *Sos*, one site *Y48* connected to an adapter not connected to *Grb2*, and a site *Y48* connected to *Sos*. In general, 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 [2,5]. This model is made of 253 kinds of bio-molecular species. If we insert information about phosphorylation, we get a model with 932 kinds of bio-molecular species. Nevertheless, enumeration engines do not scale to large combinatorial networks such as the longer version of the EGFR model (including Ras, Erk, and Mapk) that is described in [6] and that involves around 10^19 different kinds of different bio-molecular compounds [7] or the model of the interactions found in the cytoplasmic portion of the Structural Interaction Network (cSIN) [10,14] that involves an infinite number of bio-molecular compounds.

Our goal is to design an well-suited data-structures 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 explicitely. 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 proteins 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 is indeed $\frac{n!}{k!}$ elementary repeatable patterns with exactly $k+1$ proteins, for any k such that $2 \leq k \leq n$).

As a consequence, it is not possible to enumerate all the elementary repeatable patterns that are compatible with large combinatorial contact maps. In this paper, we will instead compute exactely the set of bonds that may occur in these reapeatable 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 [15] 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 [13,3].

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 consrtraints 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 potential 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. 4.5, that the graph of the potential 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. Since we focus on counting some specific occurrences of patterns, we do not introduce the full semantics of Kappa. Instead, we introduce only the notions of site-graphs and of embeddings among them, and 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 [9,12] 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) $\Sigma_{ag\text{-}st} \triangleq [A \mapsto \{b; c\}, B \mapsto \{a; c\}, C \mapsto \{a; b\}]$.

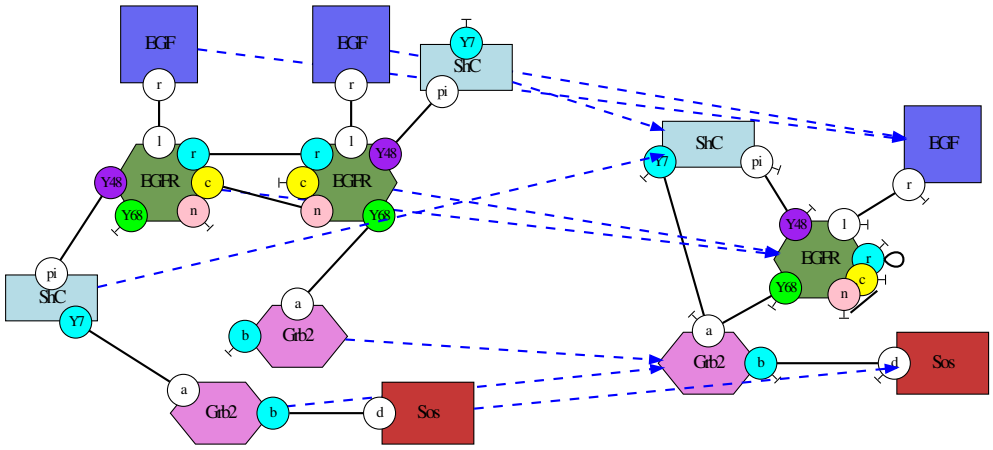
Example 3.3 (signature) *We define the signature for the model of the early events in the epidermic growth factor (e.g. see Sec. 2.4)::*

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

where:

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

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



(a) A morphism from G_Σ into G_{CM}.

Fig. 6. Two Σ-graphs G_{CM} and G_{SP}, and a morphism from G_{CM} to G_Σ. The Σ-graph G_{CM} is a contact map. It provides context-insensitive information about the potential state of each binding site. The Σ-graph G_{SP} is a bio-molecular species. It containts several instances of some proteins. Every site is documented in each protein instance and each site is either free, or bound to another site. The morphism between G_{CM} and G_{SP} smashes all the proteins of the Σ-graph G_{SP} according to their type. This is the unique morphism from the site graph G_{CM} into the site-graph G_{SP}.

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

3.2 Σ-graphs and morphisms among Σ-graphs

Σ-graphs are graphs the nodes of which 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) $\mathcal{A}_G \subseteq \mathbb{N}$ *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. Various levels of information may be given about the sites that are bound. Whenever $- \in \mathcal{L}_G(n,i)$, the site (n,i) may be bound to an unspecified site. 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. see Fig. 2.1).*

The graph that is depicted in Fig. 1(a) is the Σ-graph $\mathcal{T}_{CM}$ that is 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) $\mathcal{S}_{\mathcal{T}_{CM}} \triangleq \bigcup\{(n,i) \mid n \in \mathcal{A}_{\mathcal{T}_{CM}}, i \in \Sigma_{ag\text{-}st}(type_{\mathcal{T}_{CM}})\}$;

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

(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 Fig. 2.4).*

The graph that is depicted in Fig. 5(a) is the Σ-graph G_{CM} that is 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) $\mathcal{S}_{G_{CM}} \triangleq \bigcup\{(n,i) \mid n \in \mathcal{A}_{G_{CM}}, i \in \Sigma_{ag\text{-}st}(type_{G_{CM}})\}$;

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

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

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

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

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 type disciplines [8]: they summarise the potential bonds and provide contextual conditions over them [4].

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* $i \in \Sigma_{ag\text{-}st}(type_G(n))$, $i' \in \Sigma_{ag\text{-}st}(type_G(n'))$*:*

(i) $type_G(n) = type_H(h(n))$*;*

(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* $\dashv \in \mathcal{L}_H(h(n),i)$*;*

(v) *if* $- \in \mathcal{L}_G(n,i)$*, then* $\mathcal{L}_H(h(n),i) \cap \{-\} \cup \mathcal{S}_H \neq \emptyset$*.*

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. Lastly, whenever a site may be bound to an unspecified site, it shall be mapped to a site that is bound to either an unspecified or a specified (or both) one.

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. 7. The 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. The morphism maps any agent*

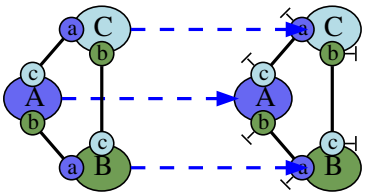


Fig. 7. The unique morphism from the Σ-graph $\mathcal{T}_\Sigma$ and the Σ-graph $\mathcal{T}_{CM}$.

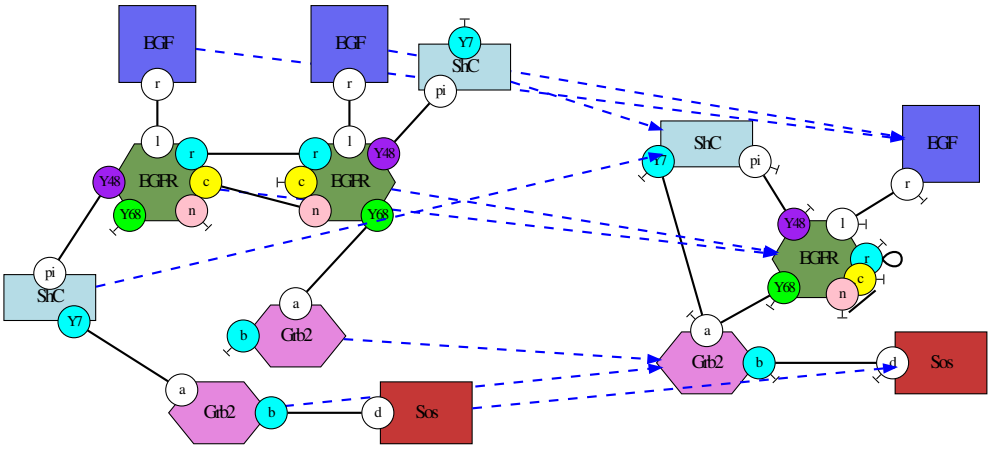


Fig. 8. The unique morphism from the Σ-graph G_Σ and the Σ-graph G_{CM}.

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 an unspecified site, 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). Depending on the choice of the semantics, the state of the system may be described either as a function from bio-molecular compound to concentrations (differential setting), or as a multi-set of bio-molecular compound (stochastic setting).

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.*

We denote as $[P, P']$ the set of the embeddings from a pattern P to a pattern P'.

As opposed to classical notions of embeddings between graphs, embeddings between patterns preserve the freeness of sites.

The composition of two embeddings is an embedding.

Two patterns E and F are said 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. We denote as $E \approx F$ whenever two patterns E and F are isomorphic. 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 patterns. Then we introduce two kinds of graph to reason about this problem.

4.1 Interpretation of a contact map

Intuitively, we want to interpret a contact map 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 in a given model. Indeed with such a definition, each model admitting at least one bio-molecular compound would always admit an infinite number of bio-molecular compounds due to isomorphisms. Thus we consider $\approx$-equivalence class of bio-molecular species instead.

Definition 4.1 (Interpretation of a contact map) *The interpretation of a contact map G_{CM} is defined as the set of the $\approx$-equivalence class of bio-molecular compound $[G]_\approx$ 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:

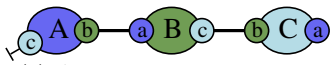
Problem 4.2 *Let G_{CM} be a contact map. We are looking for an automatic prodecure to decide whether the set $[\![G_{CM}]\!]$ is finite, or not.*

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 of a repeatable pattern compatible with the contact.

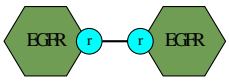
Firstly, we define properly a repeatable pattern as a chain of agents which may be itterated to form arbitraryly long patterns.

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

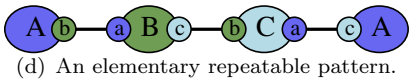


(a) A chain with one site free.

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



(c) A chain with two extremities documenting the same site.



(d) An elementary repeatable pattern.

Fig. 9. Four patterns. Each of them is a chain. But only the last one is repeatable.

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

(ii) *there is an agent with a site free or that documents at most one site (or both);*

(iii) *at most two agents does not have two sites bound.*

In particular, every chain is connected. A chain is formed either of a single agents with at most two sites each of them free, or of a linear chain of agents with exactely two extremities. In the latter case, every agent not in the extremities has two sites and these sites are bound. The agents on the extremity either have exactely one site that is bound. Additionaly, it may have at most one site free.

A chain is a repeatable patterns whenever it contains at least two agents and its extremities may be replug to each other. This is formalised in the following definition.

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

(iv) *both sites documented by the extremities are different.*

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

Example 4.5 *We consider four patterns in Fig. 9. All these patterns are chains. The pattern in Fig. 9(a) is not repeatable because one of its extremity has a free site. The pattern in Fig. 9(b) is not repeatable because its extremities are not of the same kind. The pattern in Fig. 9(c) is not repeatable because its extremities document the same site The pattern in Fig. 9(d) is repeatble (and elementary).*

We can now establish our pumping lemma.

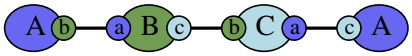
Lemma 4.6 (pumping lemma) *Let G_{CM} be a contact map. Both following assertions are equivalent:*

(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:



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. While 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 of each other site in that sequence of sites.

In the following, 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. We call this graph the graph of the sites of this contact map.

Definition 4.7 (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 that 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 cycles in the graph of the sites to the existence of repeatable patterns.

Theorem 4.8 *Let G_{CM} be a contact map.*

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

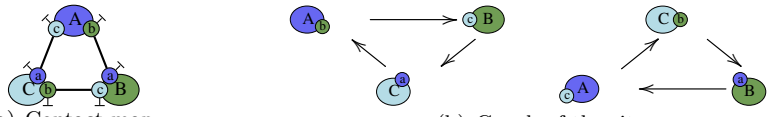
Both following properties are equivalent:

(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 via its site i'.*

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

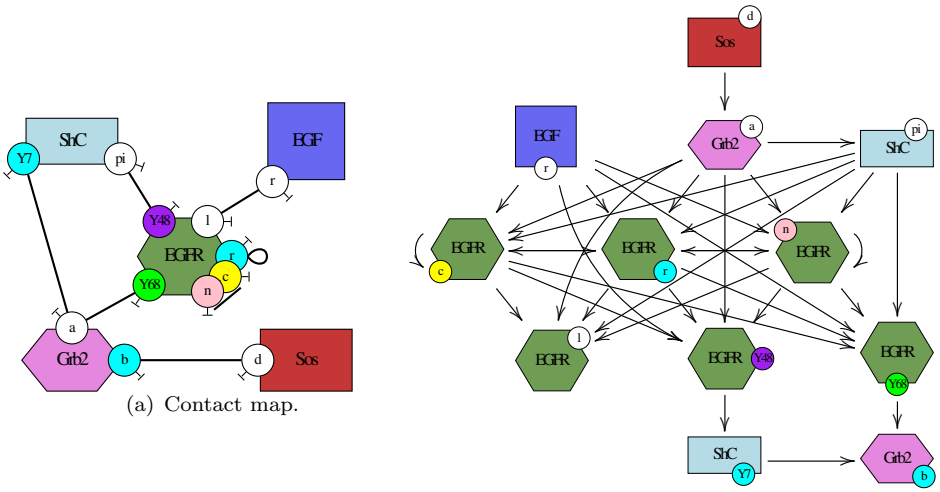
Thus, Thm. 4.8 reduces the problem of deciding whether a contact map is compatible with an infinite number of non-isomorphic bio-molecular compounds to the one of computing the strongly connected components of the graph of the sites of this contact map.

Example 4.9 (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 proteins 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 models to prove that there*



(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 a site connected in the contact map to the site s, in the same kind of protein as the site t but distinct from t.

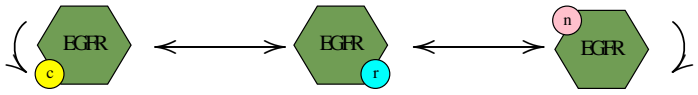


(a) Contact map. (b) Graph of the sites.

Fig. 11. EGFR 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.

is a finite amount of them (the contact map of the model is compatible with an infinite number of them).

Example 4.10 (graph of the sites (egfr 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:*

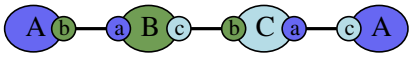


Further constraints are required on the bio-molecular compounds of the models to prove that there is a finite amount of them (the contact map of the model is compatible with an infinite number of them).

4.4 Graph of the potential links

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

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 a of B to the site c of B, from the site c of B to the site b of C, from the site b of C to the site a 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 on repeatable patterns that are compatible with a given contact map. We call this graph the graph of the potential links of this contact map.

Definition 4.11 (Graph of the potential 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 sites 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 $s' = \mathcal{L}_{G_{CM}}(s)$.*
- *$\mathcal{E}$ 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 in the contact map G_{CM} and two different site names i and i' such that $s' = (n, i)$ and $s'' = (n, i')$.*

The condition on the edges of the graph of the potential links ensures that both links may be consecutive in a repeatable pattern.

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

Theorem 4.12 *Let G_{CM} be a contact map.*

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

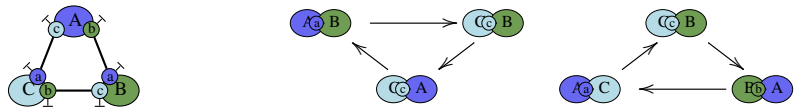
Both following properties are equivalent:

(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 via its site i'.*

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

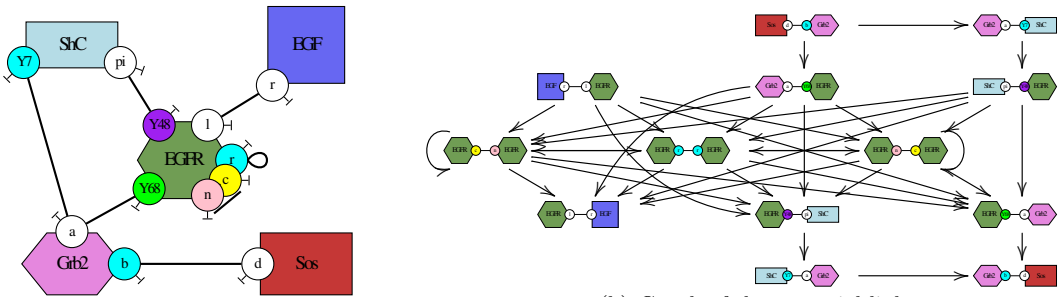
Thus, Thm. **??** reduces the problem of deciding whether a contact map is compatible with an infinite number of non-isomorphic bio-molecular compounds to the one of computing the strongly connected components of the graph of the potential links of this contact map.

Example 4.13 (graph of the potential links (ABC model)) *In Fig. 12, we compute the graph of the potential links for the contact map of the model with three proteins 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 models*



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

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 links).

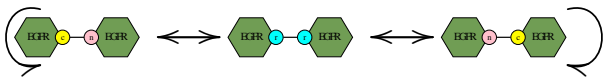


(a) Contact map. (b) Graph of the potential links.

Fig. 13. EGFR 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 links, except for the link between the site r of $EGFR$ and it self, for which there is a unique node. There is an oriented edge between a node s and a node t if and only if the target site of the node s and the source site of the node t are different while belonging to the same kind of protein.

to prove that there is a finite amount of them (the contact map of the model is compatible with an infinite number of them).

Example 4.14 (graph of the potential links (egfr model)) *In Fig.* **??**, *we compute the graph of the potential 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 one non trivial strongly connected component:*



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

4.5 Taking into account the result of a static analysis

In this section, we explain how to refine the graph of the potential links of a gioven contact map, in order to take into the constraints about the reachable bio-molecular compounds that cannot be written in the contact map. These constraints may come from a black box static analysis [13,3] and they take the form of a set of patterns that shall occur in no reachable bio-molecular species.

For instance, the analysis that is described in [13] can infer automatically, from the set of rules that describes a model and its set of initial state, that, in the early events of the integration of the epidermic growth factor, a receptor cannot be bound to two different instances of receptors. Indeed the analysis detects that the following patterns are unreachable:

The analysis that is described in [11] generalises this approach to arbitrary cycles

of proteins. In the example of the triangle, it may infer, when it is a consequence of a rules, that no two As may occur in a given connected compounds, by proving that the following pattern:

is unreachable

5 Conclusion

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, to appear.

[4] Camporesi, F., J. Feret and J. Hayman, *Context-sensitive flow analyses: a hierarchy of model reductions.*, in: A. Gupta and T. Henzinger, editors, *Proc. CMSB'13*, number 8130 in LNCS (2013).

[5] Camporesi, F., J. Feret and K. Q. Lý, *Kade: a tool to compile kappa rules into (reduced) odes models*, in: *Fifteenth International Workshop on Static Analysis and Systems Biology (SASB'17)*, LNCS/LNBI **10545**, supplementary information available at www.di.ens.fr/~feret/CMSB2017-tool-paper.

[6] 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.

[7] Danos, V., J. Feret, W. Fontana and J. Krivine, *Abstract interpretation of cellular signalling networks*, in: F. Logozzo, D. A. Peled and L. D. Zuck, editors, *Proceedings of the Ninth International Conference on Verification, Model Checking and Abstract Interpretation, VMCAI'2008*, Lecture Notes in Computer Science **4905** (2008), pp. 83–97.

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

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

[10] 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).
URL http://dx.doi.org/10.1371/journal.pone.0032032

[11] 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).

[12] 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).

[13] Feret, J. and K. Q. Lý, *Reachability analysis via orthogonal sets of patterns.*, in: *Proc. SASB'16*, ENTCS, to appear.

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

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