Reachability Analysis of Biological Signalling Pathways by Abstract Interpretation

Jérôme Feret[1,2]

École Normale Supérieure
45, rue d'Ulm
75 005 Paris, FRANCE

Abstract. Agent-based formal languages can be used to describe biological signalling networks. As for any language, this process is error prone. Thus we require static analysis tools to check whether the formal description of models matches with what the programmer (or the biologist) has in mind. However, biological networks involve a large number of non-isomorphic complexes (i.e. the number of non-isomorphic ways in which agents can connect), as a consequence static analyses must cope with this combinatorial blow up. We use the abstract interpretation framework, which is a theory of semantics approximation, to design an abstraction of the set of reachable complexes. This abstraction is both accurate and efficient. Then we show several applications. First, we use this abstraction to detect some bugs such as dead reactions (reactions that can never be triggered) and conflicting rules (distinct rules that compute the same thing). Our analysis also predicts whether two sites may bind in any context, or if this binding is controlled by other sites.

Keywords: Biological signalling pathways, κ-calculus, reachable complexes, abstract interpretation.

INTRODUCTION

Biological signalling pathways are natural, large scale, concurrent systems in charge of receiving extra-cellular signals and triggering appropriate responses in the cell (e.g. differentiation, growth, death). They involve multiple proteins, from membrane bound receptors to transcription factors. Many pathologies can be traced back to the subtle regulation of such signalling networks and, as new experimental techniques develop, so does the realisation that a thorough description is key to their understanding and control.[3] This task is very complex. That is due partly to the fact that those networks involve a large number of different agents, and partly, and perhaps more importantly, to the fact that they are *high-dimensional*, meaning that their various participants can combine and modify their internal states in a huge number of non-isomorphic ways.

We model these networks in structured concurrent languages, such as κ [2, 3, 4], or the closely related BioNetGen language [5, 6]. Models being based on rules, not flat reactions, they incorporate *bona fide* biological knowledge, and are easier to discuss and update. Nevertheless the modeling process might be error prone. Thus we require static analysis tools to help proving whether the formal description of models matches with what the programmer (or the biologist) has in mind. Such static analyses must cope with this combinatorial blow up. We use the abstract interpretation framework [7], which is a theory of semantics approximation, to design an abstraction of the set of reachable complexes. This abstraction is both accurate and efficient. We can use this abstraction to detect some bugs in the early phase of modelling. Indeed, our analysis detects some dead rules which are reactions that can never be triggered. Our analysis also checks whether rules that perform the same thing cannot be applyied with the same complexes (which would have no meaning as far as quantitative properties are considered). Last, our analysis provides useful information about the usage of binding sites. We can use our abstraction to simplify rules, in order to detect whether the binding between two sites is independent from the other sites in complexes, or whether this binding is controlled by some specific sites.

---

[1] This research has been done within the INRIA ABSTRACTION project-team (common to the CNRS and the ÉNS).

[2] These works have been done while the author was visiting Plectix Biosystems Inc. They are the result of a collaboration with Vincent Danos, Walter Fontana and Jean Krivine.

[3] A recent study explains how new mechanistic details of HER-2 signalling show that some inhibitors, currently in clinical trial, cannot work [1].

THE κ-CALCULUS

We give in this section an informal description of the κ-calculus. A more complete description of this calculus may be found in [2, 3, 4].

In the κ calculus, the state of the system is described by a solution $S = A_1(\overline{x_1}), \ldots, A_n(\overline{x_n})$ that is a (multi-)set of agents. Each agent $A(\bar{x})$ is described by a name A and an interface $\bar{x}$. This interface denotes the states of agent sites. This way, the interface is a set of site names stamped with some information. These information may encode the internal state of sites (such as activity levels) and bindings between sites. The notation $x{\sim}i!b$ denotes a site x with the internal state i and the binding state b. A site may have no internal state. A site without binding state is free. Last, binding states are integers that denote bindings. Sites having the same binding states are bound pairwise. A connected component in a solution is called complex.

For instance, the following solution:

```
EGF(r!1),EGFR(l!1,r!2,Y1148~p),EGFR(r!2,l!3),EGF(r!3)
```

describes a dimer: two receptors (`EGFR`) are bound with some ligands (`EGF`); both receptors are bound together; the site `Y1148` of one of the two receptors is phosphorylated.

The potential evolution of the main solution is described by rewriting rules. A rule is defined by a left hand side and a right hand side which are both solutions. The left hand side describes some conditions that allow the application of the rule with a part of the main solution, whereas the right hand side describes the result of the rule application.

For instance, the following rules:

```
EGF(r),EGFR(l) -> EGF(r!1),EGFR(l!1)
EGF(r!1),EGFR(l!1,r),EGFR(l!2,r),EGF(r!2) -> \
    EGF(r!1),EGFR(l!1,r!3),EGFR(l!2,r!3),EGF(r!2)
EGF(r!1),EGFR(l!1,r!3,Y1148~u),EGFR(l!2,r!3),EGF(r!2) -> \
    EGF(r!1),EGFR(l!1,r!3,Y1148~p),EGFR(l!2,r!3),EGF(r!2)
```

allow the binding of ligands `EGF` with `EGF` receptors, the dimerisation of `EGF` receptors, the cross-phosphorylation of the site `Y1148` of a `EGF` receptor by another `EGF` receptor.

We notice that in the rules, the agent interfaces do not need to be fully specified. Whenever a site is not specified in a rule, it means that this rule may apply whatever the state of the site is. Rules may be fit with kinetics rates.

VIEWS

A biological network can be described by an initial solution and a set of rules. In order to debug such a network, we would like to use the set of reachable complexes (i.e. the set of all complexes that may be built from the initial solution by applying the rules (as many times as desired). Nevertheless, due to combinatorial blow up, the computation of this set may be very costly (this set might even be infinite). So we use the abstract interpretation framework [7] which is a theory of the approximation of semantics, to compute efficiently an approximation of the set of reachable complexes.

Our approximation is inspired from the one we proposed earlier for π-calculus [8]. We first forget whole complexes. In contrast, we keep local information about each agent. Given an agent $A(\bar{x})$ in a solution S, the view associated with the agent $A(\bar{x})$ is obtained by replacing any binding state $!i$ with the type $!B.y$ of the binding, where B and y are respectively the name of the agent and the site to which the site of agent A is bound. Moreover, in order to get a finite abstraction, we forget the number of views.

This way, the following solution:

```
EGF(r!1),EGFR(l!1,r!2,Y1148~p),EGFR(r!2,l!3,Y1148~u),EGF(r!3)
```

is abstracted into the following one:

```
EGF(r!EGFR.l),EGFR(l!EGF.r,r!EGFR.r,Y1148~p),EGFR(r!EGFR.r,l!EGF.r,Y1148~u).
```

Rewriting rules can be easily lifted at the level of views. This allows the computation of a super-set[4] of the views that may occur in the system during any transition sequence. We notice that since we do not take into account the number of instances of views, a rule application may never remove a view: it can only build new views.

APPLICATIONS

Now we describe some applications for our analysis. All these applications can be done fully automatically. They all depend on the reachability analysis that we presented in the previous section: the more accurate this reachability analysis is, the more powerful these applications are.

Dead rule detection

Having a super-set of the reachable views, we can detect some rules that can never be triggered because their left hand side involves an unreachable view. For instance, the following rule:

```
EGFR(Y1148~u!1),SHC(SH2!1) -> EGFR(Y1148~u),SHC(SH2)
```

is a dead rule, because each time the site `Y1148` of a receptor is bound, then it is also phosphorylated. These information can be read from views: our super-set of views contains no view in which the site `Y1148` is both bound and not phosphorylated.

Due to the approximation, we may not detect all dead rules. Nevertheless, we know that all the rules that are detected dead, are effectively dead.

Conflicting rules detection

It might happen that several rules perform the same action. However such rules should not be applied with the same solution part, because it would be meaningless as long as kinetics is considered. Moreover, such situation is very likely to come from a mistake in the model encoding. For instance, the following rules:

```
EGFR(Y1148~p),SHC(SH2) -> EGFR(Y1148~p!1),SHC(SH2!1)
EGFR(Y1148,Y1068~p),SHC(SH2) -> EGFR(Y1148!1,Y1068~p),SHC(SH2!1)
```

are conflicting because, they can both be applied to the solution `EGFR(Y1148~p,Y1068~p),SHC(SH2)` to produce the solution `EGFR(Y1148~p!1,Y1068~p),SHC(SH2!1)`. In this case, it is obviously a bug, since there is no reason why the site `Y1068` should control the binding of the site `Y1148` in the second rule.

A super-set of views can be used to detect which pairs of rules, among those that perform a given action, can be applied with the same solution part. Due to the approximation, our analysis may warn about conflicting rules whereas the complexes with which they can both apply are not reachable. We stress out the fact that without a reachable analysis (that is, by considering that any writable complexe is reachable), we would have too many spurious warnings.

Rule decontextualization

In some rewriting rules, left hand sides can be simplified without modifying the properties of the network. There may be two trade-offs according to whether we are interested in qualitative or quantitative properties.

For instance, the following rule:

```
EGF(r!1),EGFR(l!1,r!3),EGFR(l!2,r!3),EGF(r!2) -> \
    EGF(r!1),EGFR(l!1,r),EGFR(l!2,r),EGF(r!2)
```

[4] Due to the approximation, we can only be sure that we get all actual views, however there might be fictitious views introduced because of the abstraction.

may conservatively be replaced with the following one:

```
EGFR(r!1),EGFR(r!1) -> EGFR(r),EGFR(r)
```

in any model where each `EGF` receptor is bound to a ligand `EGF` whenever its dimerisation site is bound. This transformation preserves the quantitative properties of the system: we call it the *quantitative* decontextualization.

Whenever we are only interested in qualitative properties, we can make further simplifications. For instance, the following rules:

```
EGFR(Y1068~p!1),Grb2(SH2!1,SH3),Sos(d) -> \
    EGFR(Y1068~p!1),Grb2(SH2!1,SH3!2),Sos(d!2)
Grb2(SH2,SH3),Sos(d) -> Grb2(SH2,SH3!1),Sos(d!1)
SHC(Y317~p!2),Grb2(SH2!2,SH3),Sos(d) -> \
    SHC(Y317~p!2),Grb2(SH2!2,SH3!1),Sos(d!1)
```

may be replaced with the following single rule:

```
Grb2(SH3),Sos(d) -> Grb2(SH3!1),Sos(d!1)
```

because we can prove that the rules cover all application cases with some reachable complexes. This means that both free sites `SH3` in the protein `Grb2` and `d` in the protein `Sos` may always bind together whatever the states of the complexes they belong to are. Nevertheless, the quantitative properties are not the same, if kinetics rates are different.

This *qualitative* decontextualization may fail to remove some sites for two reasons. One reason is that the reachability analysis may not be accurate enough: this happens when the analysis introduces fictitious complexes that may not be taken into account into reaction rules. The other reason is that some sites may control the binding of others. In this case, the decontextualization computes a super-set of the sites which control (qualitatively) the modification of other sites. This allows the programmer to check that his knowledge matches with the model he has encoded. For instance, in the following rule:

```
EGF(r!1),EGFR(l!1,r),EGFR(l!2,r),EGF(r!2) -> \
    EGF(r!1),EGFR(l!1,r!3),EGFR(l!2,r!3),EGF(r!2)
```

the ligands cannot be removed, which suggests that the dimerisation is controlled by the binding with ligands.

CONCLUSION

We have introduced an abstract interpretation-based static analysis for approximating the set of reachable complexes in biological networks. This abstraction collects the potential states of each protein, abstracting away the global structure of complexes. This analysis is useful in the early stages of modelling process, since it allows the debugging of networks: it detects rules which can never be triggered; it detects conflicting rules (that perform the same actions over the same complexes). It also provides useful information about bindings (whether a binding formation depends on other sites, or not).

REFERENCES

1. N. V. Sergina, M. Rausch, D. Wang, J. Blair, B. Hann, K. M. Shokat, and M. M. Moasser, *Nature* (2007), URL `http://www.nature.com/nature/journal/vaop/ncurrent/full/nature05474.html`.
2. V. Danos, and C. Laneve, "Core Formal Molecular Biology," in *Proc. ESOP'03*, Springer-Verlag, 2003, vol. 2618 of *LNCS*.
3. V. Danos, and C. Laneve, "Graphs for Formal Molecular Biology," in *Proceedings of the First International Workshop on Computational Methods in Systems Biology, CMSB'03*, Springer-Verlag, 2003, vol. 2602 of *LNCS*, pp. 34–46.
4. V. Danos, and C. Laneve, *Theoretical Computer Science* **325**, 69–110 (2004).
5. J. Faeder, M. Blinov, and W. Hlavacek, *Proc. ACM Symp. Appl. Computing* pp. 133–140 (2005).
6. J. Faeder, M. Blinov, G. B., and W. Hlavacek, *Complexity* **10**, 22–41 (2005).
7. P. Cousot, and R. Cousot, "Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints," in *Proc. POPL'77*, ACM Press, Los Angeles, California, 1977, pp. 238–252.
8. J. Feret, "Dependency analysis of Mobile Systems," in *In proc. ESOP'02*, LNCS 2305, Springer-Verlag, 2002.