Integrating time-series data on large-scale cell-based models: application to skin differentiation

Louis Fippo Fitime[1], Andrea Beica[1], Olivier Roux[1] and Carito Guziolowski[1]

[1] LUNAM Université, École Centrale de Nantes, IRCCyN UMR CNRS 6597
(Institut de Recherche en Communications et Cybernétique de Nantes)
1 rue de la Noë – B.P. 92101 – 44321 Nantes Cedex 3, France.
Louis.Fippo-Fitime@irccyn.ec-nantes.fr

Abstract

The way living organisms work and develop themselves is controlled by large and complex networks of genes, proteins, small molecules, and their interactions, called biological regulatory networks. Confronting time-series gene expression data with models may allow us to examine and characterize the dynamics of elements that compose such regulatory networks. In this work, we propose a way to model and simulate large-scale regulatory networks, by using the Process Hitting (PH) framework, in order to verify if the model can predict the experimental measures. The preliminary work presented here proposes: (1) a semi-automatic method to build a PH from a regulatory network of biochemical reactions, (2) a discretization scheme of the continuous time-series measurements, and (3) an approach to estimate the PH stochastic simulation parameters in an unbiased manner.

1 Introduction

The comprehension of the mechanisms involved in the regulation of a living cell is a fundamental issue. These mechanisms can be modeled as biological regulatory networks, which analysis requires to preliminary build a mathematical or computational model. By just considering qualitative regulatory effects between components, biological regulatory networks depict fairly well biological systems, and can be built upon public repositories such as the Pathways Interaction Database [8], and hiPathDB[10] for human regulatory knowledge.

This work aims to propose a dynamical model of large-scale systems based on the formal integration (complete validation/invalidation) of high-throughput experimental time-series data. So far this idea has been addressed separately by approaches that either: (a) focus first on modeling at small-scale the system and then on refining or improving it through the fitting with some data points, such as methods based on differential equations [9, 1, 6], or (b) integrate in an efficient and complete fashion large-scale models and high-throughput data regardless of the system dynamics [3, 5], or (c) fit dynamical data to middle-scale networks using stochastic approaches, and therefore without the guarantee of finding global optima [4]. Therefore, with this work we intend to fill the gaps between the previously cited methodologies and converge to a more realistic model of biological behavior.

For modeling and analyzing the biological system we rely on the Process Hitting (PH) framework[7], since it is especially useful for studying systems composed of biochemical interactions, and provides stochastic simulation as well as efficient static methods to model dynamical properties of the system. The PH framework uses qualitative and discrete information of the system, without requiring enormous parameter estimation tasks for its stochastic simulation. So far, this method has been successfully demonstrated only on very well-known systems and without exploiting high-throughput measures. We believe, however, that the use of high-throughput data has become unavoidable with the advent of massive, publicly available data sets in the form of well-standardized DNA microarray data and, more recently, in the form of phosphoproteomics data.

The main methodological and preliminar results of this work are: (i) semi-automatic PH generation from a biological system composed of biochemical reactions, extracted from public databases; (ii) discretization approach of time-series expression data, so we can reproduce these traces by using in a first attempt the PH stochastic simulation, and afterwards perform static reachability analyses to satisfy these data; and (iii) estimation of the the temporal and stochastic parameters of the simulation, based on statistical analyses of the full-compendium of time-series expression data. The biological system used as a case-study for this work is a cell-based model of skin differentiation, which is of key importance in wound healing.

2 Methods and data

2.1 The Process Hitting Framework

Process Hitting (PH) gathers a finite number of concurrent processes grouped into a finite set of sorts. A sort stands for a component of a biological system while a process, which belongs to a unique sort, corresponds to a unique state of the system components(sorts). At any time, exactly one process of each sort is present. A state of the PH corresponds to such a set of processes. We denote here a process by a_i where a is the sort and i is the process identifier within the sort a. The concurrent interactions between processes are defined by a set of *actions*. Actions describe the replacement of a process by another of the same sort conditioned by the presence of at most one other process in the current state. An action is denoted by $a_i \rightarrow b_j \Rsh b_k$, which is read as "$a_i$ *hits* b_j to make it bounce to b_k", where a_i, b_j, b_k are processes of sorts a and b, called respectively *hitter*, *target* and *bounce* of the action.

Definition 1 (Process Hitting) *A* Process Hitting *is a triple* $(\Sigma, L, \mathcal{H})$*, where:*

- $\Sigma = \{a, b, \dots\}$ *is the finite set of* sorts*;*
- $L = \prod_{a \in \Sigma} L_a$ *is the set of states with* $L_a = \{a_0, \dots, a_{l_a}\}$ *the finite set of* processes *of sort* $a \in \Sigma$ *and* l_a *a positive integer, with* $a \neq b \Rightarrow L_a \cap L_b = \emptyset$*;*
- $\mathcal{H} = \{a_i \rightarrow b_j \Rsh b_k \in L_a \times L_b \times L_b \mid (a, b) \in \Sigma^2 \wedge b_j \neq b_k \wedge a = b \Rightarrow a_i = b_j\}$ *is the finite set of* actions.

Given a state $s \in L$, the process of sort $a \in \Sigma$ present in s is denoted by $s[a]$. An action $h = a_i \rightarrow b_j \Rsh b_k \in \mathcal{H}$ is *playable* in $s \in L$ if and only if $s[a] = a_i$ and $s[b] = b_j$. In such a case, $(s \cdot h)$ stands for the state resulting from the play of the action h in s, with $(s \cdot h)[b] = b_k$ and $\forall c \in \Sigma, c \neq b, (s \cdot h)[c] = s[c]$.

Modeling cooperation. As described in [7], the cooperation between processes to make another process bounce can be expressed in PH by building a *cooperative sort*. Fig. 1 shows an example of a cooperative sort ab between sorts a and b, defined with 4 processes (one for each sub-state of the presence of processes a_1 and b_1). For the sake of clarity, processes of ab are indexed using the sub-state they represent. Hence, ab_{01} represents the sub-state $\langle a_0, b_1 \rangle$, and so on. Each process of sort a and b hit ab, which makes it bounce to the process reflecting the status of the sorts a and b (e.g., $a_1 \rightarrow ab_{00} \Rsh ab_{10}$ and $a_1 \rightarrow ab_{01} \Rsh ab_{11}$). Then, to represent the cooperation between processes a_1 and b_1, the process ab_{11} hits c_1 to make it bounce to c_2 instead of independent hits from a_1 and b_1. The same cooperative sort is used to make a_0 and b_0 cooperate to hit c_1 and make it bounce to c_0. Cooperation can be used to model protein-complex biochemical reaction. For instance a molecule a that cooperates with a molecule b to activate a molecule c, Fig. 1 (left), We model this interaction by four sorts Fig. 1 (right) a, b, c and ab. Sorts a, b and c represent components a, b and c. We introduce the cooperative sort ab to characterize constraints on components a and b. Cooperation can be a way to model protein-complex formation.

Example 1 *Fig. 1 represents a PH* $(\Sigma, L, \mathcal{H})$ *with* $\Sigma = \{a, b, c, ab\}$*, and:*

$$
\begin{aligned}
L_a &= \{a_0, a_1\}, & L_b &= \{b_0, b_1\},\\
L_{ab} &= \{ab_{00}, ab_{01}, ab_{10}, ab_{11}\}, & L_c &= \{c_0, c_1, c_2\}.
\end{aligned}
$$

This example models a Biological Regulatory Network (BRN) where the component c has three qualitative levels, components a and b are Boolean and ab is a cooperative sort. In this BRN, ab inhibits c at level 2 through the cooperative sort ab (e.g.$ab_{00} \rightarrow c_2 \Rsh c_1$, $ab_{00} \rightarrow c_1 \Rsh c_0$) while a and b activate c through the cooperative sort ab (e.g. $ab_{11} \rightarrow c_0 \Rsh c_1$ $ab_{11} \rightarrow c_1 \Rsh c_2$). Indeed, the reachability of c_2 and c_0 is conditioned by a cooperation of a and b, as explained above.

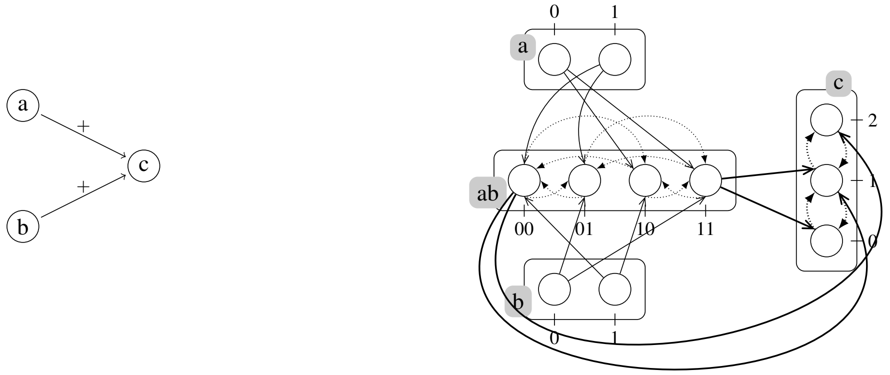


Figure 1: (left) Biological pattern example. Nodes are components and edges are interactions For instance, components a and b cooperate to activate c. (right) equivalent PH model. A PH example with four sorts: three components (a, b and c) and a cooperative sort (ab). Actions targeting processes of c are in thick lines.

2.2 Time-series microarray data

To illustrate our approach, we used the time series microarray data from calcium stimulated keratinocyte cells measured at 10 time-points. 200 transcripts were selected for their dynamic patterns, that is, their fold expression with respect to the non-stimulated cell was significant in at least one time point. We included in our model a subset of 12 of them: MKP3, MKP1, UPAR, HES5, ILB1, A20, SM22, IL8, ET1, TNF-a, TFR, DKK1. This subset was selected because we were able to automatically retrieve the regulatory mechanisms upstream of these 12 genes from public repositories of biochemical reactions. The full dataset (data not shown) was produced by the German Cancer Research Center (DKFZ) and is currently in the process of getting published.

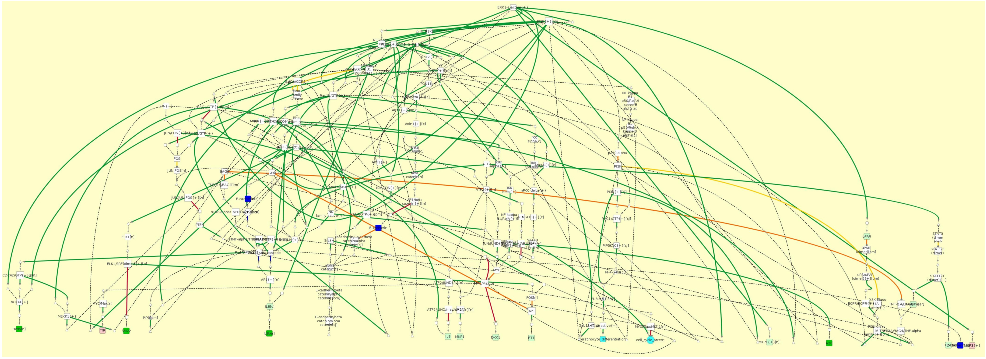


Figure 2: **RSTC network**

2.3 Interaction network

The interactions of the studied biological system were represented in a RSTC network, which stands for multi-layer receptor-signaling-transcription-cell state network, generated from the Pathway Interaction Database (PID). In order to build this network, we selected a set of seed nodes related to the biological process studied. The seed nodes for our case study were: (1) E-cadherin, which is a protein having Ca binding domains and which plays an important role in cell adhesion; (2) the 12 significantly differentially expressed genes accross the 10 time-points; and (3) the cell states of keratinocytes-differentiation and cell-cycle-arrest. The network was extracted automatically from the whole content of the NCI-PID database by using a subgraph algorithm to link the seed nodes[2]. Fig.2 shows the RSTC network obtained.

3 Results

3.1 Modeling the RSTC network as a PH model

In order to model the RSTC network with a PH model we selected known biological regulatory patterns (atomic set of biological components and their interacting roles), represented as biochemical reactions in the RSTC network, and proposed their PH representation. For instance a molecule a that cooperates with a molecule b to activate a molecule c, Fig. 1 (left), is a regulatory pattern because it is a protein-complex biochemical reaction that appears recurrent times. We model this pattern by four sorts Fig. 1 (right) a, b, c and ab. Sorts a, b and c represent components a, b and c. We introduce the cooperative sort ab to characterize constraints on components a and b. In our RSTC network, we found 11 regulatory patterns (see Appendix 4).

3.2 Integrating time-series gene expression data

3.2.1 Discretizing times-series data

Because PH simulation is discrete we need to discretize continuous experimental data, so we can compare our simulation outputs. The goal of this method was to better determine, according to the gene expression level, when a given molecule is activated or inhibited. To do this, we introduced the new analog concept of Significant Increase or Decrease to characterize the fact that a level of a molecule increases or decreases when crossing a threshold of significance; We limited the possible expression levels for a molecule to $\{0, 1, 2\}$. Algorithm 1 underlines the main steps of the proposed discretization method. For more details about the functions used in this algorithm see Appendix 4

Algorithm 1 Discretization of experimental data

```
Require: $X$ a table of experimental data
Ensure: $Y$ a table of discretized data
  for all gene $i$ in $X$ do
    $threshold \leftarrow computeThreshold(X[i,])$;
    $Y[i,0] \leftarrow initialState(threshold, X[i,])$;
    for all $j$ in $numberExpression$ do
      if $Increase(X[i,j], X[i,j+1])$ then
        $computeSignificativityOfIncrease(threshold, X[i,j], X[i,j+1])$;
        $fixSTATE(Y[i,j], Y[i,j+1])$;
      else
        $computeSignificativityOfDecrease(threshold, X[i,j], X[i,j+1])$;
        $fixSTATE(Y[i,j], Y[i,j+1])$;
      end if
    end for
  end for
```

To illustrate the result of the discretization algorithm 1 we plot in Fig. 3 the expression of the TFRC and IL8 genes from the times-series data with their respective discrete plots. On the discrete plot, one can clearly differentiate when a molecule is active or not, which is of extreme importance when modeling these steps in the PH framework since we want to have coherent simulation results.

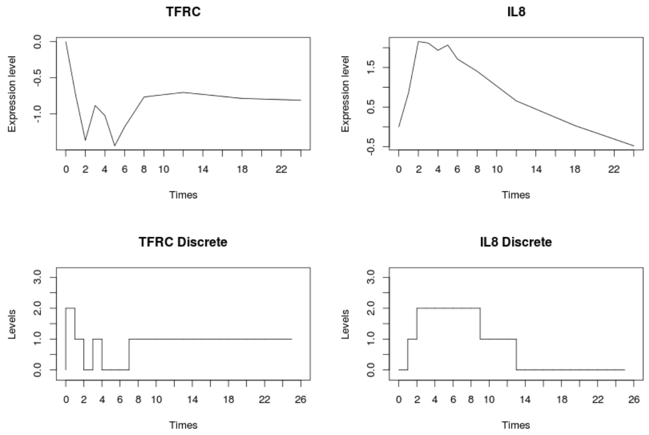


Figure 3: Illustration of discretisation of Experiment Data

3.2.2 Estimating the parameters for the PH-simulation

The simulation of the execution of the PH actions is done stochastically. Therefore, we need to relate each action with temporal and stochastic parameters, introduced into the PH framework to achieve dynamic refinement [7]. This is an important aspect of the modeling when taking into account the temporal and stochastic dimensions of biological reactions by performing simulations. On the one hand, we consider the probability of a reaction to occur, and on the other hand, we consider stochastic parameters in the aim at observing an expected behavior. In the PH framework, to play an action we need two essential parameters: the rate r or the temporal parameter because $t = r^{-1}$ and the stochasticity absorption sa. These two parameters will be estimated according to the expression profile of time series data of the experiment. To avoid overfitting in the estimation of these parameters, we propose that each component of the PH, representing a measured gene in the network, will take the estimated values of the parameters of its respective cluster in the experimental data.

1. The first step is to cluster the data set. The goal of the clustering process is to partition the genes into groups such that the profiles contained in the same group (cluster) are similar to each other and as different as possible of the profiles assigned to the other clusters. The particularity here is to choose the best clustering criteria.

2. For each cluster obtained in the previous step, estimate the value of r and sa associated to the cluster.

3. For each component of the PH model associated to the measured gene, determine its cluster, and assign it the previously estimated parameters, r and sa.

In our time-series data, the components of the PH which need to be associated specific parameters (step 3) are the 12 genes present in our RSTC network.

3.3 PH code generation

To simulate of the model, we generated a PINT code to be simulated by the PINT simulator[1]. For the PINT code generation we first list all the selected patterns in the biological reaction into a file. In this file, each line contains the name of the nodes belonging to the current reaction and the reaction type number. The list was then parsed, line by line and, after renaming the nodes using numbers (for readability and in conformity with the PINT language syntax) the corresponding PINT code for the PH process equivalent to each reaction was generated. This was implemented in the Java programming language.

4 Conclusions

This work describes the preliminary steps towards the integration of time-series data in large-scale cell-based models. We proposed a semi-automatic method to build a PH from a biological system composed of biochemical reactions, extracted automatically from public databases, relevant to keratinocyte stimulation induced by calcium. We then proposed a method to discretize time-series gene expression data, so they can be confronted to the PH simulations and logically explained by the PH static analysis. Finally we described a method to automatically estimate the temporal and stochastic parameters for the PH simulation, so this estimation process will not be biased by overfitting. As concrete perspectives of this work, we intend to *(i)* validate the RSTC network topology by confronting its *in-silico* simulation with real measurements of its components; *(ii)* compare the stochastic simulation results with reachability static analysis over the same PH components mapped to the 12 measured genes; and finally *(iii)* search for key-regulators up-stream the 12 genes which will control the dynamics of the system, to provide our biologist partners concrete hypotheses to test experimentally.

[1] Available at `http://process.hitting.free.fr`

References

[1] Grégory Batt, Delphine Ropers, Hidde De Jong, Johannes Geiselmann, Radu Mateescu, Michel Page, and Dominique Schneider. Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in escherichia coli. *Bioinformatics*, 21(suppl 1):i19–i28, 2005.

[2] Carito Guziolowski, Aristotelis Kittas, Florian Dittmann, and Niels Grabe. Automatic generation of causal networks linking growth factor stimuli to functional cell state changes. *FEBS Journal*, 279(18):3462–3474, 2012.

[3] Carito Guziolowski, Santiago Videla, Federica Eduati, Sven Thiele, Thomas Cokelaer, Anne Siegel, and Julio Saez-Rodriguez. Exhaustively characterizing feasible logic models of a signaling network using answer set programming. *Bioinformatics*, 29(18):2320–2326, 2013.

[4] Aidan MacNamara, Camille Terfve, David Henriques, Beatriz Peñalver Bernabé, and Julio Saez-Rodriguez. State–time spectrum of signal transduction logic models. *Physical Biology*, 9(4):045003, 2012.

[5] Alexander Mitsos, Ioannis N Melas, Paraskeuas Siminelakis, Aikaterini D Chairakaki, Julio Saez-Rodriguez, and Leonidas G Alexopoulos. Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS computational biology*, 5(12):e1000591, 2009.

[6] Mohammad Mobashir, Burkhart Schraven, and Tilo Beyer. Simulated evolution of signal transduction networks. *PloS one*, 7(12):e50905, 2012.

[7] Loïc Paulevé, Morgan Magnin, and Olivier Roux. Refining dynamics of gene regulatory networks in a stochastic π-calculus framework. In *Transactions on Computational Systems Biology XIII*, pages 171–191. Springer, 2011.

[8] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl 1):D674–D679, 2009.

[9] John J Tyson, Katherine C Chen, and Bela Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current opinion in cell biology*, 15(2):221–231, 2003.

[10] Namhee Yu, Jihae Seo, Kyoohyoung Rho, Yeongjun Jang, Jinah Park, Wan Kyu Kim, and Sanghyuk Lee. hipathdb: a human-integrated pathway database with facile visualization. *Nucleic acids research*, 40(D1):D797–D802, 2012.

Appendix A

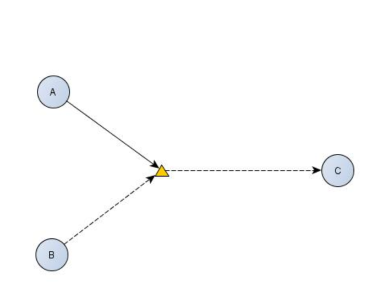
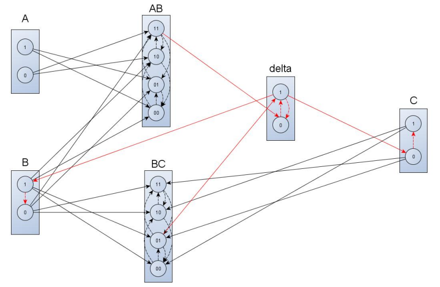
Biological Pattern no 1	PH Model
	

Figure 4: (left) Biological pattern: Molecules A and B cooperate to activate molecule C. After the activation of C, A remains active and B is desactivated. (right) equivalent PH model. AB and BC are regular sorts, while the sort delta models the reaction beginning or end.

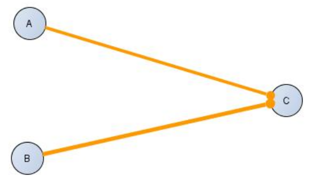
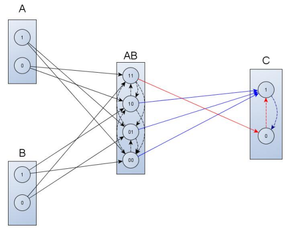
Biological Pattern no 2	PH Model
	

Figure 5: (left) Biological pattern: A and B cooperate to activate C. Both A and B remain active after end of reaction (right) equivalent PH model

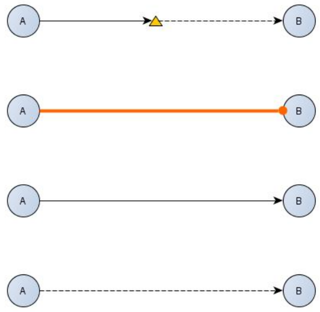
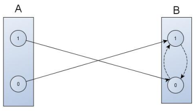
Biological Pattern no 3	PH Model
	

Figure 6: (left) Biological pattern: different types of activation. (right) equivalent PH model

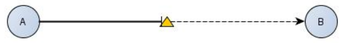
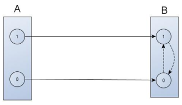
Biological Pattern no 4	PH Model
	

Figure 7: (left) Biological pattern of an inhibition reaction: the inhibitor presence leads to the desactivation of its target, while its absence leads to the activation of the target (right) equivalent PH model

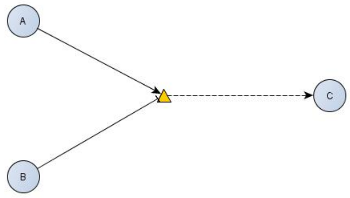
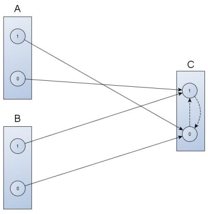
Biological Pattern no 5	PH Model
	

Figure 8: (left) Biological pattern. Molecule C is either activated by A, or inhibited by B; (right) equivalent PH model where A and B are not cooperating to modify C, each one has independent, opposite action on C.

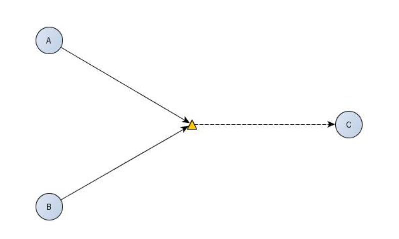
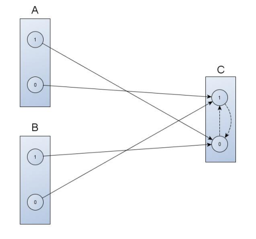
Biological Pattern no 6	PH Model
	

Figure 9: (left) Molecule C is activated by either A, or B, independantly one from other. (right) equivalent PH model

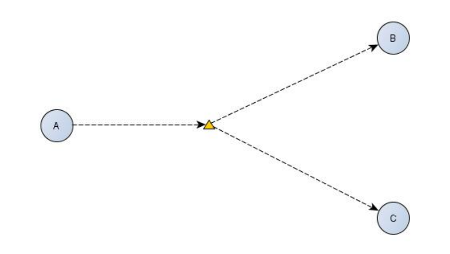
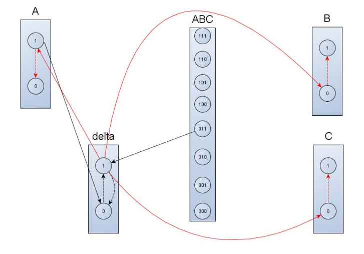
Biological Pattern no 7	PH Model
	

Figure 10: (left) Complex A decomposes in components B and C. At the end of the reaction, A no longer exists/ is no longer active. (right) equivalent PH model. ABC is a regular cooperative sort and delta models the reaction, as explained in Pattern 1. For clarity purposes, the hits from A, B and C to the cooperative sort ABC have not been drawn.

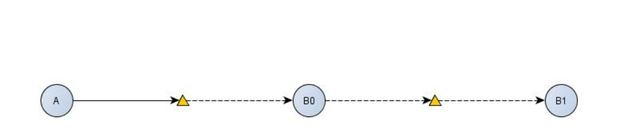
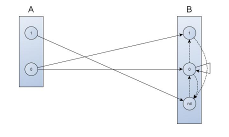
Biological Pattern no 8	PH Model
	

Figure 11: (left) $B0$ and $B1$ represent the same biological entity. (right) equivalent PH model, $B0$ and $B1$ are different process of the same sort; A create B, which then activates itself.

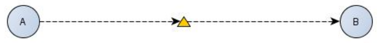
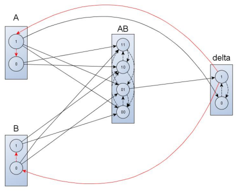
Biological Pattern no 9	PH Model
	

Figure 12: (left) A modification reaction: A activate B, then dissapears; The reaction begins when A is present, and ends when A has replaced by B. (right) equivalent PH model, AB is a cooperative sort and the delta sort models the reaction.

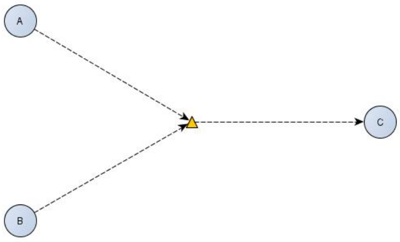
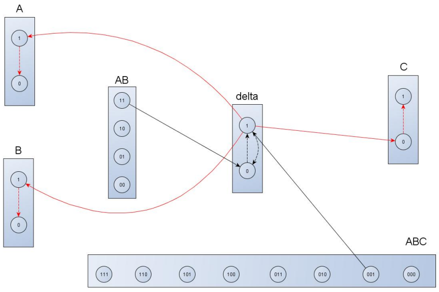
Biological Pattern no 10	PH Model
	 A AB delta C B ABC 1 0 11 10 01 00 111 110 101 100 011 010 001 000

Figure 13: (left) A composite modification: A and B cooperate to create C, then disappear. (right) equivalent PH model. For clarity purposes, hits to cooperative sorts have not been drawn.

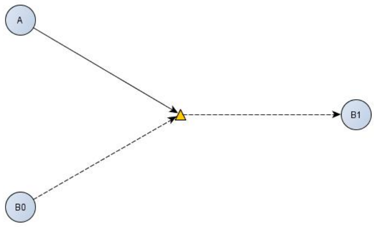
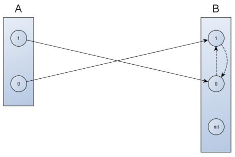
Biological Pattern no 11	PH Model
	

Figure 14: (left) Activation of non-binary sort: similar to Pattern 1, except for the non-binarity of the target source. $B0$ and $B1$ represent the same entity. Unlike pattern 8 (the other pattern dealing with non-binary sorts), entity B is already present, via the condition on $B0$, it just needs to be activates. (right) equivalent PH model.

Appendix B

Functions	Specifications
computeThreshold(X)	compute the threshold of the profile of expression represent by X
initialState(X)	fixe the initial state of the expression represent by X according to the initial value of X and the threshold
Increase(X,Y)	Test if the measure increases between the two times points X and Y
computeSignificativityOfIncrease(s,X1,X2)	compute the significance of the increase according to the threshold and X1 and X2
computeSignificativityOfDecrease(s,X1,X2)	compute the significance of the decrease according to the threshold and X1 and X2
fixSTATE(X1, X2)	fix the current state

Figure 15: Functions(first column) and Specifications(second column)

Appendix C

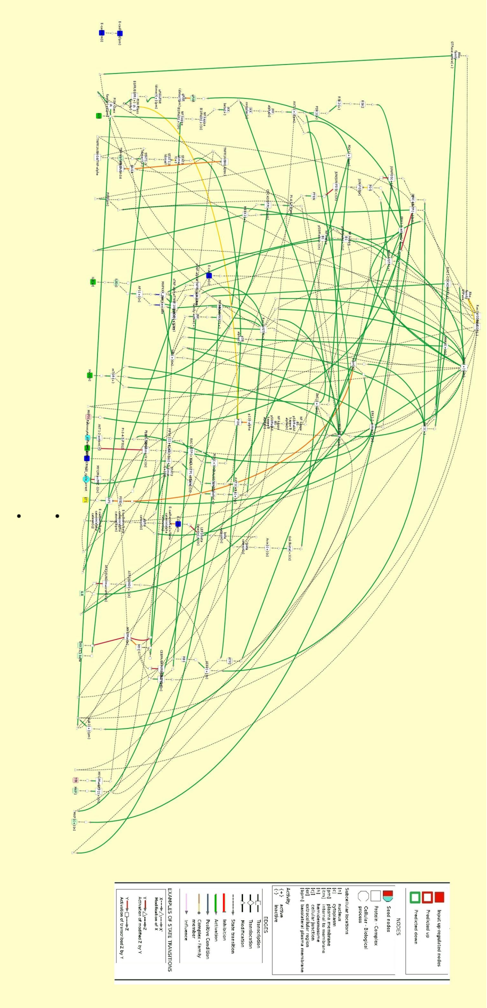


Figure 16: RSTC Network