Regulatory networks

Gilles Bernot[1], Jean-Paul Comet[1], Christine Risso – de Faverney[2]

(1) University of Nice-Sophia Antipolis, I3S laboratory, UMR 6070 CNRS, B.P. 121, 06903 Sophia Antipolis, France

(2) University of Nice-Sophia Antipolis, ECOMERS laboratory, EA 4228 Parc Valrose, 06108 Nice Cedex, France

i. Summary/Abstract

The usefulness of mathematical models for the biological regulatory networks relies on the *predictive* capability of the models, in order to suggest interesting hypotheses and suitable biological experiments. All mathematical frameworks dedicated to biological regulatory networks must manage a large number of *abstract parameters*, which are not directly measurable in the cell. The cornerstone to establish predictive models is the *identification* of the possible parameter values. Formal frameworks involve *qualitative models* with discrete values and *computer aided logic reasoning*. They can provide the biologists with an automatic identification of the parameters *via* a formalization of some biological knowledge into temporal logic formulas. For pedagogical reasons, we focus on gene regulatory networks and develop a qualitative model of the detoxification of benzo[a]pyrene in human cells to illustrate the approach.

ii. Key Words

Biological regulatory networks, Gene regulatory networks, Mathematical modeling, Systems biology, Temporal logic, Model checking, Benzo[a]pyrene, Detoxification pathway, CYP, Metabolizing enzymes.

1. Introduction

Almost all the difficult questions that involve *biological systems*, with several interacting entities, need mathematical models and computer aided reasoning in order to predict the global behavior of the system, or to establish some characteristics of the dynamics of the system. Domain oriented *formal frameworks* are then required in order to efficiently design the mathematical models, to perform low cost simulations and to extract relevant predictions from the models. The choice of the best suited formal framework is guided by the biological question(s) under consideration. For instance if the position of the entities in a two or three-dimensional space is important, as well as their trajectories, or the form of some compartments, or the diffusion speeds, *etc*, then frameworks such as cellular automata **(1)**, multi-agent systems **(2)**, discrete geometry **(3)**, and so on, may be suited. If, on the contrary, it can be

relevant to ignore the 3D arrangement of the biological objects under consideration in favor of their quantities and the evolution of these quantities along time, then frameworks dedicated to *biological regulatory networks* are relevant.

Since about ten years, the formal frameworks for regulatory networks play a central role in integrative biology and in systems biology, and they are one of the main scientific roots that initiate the new era of synthetic biology. Based on the simple idea that toxicology often needs to predict the behavior of complex biological systems, and that it has consequently a large intersection with systems biology in general, the formal frameworks for regulatory networks and their associated computer tools (which are now commonly used in the systems biology laboratories) constitute an interesting part of computational toxicology. We will show in this chapter how they can be used at the cellular and intracellular levels of description but regulatory network models can also be relevant at higher levels of description. A good model can *predict* for instance the detoxification capabilities of certain pathways, and it can be used to point out potentially dangerous configurations such as DNA damage. Nevertheless, one should always have in mind that a mathematical model never *proves* the safety of a configuration or a molecule: *in vivo* experiments constitute the only biological proof of a biological property. A formal model can only *suggest* the best promising solutions, it never establishes certified solutions. The predictive capabilities of models can also be used to suggest interesting biological experiments. They can also point out less interesting experiments that can be redundant for reasons far from being obvious for a human reasoning.

The predictive capability of a mathematical model is essentially based on a good choice of the *parameters* that drive the dynamics (the semantics) of the model. Contrarily to most of the classical models in sciences such as physics, chemistry or computer science, even a very simple biological system involves a very intricate network of interactions, and a small change in the relative strengths of these interactions can deeply modify the behavior of the system under consideration. Consequently, a formal model for biology contains a *large number of parameters* controlling these interactions and *the* problem of the modeling activity is to find and firmly establish all the possible values of those parameters. This question is known as the *parameter identification problem* and is particularly difficult in biology because the experiments that could establish the parameter values ask for indirect reasonings: a direct measure of a parameter value *in vivo* is rarely possible.

Formal logic and formal methods from computer science have proved to be very efficient to assist the identification of parameters: there are well established computer algorithms that can perform *in silico* formal reasonings, and this often leads to clever conclusions, far from being obvious at first glance. Formal methods constitute also a powerful approach to abstract (to simplify) the mathematical models in such a way that only the relevant features to answer the considered biological question(s) are retained into the models. We will see how the so called *discrete frameworks* can realize “qualitative” models dedicated to the sensible questions.

There are different kinds of biological regulatory networks and they give rise to different mathematical frameworks (signaling networks **(4)**, metabolic networks **(5)**, gene networks **(6)**...). In this chapter, we will focus on gene regulatory networks; the overall modeling method (drawing of the interaction network as a graph, identification of the parameters, predictions and

feedback to experiments) does not differ notably for the other kinds of regulatory network, only the underlying mathematical frameworks differ. Moreover, for pedagogical reasons, we will use a simple biological running example in order to give an good intuitive idea of the formal modeling approach. Section 2 explains the biological aspects of our example; Section 3 explains in details how gene regulatory networks can be formally modeled; Section 4 discusses some results obtained for our running example; Section 5 shows how formal logic and *in silico* reasonings provide a systematic way to identify parameters, and consequently predict the possible behaviors of the considered gene network.

2. Example: Detoxification induced by benzo[a]pyrene exposure

Benzo[a]pyrene (BaP) is an environmental carcinogenic polycyclic aromatic hydrocarbon (PAH) that is formed through incomplete combustion of organic materials, and common sources are tobacco smoke, automobile exhaust, and food **(7, 8)**. BaP toxicity is largely mediated through binding to the aryl hydrocarbon receptor (AhR), a ligand-activated transcription factor, found in vertebrate species from fish to humans **(9)**.

The unliganded AhR is maintained in cytoplasm in association with a chaperone complex (Hsp90/XAP/p23).

Depending upon BaP binding and concentration, the activated AhR sheds the chaperon proteins and translocates into the nucleus, where it forms a heterodimer with AhR nuclear translocator (ARNT) already present in the nucleus **(10)**. This complex recognizes an enhancer DNA element, known as the aryl hydrocarbon response element (AHRE) - also called xenobiotic response element (XRE), and dioxin response element (DRE) – in the promoter region of target genes collectively known as the *AhR gene battery*, which results in their transcriptional activation **(11, 12, 13)**.
The *AhR gene battery* includes cytochrome P450 (e.g. *Cyp 1* family), as well as non-P450 genes (e.g. a glutathione S-transferase (*Gsta1*), a UDP glucuronosyltransferase (*Ugt a6*) …) that are coordinately induced by AhR-ligands such as BaP and encode respectively phase I and II xenobiotic metabolizing enzymes involved in the detoxification of BaP **(14, 15, 16)**.
The coordinate regulation of phase I and phase II metabolizing enzymes facilitates AhR-mediated detoxification (dp1 = detoxification pathway 1)
[Fig. 1 near here]
and is necessary for cellular protection against BaP. The oxidative metabolism of BaP catalyzed by cytochrome P450 enzymes (e.g. CYP1A1, phase I enzymes) leads to the formation of reactive and electrophilic BaP metabolites (BM) that can be inactivated by phase II enzymes-catalyzed conjugation reactions (**17)**. Phase II reactions can aid in formation of water soluble metabolites that are easily excreted from the organism, thereby reducing exposure to BaP **(18, 19)**.

Although metabolism of BaP by CYP1A1 is important for detoxification, the process can lead to the formation of reactive intermediates - both reactive BaP metabolites and reactive oxygenated species (ROS) - that cause an oxidative stress signal **(12)**.
The non-P450 AhR battery genes, which are transcriptionally activated by AhR-ligand via the AHRE, are upregulated by oxidative stress via antioxidant response element (ARE) **(12, 16)**. The

ARE is a *cis*-acting sequence located in the promoter region of target genes, which encodes enzymes essential in protection against oxidative stress.
Linkage between AHRE- and ARE-controlled genes strengthens coupling between phase I and II enzymes, and attenuates oxidative stress due to AhR-controlled CYP1A1 induction (dp2 = detoxification pathway 2) **(20).**

[Fig. 2 near here]

3. Method: Thomas' framework

3.1. Mathematical models of regulatory networks

In order to design a predictive mathematical model for regulatory networks, one has to collect two kinds of biological knowledge:

(1) the sensible set of biological objects that are supposed to drive the biological system (or the biological phenomenon) under consideration, and the mutual influences between these objects; this knowledge will constitute the *structure* of the model;
(2) a sufficiently precise evaluation of the strength of each influence between objects, under any relevant situation; this knowledge, once mathematically translated into suitable parameters, will establish the *dynamics* of the model.

The ideal situation from the mathematical point of view would be when all details about the system under consideration have been biologically elucidated, providing a unique possible structure with known parameter values and leading to a unique model exhibiting a completely defined behavior. In practice the situation is far from ideal, a majority of parameters are unknown and even the structural part may be subject to different possible versions. Consequently, one has to consider a set of potential models (possibly infinite), which can exhibit different possible behaviors. This uncertainty does not imply that the modeling activity cannot be predictive because, even under partial knowledge, all the potential models can exhibit certain common behaviors under certain conditions. The price to pay is that we have to manage a huge number of unknown parameters and possible configurations: here, computers and computer science become a corner stone for regulatory networks.

There are several mathematical frameworks to model regulatory networks and they can be classified according to the way they handle dynamics **(21)**:

- *Probabilistic or stochastic frameworks* consider that the state of the regulatory network is defined by the number of molecules of each sort in the biological system (the considered biomolecules can be for example RNA or proteins in order to define the “state” of the gene that codes for them). The possible evolutions of the system are then driven by the probability for each considered object to produce new molecules, taking also into account the probabilities of degradation, see the seminal work of Gillespie **(22)**. All these probabilities constitute the parameters of the model. Unfortunately the probabilistic models are often too detailed to facilitate predictions, even with the help of computers, because they require a huge number of non deterministic simulations *in silico,* so that the precise evaluation of the parameters is incredibly time consuming.
- *Continuous frameworks* approximate the number of molecules by a concentration level for each considered object **(23)**. Concentrations are positive real numbers, so, it becomes possible to consider the derivative of the concentrations with respect to time, and the

dynamics are then modeled by a system of differential equations with parameters. This kind of approximation, that smooths the concentration levels, is of course only valid for large numbers of molecules of all sorts. A drawback of this approach is that all the trajectories become deterministic but the advantage is that simulations are less costly and consequently it is easier to identify the possible values of the parameters.

- *Discrete or qualitative frameworks* can be seen as an opposite approximation where the concentration of molecules is discontinuous and is roughly counted for each considered object (*e.g.* "low", "medium", "high"), with of course suitable thresholds. There are as many parameters as for the two previous kinds of framework and consequently the richness of possible qualitatively different behaviors is the same, but there are fewer possible values for the parameters. So, computer science with the help of formal logic becomes very efficient to identify the parameters and to extract predictions. We will explain in details the discrete approach, focusing on the approach defined for gene regulatory networks by René Thomas in the 70's **(6)** and formalized in **(24)**.
- Lastly *hybrid frameworks* try to take benefit of the qualitative approach, whilst preserving some continuous or stochastic aspects inside each discrete state of the network. Hybrid frameworks constitute currently a very active research area in theoretical biology.

3.2. Structure: regulatory graphs

All kinds of framework represent the structure of a regulatory network as a directed graph:

- The considered objects (such as genes, relevant external conditions that can vary, or some technical observation points) are represented as nodes of the graph. These nodes are called *variables* because a "level," which can vary, will be attached to them (e.g. a concentration level or an expression level).
- The possible actions from one object to another object (such as activations or inhibitions) are represented as directed *edges* of the graph, from a source node to a target node.
- Some actions can require several source nodes (e.g. when a complex of molecules is needed to act on the target) and they can also have several targets; in such cases we often add "virtual" nodes in the graph that make explicit the cooperation between source nodes. We call *multiplexes* these nodes.

Let us consider for example the graph of Figure 3. It provides a simplified view of the benzo[a] pyrene regulatory network described in Section 2.

[Fig 3 near here]

This regulatory graph contains three variables, which are conventionally surrounded by cycles.

- BaP represents the quantity of benzo[a]pyrene present in the cell.
- CYP represents the product of the CYP1A genes, *i.e.* the cytochrome P450 concentration level.
- BM represents the quantity of benzo[a]pyrene metabolites in the cell, *i.e.* the capability to start an oxidative stress.

The regulatory graph contains two multiplexes, which are conventionally rectangles, with a first line giving the name of the multiplex (dp1 and dp2) and a second line containing a formula.

- dp1 contains the formula “CYP & not(BM)” that says that CYP must be present *with a level sufficiently high* (see below), and, on the contrary, BM must not reach *a certain level* (see below) in order to reduce the quantity of BaP in the cell. The multiplex dp1 characterizes the detoxification pathway 1 because the low level of BM reflects the absence of a significant oxidative stress.
- dp2 contains the formula “CYP & BaP” that says that CYP and BaP must be both present, *with a level sufficiently high* (see below), in order to produce BM (and start detoxification pathway 2).

The edges whose target node is a variable can be:

- *activations*, conventionally represented with arrows of the form “source → target” such as the two edges from BaP to CYP or the edge from dp2 to BM;
- or *inhibitions*, conventionally represented with arrows of the form “source ⊣ target” such as:
  - the edge from dp1 to BaP, which represents the reduction of BaP level in the cell performed *via* the coordinate induction of CYP1A and non-P450 enzymes mediated through AhR-ligand(BaP)/AHRE pathway, in the absence of oxidative stress (this inhibition reflects the detoxification pathway 1 of Figure 1),
  - the edge from BM to BaP, which represents the reduction of BaP level in the cell performed *via* the induction of CYP1A mediated through AhR-ligand/AHRE pathway and also the induction of non-P450 enzymes controlled by both AhR-ligand/AHRE and oxidative stress/ARE pathways (this inhibition reflects the detoxification pathway 2 of Figure 2).

Lastly, the edges whose source node is a variable are labeled by a positive integer:

- There are two edges that start from CYP, with targets dp1 and dp2 respectively. The first one is labeled by 1 and the second one is labeled by 2. This means that the level of CYP required to participate to dp1 is lower than the level required to participate to dp2: the integers represent the order of “triggering” when we assume that CYP is increasing, starting from its lowest possible level. The integer label is called the *threshold* of the edge; it makes more precise (and above all, *edge-dependent*) the notions of “sufficiently high level” or “certain level” used before.
- The same applies to the edges starting from BaP. The quantity of BaP can be sufficient to activate the detoxification pathway 1 where CYP is activated but the oxidative stress remains low. So, there is an edge from BaP to CYP with threshold 1. Also, there is a higher level of BaP that increases again the production of CYP and also starts an oxidative stress by producing more BM. This phenomenon is represented by the two edges from BaP to CYP and from BaP to dp2, both with threshold 2.
- Let us remark that the only relevant threshold for BM is 1 because the edge from BM to dp1 is purely virtual as it serves to mutually exclude dp1 and dp2 *via* the “not(BM)” sub-formula of dp1.

For gene regulatory graphs, René Thomas has proposed a systematic way to properly define the expression levels of a gene with simple integers **(6)**. He started from the known fact that, considering solely the action of a source gene on a target gene, the curve that represents the quantity of the target gene product (at equilibrium), in function of the quantity of the source gene

product, is a *sigmoid*. When the source gene activates the target gene, the sigmoid is increasing whereas the sigmoid is decreasing if the source inhibits the target, see for example Figure 4.
[Fig 4 near here]

A gene g of the regulatory graph being given, it is sufficient to consider all its target genes $g_1,...,g_n$ and their corresponding sigmoid curves; Figure 4 shows three target genes. Once ordered increasingly, the inflection points cut the set of possible expression levels of g into $n+1$ intervals, from 0 to n. So, the positive integer i labeling each outgoing edge is the number of the first interval where g acts on g_i. Sometimes, g_i and g_{i+1} may share the same threshold, in which case the inflection points cut the set of possible expression levels of g into n intervals only, from 0 to $n-1$ (or less if there are several shared thresholds).

To summarize, a *regulatory graph* contains variables, multiplexes, activation edges on a target variable, inhibition edges on a target variable, and edges from a source variable labeled with an integer threshold. It constitutes the structural part of the model. A fully formal mathematical definition of regulatory graphs with multiplexes can be found in **(25)**.

3.3. Dynamics: state graphs

The dynamics of a regulatory network define how the biological system evolves autonomously by describing the successive "*states"* that the system shall exhibit, starting from any initial state. Following René Thomas, within a discrete model, a *state* is defined by the number of the interval (as described before) associated to each variable. Intuitively, the state of a given variable g represents the number of edges in the graph on which g is acting (as already pointed out, it can be lower if there are some shared thresholds).

A "current" state being given, it describes for each variable the targets on which the variable is "currently" acting. For example (BaP=1, CYP=1, BM=0) is a state where, according to Figure 3:

- BaP activates CYP *via* the left hand side black edge but not *via* the right hand side blue edge, and BaP is not acting on dp2;
- CYP is acting on dp1 but it is not acting on dp2;
- BM is not acting on dp1 (thus "not(BM)" is true) and BM does not repress BaP.

Consequently, a current state being given, one can make an *inventory* of the edges that are acting on a variable. For example within the current state (BaP=1, CYP=1, BM=0):

- BaP is repressed by dp1 because "CYP & not(BM)" is satisfied (because CYP passes its threshold 1 and BM does not); BaP is not repressed by BM because BM does not pass the threshold 1;
- CYP is activated by BaP at level 1 (left hand side black edge), but not at level 2 (right hand side blue edge);
- BM is not activated by dp2 because "CYP & BaP" is false (because CYP does not pass its threshold 2, and neither does BaP).

All mathematical frameworks for gene regulatory networks consider that this inventory decides what are all the possible futures from the current state in the dynamics. More precisely, in the discrete framework, we consider that the state of each variable g of the regulatory graph tries to move towards a value $K_{g,\{e1,e2,...\}}$ where $\{e1, e2, \ldots\}$ is the inventory of all the edges that currently act on g, according to the current state. The value of the parameter $K_{g,\{e1,e2,...\}}$ is called the *focal point* of g for the current state.

Consequently, to define the dynamics of a regulatory graph, one has to identify the values of a family of parameters of the form $K_{g,\{e1,e2,\ldots\}}$ where *g* is any variable of the regulatory graph and *{e1, e2, …}* is any subset of the edges whose target is *g*. For example, according to the regulatory graph of Figure 3, the following parameters should *a priori* be considered:

- $K_{BaP,\{\}}$, $K_{BaP,\{dp1\}}$, $K_{BaP,\{BM\}}$ and $K_{BaP,\{dp1,BM\}}$ whose possible values range from 0 to 2;
- $K_{CYP,\{\}}$, $K_{CYP,\{BaP1\}}$, $K_{CYP,\{BaP2\}}$ and $K_{CYP,\{BaP1,Bap2\}}$, whose possible values range from 0 to 2;
- $K_{BM,\{\}}$ and $K_{BM,\{dp2\}}$, whose possible values are 0 or 1.

In fact, the focal point $K_{CYP,\{BaP2\}}$ is useless because it is impossible for BaP to act on CYP *via* the right hand side blue edge without acting also *via* the left hand side black edge (if BaP=2 then it is greater than 1, consequently $K_{CYP,\{BaP1,Bap2\}}$ applies). Similarly, $K_{BaP,\{dp1,BM\}}$ is useless because if BM passes the threshold 1 to repress BaP, then not(BM) is false, and consequently dp1 cannot repress BaP. In other words BaP can be either repressed *via* the detoxification pathway 1 in black or repressed *via* the detoxification pathway 2 in blue, never both.

As usual, the corner stone of the modeling activity is the parameter identification process: it will be discussed in the next section. Let us assume for the moment that the parameter values are known and let us show how to deduce the *state graph*, which defines the dynamics. A state being defined by an integer value for each variable of the regulatory graph, the state space can be seen as a hyperrectangle whose dimension is the number of considered variables. So, the state space for our example is a three-dimensional box, with three values (from 0 to 2) in the BaP and CYP dimensions, and two values (0 and 1) in the BM dimension (according to the numbering of intervals mentioned previously). Consequently, it contains 18 states. Let us first consider the 6 states where BaP=0, so that we will be able to draw this subspace on a flat paper easily, and study the cell behavior without BaP. It seems reasonable in this case to assume that CYP and BM admit their intervals numbered 0 as focal points: $K_{CYP,\{\}}=0$ and $K_{BM,\{\}}=0$. As BaP is required for any action on CYP or BM, these two parameters are the only ones that are useful when BaP=0.

- Let us consider for instance the state (CYP,BM)=(0,1). Its focal state is the target state ($K_{CYP,\{\}}$,$K_{BM,\{\}}$)=(0,0) and the dynamics will simply contain a *transition* from the state (0,1) to the state (0,0).

[Fig 5 near here]

- Let us consider now the state (2,0). Its focal state is still (0,0) as shown on the left part of Figure 5 with the red arrow. Obviously a direct transition from (2,0) to (0,0) would be biologically impossible because the CYP degradation must cross the interval numbered 1 instead of jumping from 2 to 0. Consequently, the dynamics convert the red arrow into a transition of length 1, as shown with the blue arrow on the same figure.
- Lastly, let us consider the state (1,1). Its focal state is still (0,0) as shown on the middle part of Figure 5 with the red arrow. A transition from (1,1) to (0,0) would mean that both CYP and BM cross their respective sigmoidal thresholds exactly at the same time. In fact, *one* of them is likely to cross the threshold first, depending on which one is "closer" to its threshold in the real current state *in vivo*. Consequently, the dynamics replace the oblique red arrow by two transitions, as shown with the two blue arrows, one of them modifying the CYP state alone and the other one modifying the BM state alone.

These two principles (the length of a transition is 1 and a transition modifies only one variable at a time) define how to build the state graph. The right part of Figure 5 shows all the transitions

that stay in the BaP=0 plane with $K_{CYP,\{\}}=0$ and $K_{BM,\{\}}=0$. This part of the state graph shows that, in absence of BaP, the state (CYP=0, BM=0) is a *stable state* toward which all states converge and that CYP and BM can decrease in any order.

4. An example of possible parameter values, among others

When the environment brings BaP into the cell, one has to consider all possible values of BaP and consequently the state graph is three-dimensional, with three planes (one for each value of BaP) whose transitions are similarly deduced from the parameters and there are several transitions that jump between planes when BaP varies. Let us consider a first case where the quantity of intracellular BaP can be handled by the detoxification pathway 1. This case is modeled by $K_{BaP,\{\}}=1$, and except $K_{CYP,\{\}}=0$ and $K_{BM,\{\}}=0$, the other parameters are *a priori* unknown. The next section explains how the computer can help finding the parameter values. For the moment, let us arbitrarily consider the following "reasonable" values to complete Figure 3:

(1) $K_{BaP,\{dp1\}}=0$ (meaning that the detoxification pathway 1 can be sufficient to reduce BaP from 1 to 0) and $K_{BaP,\{BM\}}=1$ (following the intuition that BM characterizes the detoxification pathway 2 by the presence of oxidative stress, and that the role of the detoxification pathway 2 is to reduce BaP from 2 to 1);

(2) $K_{CYP,\{BaP1\}}=1$ (the expression of CYP when BaP is maintained at level 1) and $K_{CYP,\{BaP1,Bap2\}}=2$ (the expression of CYP when BaP is maintained at level 2);

(3) $K_{BM,\{dp2\}}=1$ (BM trigger a significant oxidative stress when both BaP and CYP are maintained at level 2).

For each of the 18 possible states, we inventory the edges that act on each variable, and this determines the parameter that plays the role of focal point. The table on the left of Figure 6 gives the 18 corresponding lines.

[Fig 6 near here]

Then, by applying the two principles explained before, we get the state graph drawn on the right of Figure 6. Of course, the lower level plane, where BaP=0, is the one obtained in Figure 5, but we can see that the state (BaP=0, CYP=0, BM=0) is not a stable state anymore, because the new value $K_{BaP,\{\}}=1$ creates the red transition $(0,0,0)\rightarrow(1,0,0)$. This transition represents the fact that the cell environment pulls BaP to level 1. In the BaP=1 plane, CYP increases to level 1 *via* the blue transition $(1,0,0)\rightarrow(1,1,0)$. Then BaP is reduced to level 0 *via* the green transition $(1,1,0)\rightarrow(0,1,0)$. In the BaP=0 plane, CYP is reduced to level 0 *via* the blue transition $(0,1,0)\rightarrow(0,0,0)$ and finally we observe that the cycle $(0,0,0)\rightarrow(1,0,0)\rightarrow(1,1,0)\rightarrow(1,1,0)\rightarrow(0,0,0)$ replaces the stable state observed in Figure 5 where $K_{baP,\{\}}$ was equal to 0. This cycle reflects the behavior of the detoxification pathway 1.

Some remarks:

- Notice that the non-P450 genes are not explicitly taken into account in this regulatory model for pedagogical reasons only (in order to avoid a four-dimensional state graphs in this chapter). A non-P450 variable should have been included into the model for a better biological credibility and, of course, the example is easy to study with the help of a computer, which has no difficulty to handle a large number of dimensions. Indeed, all the

results explained here remain valid when we hide, as we did, the non-P450 genes in the two inhibition arrows of Figure 3.

- Besides, notice that it is impossible to escape from the cycle $(0,0,0) \rightarrow (1,0,0) \rightarrow (1,1,0) \rightarrow (1,1,0) \rightarrow (0,0,0)$ which is consequently a *basin of attraction* of the state graph.
- All other states of the state graph converge toward this basin of attraction which consequently represents the only functional behavior according to this parameter setting.
- Notice also that the BaP=2 plane is unreachable in normal conditions because all vertical transitions between the two planes BaP=1 and BaP=2 are going down.

This family of parameter values seems also suitable to model the case where the cell environment brings BaP into the cell at a sufficient level to trigger a significant oxidative stress and the detoxification pathway 2: we consider the same parameter values except that $K_{baP,\{\}}=2$ instead of 1; we get the state graph of Figure 7.

[Fig 7 near here]

We see that the cycle reflecting the detoxification pathway 1 is no more a basin of attraction. Indeed, there is a red transition $(1,0,0) \rightarrow (2,0,0)$ that escapes from the cycle of pathway 1, due to the capability of the environment to pull up BaP to level 2. New cycles appear; among them, the preferentially chosen ones go through the states (2,1,0), (2,2,0), (2,2,1) and (1,2,1). These cycles denote that a larger amount of CYP is expressed, that BM are significantly produced and that a significant oxidative stress is triggered. These cycles belong to the possible behaviors of the detoxification pathway 2. The bold arrows of Figure 7 show one of those cycles. Remember that we do not take into account the DNA damage in this model, which would impose an escape of detoxification pathway 2.

5. Materials: Model checking and SMBioNet

Since the parameters are generally not measurable *in vivo*, finding a suitable class of parameters constitutes a major issue of the modeling activity. In fact, while available data on the connectivity between elements of the network are more and more numerous, the kinetic data of the associated interactions remain difficult to interpret in order to identify the strength of the gene activations or inhibitions. While it is rather easy to construct the interaction graph, the determination of the dynamics of the model is quite difficult. This parameter identification problem constitutes the cornerstone of the modeling activities. Then, it would be interesting to automatically exhibit from some biologically known behaviors or some hypothetical behaviors, parameters of the model which lead to dynamics coherent with the set of available knowledge on the behavior of the system. In the context of purely discrete modeling presented before, this problem is simpler because of the finite number of parameterizations to consider. Nevertheless this number is so enormous that a computer aided method is needed to help biologists to go further in the comprehension of the biological system under study. We show in this section how formal methods from computer science are able to perform computer-aided identification of parameters.

5.1 Temporal logic

Temporal logics are languages that allow us to formalize known biological behaviors or hypothetical behaviors in such a way that computers can automatically check if a model exhibits those behaviors or not. The building blocks of a temporal logic are atoms, connectives and temporal modalities. Let us here consider the Computation Tree Logic **(26)**, CTL for short, which is one of the most common temporal logics:

(4) Atoms in CTL are simple statements about the current state of a variable of the network: equalities (e.g., (BaP=2)) or inequalities (e.g., (CYP<1) or (CYP>1)).

(5) Connectives are the standard connectives: "$\neg$", as negation (e.g., $\neg$(BaP = 0) is the negation of the atom (BaP=0)); "$\wedge$", as "and" stands for the conjunction (e.g., (BaP=0) $\wedge$(CYP>1)); "$\vee$", as "or", stands for the disjunction (e.g., (BaP=0)$\vee$(CYP>1)); "$\Rightarrow$", as "implies", stands for implication (e.g., (BaP=0) $\Rightarrow$ (CYP>1)), and so on.

(6) Temporal modalities are combinations of two types of information:

1. Quantifiers: a formula can be checked with respect to all possible choices of path in the asynchronous state graph (universal quantifier, denoted by the character "A"), or one can check if it exists at least one path such that the formula is satisfied (existential quantifier, denoted by the character "E").
2. Discrete time elapsing: a formula can be checked at the next state (character "X"), in some future state which is not necessarily the next one (character "F"), and in all future states (character "G"). Moreover a formula can be checked until another formula becomes satisfied in the future (character "U").

In short, a CTL modality is the concatenation of two characters:

first character	**second character**
A = for All path choices	X = neXt state
	F = for some Future state
E = there Exists a choice	G = for all future states (Globally)
	U = Until

To illustrate how to use CTL to express a biological property, let us consider the formula:

$$\big(\,(\mathrm{BaP}=0)\wedge(\mathrm{CYP}=0)\wedge(\mathrm{BM}=0)\,\big)\Rightarrow \mathrm{EF}\big(\,(\mathrm{BaP}=2)\wedge(\mathrm{CYP}=2)\wedge(\mathrm{BM}=1)\wedge \mathrm{AG}(\mathrm{BaP}=2)\,\big)$$

This formula means that, starting from an initial state where (BaP=0), (CYP=0) and (BM=0), it is possible (character "E") to reach, in the future (character "F"), a state where (BaP=2) and (CYP=2) and (BM=1). Moreover, from this latest state, all trajectories (character "A") will stay for ever (character "G") in the set of states where (BaP=2).

5.2 CTL to Encode Biological Properties

CTL formulas are useful to express temporal properties of biological systems. Once such properties have been elaborated, a model of the biological system will be acceptable only if its state graph satisfies the CTL formulas, otherwise, it is not considered anymore. Considering our running example, three temporal properties seem relevant.

The first temporal property focuses on the behavior of the system when the toxic exposure level is null (K_{BaP}=0). In such a case, the system is able to reset the expression level of CYP towards its basal level, that is towards 0. Let us first denote by (x,y,z) the formula ((BaP=x) $\wedge$ (CYP=y)

$\wedge$ (BM=z)). Since (K_{BaP}=0) is equivalent to the fact that from the state (0,0,0), the increasing of BaP is not possible, this behavior is translated into CTL as:

$$\varphi_0 \equiv [\,(0,0,0) \Rightarrow \neg\, EX\,(1,0,0)\,] \wedge [\,(BaP=0) \Rightarrow AF(AG(CYP=0))\,]$$

The second property focuses on the behavior of the system when the toxic exposure level is set to 1 (K_{baP}=1). The detoxification pathway 1 is supposed to be sufficient to detoxify the cell completely. Besides BaP cannot increase up to level 2. In addition the detoxification pathway 1 (when the BaP level is decreasing from level 1 to 0) does not involve the oxidative stress/ARE pathway (in other words BM=0). Since (K_{BaP}=1) is equivalent to the fact that, on the one hand, from the state (0,0,0), the increasing of BaP is possible, and on the other hand, from the state (2,0,0), the decreasing of BaP is possible. Thus, these properties are translated in CTL as follows:

$$\varphi_1 \equiv \big(\,[(0,0,0) \Rightarrow EX(1,0,0)] \wedge [(2,0,0) \Rightarrow EX(1,0,0)]\,\big) \wedge$$
$$\big(\,(BaP>0) \Rightarrow \{EF(BaP=0) \wedge AF(AG(BaP<2)) \wedge AG[((BaP=1) \wedge EX(BaP=0)) \Rightarrow (BM=0)]\}\big)$$

The third temporal property focuses on the behavior when the toxic exposure level is set to 2 (K_{baP}=2). In such a case, a path which detoxifies completely exists:

$$\varphi_2 \equiv (BaP=2) \Rightarrow EF(BaP=0)$$

In practice, CTL formulas are sufficient to express the majority of useful biological properties even if in some cases the translation of a property is tricky.

5.3 Computer Aided Elaboration of Formal Models

To apprehend a biological system, the researchers accumulate knowledge on this system. As seen in section 2, this knowledge includes structural or dynamic knowledge. CTL is used to encode dynamic properties of the biological system, including the response to a given stress, some possible stationary states, known oscillations, etc. In general this second kind of knowledge can also be an hypothesis about the behavior of the system.

The first question focuses on consistency: *is the dynamic knowledge coherent with the structural knowledge?*

After the formalization step, formal logic and formal models allow us to test hypotheses, to check consistency, to elaborate more precise models incrementally, and to suggest new and relevant biological experiments. The classical way of testing consistency, introduced in **(24)**, consists in the following four steps:

- Draw all the sensible regulatory graphs according to the structural biological knowledge, with all the sensible, possible threshold allocations.
- Express in a formal language, CTL for example, the known behavioral properties as well as the considered biological hypotheses.
- Then, automatically generate, for each possible regulatory graph, all the possible values for all parameters. For all of them, generate the huge number of corresponding state graphs.

- Check each of these models against the CTL formulas expressing the dynamic knowledge. This step is called *model checking*.

If no model survives to the fourth step, then reconsider the hypotheses and perhaps extend model schemes.

In the context of R. Thomas' modeling, the software plateform SMBioNet **(25)** implements this way of testing consistency: it allows one to select the models that are consistent with the regulatory graph and the dynamic properties expressed in CTL. For each parameterization, SMBioNet constructs the corresponding asynchronous state graph and check if the CTL temporal formula is satisfied by this state graph. This verification step is performed by the model checker NuSMV **(27)**.

The total number of parameterizations to consider can be easily computed. Let us first remark that parameters associated with BaP ($K_{baP,\{\}}$, $K_{baP,\{dp1\}}$, $K_{baP,\{dp1,BM\}}$) can take their values in [0,1,2], that parameters associated with CYP ($K_{CYP,\{\}}$, $K_{CYP,\{bap1\}}$, $K_{CYP,\{bap1,bap2\}}$) can take their values in [0,1,2], and that parameters associated with BM ($K_{BM,\{\}}$, $K_{BM,\{dp2\}}$) can take their values in [0,1], as described in Section 3.3. Thus there exist $3^3 \times 3^3 \times 2^2$=2916 different parameter values to consider. It does *not* mean that there exist 2916 different *state graphs* to consider, for two reasons:

- It is not restrictive to consider only "monotonous" parameterizations where an activator cannot decrease a parameter (if a is an activator of a variable v and if w is a set of resources of v, then $K_{v,w} \leq K_{v,w\cup\{a\}}$) and an inhibitor cannot increase a parameter (if i is an inhibitor of v, then $K_{v,w\cup\{i\}} \leq K_{v,w}$) because the multiplexes explicit the exceptions at the structural level.
- Several parameter values can lead to a same state graph.

Consequently, the software plateform SMBioNet enumerates only the *different* state graphs that are associated with a *monotonous* parameterization. For our BaP example, SMBioNet enumerates only 420 state graphs, submits each graph to the model checker, selects only the state graphs that satisfy the CTL formulas φ_0, φ_1 or φ_2, and outputs the corresponding parameterizations.

There are 18 models (parameter valuations) which lead to a state graph which is consistent with φ_0, 14 models consistent with φ_1 and 90 models consistent with φ_2. According to Section 5.2, we are interested in the models that satisfy φ_0 when K_{BaP}=0, and φ_1 when K_{BaP}=1, and φ_2 when K_{baP}=0. Interesting models are then those that share all the values of parameters except the values of parameters $K_{baP,\ldots}$ because BaP is an environmental variable of the regulatory network (external to the cell).

Finally three different parameterizations of $K_{CYP,\ldots}$ and $K_{BM,\ldots}$ survive. To deepen our understanding of the system, we have to construct the corresponding state graph for each possible value of the parameters $K_{baP,\ldots}$ and check their biological meaning, as partly done in Section 4 for one of these three parameterizations.

6. References

(1) Kier LB, Bonchev D, Buck GA (2005) Modeling biochemical networks: a cellular-automata approach. Chem. Biodivers. 2:233-243

(2) Hoehme S, Drasdo D (2010) A cell-based simulation software for multicellular systems. Bioinformatics 26 (20): 2641-2642

(3) Poudret M, Comet J-P, Le Gall P et al (2008) Topology-based abstraction of complex biological systems: Application to the Golgi apparatus. Theory in Biosciences, 127:79-88

(4) Eungdamrong NJ, Iyengar R (2004) Modeling Cell Signaling Networks. Biol of the Cell, 96:355-362

(5) Schuster S, Hilgetag C, Woods JH et al (2002) Elementary flux modes in biochemical reaction systems: Algebraic properties, validated calculation procedure and example from nucleotide metabolism. J Math. Biol. 45:153-181

(6) Thomas R, d'Ari R (1990) Biological feedback. CRC Press

(7) Bostrom CE, Gerde P, Hanberg A et al (2002) Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambiant air. Environ. Health Perspect. 110 (Suppl. 3):451–488

(8) Phillips DH (1999) Polycyclic aromatic hydrocarbons in the diet. Mutat. Res. 443 (1–2):39–147

(9) Schmidt JV, Bradfield CA (1996). Ah receptor signaling pathways. Annu. Rev. Cell Dev. Biol. 12:55–89

(10) Hankinson O (1995) The aryl hydrocarbon receptor complex. Annu. Rev. Pharmacol. Toxicol. 35:307–340

(11) Gu YZ, Hogenesch JB, Bradfield CA (2000) The PAS superfamily: sensors of environmental and developmental signals. Annu. Rev. Pharmacol. Toxicol. 40:519–561

(12) Nebert DW, Roe AL, Dieter MZ et al (2000) Role of the aromatic hydrocarbon receptor and [Ah] gene battery in the oxidative stress response, cell cycle control, and apoptosis. Biochem. Pharmacol. 59:65–85

(13) Nebert DW, Dalton TP, Okey AB et al (2004) Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer. J. Biol. Chem. 279(23):23847–23850

(14) Nebert DW, Vasiliou V (2004) Analysis of the glutathione *S*-transferase (GST) gene family. Hum. Genomics 1(6):460–464

(15) Nioi P, Hayes JD (2004) Contribution of NAD(P)H:quinine oxidoreductase 1 to protection against carcinogenesis, and regulation of its gene by the Nrf2 basic-region leucine zipper and the arylhydrocarbon receptor basic helix–loop–helix transcription factors. Mutat.Res. 555(1–2):149–171

(16) Yoshinari K, Okino N, Sato T et al (2006) Induction of detoxifying enzymes in rodent white adipose tissue by aryl hydrocarbon receptor agonists and antioxidants. Drug Metabolism and Disp. 4:1081–1089

(17) Wills LP, Zhu S, Willett KL et al (2009) Effect of CYP1A inhibition on the biotransformation of benzo[a]pyrene in two populations of Fundulus heteroclitus with different exposure histories. Aquat Toxicol. 5; 92(3):195–201

(18) Parkinson A (1996) Biotransformation of Xenobiotics Casarett and Doull's Toxicology. In: Klaassen, CD., editor. The Basic Science of Poisons. McGraw-Hill, New York

(19) Yang SK (1988) Stereoselectivity of Cytochrome P-450 Isozyemes and Epoxide Hydrolase in the Metabolism of Polycylic Aromatic Hydrocarbons. Biochem. Pharmocol. 37:61–70

(20) Köhle C, Bock KW (2007) Coordinate regulation of Phase I and II xenobiotic metabolism by the Ah receptor and Nrf2. Biochem Pharmacol. 73:1853-62.

(21) de Jong H (2002) Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. J. Comput. Biol. 9(1):67-103.

(22) Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340-2361

(23) Tyson JJ, Othmer HG (1978) The dynamics of feedback control circuits in biochemical pathways. Prog. Theor. Biol. 5:1–62

(24) Bernot G, Comet J-P, Richard A et al (2004) Application of Formal Methods to Biological Regulatory Networks: Extending Thomas' Asynchronous Logical Approach with Temporal Logic, J Theor. Biol., 229(3): 339-347

(25) Khalis Z, Comet J-P, Richard A, et al. (2009) The SMBioNet Method for Discovering Models of Gene Regulatory Networks. Genes, Genomes and Genomics, 3(special issue 1):15-22

(26) Emerson EA (1990) Temporal and modal logic. In: Van Leeuwen, J. (ed) Handbook of theoretical computer science, MIT Press

(27) Cimatti A, Clarke EM, Giunciglia EF, et al (2002) NuSMV 2: An open source tool for symbolic model checking. In: *Proceeding of International Conference on Computer-Aided Verification (CAV 2002)*, pp 27-31

7. Captions

Figure 1: Detoxification pathway 1 (dp1)

Figure 2: Detoxification pathway 2 (dp2)

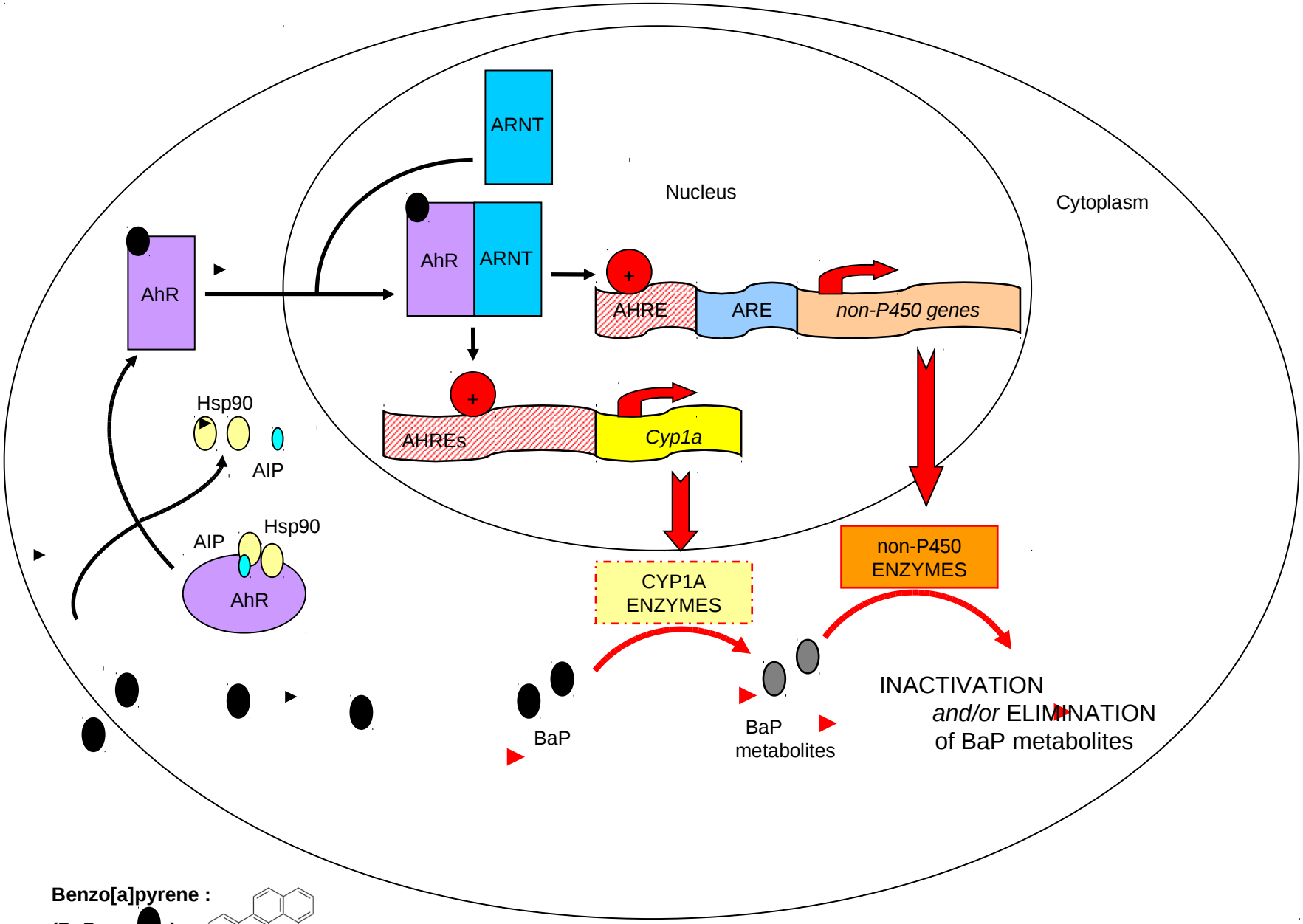
Figure 3: The simplified benzo[a]pyrene regulatory graph. Detoxification pathway 1 is black on the left and detoxification pathway 2 is blue on the right. DNA damages are not considered here.

Figure 4: Example of sigmoid shapes where a source gene g activates its target genes g1 and g3 and inhibits g2. Four qualitatively different intervals appear, numbered by the number of genes on which g is acting.

Figure 5: Construction of the state graph where BaP is fixed to 0.

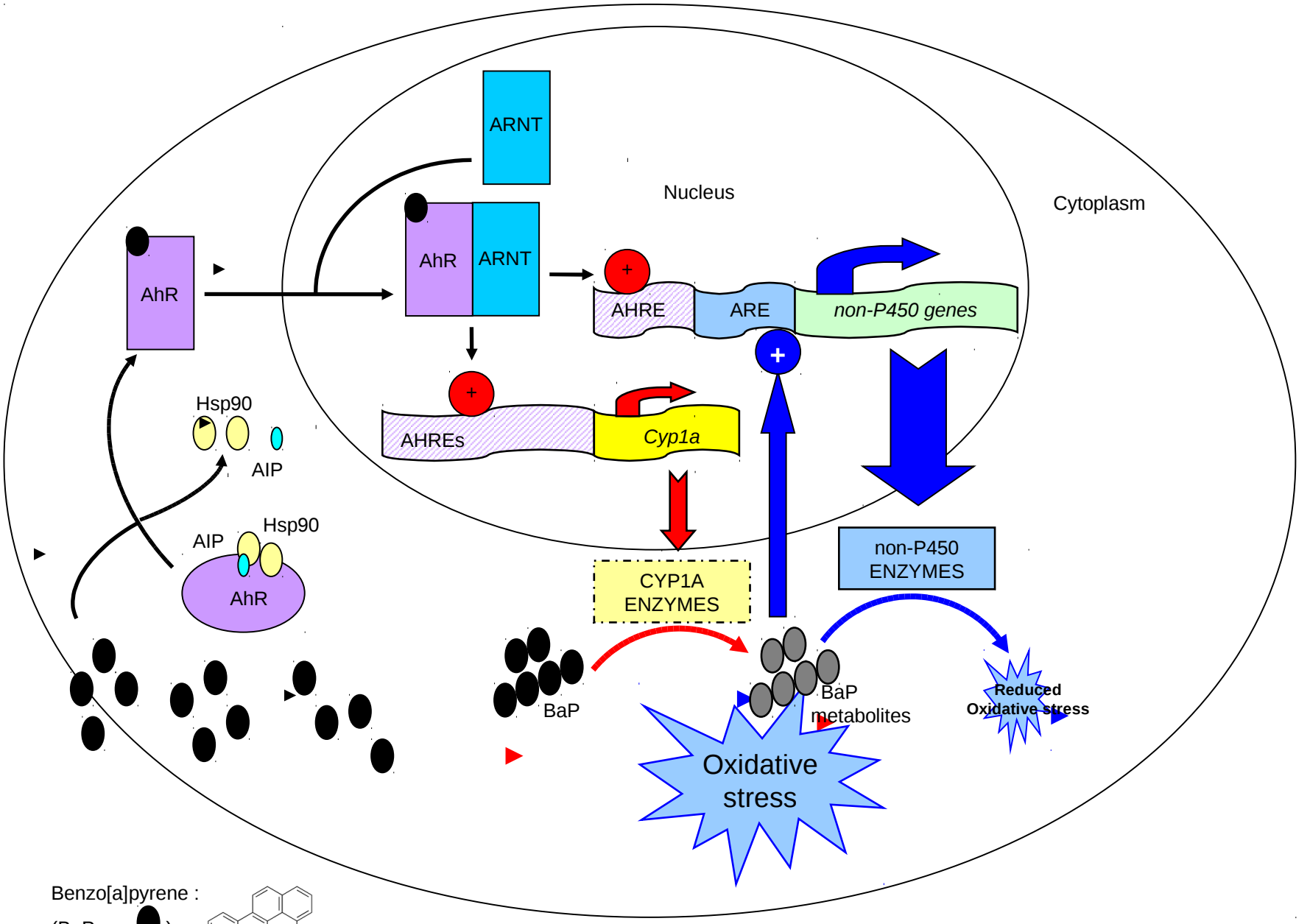
Figure 6: A discrete model of the interleaving pathways *dp1* and *dp2* when environment imposes an in-between level of BaP. **(Left)** The table shows for each possible state the set of resources of each variable and the possible evolution directions. **(Right)** The associated state graph.

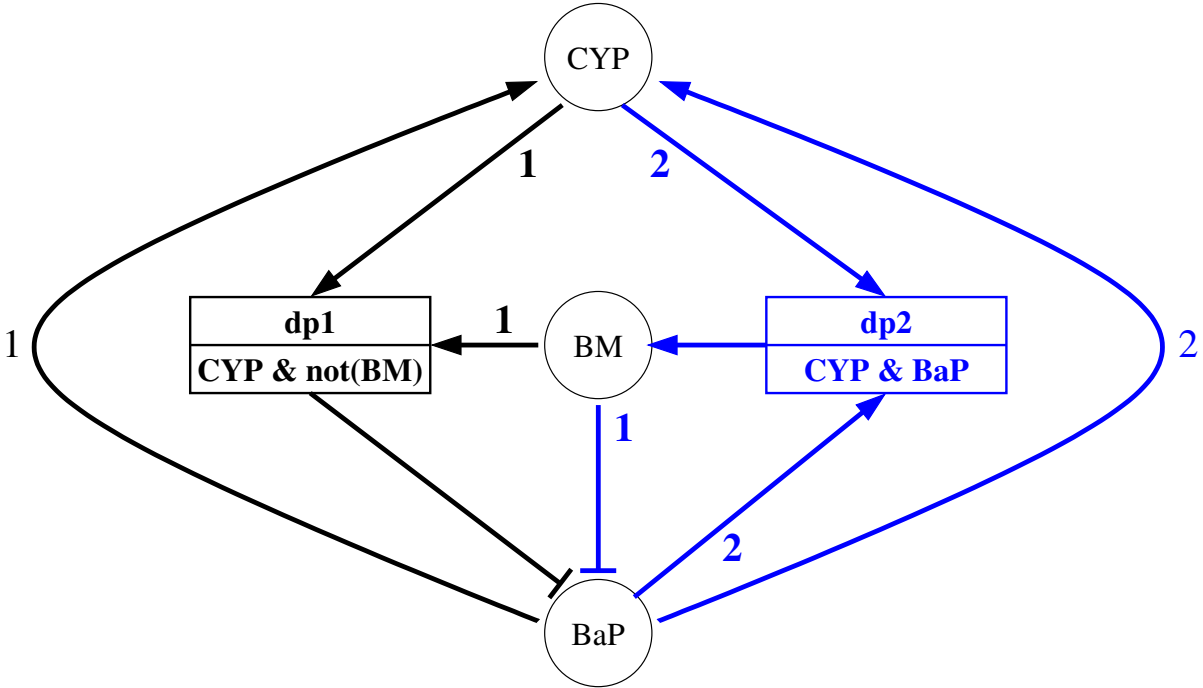
Figure 7: A discrete model of the interleaving pathways *dp1* and *dp2* when environment imposes a high level of BaP. **(Left)** The table shows for each possible state the set of resources of each variable and the possible evolution directions. **(Right)** The associated state graph.

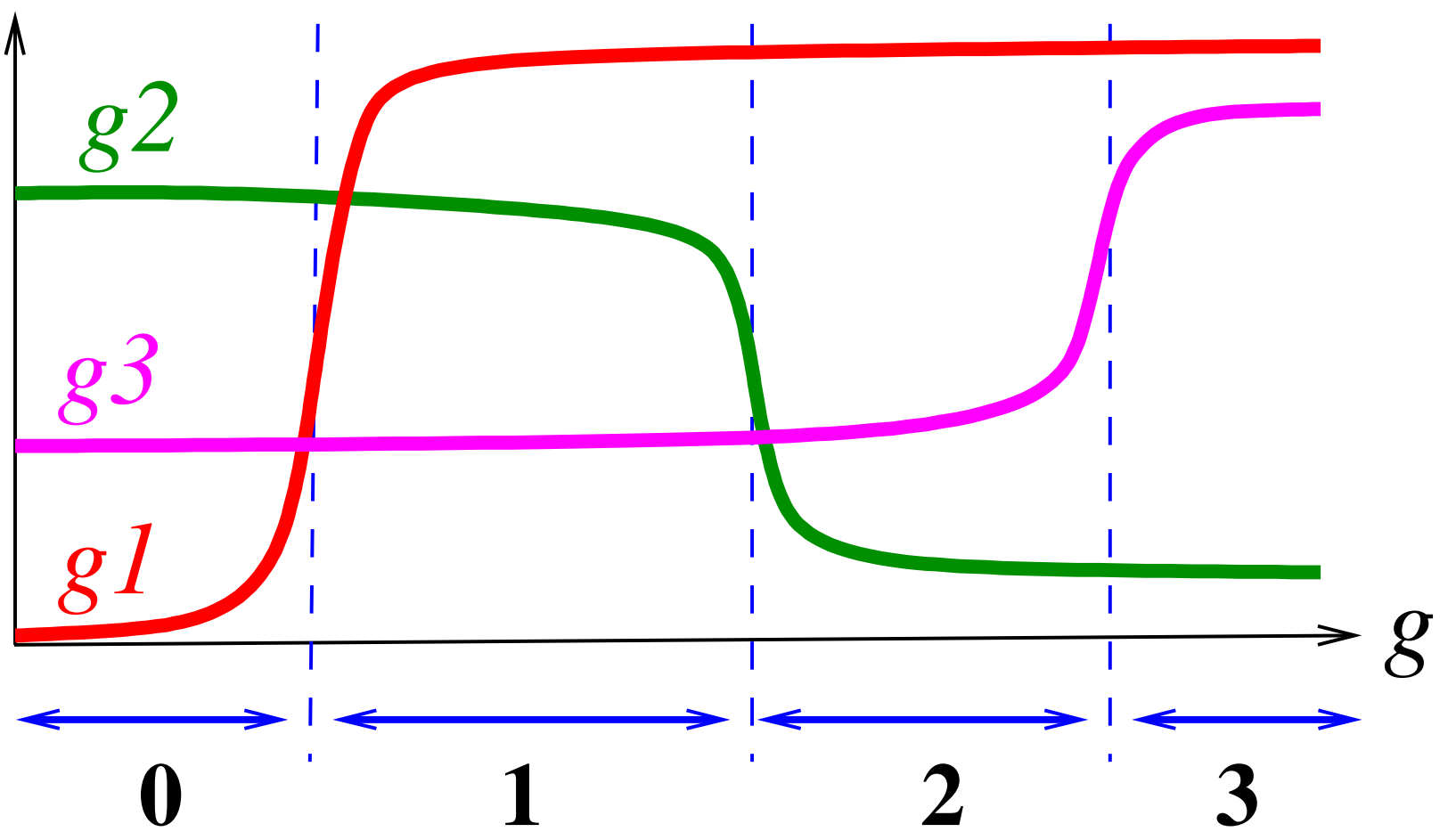


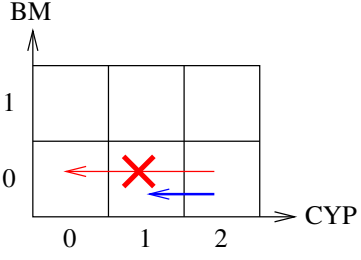
Benzo[a]pyrene :

(BaP or ●)

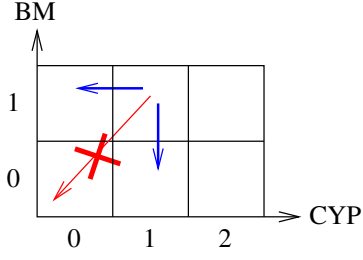

ARNT
Nucleus
Cytoplasm
AhR
AhR
ARNT
+
AHRE
ARE
non-P450 genes
+
+
AHREs
Cyp1a
Hsp90
AIP
AIP
Hsp90
AhR
CYP1A
ENZYMES
non-P450
ENZYMES
BaP
BaP
metabolites
Reduced
Oxidative stress
Oxidative
stress
Benzo[a]pyrene :
(BaP or)


CYP
1
2
dp1
CYP & not(BM)
1
BM
dp2
CYP & BaP
1
2
1
2
1
2
BaP

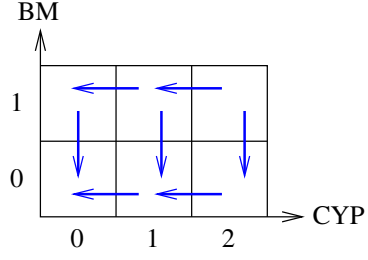

g2
g3
g1
g
0
1
2
3



transition from state (2,0)

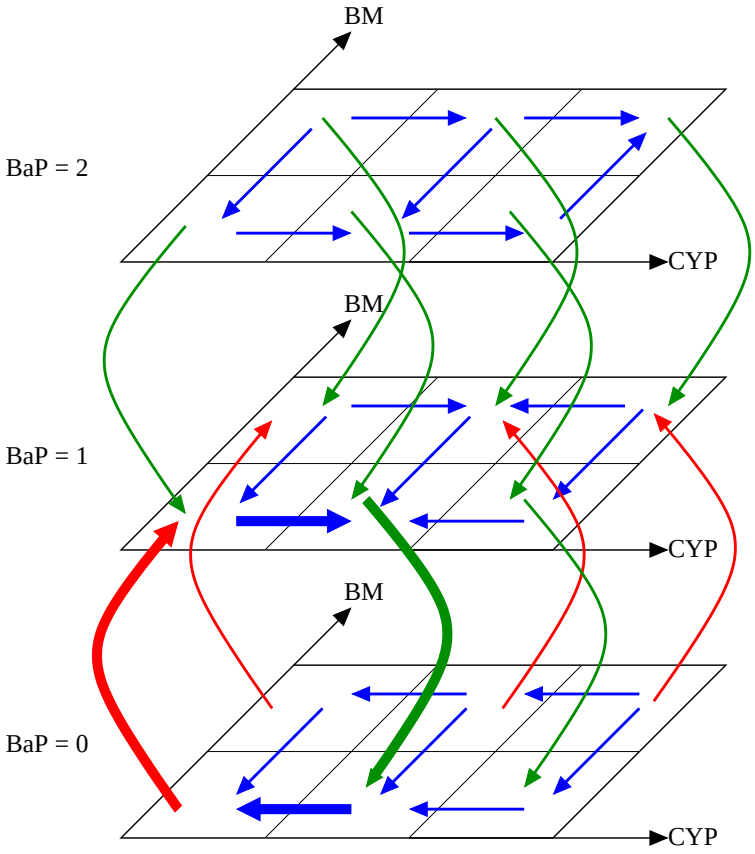


transitions from state (1,1)



state graph (for BaP=0)

State			Resources			Focal points		
BaP	CYP	BM	BaP	CYP	BM	BaP	CYP	BM
0	0	0				1	0	0
0	0	1	BM			1	0	0
0	1	0	dp1			0	0	0
0	1	1	BM			1	0	0
0	2	0	dp1			0	0	0
0	2	1	BM			1	0	0
1	0	0		BaP1		1	1	0
1	0	1	BM	BaP1		1	1	0
1	1	0	dp1	BaP1		0	1	0
1	1	1	BM	BaP1		1	1	0
1	2	0	dp1	BaP1		0	1	0
1	2	1	BM	BaP1		1	1	0
2	0	0		BaP1,BaP2		1	2	0
2	0	1	BM	BaP1,BaP2		1	2	0
2	1	0	dp1	BaP1,BaP2		0	2	0
2	1	1	BM	BaP1,BaP2		1	2	0
2	2	0	dp1	BaP1,BaP2	dp2	0	2	1
2	2	1	BM	BaP1,BaP2	dp2	1	2	1



State			Resources			Focal points		
BaP	CYP	BM	BaP	CYP	BM	BaP	CYP	BM
0	0	0				2	0	0
0	0	1	BM			1	0	0
0	1	0	dp1			0	0	0
0	1	1	BM			1	0	0
0	2	0	dp1			0	0	0
0	2	1	BM			1	0	0
1	0	0		BaP1		2	1	0
1	0	1	BM	BaP1		1	1	0
1	1	0	dp1	BaP1		0	1	0
1	1	1	BM	BaP1		1	1	0
1	2	0	dp1	BaP1		0	1	0
1	2	1	BM	BaP1		1	1	0
2	0	0		BaP1,BaP2		2	2	0
2	0	1	BM	BaP1,BaP2		1	2	0
2	1	0	dp1	BaP1,BaP2		0	2	0
2	1	1	BM	BaP1,BaP2		1	2	0
2	2	0	dp1	BaP1,BaP2	dp2	0	2	1
2	2	1	BM	BaP1,BaP2	dp2	1	2	1