Etude et Modélisation de Réseaux de Régulation Génétiques

Projet de fin d'études ENSIMAG

D.E.A. Mathématiques Appliquées

Laurent TOURNIER

22 juin 2001

laboratoire : L.M.C.
directeur de stage : Jean DELLA DORA

Abstract

Given the imminent results concerning the studies on the genome, today we seek mathematical models and computation methods in order to formalize and simulate more and more precisely biochemical systems and especially gene regulation networks. Today, there exists several models which follow mainly two ways. Some of them tempt to model directly the continuous evolution of such systems while the other ones are discrete models. These two ways are interesting because they bring very different results. The hybrid systems, as a link between discrete and continuous approaches appear to be a good mathematical tool to model such complex systems.

Table des matières

Introduction

J'ai effectué mon stage dans le Laboratoire de Modélisation et Calcul de Grenoble, au sein de l'équipe MOSAIC. Mon travail s'inscrit dans une branche assez récente des mathématiques appliquées : la modélisation de systèmes complexes, notamment de systèmes provenant des sciences du vivant. Mon stage avait pour but de comprendre les différentes modélisations qui ont été faites des réseaux de régulation génétique, afin de proposer un modèle faisant intervenir les systèmes hybrides.

Il existe aujourd'hui principalement deux approches qui tentent de modéliser ce type de systèmes biochimiques. La première est une approche discrète, dans laquelle on considère que les variables biologiques peuvent être représentées par des impulsions. Les grands noms de cette théorie sont nombreux, on citera notamment l'apport du chercheur américain Stuart Kauffman. Cette méthode a l'avantage de pouvoir manipuler de gros ensembles de variables par l'intermédiaire de circuits logiques, malheureusement elle ne permet pas de rendre compte du caractère continu de l'évolution de tels systèmes.
La seconde approche est une approche continue, elle se base sur l'étude de systèmes différentiels dont la dimension croît relativement vite lorsque le système se complexifie. Nous citerons par exemple Leon Glass et Erik Plahte. Grâce à ce modèle, nous pouvons décrire de manière plus fine l'évolution des variables, mais la résolution numérique de telles équations devient très rapidement impossible à cause du nombre trop élevé des variables.

Ce constat établi, il semble plutôt intéressant d'allier ces deux approches de manière à affiner le modèle. Ceci a déjà été fait à plusieurs reprises, notamment par Harley H. McAdams et Lucy Shapiro, qui ont fait une analogie entre les réseaux génétiques et les circuits électriques ([6]). C'est pourquoi l'idée d'associer la théorie des systèmes hybrides à la modélisation de systèmes biologiques semble judicieuse. En outre, nous verrons au cours de ce rapport que la notion même de système hybride est particulièrement bien adaptée à la notion de réseau génétique.

Avant de nous lancer dans une tentative de modélisation, nous devons d'abord faire quelques rappels biologiques de base sur le métabolisme cellulaire et le contrôle génétique. Ce sera l'objet de la première partie. Nous tenterons ensuite de décrire avec précision ce qu'est réellement un réseau génétique : ses composants, son fonctionnement etc ... Dans la troisième partie, nous verrons le modèle hybride à proprement parler.

Les deux dernières parties complèteront notre étude, en rendant compte de deux travaux effectués durant le stage. Nous examinerons tout d'abord la résolution d'une équation biochimique de base : l'équation de Michaelis-Menten, puis nous tenterons de résumer rapidement l'un des problèmes que pose la modélisation booléenne : le problème de l'identification des réseaux.

1 Éléments de biologie cellulaire

La **cellule** est l'entité organique élémentaire de tous les êtres vivants. Sa taille est variable mais elle est de l'ordre de la dizaine de micromètres. Malgré la diversité des êtres vivant sur la planète, les cellules ont des caractères anatomiques, biochimiques et fonctionnels communs. Ces caractères sont à la base même de la vie de la cellule. Mais que peut bien signifier le terme 'vie' appliqué à une si petite entité? On l'emploie en fait pour désigner **l'activité** de la cellule: une cellule est vivante si elle est capable d'échanger avec son milieu extérieur des matières premières et de l'énergie qu'elle réutilisera ensuite à différents travaux: croissance, division (c'est-à-dire reproduction), différenciation... C'est cette consommation de matière et cette transformation d'énergie que l'on appelle 'vie'.

Au fur et à mesure de l'évolution, les êtres vivants se sont développés, complexifiés; les organismes **unicellulaires** sont devenus avec le temps **pluricellulaires**. Toutefois, les mécanismes de base de la vie n'ont pas changé: les pièces qui les composent sont très semblables, comme le sont également les molécules porteuses d'information ou celles qui interviennent dans les transformations d'énergie.

Afin de bien comprendre le fonctionnement d'une cellule, il nous faut distinguer deux parties très distinctes, par le matériel biologique qu'elles utilisent ainsi que par leurs fonctionnalités. La première contient l'information nécessaire au fonctionnement de la cellule (il s'agit de l'ADN) ainsi que tout le matériel permettant d'**exprimer** cette information. La seconde regroupe tout le reste de la cellule, c'est-à-dire les différents composants métaboliques qui participent aux processus propres à la vie cellulaire. Ainsi, si l'on tente une analogie entre une cellule et une usine, le matériel génétique formerait les plans de fabrication tandis que la deuxième partie serait l'ensemble de la machinerie qui exécute ces plans. Pour des cellules évoluées dites **eucaryotes**, le matériel génétique est conservé dans une enveloppe interne appelée **noyau**. dans les cellules moins élaborées (**procaryotes**), il n'y a pas de noyau et les molécules porteuses de l'information coexistent avec les autres composants dans le milieu intracellulaire.

Ainsi, une cellule, considérée comme une unité à part, va réaliser au cours de son existence un certain nombre de travaux, guidée par son bagage génétique. La difficulté de la biologie cellulaire réside dans la compréhension de cette activité et des mécanismes qui relient la *machinerie* au *centre de commandes*. Afin de comprendre l'extrême difficulté de ce processus ainsi que des tentatives de sa modélisation mathématique, il nous faut d'abord expliquer rapidement les mécanismes de base du fonctionnement de la cellule. Il est bien entendu que ce rapport ne cherche pas à expliquer dans le détail un demi-siècle de biologie cellulaire, nous chercherons simplement à nous familiariser avec certaines notions qui nous semblent fondamentales pour comprendre les mécanismes de régulation génétique que nous nous proposons d'étudier.

Les ouvrages de biologie cellulaire sont nombreux, on pourra se référer par exemple à [1] pour plus de détails. Le site internet [2] est également assez complet et relativement accessible.

1.1 Le métabolisme cellulaire

Nous sommes restés pour le moment relativement vagues concernant les *travaux* effectués par la cellule. Il est difficile de trouver un terme adéquat car, en fait, une cellule a énormément d'activités, toutes assez différentes. Nous n'en connaissons d'ailleurs qu'une petite partie à l'heure actuelle. Certaines sont évidentes car facilement détectables, c'est le cas de la **mitose**, ou division cellulaire. Toutefois, la vie cellulaire cache beaucoup de plus petites tâches toutes aussi essentielles. Ainsi, pour se diviser, la cellule a d'abord besoin de grossir, puis de séparer sa membrane extérieure en ayant auparavant pris soin de transmettre à ses deux futures filles la totalité de son information génétique ... En outre, Toutes ces étapes demandent, comme tout processus de production, de la matière première et de l'énergie. L'absorption de matière et la production d'énergie sont donc deux autres mécanismes vitaux pour la cellule.

De quels moyens dispose la cellule pour effectuer toutes ces opérations?

Pour répondre à cette question, il nous faut étudier la cellule à une plus petite échelle. Tous les mécanismes internes peuvent s'interpréter alors comme des intéractions chimiques entre les différents composants présents dans le milieu.

1.1.1 la mécanique enzymatique

Beaucoup de ces composants se ressemblent par leur structure : ce sont des **protéines**. Les protéines sont des chaînes (qui peuvent être relativement longues) de constituants appelés acides aminés. Il y a vingt-trois acides aminés différents dans toute la biosphère, et c'est leur agencement dans la chaîne qui détermine la forme de la protéine (qui peut être très complexe si la chaîne est très longue) ainsi que sa fonction. On parle souvent des protéines comme des 'briques' de la vie car elles forment le constituant de base de la matière organique. Parmi ces protéines, il en existe certaines aux particularités très intéressantes : les **enzymes**. Ces protéines ont, par leur forme et leur agencement, la capacité d'accélérer un certain type de réaction chimique. Cette accélération peut être fulgurante. Une fonction peut alors s'expliquer par la présence (respectivement l'absence) d'une certaine enzyme qui va catalyser (respectivement bloquer) une réaction.

Prenons l'exemple de la glycolyse. Il s'agit de la chaîne d'opérations visant à transformer le sucre présent dans la cellule afin d'en tirer de l'énergie. Afin de réaliser la dégradation du sucre, il faut préalablement que celui-ci soit sous la forme de glucose. Or, il est stocké dans la cellule sous d'autres formes (notamment l'amidon, le saccharose ou le glycogène). La première étape consiste donc à transformer une partie des stocks de sucre en glucose. Pour ce faire, il existe plusieurs mécanismes de dégradation. Par exemple, la molécule d'amidon doit se dégrader en maltose, puis le maltose doit être transformé en glucose. Ces deux transformations successisves sont, à température ambiante, très lentes, quasi inexistantes. Mais grâce à la présence de deux enzymes (l'amylase pour la première réaction et la maltase pour la seconde) ces réactions peuvent se faire dans la cellule et la glycolyse peut continuer. Nous voyons bien dans cet exemple l'action directe des enzymes dans les mécanismes intracellulaires : sans ces enzymes, l'amidon ne peut pas être transformé en glucose et la glycolyse ne peut pas avoir lieu. Grâce à elles, le processus est débloqué et la cellule peut continuer son activité.

Les enzymes ne sont cependant pas les seuls *agents* de la vie cellulaire. Nous verrons par la suite d'autres molécules toutes aussi importantes : les **protéines régulatrices**. Néanmoins, la mécanique enzymatique est une clé pour comprendre la vie cellulaire. En effet, c'est en partie grâce à cette multitude d'agents que la cellule peut s'alimenter, se reproduire, en un mot, peut vivre. Si nous reprenons l'analogie avec l'usine, nous constatons que la cellule, dans son fonctionnement même, dépasse de loin ce que l'homme peut concevoir comme machinerie, car sa mécanique lui permet non seulement d'effectuer une multitude de tâches très différentes, mais également d'évoluer sans cesse au cours de son existence.

L'une des caractéristiques principales de la cellule est en effet sa formidable capacité d'**adaptation** au monde extérieur. En l'espace de quelques minutes (voire parfois quelques secondes), elle est capable de produire un nouvel agent qui serait devenu indispensable à sa survie. Considérons l'exemple de la consommation lactique (qui sera développé plus longuement par la suite). Comme nous l'avons souligné à plusieurs reprises, une condition nécessaire à la vie cellulaire est une alimentation correcte en matières carbonées et énergétiques. Toutefois lorsque le milieu en est dépourvu et que le sucre présent en réserve n'est plus suffisant, la cellule dispose d'autres mécanismes d'alimentation. Par exemple, si le milieu en contient, elle pourra consommer du lactose, qui est un sucre plus difficile à traiter mais qui peut servir de substitut tant que les autres sucres manquent. Pour pouvoir faire pénétrer le lactose, puis pour pouvoir le transformer en glucose, il faut la présence de deux enzymes quasi inexistantes en temps normal : la perméase et la β-galactosidase. plusieurs expériences, notamment celles menées par François Jacob et Jacques Monod (prix Nobel de médecine en 1965) ont montré qu'une culture de levures privée de glucose, en présence de lactose, produit une quantité massive de ces enzymes en l'espace de quelques minutes seulement. Cette expérience apporte une preuve irréfutable de l'adaptabilité cellulaire : si les conditions changent, la cellule est capable, dans une certaine mesure, de faire évoluer sa machinerie de manière à répondre aux changements.

Nous venons de voir de manière succinte comment la mécanique enzymatique permet à la cellule de répondre à son environnement et d'effectuer toutes les tâches qui composent sa vie. Nous allons à présent nous intéresser à ce mécanisme d'un point de vue plus quantitatif.

1.1.2 les équations de Michaelis-Menten (1913)

Dans les réactions catalysées par des enzymes, les concentrations en réactifs et en produits sont généralement cent ou mille fois plus élevées que la concentration en enzymes. On en déduit que chaque molécule d'enzyme participe à plusieurs transformations de réactif. La catalyse à proprement parler se fait au niveau d'un site précis de l'enzyme, appelé **site catalytique**. Ce site a une séquence d'acides aminés et une forme très spécifiques, qui lui donnent une haute affinité avec ses substrats.

Considérons par exemple une réaction ne comportant qu'un réactif (voir [3]). L'hypothèse fondamentale de Michaelis et Menten est que l'enzyme E et le substrat S commencent par se lier (sur le site catalytique) pour former un **complexe enzyme-substrat** ES. Cette réaction est supposée réversible. C'est ce complexe qui donne ensuite le produit P et qui libère l'enzyme pour une nouvelle réaction. Schématiquement :

$$S + E \underset{k_{-1}}{\overset{k_{+1}}{\rightleftarrows}} SE \xrightarrow{k_2} P + E$$

D'après la *loi d'action de masse*, la vitesse de chacune de ces réactions est proportionnelle au produit de ses réactifs. les k_i sont les constantes de proportionnalité et dépendent uniquement de la réaction chimique qui leur est associée. Ainsi, en notant de la manière suivante les concentrations en les différentes espèces :

$$s = [S], e = [E], c = [ES], p = [P]$$

on obtient les vitesses de production de ces espèces :

$$\begin{cases} \dfrac{ds}{dt} &= -k_1 es + k_{-1} c \\[2ex] \dfrac{de}{dt} &= -k_1 es + (k_{-1} + k_2) c \\[2ex] \dfrac{dc}{dt} &= k_1 es - (k_{-1} + k_2) c \\[2ex] \dfrac{dp}{dt} &= k_2 c \end{cases}$$

Nous remarquons que ce système est non-linéaire puisqu'il comporte des termes quadratiques en $[E][S]$. Afin de compléter la formulation mathématique de ce système, nous devons rajouter les conditions initiales suivantes :

$$s(0) = s_0, e(0) = e_0, c(0) = 0, p(0) = 0$$

La résolution de ce système ne peut se faire analytiquement, à cause de la présence des termes quadratiques. Nous verrons dans un chapitre ultérieur la théorie qu'ont proposée Michaelis et Menten pour étudier le comportement temporel des solutions de ces équations.

Le but de cette partie était d'étudier succintement comment fonctionne le métabolisme cellulaire à partir de la mécanique enzymatique. En d'autres termes, nous avons tenté de donner certaines clés nous permettant de comprendre comment la cellule effectuait toutes les tâches nécessaires à sa survie. Revenons à présent à notre problème initial qui est de comprendre comment la cellule régule ce métabolisme, c'est-à-dire comment son *centre de commande* génétique lui permet de synthétiser au moment adéquat les enzymes et toutes les autres protéines qui lui sont utiles.

1.2 Le contrôle génétique

La découverte de l'information génétique est encore très récente. Si la molécule d'ADN a été identifiée dès 1868 par le biologiste suisse Friedrich Miescher, ce n'est qu'en 1952 que l'on a réellement compris, grâce aux travaux d'Alfred Hershey et de Martha Chase, que cette molécule était effectivement le matériel génétique. Dès 1953, on connaissait sa structure en double brin ainsi que sa forme hélicoïdale, mais cette molécule renferme encore bien des mystères qui la placent parmi les sujets de recherche actuels (voir par exemple [4]).

Cependant, ce n'est plus seulement la molécule d'ADN elle-même qui nous intéresse, mais le fonctionnement global de la réplication et de l'interprétation du message génétique. En effet, depuis le prix Nobel de 1965 accordé aux chercheurs François Jacob et Jacques Monod pour leurs recherches sur les réseaux de régulation génétique, il semble évident que le défi de la biologie cellulaire aujourd'hui est de mettre à jour et de tenter de modéliser les nombreuses interactions qui existent entre le matériel génétique d'une part et l'ensemble du métabolisme cellulaire d'autre part. Mais avant de nous intéresser à ces réseaux, une mise au point sur les principales notions de génétique nous semble importante. Après quelques brefs rappels sur la molécule d'ADN elle-même, nous évoquerons le dogme qui est au centre de l'étude de la génétique (pour une desription plus détaillée, on se reportera à [1] et [2]).

1.2.1 la molécule d'ADN

L'**ADN**, ou **acide désoxyribo-nucléique**, a une structure très particulière qui lui permet de répondre aux principales exigences de sa fonction de porteur d'information. La molécule d'ADN est une molécule extrêmement longue, formant une double hélice. La plupart du temps, la molécule forme une pelote microscopique, enroulée (dans les cellules eucaryotes) dans le noyau, ou directement dans la cellule (pour les procaryotes). On appelle **chromosome** le molécule enroulée. Déroulée, la molécule d'ADN est un très long fil constitué par un enchaînement de bases, ou **nucléotides**. Il y a quatre bases, représentées par une lettre : l'adénine **A**, la cytosine **C**, la guanine **G** et la thymine **T**. Ces bases sont deux à deux complémentaires : **A** et **T** d'une part et **C** et **G** d'autre part se lient l'une à l'autre par des liaisons hydrogène faibles. la molécule est donc constituée de deux brins complémentaires, formés par un long squelette sur lequel s'attachent une suite de nucléotides. Chacun de ces nucléotides se lie au nucléotide correspondant, sur l'autre brin, par liaison hydrogène.

Grâce à sa structure de chaîne de nucléotides, l'ADN contient l'information nécessaire pour coder la synthèse de protéines (qui sont des chaînes d'acides aminés). Sa structure lui permet également de se répliquer facilement et de rendre la molécule relativement stable (grâce aux deux brins complémentaires).

La figure 1 représente schématiquement un bout de molécule d'ADN. Elle montre notamment l'enchaînement des nucléotides complémentaires. La figure 2 représente la double hélice d'ADN. Un tour d'hélice contient approximativement 10 paires de bases (plus précisément autour de 10.5). Le *squelette* de la molécule entoure cette dernière, tandis que les paires de base se trouvent à l'intérieur.

Bien que l'ADN ne représente que 5% environ de la masse sèche totale d'une cellule classique, une molécule compte plusieurs milliers de paires de

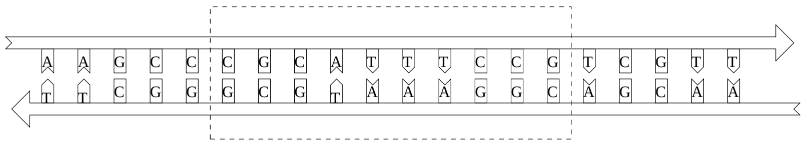


FIG. 1 – *schéma de structure de l'ADN : enchaînement des nucléotides*

base. Toute cette quantité d'information n'est pas codante. En fait, on appelle **gènes** les segments de la molécule qui vont effectivement servir au codage d'une protéine (cette définition est discutable car nous nous sommes rendus compte qu'un gène pouvait coder pour plusieurs protéines). Ces gènes sont séparés par plusieurs sites, que l'on étudiera plus précisément par la suite, qui servent de *signaux* pour, par exemple, délimiter les gènes, pour bloquer ou au contraire activer le codage de certains gènes etc... Tous ces segments que l'on nomme de façon générale **exons** ne forment pas la totalité de l'ADN. Il existe d'autres parties non codantes qui sont généralement éliminées au cours de la traduction des gènes : les **introns**. Souvent négligées, ces parties ont toutefois une importance, mais leurs rôles sont encore mal connus.

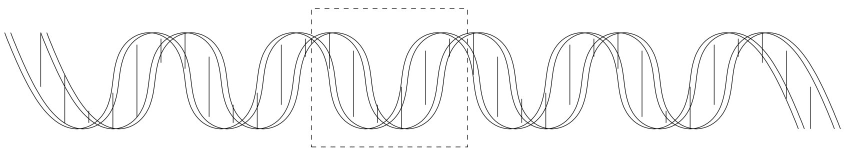


FIG. 2 – *schéma de la double hélice*

Dans le processus de traduction, expliqué plus bas, l'ADN n'est pas directement traduit en une protéine mais doit d'abord être recopié en une molécule d'**ARN**, ou **acide ribo-nucléique**, qui a la même structure de chaîne de nucléotides que l'ADN mais qui cette fois se compose d'un seul brin. C'est cette molécule qui servira d'intermédiaire entre le matériel génétique et le milieu intracellulaire. Voyons à présent plus en détail comment se fait la traduction d'un gène pour synthétiser la protéine qui lui correspond.

1.2.2 le dogme central

Le **dogme** central de la biologie cellulaire établit la liaison qui existe entre le matériel génétique contenu dans la cellule et les protéines que cette cellule synthétise.

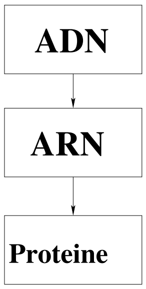


FIG. 3 – *dogme central de la biologie cellulaire*

Le passage du gène à la protéine se fait en deux étapes : tout d'abord le segment de la molécule d'ADN correspondant au gène est copié sur un brin d'ARN, que l'on appelle ARN messager (ou ARNm), c'est la **transcription**. Puis ce brin d'ARNm est à son tour recopié, mais dans un langage différent, celui des acides aminés, pour donner la séquence correspondant à la protéine synthétisée, c'est la **traduction**. La séquence polypeptidique ainsi constituée prend alors une forme spatiale qui lui est spécifique pour devenir à proprement parler la protéine.

Remarque : il existe également des traitements qui modifient la forme de l'ARN-m. La molécule peut se tordre et la séquence de bases peut se recouper par endroit formant ainsi des boucles ou des nœuds. Ces structures sont importantes car elles peuvent servir de signal d'arrêt durant la traduction.

D'un point de vue plus formel, ces deux étapes sont très différentes. En effet pour la première on garde le même alphabet, il s'agit donc simplement d'une recopie du gène. Cette recopie a essentiellement deux intérêts : elle permet de recopier le gène sur un seul brin, forme certes moins stable mais qui rendra plus simple le processus de traduction. En outre, chez les cellules eucaryotes, seul l'ARNm se déplace hors du noyau ce qui protège la molécule d'ADN. L'étape de recopie peut donc être vue comme une préservation de la molécule d'ADN, vitale pour la cellule. La seconde étape, par contre, est une vraie traduction car on passe de l'alphabet à quatre lettres de l'ADN à l'alphabet à vingt lettres des acides aminés. Voyons à présent plus en détail ces deux étapes.

– **la transcription**

Comme nous l'avons déjà évoqué plus haut, l'étape de transcription consiste en la recopie d'un segment de brin d'ADN sur une molécule d'ARNm. Cette recopie se fait par l'intermédiaire d'une enzyme appelée **ARN-polymérase**. Les fonctions de cette enzyme sont multiples. Tout d'abord, elle doit reconnaître le commencement du gène portant l'information permettant de synthétiser la protéine voulue. Ceci se fait grâce à un site spécifique sur la molécule d'ADN appelé **promoteur**, qui se situe juste avant le début du gène. Un promoteur typique comporte une soixantaine de paires de base. Une fois que la polymérase s'est liée au promoteur, elle commence la transcription du gène : elle ouvre d'abord la double hélice (au niveau des liaisons hydrogène qui sont des liaisons faibles) puis elle avance le long d'un des deux brins en synthétisant au fur et à mesure le brin complémentaire. Arrivée à la fin du gène, une séquence spécifique provoque le détachement de la polymérase et le brin d'ARN ainsi construit est libéré. La figure 4 résume l'étape de transcription.

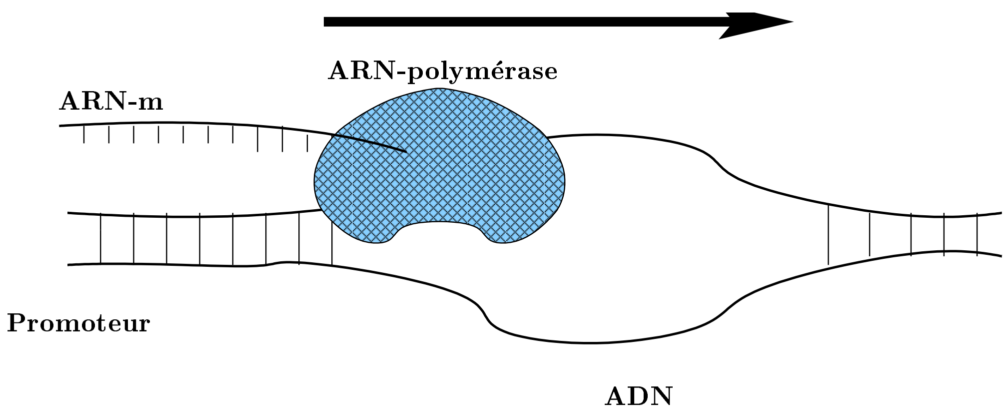


FIG. 4 – *schéma simplifié de la transcription*

La molécule d'ARNm ainsi constituée sort ensuite du noyau (chez les cellules eucaryotes) pour subir l'étape de traduction dans le milieu intracellulaire. La transcription permet de minimiser les risques d'endommagement de la molécule d'ADN, unique dans la cellule. L'alphabet utilisé reste identique, mis-à-part la base **T** qui est remplacée par l'**uracile U**. Selon la polarité de la molécule d'ADN, le brin recopié ainsi que le sens de recopie sont fixés pour chaque gène.

Notons ici que la transcription est dans les faits plus complexe car elle est suivie d'autres traitements, qui ne retiennent du gène que sa partie effectivement codante (on parle d'épissage)

– **la traduction**

La traduction est une étape légèrement plus complexe, au moins d'un point de vue formel, car il s'agit de passer de l'alphabet à quatre lettres $\{A, U, C, G\}$ à l'alphabet des acides aminés qui sont plus d'une vingtaine. Il faut donc plusieurs bases pour pouvoir coder un acide aminé. Un rapide calcul montre qu'utiliser des triplets de bases suffit à tous les déterminer. On code alors $4^3 = 64$ mots ou **codons**, ce qui implique que ce codage est redondant.

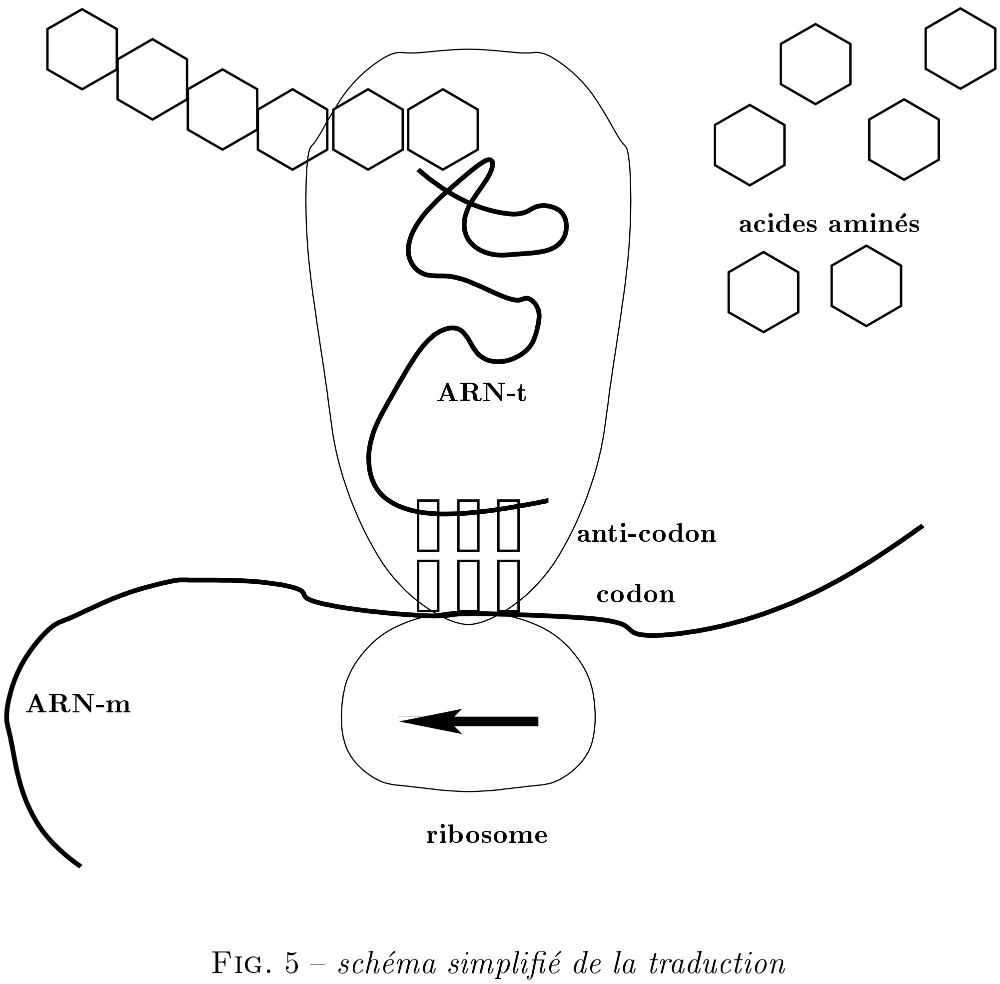


FIG. 5 – *schéma simplifié de la traduction*

La traduction se déroule dans le milieu intracellulaire, grâce à des agents appelés **ribosomes**. l'ARN messager est *lu* par le ribosome qui associe à chaque codon, par l'intermédiaire d'une molécule d'ARN particulière (l'ARN de transfert), l'acide aminé qui lui correspond. La chaîne polypeptidique ainsi constituée subit à son tour quelques traitements (dont la prise de sa forme, souvent très complexe) pour devenir la protéine désirée. La figure 5 résume la traduction en montrant ses principaux acteurs.

A présent que nous connaissons mieux le métabolisme cellulaire ainsi que le mécanisme qui permet à la cellule de le diriger, une question se pose :

Comment la cellule procède-t-elle pour contrôler toute cette machinerie?

En effet, il paraît judicieux de considérer toutes les *fonctions* effectuées par la cellule comme les différentes unités d'un gigantesque programme ayant pour but final la vie de la cellule (c'est-à-dire typiquement sa croissance, jusqu'à sa division). Mais alors comment la cellule arrive-t-elle à gérer, dans de si petites dimensions, cette multitude de tâches différentes? Il s'agit bien entendu d'une question qui est encore loin d'être élucidée. Néanmoins, de nombreux chercheurs estiment aujourd'hui qu'il faut aller plus loin dans l'étude des mécanismes précédemment décrits de manière à comprendre le contrôle qui peut être exercé sur la transcription et sur la traduction. C'est ainsi que l'on arrive à s'interroger sur le sujet de ce rapport, comment les **réseaux génétiques** permettent-ils d'expliquer la vie cellulaire? Afin d'apporter des éléments de réponse à cette question, il nous faut tout d'abord comprendre ce qu'est un réseau génétique et donner quelques exemples de réseaux simples. C'est l'objet de la partie qui suit.

2 Etude d'un réseau génétique

L'expression *réseau génétique* employée pour les mécanismes régulant la transcription et la traduction de l'ADN est à mettre en relation avec les *réseaux de régulation*. En effet, comme nous allons le voir à travers l'exemple du bactériophage λ, ces *réseaux* sont de véritables systèmes de production, incluant des termes de contrôle et de régulation, exactement comme pour une chaîne de production dans une usine. Afin de bien comprendre les mécanismes mis en œuvre, nous allons étudier en détail un réseau simple : l'infection d'une population de bactéries par un type de virus appelés phages λ. Pour plus d'informations sur le sujet, on se reportera au livre [5] de Mark Ptashne qui est très complet.

Certaines expériences dirigées par André Lwoff il y a quarante ans à l'institut Pasteur à Paris ont montré une particularité étonnante de la bactérie connue sous le nom d'*escherichia coli*. Certaines de ces bactéries, une fois irradiées par de la lumière ultraviolette, stoppent leur croissance et après une période d'environ 90 minutes, explosent littéralement (on utilise souvent en biologie le verbe *lyser*) en libérant une large quantité de phages λ dans le milieu. Les phages ainsi libérés se multiplient en allant infecter d'autres bactéries. Ces dernières ont alors deux comportements possibles : soit elles se lysent et libèrent à leur tour de nouveaux phages, soit elles survivent en conservant le phage sous une forme "endormie". Celles-ci continuent à se diviser naturellement jusqu'à ce que la culture soit à nouveau irradiée. Le phage se réveille alors et l'infection redémarre.

A. Lwoff, aidé de ses collègues F. Jacob et J. monod, a remarqué que ce phénomène était une preuve flagrante de l'existence de réseaux de régulation des gènes : selon certains facteurs intra ou extracellulaires, certains gènes stoppent (respectivement débutent) leur processus d'expression. En fait, à un moment précis, seules certaines protéines sont effectivement synthétisées. Toute la question est de comprendre les processus qui activent ou inhibent cette synthèse.

Cette question est fondamentale aujourd'hui. En effet, les maladies dites génétiques sont justement les conséquences du dysfonctionnement de certains de ces réseaux : les gènes concernés sont alors anormalement activés, ou au contraire sont complètement inhibés et ne produisent pas une protéine au rôle fondamental. Comprendre les mécanismes qui commandent l'expression des gènes fait donc l'objet d'études nombreuses, aux implications biologiques évidentes.

2.1 Le phage λ comme *bascule* biologique

2.1.1 l'état lytique et l'état lysogénique

L'enveloppe du phage est constituée de deux parties, une tête et une queue, comme le montre la figure 6. Il contient un unique chromosome qui code, entre autres, pour les quinze protéines qui constituent son enveloppe et qui se situe dans la tête du virus. Comme tout parasite, le phage est obligé d'infecter un hôte pour pouvoir se reproduire. Il s'y prend alors de la manière suivante : grâce à sa queue, il creuse littéralement un trou dans la membrane de l'*E-coli* et injecte son chromosome à l'intérieur de la cellule. *E-coli* étant une cellule procaryote, le chromosome du phage n'a pas besoin de pénétrer un noyau, il fait directement partie du matériel génétique de la bactérie. La cellule infectée a alors deux comportements possibles. Elle peut rentrer dans le **cycle lytique**, c'est-à-dire que selon un programme de régulation précis, le chromosome λ est répliqué de façon intensive et la cellule synthétise en grande quantité les protéines constituant l'enveloppe du phage. Pendant une durée d'environ 45 minutes, une centaine de nouveaux virus sont ainsi fabriqués à l'intérieur de la cellule. Au bout de cette période, la cellule meurt et sa membrane se perce pour libérer les nouveaux virus qui vont ainsi chercher à infecter de nouveaux hôtes. La deuxième possibilité pour la bactérie est de rentrer dans le **cycle lysogénique** : tous les gènes du chromosome sont inactivés sauf un (nous verrons par la suite la fonction de ce gène). Le chromosome, appelé **prophage** fait alors partie intégrante du matériel génétique de la cellule mais reste sous une forme latente. La bactérie, qu'on appelle alors **lysogène** est immunisée contre toute nouvelle infection par un autre phage; elle continue à vivre tout-à-fait normalement. En revanche, un lysogène soumis à de la lumière ultra-violette quitte son état d'équilibre et rentre à son tour dans le cycle lytique. Il propage de cette manière une nouvelle infection dans son milieu.

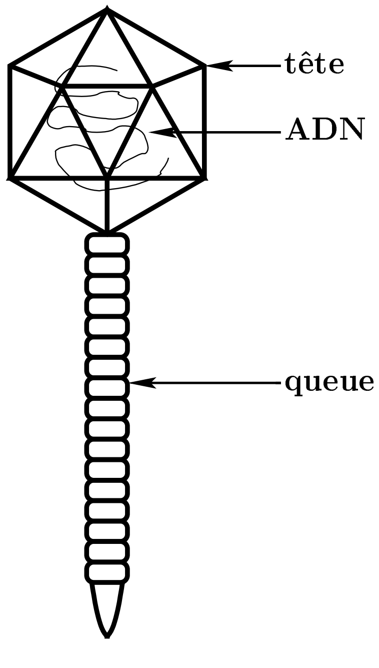


FIG. 6 – *le λ-phage*

Ce comportement présente une caractéristique fréquente des réseaux génétiques : une **bascule** (ou *switch*). En effet, à l'échelle de la cellule, les molécules synthétisées dans les états lytiques et lysogéniques sont très différentes et pourtant il suffit d'une variation minime dans la concentration d'une espèce critique pour basculer d'un état vers l'autre.

Comment s'explique cette bascule entre état lysogénique et état lytique?

L'unique gène qui est activé dans le lysogène code pour une protéine ayant une propriété bien particulière : le **répresseur** λ. Cette protéine est une **molécule régulatrice** en ce sens que son seul but est d'agir directement sur l'expression d'un ou de plusieurs gènes. Cette molécule est appelée répresseur car elle a la faculté de bloquer la transcription de tous les autres gènes du prophage (nous verrons plus tard comment se fait cette répression). Cette molécule a toutefois une autre action, celle-la positive, sur son propre gène. La production de cette molécule est telle qu'en régime permanent, on compte environ une centaine de répresseurs dans la bactérie (alors que deux suffisent pour bloquer la production). Grâce à cet excès, la bactérie est protégée contre toute autre invasion par un chromosome λ : elle ne pourra pas entrer dans le cycle lytique. La bactérie est dite alors immunisée contre l'infection λ.

En revanche, une irradiation par lumière ultraviolette a pour effet d'annuler l'action du répresseur. Il s'en suit la production d'une autre protéine régulatrice *cro* qui a l'effet inverse du répresseur : elle favorise la production des protéines nécessaires à la construction du virus (la bactérie entre dans l'état lytique) tout en bloquant la production de répresseur λ (l'état lytique est donc irréversible).

Nous voyons bien les deux positions de la bascule : d'un côté le répresseur λ inhibe toute production du prophage et de l'autre côté la protéine *Cro* déclenche la phase lytique irréversible. Nous allons étudier à présent comment fonctionne ce switch à l'échelle moléculaire.

2.1.2 Les composants de la bascule

– <u>L'ADN</u>

Le fonctionnement de la bascule est déterminé, comme nous l'avons évoqué plus haut par deux protéines régulatrices : le répresseur λ, ou *cI* et la protéine *Cro*. Afin de bien comprendre leurs rôles respectifs, nous devons nous familiariser avec les gènes codant pour ces protéines : cI et Cro. Comme nous le montre la figure 7, ces deux gènes sont adjacents sur la molécule d'ADN. Ils ne sont séparés que par une courte région regroupant les deux **promoteurs** P_{RM} et P_R. Le but de ces séquences est de se lier à l'ARN-polymérase de manière à initier la transcription d'un gène. L'opérateur P_{RM} permet d'initier la transcription du gène cI en dirigeant la polymérase vers la gauche, tandis que P_R permet d'initier la transcription du gène Cro en dirigeant la polymérase vers la droite.

Outre ces deux promoteurs, cette région d'ADN comporte trois sites aux propriétés bien particulières : les **opérateurs** OR_1, OR_2 et OR_3 (voir figure 7). Ces sites comportent 17 paires de bases, leurs séquences se ressemblent mais diffèrent légérement. Ces séquences déterminent l'affinité avec les protéines *cro* et cI.

– Le répresseur cI et la molécule Cro

Le répresseur cI est une chaîne de 236 acides aminés composée de deux groupes compacts reliés par une chaîne de 40 acides aminés. Ces deux groupes portent le nom de carboxyle et de aminé. Dans un lysogène, une majeure partie de ces molécules (environ 95%) est sous forme de dimères, c'est-à-dire qu'elles se lient deux par deux pour former un ensemble plus gros. Cette liaison protéine-protéine se fait au niveau des groupements carboxyles (voir figure 8). Au niveau des groupements aminés, le dimère a une affinité pour les trois opérateurs OR_i du chromosome λ.

La protéine Cro est une chaîne de 66 acides aminés, elle n'est formée que d'un seul groupe. Cette molécule ne se trouve présente dans la cellule que sous forme de dimère. Comme pour le répresseur, cette molécule a une affinité naturelle pour les trois opérateurs de la molécule d'ADN.

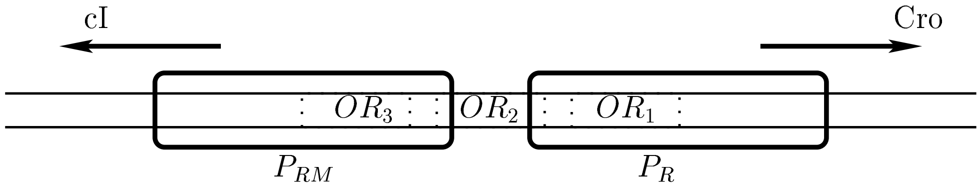


FIG. 7 – *les gènes cI et Cro*

L'action du répresseur est double. Son nom de *répresseur* lui vient du contrôle négatif qu'il impose sur la transcription du gène Cro. Mais on a découvert récemment que le répresseur exerçait un autre contrôle, celui-là positif, sur son propre gène. De par sa séquence spécifique, l'opérateur OR_1 est celui qui présente la plus grande affinité avec le dimère. Mais lorsqu'un dimère se lie au site OR_1, l'affinité du site OR_2 se trouve augmentée et on constate donc que les deux sites OR_1 et OR_2 se remplissent presque simultanément. A 90%, on peut donc dire que les deux sites OR_1 et OR_2 sont remplis et que le site OR_3 est vide. L'effet de cette liaison avec la molécule d'ADN est alors double. En effet, la liaison du répresseur à OR_2 empêche l'ARN-polymérase de se lier au promoteur P_R et bloque donc la transcription du gène Cro. D'autre part, cette liaison a pour effet d'augmenter l'affinité de la polymérase avec le promoteur P_{RM} (car en plus de la liaison ADN-protéine il y a une liaison protéine-protéine entre le dimère et la polymérase) et donc d'encourager la production de répresseur. Ce mécanisme est bien sûr contrôlé, sinon la production de cI n'aurait pas de raison de stopper. Ce contrôle se fait grâce au troisième opérateur : OR_3. En effet, nous avons dit plus haut que l'affinité de ce site était moindre. Toutefois, lorsque la concentration en dimère augmente et dépasse un certain seuil, son effet l'emporte et la liaison dimère-OR_3 se fait. Cela a alors pour effet de bloquer la transcription du gène cI en empéchant la polymérase de se lier au promoteur P_{RM}.

Dans un lysogène, la production de répresseur est donc régulée de manière à obtenir une quantité moyenne d'environ cent molécules. Ce répresseur bloquant le gène Cro, le cycle lytique est impossible et la bactérie est maintenue dans l'état lysogénique. Il est important de noter dans ce mécanisme la coopération qui existe entre les trois opérateurs. OR_1 est occupé en premier de par son affinité intrinsèque au dimère. Par contre, OR_2 et OR_3 ont la même affinité mais grâce à la liaison OR_1-dimère, c'est OR_2 qui est privilégié (OR_3 devant se lier indépendemment).

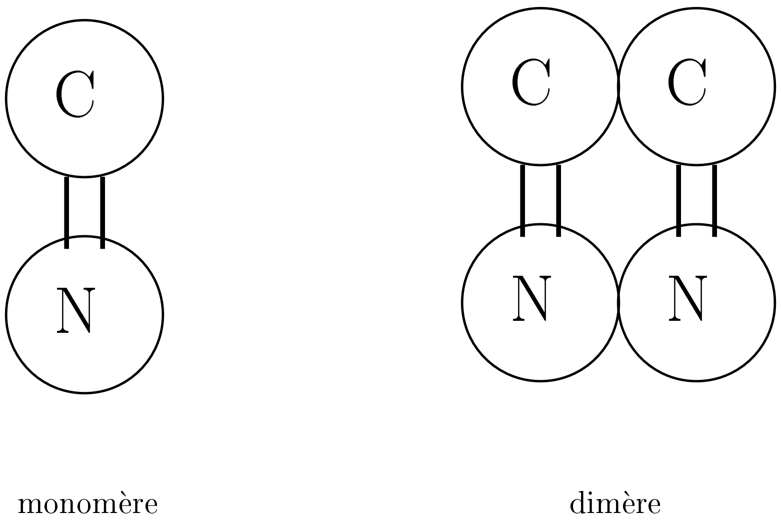


FIG. 8 – *la molécule cI*

L'action de *Cro* est plus simple car il n'y a pas de coopération. Le dimère se lie d'abord à OR_3 (qui a cette fois l'affinité la plus grande) puis, lorsque la concentration augmente, aux sites OR_2 et OR_1. La liaison avec OR_3 bloque tout d'abord la production de répresseur (c'est le début de la phase lytique), puis la concentration de Cro augmentant, les liaisons avec OR_2 et OR_1 bloquent à leur tour la production de Cro (il s'agit là de la deuxième partie de la phase lytique : le processus de lyse est démarré et le phage n'a plus besoin de produire de Cro).

Nous allons maintenant tenter de comprendre pourquoi, suite à une irradiation par lumière ultra-violette, la cellule bascule de l'état lysogénique vers l'état lytique.

2.1.3 Le déclenchement de la bascule

Les rayons UV ont des effets dramatiques sur les cellules en général. Cet agent endommage généralement la molécule d'ADN de manière encore peu connue. Ceci a pour effet de modifier une protéine bien spécifique : RecA. En temps normal, cette protéine a pour rôle la recombinaison de l'ADN mais après irradiation, elle devient une redoutable protéase qui coupe les monomères de cI. Le répresseur devient alors inactif, il est de moins en moins produit (car il est son propre activateur) et le gène Cro est bientôt transcrit, ce qui a pour effet de lancer la phase lytique.

Des expériences ont été menées afin d'étudier cette bascule d'un point de vue quantitatif. Dans l'état lysogénique, la molécule cI réprime l'expression de Cro à environ un millième de fois son taux maximal. Si l'on divise par environ deux ou trois la quantité de répresseur, il reste une quantité suffisante de cI pour maintenir l'état lysogénique. En revanche, une chute d'environ 5 fois la quantité de cI suffit à ramener la production de Cro à environ 50% de son taux maximal. Ceci crée suffisamment de Cro pour basculer vers l'état lytique.

Nous assistons à un phénomène fréquent dans les réseaux génétiques : l'état macroscopique du système ne dépend que de la concentration d'une espèce, ici le répresseur cI. En outre, comme dans beaucoup de bascules génétiques que nous connaissons, la concentration critique est très faible et pourtant son dépassement a des effets notables au niveau de la cellule même. Dans l'exemple du phage λ, cette sensibilité au répresseur s'explique par deux choses : tout d'abord le répresseur a une action positive sur sa propre production, ensuite le mécanisme d'autorégulation que nous avons vu (au niveau de l'opérateur OR_3) fait que la quantité de molécules en régime permanent est maintenue à un certain taux. Ce taux se situe suffisamment au-dessus du taux critique pour permettre à la fois l'immunisation de l'hôte à une nouvelle infection et la possibilité de déclencher la bascule en cas d'irradiation.

L'étude du phage λ constitue un excellent exemple de réseaux génétiques tels que nous les connaissons aujourd'hui. Nous venons de voir l'effet de bascule, qui permet de passer d'un état à un autre, nous allons maintenat étudier plus précisément le réseau qui sous-tend linfection λ. Nous allons en particulier tenter de répondre à la question : comment une bactérie décide-t-elle son chemin métabolique? Ou, en d'autres termes, quels sont les facteurs qui font qu'elle entre initialement dans la phase lytique ou dans la phase lysogénique?

2.2 Étude du circuit génétique du phage λ

Dans la partie précédente, nous avons présenté un phénomène particulier : la bascule de l'état lysogénique vers l'état lytique. Toutefois le comportement du phage λ est beaucoup plus complexe que cela. Après plusieurs années d'études et d'expériences, les connaissances sur ce bactériophage sont bien avancées et l'on commence à avoir une idée assez précise du *programme génétique* qui sous-tend son comportement. Dans cette partie, nous nous proposons d'étudier comment se fait l'infection d'un hôte et comment le phage *décide* du chemin métabolique qu'il va emprunter.

2.2.1 le chromosome λ

Nous connaissons bien aujourd'hui le chromosome λ. La figure 9 représente ses principaux gènes. Nous remarquons sur cette figure qu'il existe deux types de gènes. Les premiers sont les gènes qui codent pour des protéines directement utilisées dans le métabolisme cellulaire. Généralement (et c'est le cas pour le phage), ces derniers peuvent être regroupés selon leur fonction. Ainsi, la partie nommée *tête* dans la figure 9 contient les 10 gènes codant pour les protéines qui forment la tête du phage. La partie *recombinaison* regroupe les enzymes qui codent pour les opérations touchant la molécule d'ADN, notamment celles qui intègrent le chromosome λ dans le chromosome de l'hôte (pour la lysogénisation) ou au contraire qui le détachent (pour la lyse). La partie *réplication* contient les gènes codant pour la recopie de la molécule d'ADN et la partie *lyse* contient trois gènes codant pour la desctruction de la cellule. L'avantage d'avoir ces gènes regroupés par fonction est évident : ils sont régulés de la même manière et donc une fois que l'ARN-polymérase commence la transcription, elle transcrit tout le segment correspondant.

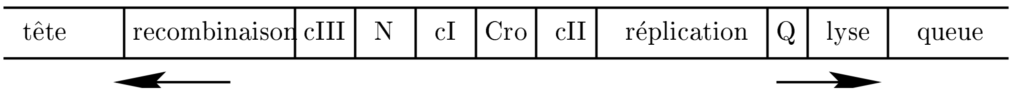


Fig. 9 – *le chromosome λ (seulement les gènes)*

Le deuxième type de gènes rassemble tous les gènes codant pour des protéines régulatrices. Contrairement aux précédents, ils sont souvent seuls car leur fonction est unique. Ce sont eux qui contrôlent la transcription des gènes précédents. On parle souvent de réactions *en cascade* car les protéines régulatrices déclenchent ou inhibent la transcription de plusieurs gènes.

2.2.2 les premières étapes de l'infection

Nous allons étudier à présent les toutes premières étapes de la vie du phage à l'intérieur de son hôte. Tout d'abord, le chromosome λ se circularise : il restera sous cette forme tout le temps de l'infection. A ce stade il commence à créer ses propres protéines. L'étude du **circuit génétique** commence.

L'ARN-polymérase de l'hôte se fixe tout d'abord sur deux promoteurs afin de transcrire les gènes N et Cro. Le rôle de Cro est connu. Intéressons-nous plutôt à la molécule N. Cette protéine régulatrice a une action positive différente de celle que nous connaissons. Elle se lie sur certains sites de l'ADN et au passage de la polymérase, elle modifie cette dernière de telle façon qu'elle ignore certains signaux d'arrêt qu'elle va rencontrer. Ainsi, dès que suffisamment de protéine N est présente dans le milieu, la polymérase continue son chemin vers la gauche, de N jusqu'aux gènes de recombinaison, ainsi que vers la droite, de Cro jusqu'à Q.

A ce niveau d'avancement, la cellule entre dans l'un des deux **chemins métaboliques** : lytique ou lysogénique. Nous allons rapidement évoquer les opérations relatives à ces deux états, puis, dans la partie suivante, nous essaierons de comprendre comment se fait le choix du chemin.

Si la cellule prend le chemin de la lyse, la molécule Q, qui a le même rôle d'anti-terminaison que N, perpétue la transcription vers la droite de manière à transcrire les gènes commandant la lyse, la production de la tête et de la queue. Parallèlement à ceci, la protéine Cro annule la transcription de N et de ses suivants, puis de Cro lui-même et de ses suivants. Il n'y a alors plus que *tête* et *queue* qui produisent de nouveaux phages, tandis que le chromosome λ est intensivement répliqué. Au bout d'un certain temps la bactérie contient de nombreux nouveaux virus, les molécules de lyse cassent alors la membrane cellulaire et les phages s'échappent pour aller infecter les bactéries voisines.

Au contraire, si elle prend le chemin lysogénique, il faut qu'elle produise une grande quantité de répresseur. Or, nous venons de voir qu'elle produisait la molécule Cro qui empêche sa production. En fait, la production de cI ne se fait pas comme nous l'avons vu plus haut : puisque le promoteur P_{RM} est occupé par des molécules Cro, le gène cI est transcrit à partir d'un autre promoteur P_{RE} (RM veut dire *repressor maintenance* tandis que RE signifie *repressor establishment*). La liaison de la polymérase avec P_{RE} se fait grâce à la molécule cII, présente dans le milieu. Une fois que suffisamment de cI est produite, la concentration de Cro diminue et cI peut contribuer à sa propre transcription et, en meme temps, bloquer toute autre transcription. Le régime lysogénique s'installe.

<u>Remarque</u> : *juste avant l'établissement du régime lysogénique, le gène Int qui fait partie de la recombinaison permet d'intégrer le chromosome λ dans le chromosome de l'E-coli. Il fera dès lors partie intégrante du matériel géné-*

tique de l'hôte et se transmettra ainsi à tous ses descendants.

2.2.3 le choix : lytique / lysogénique

Comme la description précédente nous le laisse pressentir, c'est la molécule cII qui a un rôle capital dans ce choix. En effet, outre son action de démarreur pour le répresseur cI, cette molécule a plusieurs fonctions inhibitrices, elle bloque notamment la production de Cro et de Q. Le choix du chemin réside donc dans l'activité de cette molécule.

L'activité de cII dépend beaucoup de facteurs environnementaux, notamment de l'activité des protéases qui détruisent cette molécule. L'activité de ces protéases dépend à son tour de nombreux facteurs. Ainsi, dans une bactérie qui s'est développée dans un milieu pauvre en alimentation, ces protéases sont relativement faibles, ce qui explique que la lysogénie est plus fréquente dans les cultures pauvres en alimentation. Ceci paraît en outre assez logique car les cellules peu alimentées ne possèdent pas suffisamment de composants permettant une phase lytique efficace.

A travers l'exemple du phage, nous avons tenté de décrire avec précision plusieurs des mécanismes qui nous semblent essentiels pour l'étude des réseaux de régulation génétique. Le premier est le mécanisme de bascule : il se situe généralement à l'échelle de la cellule et consiste en un changement, plus ou moins radical, d'une partie du matériel enzymatique de la cellule. Ce phénomène s'explique par un enchaînement généralement complexe d'opérations de contrôle génétique. Selon un programme précis, des protéines régulatrices déclenchent à certains moments la transcription de groupes de gènes ou au contraire inhibent certains gènes et ce processus se continue indéfiniment jusqu'à la mort de la cellule. C'est grâce à l'enchaînement en cascade de toutes ces transcriptions que la cellule peut répondre de façon adaptée aux changements parfois brusques de l'environnement.

La dernière partie évoque néanmoins une certaine adaptabilité de ces systèmes à l'environnement et aux conditions initiales. Une modification minime dans une concentration ou dans l'activité de différents enzymes peut être à l'origine d'un changement structurel important de la cellule. Ces systèmes s'adaptent à leur milieu extérieur, mais ils renferment également une grande part d'incertitude qui leur donne une instabilité relative.

Nous touchons là une subtilité propre aux systèmes biologiques. Il est souvent compliqué de modéliser de tels systèmes car il existe un équilibre subtil entre l'ordre et le chaos. En effet, une trop grande stabilité du système lui fait perdre sa faculté de s'adapter aux conditions extérieures. En revanche, si trop de gènes changent en même temps, l'instabilité qui en résulte a des conséquences désastreuses. Ainsi une cellule trop sensible aux perturbations extérieures évolue rapidement vers un état pathologique. La sélection naturelle laisse donc évoluer les organismes qui ont un équilibre subtil entre ordre et chaos.

3 Modélisation des réseaux génétiques à l'aide de systèmes hybrides

Les différents modèles qui ont été proposés pour représenter formellement les réseaux génétiques se rangent pour l'essentiel dans deux catégories : les systèmes discrets et les systèmes continus.

Les systèmes discrets semblent plutôt bien adaptés aux systèmes biologiques car ces derniers présentent des phénomènes facilement modélisables par des variables discrètes, notamment le phénomène de bascule précédemment décrit. Dans le système booléen, dont le précurseur est Stuart Kauffman ([8], [9],[10]), un gène est soit allumé, soit éteint, il est donc représenté par une variable booléenne qui vaut 1 ou 0. On modélise ensuite les influences positives et négatives des différents gènes par des fonctions booléennes. L'avantage d'une telle représentation est bien sûr sa simplicité. Elle nous permet notamment de trouver et d'étudier le circuit logique sous-jacent au réseau. Nous verrons dans un chapitre ultérieur l'intérêt de cette modélisation dans la recherche de réseaux génétiques.

Néanmoins, si l'on veut tenter de simuler un réseau génétique, l'approche booléenne, bien que relativement simple, est trop restrictive. Il devient alors intéressant de représenter l'évolution continue du système. Pour cela, on modélise le réseau par un système différentiel où les variables sont les concentrations en les différentes espèces produites. Cette modélisation a bien sûr ses limites, l'inconvénient majeur restant la dimension du système qui peut croître très vite.

Un autre inconvénient réside dans le fait que l'évolution temporelle de tels réseaux est complexe : de longs intervalles de temps où la variation est de faible amplitude sont ponctués par de brusques changements. C'est le cas des **bascules** dont nous avons donné un exemple pour le bactériophage. Ces comportements sont difficiles à visualiser sur des simulations numériques, surtout lorsque l'on veut simuler l'évolution temporelle sur une longue période de temps. L'idée d'utiliser des systèmes hybrides pour étudier de telles systèmes paraît alors très intéressante, car elle permet justement d'inclure dans le modèle ces phénomènes de *switch*.

3.1 Définition du modèle

L'autre avantage de la modélisation par système hybride est qu'elle permet d'allier l'approche discrète et l'approche continue. En effet, plus on étudie les réseaux génétiques, plus on se rend compte de leur dualité : si, à grande échelle, leur comportement temporel est continu, on conçoit assez bien que derrière chaque réseau se trouve une architecture discrète, qui ressemble à un circuit logique.

La définition du modèle se fait en plusieurs étapes :

3.1.1 première étape : isoler le réseau étudié

La première étape dans l'étude d'un réseau génétique est de bien isoler les variables d'intérêt. Les variables du système seront donc les concentrations en les différentes espèces qui interviennent dans la régulation globale du réseau. Il est important de noter ici que les espèces qui entrent en jeu ne sont pas seulement des protéines.

En étudiant chaque espèce d'un peu plus près, on se rend compte que certaines d'entre elles ont un rôle particulier. Considérons le cas d'une induction, par exemple celle de la molécule Cro dans le phage λ. En présence de répresseur, le gène Cro est bloqué. L'expression ne se fait pas du tout et la protéine n'est donc pas produite. Lorsque la concentration en répresseur diminue, les molécules de répresseur ont moins de chances de se fixer à l'opérateur et donc l'amorce de la transcription peut se faire. En première approximation, on peut considérer que lorsque la concentration en répresseur passe au-dessous d'un certain seuil, l'induction commence.

La molécule cI a donc un rôle régulateur. En revanche, pour les protéines servant à construire le phage (tête et queue), seule la valeur de la concentration est importante. On distinguera donc les espèces au rôle purement régulateur des espèces qui ont un rôle différent.

Les **variables continues** du système hybride sont donc :

$$\begin{array}{rcll} \mathcal{X} & = & \{x_1, x_2, \ldots, x_n\} & : \text{espèces régulatrices} \\ \mathcal{Y} & = & \{y_1, y_2, \ldots, y_m\} & : \text{autres espèces} \end{array}$$

Comme ce sont toutes des concentrations, nous avons bien sûr :

$$\forall i, j : \quad x_i \geq 0 \text{ et } y_j \geq 0$$

3.1.2 deuxième étape : partitionner l'espace des phases

Afin de modéliser le circuit logique sous-jacent, il nous faut donner une valeur logique aux espèces régulatrices. En première approximation, nous pouvons considérer qu'une espèce i est *présente* si elle est au-dessus d'un certain seuil θ_i et *absente* si elle est en-dessous. Une description plus fine prendrait en compte plusieurs seuils pour désigner plusieurs niveaux de présence de la molécule, mais cela n'est pas nécessaire pour des réseaux simples.

Il faut également définir pour chaque espèce i une concentration maximale max_i qu'elle ne peut pas dépasser sans être nocive pour la cellule. Nous avons donc :

$$\forall i, \quad x_i \in [0, \theta_i] \text{ ou } x_i \in [\theta_i, max_i]$$

On a ainsi une partition (P_r) de l'espace des phases avec :

$$P_r = \left(\prod_{j \in I_r} [0, \theta_j]\right) \times \left(\prod_{j \notin I_r} [\theta_j, max_j]\right)$$

et nous associons à chacune de ces régions un **état discret** q_r du système hybride. Ainsi, à un état donné, on sait pour chaque espèce régulatrice si elle est *allumée* ou *éteinte*. Nous noterons Q l'ensemble des états q_r.

3.1.3 troisième étape : Les équations différentielles

Les **systèmes dynamiques** associés à chaque état sont définis sur :

$$\Omega = \left(\prod_{j=1}^{n} [0, max_j]\right) \times \left(\prod_{j=1}^{m} [0, max_j]\right)$$

Ils doivent de plus respecter des **contraintes** définies pour chaque état de la manière suivante :

$$\forall q_r \in Q, \quad (x_1(q_r), x_2(q_r), \ldots, x_n(q_r)) \in P_r$$

On notera $\mathcal{H}(q_r) = P_r$ et $\mathcal{H} = \bigcup_{q \in Q} \mathcal{H}(q)$.

Pour déterminer les systèmes dynamiques associés à chaque état, on utilise les équations différentielles données par la cinétique chimique. Pour cela, il nous faut d'abord établir les interactions entre les différents composants présents dans le milieu.

Il nous faudra également introduire des termes de production sous forme de constantes de production γ_{esp} pour modéliser la fabrication de protéines à partir de l'ADN. Enfin, on prévoira des termes de dégradation (termes linéaires avec un taux ν_{esp}) qui correspondent à l'usure naturelle des molécules organiques au cours du temps. Ces derniers seront présents pour toutes les protéines et dans toutes les équations; en revanche, les termes de production dépendront de l'état dans lequel le système se trouve.

Comme nous l'avons vu plus haut, les équations de la cinétique enzymatique, même les plus simples, introduisent des termes au moins quadratiques qui nous empêchent une résolution exacte des systèmes différentiels. Une résolution numérique sera donc nécessaire afin d'approcher les trajectoires réelles.

3.1.4 quatrième étape : Les transitions

Les transitions sont un point délicat. Dès que la trajectoire parvient à la frontière d'un état, elle *saute* dans l'état suivant en prenant pour conditions initiales les valeurs du moment où la frontière a été atteinte. Pour reprendre les termes de la théorie des systèmes hybrides, il existe une **transition** entre tous les états voisins : $e = q \to q'$ et la frontière correspondante constitue la **garde** $G(e)$. La fonction $Reset$ associée à la garde $G(e)$ est l'identité :

$$\forall X \in G(e), \quad Reset_{q \to q'}(t_-, X(t_-)) = (t_+, X(t_+))$$

On notera $\mathcal{E}$ l'ensemble des transitions.

En notant $Init$ l'ensemble des conditions initiales possibles de notre système, nous pouvons complètement le déterminer :

$$\boxed{\mathcal{S} = (Q, \{\mathcal{X}, \mathcal{Y}\}, \Omega, \mathcal{H}, \mathcal{E}, G, Reset, Init)}$$

Il est important de noter ici que la compréhension biologique du système étudié est très importante car elle permet dans la pratique de simplifier cette définition. En particulier, certaines transitions ne sont pas plausibles biologiquement, ou encore certaines conditions initiales sont aberrantes car impossibles à trouver dans la nature.

3.2 Un exemple de modélisation : l'opéron lactose

Afin de tester notre modèle sur de vrais réseaux génétiques, nous avons décidé de l'appliquer à un réseau simple et relativement connu, **l'opéron lactose** découvert en 1965 par Monod et Jacob. L'étude de ce modèle nous permettra en outre d'évoquer la notion d'opéron qui est capitale dans l'étude des réseaux génétiques. Nous commencerons par donner une brève description du mécanisme principal de l'opéron puis nous verrons la modélisation qui en a été faite.

3.2.1 description de l'opéron lactose

L'opéron lactose est un réseau simple mais fondamental car il permet, en cas de pénurie de glucose, d'exploiter un autre sucre, certes moins efficace : le lactose. Pour cela, la cellule a besoin principalement de deux enzymes : la **perméase** qui permet de faire pénétrer le lactose à l'intérieur de la cellule et la **β-galactosidase** qui permet de le dégrader en glucose. Chez la bactérie *Escherichia-coli*, les deux gènes qui codent pour ces enzymes sont côte à côte, le long du chromosome unique de la bactérie (voir figure 10). Un opérateur les précède, sur lequel peut se lier un répresseur, produit de façon constante par un autre gène : I.

S'il n'y a pas de lactose, il ne sert à rien d'entamer la production des enzymes, l'opéron est donc bloqué grâce au répresseur (c'est le cas sur la figure 10. En revanche, dès que du lactose est présent, les quelques molécules de perméase présentes dans la cellule font entrer quelques molécules de lactose. Le lactose, qui est **l'inducteur** de l'opéron se lie alors au répresseur pour lui ôter son affinité avec l'opérateur. La polymérase peut alors transcrire les deux gènes : l'opéron est débloqué ou **induit**. Grâce à la perméase produite, le lactose entre alors en masse dans la cellule et se dégrade en glucose en vue de fournir de l'énergie. Cette boucle constitue la boucle **négative** de l'opéron. Il existe également une boucle **positive**.

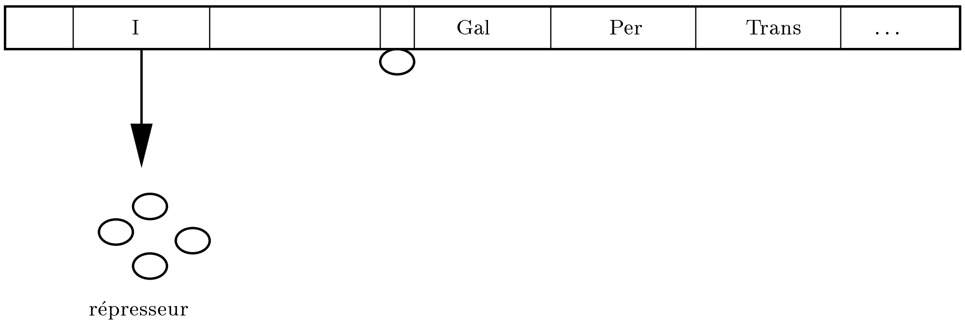


FIG. 10 – *opéron lactose bloqué*

En effet, même si la cellule a à sa disposition du lactose, il ne lui sert à rien si elle dispose de meilleures sources d'énergie (c'est-à-dire du glucose). Afin de pallier à ce problème, il existe une molécule qui active la production : la CAP (*catabolite activator protein*). Pour que cette molécule se fixe sur l'ADN, elle a besoin de créer un complexe avec l'AMP cyclique (*c-AMP*). Ce complexe prend alors une forme particulière qui lui permet de se fixer à l'ADN au niveau du promoteur activant ainsi la transcription. Or, l'AMP-c est une molécule qui n'est présente que lorsque le glucose vient à manquer. Cette régulation positive permet d'accélérer la transcription s'il n'y a plus de glucose. Cette accélération (expérimentalement d'un facteur 50) permet ainsi de contrôler l'induction de l'opéron : l'induction démarre faiblement si le milieu contient du lactose (la transcription reste relativement basse) et elle s'accélère vivement si, en plus, le glucose passe au-dessous d'un certain seuil.

Remarque : le fonctionnement complet de l'opéron n'est pas encore connu de façon certaine aujourd'hui. Il existe notamment un troisième gène qui suit le gène de la perméase et qui code pour une autre enzyme, la transacétylase. Le rôle de cette enzyme dans le processus de la consommation lactique n'est pas encore établi.

3.2.2 modèle hybride de l'opéron lactose

– Les variables

Le premier travail est de recenser les variables continues utiles pour notre problème. Il faut aussi repérer parmi ces variables celles qui font partie du circuit logique sous-jacent (les molécules au rôle régulateur). On note r la concentration du répresseur, a celle de l'activateur (on considère en première approximation que l'activateur est représenté par l'AMP-c et que l'AMP-c est produite avec un taux constant en absence de glucose). x et y représente respectivement les concentrations de β-galactosidase et de perméase (la trans-acétylase n'intervenant pas, il est inutile de la représenter). Enfin, les deux variables qui nous servent pour le circuit logique sont la concentration en lactose interne l et en glucose interne g.

– Les états

Les états discrets du système sont dictés par les concentrations en glucose et en lactose. Il y a donc quatre états. Notons G (respectivement L) si on est en présence de glucose (respectivement de lactose), c'est-à-dire si $g > \theta_g$ et $\bar{G}$ (respectivement $\bar{L}$) sinon. Les quatre états discrets se notent donc : $GL,\ \bar{G}L,\ G\bar{L},\ \bar{G}\bar{L}$. Etudions les plus précisément.

1. $\bar{G}\bar{L}$: signifie que la cellule est à cours de glucose et de lactose. Ceci est un état mortel car elle ne dispose plus d'énergie. En fait, il existe d'autres sources possibles d'énergie mais pour cela elle doit allumer d'autres opérons et nous sortons donc du cadre de notre modèle. Dans notre simulation, la cellule reste dans cet état jusqu'à ce qu'elle obtienne à nouveau du glucose ou du lactose (par une source extérieure).

2. $G\bar{L}$: c'est l'état le plus fréquent : la cellule dispose de ressources suffisantes en glucose et il n'y a pas de lactose dans le milieu. Dans ce cas l'opéron est **bloqué**.

3. GL : dans cet état, la cellule dispose des deux sucres. Elle choisira alors préférentiellement le glucose qui est plus facile d'utilisation. L'opéron est donc débloqué (puisqu'il y a du lactose dans le cellule) mais sa transcription est désactivée (la boucle positive n'a pas lieu). On parle d'**induction modérée**.

4. $\bar{G}L$: c'est l'état d'**induction activée**: le lactose est présent, par contre il n'y a plus de glucose, l'opéron *tourne* donc à plein régime.

– Les équations

Afin de ne pas surcharger le rapport, on se reportera en annexe pour voir le système dynamique associé à l'état de répression (par exemple). Nous ferons ici quelques remarques importantes :

Tout d'abord, il nous faut introduire deux variables de plus qui représentent les sources extérieures de sucre: g_{ext} et l_{ext}. Ces variables ne font pas partie des variables intérieures précédemment décrites car elles sont entrées par l'utilisateur. Elles peuvent être vues comme des variables de contrôle : c'est en imposant les ressources extérieures dont dispose la cellule que l'on peut influer sur son comportement.

Ensuite, afin de modéliser les réactions enzymatiques suivantes:

$$\begin{array}{lcl} \text{lactose extérieur} & \xrightarrow{\text{perméase}} & \text{lactose intérieur} \\ \text{lactose} & \xrightarrow{\beta\text{-galactosidase}} & \text{glucose} + \text{galactose} \end{array}$$

il nous faut introduire deux autres variables internes qui correspondent aux complexes enzyme-substrat : c_x pour le complexe lactose-galactosidase et c_y pour le complexe lactose extérieur-perméase. Nous pouvons alors utiliser les équations de la cinétique que nous avons déjà évoquées plus haut.

Remarque : un problème se pose alors car les équations font intervenir la concentration en substrat, or le substrat de la première réaction n'est pas une variable du problème mais une donnée. Afin de modéliser la pénétration du lactose dans la cellule, nous avons introduit une autre quantité : l_m qui correspond au substrat en question. Cette quantité est directement alimentée par l'extérieur mais c'est une variable du problème et elle entre donc dans les équations différentielles.

3.2.3 résultats

Dans cette partie, nous allons présenter quelques résultats auxquels nous a conduit ce modèle. Notons que ces résultats donnent lieu à des interprétations qualitatives, mais ne permettent pas de tirer de conséquences quantitatives. Il faudrait pour cela disposer de plusieurs constantes chimiques (notamment des constantes de cinétique) avec précision. Toutefois, les résultats obtenus permettent d'illustrer les phénomènes de régulation que nous avons décrits.

Afin d'obtenir ces résultats, nous avons résolu les systèmes différentiels numériquement, la résolution exacte étant impossible. Nous avons donc utilisé la méthode de Runge-Kutta d'ordre 4. Nous rappelons la formule. Pour un système différentiel vectoriel

$$\dot{X} = F(t, X) \quad \text{où } X \in \mathbb{R}^n$$

l'itération de Runge-Kutta d'ordre 4 s'écrit :

$$X_{n+1} = X_n + \frac{\Delta t}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

avec :

$$\begin{cases} k_1 &= F\left(t_n, X_n\right) \\ k_2 &= F\left(t_n + \frac{\Delta t}{2}, X_n + k_1 \frac{\Delta t}{2}\right) \\ k_3 &= F\left(t_n + \frac{\Delta t}{2}, X_n + k_2 \frac{\Delta t}{2}\right) \\ k_4 &= F\left(t_n + \Delta t, X_n + k_3 \Delta t\right) \end{cases}$$

Les courbes obtenues grâce au logiciel MATLAB nous permettent notamment de retrouver les caractéristiques des réseaux génétiques. Nous voyons par exemple sur la figure 11 le phénomène d'induction : le lactose rentre d'abord doucement dans la cellule (la variation n'est presque pas visible sur le graphe). Le répresseur commence alors à augmenter. Parallèlement, étant donné que la cellule est privée de glucose, l'augmentation du taux de lactose interne fait augmenter la concentration d'activateur. Lorsque le répresseur passe au-dessous d'un certain seuil, l'opéron démarre et le lactose entre en masse dans la cellule. Ce démarrage est activé en plus par le fait que l'activateur monte : c'est le phénomène d'induction activée. Le premier graphe nous donne les concentrations d'activateur et de répresseur, le deuxième nous donne les concentrations en lactose extérieur et intérieur tandis que le troisième nous donne les concentrations en enzymes.

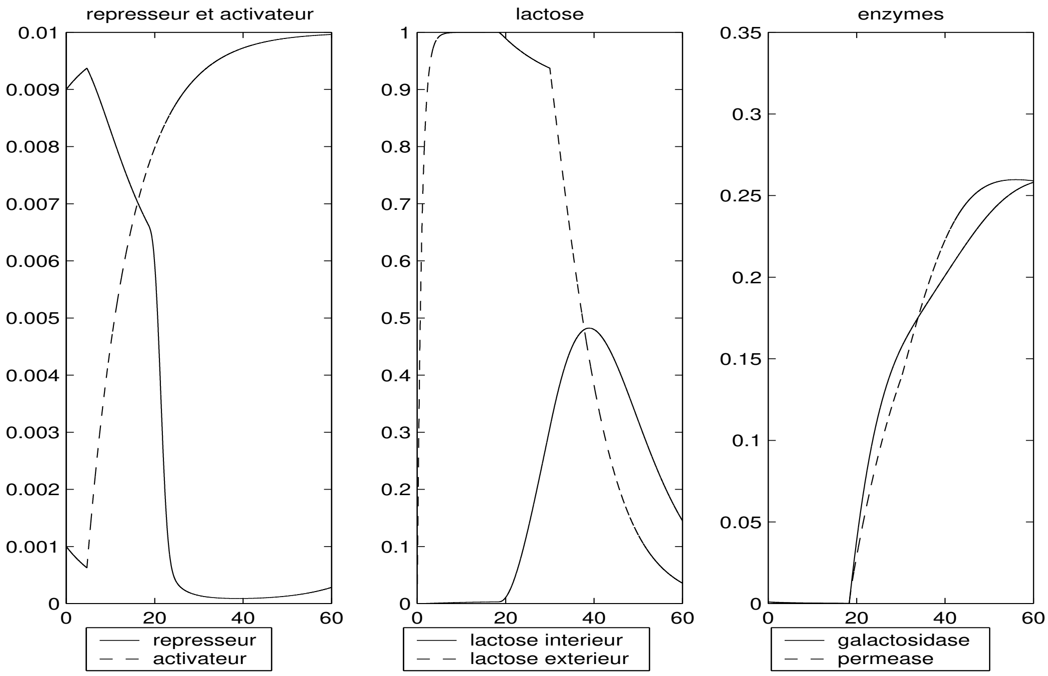


FIG. 11 – *simulation de l'induction lactique : induction activée*

La figure 12 illustre le phénomène d'induction modérée. Il s'agit de la même expérience que pour la figure 11, mais cette fois on introduit du glucose entre les temps $t = 40$ et $t = 50$. Ceci provoque une chute de l'activateur ce qui a pour effet de faire tomber la concentration en enzymes. Les taux de dégradation sont ici exagérés de manière à montrer ce phénomène de façon nette. Dans la réalité, les enzymes mettent plus de temps à se dégrader. On n'observe donc pas de chute de la concentration mais une stabilisation.

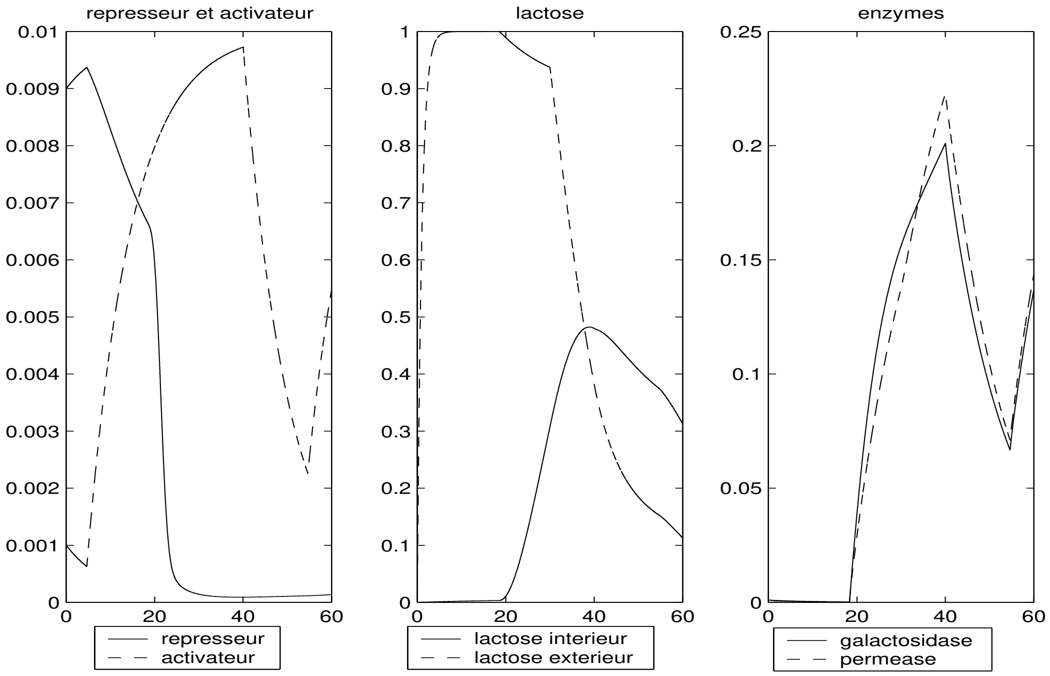


FIG. 12 – *simulation de l'induction lactique : induction modérée*

Nous venons de voir comment les réseaux génétiques peuvent se modéliser à l'aide de systèmes hybrides. Pour cela, nous avons utilisé une formule d'intégration numérique afin de résoudre de manière approchée les équations de Michaelis-Menten. Dans la partie suivante, nous allons nous intéresser à ces équations d'un point de vue théorique. Nous allons notamment étudier l'hypothèse de quasi-stationnarité établie par Michaelis et Menten en 1913.

4 Les équations de Michaelis et Menten (1913)

Cette partie est consacrée à l'étude des équations de Michaelis et Menten qui représentent l'activité enzymatique. Nous nous proposons d'en apporter une analyse théorique en vue de comprendre le comportement exact des solutions de ces équations (voir [3]).

4.1 Présentation du problème

Schématiquement, comme nous l'avons déjà vu au début de ce rapport, la réaction de catalyse biochimique la plus simple que nous puissions imaginer s'exprime par l'équation chimique suivante :

$$S + E \underset{k_{-1}}{\overset{k_{+1}}{\rightleftarrows}} SE \xrightarrow{k_2} P + E$$

C'est l'équation de catalyse d'un substrat unique par un enzyme. Grâce aux lois de base de la cinétique chimique, nous arrivons au système différentiel suivant :

$$\begin{cases} \dfrac{ds}{dt} = -k_1 es + k_{-1}c & , \quad s(0) = s_0 \\[2ex] \dfrac{de}{dt} = -k_1 es + (k_{-1} + k_2)c & , \quad e(0) = e_0 \\[2ex] \dfrac{dc}{dt} = k_1 es - (k_{-1} + k_2)c & , \quad c(0) = 0 \\[2ex] \dfrac{dp}{dt} = k_2 c & , \quad p(0) = 0 \end{cases}$$

où s, e, c et p représentent les concentrations en substrat, enzyme, complexe enzyme-substrat et produit. La quatrième équation nous fournit p connaissant c par la formule :

$$p(t) = k_2 \int_0^t c(\tau)d\tau$$

De plus, en sommant les deux équations du milieu, on obtient :

$$\frac{d}{dt}(e + c) = 0$$

Ceci s'interprète tout simplement par le fait que l'enzyme n'intervient pas dans la réaction globale, sa concentration (égale à la concentration en enzyme libre plus celle de l'enzyme lié) reste constante au cours du temps. D'où :

$$e(t) = -c(t) + e_0$$

Nous sommes donc ramenés à la résolution du système :

$$\begin{cases} \dfrac{ds}{dt} = -k_1 e_0 s + (k_1 s + k_{-1}) c & , \quad s(0) = s_0 \\ \\ \dfrac{dc}{dt} = k_1 e_0 s - (k_1 s + k_{-1} + k_2) c & , \quad c(0) = 0 \end{cases}$$

Grâce à l'adimensionalisation suivante :

$$\begin{cases} u(\tau) = \dfrac{s(t)}{s_0} \\ v(\tau) = \dfrac{c(t)}{e_0} \end{cases}$$

$$\tau = k_1 e_0 t, \quad \lambda = \frac{k_2}{k_1 s_0}, \quad K = \frac{k_{-1} + k_2}{k_1 s_0} et \quad \epsilon = \frac{e_0}{s_0}$$

on obtient le système sans dimension suivant :

$$(\mathcal{S}) \begin{cases} \dfrac{du}{d\tau} = -u + (u + K - \lambda) v & , \quad u(0) = 1 \\ \\ \epsilon \dfrac{dv}{d\tau} = u - (u + K) v & , \quad v(0) = 0 \end{cases}$$

Ce système est non linéaire et ses solutions ne peuvent s'exprimer par formules. En revanche, on peut le simuler numériquement, ce que nous avons fait grâce au logiciel MATLAB. La figure 13 présente les résultats. On remarque un phénomène intéressant : le système semble avoir un comportement particulier à l'origine. En effet, bien que les solutions soient régulières, elles présentent une brusque variation au voisinage de 0 qui n'est visible que grâce à un changement de l'échelle de temps. Ce phénomène est connu sous le nom de **couche limite**.

4.2 L'hypothèse de quasi-stationnarité

Le phénomène vient du paramètre $\epsilon = \frac{e_0}{s_0}$ qui est dans la pratique très petit. Ceci s'explique par le fait que dans la nature, les réactions biochimiques nécessitent une concentration en enzymes faible devant celle du substrat. C'est là un grand avantage des enzymes biologiques : ce sont des catalyseurs très efficaces et il suffit d'en produire peu pour accélérer de manière notoire une réaction. En pratique, on peut considérer généralement que ϵ est de l'ordre de 10^{-2} à 10^{-7}.

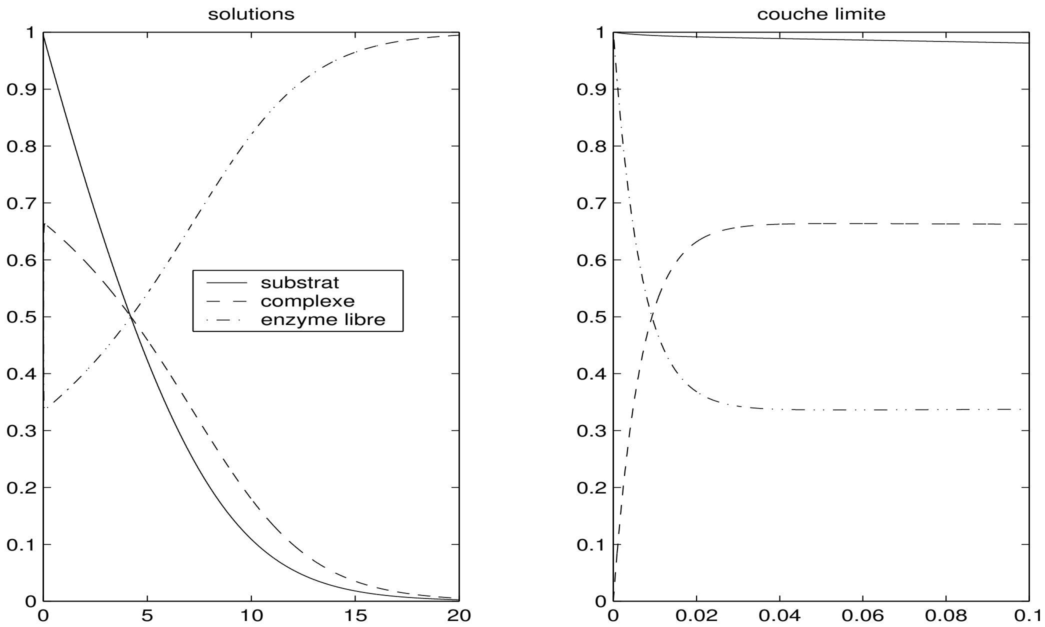


FIG. 13 – *simulation numérique du système : visualisation de la couche limite. Le deuxième graphe représente le phénomène de couche limite que l'on peut observer au voisinage de l'origine.*

4.2.1 approximation extérieure

En supposant que u et v admettent un développement asymptotique, nous pouvons écrire :

$$\left\{\begin{array}{lcl} u(\tau,\epsilon) & = & \displaystyle\sum_{n=0}^{N_1} \epsilon^n u_n(\tau) \\ & & \\ v(\tau,\epsilon) & = & \displaystyle\sum_{n=0}^{N_2} \epsilon^n v_n(\tau) \end{array}\right.$$

En remplaçant u et v par leurs développements asymptotiques dans le système $(\mathcal{S})$, puis en ne retenant que les termes d'ordre 0 (en ϵ) nous obtenons :

$$(\mathcal{E})\left\{\begin{array}{lcll} \dfrac{du_0}{d\tau} & = & -u_0+(u_0+K-\lambda)v_0 & ,\quad u_0(0)=1 \\ & & & \\ 0 & = & u_0-(u_0+K)v_0 & ,\quad v_0(0)=0 \end{array}\right.$$

Nous remarquons que la deuxième équation perd un ordre : ce n'est plus un équation différentielle mais algébrique. C'est cette équation algébrique qui traduit le quasi-équilibre. Ceci pose un problème car le système obtenu ne vérifie pas les conditions initiales :

$$u(0)=1 \quad \text{mais} \quad v(0)=\frac{1}{1+K}\neq 0$$

Nous voyons la simulation numérique du système approché $(\mathcal{E})$ sur la figure 14. Nous remarquons en les superposant avec les courbes de la figure 13 que l'approximation est bonne, sauf au voisinage de l'origine (c'est-à-dire hors de la couche limite). Le système $(\mathcal{E})$ traduit donc le quasi-équilibre du système initial $(\mathcal{S})$. On appelle ce système **approximation asymptotique extérieure** ou **non singulière** du système initial. La résolution de ce système conduit à l'équation scalaire suivante, dépendant d'une constante d'intégration A :

$$u_0(\tau)+K\ln(u_0(\tau))=A-\lambda\tau$$

4.2.2 approximation de la couche limite

Afin d'approximer les solutions du système initial à l'intérieur de la couche limite, nous allons encore procéder grâce aux développements asymptotiques. Mais comme nous venons de le voir, un développement asymptotique direct de u et de v ne permet pas de rendre compte de la couche limite. Afin de la visualiser, nous allons auparavant procéder à un changement de variable afin de dilater le temps au voisinage de l'origine. En posant :

$$\sigma=\frac{\tau}{\epsilon}$$
$$U(\sigma,\epsilon)=u(\tau,\epsilon) \quad et \quad V(\sigma,\epsilon)=v(\tau,\epsilon)$$

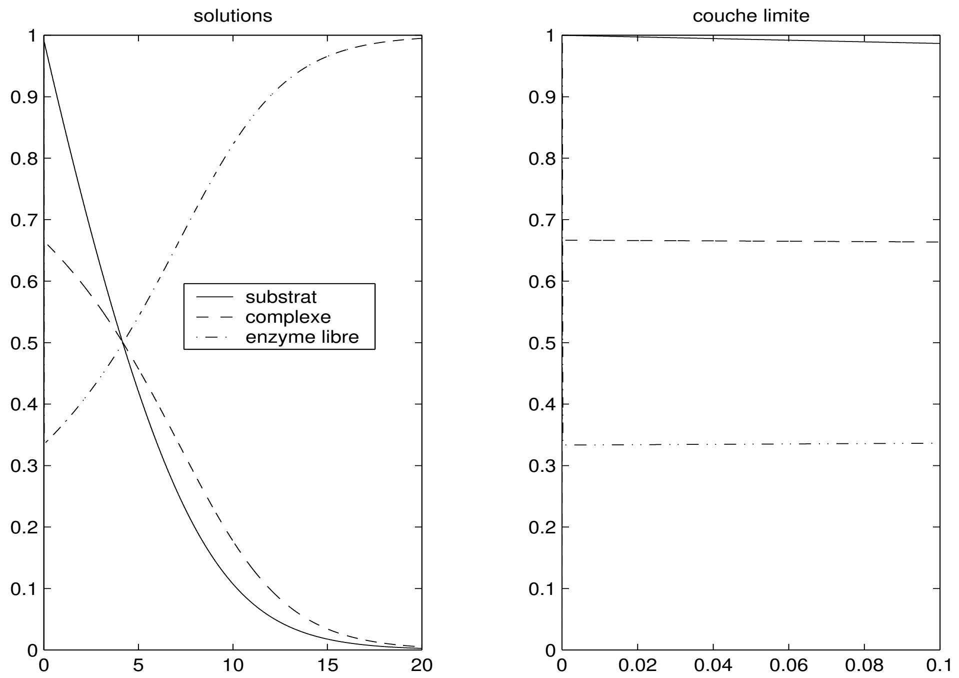


FIG. 14 – *simulation numérique du système extérieur: c'est une bonne approximation du système $(\mathcal{S})$ sauf à l'intérieur de la couche limite.*

le système $(\mathcal{S})$ se récrit alors:

$$\left\{\begin{array}{lll} \dfrac{dU}{d\sigma} & = & -\epsilon U + \epsilon(U + K - \lambda)V \quad , \quad U(0) = 1 \\ \dfrac{dV}{d\sigma} & = & U - (U + K)V \qquad\qquad\quad , \quad V(0) = 0 \end{array}\right.$$

Nous remarquons que le changement de variable a eu pour effet d'ôter la singularité de la deuxième équation. En utilisant un développement asymptotique de U et de V, nous obtenons alors au premier ordre des équations différentielles et non algébriques:

$$(\mathcal{I}) \left\{\begin{array}{lll} \dfrac{dU_0}{d\sigma} & = & 0 \qquad\qquad\qquad\qquad , \quad U_0(0) = 1 \\ \dfrac{dV_0}{d\sigma} & = & U_0 - (U_0 + K)V_0 \quad , \quad V_0(0) = 0 \end{array}\right.$$

Ce système est nommé **approximation asymptotique intérieure** ou **singulière** du système $(\mathcal{S})$. Une résolution facile nous conduit aux formules:

$$U_0(\sigma) = 1 \quad et \quad V_0(\sigma) = \frac{1 - e^{-(1+K)\sigma}}{1 + K}$$

Cette fois, le système $(\mathcal{I})$ nous donne une bonne approximation du système initial, mais uniquement dans la couche limite. Nous disposons donc de deux approximations du système $(\mathcal{S})$, chacune étant valide sur une partie différente du domaine $\mathbb{R}_+$. Il va nous falloir à présent raccorder les solutions de ces deux systèmes. Pour cela, il nous faut définir une notion de raccordement.

4.2.3 le raccordement

Définition : *Deux développements asymptotiques, l'un intérieur et l'autre extérieur se raccordent à l'ordre k s'il existe un segment $[\eta_1, \eta_2]$ d'intérieur non vide sur lequel les deux développements coïncident avec une erreur en $\mathcal{O}(\epsilon^{k+1})$.*

Nous allons chercher un raccordement à l'ordre 0 (c'est-à-dire en $\mathcal{O}(\epsilon)$. Pour cela, nous disposons de trois degrés de liberté : le choix de A et des η_i. Le raccord nécessite la condition :

$$\lim_{\sigma\to+\infty}[U_0(\sigma), V_0(\sigma)] = \lim_{\tau\to 0}[u_0(\tau), v_0(\tau)]$$

d'où :

$$\begin{cases} \lim\limits_{\tau\to 0} u_0(\tau) &= 1 \\ \lim\limits_{\tau\to 0} v_0(\tau) &= \dfrac{1}{1+K} \end{cases}$$

ce qui entraîne :

$$A = 1$$

Nous obtenons finalement une expression pour les solutions du système $\mathcal{S}$:

$$\begin{cases} u(\tau,\epsilon) &= u_0(\tau) + \mathcal{O}(\epsilon) \\ v(\tau,\epsilon) &= \dfrac{1-e^{\frac{1+K}{\epsilon}\tau}}{1+K} + \mathcal{O}(\epsilon) \qquad \textit{pour} \quad 0 \le \tau \ll 1 \\ &= \dfrac{u_0(\tau)}{u_0(\tau)+K} + \mathcal{O}(\epsilon) \qquad \textit{pour} \quad 0 < \epsilon \ll \tau \end{cases}$$

où $u_0(\tau)$ vérifie :

$$u_0(\tau) + K\ln(u_0(\tau)) = 1 - \lambda\tau$$

4.3 conséquences pour le modèle

Nous ne cherchons pas dans cette partie à étudier complètement l'équation de Michaelis-Menten. Les développements asymptotiques et l'existence d'un raccordement sont donc admis, ils résultent de la théorie des perturbations singulières (voir [7]). Nous déduirons simplement de cette étude que les solutions de l'équation de catalyse enzymatique peuvent être considérées, dans la mesure où la valeur ϵ reste négligeable, comme en quasi-équilibre. La composante singulière n'apparaît alors pas et la solution extérieure $(u_o(\tau), v_0(\tau))$ est la solution pertinente. La seule condition initiale est $u_0(0) = 1$ et la concentration $v(\tau)$ est dans un état stationnaire. C'est ce qu'on appelle *l'hypothèse de quasi-stationnarité de Michaelis et Menten.*

Sous cette hypothèse, on peut considérer que les variables adimensionnelles u et v suivent le système :

$$\begin{aligned} \frac{du}{d\tau} &= -u + (u + K - \lambda)\frac{u}{u+K} \quad &, \quad u(0) = 1 \\ &= -\lambda \frac{u}{u+K} \quad &, \quad u(0) = 1 \end{aligned}$$

Après redimensionalisation, on trouve que la vitesse de disparition du substrat suit la loi :

$$\boxed{\frac{ds}{dt} = -\frac{Vs}{s + K_m}}$$

avec :

$$\begin{aligned} V &= k_2 e_0 && : \text{vitesse maximale de la réaction} \\ K_m &= K s_0 = \frac{k_{-1} + k_2}{k_1} && : \text{constante de Michaelis} \end{aligned}$$

La figure 15 montre l'allure de la courbe représentant la vitesse de réaction en fonction du substrat s obtenue pour $V = 1$ et $K_m = 0.5$.

Cette partie achève l'étude des réseaux génétiques à l'aide de la modélisation continue. Comme nous l'avons déjà évoqué plus haut, il existe également une approche booléenne de ces réseaux. Par leurs natures fondamentalement différentes, ces deux approches présentent des aspects très différents du même problème. Nous allons voir dans la partie qui suit en quoi la modélisation booléenne permet de compléter notre étude sur les réseaux de régulation génétique.

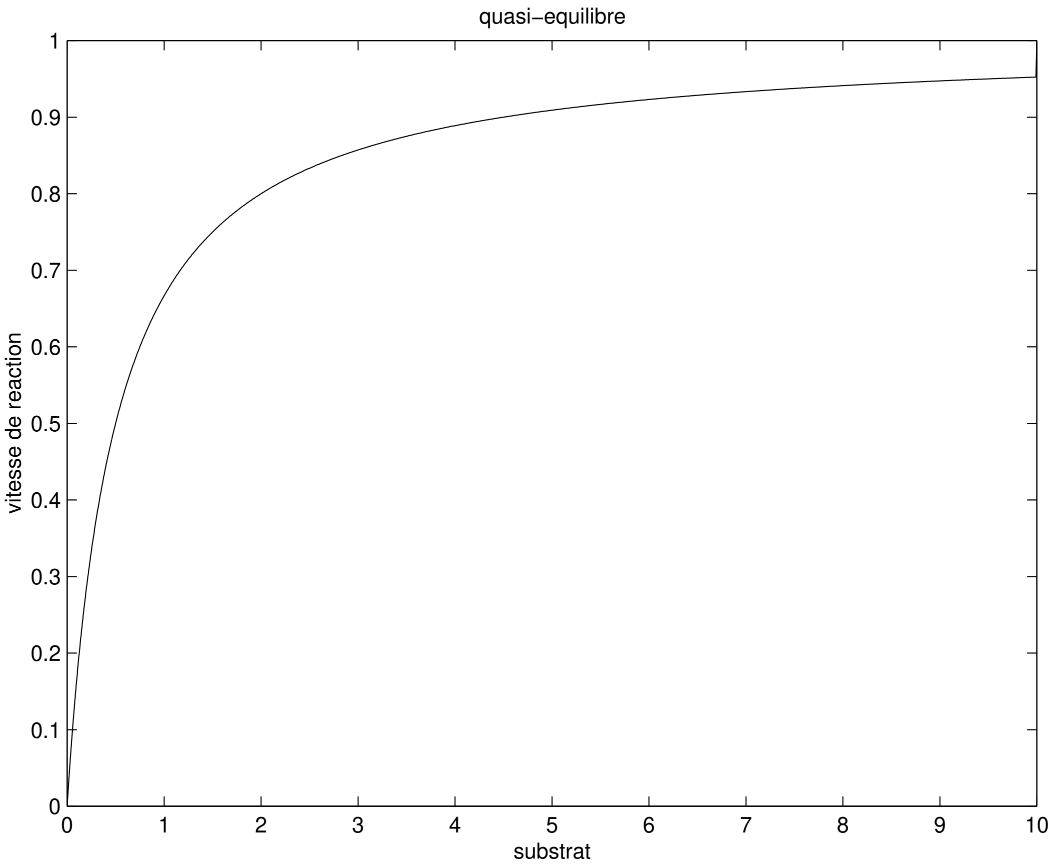


FIG. 15 – *L'hypothèse de quasi-équilibre (sur cette figure on voit la vitesse de la réaction en fonction de la concentration de substrat).*

5 Le modèle booléen

Dans la modélisation booléenne, nous oublions les concentrations et les réactions chimiques. Une espèce est *présente* ou *absente*, de la même manière un gène est *allumé* ou *éteint*. Dans ce modèle il n'y a plus d'équations différentielles, elles sont remplacées par des fonctions de transition booléennes. Il semble alors évident que nous allons perdre de la précision quant à la simulation d'un réseau. Néanmoins, il nous semble fondamental de décrire ce modèle car il présente deux avantages. Le premier est historique : les premières modélisations complètes de réseaux génétiques sont booléennes et sont dues à Stuart Kauffman dans les années 70 ([8], [9], [10]). Aujourd'hui encore, seuls ces modèles permettent de simuler des réseaux de très grandes tailles par l'intermédiaire de **réseaux booléens aléatoires**. Le deuxième avantage de l'approche booléenne vient du fait qu'aujourd'hui, l'étude des réseaux génétiques en est encore à ses premiers pas et la facilité de manipulation du modèle booléen est un sérieux atout, notamment pour identifier les gènes et les réseaux eux-mêmes.

En effet, si le modèle continu donne des résultats tout-à-fait satisfaisants, il n'en demeure pas moins que la découverte de nouveaux réseaux nécessite de nombreuses expériences biologiques qui nous permettent de trouver de nouvelles interactions entre les différents composants des réseaux. A l'heure d'aujourd'hui, le séquençage du génôme humain est établi, mais si nous savons que le nombre de gènes de l'être humain ne s'élève quà environ 30000, nous sommes encore loin d'avoir compris les centaines de milliers d'interactions entre tous ces gènes. Grâce aux nouveaux outils que les biologistes ont mis au point, notamment les récentes **biopuces**, de nouvelles perspectives s'ouvrent à la Recherche ([4]). Il est notamment capital aujourd'hui de trouver des algorithmes rapides et efficaces pour exploiter les données trouvées par l'expérience.

Afin de mieux comprendre ces différents arguments, nous avons décidé de consacrer la dernière partie de ce rapport à l'étude du modèle booléen et de ses propriétés.

5.1 Description du modèle

5.1.1 notion de réseau booléen

Un réseau génétique booléen se modélise comme un réseau booléen. C'est un **graphe orienté** $G = (V, F)$ où $V = \{v_1, v_2, \ldots, v_n\}$ est un ensemble fini de **nœuds** et où F est un ensemble de **fonctions booléennes** qui définit une topologie d'arêtes. n est appelé *taille* ou *dimension* du réseau (voir [12]).

Un nœud représente soit un gène, soit un **stimulus biologique**, c'est-à-dire un facteur physique ou chimique qui fait partie du réseau mais qui n'est ni un gène, ni le produit d'un gène. Ainsi dans le modèle du phage, l'irradiation par des rayons ultra-violets constitue un stimulus biologique qui influe sur le réseau sans être pour autant lié à l'expression d'un gène. A chaque nœud v, on associe une valeur booléenne $x(v)$ qui représente le **niveau d'expression** du stimulus correspondant. Ainsi pour un gène, la valeur de x sera 0 si le gène est *éteint* (ou non transcrit) et 1 si le gène est *allumé* (ou transcrit). De même, pour une espèce chimique, la valeur sera 1 si l'espèce est considérée comme *présente* (c'est-à-dire si elle est au-dessus d'un certain seuil) et elle sera 0 si l'espèce est *absente* (au-dessous du seuil).

L'ensemble F est un ensemble de fonctions booléennes, appelées règles, chacune d'entre elles étant associée à un nœud : $F = \{f_v |\ v \in V\}$, avec :

$$\forall v \in V, \quad f_v : \begin{array}{ccc} \{0,1\}^k & \longrightarrow & \{0,1\} \\ (x_1, \ldots, x_k) & \longmapsto & f_v(x) \end{array}$$

k s'appelle la **connectivité** de la fonction f_v.

Le fonctionnement d'un tel réseau se fait bien entendu en temps discret. A chaque top d'horloge, L'ensemble F définit pour chaque nœud la valeur de la variable $x(v)$:

$$\forall v \in V, \quad x(v)^{(t+1)} = f_v(x(v_{i_1})^{(t)}, \ldots, x(v_{i_k})^{(t)})$$

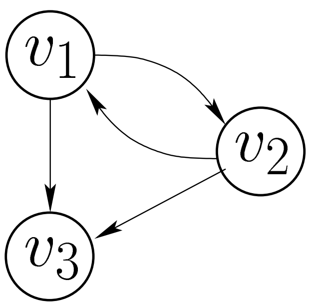
On définit l'**état** du réseau au temps discret t par le vecteur $(x_1^{(t)}, \ldots, x_n^{(t)})$ des niveaux expressions au temps t. (on notera par commodité $x_i^{(t)}$ au lieu de $x(v_i)^{(t)}$). Une **itération** du réseau se formalise donc par l'application de la fonction ψ :

$$(x_1^{(t+1)}, \ldots, x_n^{(t+1)}) = \psi(x_1^{(t)}, \ldots, x_n^{(t)})$$

avec :

$$\forall 1 \leq i \leq n \ , \quad \psi_i(x_1, \ldots, x_n) = f_{v_i}(x_{i_1}, \ldots, x_{i_k})$$

5.1.2 représentation graphique



v_2	v_1
0	1
1	0

v_1	v_2
0	0
1	1

v_1	v_2	v_3
0	0	0
0	1	1
1	0	0
1	1	0

FIG. 16 – *représentation graphique d'un réseau booléen*

Pour représenter un réseau booléen, on peut utiliser un graphe orienté pour voir la topologie du réseau (c'est-à-dire quel nœud influe sur quel autre nœud) ainsi que des tables de vérité pour expliciter les règles. La figure 16 présente un exemple de réseau, avec $n = 3$.

Nous pouvons également représenter un réseau par son **graphe d'état**. L'état du système est donné par les valeurs $x(v)$ pour chaque v à un instant donné. Nous voyons grâce à cette représentation les **trajectoires** que le réseau emprunte (voir figure 17). Le nombre d'états étant fini, il existe au moins un **attracteur** qui est soit un **cycle**, soit un **point fixe**. Ainsi dans l'exemple proposé figure 16, le réseau admet un cycle de longueur 4 (fig 17).

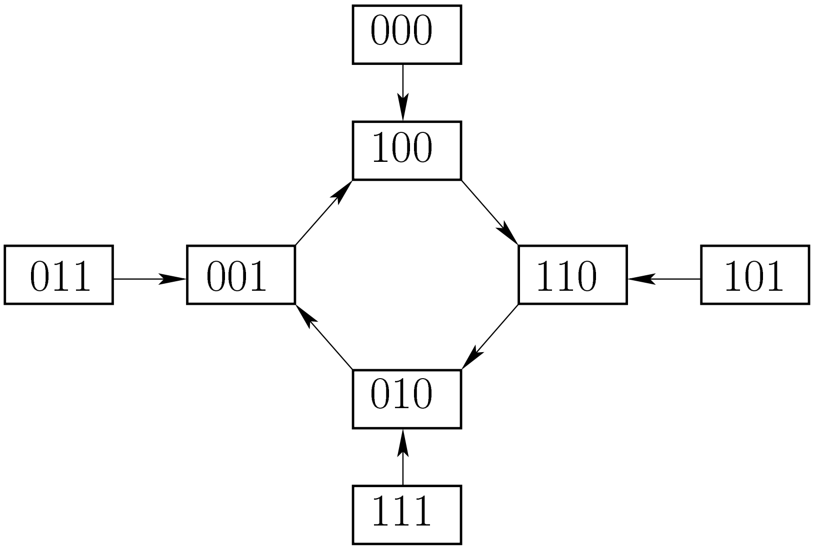


FIG. 17 – *représentation graphique d'un réseau booléen : graphe d'état*

Nous rappelons ici la définition d'un **bassin d'attraction** : le bassin associé à un attracteur est l'ensemble de toutes les trajectoires qui mènent à cet attracteur. Plus formellement, soit $\mathcal{A}$ un attracteur, le bassin d'attraction associé à $\mathcal{A}$ est l'ensemble $\mathcal{B}(\mathcal{A})$ défini par :

$$\mathcal{B}(\mathcal{A}) = \left\{ x \in \{0,1\}^n \mid \exists i \ , \ \psi^{(i)}(x) \in \mathcal{A} \right\}$$

où $\psi^{(i)}$ est la composée i-ème de ψ.

5.2 Applications du modèle : identification des réseaux

5.2.1 le probème de l'identification

Le problème de l'identification des réseaux est simple. Nous cherchons à retrouver les interactions entre les différents gènes. Afin d'être le plus efficace et le plus complet possible, il nous faut prendre en compte un nombre important de gènes. Pour cela, il nous faut des données biologiques de taille importante. Ceci est rendu possible grâce à la technique des **biopuces** (*ADN microarrays*). Cette technique nous permet notamment de repérer l'activité de plusieurs milliers de gènes à un instant donné. Les algorithmes d'identification cherchent ensuite à isoler les paquets de gènes qui marchent simultanément, puis, en raffinant peu à peu les mesures, à identifier le réseau booléen sous-jacent par interpolation (voir [11]).

5.2.2 les biopuces

Afin de mieux comprendre le type des données qui sont produites, il nous faut parler brièvement des biopuces. Ce sont de grandes plaques sur lesquels sont disposés plusieurs compartiments. Dans chacun de ces compartiments est déposée une certaine quantité de molécules d'ARN messager dont la séquence est la complémentaire de celle de l'ARN-m du gène étudié. On fait évoluer alors une culture en relevant précisément les différents facteurs extérieurs qui entrent en jeu (température, irradiation, *etc*...). Puis, à un moment précis de l'évolution de la culture, on effectue des prélèvements d'ARN que l'on dispose sur la plaque. Après un protocole expérimental complexe, on arrive donc à déterminer si, à l'instant du prélèvement, un gène donné était allumé ou éteint (grâce à une technique d'hybridation puis de coloration, les compartiments se colorent d'une teine spécifique relative au taux d'activité du gène). Chaque plaque nous donne ainsi un *instantané* et nous procure un état du réseau booléen à un moment donné (voir [4]).

Les données dont nous disposons sont donc d'immenses matrices de 0 et de 1 représentant l'état du réseau à divers moments. Il s'agit d'une série temporelle d'états successifs. On considèrera donc dans l'algorithme qui suit que l'on dispose d'un certain nombre de paires (I_j, O_j) d'entrées-sorties où $I_j, O_j \in \{0,1\}^n$, et l'on cherchera à identifier le réseau, c'est-à-dire sa topologie (qui influe sur qui ?) ainsi que les règles associées f_{v_i}.

5.2.3 l'algorithme d'identification

Pour écrire le programme d'identification, nous nous sommes inspirés de l'article [11] qui présente le principe général de l'algorithme REVEAL (REverse Engineering ALgorithm). Cet algorithme, qui prend en entrée un ensemble de séries temporelles, recherche la **connectivité** entre les différents éléments du réseau, puis les fonctions booléennes qui déterminent l'évolution de ce dernier.

Pour cela, l'algorithme utilise la théorie de l'information, notamment **l'entropie de Shannon**. De cette manière, on considère donc l'organisme comme un système qui nous fournit de l'information, cette information prenant la forme de séries de 0 et de 1. L'entropie permet de quantifier l'information portée par un flot d'informations en calculant les probabilités d'occurences des évènements élémentaires. Ainsi, si X est un flot, on calcule l'entropie de X, $H(X)$ par la formule :

$$H(X) = -\sum p_X \log p_X$$

Dans notre cas, X est un vecteur de 0 et de 1 et $H(X)$ s'écrit :

$$H(X) = -p_0 \log(p_0) - p_1 \log(p_1)$$

où p_0 est la proportion de 0 dans X et p_1 la proportion de 1. De même si X et Y sont deux flots, on définit l'entropie du couple (X, Y) par :

$$H(X,Y) = -p_{0,0} \log(p_{0,0}) - p_{0,1} \log(p_{0,1}) - p_{1,0} \log(p_{1,0}) - p_{1,1} \log(p_{1,1})$$

Grâce à la notion d'entropie, nous pouvons définir **l'information mutuelle** :

Définition : *Soient deux vecteurs booléens X et Y. On définit l'information mutuelle portée par le couple (X, Y) par la formule :*

$$M(X,Y) = H(X) + H(Y) - H(X,Y)$$

Cette quantité quantifie l'information que X et Y portent en commun.

Nous pouvons à présent décrire le principe de l'algorithme d'identification. Dans toute la suite, les lettres A, B, ... dénoteront des vecteurs booléens qui représentent l'état des nœuds a, b, ... **en entrée**. Les lettres primées A', B', ... dénoteront les vecteurs qui représentent l'état des mêmes nœuds **en sortie**.

Il s'agit dans un premier temps de déterminer les connexions du réseau, puis, dans un deuxième temps, de déterminer les règles.

1. <u>première étape</u> : identification des liens de connectivité $k = 1$

   Nous cherchons dans cette étape à déterminer si le réseau comporte des nœuds qui n'ont qu'une seule entrée. Pour cela, nous cherchons s'il existe un nœud X tel que :

   $$M(A', X) = H(A')$$

   Ceci signifie alors que X détermine exactement A'. Il existe donc une arête qui part de x et qui va vers a dans le graphe du réseau.

2. <u>deuxième étape</u> : détermination des règles associées

   Afin de déterminer la règle, il suffit de regarder le tableau des entrées-sorties et d'enlever les paires d'entrées-sorties redondantes. On peut désigner une règle par le mot binaire des sorties (en rangeant bien sûr les entrees par ordre croissant) ou encore par le décimal correspondant (c'est ainsi que nous avons procédé dans notre programme).

3. troisième étape : identification des liens de connectivité $k = 2$

   Si toutes les transitions ne peuvent être expliquées en termes de connectivité 1, nous recherchons les couples de nœuds (X, Y) tels que :

   $$M(A', [X, Y]) = H(A')$$

   Ceci signifie que le couple (X, Y) détermine entièrement A'. Nous pouvons donc tracer les deux arêtes $x \to a$ et $y \to a$ dans le graphe.

4. quatrième étape : détermination des règles associées

   Nous procédons comme dans l'étape 2.

Nous continuons ainsi en incrémentant la valeur de k tant que nous n'avons pas trouvé les entrées de tous les nœuds. L'algorithme s'arrête de toute façon lorsque $k = n$. Ecrit de cette manière, nous remarquons que l'algorithme est censé trouver tous les réseaux booléens possibles en accord avec les paires d'entrées-sorties proposées.

Afin d'implémenter cet algorithme, nous avons utilisé le logiciel MATLAB. Le programme prend en entrée deux fichiers : entrees et sorties et retourne un fichier resultat dans lequel est écrit le réseau.

Afin d'aller plus vite, il est également possible de fixer une connectivité maximale K au-delà de laquelle on arrête la recherche. Le réseau trouvé peut être alors incomplet. Nous avons également une version qui génère elle-même les fichiers entrees et sorties aléatoirement. Selon le nombre de paires d'entrées-sorties, le réseau peut également être incomplet.

En annexe se trouve un exemple d'application du programme pour la recherche d'un réseau simple ($n = 4$).

5.2.4 analyse

Nous reprenons brièvement dans cette partie les résultats de [12] sur les performances de l'algorithme REVEAL. Ce qui me semble important est de donner une borne supérieure d'entrées-sorties nécessaires pour identifier un réseau. En effet, cet algorithme étant particulièrement gourmand en données, et ces données étant particulièrement chères, il devient capital de minimiser la taille des données nécessaires.

Afin de trouver cette borne, on se rend vite compte que la connectivité est importante. Lorsque la taille du réseau n augmente, la complexité de l'algorithme explose. Il nous faut donc imposer une connectivité maximum K. Cette borne est donnée par le théorème suivant :

Théorème : *si l'on donne en entrée* $\mathcal{O}\left(2^{2K}(2K + \alpha)\log n\right)\}$ *paires d'entrées-sorties réparties uniformément aléatoirement, le résultat suivant est vrai avec une probabilité supérieure à* $1 - \frac{1}{n^\alpha}$ *: il existe au moins un réseau booléen de dimension* n *avec une connectivité maximale* K *qui vérifie les entrées-sorties* *proposées.*

Ce théorème nous révèle donc que pour une connectivité bornée par une constante, le problème de l'identification demande un jeu de données dont la taille varie avec $\log n$. Ceci est rassurant quant aux expérimentations futures de cet algorithme. La preuve de ce théorème est dans akutstu...

Remarque : il faut noter que l'expérimentation de ce théorème nécessite de nombreux calculs, effectués à partir de réseaux booléens générés aléatoirement. En générant ces réseaux, on peut imposer la connectivité maximale. En revanche, si l'on génère des données aléatoires en vue de déterminer un réseau (ce qui était mon cas), il est très difficile d'observer un réseau avec une connectivité inférieure à n.

Conclusion et perspectives

La modélisation de systèmes biologiques complexes est devenue aujourd'hui un domaine de recherche très fertile. De nombreux chercheurs, mathématiciens, biologistes et même informaticiens tentent aujourd'hui de comprendre ces systèmes et de les modéliser formellement afin de procéder à des simulations. Les applications de cette théorie sont très importantes, notamment d'un point de vue biologique car elles ont pour but de prévoir l'évolution de variables biologiques souvent difficiles à observer par l'expérience.

Plus précisément, l'étude des réseaux génétiques vise à mieux comprendre l'impact du facteur génétique dans la vie de nos cellules. Cette étude est donc fondamentale et doit absolument suivre, à mon avis, les études récentes sur le séquençage du génôme humain.

Nous avons vu qu'il existe en fait deux voies dans ce nouveau champ de recherche. La première consiste à modéliser directement l'évolution continue des réseaux. Pour cela, l'apport de la théorie des systèmes hybrides semble tout-à-fait intéressant. Les premiers résultats concernant l'opéron lactose sont plutôt satisfaisants et il paraît raisonnable d'imaginer une continuité dans ces travaux. Il reste en effet à modéliser des réseaux plus complexes, et pour cela, générer de façon automatique les systèmes, en prenant en compte les constantes chimiques du problème et le réseau logique sous-jacent.

Une étape future sera d'augmenter le nombre de variables, en trouvant un moyen d'effectuer la résolution numérique malgré les dimensions très élevées. Pour cela, on pourra par exemple appliquer les techniques de réduction de modèles afin de projeter les solutions sur des espaces de dimension inférieure. On pourra également s'intéresser plus précisément au problème de l'identification des réseaux, qui est aujourd'hui crucial puisque c'est cette identification qui nous permet d'interpréter toutes les données issues des programmes de recherche concernant le génôme.

Afin d'améliorer la qualité d'approximation de notre modèle, on sera également amenés à étudier plus précisément les réactions de catalyse enzymatique, en considérant notamment les réactions à plusieurs enzymes. Celles-ci amènent de nouvelles difficultés comme le phénomène de coopérativité.

Les sujets de recherche dans le domaine des bio-mathématiques sont donc nombreux. Je trouve personnellement qu'il est très agréable de travailler dans ce domaine car c'est un domaine pluri-disciplinaire aux applications variées et très intéressantes. Je trouve en outre que la modélisation de systèmes biologiques est une discipline particulièrement enrichissante. Elle permet en effet de comprendre des théories mathématiques complexes et de les manipuler en vue de simuler des phénomènes concrets. Elle allie donc la théorie et ses applications, ce qui me semble très intéressant.

Annexes

Annexe 1 : définition des équations de l'opéron lactose

Les variables sont :

r	: répresseur
a	: activateur
x	: β-galactosidase
y	: perméase
g	: glucose intérieur
l	: lactose intérieur
c_1	: complexe lactose - perméase
c_2	: complexe lactose intérieur - galactosidase
l_m	: lactose

l_{ext} et g_{ext} sont des variables de contrôle (typiquement, $l_{ext}, g_{ext} \in \{0,1\}$) qui représentent les sources extérieures en sucres.

Les k_i représentent les constantes relatives au complexe lactose - perméase, alors que les k'_i représentent les constantes relatives au complexe lactose intérieur - galactosidase.

K est la constante de dégradation du glucose.

état 1 : $G\bar{L}$, répression

$$
\begin{aligned}
\dot{r} &= -\nu_r r + \gamma_r r \\
\dot{a} &= -\nu_a a \\
\dot{x} &= -\nu_x x \\
\dot{y} &= -\nu_y y - k_1 y l_m + (k_{-1} + k_2) c_1 \\
\dot{g} &= -Kg + g_{ext} \\
\dot{l} &= k_2 c_1 \\
\dot{c_1} &= k_1 y l_m - (k_{-1} + k_2) c_1 \\
\dot{c_2} &= 0
\end{aligned}
$$

$$
\begin{cases}
\dot{l_m} = -k_1 y l_m + k_{-1} c_1 + (1 - l_m) \text{ si } l_{ext} = 1 \\
\dot{l_m} = -k_1 y l_m + k_{-1} c_1) \text{ sinon}
\end{cases}
$$

Annexe 2 : identification d'un réseau booléen de taille 5

Après avoir mis en entrée de notre programme les fichiers entrees et sorties générés aléatoirement, nous trouvons le réseau booléen sous-jacent :

simulation, 21-Jun-2001
taille du reseau : 5
noeud 1:
* entrees: 1 2 3 4 5
* regle: 2.262784e+09
noeud 2:
* entrees: 1 2 3 4 5
* regle: 705771049
noeud 3:
* entrees: 1 2 3 4 5
* regle: 1286589284
noeud 4:
* entrees: 1 2 3 4 5
* regle: 929311702
noeud 5:
* entrees: 1 2 3 4 5
* regle: 1623101801

Les règles sont données en décimal.

Nous voyons sur cet exemple l'illustration de la remarque faite plus haut : il est extrêmement rare de trouver un réseau aléatoirement de taille supérieure à 4 qui ait des nœuds de connectivité inférieure à n. Ici, les cinq nœuds sont de connectivité $k = 5$.

Références

Articles cités dans le rapport :

[1] D. Robert, B. Vian, *Eléments de Biologie Cellulaire*, doin ed. (1994)

[2] site web : *http://esg-www.mit.edu:8001/esgbio/chapters.html*

[3] J.D. Murray, *Mathematical Biology*, Springer - Verlag, New York (1989) chap. 5

[4] J.P. Fitch, B Sokhansanj, *Genomic Engineering : Moving Beyond DNA Sequence to Function*, proceedins of the IEEE (dec 2000)

[5] M. Ptashne, *A Genetic Switch, second edition Phage λ and Higher Organisms*, Cell Press & Blackwell scientific publications (1992)

[6] H.H. McAdams, L. Shapiro, *Circuit Simulation of Genetic Networks*, Science 269:650-656 (1995)

[7] M.H. Holmes, *Introduction to Perturbation Methods*, Texts in applied mathematics 20, Springer - Verlag (1995)

[8] S. Kauffman, *The Origins of Order*, Oxford university press (1993)

[9] S. Kauffman, *Gene Regulation Networks: A theory for their global structure and behavior*, Current topics in Dev. Biol. 6: 145 (1971)

[10] S. Kauffman, *Antichaos et Adaptation*, Pour la science, hors-série Le chaos, 104-110 (1994)

[11] P. D'haeseleer, S. Liang, R. Somogyi, *Tutorial : Gene Expression Data Analysis and Modeling*, Session on Gene Expression and Genetic Networks, Pacific symposium on Biocomputing, Hawaii (1999)

[12] T.Akutsu, S. Miyano, S. Kuhara, *Identification of Genetic Networks from a small number of gene expression patterns under the boolean network model* Pacific symposium on Biocomputing, Hawaii (1999)

Autres lectures sur les réseaux génétiques :

[13] M.A. Gibson, E. Mjolness, *Modeling the activity of single genes*, à paraître dans : Computational Methods for Molecular and Cellular Biology, ed. J. M. Bower and H. Bolouri, MIT Press.

[14] H. de Jong, *Modeling and simulation of genetic regulatory systems : a literature review*, Rapport de recherche INRIA, 4032 (2000)