UNIVERSITÉ D'ÉVRY – VAL D'ESSONNE

De la bio-informatique textuelle à une approche formelle de la biologie des systèmes

mémoire présenté par

Jean-Paul COMET

en vue de l'obtention du diplôme d'

Habilitation à Diriger des Recherches

Spécialité : Informatique

– soutenue le 20 novembre 2006 devant le jury composé de –

M.	Gilles	BERNOT	Directeur de recherche
M.	Jacques	DEMONGEOT	Rapporteur
M.	Alain	DENISE	Rapporteur
M.	Hidde	DE JONG	Rapporteur
M.	Didier	ARQUÈS	Examinateur
M.	Erol	GELENBE	Examinateur
M.	Michel	THELLIER	Examinateur
M.	René	THOMAS	Examinateur

Table des matières

Introduction

Modélisation

Modélisation : ce mot a aujourd'hui tant d'acceptions que presque tout scientifique affirme que son métier est de modéliser, de construire des modèles de son objet d'étude. On peut évidement essayer de faire une typographie des acceptions de ce mot, ce qui mènerait encore à un modèle des différents sens de ce terme.

Initialement, le modèle désigne la référence à imiter ou à reproduire, comme par exemple le modèle que le photographe immortalise ou encore l'organisme-modèle que le biologiste étudie. Puis ce mot a progressivement désigné le résultat de cette imitation. Le modèle est ainsi devenu la représentation d'un objet. La recherche théorique s'est progressivement appropriée ce concept de modèle, elle l'utilise couramment aujourd'hui pour désigner les « représentations » des phénomènes qu'elle cherche à comprendre ou à expliquer. Le modèle peut alors être compris comme un « système de symboles », qui peuvent être textuels, mathématiques, logiques, schématiques, graphiques... Le modèle doit cependant représenter l'ensemble des caractéristiques de l'objet étudié. Ainsi le modèle est dit cohérent lorsqu'il rend compte de la plupart des informations dont on dispose (ou dont on veut tenir compte) lorsqu'on souhaite décrire un phénomène. Comme ces informations concernent non seulement les objets représentés par les symboles mais aussi les conséquences de leurs cohabitations dans le système, les notations mathématiques, formelles, les systèmes de notation musicale, chorégraphique, chimique... se sont dotés de règles d'articulation ou grammaire qui permettent de composer ces représentations. Le modèle doit aussi être interprétable facilement, la sémantique de chacun des symboles et des règles de manipulation syntaxique étant partagée par le plus grand nombre.

Ces règles grammaticales doivent mener à des raisonnements sur le modèle dont le résultat est cohérent avec l'ensemble des observations sur l'objet d'étude. La science informatique s'est naturellement focalisée sur le lien qui doit exister entre modèles, règles d'articulation et calcul sur ces règles. Ces trois éléments forment un tout qui doit être intelligible, reproductible et communicable. Pour cela, il est nécessaire d'expliciter ce que veut montrer le modélisateur qui élabore et interprète le modèle considéré : un même système pourra être modélisé de différentes manières. La modélisation est ainsi devenue l'opération par laquelle on établit un modèle d'un phénomène, afin d'en proposer une représentation interprétable, reproductible et simulable, ou mieux encore, sur laquelle un raisonnement peut amener à de nouvelles connaissances.

Dans le cadre de la biologie, parallèlement à l'accumulation de données génomiques, transcriptomiques et protéomiques... les modèles biologiques cherchent à identifier, au delà des corrélations, les liens de causalité susceptibles d'exister entre des phénomènes en apparence disjoints. La complexité des systèmes biologiques a amené les biologistes à développer une expertise pointue sur les raisonnements à propos de ces systèmes. La formalisation ne fait que prolonger le raisonnement en langage naturel ; elle s'avère indispensable lorsque la logique sous-jacente devient trop complexe pour pouvoir être perçue sans ambiguïté par de simples mots.

La bio-informatique

La bio-informatique est née de la nécessité d'utiliser des méthodes provenant de l'informatique pour le traitement des informations dans le cadre de la biologie. Elle constitue un large domaine de recherche pluridisciplinaire où la biologie, les mathématiques, l'informatique, la physique théorique apportent des éléments cruciaux pour la recherche. Elle est en pleine mutation depuis sa naissance suivant ainsi les évolutions des méthodes de la biologie. L'origine de la bio-informatique est certainement liée à l'analyse de la séquence afin de comparer les séquences entre elles, puis d'annoter les génomes, prédire les gènes, leurs structures et leurs régions régulatrices, permettant ainsi la construction de petits réseaux génétiques.
L'apparition des puces à ADN a permis une première évolution des problématiques de la bio-informatique. Par la quantification possible de milliers de transcrits, la biologie fonctionnelle change d'échelle. Il faut donc créer les outils pour analyser ces données et comprendre le fonctionnement de l'organisme. Il s'agit là d'une première tentative de compréhension de la dynamique des réseaux de régulation. Apparaît aussi l'espoir de pouvoir reconstruire une partie de ces réseaux au travers des dynamiques de chacun des gènes.
La biologie des systèmes (*systems biology*), comme l'ont bien décrit O. Wolkenhauer [118] et T. Ideker [52], cherche à appréhender un système biologique non pas comme la somme de ses éléments constitutifs mais comme un système à part entière. Elle cherche à rendre compte de l'émergence de certains phénomènes (un système est plus que la somme de ses parties), de la sensibilité à certains paramètres, bref, de la complexité des systèmes biologiques. Pour élucider les principes qui gouvernent la complexité biologique, S. Huang [50] propose de dépasser les explications *ad-hoc* pour faire émerger de nouveaux concepts abstraits. D'après H. Kitano [62], la biologie des systèmes essaye d'établir des méthodes et techniques qui nous permettent de comprendre les systèmes biologiques comme des systèmes à part entière et de comprendre leur complexité. Parallèlement, J. Hasty, D. McMillen et J.J. Collins [48] ont prouvé qu'il était possible d'en construire et de les manipuler, ce qui élargit le champs de la biologie des systèmes. Ainsi, comprendre la structure des réseaux génétiques, des réseaux métaboliques et des réseaux de transduction du signal, est une première étape pour comprendre leurs dynamiques, et pour faire émerger des méthodes pour contrôler, construire et modifier ces systèmes afin d'assurer qu'ils possèdent certaines propriétés [63]. La technique émergeante des puces à cellules permet d'étudier la réponse fonctionnelle d'une cellule à un stimulus, et permet de se focaliser sur l'initiation de la transcription. Elle devient à la fois un outil de validation des prédictions issues des modèles, et un outil de construction de ces mêmes modèles. Il s'agit là encore d'un changement d'échelle de la modélisation.

Parcours scientifique

L'ensemble des travaux présentés dans cette Habilitation à Diriger des Recherches se situe dans le cadre de la bio-informatique. D'un point de vue chronologique, mon activité scientifique a commencé dans le domaine de la comparaison de séquences : reformulation des algorithmes fondateurs de la bio-informatique (algorithmes d'alignement de séquences de Needleman & Wunsch et de Smith & Waterman), estimation de la significativité d'un alignement par une méthode de Monte-Carlo, prise en compte d'information autre que la séquence primaire dans un alignement. Puis avec l'avènement de la technologie des puces à ADN, je me suis intéressé à l'analyse statistique de données issues de ces puces. Enfin, la modélisation s'est avérée incontournable lorsque les biologistes nous ont posé des questions relatives à la dynamique des réseaux de régulation : modélisation des réseaux de régulation génétiques, automatisation partielle d'un processus d'aide à la modélisation, lien entre graphe d'interactions et dynamiques, variation sur le model-checking... Toutes ces activités forment un tout, la modélisation en bio-informatique.

- Du point de vue de la biologie, mes activités ont d'abord évolué de la bio-informatique textuelle (recherche de liens syntaxiques entre des objets de type chaînes de caractères) à l'analyse statistique de données de puces à ADN, technique permettant d'associer à des gènes des concentrations voire des dynamiques (un embryon de sémantique). Depuis elles se focalisent sur la biologie des systèmes.
- Du point de vue de la pluridisciplinarité, j'ai commencé par l'application et l'adaptation de techniques provenant de l'informatique et des mathématiques appliquées (programmation dynamique, algèbre max-plus, statistiques) dans le contexte de collaborations avec des biologistes. L'analyse du transcriptome m'a permis de collaborer avec une équipe de biologistes sur l'interprétation biologique de ces données. Enfin je me suis tourné vers un travail en équipe autour de la formalisation des réseaux de régulation génétique. Un tel travail est favorisé par les échanges entre plusieurs disciplines car chacun, en gardant ses compétences propres enrichit les compétences de l'autre.
- D'un point de vue de la complexité, je suis passé de l'étude de phénomènes statiques à l'étude de phénomènes dynamiques. La difficulté de cette étape réside non seulement dans le changement des outils et méthodes à développer mais aussi dans le changement de niveau de complexité des systèmes d'étude. La dynamique fait naturellement apparaître des phénomènes émergeants caractérisant les systèmes complexes.
- Au cours de cette évolution, les méthodes informatiques que j'ai utilisées ont changé. Après l'algorithmique, j'ai utilisé des méthodes de probabilités/statistiques, puis des méthodes discrètes et enfin des méthodes formelles.

Parallèlement à ces différents aspects de mon parcours, la modélisation en bio-informatique a toujours été présente, que ce soit pour la comparaison de séquences, l'analyse du transcriptome ou encore la dynamique des réseaux de régulation génétique.

Guide de lecture

Le mot « *modélisation* » est finalement le maître-mot de mon travail autour de la bio-informatique. Nous montrons dans ce document comment le terme modélisation peut être compris dans ses différentes acceptions dans les différents thèmes abordés que sont l'analyse de séquences, l'analyse du transcriptome ou encore la biologie des systèmes. Notons cependant comme nous l'avons déjà souligné, que le mot « *modélisation* » est utilisé par toutes les disciplines et sous-disciplines puisqu'il représente le cœur même de l'activité scientifique. Ainsi, d'autres acceptions n'ont pas du tout été abordées au cours de ce travail en particulier celles qui font référence à la représentation des connaissances, au stockage de ces dernières et à toute l'activité de recherche autour des bases de données.

Le premier chapitre montre comment la modélisation du comportement du Z-score, indice de significativité d'alignements de séquences protéiques, par une loi de Gumbel permet de donner un sens bien précis à l'homologie de séquences. Cette première étape d'estimation de la significativité d'un alignement est fondamentale pour pouvoir ensuite construire des familles de protéines, pour pouvoir associer à des protéines nouvellement séquencées des fonctions biologiques par analogie, pour reconstruire des arbres phylogénétiques... La modélisation de la loi du Z-score donne ainsi les bases pour un grand nombre de travaux en génomique.

Le deuxième chapitre se focalise sur l'analyse des données du transcriptome et propose une modélisation statistique des niveaux d'expression (modèle ANOVA) qui permet de tisser encore des liens, cette fois-ci entre observation du transcriptome et comportement macroscopique ou même phénotype. Pour cela nous nous sommes intéressés à la réponse d'un organisme à un stress car cette réponse peut nous mettre sur la voie pour comprendre la fonction d'un gène ou d'une

famille de gènes. D'un coté, la classification non supervisée a permis de construire des classes de gènes dont les profils d'expression sont similaires, elle met en exergue les gènes co-exprimés. D'un autre coté, la classification supervisée permet de se focaliser sur les différences qui peuvent exister entre différents échantillons. Dans cette perspective, nous nous sommes intéressés à la détermination de classifieurs en croisant deux méthodes, l'une provenant des statistiques (analyse de la variance) et l'autre mettant en œuvre un algorithme d'estimation d'attributs.
Les chapitres 3 et 4 portent tous deux sur la modélisation des réseaux de régulation génétique. Le premier d'entre eux donne les fondements sémantiques de tels réseaux, en s'appuyant sur la modélisation de R. Thomas, puis en reformulant cette théorie en termes de réseaux de Petri de haut niveau. Des extensions sont proposées pour affiner la compréhension de certains points. Cette modélisation donne un cadre formel pour le calcul de la dynamique associée à ces réseaux, elle est complétée par un chapitre présentant comment la formalisation de la dynamique et de ses paramètres permet de concevoir une méthodologie de conception de modèles. Cette méthodologie consiste à modéliser des propriétés temporelles du système à étudier sous forme de *spécifications*, puis à confronter de manière exhaustive l'ensemble des modèles envisageables pour le système d'étude à ces spécifications. Elle peut éventuellement être complétée par un retour à l'expérimentation pour affiner l'ensemble des modèles jusque là cohérents. Dans ces chapitres, la modélisation a permis non seulement de comprendre très finement la dynamique de réseaux de régulation, mais a aussi permis la mise en place d'une méthodologie pour automatiser la construction de modèles.
Le chapitre 5 montre comment la modélisation des réseaux de régulation a pu faire émerger des relations entre la représentation statique d'un réseau et sa dynamique associée. Cette démarche, *modéliser pour raisonner*, est illustrée en donnant les éléments de preuve des deux conjectures énoncées dès 1981 par R. Thomas [107] : une boucle de rétroaction positive est nécessaire pour observer une multi-stationnarité et de manière presque symétrique, une boucle de rétroaction négative est nécessaire pour observer un comportement cyclique stationnaire. Pour traiter ces conjectures, il a fallu envisager le graphe d'interactions comme déduit de la dynamique et non pas la dynamique comme déduite du graphe d'interactions. Le modèle proposé expliquant comment le graphe d'interactions est construit à partir de la dynamique a permis de raisonner sur un ensemble de dynamiques beaucoup plus grand que celui atteignable par des modèles de R. Thomas. En particulier, on peut considérer des dynamiques de $\mathbb{Z}^n$ dans $\mathbb{Z}^n$. Ces deux résultats n'ont en fait pas le même statut :

- dans le cas de la multi-stationnarité, l'existence de la boucle positive est démontrée pour au moins un état dans l'espace des phases,
- alors que pour le comportement cyclique, l'existence de la boucle négative est pour l'instant prouvée pour une superposition de graphes d'interactions différents dans l'espace des phases, ce qui en fait un résultat un peu plus faible.

Les prolongements de ces travaux autour de la modélisation des réseaux de régulation génétique sont nombreux. Deux raffinements de la modélisation de R. Thomas sont facilement envisageables. Le premier permettrait de découper chacun des domaines de l'espace des phases en fonction des domaines atteints par les trajectoires des solutions du système d'équations différentielles sous-jacent, l'autre basé sur une approximation des trajectoires par des lignes brisées pourrait mener à une vérification de certaines propriétés temporelles grâce au model-checker `HyTech`. Même si on reste dans le cadre d'une modélisation discrète, les paramètres gérant la dynamique ne sont pas toujours connus. La vérification exhaustive de chacun des modèles n'est pas toujours envisageable à cause de l'explosion combinatoire. Pour faire face à ce défit, plusieurs pistes ont été envisagées.

- La prise en compte d'information de coopération ou de concurrence dès le début du processus de modélisation, c'est à dire dans le schéma des interactions, est une première voie pour

éliminer plus rapidement des modèles qui ne correspondent pas aux interprétations biologiques. Nous proposons de représenter ces connaissances dans les transitions d'un graphe biparti dont les nœuds représenteraient les entités biologiques d'intérêt et les transitions les processus régulatoires.
- La deuxième possibilité se focalise sur la phase de vérification de l'ensemble des modèles à considérer. Tous ces modèles partagent certaines informations, comme par exemple, les états. Il serait intéressant de développer une méthode de model-checking qui travaille sur des structures de Kripke symboliques qui abstraient un ensemble de structures de Kripke classiques.

Tous ces travaux ont pour objectif d'aider à la découverte en biologie, soit en définissant des mesures plus précises pour l'analyse des liens qui peuvent exister entre séquences, soit pour comprendre les liens fonctionnels entre gènes, soit enfin pour comprendre la dynamique d'un réseau de régulation dont toutes les caractéristiques ne sont pas connues. Ce dernier domaine de recherche a pour objectif d'aider le biologiste dans ses différentes phases de modélisation du monde vivant en utilisant des méthodes bien connues de l'informatique pour la *vérification* et la *validation* d'un modèle.

Chapitre 1

Séquences

Résumé

La bio-informatique est née de la nécessité d'extraire de manière automatique les ressemblances entre sous-séquences biologiques nucléiques ou protéiques. La programmation dynamique permet le calcul de l'alignement optimal entre deux séquences en temps quadratique, mais la question de la significativité de cet alignement reste une question ouverte.

Plusieurs approches d'estimation de cette significativité ont été proposées, comme par exemple le Z-score, variable aléatoire construite sur une approche de Monte-Carlo. Notre travail de modélisation nous a amené à montrer que la loi du Z-score pouvait être approximée par une loi de Gumbel. Ce résultat a permis de proposer le Z-score comme indice de similarité robuste. Depuis, ces travaux ont été continués et ont donné lieu à des résultats exploitables dans des cas où les autres approches d'estimation de la significativité ne l'étaient pas.

1.1 Modéliser pour annoter par homologie

En 2006, les banques de données électroniques publiques de séquences biologiques contiennent plus de 3 millions de séquences protéiques, obtenues par traduction automatique de séquences génomiques. On pourra se référer aux statistiques mises à disposition sur les sites internet de Genbank[1] et de SwissProt[2]. Plus de 300 génomes sont maintenant disponibles, et le séquençage de centaines d'autres sont en cours. Cette quantité d'information est encore en pleine croissance et la courbe représentant le nombre d'entrées en fonction du temps semble être une fonction exponentielle (en 2006, plus de 100 milliards de bases d'ADN sont séquencées) et ce processus s'accélèrera encore prochainement grâce à l'avancée technologique rapportée par M. Margulies *et al.* [72] permettant un débit 100 fois supérieur par rapport aux techniques de séquençage classiques.

Vu ces débits de production de données, il ne faut pas s'attendre à ce qu'elles soient confirmées expérimentalement ni même que la fonction biologique associée à ces séquences soit analysée directement. Notre faculté à prédire des propriétés biologiques pertinentes pour la biologie ne repose donc que sur des procédures d'analyse *in silico* dont la fiabilité devient un enjeu important.

L'information biologique contenue dans les séquences protéiques est issue des séquences nucléiques. De par sa biosynthèse, une protéine est d'abord une séquence d'acides aminés qui se replie spontanément, éventuellement avec l'aide d'une protéine chaperonne. L'analyse des séquences repose donc sur un postulat exprimé dès 1974 par T.T. Wu, W.M. Fitch et E. Margoliash [119] : l'enchaînement des acides aminés contient l'information nécessaire et suffisante pour permettre la structuration tridimensionnelle très complexe qui mène aux fonctions moléculaires.

[1] http ://www.ncbi.nlm.nih.gov/Genbank/
[2] http ://www.ebi.ac.uk/swissprot/sptr_stats/index.html

Il est admis que le processus de diversification a lieu au niveau de l'information génétique (théorie synthétique de l'évolution des systèmes vivants) et que les événements sont de deux types : mutation et recombinaison, voir par exemple les travaux de E. Mayr [75, 76]. Les séquences de différents individus ou différentes espèces, qui ont évolué à partir d'une séquence ancestrale commune se ressemblent structurellement. Le maintien de ces propriétés structurales, et donc la ressemblance des séquences, dépend non seulement des mutations/recombinaisons qui caractérisent les évolutions des différentes espèces mais aussi de la pression évolutive exercée sur ces séquences qui permet le maintien d'une fonctionnalité utile.

Lorsqu'on est face à une nouvelle séquence obtenue par séquençage automatique, on cherche à inférer la fonction de la protéine en question le plus souvent grâce à la recherche de *ressemblances* de ladite protéine avec des séquences déjà caractérisées des banques de données. On peut caractériser plusieurs types de ressemblances. Deux séquences seront dites *similaires* si elles sont proches par rapport à un critère de ressemblance structurale (ici si leurs séquences primaires sont proches), elles seront dites *homologues* si elles sont similaires et si elles possèdent une séquence ancestrale commune voire dans un sens plus strict en partageant aussi un certain degré de fonctionnalité, et *analogues* si elles partagent des propriétés biologiques communes sans pour autant partager une séquence ancestrale commune.

L'*annotation par homologie* a pour but de profiter de l'annotation fonctionnelle de certaines séquences pour annoter de nouvelles séquences sur des critères calculés sur la représentation desdites séquences. Autrement dit, la modélisation des séquences qui repose sur la représentation de celles-ci (mots sur un alphabet donné) se dote de règles de manipulation de ces séquences dont le résultat (mesure de ressemblance) est en accord avec le niveau de ressemblance biologique (fonctionnelle par exemple) des séquences biologiques.

L'annotation par homologie nécessite un critère de ressemblance entre séquences permettant éventuellement une mesure du temps de divergence entre les espèces. Les premières méthodes automatiques d'*alignement de séquences* ont été développées sur des critères algorithmiques. Aujourd'hui encore la comparaison de séquences reste une sous-discipline majeure de la bioinformatique dont les enjeux sont le temps de réponse lors de la comparaison d'une séquence à une banque de données, l'annotation globale de génomes nouvellement séquencés, la classification ou clustering de séquences, la phylogénie moléculaire...

La recherche dans le cadre de la comparaison de séquences deux à deux traite de deux problèmes. Le premier concerne le calcul de l'alignement, il s'agit de problèmes essentiellement algorithmiques. Étant données deux séquences et une mesure de ressemblance, quel est l'alignement qui optimise cette ressemblance ? C'est dans ce cadre que j'ai proposé une reformulation de l'algorithme d'alignement de séquences par programmation dynamique dans l'algèbre max-plus [23]. Cette reformulation a permis de montrer que pour une séquence cible fixée A, l'ensemble des scores d'alignement atteignables en comparant A avec n'importe quelle séquence est projectivement fini pour le score d'alignement global, et fini pour le score d'alignement local. Ce travail a aussi mené à un algorithme linéaire qui permet de comparer une séquence cible à une banque de données de séquences. Malheureusement, la taille de l'automate sous-jacent empêche la mise en œuvre de cet algorithme.
Le deuxième problème abordé concerne la significativité des scores d'alignements obtenus de manière algorithmique. Ce problème revient à savoir si le score d'alignement est représentatif d'une ressemblance biologique entre les objets comparés. Autrement dit, étant donné un score d'alignement entre deux séquences, quelle est la probabilité d'observer ce score entre deux séquences quelconques ? Nous verrons dans ce chapitre que la recherche de cette probabilité dépend entre autres de la modélisation d'une séquence ne comportant pas d'information biologique.

Pour traiter cette question de la significativité d'un alignement, nous avons d'abord étudié statistiquement le comportement du Z-score, mesure empirique de cette significativité. Cette étude [24] menée avec des biologistes, J.-L. Risler, A. Hénaut et P. Slonimski de l'université de Versailles et du CGM de Gif-sur-Yvette, et des informaticiens de l'INRIA Rocquencourt, a permis de montrer que le Z-score suivait empiriquement une distribution des valeurs extrêmes même si on observait une sur-représentation des grandes valeurs. Pour expliquer cette distribution des valeurs extrêmes, j'ai collaboré avec J.-N. Bacro de l'INA-PG et nous avons donné une approximation asymptotique de la loi théorique du Z-score [12] qui se fonde sur les travaux de M.S. Waterman et M. Vingron [116].

Dans ce chapitre, nous expliquons comment nous avons obtenu cette approximation asymptotique de la loi de distribution du Z-score. Nous donnons aussi une estimation de la pertinence de notre approximation en utilisant des séquences « *quasi-réelles* » (*i.e.* des séquences permutées de mêmes longueurs et mêmes compositions en acides aminés). Ainsi nous montrons que l'approche de Monte-Carlo génère un biais sur l'estimation de l'un des paramètres du modèle, et nous proposons une procédure de correction. Nous montrons enfin que cette approximation asymptotique est applicable sur des séquences réelles.

1.2 Alignement de séquences

La structure primaire d'une séquence biologique est classiquement représentée par une suite finie de caractères, représentant les acides nucléiques (dans le cas d'ADN) ou acides aminés (dans le cas de protéines) constituant la séquence. L'alphabet utilisé pour représenter les séquences biologiques dépend donc de la nature des séquences : $\{A, T, G, C\}$ pour les séquences d'ADN, $\{A, U, G, C\}$ pour les séquences d'ARN, et un alphabet de 20 lettres pour les protéines (possiblement plus pour représenter des incertitudes). Une séquence est alors une suite ordonnée finie de caractères de l'alphabet considéré.

Comme les sites actifs d'une séquence tendent à être conservés au cours de l'évolution, on peut trouver des similitudes entre les protéines portant les mêmes fonctions. Ceci explique tout l'intérêt de découvrir les régions similaires entre deux protéines. Les méthodes de comparaison tentent toutes de résoudre un problème de minimisation d'une fonction *coût*. Lorsque la fonction *coût* est simple, la programmation dynamique résout efficacement le problème, mais, reste un algorithme coûteux en temps.

Schématiquement, on se donne une mesure de proximité entre les lettres de l'alphabet (matrice de similarité), et on cherche un alignement entre deux sous-séquences qui maximise le nombre de paires de lettres identiques ou proches au sens de la mesure sus-mentionnée et qui minimise le nombre de paires de caractères qui sont éloignés. Ce problème d'optimisation peut être traité par des algorithmes de programmation dynamique dont les plus connus sont ceux de Needleman & Wunsch [80] pour l'alignement global (NW) et celui de Smith & Waterman [94, 95, 96] pour l'alignement local (SW).

L'algorithme de NW répond parfaitement au problème initial de l'alignement global. On montre que le problème de la recherche de l'alignement global optimal est équivalent à la recherche d'un chemin optimal entre deux sommets dans un graphe bien choisi (cf. figure 1.1). Pour une fonction de pénalisation affine des insertions/délétions et pour deux séquences de longueur m et n, le graphe associé a exactement $(n+1) \times (m+1)$ sommets (cf. figure 1.1-a). Lorsqu'on passe à une fonction affine, le graphe a trois fois plus de sommets (cf. figure 1.1-b). La multiplication des sommets est due au fait que la pénalisation d'une extension d'une insertion/délétion n'est pas égale à celle de la création de l'insertion/délétion. Pour construire ce graphe, on distingue 6 types d'arcs :

- les arcs « *diagonaux* » modélisant les appariements,

– les arcs du type $D(i,j) \longrightarrow V(i+1,j)$ (resp. $D(i,j) \longrightarrow H(i,j+1)$) modélisant la création d'une insertion/délétion, le coût associé à cet arc est *gapopen*,
– les arcs du type $V(i,j) \longrightarrow V(i+1,j)$ (resp. $H(i,j) \longrightarrow H(i,j+1)$) modélisant l'extension d'une insertion/délétion, le coût associé à cet arc est *gapextend*,
– les arcs du type $H(i,j) \longrightarrow D(i,j)$, ou $V(i,j) \longrightarrow D(i,j)$ modélisant la fin d'une insertion/délétion dont le coût est nul.
  Ce n'est pas la seule manière de modéliser la fin d'une insertion/délétion. On pourrait par exemple rajouter les arcs $H(i,j) \longrightarrow D(i,j+1)$ et $V(i,j) \longrightarrow D(i+1,j)$. Cependant, la représentation choisie a l'avantage de pouvoir intégrer une pénalisation de fin d'insertion/délétion.

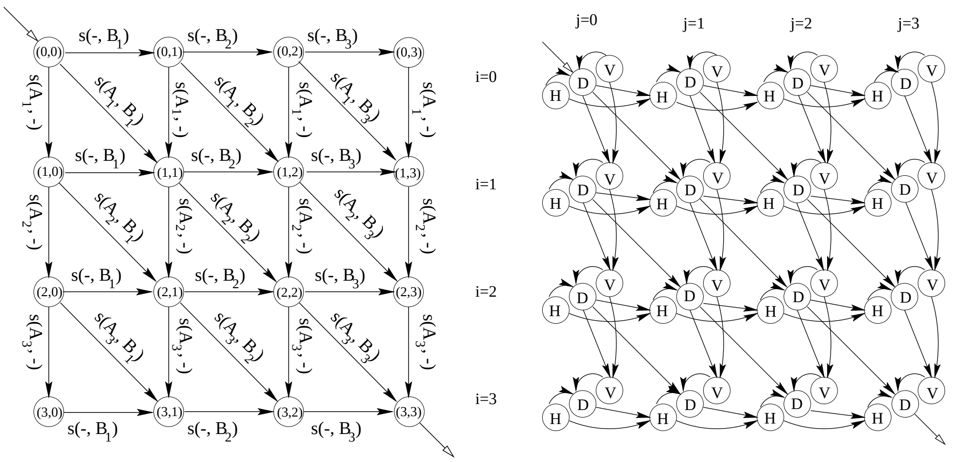


FIG. 1.1 – *Graphes associés à l'algorithme de l'alignement global dans les cas d'une fonction linéaire et affine de pénalisation des insertions/délétions.*

L'algorithme de Needleman & Wunsch est strictement équivalent à la recherche du plus court chemin dans un graphe valué *sans circuit*. La complexité de l'algorithme NW est ainsi la même que celle des meilleurs algorithmes de recherche de plus court chemin dans un graphe valué *sans circuit*. Dans le cas général, la complexité de l'algorithme de Bellman pour la recherche du plus court chemin est en $O(S \times A)$, où S et A sont respectivement le nombre de sommets et le nombre d'arcs du graphe. S'il n'y a pas de circuit, l'algorithme passe en $O(S)$, voir par exemple le livre de M. Gondran et M. Minoux [43].

Le problème de l'alignement local est un problème très similaire. Le graphe est légèrement différent en ce qui concerne l'ensemble des sommets initiaux et finaux. Lorsque la fonction de pénalisation des insertions/délétions est linéaire, le graphe associé à l'algorithme NW n'a qu'un seul sommet initial et qu'un seul sommet final (cf. figure 1.1). Dans le graphe associé à l'algorithme SW, tous les sommets sont à la fois initiaux et finaux. Lorsque la fonction de pénalisation des insertions/délétions est affine, seuls les sommets correspondant à des appariements doivent être considérés comme initiaux et finaux, puisque, des autres sommets, ne partent et n'arrivent que des arcs de coût négatif ou nul.

La théorie des graphes appelle *distancier* l'ensemble des coûts des chemins de poids optimaux entre deux sommets quelconques d'un graphe valué. L'algorithme de Smith & Waterman ne détermine que *l'unique* score optimal, et éventuellement le chemin associé. Cet algorithme n'a pas d'équivalent pour la théorie des graphes, bien qu'il soit tout à fait applicable au problème de la recherche *du* chemin de coût minimum (quels que soient les sommets de début et de fin) dans un graphe valué *sans circuit*.

1.3 Modèles statistiques pour la significativité d'un alignement

L'estimation de la significativité d'un alignement est l'un des problèmes les plus importants de la théorie de l'alignement de séquences. De cette estimation dépend la pertinence des conclusions biologiques, que ce soit en termes d'annotation des séquences, de classification de séquences, ou de reconstruction phylogénétique.

Le modèle de Karlin-Altschul. S. Karlin et S. Altschul ont proposé dès 1990 [54, 55] un modèle statistique pour exprimer la significativité d'un score d'alignement. Considérant que le score calculé était un extrême parmi les scores d'alignement possibles (la procédure recherche un optimum) ces auteurs ont proposé un modèle dérivé des travaux de E.J. Gumbel [47] sur les lois régissant les valeurs extrêmes.

Soient deux séquences A et B aléatoires indépendantes et identiquement distribuées (i.i.d.) de longueurs respectives m et n. S. Karlin et S. Altschul modélisent le nombre de régions présentant un score dépassant un certain seuil par une distribution de Poisson. Pour deux séquences i.i.d. et une matrice de similarité, le nombre d'alignements sans insertion/délétion de score supérieur à s est approximativement une loi de Poisson de moyenne :

$$E(s) = K.m.n.exp(-\lambda.s)$$

où K et λ peuvent être déduits de la matrice de similarité et de la composition en acides aminés des deux séquences. Cette valeur est appelée la E-value. Ainsi le score d'alignement sans insertion/délétion entre deux séquences i.i.d. suit une loi de Gumbel et on a :

$$P(S(A,B) \leq s) = exp(-K.m.n.exp(-\lambda.s))$$

De cette expression on peut calculer la probabilité pour que le score dépasse une certaine valeur, $p(s) = P(S(A,B) > s)$, appelée p-value. Notons d'une part que la p-value et la E-value sont équivalentes pour des scores suffisamment grands (développement limité de $1 - exp(x)$ au voisinage de 0) et d'autre part qu'il s'agit d'une approximation asymptotique (s doit être suffisamment grand). Ces remarques limitent l'utilisation de ce modèle puisque, pour être valable, il est nécessaire que les deux séquences aient des distributions similaires et que les longueurs ne soient pas trop éloignées.

Le modèle du Z-score. Le Z-score, appelé aussi Z-value, a été introduit en 1984 par D.J. Lipman *et al.* [69], il permet d'estimer la significativité statistique d'un score d'alignement obtenu par programmation dynamique (Smith & Waterman) en utilisant une procédure de type Monte-Carlo. Le principe de base de la Z-value est d'effectuer des comparaisons du score obtenu par programmation dynamique avec des scores obtenus avec des séquences aléatoires. Comme on recherche des propriétés biologiques, on va normaliser le score obtenu par rapport à des scores d'alignements entre des séquences qui ne portent que peu d'information biologique mais qui ressemblent tout de même à de vraies séquences biologiques. Le premier *modèle de séquences aléatoires* qui a été retenu est celui de séquences i.i.d. de même longueur que les séquences réelles et de même composition en acides aminés. Certains auteurs ont voulu prendre en compte aussi d'autres types d'information comme par exemple la composition en di-acides aminés voire en tri-acides aminés. Cependant le modèle le plus largement utilisé reste celui de séquences i.i.d. de même longueur et même composition que les séquences réelles. Une telle séquence peut donc être vue comme une permutation d'une séquence réelle.

Pragmatiquement, pour estimer la significativité d'un alignement entre les séquences a et b, on calcule non seulement le score d'alignement entre a et b mais aussi un grand nombre de scores d'alignement entre a et des permutations de b. On en déduit la moyenne empirique μ et l'écart-type σ de la variable aléatoire $S(a, X)$ où X est une séquence aléatoire de même longueur et composition en acides aminés que b. Le Z-score est alors défini par :

$$Z(a, b) = \frac{S(a, b) - \mu}{\sigma}$$

Puisque cet indice tient compte de la moyenne et de l'étalement, son utilisation permet de ramener à une même échelle des objets différents et du même coup, faciliter leur comparaison.

1.4 Approximation asymptotique de la loi du Z-score

Considérons $\mathbf{X} = X_1 \dots X_n$ et $\mathbf{Y} = Y_1 \dots Y_m$ deux séquences aléatoires pour lesquelles X_i et Y_j sont indépendantes identiquement distribuées. M.S. Waterman et M. Vingron [116] ont proposé une procédure pour estimer la significativité statistique du score d'alignement $S(\mathbf{X}, \mathbf{Y})$ entre $\mathbf{X}$ et $\mathbf{Y}$: une valeur approchée de la p-value pour la comparaison de $\mathbf{X}$ et $\mathbf{Y}$ peut être calculée en utilisant $1 - e^{-\gamma mnp^{S(\mathbf{X},\mathbf{Y})}}$ où γ et p sont deux paramètres à estimer. Le résultat de M.S. Waterman et M. Vingron [116] est basé sur l'approximation

$$P\left(S(\mathbf{X}, \mathbf{Y}) < t = \frac{\log nm}{|\log p|} + c\right) \simeq e^{-\gamma mnp^t} \tag{1.1}$$

qui étend l'approximation de Poisson présentée par S. Karlin et S. Altschul pour des alignements sans insertion/délétion [54]. L'approximation 1.1 a été obtenue en faisant deux étapes consécutives :

(a) Une approximation de Poisson pour la distribution du score local optimal en utilisant une heuristique dite de D. Aldous [3] : pour m et n suffisamment grands

$$P\left(S(\mathbf{X}, \mathbf{Y}) < t = \frac{\log mn}{|\log p|} + c\right) \simeq e^{-\alpha p^t} \tag{1.2}$$

où $\alpha \equiv \alpha(\mathbf{X}, \mathbf{Y})$ et $p \equiv p(\mathbf{X}, \mathbf{Y})$ sont deux paramètres positifs ($p < 1$),

(b) une normalisation par rapport aux longueurs des séquences en posant $\alpha = \gamma mn$.

Nous pouvons déduire de l'équation 1.2 que pour m et n suffisamment grands,

$$P\left(S(\mathbf{X}, \mathbf{Y}) - \frac{\log nm}{|\log p|} < c\right) \simeq \exp\left(-\exp\left(-|\log p|\left(c + \frac{\log\left(\frac{mn}{\alpha}\right)}{|\log p|}\right)\right)\right)$$

qui stipule que la distribution de $S(\mathbf{X}, \mathbf{Y}) - \frac{\log nm}{|\log p|}$ peut être approximée, pour m et n suffisamment grands, par une distribution de Gumbel de paramètres $\frac{-\log\left(\frac{mn}{\alpha}\right)}{|\log p|}$ et $\frac{1}{|\log p|}$. En utilisant les propriétés classiques de la distribution de Gumbel[3], nous déduisons que

$$E(S(\mathbf{X}, \mathbf{Y})) \simeq \frac{K + \log \alpha}{|\log p|} \tag{1.3}$$

$$\sigma^2_{S(\mathbf{X},\mathbf{Y})} \simeq \frac{\pi^2}{6(\log p)^2} \tag{1.4}$$

[3]Si U suit une loi de Gumbel de paramètres a et b, alors on a $E(U) = a + b.K$ et $\sigma_U^2 = \frac{\pi^2}{6}b^2$. On a de manière équivalente $a = E(U) - b.K$ et $b = \frac{\sqrt{6}}{\pi}\sigma_U$.

où $K = 0,57721\ldots$ est la constante d'Euler. Il est maintenant facile d'obtenir une approximation de la loi de la variable $Z(\mathbf{X}, \mathbf{Y})$: pour m et n suffisamment grands, et sous l'hypothèse (a), nous avons :

$$\frac{S(\mathbf{X}, \mathbf{Y}) - E(S(\mathbf{X}, \mathbf{Y}))}{\sigma_{S(\mathbf{X}, \mathbf{Y})}} \overset{\mathcal{D}}{\approx} \frac{\sqrt{6}|\log p|}{\pi} G(-\frac{K}{|\log p|}, \frac{1}{|\log p|}) \tag{1.5}$$

qui peut s'écrire comme

$$\frac{\pi}{\sqrt{6}} Z(\mathbf{X}, \mathbf{Y}) \overset{\mathcal{D}}{\approx} G(-K, 1). \tag{1.6}$$

En d'autres termes, notre approximation est indépendante des séquences : dans (1.6), l'approximation de la distribution de la Z-value ne dépend pas des longueurs et compositions en acides aminés des séquences alors que j'avais montré au cours de ma thèse [22] que cette propriété n'était pas vérifiée pour les scores $S(.,.)$. Bien que la dépendance des scores d'alignement en fonction de la longueur ait été largement discutée dans la littérature [6, 7, 9, 8, 54, 57, 30, 31, 56, 41, 10, 42, 115, 114, 116], il n'y a pas de résultat disponible de la dépendance en fonction de la composition. En particulier, la normalisation faite en (b) est une tentative pour prendre en compte les différentes longueurs des séquences considérées, mais ne semble correspondre que faiblement aux situations observées comme l'ont montré M.S. Waterman et M. Vingron [116].

Notons que les approximations (1.2) et (1.6) sont toutes les deux obtenues dans le cas de séquences i.i.d. Mais notre résultat ne nécessite aucune normalisation concernant les longueurs et la composition en acides aminés puisque ces deux éléments sont pris en compte à travers le processus d'évaluation de la Z-value. De ce point de vue, la Z-value est clairement utile. Mais la difficulté vient maintenant d'un point de vue pratique : comment peut on obtenir une évaluation directe des Z-values ? L'idée est d'utiliser une procédure de permutation, comme nous l'avons présenté dans notre article sur la significativité de la Z-value [24]. Schématiquement, on calcule deux Z-values différentes $\widehat{Z_1}(\mathbf{X}, \mathbf{Y})$ et $\widehat{Z_2}(\mathbf{X}, \mathbf{Y})$ par permutation de la première et de la seconde séquence respectivement et on utilise la valeur minimale pour estimer $Z(\mathbf{X}, \mathbf{Y})$. Le choix de la valeur minimale comme estimateur de $Z(X, Y)$ mène à une approche conservative (on minimise le nombre d'alignements ayant une Z-value élevée).

Comme pour obtenir notre approximation nous avons considéré que $\mathbf{X}$ et $\mathbf{Y}$ sont des séquences aléatoires i.i.d., il est naturel de tester notre résultat sur des séquences i.i.d. D'un point de vue pratique, l'hypothèse i.i.d. n'est clairement pas réaliste (et c'est pour cela que seules les toutes petites p-values sont considérées pour caractériser les valeurs de score d'alignements significatifs). Mais il n'y a pas de résultat théorique permettant d'apprécier la robustesse de l'approche de M.S. Waterman et M. Vingron ou la robustesse de notre approximation en fonction de l'hypothèse i.i.d. Un manque de robustesse pour notre approximation relatif à l'hypothèse i.i.d. serait clairement un inconvénient majeur pour les applications pratiques. Pour apprécier la robustesse de notre résultat nous avons donc décidé de tester notre approximation sur des séquences permutées construites à partir de séquences réelles.

Dans la suite nous considérons les deux ensembles de séquences que nous avons décrits dans [24] : l'ensemble des séquences réelles et l'ensemble des séquences « quasi-réelles » qui sont des séquences obtenues par permutation de séquences réelles. En dehors de la composition en acides aminés qui correspond à celle des séquences réelles, aucune autre structure particulière n'est introduite dans les séquences quasi-réelles. Les séquences quasi-réelles seront ensuite permutées pour permettre l'évaluation des Z-values.

1.5 Test de l'approximation asymptotique

1.5.1 Estimation des paramètres de la loi de Gumbel

La fonction de distribution $F_{\lambda,\delta}$ d'une variable T suivant une loi de Gumbel $G(\lambda, \delta)$ dépend de deux paramètres dont le premier est appelé *paramètre de décalage* et le second la *valeur caractéristique*. Pour évaluer la pertinence de notre approximation de Gumbel $G(-K\sqrt{6}/\pi, \sqrt{6}/\pi)$ (équation 1.6), nous considérons trois échantillons différents de Z-values et estimons les paramètres par la méthode de maximum de vraisemblance, voir par exemple le livre de N.L. Johnson et S. Kotz [53].

Les données. A partir de cinq génomes complètement séquencés, nous avons construit une banque de données de 16 956 séquences réelles [24] que nous avons ensuite permutées. Après avoir ordonné aléatoirement les séquences de cette banque, nous avons calculé tous les Z-scores entre deux séquences consécutives, obtenant ainsi 16 955 Z-values. Pour éliminer les dernières dépendances, nous avons découpé l'échantillon précédent en deux :

- le premier échantillon contient les Z-values entre la première séquence et la seconde, entre la troisième et la quatrième, et ainsi de suite (cet échantillon contient 8 478 Z-values),
- le deuxième contient les Z-values entre la deuxième séquence et la troisième, entre la quatrième et la cinquième et ainsi de suite (cet échantillon contient 8 477 Z-values).

Le tableau 1.1 donne les valeurs des estimateurs de maximum de vraisemblance pour ces deux échantillons. Un autre échantillon plus petit est aussi considéré pour essayer d'évaluer l'effet de la taille de l'échantillon. Ce dernier est construit à partir de *Saccharomyces cerevisiae* : nous avons choisi 1000 séquences au hasard, les avons permutées et avons calculé les Z-values de la première contre la seconde, de la troisième contre la quatrième et ainsi de suite.

	λ	δ
8478 Z-values	-0.549	0.789677
8477 Z-values	-0.527	0.789987
500 Z-values	-0.535	0.796190
Modèle de Gumbel	$-K\sqrt{6}/\pi = -0.45$	$\sqrt{6}/\pi = 0.7797$

TAB. 1.1 – *Estimation par maximum de vraisemblance des paramètres de la loi de Gumbel.*

Résultats. Les résultats sont légèrement différents de ceux attendus, particulièrement pour le paramètre de décalage λ. A part le biais résultant de l'estimation par la méthode du maximum de vraisemblance, un autre biais doit être exploré : celui dû au processus de permutation dans l'estimation du Z-score.

1.5.2 Permutations et estimation du biais

Les deux paramètres $\alpha \equiv \alpha(\mathbf{X}, \mathbf{Y})$ et $p \equiv p(\mathbf{X}, \mathbf{Y})$ considérés dans l'approximation poissonienne (équation 1.2) sont de différentes natures. Dans le cas i.i.d., le paramètre p ne dépend pas des positions des lettres dans les séquences mais uniquement de la composition des séquences. On en déduit que pour tout alignement entre des séquences i.i.d. de même loi, le paramètre p peut être considéré comme une constante. Mais certains scores peuvent être bien différents des autres. En conséquence, lorsqu'on compare deux séquences, il semble que le paramètre α utilisé dans l'approximation (1.2) soit dépendant non seulement des longueurs mais aussi de la structure des séquences. Puisque la procédure de permutation casse les structures mais respecte les compositions des séquences, il est naturel de considérer qu'un effet possible de la procédure de permutation affecte particulièrement le paramètre α. Ainsi, si nous supposons que le processus de permutation est appliqué à $\mathbf{Y}$, pour toutes les comparaisons $(\mathbf{X}, \mathbf{Y_i})_{i=1..N}$, les paramètres $p(\mathbf{X}, \mathbf{Y_i})$ peuvent être considérés comme une constante p alors que le rôle des paramètres $\alpha(\mathbf{X}, \mathbf{Y_i})$ doit être pris en compte.

Pour la comparaison des séquences $\mathbf{X}$ et $\mathbf{Y}_i$, d'àprès l'équation 1.5 nous avons

$$S(\mathbf{X}, \mathbf{Y}_i) \overset{\mathcal{D}}{\approx} \frac{\log \alpha(\mathbf{X}, \mathbf{Y}_i)}{|\log p|} + \frac{\Lambda}{|\log p|}$$

où Λ est une variable suivant une loi de Gumbel $G(0,1)$. Il s'en suit que

$$|\log p| \;\; E(\overline{S}_2(\mathbf{X}, \mathbf{Y})) \simeq K + \frac{1}{N} \log \left(\prod_{i=1}^{N} \alpha(\mathbf{X}, \mathbf{Y}_i) \right)$$

où $\overline{S}_2(\mathbf{X}, \mathbf{Y}) = \frac{1}{N} \sum_{i=1}^{N} S(\mathbf{X}, \mathbf{Y}_i)$. En utilisant (1.3), on obtient

$$|\log p| \Big(E(S(\mathbf{X}, \mathbf{Y})) - E(\overline{S}_2(\mathbf{X}, \mathbf{Y})) \Big) \simeq \log \alpha(\mathbf{X}, \mathbf{Y}) - \frac{1}{N} \log \left(\prod_{i=1}^{N} \alpha(\mathbf{X}, \mathbf{Y}_i) \right) \qquad (1.7)$$

qui caractérise l'estimation du biais pour la moyenne lorsqu'on permute $\mathbf{Y}$. Puisque

$$\begin{aligned} \frac{\sqrt{6}}{\pi} |\log p| \left(S(\mathbf{X}, \mathbf{Y}) - \overline{S}_2(\mathbf{X}, \mathbf{Y}) \right) \;\; &= \;\; \frac{\sqrt{6}}{\pi} |\log p| \Big(S(\mathbf{X}, \mathbf{Y}) - E(S(\mathbf{X}, \mathbf{Y})) \Big) + \\ &\quad\;\; \frac{\sqrt{6}}{\pi} |\log p| \Big(E(S(\mathbf{X}, \mathbf{Y})) - \overline{S}_2(\mathbf{X}, \mathbf{Y}) \Big) \qquad (1.8) \end{aligned}$$

nous déduisons de (1.7) que pour N suffisamment grand

$$\widehat{Z_2} \;\approx\; Z + \frac{\sqrt{6}}{\pi} \log \frac{\alpha(\mathbf{X}, \mathbf{Y})}{\left(\prod_{i=1}^{N} \alpha(\mathbf{X}, \mathbf{Y}_i) \right)^{1/N}} \equiv Z + a_2 \qquad (1.9)$$

où a_2 désigne une constante. Notons que si $\forall i \; \alpha(\mathbf{X}, \mathbf{Y}_i) \equiv \alpha(\mathbf{X}, \mathbf{Y})$, alors $a_2 = 0$.

Lorsqu'on permute la séquence $\mathbf{X}$, on arrive au même type de résultat et on a finalement : $\widehat{Z} \;=\; \min(\widehat{Z_1}, \widehat{Z_2}) \;\approx\; Z + a$ où

$$a = \frac{\sqrt{6}}{\pi} \min \Bigg(\log \frac{\alpha(\mathbf{X}, \mathbf{Y})}{\left(\prod_{i=1}^{N} \alpha(\mathbf{X}, \mathbf{Y}_i) \right)^{1/N}} \; , \; \log \frac{\alpha(\mathbf{X}, \mathbf{Y})}{\left(\prod_{i=1}^{N} \alpha(\mathbf{X}_i, \mathbf{Y}) \right)^{1/N}} \Bigg).$$

Cette étude montre clairement que l'écart entre la distribution observée et l'approximation proposée peut être simplement relié au processus de permutation.

1.5.3 Réduction du biais

Considérons la variable aléatoire $U = F(Z) = exp(-exp(-\frac{Z-\lambda_0}{\delta_0}))$ où $\lambda_0 = -K\sqrt{6}/\pi$ et $\delta_0 = \sqrt{6}/\pi$. Puisque F est la fonction de distribution de la variable Z, la variable U est distribuée uniformément sur $[0,1]$. En définissant l'échantillon ordonné $U_{(1)} \leq \ldots \leq U_{(N)}$, nous avons $U_{(i)} = F(Z_{(i)})$ et $\mathrm{E}(U_{(i)}) = \frac{i}{N+1}$. Le graphe $\left\{ \left(U_{(i)}, \frac{i}{N+1} \right), i = 1, \ldots, N \right\}$ est approximativement linéaire et il est habituel de tracer le graphe $\left\{\left(Z_{(i)}, F^{-1}\left(\frac{i}{N+1}\right)\right), i = 1, \ldots, N\right\}$, souvent appelé QQ-plot, pour tester si l'échantillon $Z_1, \ldots, Z_N$ provient d'une distribution F (on pourra se référer au livre de G.R. Shorack et J.A. Wellner [91]). Si les données sont générées par la distribution F, le QQ-plot doit être proche de la droite $y = x$; de plus, un changement de la distribution par une transformation linéaire transforme simplement le QQ-plot par la même transformation.

Séquences quasi-réelles. On teste graphiquement si la distribution F est la distribution des Z-values en traçant le QQ-plot sur les données des séquences quasi-réelles. Si notre approximation est correcte, tous les points doivent être proches de la droite $y = x$. Si la pente du QQ-plot est proche de 1, l'intersection de la droite de régression à l'origine donne une approximation a_0 du biais a. Si la pente est loin de 1, notre approximation (1.5) doit être mise en doute. Nous présentons ci-dessous le QQ-plot uniquement pour le premier échantillon (fig. 1.2). Des figures similaires sont observées avec les deux autres échantillons.

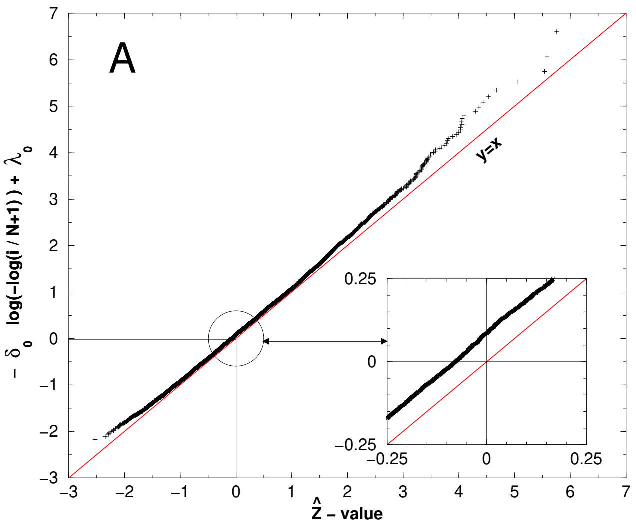


FIG. 1.2 – *QQ-plot sur les séquences* quasi-réelles *(premier échantillon : 8478 alignments). Cette figure permet une approximation du biais a_0 induit par la procédure de permutation.*

Pour tester notre loi $G(\lambda_0, \delta_0)$ nous considérons la $\widetilde{Z}$-value définie en retranchant à la Z-value l'estimation du biais introduit par le processus de permutation : $\widetilde{Z} = \widehat{Z} - a_0$. La table 1.2 donne les valeurs des estimateurs par maximum de vraisemblance lorsqu'on utilise les $\widetilde{Z}$-values. Les résultats obtenus sont proches des valeurs attendues, ce qui valide notre approximation asymptotique.

	λ	δ
8478 Z-values	-0.454	0.789668
8477 Z-values	-0.432	0.789974
500 Z-values	-0.441	0.796196
Modèle de Gumbel	$-K\sqrt{6}/\pi = -0.45$	$\sqrt{6}/\pi = 0.7797$

TAB. 1.2 – *Estimateurs des paramètres de la loi de Gumbel par maximum de vraisemblance - $\widetilde{Z}$-values.*

Séquences réelles. L'approximation de Gumbel concerne la comparaison de séquences aléatoires, c'est à dire sans structure intrinsèque. Cette hypothèse sous-jacente n'est jamais vérifiée lorsqu'on compare des séquences réelles, et des écarts à la loi de Gumbel pourraient être observés même pour des séquences qui ne sont pas reliées biologiquement. La première étape pour tester l'approximation asymptotique de la loi de la Z-value dans le cas de séquences réelles consiste donc à estimer le biais introduit par le processus de permutations.

1. On peut considérer que la valeur du biais a_0 obtenue pour des séquences quasi-réelles est une bonne estimation de ce biais pour des séquences réelles. Soit on utilise une valeur du biais « universelle » estimée sur un ensemble fixé de séquences quasi-réelles, soit on implémente la procédure complète pour les séquences réelles (nouvelle estimation du biais a_0 sur des séquences quasi-réelles obtenues à partir des séquences réelles). Dans les deux cas, on pose $\widetilde{Z} = \widehat{Z} - a_0$.
2. On pourrait aussi considérer que la valeur du biais a ne peut pas être correctement estimée : la seule information que nous ayons est donnée par les $\widehat{Z}$-values. Mais si le nombre de permutations est suffisamment grand, nous avons $a_0 \leq 0$. En effet, sous l'hypothèse nulle des séquences i.i.d., plus les séquences $\mathbf{X}$ et $\mathbf{Y}$ sont proches, plus petite est la p-value[4]. Ainsi, la fonction $\alpha(\mathbf{X}, \mathbf{Y})$ décroît en fonction de la similarité entre $\mathbf{X}$ et $\mathbf{Y}$. On s'attend donc à ce que $\alpha(\mathbf{X}, \mathbf{Y_i}) \geq \alpha(\mathbf{X}, \mathbf{Y})$ pour tout i. Dans un tel cas, l'approximation de Z par $\widehat{Z}$ mène à surestimer les Z-values.

Dans la suite, nous considérons que le biais a est bien approximé par la procédure sur les séquences quasi-réelles et nous calculons $\widetilde{Z}$ avec la valeur $a = a_0$.

Databank scanning. Les différentes banques de données de séquences comportent des séquences de longueurs vraiment différentes ainsi que des séquences qui appartiennent à la même famille, et parfois même, des séquences dupliquées. Le modèle i.i.d. n'est ainsi pas valable pour ces séquences réelles. Pour évacuer les effets dus à la duplication des séquences, nous avons construit une banque de séquences de protéines qui ne comporte qu'un seul représentant de chaque famille de protéines. Les données ont été extraites de la banque de clusters décrite par J. Park et S.A. Teichman [82] en ne retenant qu'une seule séquence de chaque cluster. Cette banque contient 618 séquences non-redondantes.

Nous avons choisi une séquence particulière (EC1003) et l'avons comparée à toutes les autres en calculant toutes les $\widetilde{Z}$-values. Le QQ-plot de ces données est montré en figure 1.3. Le modèle correspond bien aux données empiriques sur ces séquences réelles bien que les $\widetilde{Z}$-values ne constituent pas un échantillon au sens strict du terme puisque la séquence choisie est partagée par tous les alignements. Cette séquence représente le lien entre chaque alignement.

Analyse globale de génomes. Maintenant que de nombreux génomes ont été séquencés, un domaine de recherche intensif concerne la classification de séquences au sein d'un seul génome ou bien entre plusieurs génomes. On recherche les liens biologiques qui sont dus au phénomène de duplication. L'hypothèse d'indépendance des séquences n'est forcément pas vérifiée. Pour cela, on construit d'abord tous les scores d'alignements 2 à 2 et on induit une matrice de dissimilarité entre séquences. Puisque le nombre de séquences est trop grand pour simplement utiliser les méthodes classiques de classification, on a souvent recours dans une première étape à une clusterisation simple construite par *single link clustering* (deux séquences sont dans un même cluster si leur indice de similarité dépasse un seuil arbitraire). Au sein de chacun de ces clusters, une analyse hiérarchique peut être appliquée. Pour un tel but, le point le plus important

[4]Rappelons que la p-value est la probabilité que le score d'alignement ait une valeur au moins égale à celle mesurée : $P(S(\mathbf{X}, \mathbf{Y}) \geq t)$.

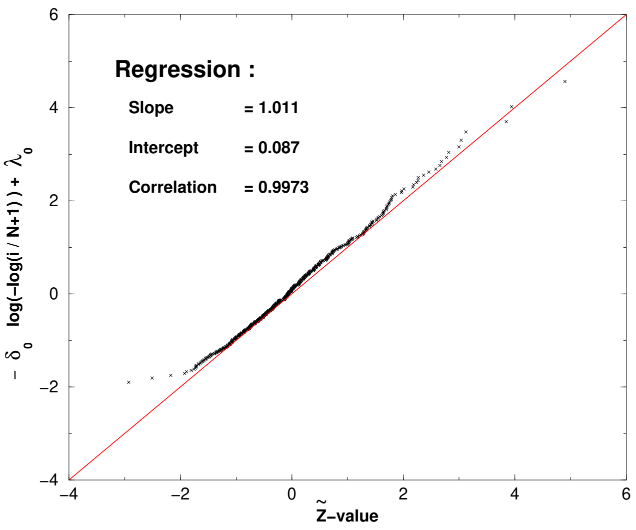


FIG. 1.3 – *QQ-plot des $\widetilde{Z}$-values obtenues lors du databank scanning : comparaison d'une séquence contre une banque non redondante de séquences.*

est d'avoir un indice global qui ne dépende pas des séquences individuelles, y compris de leurs longueurs. La Z-value est ici très utile.

A partir du génome complet de *Saccharomyces cerevisiae* nous avons choisi de manière aléatoire 1000 séquences. Toutes ces séquences ont été permutées pour construire une banque de séquences quasi-réelles. Sur les deux ensembles de séquences (réelles et quasi-réelles) toutes les comparaisons 2 à 2 ont été calculées et toutes les $\widetilde{Z}$-values calculées. La figure 1.4 montre le QQ-plots pour les deux ensembles de $\widetilde{Z}$-values non indépendantes.

En dépit de la dépendance entre les scores $\widetilde{Z}(\mathbf{X}, \mathbf{Y})$, la distribution de Gumbel correspond bien à l'observation dans le cas des séquences quasi-réelles. Dans le cas des séquences réelles, on observe un comportement bien différent : les $\widetilde{Z}$-values s'écartent significativement de la loi de Gumbel à partir d'un certain seuil comme nous l'avions déjà remarqué dans une étude empirique à grande échelle [24]. Pour des valeurs en dessous du seuil, le modèle de Gumbel semble rester valide. La valeur seuil v relative au quantile 0.9999 de la loi $G\left(-\frac{K\sqrt{6}}{\pi}; \frac{\sqrt{6}}{\pi}\right)$ (environ 6.7) peut être reliée au seuil empirique utilisé par les biologistes : en pratique, une Z-value supérieure à 8 leur indique que l'alignement est significatif.

1.6 Conclusion

Ce travail de modélisation de la loi du Z-score justifie l'utilisation de simulations pour évaluer la significativité des alignements avec insertions/délétions. Il est bien connu que la loi des scores d'alignement par l'algorithme de Smith et Waterman dépend de la longueur et de la composition en acides aminés des séquences. Cette étude donne un modèle pour la Z-value, ici une loi asymptotique, qui est indépendant des séquences. Cette propriété est fondamentale pour l'analyse ou la comparaison de génomes complets, puisqu'il est nécessaire de pouvoir comparer les scores d'alignement entre deux paires de séquences.

Dans les applications pratiques, on observe une déviation des Z-values par rapport à la distribution de Gumbel supposée. Cette divergence souligne les liens biologiques : si une Z-value est supérieure à un seuil, alors l'hypothèse nulle de séquences aléatoires est rejetée, et on peut conclure à l'existence d'un lien biologique.

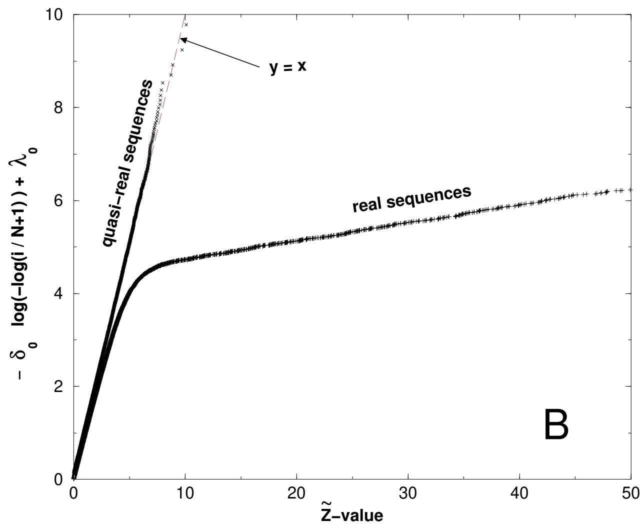


FIG. 1.4 – *QQ-plot des Z-scores pour les séquences quasi-réelles et réelles extraites de Yeast.* **Séquences réelles :** 1000 *séquences réelles ont été choisies aléatoirement dans le génome de* Saccharomyces cerevisiae. *Toutes les $\widetilde{Z}$-values ont été calculées (*499500 *$\widetilde{Z}$-values).*
Séquences quasi-réelles : 1000 *séquences quasi-réelles ont été construites par permutation des séquences réelles. Les* 499500 *$\widetilde{Z}$-values ont été calculées.*

Bien que cette méthode d'évaluation de la significativité des scores ait été développée il y a plus de 20 ans, la communauté scientifique lui a préféré le modèle d'évaluation de S. Karlin et S. Altschul même si le domaine de validité de ce dernier est plus faible puisqu'il suppose d'avoir des longueurs de séquences du même ordre de grandeur et des compositions en acides aminés identiques. Avec l'analyse de génomes fortement biaisés, un regain d'intérêt s'est porté sur la méthode du Z-score. Par exemple, O. Bastien s'est intéressé au génome de *Plasmodium falciparum* [13]. Ce génome fortement biaisé compositionnellement n'a pas pu être bien annoté par la méthodologie classique, car les séquences homologues d'autres organismes n'ont pas pu être trouvées. La raison en est simple : le modèle de significativité des scores d'alignement utilisé ne tient pas compte des biais en composition.

Ainsi, l'étude probabiliste du Z-score initiée il y a une dizaine d'années continue à être d'actualité dans des cas particuliers où les autres modèles statistiques ne sont pas applicables. La *modélisation statistique* du Z-score a été une première étape vers l'estimation de la significativité d'un alignement de séquences dont les longueurs ne sont pas comparables et qui ont des compositions en acides aminés différentes. Depuis la fin de mes travaux sur le sujet, certains auteurs comme O. Bastien et ses co-auteurs [15, 13] se sont intéressés à la définition de nouvelles méthodologies pour annoter ce type de génome. D'abord, il a fallu trouver un lien entre le Z-score et une p-value (probabilité que le score d'alignement de deux séquences aléatoires soit supérieur ou égal au score calculé). O. Bastien *et al.* [14] ont utilisé l'inégalité de Bienaymé-Chebyshev pour démontrer un théorème de limite supérieure de la p-value associée au score d'alignement. Ce théorème implique que le Z-score calculé est un test statistique et que $1/Z^2$ est une limite supérieure de la probabilité du score d'alignement et ce, quelle que soit la réelle loi de probabilité des alignements.

Plus généralement, plusieurs études à grande échelle se basent sur l'approche du Z-score pour construire des classifications de séquences protéiques. R. Apweiler, E.V. Kriventseva et leurs co-auteurs offrent une classification des séquences de SWISS-PROT et TrEMBL appelée `CluStr` [65, 5, 66] ; A. Louis, E. Ollivier, J.-C. Aude et J.-L. Risler [70] proposent d'utiliser une classification pyramidale pour aider à l'annotation des séquences génomiques ; S. Mohseni-Zadeh, A. Louis et P. Brezellec et J.-L. Risler [79] utilisent la Z-value pour construire une banque de données de clusters de protéines de plantes. Enfin, C. Lefebvre *et al.* [67] se basent entre autres sur le Z-score pour trouver des liens entre la conservation des séquences protéiques et la diversité développementale chez les invertébrés.

Chapitre 2

Puces à ADN

Résumé

L'avènement de la technologie des puces à ADN a donné de nouveaux objectifs à la recherche en bio-informatique. Non seulement, il a fallu s'intéresser au stockage et à la manipulation de cette énorme quantité de données produites, mais surtout développer de nouvelles méthodes pour extraire de ces données largement bruitées des informations biologiques pertinentes. Les premières études sur le transcriptome ont porté sur la recherche de gènes co-exprimés, ou encore sur la recherche de gènes différentiellement exprimés.

L'observation au niveau du transcriptome de la réponse d'un organisme à un stress intéresse le biologiste car elle peut le mettre sur la voie pour comprendre la fonction d'un gène ou d'une famille de gènes. D'un coté, la classification non supervisée a permis de construire des classes de gènes dont les profils d'expression sont similaires, elle met en exergue les gènes co-exprimés. D'un autre coté, la classification supervisée permet de se focaliser sur les différences qui peuvent exister entre différents échantillons. Dans cette perspective, nous nous sommes intéressés à la détermination du nombre de gènes à prendre en compte pour construire un classifieur pour distinguer deux classes. Nous avons combiné une approche statistique fondée sur l'utilisation de l'ANOVA (analyse de la variance) avec une autre méthode mettant en œuvre un algorithme d'estimation d'attributs (Relief).

Cette recherche de corrélations permet de se focaliser sur un sous-ensemble de gènes qui participent à un processus biologique. Les résultats confortent le biologiste dans ses hypothèses, ou éventuellement l'amènent sur de nouvelles pistes pour l'explication des phénomènes étudiés. Mais la biologie ne peut se contenter de corrélations, elle doit aborder les questions de causalité, *qui ne sont qu'effleurées par les outils statistiques ou algorithmiques.*

2.1 Modéliser pour typer *via* l'observation du transcriptome

La technologie des puces à ADN, ou plutôt les différentes techniques de puces, ont permis de mesurer à un moment donné les niveaux d'expression des différents gènes d'intérêt. Alors que certaines technologies permettent de déposer sur le support de la puce les gènes d'intérêt sélectionnés par l'utilisateur, d'autres font le pari d'une mesure du niveau d'expression de tous les gènes d'un organisme donné. L'exhaustivité des gènes déposés sur la puce à ADN doit permettre une recherche à l'aveugle des gènes impliqués dans la réponse de l'organisme à un stress donné.

Deux types d'approches ont été développées pour essayer de comprendre les liens entre observation du transcriptome et phénotype. La première repose sur la *classification supervisée* : on compare dans une première phase les mesures du transcriptome obtenues d'une part sur des individus du phénotype 1 et d'autre part sur des individus du phénotype 2, puis on construit un test statistique qui permet d'associer un nouvel échantillon au phénotype 1 ou 2. La deuxième approche repose sur la *classification non supervisée* : on cherche à regrouper les différents gènes en différents clusters en fonction de leur profil d'expression.

Typage par observation du transcriptome. Une des premières applications envisagées de la technologie des puces à ADN a certainement été de se servir de cette observation très fine du transcriptome pour essayer de chercher des *signatures* de maladies ou simplement d'un phénotype, voir par exemple l'article de D.K. Slonim *et al.* [93] et celui de Furey *et al.* [39]. Si le niveau d'observation des transcrits est corrélé au phénotype, alors l'observation du transcriptome est un bon moyen de signer une maladie. La notion de diagnostic assisté par puces à ADN découle naturellement de cette approche, et pourrait éventuellement permettre des diagnostics très précoces. Pour aller encore un peu plus loin, des pronostics pourraient aussi permettre de prédire l'évolution d'une maladie. Ce typage par observation du transcriptome est basé sur des observations antérieures, sur une classification supervisée des classes prédéfinies (*malade/pas malade* ou alors *malade de type A/malade de type B*) et sur les statistiques qui y sont attachées. Il ne fait pas intervenir de notion de causalité et ne se fonde que sur des corrélations statistiques. Cependant, cette méthodologie permet aux chercheurs de se focaliser sur ces gènes afin de comprendre les relations de causalité entre les niveaux d'expression observés et les phénotypes. Ces technologies sont de bons guides pour les futures expériences humides qui expliciteront les liens entre *transcripto-type* et phénotype.

Mon séjour post-doctoral au sein du Whitehead Institute de Cambridge (MA) m'avait permis de comprendre les différentes problématiques de ce typage par observation du transcriptome, mais les méthodes développées par D. Slonim *et al* [93] au sein de mon laboratoire d'accueil me semblaient essentiellement empiriques. A mon retour en France, j'ai voulu approfondir ces questions et proposer un cadre théorique à ces études statistiques. L'animation scientifique sur le site d'Evry m'a donné l'occasion de participer à un groupe de travail sur l'analyse des données d'expression puis de collaborer avec l'équipe de M. Dutreix de l'institut Curie d'Orsay. Cette équipe s'intéressait aux effets de radiations de faibles doses sur un organisme modèle et étudiait ces effets *via* des puces à ADN. Le lien entre le transcripto-type et l'irradiation de faibles doses que nous avons démontré [78], est le résultat d'un travail pluridisciplinaire avec des statisticiens (A. Antoniadis), des biologistes de l'équipe de M. Dutreix et des informaticiens (A. Cornuejols, C. Froidevaux et moi-même).

Dans ce chapitre, je défend l'idée que la modélisation statistique des niveaux d'expression peut être utile pour bien choisir les gènes témoins qui serviront au classifieur. Ici, l'approche *via* l'analyse de la variance propose un modèle où chacun des gènes subit un effet de classe. L'étude de ces effets permet de sélectionner uniquement les gènes qui subissent des effets fortement différents dans les deux classes. A cause du petit nombre de lames et du grand nombre de gènes, nous avons croisé une méthode statistique avec une autre méthode d'estimation d'attributs provenant de l'informatique.

2.2 Les données du transcriptome

De nombreux auteurs comme O. Gandrillon et R. Houlgatte [40] ont remarqué qu'il n'était pas évident de déduire des approches transcriptomiques des interprétations précises des régulations observées. D'une part, l'état biologique, et particulièrement le degré de pureté de la population de départ est un critère déterminant. La mesure du transcriptome ne permet pas de distinguer une population homogène dans laquelle toutes les cellules expriment un gène à un niveau moyen, d'une population dans laquelle la moitié des cellules expriment le gène très fortement et l'autre moitié très faiblement. D'autre part, il n'est pas très sûr que le comportement global d'une population de cellules puisse servir de base à la compréhension du comportement transcriptionnel moyen d'une cellule puisque la transcription des gènes semble être un phénomène essentiellement stochastique, voir par exemple les articles de M.B. Elowitz [33] et de M.J. Levsky [68].

Tout cela implique une grande variabilité biologique à laquelle s'ajoute une grande variabilité expérimentale. Par exemple, la fabrication de la lame support combine des problèmes de chimie et de mécanique, pouvant conduire à des disparités spatiales sur la lame. Le processus de marquage des ARNm par accrochage de molécules fluorescentes est très sensible à de nombreux paramètres comme la taille ou la composition de la molécule à marquer, entraînant un marquage différentiel de chacun des gènes. L'étape d'hybridation n'est pas complètement contrôlée, les problèmes de stabilité des fluorophores peuvent être observés, une saturation du signal lors de sa lecture peut entraîner une distribution tronquée des mesures. Les choix de pré-traitements comme la mesure du bruit de fond (niveau zéro d'expression) et la normalisation des données peuvent avoir des effets sur l'analyse des résultats.

Pour analyser de telles données, on est donc confronté à la qualité des données et à la présence de bruit. Nous nous intéressons dans ce chapitre à rechercher la significativité d'une différence de mesure d'expressions de gènes entre différents échantillons.

2.3 Analyse de l'effet de faibles doses de polluants

La détermination exacte de l'effet biologique sur l'organisme de faibles doses de polluants est d'un intérêt majeur pour la santé publique. Par exemple, l'estimation de « limite de sûreté » sur l'exposition à des radiations ionisantes est l'objet de débats permanents pour les nouvelles lois sur l'environnement. Les puces à ADN donnent un bon moyen d'étudier les changements intracellulaires à faible effet macroscopique. Cependant la faible fiabilité inhérente à ces techniques, le faible nombre de doublons de ces mesures et le manque de méthodes statistiques adaptées pour l'analyse de ce grand nombre d'attributs (gènes) affaiblissent l'interprétation des données.

Pour dépasser ce problème, nous avons combiné les résultats de deux méthodes d'analyse indépendantes (une méthode statistique et une méthode algorithmique). Nous avons appliqué ce protocole à l'analyse des profils d'expression des gènes de la levure *Saccharomyces cerevisiae* dans deux conditions expérimentales différentes : un environnement sans aucune radiation et un environnement avec radiations à faibles doses (plusieurs niveaux de radiations ont été envisagés). D'une manière générale, cette étude a montré que les radiations à faibles doses entraînaient une réponse de l'organisme importante au niveau des membranes mitochondriales. Nous démontrons aussi que les puces à ADN détectent des changements au sein de la cellule induits par des doses d'irradiations qui sont 1000 fois plus faibles que les doses minimales associées à des effets mutagénétiques. Notons enfin que de telles méthodes pourraient aussi permettre de contrôler l'environnement pour éventuellement lever des alertes.

2.3.1 Méthode fondée sur l'utilisation de l'ANOVA

Rappel sur l'analyse de la variance. L'ANOVA (Analysis Of Variance) est une méthode paramétrique fondée sur un modèle statistique particulier. Supposons que l'on observe n individus répartis en a classes. Supposons de plus que l'on connaisse pour chaque classe i, le nombre d'individus n_i appartenant à cette classe. Le modèle de l'ANOVA d'une variable mesurée Y s'écrit

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{2.1}$$

où α_i représente l'effet de la classe i et les ϵ_{ij} représentent les effets individuels et sont considérés i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. L'ANOVA consiste à comparer deux sources de variabilité :

- la variabilité intra-classes, due à la variance individuelle,
- la variabilité inter-classes, due à la fois à la variance attribuable à la classe des individus, et à la variance individuelle.

Si les classes n'ont pas d'effet sur Y, variabilités inter et intra-classes doivent être égales. Si, au contraire, les classes ont un effet, alors la variabilité inter-classes doit être supérieure à la variabilité intra-classes. Aussi l'ANOVA consiste, *in fine*, à comparer ces deux sources de variabilités et à rejeter l'hypothèse nulle (la classe n'a pas d'effet) lorsque la variance inter-classes est substantiellement plus importante que la variance intra-classes. En introduisant les notations :

$$\overline{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \overline{y}_{..} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij} \tag{2.2}$$

on peut définir les différentes notions suivantes :

Contribution des classes	Contribution individuelle E	Contribution totale T
$(a-1)$ degré de liberté (ddl)	$(n-a)$ ddl	$(n-1)$ ddl
$SC_C = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{y}_{i.} - \overline{y}_{..})^2$	$SC_I = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2$	$SCT = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{..})^2$

La statistique employée pour tester l'effet du facteur de classe est

$$F = \frac{SC_C/(a-1)}{SC_I/(n-a)} \tag{2.3}$$

qui suit, sous l'hypothèse nulle d'égalité de tous les α_i, une loi de Fisher–Snedecor à $(a-1, n-a)$ ddl. Pour un niveau α, la table de la loi de Fisher–Snedecor donne u_α tel que

$$\mathbb{P}(F(a-1, n-a) > u_\alpha) = \alpha \tag{2.4}$$

En termes de test, on a :
- $\mathbf{H}_0$: hypothèse nulle : les α_i sont tous égaux
- $\mathbf{H}_1$: il y a au moins 2 valeurs différentes parmi les α_i
- règle de décision : rejet de $\mathbf{H}_0$ pour le risque $\alpha \iff F > u_\alpha$

Application pour la recherche de gènes témoins. Dans le cadre de notre étude, on mesure pour chaque gène de la levure son niveau d'expression dans différentes conditions (radiations faibles doses / sans aucune radiation). Pour plus de robustesse, plusieurs puces à ADN sont utilisées dans chacune des conditions. Ainsi, pour chaque gène et pour chaque condition, on a un ensemble de mesures de son niveau d'expression. On peut ainsi effectuer une ANOVA pour chacun des gènes.

L'hypothèse nulle de ces analyses de la variance stipule que les moyennes des niveaux d'expression du gène considéré sont égales quelle que soit la condition expérimentale. Chaque ANOVA va donc rejeter ou non, pour un risque donné, l'hypothèse nulle sans pour autant donner les gènes qui réagissent le plus au changement d'environnement. Afin de rechercher les gènes les plus différentiellement exprimés, on évalue la statistique F_g pour tout gène g. On classe les gènes par valeur décroissante de F_g. Les gènes classés en premier, qui ont une statistique élevée, représentent donc les gènes pour lesquels l'effet environnemental est le plus important. Comme on connaît la loi de la statistique employée, on peut aussi limiter la liste aux gènes qui ont un risque inférieur à une valeur seuil.

Puisque l'on effectue des ANOVA indépendantes les unes des autres, on peut se permettre d'éliminer certaines valeurs si certains indicateurs nous le suggèrent, sans pour autant affecter la cohérence globale de l'étude. Dans de tels cas, le nombre de mesures disponibles dans chacune des conditions n'est pas le même pour tous les gènes, il en est de même pour le degré de liberté de chacune des ANOVA.

Les résultats. Avant de regarder les gènes que la méthode fait ressortir, il faut se demander si la différence de classes est observable à travers le transcriptome et à travers les statistiques de Fisher. Pour cela on regarde la répartition des gènes en fonction de la probabilité d'observer une statistique de Fisher supérieure ou égale à la valeur mesurée. Cette répartition est-elle vraiment caractéristique de la différence de classe ? Aurait-on observé le même type de distribution si les classes avaient été différentes ? Pour répondre à ces questions, on compare la répartition précédente avec la répartition qu'on aurait obtenue si les classes avaient été différentes.

- On construit des partitions aléatoires des lames, chacune de ces partitions ayant 2 classes, la première classe comportant autant de lames qu'il y en a pour les conditions normales, et la deuxième classe comportant autant de lames qu'il y en a pour l'exposition au polluant.
- On effectue ensuite, sur les mêmes données d'expression, pour chacune de ces partitions aléatoires, la même analyse par ANOVA.
- Dans le cas d'une partition aléatoire, on observe une répartition des gènes quasi-uniforme, contrairement au cas de la partition réelle (voir figure 2.1).

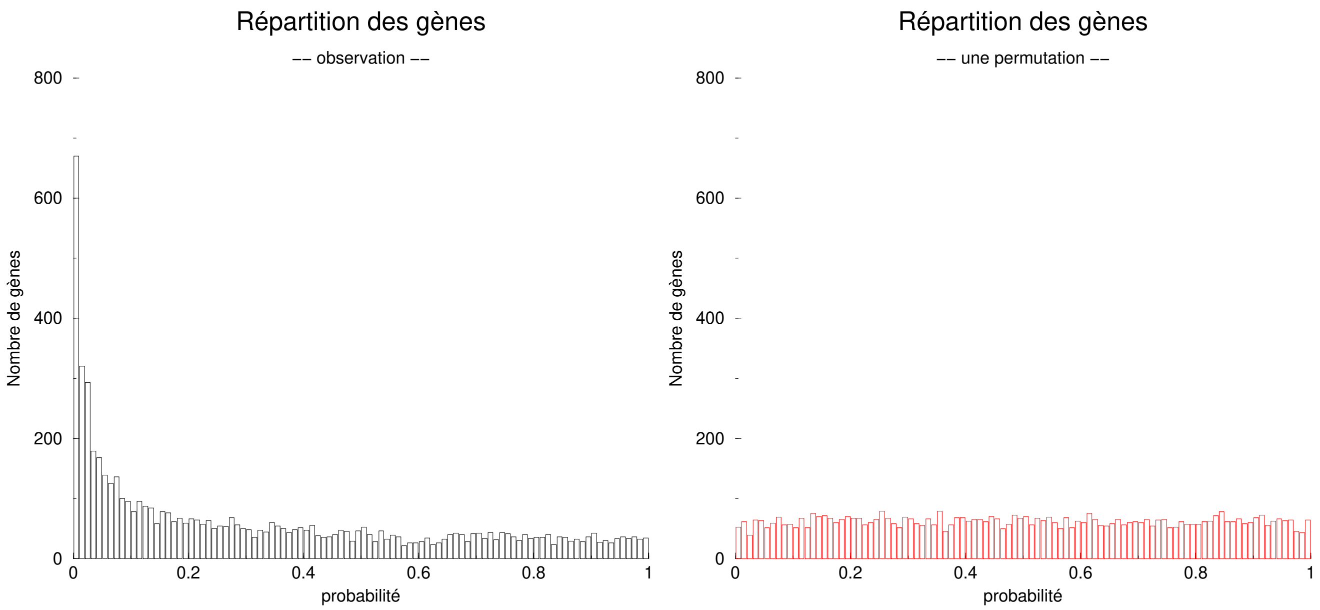


FIG. 2.1 – *Répartition des gènes en fonction de la probabilité associée à la statistique de Fisher. A gauche : la partition réelle. A droite : une partition aléatoire.*

La différence de classes est ainsi statistiquement visible à travers les statistiques de Fisher, puisqu'il y a dans le cas réel, beaucoup plus de gènes avec une faible probabilité associée. La répartition dans le cas réel donne l'impression qu'au dessus d'une probabilité de 0.1, la distribution est quasi-uniforme. Les gènes caractéristiques de la différence de classes doivent ainsi être choisis parmi les gènes ayant les plus petites probabilités (inférieures à 0.1), ou de manière équivalente, parmi les gènes ayant les plus grandes valeurs de la statistique de Fisher. Il est difficile d'en dire plus puisque si l'on considère que la distribution est uniforme dans le cas d'une partition aléatoire, alors parmi les gènes ayant une probabilité associée inférieure à 0.1, certains proviennent de cette distribution uniforme.

Si l'on génère un grand nombre de partitions aléatoires, certaines seront proches de la partition réelle, et on observera une répartition des gènes non uniforme. On compare alors la courbe représentant le nombre de gènes témoins pour un risque α, c'est à dire le nombre de gènes ayant une statistique de Fisher supérieure au quantile de probabilité α, à un grand nombre de courbes

obtenues avec des partitions aléatoires (cf. figure 2.2). La procédure est la suivante :
- Soit $C_{obs}(\alpha)$ la courbe représentant le nombre de gènes témoins sur les données initiales en fonction du risque α.
- On génère un grand nombre de partitions. Pour chaque partition p et chaque gène g, on calcule la statistique $F_{g,p}$. Pour tout α et tout p, on calcule $N_p(\alpha)$ le nombre de gènes témoins sur la partition p pour le risque α.
- Enfin, on calcule pour tout α, la distribution des $N_p(\alpha)$. On trace les courbes des quantiles 50%, 5%, 1%, 0.1%, appelées respectivement $C_{50\%}$, $C_{5\%}$, $C_{1\%}$, $C_{0.1\%}$.

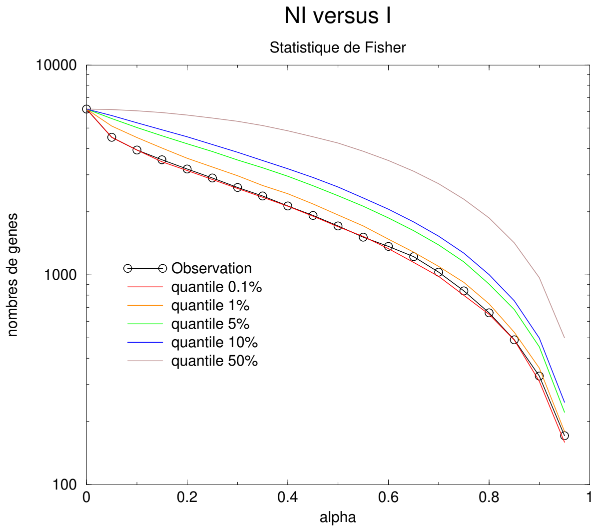


FIG. 2.2 – *Nombre de gènes dont la probabilité associée à la statistique de Fisher est supérieure à α en fonction de α. Observation : la partition réelle est utilisée. Les autres courbes correspondent aux différents quantiles pour un grand nombre de partitions aléatoires (1000).*

Cette figure montre que le nombre de gènes témoins pour un risque α sur la partition réelle est loin du nombre moyen de gènes témoins sur une partition aléatoire. La répartition des gènes est proche de celle du quantile 0.1%, ce qui signifie qu'environ une partition sur mille présente une répartition des gènes similaire à celle observée pour la partition réelle. Si la courbe réelle avait été plus proche des courbes aléatoires, on aurait dû remettre en question l'utilisation du transcriptome ou du modèle ANOVA pour étudier la différence entre les deux classes de la partition réelle.

2.3.2 Algorithme d'estimation d'attributs – Relief

La méthode présentée dans la section précédente a permis de traiter un problème d'*estimation d'attributs* qui peut être abordé par de nombreuses techniques [61, 83, 112, 120]. Les caractéristiques des données (grand nombre d'attributs, *i.e.* gènes, et peu d'exemples) nous ont amené à vouloir comparer la méthode statistique précédemment exposée avec une autre méthode provenant de l'informatique : une adaptation de la méthode Relief qui a été introduite dès 1992 par K. Kira et L. Rendell [61] et améliorée par I. Kononenko [64]. Cette méthode d'estimation d'attributs recherche les attributs qui semblent les plus corrélés aux classes.

Le principe général repose sur le calcul d'un poids normalisé pour chacun des gènes. Ces poids sont compris entre -1 et $+1$, les poids positifs indiquant une corrélation positive entre l'expression du gène sur la lame et la classe de celle-ci. Le poids d'un gène est fonction de la variation de son niveau d'expression au sein d'une même classe, comparée à la variation de ce

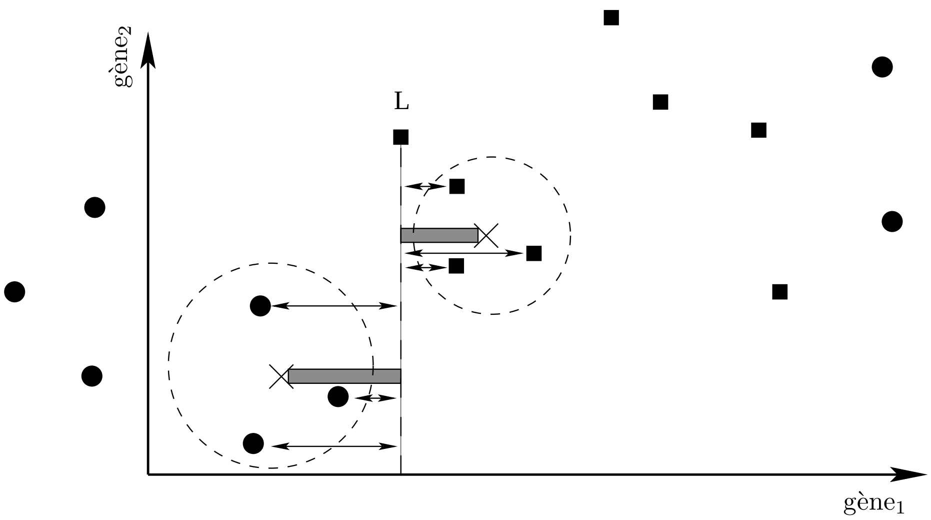


FIG. 2.3 – *Illustration de l'algorithme Relief dans le cas simple de deux gènes seulement, avec un voisinage de $k = 3$ voisins. Les lames sont représentées par des points dans le plan des niveaux d'expression des deux gènes. Pour évaluer la pertinence du gène$_1$, on prend successivement toutes les lames en compte. Ainsi pour chaque lame L, on détermine ses 3 plus proches voisins de la même classe (carrés), ainsi que ses 3 plus proches voisins de l'autre classe (ronds). On compte comme contribution positive la longueur de la projection sur gène$_1$ du segment allant de L au barycentre des 3 plus proches voisins de la classe opposée. On compte ensuite comme contribution négative la longueur de la projection du segment allant de L au barycentre des 3 plus proches voisins de la même classe. La différence des longueurs des deux gros segments mesure la contribution résultante de la lame L à la pertinence du gène$_1$ pour différencier les deux classes.*

même niveau entre les deux classes : l'expression d'un gène semble d'autant plus corrélée avec la classe de l'instance, que la variation intra-classe de l'expression du gène est petite comparée à sa variation entre classes.

La procédure de calcul du poids associé à un gène (attribut) repose sur l'estimation de deux probabilités. Une lame L est considérée comme un point dans un espace à n dimensions où n est le nombre de gènes. Pour chacune des m lames de la classe *irradiée* ou *non irradiée*, on calcule ses k plus proches voisins dans la même classe dont on calcule le barycentre H, ainsi que ses k plus proches voisins dans l'autre classe dont on calcule aussi le barycentre M. On considère alors les projections des points L, H et M selon le gène dont on calcule le poids. On détermine ensuite, d'une part, la distance entre la projection de L et la projection de M, et, d'autre part, la distance entre la projection de L et la projection de H (voir figure 2.3). La différence entre ces deux distances fournit la contribution de cette lame au calcul du poids du gène considéré. On répète ce calcul pour toutes les lames et on somme les contributions obtenues en normalisant par le nombre de lames. Pour $k = 1$, ce calcul correspond à la formule suivante :

$$\text{poids}(g) = \frac{1}{m}\sum_{L=1}^{m}\{|\text{expr}_g(L) - \text{expr}_g(M_1)| - |\text{expr}_g(L) - \text{expr}_g(H_1)|\}$$

où $\text{expr}_g(x)$ est la projection selon le gène g du point x, et m est le nombre total de lames.

Le paramètre k (nombre de voisins utilisés) permet de contrôler le compromis entre sensibilité et robustesse au bruit. La version initiale que nous avons publiée [73] ne nous permettait pas de choisir un paramètre k différent selon la classe. Comme cette fonctionnalité peut se révéler utile si l'une des classes comporte beaucoup plus d'instances que les autres, nous avons choisi d'intégrer la possibilité de choisir une valeur de k différente pour chacune des classes. Une autre variante, BioRelief, a ainsi été proposée.

L'existence de données manquantes fait émerger un autre point délicat. En pratique, pour calculer la distance entre deux lames, nous commençons par supprimer tous les gènes (attributs) pour lesquels une des valeurs manque. D'autres choix sont possibles, comme par exemple, remplacer les valeurs manquantes par 0. Ce choix est dû à la forme des données : il s'agit d'un log-ratio, donc biologiquement, si la valeur est à 0, c'est que le gène s'est exprimé de la même manière dans l'échantillon test que dans l'échantillon de référence. Une autre méthode consiste à estimer la valeur manquante à partir des données.

Il reste à choisir les « meilleures » valeurs des paramètres k_1 (classe *non irradiée*) et k_2 (classe *irradiée*) en fonction des données. Augmenter la valeur de k permet d'augmenter la résistance au bruit puisqu'un point aberrant sélectionné comme plus proche voisin aura une influence tempérée par d'autres points. En revanche, une faible valeur de k, en prenant moins en compte la moyenne des points d'une classe, conduit à une plus grande sensibilité aux particularités de la classe considérée. Nous pensons qu'une bonne solution à ce compromis consiste à se comparer au hasard et à essayer d'en être le plus éloigné possible. Cette notion d'éloignement est calculée en mesurant l'aire entre la courbe de répartition des poids pour k_1 et k_2 fixés et la courbe de répartition des poids pour des classes aléatoires. Maximiser cette surface nous suggère de prendre $k_1 = 2$ et $k_2 = 5$ [74]. Cependant, les résultats s'appuyant sur les valeurs $k_1 = 3$ et $k_2 = 3$ sont peu éloignés de l'optimum, et nous avons utilisé ces valeurs dans le cadre de la détection des radiations de faibles doses [78].

2.3.3 Les deux méthodes font-elles ressortir les mêmes gènes témoins ?

Les deux méthodes précédentes ne font pas ressortir les mêmes gènes significatifs de la différence de classes. Par exemple, si on considère les deux ensembles de 500 gènes témoins (obtenus par une ANOVA simple et obtenus par la méthode BioRelief), leur intersection représente à peu près la moitié des gènes. Ce recouvrement est-il dû au hasard ou bien montre-t-il que les 2 méthodes font ressortir les mêmes gènes ?

Loi de probabilité. Pour répondre à cette question, on considère que chacune des deux méthodes définit deux classes de gènes : les gènes choisis comme significatifs de la différence de classes (il y en a 500) et les autres.

On souhaite estimer la significativité du recouvrement des deux ensembles de gènes sélectionnés en estimant la probabilité qu'une méthode aléatoire choisisse autant de gènes parmi les gènes choisis par la première méthode. Une méthode aléatoire consisterait à choisir au hasard 500 gènes parmi les N gènes considérés (dans cette étude $N = 6157$). Le nombre de gènes choisis par la méthode aléatoire ayant déjà été choisis par la première méthode suit une *loi hypergéométrique* $\mathcal{H}(n, N - n, n)$ où n est le nombre de gènes déjà choisis par la première méthode (500), N le nombre total de gènes (6157), et où le troisième paramètre est le nombre de gènes à choisir par la méthode aléatoire (500).

On calcule alors la probabilité P pour que la méthode aléatoire choisisse au moins 250 gènes parmi les 500 gènes déjà choisis par la première méthode.

$$P \quad = \quad \sum_{i=250}^{500} \mathbb{P}_{hypergeometric}(X = i) \quad = \quad \sum_{i=250}^{500} \frac{C_n^i C_{N-n}^{n-i}}{C_N^n} \quad = \quad 5.6 \times 10^{-160}$$

Cette probabilité reste du même ordre pour différentes tailles d'ensembles de gènes sélectionnés (pour un nombre de gènes compris entre 100 et 2000, la taille de l'intersection des deux ensembles de gènes est supérieure à la moitié du nombre de gènes considérés). On en conclut que les méthodes BioRelief et ANOVA produisent en partie la même information.

Une comparaison similaire des ensembles de gènes sélectionnés par BioRelief, ANOVA et SAM (méthode d'analyse de la significativité des données de microarrays introduite par V. Tusher *et al.* [112]) donne les résultats suivants : 409 gènes communs dans les 500 premiers classés par ANOVA et SAM, soit une proportion de 82%. Parmi les 35 premiers gènes classés par BioRelief, 8 ont été classés parmi les 35 premiers par SAM, et les 8 mêmes ont aussi été classés parmi les 35 premiers par ANOVA (la probabilité est de 10^{-12}). Compte tenu de l'absence d'hypothèse d'indépendance entre les expressions de gènes dans BioRelief et de sa bonne résistance au bruit, l'interprétation biologique que nous avons faite dans le cadre de l'étude des radiations de faibles doses [78] porte sur les gènes obtenus par BioRelief.

BioRelief et les méthodes non paramétriques. L'archétype de la méthode paramétrique pour la détection des gènes impliqués dans la réponse biologique à un stress donné est la méthode de l'analyse de la variance. Afin de mesurer le degré d'implication de chaque gène, cette méthode compare l'hypothèse H_1 selon laquelle les moyennes d'expression des gènes diffèrent suivant la classe de la lame, avec l'hypothèse nulle H_0 selon laquelle ces moyennes sont égales. Pour ce faire, le test de Student suppose que les niveaux d'expression des gènes suivent des lois normales, de variances égales ou non entre les classes. On peut alors ordonner les gènes en fonction du seuil de rejet de l'hypothèse H_0 leur correspondant. SAM est aussi une méthode paramétrique puisque elle suppose aussi que les données suivent une loi normale. A l'opposé, BioRelief est une méthode non paramétrique.

En s'affranchissant de toute hypothèse sur le type de distribution des données, la méthode BioRelief est insensible à leur adéquation aux données disponibles. Cela est particulièrement intéressant pour l'analyse du transcriptome pour laquelle les hypothèses paramétriques semblent souvent trop fortes. Aucune méthode ne pouvant cependant être uniformément supérieure aux autres, il reste à déterminer les champs d'application les plus favorables à BioRelief.

2.4 Modélisation, inférence de causalité, réseaux biologiques

La recherche des corrélations permet comme nous venons de le voir de se focaliser sur un sous-ensemble de gènes qui participent à une réponse à un stress donné. Le résultat de cette étude conforte le biologiste dans ses hypothèses, ou éventuellement l'amène sur de nouvelles pistes pour l'explication des phénomènes étudiés. Mais la biologie ne peut se contenter de corrélations, car son objet d'étude est le phénomène biologique lui-même, elle doit expliciter les raisons de ces corrélations observées, c'est à dire elle doit aborder les questions de *causalité*. Les outils de statistique paramétrique ou algorithmique, ne permettent pas d'aborder ces aspects, et c'est en ce sens qu'ils permettent de se faire des idées tout en cachant l'essentiel.

La causalité de la réponse à un stress se focalise sur l'enchaînement des réactions qui se produisent au sein de la cellule pour l'amener à répondre de manière adéquate au stress. Il s'agit donc de comprendre le fonctionnement des réseaux d'interactions biologiques, qu'ils soient génétiques, métaboliques, qu'il s'agisse d'interactions macro-moléculaires ou de la transduction du signal... Cette question a donc fait émerger un nouvel intérêt pour la biologie des systèmes.

La recherche post-génome a pour but de donner un sens à toutes les informations générées par le séquençage du génome, par les données moléculaires et par les énormes bases de données transcriptomiques. Elle s'intéresse entre autres au fonctionnement du génome pris comme un système à part entière, dont les variables élémentaires sont les gènes, les protéines, et les métabolites... Les parties qui suivent se focalisent sur l'étude des réseaux génétiques, dont un premier modèle discret a été développé par R. Thomas. En reprenant ses travaux, nous avons montré comment l'informatique et certains développements récents en vérification et génie logiciel peuvent aider le biologiste à décrire la dynamique des systèmes qu'il étudie.

Chapitre 3

Fondements sémantiques des réseaux de régulation génétique

Résumé

La recherche post-génome a pour but de donner un sens à toutes les informations générées par le séquençage du génome et toutes les techniques transcriptomiques et protéomiques. Elle s'intéresse au fonctionnement du génome pris comme un système à part entière, dont les variables élémentaires sont les gènes, les protéines, et les métabolites... La modélisation en biologie se tourne ainsi vers les aspects dynamiques en représentant l'enchaînement des causes qui produisent un comportement observable.

Dans ce chapitre, nous nous focalisons sur les réseaux de régulation génétique et montrons que la modélisation discrète de R. Thomas permet de définir leur dynamique. Après avoir définis précisement sa sémantique, nous montrons que cette modélisation est reformulable en termes de réseau de Petri de haut niveau. Différentes extensions permettant d'élargir l'ensemble des comportements pris en compte dans la modélisation sont aussi présentées.

3.1 Modéliser pour comprendre

La recherche post-génome a pour but de comprendre les mécanismes sous-jacents qui gouvernent les informations générées par le séquençage du génome et toutes les techniques transcriptomiques et protéomiques. La biologie moléculaire s'intéresse entre autres au fonctionnement du génome pris comme un système dynamique à part entière, dont les variables élémentaires sont les gènes, les protéines, et les métabolites. La compréhension des processus biologiques complexes nécessite le développement de modèles qui combinent des informations quantitatives avec des détails moléculaires. Par exemple, les réactions élémentaires (associations, transformations biochimiques, diffusions, *etc.*) sont exprimées à l'aide d'équations (souvent des systèmes d'équations différentielles ordinaires). Les modèles numériques résultants, sont basés sur les données de la littérature lorsqu'elles sont disponibles. Ils peuvent alors être utilisés pour faire des prédictions qui peuvent à leur tour être confrontées à des expériences. Tout cela constitue un champs disciplinaire qualifié d'« épigénomique ».

Le choix de la méthode mathématique utilisée dépend non seulement des données disponibles à incorporer dans le modèle mais aussi de l'objectif de la modélisation. Par exemple, les modèles numériques sont bien adaptés pour calculer les traces dans l'espace des phases une fois que tous les paramètres du système d'équations différentielles sont fixés. Les simulations qui en découlent font ressortir des caractéristiques précises. Si certains paramètres du système ne sont pas connus, alors la majeure partie de l'effort de modélisation est consacrée à la recherche de paramètres menant à des simulations cohérentes avec les connaissances disponibles sur le système biologique.

De nombreuses modélisations mathématiques différentes ont été proposées pour modéliser les réseaux de régulation génétique, voir par exemple la synthèse de H. de Jong [27], l'article de P. Smolen *et al.* [97] et le livre de D.J. Wilkinson [117]. Les approches les plus couramment utilisées sont certainement les approches quantitatives à base d'équations différentielles ou stochastiques. Ces approches permettent d'obtenir des prédictions numériques précises sur les propriétés dynamiques du système. Cette précision n'est encore malheureusement que numérique car les paramètres de ces systèmes sont peu ou pas du tout connus. D'autres types de modélisation, d'inspiration qualitative, ont été développés ; on peut citer les réseaux booléens et leurs généralisations, les réseaux de Petri, et les automates hybrides. Naturellement, chaque cadre de modélisation met l'accent sur certains éléments et permet soit de détailler, soit d'abstraire, différents aspects. Ces modélisations peuvent être classées en fonction de différentes caractéristiques : certaines introduisent un formalisme discret, d'autres un formalisme continu ou hybride (*i.e.* qui mélange des variables continues et des variables discrètes) ; certaines utilisent des mécanismes déterministes, d'autres des mécanismes stochastiques ; certaines reposent sur une description qualitative et d'autres sur une description quantitative. Cette abondante diversité de modélisations possibles nous amène à définir un critère pour nous aider à en choisir une.

Le Rasoir d'Occam. Il s'agit du principe de raisonnement que l'on attribue au moine franciscain et philosophe logicien Guillaume d'Ockham[1], mais qui était connu et formulé avant lui. L'énoncé du XVe siècle de ce principe est le suivant : « Les choses essentielles ne doivent pas être multipliées sans nécessité »(*pluralitas non est ponenda sine necessitate*).

Aussi appelé « principe de simplicité », « principe de parcimonie », ou encore « principe d'économie », il exclut la multiplication des raisons et des démonstrations à l'intérieur d'une construction logique. Le principe du rasoir d'Occam consiste à ne pas multiplier les hypothèses au-delà du nécessaire, en d'autres termes à toujours privilégier aussi longtemps que cela reste compatible avec les observations, l'hypothèse la plus simple parmi toutes celles qui peuvent être envisagées. Bien que ce principe ait une apparence fort raisonnable, certaines difficultés surgissent puisque ce qui est considéré comme simple dépend du cadre de référence utilisé, et par voie de conséquence du langage et des habitudes.

De nombreux chercheurs qui s'intéressent à la modélisation de systèmes complexes en biologie, utilisent comme principe directeur important le rasoir d'Occam d'origine : ce principe impose de retenir dans un cadre informel l'explication la plus simple consistante avec les données, et dans un cadre rigoureux, voire formel, il impose de retenir le modèle le plus simple consistant avec les données. Il est ainsi préférable de choisir le modèle qui donne la « meilleure » adéquation aux données en utilisant le plus petit nombre de paramètres. La théorie la plus simple qui représente les faits d'un objet d'étude, avec le moindre nombre de suppositions est celle qui doit être sélectionnée. Les observations qui ne peuvent pas être expliquées grâce au modèle remettent en cause le modèle lui-même, en d'autres termes, le modèle n'accepte pas ou peu de faux négatifs. Le challenge est donc de construire un modèle qui permette d'expliquer le plus de données disponibles tout en réduisant le nombre d'hypothèses faites à propos du système.

Les modèles ne font souvent qu'imiter des phénomènes particuliers et isolés, on ne cherche bien souvent qu'à modéliser un petit sous-ensemble de l'espace des comportements potentiels. Ainsi le choix des données les plus riches mène à sélectionner des modèles qui représentent bien des données spécifiques mais qui sont limités dans leur utilité et leur pouvoir prédictif.

La théorie discrète des réseaux génétiques. La théorie des réseaux génétiques développée par R. Thomas et ses différents co-auteurs [108, 106, 100, 111, 104, 105] permet de représenter

[1] 1285–1349

avec un graphe les influences mutuelles d'un groupe de gènes au sein d'une cellule. Chaque interaction est décrite par un signe (activation ou répression) et par un seuil de déclenchement. Il a été conjecturé puis démontré que les cycles de ce graphe jouent un rôle majeur pour l'identification des états stables possibles. L'existence de plusieurs états stables implique des comportements différents de la cellule, tout en conservant le même génome. L'environnement peut alors entraîner des changements d'états stables d'une même cellule sans mutation (phénomène épigénétique). Lorsqu'un gène appartient à deux cycles et que les influences de ces cycles sur ce gène sont opposées, la prédiction « intuitive » du comportement du système devient impossible. Elle repose alors essentiellement sur l'expertise des biologistes, et le passage à l'échelle semble très difficile sans cadre théorique plus approfondi.

Les travaux impressionnants de R. Thomas et E.H. Snoussi [98, 100] ont permis de faire le lien entre la représentation des traces possibles d'un modèle « logique » et les dynamiques représentées par des systèmes d'équations différentielles linéaires par morceaux. Nous avons voulu donner un cadre formel à cette théorie afin de pouvoir automatiser les raisonnements sur l'ensemble des dynamiques possibles. Les discussions régulières avec D. Thieffry de l'ESIL puis avec M. Kaufman de l'université libre de Bruxelles nous ont d'abord permis d'avoir une bonne connaissance des travaux de l'équipe de R. Thomas. J'ai ensuite rejoint le groupe de travail du *programme d'épigénomique* animé par G. Bernot et J. Guespin portant sur le lien entre modèles et observabilité. Ce groupe de travail a mené à une collaboration avec J. Guespin (laboratoire de microbiologie du froid, université de Rouen) sur l'étude de la production de mucus chez *Pseudomonas aeruginosa* [17] et sur la cytotoxicité dans la même bactérie [38]. Afin de bénéficier de plateformes logicielles développées dans le cadre de l'informatique, j'ai collaboré avec H. Klaudel, spécialiste des réseaux de Petri, afin de reformuler la formalisation déjà faite en termes de réseaux de Petri [25]. Nous avons développé avec G. Bernot et F. Delaplace du laboratoire IBISC, O. Roux et F. Cassez de l'IRCCyN, un modèle permettant plus de souplesse dans la description des domaines d'activités des interactions [16] ainsi qu'un modèle avec délais d'activation/inhibition [2]. Enfin, c'est avec A. Richard qu'a été introduit un modèle formel représentant dans la dynamique les états sur seuils [86].

Ainsi, nous présentons dans ce chapitre la théorie de R. Thomas et en donnons un cadre sémantique. Nous montrons ensuite que la puissance d'expression des réseaux de Petri est suffisante pour exprimer cette théorie, ce qui nous permet de raisonner sur ces modèles et de prévoir les différentes évolutions possibles. Enfin, nous donnons quelques extensions possibles de la théorie de R. Thomas.

3.2 La théorie de R. Thomas

Les éléments de bases de l'organisme, comme les protéines, sont codés dans le génome, information qui se transmet de génération en génération et qui peut évoluer selon le principe de mutation. Ces protéines participent activement aux processus biologiques qui sont sensibles à la concentration de chacune d'entre elles. La synthèse des protéines et leur régulation sont des phénomènes complexes dans lesquels de nombreuses autres protéines et ARN jouent un rôle important. Les macromolécules se régulent ainsi mutuellement et forment ce que l'on appelle un réseau de régulation.

L'état à un instant donné d'une cellule est principalement déterminé par la concentration de l'ensemble des macromolécules du réseau. Deux états différents pourront éventuellement mener à des comportements différents, ce qui ouvre de nouvelles perspectives puisqu'on peut envisager que certains phénotypes différents proviennent d'un changement de l'état du réseau et non du changement de la nature des macromolécules (mutation). Ce type de phénomène est qualifié d'« épigénétique ».

Les réseaux de régulation génétique modélisent les interactions entre objets biologiques (gènes et protéines essentiellement). Les biologistes établissent des modèles moléculaires représentant l'ensemble des relations connues pour les gènes et protéines régulatrices qui, en se liant avec ces gènes favorisent (effet positif) ou empêchent (effet négatif) leur « expression » c'est-à-dire *in fine* la synthèse des protéines codées par ces gènes (la régulation post-transcriptionnelle et post-traductionnelle est souvent prise en compte dans l'abstraction faite de ces régulations). Une même protéine peut être régulatrice de plusieurs gènes et avoir un effet positif ou négatif selon la cible considérée. Il est ainsi possible de considérer de tels modèles comme des graphes dont les nœuds représentent ces objets biologiques et les arcs leurs interactions. L'aspect statique des relations entre les gènes et les protéines a été jusqu'à présent un centre d'intérêt majeur de la biologie moléculaire, mais on se rend compte que cet aspect statique ne suffit pas à rendre compte de phénomènes dynamiques. En associant à chaque instant et à chaque nœud une valeur numérique qui décrit le niveau de concentration de son produit dans la cellule, les aspects dynamiques deviennent une nouvelle voie d'étude de l'organisme.

Les premières modélisations de ces dynamiques étaient basées sur les systèmes d'équations différentielles, dans lesquels chaque équation décrit l'évolution de la concentration d'un type de macromolécules. Dans les années 1970, R. Thomas a introduit une approche booléenne dans laquelle chaque entité est allumée ou éteinte et qui capture qualitativement la dynamique [106]. Il a ensuite montré que la présence de circuits de rétroaction dans ces réseaux avait une importance capitale pour le comportement du système [107] : pour avoir multi-stationnarité, le système doit présenter un circuit positif (chaque entité du circuit a une influence positive sur sa propre évolution) et pour avoir homéostasie, état d'équilibre vers lequel le système converge où autour duquel il oscille, il doit présenter un circuit négatif (chaque entité du circuit a une influence négative sur sa propre évolution). Nous présentons dans cette section cette approche discrète qui est souvent plus favorable car les données biologiques obtenues sont rarement quantitatives.

3.2.1 Graphes d'interactions

La figure 3.1 considère un gène u qui agit positivement sur une cible v et négativement sur une autre cible v'. Les deux courbes donnent les concentrations de v et v' en fonction de la concentration de u, après un délai suffisamment long pour que v et v' se stabilisent. Comme ces interactions sont presque toujours de nature sigmoïdale, on peut isoler 3 régions intéressantes pour le comportement de u :

- dans la première région, u n'agit ni sur v ni sur v',
- dans la seconde, u agit sur v mais n'agit pas sur v', et,
- dans la dernière, u agit à la fois sur v et sur v'.

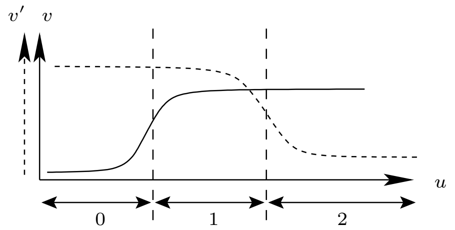


FIG. 3.1 – *Définition des niveaux de concentration abstraits*

Ceci justifie la discrétisation de la concentration de u : les trois niveaux de concentration abstraits (0, 1 et 2) correspondant aux trois régions précédemment décrites constituent l'information minimale utile d'un point de vue qualitatif. Plus généralement, si une variable agit sur n variables, au plus $n+1$ niveaux abstraits de concentration sont considérés (chaque valeur comprise entre 0 et n), éventuellement moins car deux seuils pour deux cibles peuvent être égaux.

Pour définir de manière formelle un réseau de régulation génétique, nous utilisons un graphe dirigé étiqueté. Un sommet représente un gène (il abstrait en fait le gène et son produit) et il est muni d'une borne qui correspond au maximum des valeurs que son niveau de concentration abstrait peut prendre. Un arc $(u \to v)$ est étiqueté par un seuil de déclenchement et par un signe : $+$ si u active v et $-$ s'il l'inhibe.

Définition 1 *Un* graphe de régulation génétique *est un graphe dirigé étiqueté $G = (V, E)$ où :*
- *chaque sommet v de V, appelé* variable*, est muni d'une* borne $b_v \in \mathbb{N}^*$ *plus petite ou égale au degré sortant de v dans G, sauf si le degré sortant est nul auquel cas $b_v = 1$,*
- *chaque arc $(u \to v)$ de E est étiqueté par un couple (t, ε) où t, appelé* seuil*, est un entier compris entre 1 et b_u, et $\varepsilon \in \{-, +\}$.*

3.2.2 Les paramètres qualitatifs

Nous avons déduit les niveaux de concentration à partir des interactions et de leurs seuils. Ils apparaissent par exemple sur l'axe horizontal de la figure 3.1. Il reste à considérer l'axe vertical : en supposant que $u_1 \cdots u_n$ ont une influence sur v (arcs entrants $u_i \to v$), quel est le niveau de concentration vers lequel v est attiré ? Cependant, tous les gènes prédécesseurs d'une cible donnée v, ne sont pas actifs pour v. Ainsi la cible v peut être régulée par différents sous-ensembles ω de ses prédécesseurs et nous noterons $k_{v,\omega}$ les niveaux de concentration abstraits vers lequel v est attiré lorsque ω est l'ensemble des gènes qui aident v à s'exprimer (figure 3.2). Les réseaux

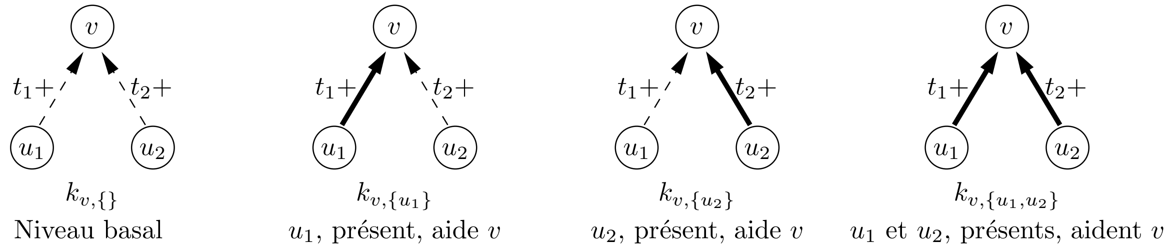


FIG. 3.2 – *Lorsque la concentration d'un prédécesseur de v ne dépasse pas son seuil d'activation, l'interaction est représentée en pointillé.*

de régulation génétique sont alors définis comme des graphes de régulation génétique munis des paramètres $k_{v,\omega}$ (définition 2).

Définition 2 *Un* réseau de régulation génétique *est un couple $\mathcal{R} = (G, \mathcal{K})$ où $G = (V, E)$ est un graphe de régulation génétique et $\mathcal{K} = \{k_{v,\omega}\}$ est une famille d'entiers positifs indexés sur l'ensemble des couples (v, ω) tels que*
- *v appartient à V,*
- *ω est un sous-ensemble de $G^{-1}(v)$,*
- *$k_{v,\omega} \leq b_v$.*

où $G^{-1}(v)$ désigne l'ensemble de tous les prédécesseurs d'un gène v dans le graphe G.

3.2.3 Construction de la dynamique (sémantique)

Pour définir la dynamique associée à un réseau de régulation génétique, nous commençons par définir un état du système comme la donnée des niveaux de concentration de chacun des gènes. Pour un état donné, l'ensemble des ressources d'un gène peut être déduit des seuils de chacun de ses arcs entrants ainsi que de l'état du système.

Définition 3 *Étant donné un réseau de régulation génétique, un état* $(n_{v_1}, ..., n_{v_p})$ *et un arc* $(v_i \rightarrow v_j)$ *étiqueté par* (t, ϵ)*, le gène* v_i *est une* ressource *de* v_j *si :*
- $n_{v_i} \geq t$ *lorsque* $\epsilon = +$
- $n_{v_i} < t$ *lorsque* $\epsilon = -$

Ceci définit l'ensemble des ressources ω *du gène* v *à un état donné comme un sous-ensemble de* $G^{-1}(v)$*.*

La figure 3.3 illustre comment sont calculées les ressources lorsqu'il y a à la fois des activateurs et des inhibiteurs : l'absence d'un inhibiteur joue un rôle similaire à la présence d'un activateur.

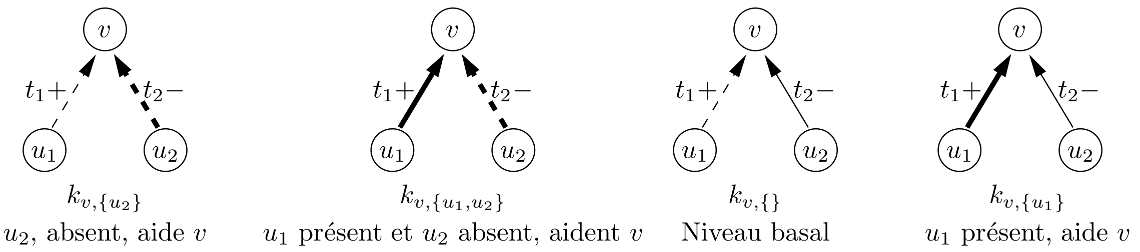


FIG. 3.3 – *Lorsque la concentration d'un prédécesseur de* v *ne dépasse pas son seuil d'activation/inhibition, l'interaction est représentée en pointillé.*

Les paramètres de $\mathcal{K}$ définissent de manière triviale un système de transitions, appelé *système de transitions synchrones*, qui est une étape technique intermédiaire pour définir la sémantique.

Définition 4 *Soit* $\mathcal{R} = ((V, E), \mathcal{K})$ *un réseau de régulation génétique, son* système de transitions synchrones *(STS)* $\mathcal{S} = (S, T)$ *est défini comme suit :*
- *l'ensemble des états* S *contient tous les états possibles,* i.e. $\prod_{v \in V} [0, b_v]$.
- *l'ensemble* T *contient les transitions de la forme :* $(n_{v_1}, \cdots, n_{v_p}) \rightarrow (k_{v_1,\omega_1}, \cdots, k_{v_p,\omega_p})$ *où pour tout* i*,* ω_i *est l'ensemble des ressources de* v_i *à l'état* $(n_{v_1}, \cdots, n_{v_p})$*.*

Comme le degré sortant de chaque état est exactement 1, ce système de transitions peut être représenté par un graphe dans une grille de dimension $|V|$ où de chaque état part un unique arc vers l'état suivant (voir figure 3.4). L'état suivant est parfois appelé *point focal* car il correspond à un point vers lequel sont attirées toutes les trajectoires du système d'équations différentielles linéaires par morceaux sous-jacent (voir sous-section 3.2.4). Cependant cette dynamique synchrone présente au moins deux inconvénients comme l'ont remarqué R. Thomas et M. Kaufman [109] :
- une flèche diagonale dans le système de transitions synchrones est une transition qui change *simultanément* la concentration d'au moins deux variables. La probabilité que deux variables passent leurs seuils respectifs au même moment est négligeable *in vivo*, mais nous ne connaissons pas laquelle passera son seuil en premier.
- Une flèche de longueur supérieure à 2 implique qu'il existe au moins une variable qui augmente sa concentration de manière abrupte et saute plusieurs seuils en une seule étape.

Nous introduisons donc une autre dynamique, appelée dynamique asynchrone, dans laquelle toute transition diagonale a été remplacée par la collection des transitions qui modifient seulement l'une des variables impliquées à la fois, et dans laquelle chaque transition fait évoluer les

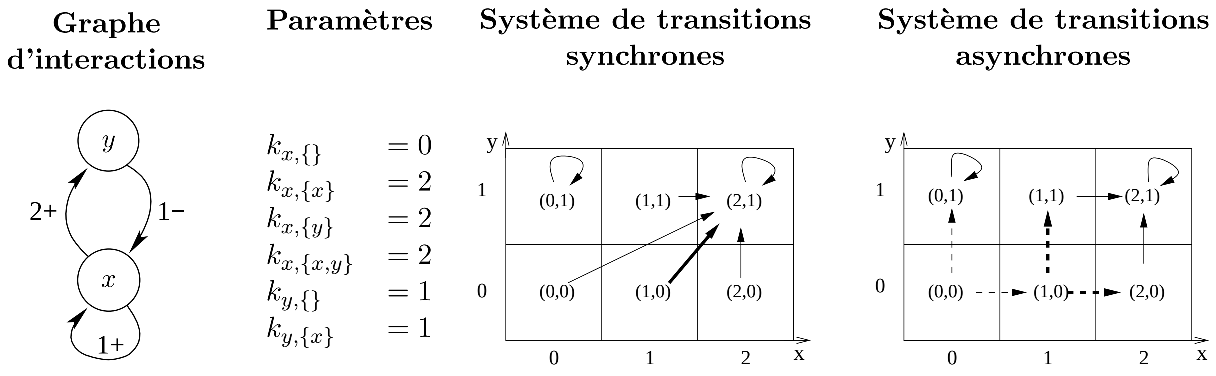


FIG. 3.4 – *Dynamiques déduites des valeurs des paramètres.*

concentrations de manière graduelle. Pour définir formellement la dynamique asynchrone, nous introduisons d'abord la notion de désynchronisation d'une transition synchrone.

Définition 5 *Soit τ une transition du système de transitions synchrones $(n_{v_1}, \cdots, n_{v_p}) \rightarrow (n'_{v_1}, \cdots, n'_{v_p})$. Une* désynchronisation *de τ est une transition de la forme*
$(n_{v_1}, \cdots, n_{v_{i-1}}, n_{v_i}, n_{v_{i+1}}, \cdots, n_{v_p}) \rightarrow (n_{v_1}, \cdots, n_{v_{i-1}}, n_{v_i} + \delta, n_{v_{i+1}}, \cdots, n_{v_p})$ où :
- *i est tel que $n_{v_i} \neq n'_{v_i}$*
- *$\delta = 1$ si $n_{v_i} < n'_{v_i}$, sinon $\delta = -1$.*

La dynamique asynchrone est alors construite à partir du graphe de transitions synchrones en remplaçant chacune des transitions synchrones par *l'ensemble de ses désynchronisations* (cf. figure 3.4). En introduisant un opérateur d'évolution $\blacktriangleright$ (pour x et k deux entiers quelconques, $x \blacktriangleright k$ est égal à $x+1$ si $x < k$, à $x-1$ si $x > k$, et à x si $x = k$), le système de transitions asynchrones peut être défini directement.

Définition 6 *Soit $\mathcal{R} = ((V,E),\mathcal{K})$ un réseau de régulation génétique contenant p gènes. Son* système de transitions asynchrones *(STA) est défini par :*
- *l'ensemble des sommets est l'ensemble des états $\Pi_{i \in V}[0, \beta_i]$*
- *il existe une transition de l'état $n = (n_1, ..., n_p)$ à $m = (m_1, ..., m_p)$ si*

$$\left\{ \begin{array}{l} \exists \textit{ unique } i \in [1,p] \textit{ tel que } m_i \neq n_i \\ m_i = (n_i \blacktriangleright k_{i,R_i(n)}) \end{array} \right. \quad \textit{ou} \quad \left\{ \begin{array}{l} m = n \\ \forall i \in [1,p], \;\; n_i = (n_i \blacktriangleright k_{i,R_i(n)}) \end{array} \right.$$

où $R_i(n)$ est l'ensemble des ressources de i à l'état n.

Lorsque plusieurs transitions partent d'un même état, elles sont concurrentes et chacune d'entre elles peut être choisie. Les points focaux du système de transitions synchrones restent des points focaux du système de transitions asynchrones, mais les chemins partant de l'état courant en direction de son point focal diffèrent. Dans le cas synchrone, une seule transition existe : elle atteint directement le point focal. Dans le cas asynchrone, plusieurs transitions peuvent coexister, elles permettent d'atteindre seulement des états voisins qui eux-mêmes ont leurs propres points focaux. Le comportement du modèle asynchrone à partir d'un état initial diffère ainsi beaucoup de celui du modèle synchrone : alors que le point focal dans le modèle synchrone est toujours atteignable, dans le modèle asynchrone, il ne l'est pas toujours.

3.2.4 Cohérence de la dynamique

E.H. Snoussi [98] a montré que l'approche multi-valuée de R. Thomas était construite comme une discrétisation de systèmes d'équations différentielles linéaires par morceaux où les domaines linéaires sont définis par les domaines où les ressources sont constantes. Ces domaines sont des pavés parallèles aux axes dans l'espace des phases et correspondent aux états abstraits introduits précédemment. M. Kaufman, J. Urbain et R. Thomas ont ensuite confronté cette discrétisation aux modèles différentiels [60, 59]. En prenant en compte les *états singuliers*, *i.e.*, les états sur seuils, R. Thomas et E.H. Snoussi ont montré que tous les états stationnaires pouvaient être trouvés *via* l'approche discrète [100]. Plus récemment, R. Thomas et M. Kaufman ont montré que la modélisation discrète proposait une description qui était cohérente avec les équations différentielles avec un nombre de combinaisons de valeurs pour les paramètres beaucoup plus faible [109, 110]. Comme les comportements observables ont pendant longtemps été réduits à des traces cycliques ou à des traces atteignant des états stables, la modélisation de R. Thomas rejoint l'objectif du rasoir d'Occam. Ainsi même si l'approche discrète n'est pas nécessaire pour étudier les réseaux de régulation, elle est souvent commode au moins parce que les données biologiques ne sont pas toujours quantitatives.

3.3 Formulation en termes de réseaux de Petri de haut niveau

La section précédente ayant permis de définir formellement la sémantique des réseaux de régulation génétique, il est assez naturel, dans une première approche, de reformuler cette sémantique avec des théories classiques de l'informatique afin de tenter de bénéficier des plate-formes logicielles qui les accompagnent.

La modélisation de R. Thomas « compilée » vers les réseaux de Petri. La gestion des seuils dans la théorie de R. Thomas fait penser aux réseaux de Petri : les régulations s'activent à partir de seuils, comme les transitions d'un réseau de Petri. Notons néanmoins que l'existence des régulations négatives, la gestion de la combinatoire des ressources et la non consommation des ressources l'en éloignent.

Un moyen naturel de définir la dynamique asynchrone de R. Thomas en termes de réseau de Petri consiste à introduire des arcs inhibiteurs. Une place est associée à chaque gène, et une transition à chaque paramètre $k_{v_i,\omega}$. Les places d'entrée de la transition correspondant à $k_{i,\omega}$ sont tous les prédécesseurs de i : les prédécesseurs de i appartenant à ω sont connectés à la transition par des arcs standards, et les prédécesseurs de i n'appartenant pas à ω le sont par des arcs inhibiteurs. Malheureusement, une telle modélisation fait apparaître autant de transitions que de paramètres, et mène à un réseau de Petri qui est difficile à exploiter. De plus, en général, l'utilisation des arcs inhibiteurs change la classe des réseaux de Petri, la décidabilité des propriétés courantes et la complexité de preuve de certaines propriétés.

C. Chaouiya et ses co-auteurs ont fait un autre choix [19]. Ils transcrivent toute la dynamique asynchrone de R. Thomas en termes de réseaux de Petri classique. Cette approche multiplie les places et les transitions et il n'y a pas d'interprétation immédiate des places en termes de variables. L'avantage est de rester dans les réseaux de Petri classiques et de ne rien perdre du point de vue de la décidabilité.

La compilation de la théorie de R. Thomas vers les réseaux de Petri ne pose ainsi pas de difficulté. On peut donc l'implémenter pour tester si cette formulation combinée avec les plateformes logicielles qui accompagnent les réseaux de Petri, est à même de répondre de manière pertinente à des questions biologiques. La difficulté réside dans le lien qui doit être fait entre réseaux de Petri et propriétés biologiques :

– La formulation des propriétés utiles en biologie en termes de propriétés sur le réseau de Petri dépend largement du type de réseaux de Petri. Par exemple, le choix des réseaux de Petri classiques a mené à déplier complètement le réseau et à multiplier les places. Les propriétés doivent donc être d'abord transcrites en termes de jetons dans ces places.
– Les questions habituellement posées aux réseaux de Petri comme la vivacité des transitions ne sont pas directement interprétables ou pertinentes du point de vue biologique.

Choix des réseaux de Petri de haut niveau. Nous avons pour notre part choisi, parmi toutes les variantes de réseaux de Petri, les « réseaux de Petri de haut niveau » ou « High Level Petri Nets »(HLPN), dans lesquels on peut faire le choix d'une représentation du réseau de régulation directement interprétable dans le domaine d'application. La représentation en est compacte, et le formalisme intuitif : les jetons représentent les niveaux d'expression de chacun des gènes et leur évolution dépend des valeurs d'un certain nombre de paramètres. La transcription de la modélisation de R. Thomas en réseaux de Petri de haut niveau nous permet aussi de profiter d'un large ensemble de développements de ce domaine de recherche, en particulier de certains outils de vérification de propriétés temporelles.

3.3.1 Les réseaux de Petri de haut niveau (HLPN)

Syntaxe. On se donne un ensemble de variables Var ainsi qu'une signature $\Sigma = (F, P)$ où F est un ensemble fini de symboles fonctionnels munis de leur arité et P un ensemble fini de symboles de prédicats munis de leur arité.

L'ensemble des termes $T_{\Sigma(Var)}$ sur cette signature est construit inductivement : $Var \subset T_{\Sigma(Var)}$, si f^n est un symbole fonctionnel d'arité n et si $t_1, t_2, \ldots t_n$ sont des termes, alors $f(t_1, t_2, \ldots t_n)$ est un terme. L'ensemble des atomes est défini à partir des prédicats P et des termes : si p^n est un prédicat d'arité n et si $t_1, t_2, \ldots t_n$ sont des termes, alors $p(t_1, t_2, \ldots t_n)$ est un atome. L'ensemble des formules $For_{\Sigma(Var)}$ sur cette signature est construit sur la signature Σ et sur les variables de Var avec les connecteurs et quantificateurs habituels de la logique du premier ordre. Un Σ-modèle A est un ensemble muni de fonctions et relations :
– le Σ-modèle A associe pour tout symbole fonctionnel $f^n \in F$ une application $f_A : A^n \to A$
– et pour tout symbole de prédicat $p^n \in P$, une relation p_A sur l'ensemble A^n.

Pour toute substitution $\sigma : Var \to A$, on dit que le modèle satisfait l'atome $p(t_1, \ldots, t_n)$ pour la substitution σ, et on note $A \vDash_\sigma p(t_1, \ldots, t_n)$, si $(\sigma(t_1), \ldots, \sigma(t_n)) \in p_A$. On étend de manière naturelle cette relation de satisfaction à $A \vDash_\sigma \varphi$ pour $\varphi \in For_{\Sigma(Var)}$.

Enfin, nous notons $\mathrm{M}_f(E)$ l'ensemble des multi-ensembles finis sur E (un *multi-ensemble* sur un ensemble E est une fonction $\mu : E \to \mathbb{N}$; μ est fini si son support $\{e \in E \mid \mu(e) > 0\}$ est fini).

Définition 7 *Un réseau de Petri de haut niveau, ou high-level Petri net (HLPN), est un 9-uplet* $(S, T, Var, \Sigma, A, \iota_1, \iota_2, \iota_3, \iota_4)$ *où*
– *S et T sont des ensembles disjoints de* places *et de* transitions *;*
– *Var un ensemble de variables ;*
– *Σ est une signature sur laquelle on peut construire l'ensemble des formules $For_{\Sigma(Var)}$;*
– *A est un Σ-modèle ;*
– *ι_1 est une fonction de S dans l'ensemble $\mathcal{P}^f(A)$ des parties finies de A ; $\iota_1(s)$ définit le* type *de s,* i.e., *l'ensemble des valeurs possibles que la place s peut accepter ;*
– *ι_2 est une fonction de T à valeur dans $For_{\Sigma(Var)}$, $\iota_2(t)$ est appelée la* garde *de t ;*
– *ι_3 (resp. ι_4) est une fonction de $(S \times T)$ (resp. $(T \times S)$) à valeur dans $\mathrm{M}_f(A \cup Var)$ qui définit un multi-ensemble de variables ou valeurs. Les arcs (s, t) ou (t, s) avec une valeur de $\iota_3(s, t)$ ou $\iota_4(t, s)$ de support vide sont omis.*

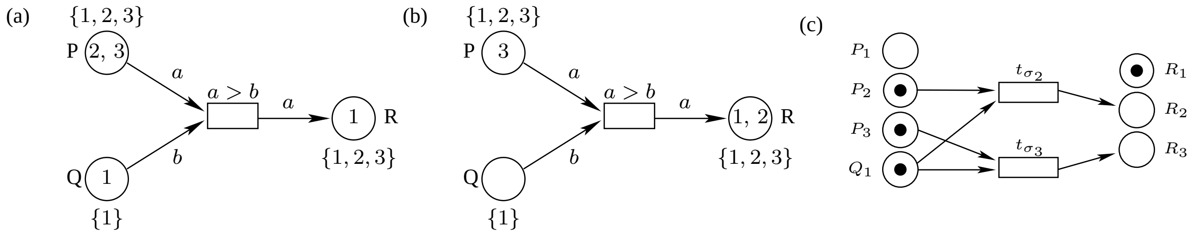


FIG. 3.5 – *Un réseau de Petri de haut niveau simple avant (a) et après (b) le tirage de la transition. (c) Réseau de Petri déplié.*

Par abus de notation, nous introduisons une fonction dite inscription *qui dénote selon le contexte l'une des quatre fonctions* ι_1,ι_2,ι_3,ι_4. *Un HLPN est alors défini par un sextuplet* $(S, T, Var, \Sigma, A, \iota)$.

Sémantique. Un *marquage* d'un réseau de Petri de haut niveau $(S, T, Var, \Sigma, A, \iota)$ est une application $M : S \rightarrow \mathrm{M}_f(A)$ qui associe à chaque place $s \in S$ un multi-ensemble de support inclus dans $\iota(s)$. Par analogie avec les réseaux de Petri classique, on dit que, pour toute valeur a de $\iota(s)$ et pour tout marquage M, il y a $M(s)(a)$ jetons de valeur a dans la place s.

Une *instantiation* (*binding* en anglais) est une application $\sigma : Var \rightarrow A$ qui préserve les types, *i.e.* telle que pour chaque arc (s, t) (resp. (t, s)) et chaque variable $a \in Var$ telle que $\iota_3(s, t)(a) > 0$ (resp. $\iota_4(t, s)(a) > 0$), on a $\sigma(a) \in \iota_1(s)$. L'*évaluation* d'un terme t (resp. d'une formule f) à travers σ, notée $\sigma(t)$ (resp. $\sigma(f)$), est définie de manière inductive pour tout terme de $T_{\Sigma(Var)}$ (resp. pour toute formule de $For_{\Sigma(Var)}$). On étend cette *évaluation* à tout vecteur ou (multi-)ensemble de variables, etc.
La règle de transition spécifie les circonstances dans lesquelles un marquage M' est atteignable à partir d'un marquage M.

- Une instantiation σ est dite *acceptable* (*enabling binding*) pour la transition t si l'évaluation de la garde $\sigma(\iota(t))$ est vraie, autrement dit si $A \vDash_\sigma \iota(t)$. Les instantiations acceptables dépendent donc du Σ-modèle choisi.
- Une transition t est *active* pour le marquage M s'il y a une *instantiation acceptable* σ telle que $\forall s \in S : \sigma(\iota(s, t)) \leq^2 M(s)$, *i.e.*, telle qu'il y ait assez de jetons de chaque valeur dans les places d'entrée pour satisfaire la garde de la transition.
- L'effet du tirage de la transition t, sous une instantiation acceptable σ, est d'enlever des jetons des places d'entrée et d'ajouter des jetons dans ses places de sortie, en fonction de l'évaluation des annotations des arcs à travers σ.

La signature à définir (les symboles fonctionnels et de prédicats) ainsi que le modèle sémantique A dépendent du domaine d'application. Nous verrons que plusieurs choix pour la signature et sa sémantique sont possibles.

Pour l'exemple de la figure 3.5-(a), l'ensemble des places est $\{P, Q, R\}$, celui des transitions est réduit à une seule transition ; l'ensemble *Var* contient les variables a et b. La signature choisie est $\Sigma = (F, P)$ où F est l'ensemble vide et où l'ensemble des symboles de prédicats P contient le prédicat $<$ d'arité 2. La sémantique des formules de $For_{\Sigma(Var)}$, est donnée par le choix du Σ-modèle A : l'ensemble des entiers naturels muni de la relation $<$ interprétée de manière usuelle sur les entiers naturels.
Le marquage présenté dans la figure 3.5-(a) est $M(P) = \{2, 3\}$, $M(Q) = \{1\}$, $M(R) = \{1\}$. Les instantiations possibles sont $\sigma_1 = \{a \rightarrow 1, b \rightarrow 1\}$, $\sigma_2 = \{a \rightarrow 2, b \rightarrow 1\}$, et $\sigma_3 = \{a \rightarrow 3, b \rightarrow 1\}$.

[2]Pour deux multi-ensembles μ_1 et μ_2 sur le même ensemble E, on dit que $\mu_1 \leq \mu_2$ si pour tout élément e du support de μ_1 on a $\mu_1(e) \leq \mu_2(e)$.

Seules σ_2 et σ_3 sont acceptables (σ_1 ne rend pas la garde vraie avec la sémantique choisie). Pour le marquage M, la transition t est active pour σ_2 et pour σ_3. La figure 3.5-(b) montre le nouveau marquage si σ_2 est choisie.

Dépliage. Une propriété importante des réseaux de Petri de haut niveau est qu'ils peuvent être dépliés en réseau de Petri classiques. L'opération de dépliage associe un réseau de Petri $N_d = (S_d, T_d, W)$ à chaque réseau de Petri de haut niveau $N = (S, T, Var, \Sigma, A, \iota)$, ainsi qu'un marquage M_d de N_d à chaque marquage M de N :

- $S_d = \{s_v \mid s \in S \text{ et } v \in \iota(s)\}$, il y a une place pour chaque valeur de chaque place de N ;
- $T_d = \{t_\sigma \mid t \in T \text{ et } \sigma \text{ est une instantiation acceptable pour } t\}$, il y a une transition pour chaque instantiation acceptable de chaque transition de N ;
- le poids $W(s_v, t_\sigma)$ de chaque arc (s_v, t_σ) est donné par : $W(s_v, t_\sigma) = \sum_{\substack{a \in \text{support}(\iota(s,t)) \\ \sigma(a) = v}} \iota(s,t)(a)$

  autrement dit, le poids de l'arc (s_v, t_σ) est le nombre total de jetons de valeurs v que le tirage de la transition t enlève à la place s sous l'instantiation acceptable σ ;
- le poids $W(t_\sigma, s_v)$ de chaque arc (t_σ, s_v) est donné de manière analogue ;
- le marquage $M_d(s_v)$ est déterminé par le marquage $M(s)$: chaque place élémentaire $s_v \in S_d$ contient autant de jetons que d'occurrences de v dans le marquage $M(s)$.

Notons que la définition des HLPN impose que $\iota(s)$ soit fini pour toute place s du réseau et que $\iota(s,t)$ (resp. $\iota(t,s)$) soit un multi-ensemble fini pour tout (s,t). Ceci assure d'une part que le réseau déplié ait un nombre fini de places et transitions et d'autre part que les poids du réseau déplié soient finis. La figure 3.5-(c) présente le réseau de Petri obtenu par dépliage du réseau de Petri de haut niveau de la figure 3.5-(a).

3.3.2 Modélisation des réseaux de régulation génétique

Nous pouvons représenter un réseau de régulation génétique par un réseau de Petri de haut niveau ne contenant qu'une seule transition et autant de places qu'il y a de gènes dans le réseau d'interactions. Chaque place correspond à un gène et contient un et un seul jeton de type entier : son niveau de concentration abstrait. Pour exprimer la sémantique de R. Thomas, on doit pouvoir comparer un niveau d'expression à une coordonnée d'un point focal. On introduit donc dans la signature les différentes fonctions $k_i()$ d'arité égale au nombre de gènes, qui renvoient pour chaque état la coordonnée i du point focal correspondant. Les prédicats de comparaisons nécessaires sont représentés par les symboles $=$ et $\neq$ (avec leur sémantiques habituelles sur les entiers). L'évolution des concentrations se fait par pas de un, il faut donc aussi introduire le symbole fonctionnel $\blacktriangleright$ d'arité 2 (cf. définition de sa sémantique, page 41). Le réseau de Petri de haut niveau est donc défini ainsi :

- l'ensemble des places $S = \{s_1, \dots, s_n\}$ contient autant de places qu'il y a de gènes,
- l'ensemble de transitions T est réduit à une transition t,
- l'ensemble de variables contient deux fois plus de variables qu'il y a de gènes $Var = \{v_1, v_1', v_2, v_2', \dots, v_n, v_n'\}$,
- la signature est définie par son ensemble de symboles fonctionnels contenant les symboles k_i et $\blacktriangleright$, et par celui des prédicats qui contient les symboles $=$ et $\neq$;
- A est l'ensemble des entiers naturels muni des prédicats dont la sémantique a été définie plus haut,
- l'inscription d'une place est l'ensemble de valeurs entières que peut prendre le gène associé,

– l'inscription de l'arc $s_i \to t$ est le multi-ensemble de support réduit à v_i ($\iota(s_i,t)(v_i) = 1$), l'inscription de l'arc $t \to s_i$ est le multi-ensemble de support réduit à v'_i ($\iota(s_i,t)(v'_i) = 1$),
– la garde de l'unique transition est la traduction dans le langage $For_{\Sigma(Var)}$ de l'expression :

$$\text{asyn\_guard}(v_1, v'_1, v_2, v'_2, \ldots, v_n, v'_n) = \left\{ \begin{array}{c} \Big(\exists i \in [1,n],\ \ (v_i \neq k_i(v_1,\ldots,v_n)) \wedge (v'_i = v_i \blacktriangleright k_i(v_1,\ldots,v_n)) \wedge (\forall j \neq i, v'_j = v_j)\Big) \\ \bigvee \\ \Big(\forall i \in [1,n],\ \ (v_i = k_i(v_1,\ldots,v_n)) \wedge (v'_i = v_i)\Big) \end{array} \right\}$$

Cette garde traduit directement la sémantique asynchrone de R. Thomas. En effet, le marquage représente un état stationnaire stable pour la sémantique asynchrone lorsque chacune des coordonnées du point focal $k_i(v_1, \ldots, v_n)$ est égale à la concentration v_i. La garde donne dans ce cas $v_i = v'_i$ pour tout i et la stabilité du système est représentée dans la dynamique par une boucle[3]. Si le marquage ne correspond pas à un état stationnaire stable, alors l'un des gènes a un niveau de concentration différent de la coordonnée du point focal lui correspondant. Un marquage possible pour l'état suivant est celui dans lequel une seule variable a augmenté ou diminué d'une unité en direction du point focal. Le schéma de gauche de la figure 3.6 représente le réseau de Petri de haut niveau pour le graphe d'interactions de la figure 3.4.

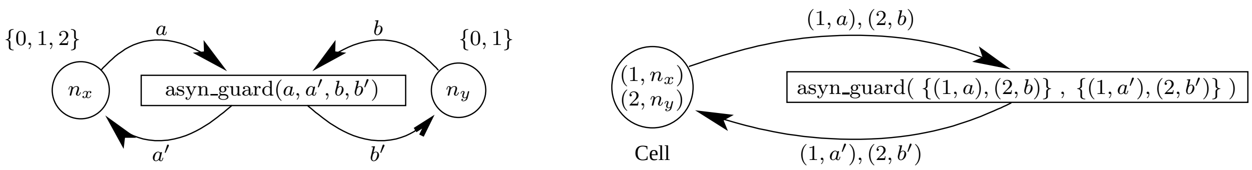


FIG. 3.6 – *Deux formulations possibles d'un réseau de régulation génétique avec deux gènes. A gauche, chaque place abstrait un gène et contient un jeton : son niveau de concentration. A droite, la seule place contient deux jetons représentant les niveaux de concentration des deux gènes. La garde exprime la relation entre le marquage courant et un possible marquage suivant.*

Une autre modélisation plus compacte est possible, elle s'exprime comme un repliage de la représentation précédente. Dans cette représentation, il n'y a qu'une seule place qui abstrait la cellule biologique, dans laquelle chaque jeton représente simultanément un gène spécifique et son niveau d'expression. L'inscription de la place contient donc les couples $(g_i, level_i)$ où $g_i \in \{1, \ldots, n\}$ est le numéro d'un gène et où $level_i$ est un niveau de concentration valide pour le gène i. Les inscriptions des arcs sont aussi modifiées : l'inscription en entrée de la transition devient $\{(1, a_1) \to 1, \ldots, (n, a_n) \to 1\}$, et celle en sortie devient $\{(1, a'_1) \to 1, \ldots, (n, a'_n) \to 1\}$, voir schéma de droite de la figure 3.6. L'état du système est maintenant représenté par l'ensemble des jetons présents dans l'unique place.

Ainsi, pour tout réseau de régulation, le réseau de Petri de haut niveau a une unique place et une unique transition. Les seuls éléments qui distinguent les différents réseaux de régulation sont le nombre et le type des jetons ainsi que les fonctions k_i qui sont présents dans la garde.

3.4 Compléments et extensions

Nous proposons ici quelques reflexions sur la modélisation des réseaux de régulation génétique de R. Thomas. Les signes des interactions ont été introduits pour être en accord avec la descrip-

[3]Une boucle (circuit de longueur 1) est différente d'un *deadlock*, particulièrement lorsqu'on cherche à vérifier des propriétés temporelles.

tion continue, mais il est possible d'enrichir la signification de ces signes afin de contraindre la dynamique sous-jacente. Les délais d'activation / inhibition peuvent être utiles pour chercher à quantifier les durées des trajectoires. Enfin, on peut envisager de détailler les domaines d'activité des ressources afin de raffiner le modèle.

3.4.1 Sémantiques des signes

Les signes associés aux interactions ne servent qu'au calcul des ressources de chacune des variables. Une fois les ressources calculées, les valeurs des paramètres peuvent entraîner des comportements contre-intuitifs. Par exemple, l'ajout d'une ressource pour x peut entraîner la diminution de la composante x du point focal.

Les travaux de E.H. Snoussi et R. Thomas [100] permettent de préciser quelles sont les relations entre les valeurs des points focaux en fonction des signes des interactions mises en jeu. E.H. Snoussi montre en effet comment la modélisation de R. Thomas peut être vue comme une discrétisation d'une classe particulière de systèmes d'équations différentielles linéaires par morceaux. Dans ce cas, les valeurs des paramètres $k_{i,\omega}$ reflètent une discrétisation de sommes de ratios de constantes positives. Il s'en suit que les valeurs des paramètres de l'ensemble $\{k_{i,\omega} \mid i \in V \text{ et } \omega \subseteq G^{-1}(i)\}$ doivent satisfaire les contraintes suivantes :

$$k_{i,\emptyset} = 0 \qquad \text{et} \qquad \omega \subset \omega' \Longrightarrow k_{i,\omega} \leq k_{i,\omega'}.$$

Cependant, il est possible d'élargir l'ensemble des modèles qui peuvent être décrits avec ce formalisme discret, ce qui mène à relâcher les contraintes précédentes. Le tableau 3.1 donne plusieurs sémantiques possibles pour le signe d'une interaction de y sur x.

Sémantique	Notation	Définition
Activation observée	$+_{obs}$	$\exists \mathcal{R}$ tel que $K_{x,\mathcal{R}} < K_{x,\mathcal{R}\cup\{y\}}$
Inhibition observée	$-_{obs}$	$\exists \mathcal{R}$ tel que $K_{x,\mathcal{R}} > K_{x,\mathcal{R}\cup\{y\}}$
Activation avec modèle de Snoussi	$+_{Snoussi}$	$\forall \mathcal{R}\ \ K_{x,\mathcal{R}} \leq K_{x,\mathcal{R}\cup\{y\}}$
Inhibition avec modèle de Snoussi	$-_{Snoussi}$	$\forall \mathcal{R}\ \ K_{x,\mathcal{R}} \geq K_{x,\mathcal{R}\cup\{y\}}$
Activation observée avec modèle de Snoussi	$++$	$+_{obs} \wedge +_{Snoussi}$
Inhibition observée avec modèle de Snoussi	$--$	$-_{obs} \wedge -_{Snoussi}$
Régulation observée	$\pm$	$+_{obs} \vee -_{obs}$
Inhibition et activation	$\pm_{obs}$	$+_{obs} \wedge -_{obs}$

TAB. 3.1 – *Sémantiques possibles pour le signe d'une interaction $y \rightarrow x$.*

Ces différentes sémantiques montrent qu'un grand nombre d'informations sur les régulations peuvent être traduites en contraintes sur les paramètres. Ainsi, la définition formelle des signes permet de contraindre les paramètres.

3.4.2 Délais d'activation / d'inhibition

Les systèmes de transitions de la théorie de R. Thomas ne sont pas déterministes (l'opération de désynchronisation introduit plusieurs futurs possibles à partir de certains états). Ils abstraient toutes les trajectoires continues du système d'équations différentielles linéaires par morceaux sous-jacent, mais, comme l'ont montré H. de Jong *et al.* [28], ils introduisent aussi des traces qui ne correspondent pas à des trajectoires continues. En effet une transition du système de transitions de R. Thomas de a vers b signifie qu'il existe une trace continue partant d'un état du pavé associé à a vers un état de la frontière qui sépare les pavés associés à a et à b. Si maintenant on considère une transition de b vers c, on obtient dans le système de transitions de R. Thomas la trace $a \to b \to c$. Dans l'espace continu, il n'est cependant pas assuré qu'il existe une trace provenant de a, traversant le pavé b pour atteindre le pavé c : il se peut que les traces continues dans le pavé b provenant de a aillent toutes vers un état autre que c, voir figure 3.7.

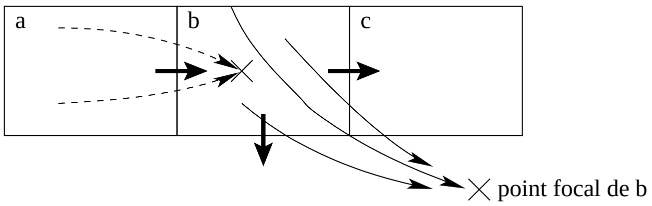


FIG. 3.7 – *La trace abstraite $a \to b \to c$ est incompatible avec le modèle continu.*

Cet indéterminisme « abusif » est dû à une grossière abstraction discrète des variations de concentration par pavé et à la non représentation du temps. Pour contrôler ces variables, M. Adélaïde et G. Sutre [1] ont montré que sous certaines conditions d'égalité des constantes de dégradation, il est facile de raffiner cette abstraction pour proposer une dynamique qui ne présente pas les mêmes inconvénients.

Nous pouvons choisir une autre voie, plus générale, qui consiste à intégrer dans la modélisation les délais d'activation et d'inhibition en s'inpirant des travaux de M. Kaufman et R. Thomas [110]. Ces délais correspondent aux temps que le système prend pour passer d'un état à un autre. Ainsi, si à partir de l'état a, il est possible d'atteindre à la fois b et c, une connaissance des délais de passage peut permettre de mieux « déterminiser » le système. Nous avons proposé, avec Olivier Roux de l'IRCCyN [2], la formalisation de ces délais ainsi qu'une méthodologie pour déduire des contraintes sur ces délais à partir de l'existence d'un chemin dans la dynamique.

3.4.3 Réseaux de régulation génétique étendus

En suivant les mêmes idées directrices, nous avons proposé avec G. Bernot et F. Delaplace du laboratoire IBISC, O. Roux et F. Cassez de l'IRCCyN, une autre modélisation des réseaux de régulation permettant plus de souplesse pour décrire les domaines d'activité des interactions [16]. L'idée initiale repose sur la séparation des ressources qui correspondent à la présence d'un activateur de celles qui correspondent à l'absence d'un inhibiteur. Les règles de définition des domaines d'activités d'un gène sur un autre ont été assouplies : un gène peut être un activateur d'une cible à un certain niveau et être un inhibiteur de cette même cible à un autre niveau.

Ainsi, nous définissons un *Réseau de régulation génétique étendu* (RRGE) comme un triplet $\mathcal{R} = (V, E, \pi)$ où (V, E) est un graphe orienté, et où $\pi : E \to 2^{\mathbb{N}} \times 2^{\mathbb{N}}$ spécifie les domaines d'activation et d'inhibition de chacun des gènes sur leurs cibles.

La figure 3.8 donne un exemple de réseaux de régulation : $V = \{x, y, z\}$ et $\pi(x, y) = ([1, 1], [])$: x active y lorsque x a un niveau égal à 1 ; $\pi(y, x) = ([], [1, 1])$: y inhibe x lorsque y a un niveau égal à 1 ; $\pi(y, z) = ([1, 1], [])$: y active z lorsque y a un niveau égal à 1.

Comme dans la théorie de R. Thomas, le comportement d'un réseau de régulation étendu

$$\begin{tikzcd} x \arrow[r, bend left, "{([1,1],[])}"] & y \arrow[l, bend left, "{([],[1,1])}"] \arrow[r, "{([1,1],[])}"] & z \end{tikzcd}$$

FIG. 3.8 – *Exemple de RRGE*

dépend de certains paramètres représentant les directions d'évolution de chacun des niveaux de concentration en fonction des activateurs/inhibiteurs actifs dans l'état courant du système. L'ensemble $Para(\mathcal{R})$ des *paramètres* de $\mathcal{R}$ est défini par

$$Para(\mathcal{R}) = \{K_{x,A,B} | A \subseteq R_+(x), B \subseteq R_-(x)\}$$

où $R_+(x)$ (resp. $R_-(x)$) est l'ensemble des activateurs (resp. inhibiteurs) potentiels. L'évolution d'un réseau dépend alors de la concentration des gènes et de l'ensemble des paramètres. Si, pour un gène donné, X_+ et X_- sont respectivement l'ensemble des gènes qui l'activent et qui l'inhibent, alors le niveau de concentration de x évolue vers la valeur définie par le paramètre K_{x,X_+,X_-}.

Ce formalisme permet d'utiliser une approche discrète dans le cas où les activateurs et inhibiteurs sont traités séparément et où les interactions ont des domaines d'activité qui ne sont pas des demi-droites. La description de ces domaines d'activité nécessite l'introduction de nouveaux paramètres (les paramètres sont indicés par un sous ensemble d'activateurs et un sous ensemble d'inhibiteurs) ce qui permet de décrire un ensemble de dynamiques plus large.

3.5 Introduction explicite des seuils

La théorie de R. Thomas ne s'arrète pas à la modélisation discrète du système d'équations différentielles linéaires par morceaux. Elle décrit aussi partiellement les phénomènes qui ont lieu au niveau des seuils appelés états singuliers. René Thomas a montré par exemple que certains états stationnaires du système étaient sur des seuils et que l'approche qu'il a développée permettait de les capturer tous. Dans un travail avec Adrien Richard [86] que nous présentons brièvement ici, nous avons introduit explicitement les seuils dans la discrétisation des concentrations. Ce travail nous a surtout permis de bien comprendre les conditions de stationnarité des états singuliers en termes de contraintes sur les paramètres.

3.5.1 Les intuitions

Les états singuliers peuvent être caractérisés par leur ordre ou de manière équivalente par la dimension du domaine continu qu'ils abstraient. Pour un réseau avec n gènes, un état singulier qui n'a qu'une seule composante sur seuil, est un état singulier d'ordre 1 (sa dimension est $n-1$), et de manière générale, un état singulier qui a k composantes sur seuils est d'ordre k.

Pour définir les transitions entre états réguliers, *i.e.* qui ne sont pas sur seuils, et états singuliers, plusieurs cas sont à considérer. Le système possède une transition d'un état régulier vers un état singulier d'ordre 1 si l'état singulier sépare l'état régulier de son point focal [86]. Si pour un état singulier d'ordre 1, les points focaux des deux états réguliers adjacents sont du même coté de l'état singulier, alors il y a une transition de l'état singulier vers l'état régulier adéquat. Si par contre pour ce même état singulier, les points focaux sont de part et d'autre de l'état singulier, alors l'état singulier est considéré stationnaire (au moins dans la direction considérée). La trajectoire peut éventuellement glisser le long de l'état singulier d'ordre 1 pour atteindre suivant des règles analogues, un état singulier d'ordre 2. Ainsi dans le modèle présenté, les transitions ne sont possibles qu'entre domaines dont les ordres diffèrent de 1.

Le cas où un état singulier d'ordre 1 est entre les deux points focaux de ses voisins réguliers, peut provenir de deux situations. Dans la première, chacun des domaines réguliers adjacents a son point focal de l'autre coté de l'état singulier et donc entraîne les trajectoires en direction de cet état singulier. Dans la deuxième, chacun des points focaux de ces domaines entraîne les trajectoires loin de l'état singulier, cf. figure 3.9. Il faut donc faire la distinction entre plusieurs types d'états singuliers. Un état singulier sera appelé *mur transparent*, si les points focaux des

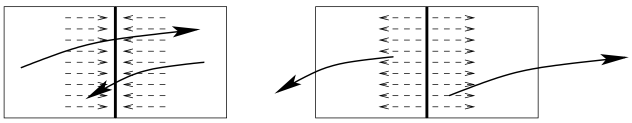


FIG. 3.9 – *Etats singuliers stationnaires stables et instables.*

états adjacents sont du même coté du mur. Il sera appelé *mur noir*, si le point focal de chacun des deux domaines adjacents est de l'autre coté du mur. Enfin il sera appelé *mur blanc*, si le point focal de chacun des deux domaines adjacents est du même coté du mur que le domaine associé, voir figure 3.10.

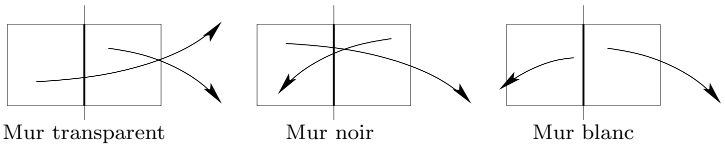


FIG. 3.10 – *Les différents types de murs.*

3.5.2 Modélisation de H. de Jong

H. de Jong et ses collaborateurs ont proposé une analyse qualitative des réseaux de régulation génétique incorporant les états singuliers [29, 28]. Pour cela, ils se sont basés sur la résolution d'équations différentielles discontinues par morceaux proposée par Filippov [37].

- Pour qu'une transition existe entre un domaine a d'ordre k et un autre domaine b d'ordre $k' > k$, il faut que b soit un domaine de la frontière de a, et qu'il existe une solution du système d'inclusions différentielles au sens de Filippov partant du domaine a et atteignant le domaine b.
- Pour qu'une transition existe entre un domaine a d'ordre k et un autre domaine b d'ordre $k' < k$, il faut qu'il existe une solution du système d'inclusions différentielles au sens de Filippov dont seul le point initial appartient au domaine a et tous les autres points appartiennent au domaine b.

Cette modélisation rend compte des trajectoires sur lesquelles plusieurs seuils sont franchis de manière simultanée, elle permet de distinguer les états singuliers stationnaires stables des états singuliers stationnaires instables, et elle explicite vraiment comment les trajectoires sortent des états singuliers.

Chapitre 4

Méthodologie pour une modélisation itérative

Résumé

La sémantique des réseaux de régulation décrite dans le chapitre précédent nous permet d'envisager l'automatisation partielle de la construction de modèles d'un réseau de régulation génétique. Étant données des observations sur l'objet d'étude, il est possible de chercher parmi toutes les sémantiques possibles d'une classe finie de modèles, celles qui sont cohérentes avec les observations. Il s'agit ainsi d'élaguer l'ensemble des modèles possibles en prenant en compte une formalisation des observations biologiques.

Cette idée était déjà présente dans les travaux de E.H. Snoussi et R. Thomas lorsqu'ils se sont intéressés à la stabilité des états caractéristiques des boucles de rétroactions. Ils ont montré qu'une condition nécessaire pour avoir multi-stationnarité (resp. homéostasie) était la stabilité d'un état caractéristique d'une boucle positive (resp. négative). Nous étendons ce filtrage des modèles à toute propriété biologique exprimable dans un langage formel.

Nous nous sommes aussi intéressés à la classification de modèles d'un point de vue apprentissage statistique afin de révéler les caractéristiques qui peuvent les différencier. Cette classification permet d'avoir une vision synthétique de l'ensemble des modèles « survivants ». On peut alors chercher à construire une expérimentation pour distinguer différentes classes de modèles et la réalisation in vivo *de cette expérimentation réduit l'ensemble de modèles à considérer.*

4.1 Modéliser formellement pour assister par ordinateur la découverte d'un modèle

Comme nous l'avons vu dans le chapitre précédent, la théorie de R. Thomas permet de décrire qualitativement la dynamique d'un réseau de régulation génétique. Le schéma d'interactions entre les différents gènes d'intérêt ne donne aucune information sur la dynamique sous-jacente, puisque ce sont les valeurs des paramètres qui pilotent la dynamique. Malheureusement, le nombre de paramètres associés à chacun des nœuds du graphe d'interactions, croît de manière exponentielle avec le nombre de ses prédécesseurs. Et même si les contraintes de signes sont prises en compte, le nombre de paramétrages envisageables reste énorme. De plus, il n'est pas rare que le graphe d'interactions ne spécifie pas (ou partiellement) les seuils d'interactions. Il est alors nécessaire de considérer les différentes possibilités pour ces seuils, ce qui décuple bien évidement le nombre de paramétrages à considérer.

Si le graphe d'interactions peut être construit sur des informations provenant de la biologie moléculaire et s'il peut être considéré comme fiable, il en va tout autrement des seuils d'interactions et des valeurs des différents paramètres. Il n'est déjà pas facile de déterminer à partir de quelle concentration le produit d'un gène agit sur un autre, mais la détermination de la position relative de plusieurs seuils est encore plus difficile. Quant aux valeurs des points focaux, autrement dit des paramètres du modèle discret, ils ne sont que très rarement mesurables expérimentalement *in vivo*, car ils ne sont pas toujours atteignables. La recherche des valeurs de ces paramètres constitue donc un point clef de l'activité de modélisation.

Pour réduire l'ensemble des dynamiques à considérer, des connaissances ou hypothèses sur le comportement peuvent être utilisées comme un critère indirect. Par exemple, l'homéostasie (resp. la multi-stationnarité) est expérimentalement observable et peut être reliée à l'existence d'un circuit négatif (resp. positif) fonctionnel[1], c'est à dire un circuit négatif (resp. positif) entraînant une homéostasie (resp. multi-stationnarité). Grâce à la notion d'état caractéristique, R. Thomas et ses collaborateurs [111] ont montré que la fonctionnalité d'un circuit s'exprimait en termes de contraintes sur les paramètres, ce qui permet de limiter l'ensemble des paramétrages à considérer.

Dans ce chapitre, nous défendons l'idée selon laquelle la réduction de l'espace des paramètrages possibles, qui est fini mais très grand, peut être renforcée et systématisée. L'idée initiale a émergé lors de réunions du groupe de travail du *programme d'épigénomique* sur les liens entre modélisation et observabilité. G. Bernot, J. Guespin, A. Zemerline, P. Amar, F. Delaplace, P. Ballet et moi-même avons remarqué qu'on pouvait faire une analogie entre les spécifications de systèmes informatiques et la modélisation en biologie [18]. Cette rencontre entre biologie et informatique [46] nous a permis, lors de notre étude avec J. Guespin sur la production de mucus chez *Pseudomonas aeruginosa* [17], de montrer que lorsqu'une connaissance ou hypothèse biologique était exprimable sous forme de formule de logique temporelle, elle pouvait aussi être utilisée comme un critère indirect pour éliminer les modèles qui ne la satisfont pas. Pour chacun des paramétrages possibles, on construit la dynamique et on utilise le model-checking pour tester si la formule temporelle est satisfaite ou non. C'est pour cela que les formules exprimant les connaissances ou hypothèses biologiques sont appelées *spécifications* du modèle. Cette approche peut être considérée comme une approche « force brute » puisque toutes les sémantiques d'intérêt doivent être considérées et confrontées par model-checking à la formule temporelle.

Cette procédure d'élagage permet de générer l'ensemble des sémantiques cohérentes avec les spécifications. Si cet ensemble est vide, alors aucun modèle ne valide les spécifications et on doit remettre en question soit le graphe d'interactions soit les spécifications. Si au contraire cet ensemble est très grand, il est nécessaire de fournir des éléments d'analyse de cet ensemble. Nous nous sommes donc intéressés, avec F. d'Alché-Buc du laboratoire IBISC et C. Damon [26], à la classification de modèles d'un point de vue apprentissage statistique afin de révéler les caractéristiques qui peuvent les différencier.

Lorsque cet ensemble de modèles est trop grand pour être exploité à la main, il est nécessaire de proposer au biologiste une expérimentation dont la mise en œuvre permet d'affiner la connaissance sur le modèle. L'expérience à faire consiste à mettre le système dans un état donné, d'attendre un certain temps et d'observer sa réponse (du type oui/non). Ainsi, une expérience peut se traduire dans le monde des modèles par une formule temporelle. Si le système « répond positivement », alors on pourra mettre de coté tous les modèles pour lesquels l'expérience aurait mené à une réponse négative (et réciproquement). La difficulté de cette démarche réside dans le choix de la formule et donc de l'expérience à proposer au biologiste. Tout schéma d'expérience n'est pas forcément compatible avec les contraintes d'expérimentation, mais il serait toutefois utile de proposer une expérience dont le résultat quel qu'il soit apporte un maximum d'informations sur le système. Nous nous sommes alors intéressés, avec S. Vial du laboratoire IBISC et F. Quessette

[1] Nous reviendrons dans le chapitre 5 sur les deux conjectures de R. Thomas qui font un lien entre dynamique et représentation des interactions.

du PRiSM, au problème de l'optimisation du choix des formules pour obtenir au plus vite un maximum d'informations sur le système.

4.2 Élagage de l'espace des possibles

Les exécutions des modèles représentant les réseaux de régulation dépendent non seulement des états initiaux qui sont malheureusement souvent inconnus, mais aussi des paramètres du modèle, dont la valeur n'est, la plupart du temps, ni mesurable ni déductible directement des expériences. Le modélisateur doit alors considérer les différentes valeurs possibles des différents paramètres.

Pour un réseau de régulation génétique, le nombre de paramètres est exponentiel en fonction du nombre de prédécesseurs pour chacun des gènes. Plus précisément, ce nombre est égal à $\prod_{i \in V} 2^{|G^{-1}(i)|}$ où $G^{-1}(i)$ représente l'ensemble des prédécesseurs de i. Cette remarque nous empêche de construire toutes les sémantiques possibles et de laisser le biologiste sélectionner « à la main » celles qui lui paraissent intéressantes. Nous sommes donc contraints de mettre en place une étape d'« élagage de l'espace des possibles » pour limiter le nombre de modèles que l'expert doit étudier.

Ainsi, le processus de modélisation revient à la détermination des modèles à considérer, *i.e.*, dans notre cas, les valeurs des paramètres qui mènent à une sémantique cohérente avec les connaissances expérimentales. Nous souhaitons de plus ne pas nous limiter à la détermination d'un seul paramétrage, puisque ce dernier pourrait être remis en question par une nouvelle connaissance expérimentale. Nous cherchons donc de manière exhaustive tous les modèles cohérents avec les connaissances expérimentales.

4.2.1 Sélection des dynamiques avec circuits de rétroaction fonctionnels

Les circuits de rétroaction jouent un rôle majeur dans la dynamique des réseaux de régulation. Ce sont eux qui permettent d'avoir un comportement périodique et/ou qui permettent d'avoir plusieurs états stables. Dans un circuit de rétroaction, chaque élément exerce un effet sur l'élément suivant, mais aussi par transitivité sur tous les éléments suivants y compris lui-même. Si chaque élément a un effet indirect activateur sur lui-même, le circuit sera dit positif et négatif dans le cas inverse. C'est la parité du nombre d'interactions négatives du circuit qui permet de donner un signe au circuit. Un circuit est *positif* s'il possède un nombre pair d'inhibitions et *négatif* s'il en possède un nombre impair.

Nous allons maintenant définir la fonctionnalité des circuits de rétroaction qui permet de savoir comment un circuit participe à la dynamique du réseau. Un circuit est dit *fonctionnel* s'il engendre l'homéostasie ou la multi-stationnarité. Deux cas sont en effet à considérer.

1. **Influence des circuits négatifs.** Dans un circuit négatif fonctionnel, l'augmentation (resp. diminution) de la concentration d'une variable entraîne l'augmentation (resp. diminution) de son effet négatif sur lui-même ce qui conduit à une diminution (resp. augmentation) de sa concentration. Le circuit fait donc tendre (ou osciller) chacune de ses variables vers une concentration d'équilibre, qui correspond au seuil de l'interaction sur l'élément suivant dans le circuit. Les circuits négatifs sont donc à l'origine de l'oscillation ou de l'amortissement qui correspondent à l'homéostasie en biologie (glycémie, cycle cellulaire...).
2. **Influence des circuits positifs.** Dans un circuit positif fonctionnel, l'augmentation (resp. diminution) de la concentration d'une variable augmente (resp. diminue) son effet positif sur elle même, ce qui conduit à une augmentation (resp. diminution) de sa concentration.

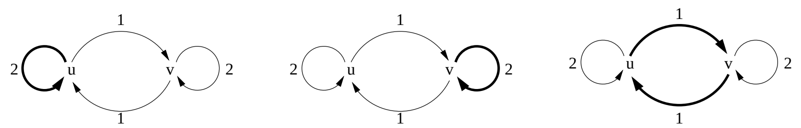
Lorsqu'une variable du circuit est inférieure (resp. supérieure) au seuil de son interaction sur l'élément suivant, elle le restera. Les circuits positifs sont donc à l'origine de la multistationnarité (différentiation et mémoire), composée de deux états stationnaires stables et d'un état stationnaire instable correspondant à la crête entre les deux états stables.

Ces comportements sont intimement liés aux paramètres du système. Pour savoir si un circuit de rétroaction est fonctionnel, il faut regarder l'influence de ses gènes sur eux-mêmes. Comme cette influence s'exprime à l'aide des paramètres, il est possible d'exprimer la fonctionnalité d'un circuit comme des contraintes sur les paramètres. Définissons d'abord les interactions singulières et les états caractéristiques des circuits.

Définition 8 (Interactions singulières) *Soit un réseau de régulation construit sur le graphe d'interaction* $G = (V, E)$ *et soit* $\mathcal{X}$ *l'ensemble des états réguliers et singuliers du système. On définit la fonction* $\mathrm{S}(\mathrm{x})$ *qui donne, lorsque l'état du réseau est* x*, l'ensemble des* interactions singulières*, c'est à dire l'ensemble des interactions* $u \to v$ *dont le régulateur a une concentration égale au seuil de l'interaction :* $\mathrm{S}(\mathrm{x}) = \{u \to v \in E : \mathrm{x}_u = \mathrm{s}_{uv}\}$.
Si $\mathrm{S}(\mathrm{x}) = \emptyset$ *alors* x *est un état régulier, sinon, c'est un état singulier.*

Définition 9 (État caractéristique d'un circuit) *Soit un réseau de régulation construit sur le graphe d'interaction* $G = (V, E)$ *et soit* C *l'ensemble des gènes d'un circuit* $\mathcal{C}$ *du réseau. L'état* x *du réseau est caractéristique du circuit* $\mathcal{C}$ *lorsque* $C = \mathrm{S}(\mathrm{x})$.

Un état caractéristique est donc un état singulier. Pour un circuit impliquant tous les gènes du réseau, l'état caractéristique est unique puisque chacun des gènes doit avoir un niveau de concentration sur un seuil d'interaction bien défini. Lorsqu'un circuit ne contient pas tous les éléments du réseau auquel il appartient, il existe plusieurs états caractéristiques. Voici quelques exemples :



Etats caractéristiques :	États caractéristiques :	États caractéristiques :
$(\mathrm{s}_{uu}, 0)$	$(0, \mathrm{s}_{vv})$	
$(\mathrm{s}_{uu}, 1)$	$(1, \mathrm{s}_{vv})$	$(\mathrm{s}_{uv}, \mathrm{s}_{vu})$
$(\mathrm{s}_{uu}, 2)$	$(2, \mathrm{s}_{vv})$	

L'intéret des états caractéristiques réside dans le théorème suivant qui fait le lien entre fonctionnalité et états caractéristiques : pour qu'un circuit soit fonctionnel, il est nécessaire que l'un de ses états caractéristiques soit stationnaire. De façon intuitive, comme dans un circuit négatif fonctionnel chaque variable d'un élément du circuit tend vers le seuil de l'interaction sur l'élément suivant, on comprend qu'un état qui le caractérise soit stationnaire et agisse comme un attracteur. Dans un circuit positif fonctionnel, les seuils des interactions du circuit agissent comme des frontières. Un état caractéristique stationnaire d'un circuit positif est instable car le moindre écart d'une des variables impliquée dans le circuit par rapport au seuil de l'interaction sur la variable suivante, suffit pour la faire basculer en dessous ou au dessus du seuil.

Sélection des boucles fonctionnelles. E.H. Snoussi et R. Thomas ont montré que la stationnarité d'un état caractéristique d'un circuit pouvait s'exprimer en termes de contraintes sur les paramètres du système. Pour cela, ils considèrent qu'un état est stationnaire lorsqu'il est égal à son point focal. Cette définition est suffisante pour les états réguliers, mais ne l'est pas pour les états singuliers, puisque pour ces états, les points focaux ne sont pas définis. Ils définissent donc les états singuliers stationnaires comme étant les états singuliers pour lesquels les points focaux des états adjacents forment un pavé qui contient l'état lui-même. Dans [86], nous revisitons le travail de E.H. Snoussi et R. Thomas en introduisant explicitement les états singuliers. La définition des états stationnaires combinée avec celle des ressources singulières (prédécesseurs dont la concentration est sur le seuil d'activation) nous permet de donner une nouvelle preuve du théorème suivant : parmi les états singuliers, seuls les états caractéristiques d'un circuit peuvent être stationnaires.

Enfin, R. Thomas et E.H. Snoussi ont montré [100] que la stationnarité d'un état caractéristique pouvait se tester sur certains états réguliers adjacents à l'état singulier considéré. Ces contraintes s'expriment simplement en termes de ressources : pour tout gène dans le circuit, le seuil d'interaction de ce gène sur son suivant dans le circuit, doit être compris entre la valeur du paramètre comprenant uniquement les ressources habituelles et celui comprenant aussi les ressources singulières.

4.2.2 Formalisation des propriétés biologiques

La théorie des circuits de R. Thomas permet de construire des contraintes sur les paramètres de la modélisation afin qu'un circuit de rétroaction puisse être fonctionnel. Cependant, certaines propriétés comportementales du système ne peuvent pas toujours se transcrire en fonctionnalité de circuits de rétroaction. Par exemple, il est parfois connu que si l'on augmente la concentration d'un produit d'un gène, alors plus tard la production d'un autre produit a lieu. Ce type de trace comportementale peut aider à construire plus rapidement des modèles cohérents avec les connaissances biologiques.

Les propriétés biologiques sont souvent exprimables dans un langage formel. Par exemple, les logiques temporelles CTL et/ou LTL permettent d'exprimer de manière formelle que certains événements sont antérieurs à d'autres, que le système doit passer par tel événement pour atteindre un état donné, qu'il est impossible d'atteindre tel autre, qu'il est possible qu'un événement arrive... Ainsi, si une propriété biologique est exprimée de manière formelle, tester la cohérence entre un modèle et cette propriété, revient à la vérification de spécifications formelles. Les méthodes comme le model-checking peuvent alors aider le modélisateur à choisir parmi l'ensemble des modèles ceux qui satisfont ces spécifications.

4.2.3 Utilisation intensive du model-checking

L'approche proposée se résume ainsi :

1. Construction de l'ensemble $\mathcal{G}$ des graphes d'interactions à envisager dans la modélisation du système étudié.
2. Transcription des propriétés temporelles en langage formel. La formule φ représente alors toutes les connaissances (ou hypothèses dont on cherche à tester la cohérence) sur le comportement du système. Cette étape doit être faite avec beaucoup de soin car l'oubli d'une propriété « évidente » peut entraîner la sélection de modèles incohérents.
3. Construction d'un graphe d'états possible (un modèle) à partir d'un graphe d'interactions et d'un paramétrage possible.
4. Utilisation du model-checking pour tester si le modèle satisfait la formule φ. Retenir ce modèle uniquement si la formule est satisfaite.
5. Revenir à l'étape 3 pour générer un autre modèle, jusqu'à épuisement des modèles possibles.
6. Si l'ensemble des modèles retenus $\mathcal{M}_\varphi$ n'est pas vide, alors la cohérence entre les graphes d'interactions et la formule φ est confirmée, et les modèles à considérer pour la suite de l'étude sont ceux de $\mathcal{M}_\varphi$

7. Si l'ensemble $\mathcal{M}_\varphi$ est vide, alors il y a inconsistance entre la formule φ et l'ensemble des modèles possibles. Le graphe d'interactions ou la formule φ doit être remis en question.

Deux prototypes de plate-forme d'aide à la modélisation ont été développés pour permettre au biologiste de vérifier systématiquement la cohérence du système sous différentes hypothèses. La première repose sur la logique temporelle CTL et sur la représentation des dynamiques sous forme de structure de Kripke (voir figure 4.1), la seconde repose sur l'utilisation de la logique LTL et l'utilisation des réseaux de Petri de Haut Niveau.

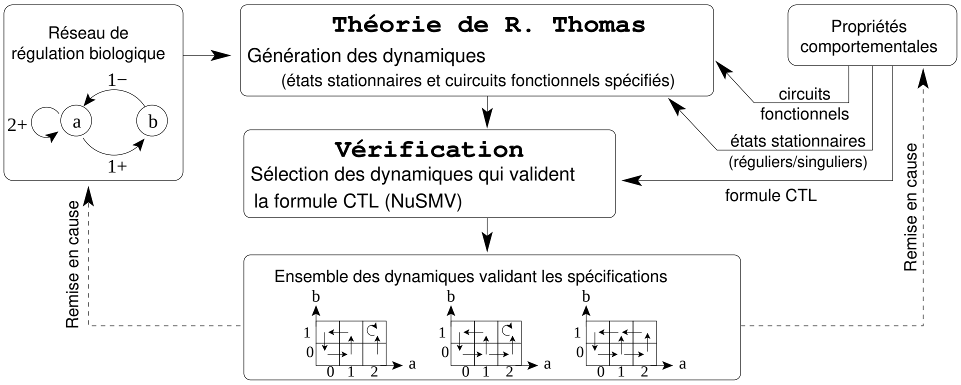


FIG. 4.1 – *Utilisation du model-checking pour sélectionner les modèles cohérents. Les graphes d'interactions ainsi que les propriétés comportementales sont des données d'entrée. Si l'ensemble des modèles sélectionnés est vide, il y a inconsistance entre les formules représentant les propriétés dynamiques et les graphes d'interactions.*

Prototype `SMBioNet`

La première plateforme, `SMBioNet`[2], est basée sur la logique temporelle *Computational tree Logic* [34] et le model-checker NuSMV développé par A. Cimatti *et al.* [20]. Nous avons démontré son applicabilité dans l'étude de la cytotoxicité chez *Pseudomonas aeruginosa* [17, 45, 38]. Cette approche est cependant limitée par le phénomène d'explosion combinatoire à deux niveaux. D'une part, l'espace des états, qu'il est nécessaire de construire avant de vérifier la satisfiabilité d'une formule CTL, peut être important. Cependant l'utilisation du model-checker symbolique NuSMV permet de traiter des systèmes de grande taille si ces derniers sont décrits de manière symbolique. D'autre part, le nombre de modèles à construire est lui-même énorme. C'est pour cela que cette plateforme permet de spécifier d'autres éléments qui contraignent fortement les valeurs de paramètres : les boucles fonctionnelles et les états stationnaires connus.

Prototype `Marianne`

La deuxième plate-forme a été développée pour éviter de construire la totalité du système de transitions, car les propriétés temporelles concernent souvent l'évolution du système à partir d'un état initial particulier. Cette deuxième plate-forme utilise la logique temporelle *Linear Temporal Logic* et une représentation du système de transitions sous forme de réseau de Petri de haut niveau. Le model checker utilisé est Maria (Modular Reachability Analyzer for Algebraic System Nets) développé par M. Mäkelä [71] .

[2] http ://smbionet.lami.univ-evry.fr

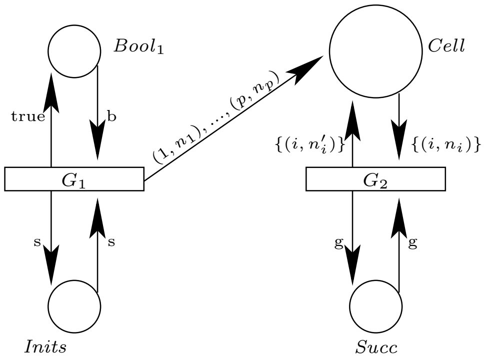


FIG. 4.2 – *Implémentation en MARIA des HLPN. La transition* $G_1 = (b = false) \wedge (s = ((1, n_1), (2, n_2), ..., (p, n_p)))$ *est franchie seulement après l'étape d'initialisation.* G_2 *est la garde asynchrone complétée avec le nombre de gènes qui évoluent (voir texte).*

Implémentation avec Maria. Le model-checker Maria [71] prend en entrée un réseau de Petri de haut niveau décrit à l'aide d'un langage particulier, une formule LTL et effectue le model-checking à partir d'un état initial donné. Malheureusement, Maria ne permet de spécifier qu'un seul état initial. Il est donc nécessaire d'inclure dans la modélisation un mécanisme permettant de prendre en compte tous les marquages initiaux intéressants. Pour cela nous introduisons par rapport à la modélisation présentée en figure 3.6 page 46, deux nouvelles places et une nouvelle transition (voir figure 4.2).

- La première place, *Inits*, contient tous les états initiaux qui doivent être considérés.
- La seconde place, $Bool_1$, contient un jeton booléen initialement positionné à faux, signifiant que l'état initial n'a pas été choisi. Après ce choix, le jeton est de type vrai.
- La transition G_1 effectue ce choix et génère les jetons correspondants dans la place *Cell*.

Le réseau de Petri de haut niveau doit aussi permettre de représenter le non-déterminisme lorsque plusieurs états successeurs peuvent être atteints. Pour cela, on rajoute une place *Succ* initialisée avec tous les nombres entiers compris entre 0 et p où p est le nombre de gènes présents dans le graphe de régulation. La transition G_2 lit l'état courant, choisit un jeton de cette place *Succ* et génère l'état suivant selon le jeton choisi : si le jeton j est choisi, le nouveau marquage correspond à l'état obtenu si c'est le gène j qui évolue.

Cette modélisation permet d'implémenter l'approche globale « génération–vérification » présentée plus haut. Mais il serait plus intéressant de tester directement un ensemble de modèles sans avoir à appeler le model-checker pour chacun des modèles. Cela est possible en rajoutant trois places supplémentaires et une transition (voir figure 4.3). La place *Models* est initialisée avec tous les modèles considérés, la place *Model* est initialement vide mais comporte ensuite un jeton correspondant à un modèle particulier, et la place $Bool_2$ contient un jeton booléen qui est faux si *Model* est vide. La transition G'_3 effectue le choix du modèle et la transition G'_2 qui simule l'évolution d'un réseau de régulation, lit les paramètres dans la place *Model*.

Cette approche peut être utilisée pour trouver un premier modèle qui est compatible avec des propriétés temporelles. Supposons que nous cherchons un modèle qui satisfait un ensemble de formules $\varphi_1, \varphi_2, ..., \varphi_n$. Le model-checker Maria essaye de valider les formules pour tous les chemins possibles de tous les modèles. Si aucun chemin ne contredit les formules, Maria répond que le réseau de Petri les satisfait. Cela signifie que tous les modèles satisfont les propriétés temporelles spécifiées. Initialisons maintenant le réseau de Petri avec tous les modèles et considérons la formule $XX(\neg(\varphi_1 \wedge \varphi_2 \wedge ... \wedge \varphi_n))$. La double occurence de X exprime que les formules sont vérifiées après l'initialisation du modèle et l'initialisation du marquage initial. Le model-checker

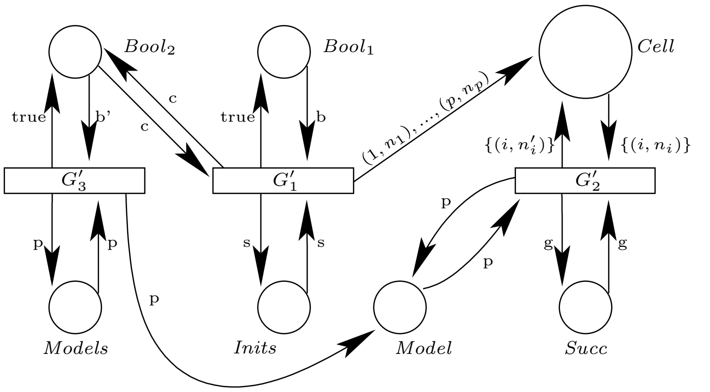


FIG. 4.3 – *Implémentation d'un HLPN manipulant un ensemble de modèles associés à un réseau de régulation.*

recherche un modèle qui contredit la négation de la conjonction des formules. Si un tel chemin existe, Maria s'arrête et donne les paramètres du modèle associé.

Malheureusement, la logique temporelle utilisée limite cette approche. Supposons que nous cherchons les modèles qui satisfont φ et passons au model-checker la négation de la formule. Maria essaye d'exhiber un chemin qui contredit la négation, *i.e.*, qui satisfait la formule. De ce chemin on déduit un modèle, mais pour ce modèle il pourrait exister un autre chemin qui réfute la formule. En fait, pour développer un tel outil, on a besoin d'exprimer que quelque chose est possible ou nécessaire, en d'autres termes CTL est une logique temporelle plus adaptée. Par exemple, si on doit verifier la formule pour tous les chemins et pour tout état initial, nous cherchons un modèle qui satisfait la formule $EX\ AX\ \varphi$: il existe un modèle pour lequel tout état initial mène à un état où la formule φ est satisfaite. La négation de la formule est $AX\ EX\ \neg\varphi$ et le model checker renvoie, s'il en existe un, un chemin qui réfute la formule, *i.e.*, un modèle tel que pour tout marquage initial, on a φ.

Cette modélisation permet d'exhiber un modèle qui satisfait les spécifications temporelles. L'existence d'un tel modèle prouve la cohérence entre le graphe de régulation et les propriétés temporelles, mais il n'est pas sûr que le système biologique fonctionne de la même manière. D'un point de vue biologique il est donc plus utile d'exhiber tous les modèles qui sont cohérents avec les propriétés temporelles, puisque cela permet d'expliquer les comportements par différents modèles.

Lisibilité des HLPN. La reformulation de la dynamique des réseaux de régulation en termes de HLPN nous a semblé dans un premier temps bien simple puisque la structure du HLPN était la même pour tout réseau de régulation génétique et que seule la garde était spécifique. L'implémentation a fait émerger la nécessité d'ajouter des places et transitions pour gérer les non-déterminismes de la dynamique et de l'état initial. La structure du réseau de Petri n'illustre plus aucune information biologique mais plutôt des artefacts de la modélisation. Tout le corpus des outils développés dans le cadre des réseaux de Petri n'est donc utilisable que si l'on accepte de perdre une certaine lisibilité du modèle.

Aide à la découverte de modèles

Ces deux prototypes de plateforme d'aide à la modélisation des réseaux de régulation génétique sont fondés sur la vérification formelle de propriétés sur un ensemble fini de modèles. En effet, l'utilisation intensive de la vérification dans un cadre fini permet de faire de la déduction, il s'agit tout simplement d'un raisonnement par cas.

4.2.4 Utilisation de la programmation logique avec contraintes

Pour éviter cette recherche exhaustive de type force brute, nous pouvons essayer de raisonner sur l'ensemble des modèles à considérer. Comme les règles de construction de la dynamique asynchrone s'écrivent facilement de manière logique, il est aisé d'écrire en prolog un programme qui permet d'exprimer la dynamique d'un modèle associé au système, voir figure 1 de l'annexe scientifique.

Une telle implémentation permet de vérifier si un jeu de paramètres mène à une dynamique cohérente avec une propriété temporelle simple. Pour trouver tous les jeux de paramètres qui mènent à une dynamique cohérente avec une propriété, il faut s'intéresser à la *programmation logique par contraintes*[3], dont les applications en bio-informatique voient le jour comme le montre la création de workshops spécialisés [11]. Comme la modélisation de R. Thomas fait intervenir des domaines finis, on peut utiliser Gnu Prolog développé par D. Diaz et P. Codognet [32] pour tester l'approche. La figure 2 de l'annexe scientifique donne le code d'un programme Gnu Prolog naïf qui renvoie l'ensemble des jeux de paramètres qui mènent à une dynamique comportant un chemin spécifié. Nous pensons que cette approche est prometteuse comme le montrent les travaux de L. Trilling et de son équipe [36], qui se sont intéressés à l'inférence de paramétrages de modèles de réseaux de régulation génétique.

4.3 Cycle de développement de modèles et classes d'équivalence

Nous avons montré jusque là comment construire de façon exhaustive l'ensemble des modèles qui répondent aux spécifications. Pour affiner le modèle, il est nécessaire de confronter l'ensemble des modèles retenus à de nouvelles expériences *in vivo*. On est ainsi amené à définir un *Cycle de développement de modèles* qui est basé sur un cycle *expérimentation–modélisation* :

1. avec un graphe d'interactions, on construit tous les modèles possibles, et si de plus on connaît une propriété biologique du système, on peut élaguer cet ensemble en enlevant les modèles ne satisfaisant pas la propriété,
2. l'analyse de l'ensemble des modèles restants peut permettre de proposer une expérience à réaliser afin de mieux connaître le système biologique. L'expérience une fois faite permet de réduire l'ensemble des modèles à considérer. Cette deuxième phase peut être répétée autant de fois que nécessaire.

La deuxième phase peut être facilitée par la classification des différents modèles retenus par l'utilisation intensive du model checking. Cette classification doit se baser sur les différentes traces des différents modèles : on ne peut distinguer deux modèles théoriques que s'il existe une trace observable qui correspond à une exécution possible de l'un des deux modèles et qui ne correspond à aucune exécution possible pour l'autre modèle. Pour définir ces traces, nous devons d'abord préciser quelles sont les expériences possibles.

4.3.1 Opérabilité / observabilité

L'ensemble des expériences que les biologistes peuvent mettre en œuvre sont soumises à plusieurs types de contraintes.

[3] Voir par exemple le livre de F. Fages [35].

- L'*opérabilité* est la capacité de mettre le système biologique dans un état spécifié ou partiellement spécifié. Il est inutile de proposer au biologiste une expérience qui consiste à regarder le comportement du système à partir d'un état initial qu'il n'a pas les moyens d'imposer au système. Les raisons de la non-opérabilité peuvent être technologiques, budgétaires, de délais ou encore d'ordre éthique.
- L'*observabilité* est la capacité à observer les conséquences des modifications de l'état afin de vérifier leur adéquation avec celles prévues par le modèle. Il est inutile de proposer une expérience pour laquelle le modèle prévoit une trajectoire vers un état que les moyens d'observation ne permettent pas d'identifier.

Certains modèles théoriques pourraient ne pas être validables soit par un manque d'opérabilité soit par un manque d'observabilité du système biologique correspondant. De plus, puisqu'on observe les conséquences de la manipulation qui amène le système dans un état initial partiellement spécifié, on particularise les types d'expériences réalisables. Une fois l'état initial atteint, non seulement il faut attendre que le système biologique réagisse aux stimuli mais aussi que les changements prévus amènent le système dans un état observable. La représentation générique d'une expérience sera donc de la forme :

$$\text{quelque chose opérable} \quad \Longrightarrow \quad \text{dans le futur (quelque chose observable)}$$

4.3.2 Classes d'équivalence

Pour étudier l'ensemble des modèles possibles associés à un graphe d'interactions, on a envie de regrouper les modèles en fonction de leurs dynamiques. Il est évident que l'isomorphisme entre deux structures de Kripke associées à un même graphe d'interactions définit une relation d'équivalence entre modèles (*identité entre dynamiques*) : deux paramétrages différents ne pourront pas être distingués.

Une autre relation d'équivalence sur l'ensemble des modèles peut être définie en se fondant sur les capacités expérimentales. Cette *relation d'équivalence* via *l'observabilité* met dans la même classe d'équivalence deux modèles qui ne sont pas expérimentalement distinguables. La définition de la capacité expérimentale dans un contexte technique, éthique ou budgétaire donné, est un arbitraire qui gouverne la relation d'équivalence. Il est donc nécessaire de définir précisément l'ensemble des expériences possibles pour un système biologique donné. Dans notre contexte théorique, cet ensemble s'exprime sous forme d'un ensemble de formules temporelles CTL dont la forme générique a été décrite plus haut.

Nous voyons ici que la modélisation a besoin de connaître les capacités expérimentales, puisqu'il est inutile de modéliser des comportements que l'on ne pourra pas observer. La formalisation des expériences implémentables *in vivo* permet de limiter les modèles d'intérêt tout en permettant de renforcer le poids des aspects expérimentaux dans le processus de modélisation.

4.4 Classification de modèles

La démarche proposée de sélection des modèles menant à une dynamique cohérente avec les spécifications provenant de la biologie peut amener à un grand nombre de modèles. Pour l'exemple de la cytotoxicité de *Pseudomonas aeruginosa* [45, 38], dont le graphe d'interactions contient deux gènes, on peut générer 648 modèles dont uniquement 32 vérifient les spécifications exprimées sous forme de formules CTL. On peut encore envisager de traiter à la main ces 32 modèles. Par contre dans d'autres cas, l'élagage de modèles peut mener à considérer plusieurs centaines, voire plusieurs dizaines de milliers de modèles. Il devient donc impossible de regarder à la main chacun de ces modèles et de laisser le biologiste avoir un regard d'expert sur chacun d'eux. Nous sommes donc confrontés à l'interprétation de l'ensemble des modèles dont certains se ressemblent du point de vue de la dynamique sous-jacente et d'autres ont des comportements complètement différents. Cette hétérogénéïté peut être mise à profit pour rechercher une représentation compacte de l'ensemble de modèles et faciliter le travail de l'expert qui pourra éventuellement sélectionner voire éliminer des classes entières de modèles.

Nous nous sommes donc intéressés à la classification de modèles d'un point de vue apprentissage statistique afin de révéler les caractéristiques qui peuvent les différencier. Un individu est ici un système de transitions, structure extrêmement riche, non numérique, vu comme un graphe. Lorsqu'on compare deux systèmes de transitions, les sommets sont identiques et seules les transitions diffèrent d'un système de transitions à un autre. La nature très riche et non numérique des données nous amène à utiliser des méthodes d'apprentissage pouvant prendre en compte des données structurées, comme les méthodes à noyaux dont le but est de transformer les données dans un espace de dimension généralement plus importante afin de rendre le problème de la classification plus simple.

4.4.1 Méthodes à noyaux

Les noyaux définissent une mesure de similarité entre objets de structure arbitraire. Ils permettent en outre d'incorporer lors de la conception du noyau, des connaissances *a priori* sur le problème, de transformer implicitement des objets dans un autre espace où la résolution du problème s'avère plus facile, et de travailler efficacement dans ce nouvel espace (de dimension très grande voire infinie).

Les objets que nous traitons ici sont des graphes orientés dont les sommets sont étiquetés. Ces données n'ont donc pas de représentation vectorielle. Il est donc nécessaire de changer d'espace *via* un noyau pour représenter les données de manière vectorielle. Le nouvel espace sera appelé *espace des caractéristiques*. Pour effectuer ce changement d'espace, nous utilisons la fonction noyau qui met en jeu la similarité entre les données : $K(x_i, x_j) \;\; = < \varphi(x_i), \varphi(x_j) >$ où φ correspond à une projection des données.

4.4.2 Classification spectrale

Nous avons choisi d'utiliser, parmi les différentes possibilités, une méthode de *spectral clustering*, facile à implémenter et raisonnablement rapide. La classification spectrale repose sur l'utilisation des vecteurs propres de la matrice de Gram (matrice contenant toutes les mesures de similarité entre les données). Il existe de nombreux algorithmes bien décrits par D. Verma et M. Meila [113], ils diffèrent sur la manière d'utiliser les vecteurs propres pour créer les groupes de données. Brièvement, il existe des méthodes itératives utilisant un vecteur propre à la fois pour effectuer une partition en 2 ; d'autres méthodes dites globales utilisent k vecteurs propres pour réaliser directement la partition en k clusters. Cette deuxième classe de méthodes est reconnue comme plus performante. Elle comporte deux algorithmes classiques : algorithme de A. Ng, M. Jordan et Y. Weiss [81] et l'algorithme multi-cut de M. Meila et J. Shi [77]. Ces algorithmes de classification spectrale reposent sur un même schéma qui peut se résumer en quatre étapes.

1. La première phase de l'algorithme est la construction de la matrice de Gram A (matrice contenant toutes les mesures de similarité entre les données), puis la construction de la matrice diagonale D dont l'élément D_{ii} est la somme des éléments de la ligne i de la matrice de Gram.
2. On normalise la matrice de Gram en posant : $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$. Pour deux individus i et j, L_{ij} est le rapport entre la similarité A_{ij} et le produit $\sqrt{D}_{ii} \times \sqrt{D}_{jj}$.
3. On forme la matrice $X = [x_1, \ldots, x_k]$ des k plus grands vecteurs propres, puis on la normalise : $Y_{ij} = \frac{X_{ij}}{(\Sigma_j X_{ij}^2)^{-\frac{1}{2}}}$

4. Le clustering consiste à traiter chaque ligne de Y comme un point de $\mathbb{R}^k$ et à classer les n points *via* un algorithme classique de classification (K-means ou classification hiérarchique ascendante).

4.4.3 Construction d'un noyau adapté (classification non supervisée)

Le noyau doit être choisi judicieusement car il va influer sur la matrice de Gram et sur toute la classification. Il doit prendre en compte les différents éléments pertinents pour la classification. Une manière simple de comparer deux graphes représentés sous la forme de matrices d'adjacence, est de compter le nombre de transitions identiques présentes dans les deux graphes. Cela revient à calculer le produit scalaire matriciel des deux matrices (la somme des produits terme à terme des deux matrices). Mais cette approche ne préserve pas la structure du graphe. Il faut donc revenir à ce que l'on veut prendre en compte lors de la comparaison de deux graphes.

1. Le premier critère consiste à dire que deux graphes sont similaires s'ils partagent des chemins de longueurs k identiques. Ce critère n'est pas forcément bien adapté, car si on considère deux cycles entre 4 éléments dont la direction des flèches est inversée, les deux graphes ne partagent aucun chemin, et donc auraient une similarité nulle. Pourtant de chacun des états on peut atteindre en un nombre fini de transitions n'importe quel autre sommet.
2. Cette notion d'atteignabilité des états doit donc aussi être prise en compte. Évidemment si les chemins pour atteindre un sommet sont les mêmes dans les deux graphes, la similarité doit être plus forte. On doit donc prendre en compte le critère précédent de partage de chemins tout en rajoutant un critère d'atteignabilité.

Expression du noyau. Le premier critère est bien représenté par les différentes puissances de la matrice de transition du graphe produit. De manière générale le graphe produit $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ de deux graphes $G = (V, E)$ et $G' = (V', E')$ est le graphe dont l'ensemble des sommets est le produit cartésien des sommets, et l'ensemble des arcs est $\{(a, a') \to (b, b') | (a, b) \in E, (a', b') \in E'\}$. Ici, les sommets de G et G' sont les mêmes, et on ne considérera dans le graphe produit que les sommets de la forme (a, a). Le graphe produit $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ est donc défini par :

$$\begin{aligned} \mathcal{V} &= V \quad (= V' \ ) \\ \mathcal{E} &= \{(a \to b) | (a, b) \in E \cap E'\} \end{aligned}$$

Si on calcule la matrice d'adjacence $\mathcal{T}$ du graphe produit comme le produit terme à terme des deux matrices T et T' des deux graphes G et G', la puissance n de $\mathcal{T}$ donne le nombre de chemins de longueur n dans le graphe produit, et donc le nombre de chemins de longueur n communs aux deux graphes. Un premier noyau pour les graphes basés sur le dénombrement des chemins de longueur inférieure à n en communs peut alors être défini :

$$K_n(G, G') = \sum_{i,j} \left(\sum_{l=0}^{n} \lambda_l \mathcal{T}^l \right)_{ij}$$

où λ_l est un coefficient pondérateur des chemins de longueur l.

Pour prendre en compte le deuxième critère, on calcule les matrices $T_{ind} = \sum_{l=2}^{n} T^l$ et $T'_{ind} = \sum_{l=2}^{n} T'^l$ qui représentent respectivement le graphe des sommets atteignables dans chacun des deux systèmes de transitions, on calcule la matrice du graphe produit $\mathcal{T}_{ind}$ et on complète le noyau précédement calculé :

$$K_n(G, G') = \sum_{i,j} \left(\sum_{l=0}^{n} \lambda_l (\mathcal{T}^l + \mathcal{T}_{ind}^l) \right)_{ij}$$

On peut généraliser le noyau précédent pour prendre en compte les chemins de taille infinie (important pour prendre en compte des circuits) : $K_\infty(G, G') = lim_{n\to\infty} K_n(G, G')$. Enfin, pour assurer la convergence des séries, on pose $\lambda_l = (\frac{\beta^l}{l!})$. En notant $exp^{\beta A} = lim_{n\to\infty} \sum_{l=0}^{n}(\frac{\beta^l}{l!})A^l$, on a :

$$K_\infty(G, G') = \sum_{i,j}\left(exp^{\beta\mathcal{T}} + exp^{\beta\mathcal{T}_{ind}}\right)_{ij}$$

4.4.4 Pré-classification pour le model-checker (classification supervisée)

Dans le schéma général qui consiste à construire chacun des modèles envisageables et à appeler pour chacun d'eux le model-checker, l'étape de model-checking est coûteuse. Une étape de classification supervisée qui apprendrait à partir de quelques exemples, à séparer les systèmes de transitions qui satisfont la propriété temporelle de ceux qui ne la satisfont pas, serait utile (sous condition que cette étape soit plus rapide que celle de l'utilisation intensive du model-checking). Pour une telle classification, il est raisonnable de vouloir prendre en compte des connaissances *a priori* qui donnent un poids à certains chemins. Ces chemins sont des variables cachées à distinguer des variables visibles du problème, ici les systèmes de transitions.

Noyau marginalisé de K. Tsuda [58]. Ce noyau est une mesure de similarité basée sur l'existence de chemins partagés par deux graphes. Ces chemins sont les variables cachées du problème. Pour définir le noyau marginalisé, K. Tsuda utilise les probabilités $p(h|G)$ de ces chemins :

$$K(G, G') = \sum_{h\in H}\sum_{h'\in H} p(h|G)p(h'|G')K'((G, h), (G', h'))$$

où K' est une mesure de similarité entre deux couples de type (graphe, chemin dans ce graphe). Comme on s'intéresse aux chemins identiques dans les deux graphes, $K'((G, h), (G', h'))$ est égal à 1 si $h = h'$ et 0 sinon. Le noyau s'écrit alors : $K(G, G') = \sum_{h\in H} p(h|G)p(h|G')$.

Pour définir les probabilités associées aux chemins, on introduit les paramètres p_s (probabilités d'émission), p_q (probabilités de terminaison) et p_t (probabilités de transition). La probabilité d'un chemin $h = (h_1, h_2, ..., h_l)$ se calcule par :

$$p(h|G) = p_s(h_1)\left(\prod_{i=2}^{l} p_t(h_i|h_{i-1})\right)p_q(h_l)$$

Toute la connaissance *a priori* que l'on veut prendre en compte devra être codée dans ces probabilités. Si l'on ne dispose pas de connaissances *a priori*, ces probabilités pourront être définies de façon uniforme.

Notion de graphes comparables pour une propriété temporelle. Si l'on veut faire de la classification supervisée pour distinguer la classe des systèmes de transitions qui vérifient une propriété temporelle de celle des systèmes de transitions qui ne la vérifient pas, il est nécessaire de bien définir la règle d'affectation d'une nouvelle instance à une classe. On souhaite évidemment que si l'un des deux systèmes de transitions vérifie la propriété et pas l'autre, alors les deux graphes n'appartiennent pas à la même classe. Mais *a contrario*, il n'est pas nécessaire de mettre les deux systèmes de transition dans le même cluster même si les deux systèmes de transitions vérifient la propriété, car les similarités entre graphes peuvent être plus ou moins fortes.

Par conséquent, si deux graphes sont *comparables*, c'est à dire s'ils se comportent de la même manière pour la propriété temporelle en question, alors on les compare *via* des connaissances *a priori* (qu'il reste à définir), et on associe une similarité nulle pour tout couple de graphes non comparables.

4.4.5 Limitation de l'approche

Que ce soit pour la classification supervisée ou non supervisée, les méthodes développées ne permettent pas encore une mise en œuvre à grande échelle. Pour le problème de classification non supervisée, c'est à dire sans connaissances *a priori*, l'approche est trop longue car le calcul des exponentielles de matrices est coûteux. Pour le problème de classification supervisée, avec connaissances *a priori*, on est amené à vérifier si un système de transitions satisfait une formule donnée de logique temporelle, ce qui revient à un appel au model checking. Si on se focalise sur des propriétés temporelles extrêmement simples comme par exemple l'existence ou la non existence d'un chemin dans le graphe, alors la représentation matricielle peut donner des résultats mais l'approche reste lourde.

4.5 Confrontation des modèles théoriques à la biologie

Dans le développement d'un modèle, une étape importante reste à envisager, celle de la confrontation du ou des modèles théoriques à la biologie. Cette phase consiste à confronter les simulations ou plus généralement les propriétés obtenues à partir du modèle théorique aux résultats d'expériences humides, elle est le parallèle avec la vérification de systèmes en informatique (le système correspond-il à ces spécifications ?). Dans le cadre du génie logiciel, le but est de proposer un ensemble d'expérimentations sur le logiciel qui sont suffisantes pour établir par extrapolation que le logiciel est compatible avec son modèle.
Le travail présenté jusqu'à présent dans ce chapitre a permis, dans le cadre de la biologie des systèmes, de construire un ensemble de modèles qui sont cohérents avec des connaissances biologiques et des hypothèses. Il s'agit maintenant de proposer des expériences de biologie humide (on teste le vivant) pour valider ou réfuter un ou plusieurs de ces modèles, et par conséquent les hypothèses introduites dans ces modèles. Nous cherchons ainsi quelles sont les expériences suffisantes à faire pour valider ou réfuter un ou plusieurs modèles. Nous cherchons à aider le modélisateur dans cette phase de confrontation des modèles théoriques au modèle vivant, en s'appuyant sur un retour à des expériences humides.

Pour montrer la faisabilité de ce « retour paillasse », nous nous focalisons sur la modélisation de R. Thomas et nous nous intéressons à la recherche des expériences à mettre en œuvre pour pouvoir préciser le modèle du système biologique. Puisqu'on ne connaît pas *a priori* le résultat de chacune des expériences, il faut envisager de faire une expérience à la fois, et que le résultat de chaque expérience permette d'éliminer un certain nombre de modèles. On itère le procédé soit jusqu'à montrer l'incohérence des modèles avec les expériences mises en œuvre, soit jusqu'à obtenir un ensemble de modèles qui ne sont pas expérimentalement distinguables.

4.5.1 Notations, définitions et position du problème

Nous noterons, dans cette section, $\mathcal{M}$ l'ensemble des modèles à considérer, et $\mathcal{F}$ l'ensemble des formules implémentables expérimentalement. $\mathcal{F}$ est l'ensemble des formules qui transcrivent toutes les expériences qui sont à la fois opérables et observables dans un contexte donné.

- Soit $m \in \mathcal{M}$ et $f \in \mathcal{F}$. On note $m \models f$, si la propriété f est satisfaite par le modèle m.
- Soit $M \subset \mathcal{M}$ un sous-ensemble de modèles. On note $M(f)$ l'ensemble des modèles de M qui satisfont la propriété f : $M(f) = \{m \in M \mid m \models f\}$. On note $f_1 \equiv_M f_2$ lorsque $M(f_1) = M(f_2)$.
- Soit $F \subset \mathcal{F}$ un sous-ensemble de formules. On note $F(m)$ l'ensemble des formules de F qui sont satisfaites par le modèle m : $F(m) = \{f \in F \mid m \models f\}$. On note $m_1 \equiv_F m_2$ lorsque $F(m_1) = F(m_2)$.

Enfin, nous dirons qu'une formule est *discriminante* pour un ensemble de modèle M si $M(f) \neq \{\}$ et $M(\neg f) \neq \{\}$, autrement dit si la réalisation de l'expérience associée à la formule apporte une information sur les modèles du système biologique associé.

Quotient de l'ensemble des modèles par la relation d'équivalence $\equiv_{\mathcal{F}}$. Plusieurs modèles peuvent répondre de la même manière à l'ensemble des formules de $\mathcal{F}$. Dans ce cas, ces modèles ne sont pas distinguables, et la relation d'équivalence $\equiv_{\mathcal{F}}$ n'est pas réduite à l'identité. Nous nous intéressons donc à l'ensemble quotient $\mathcal{M}/\equiv_{\mathcal{F}}$ des classes d'équivalence de $\mathcal{M}$: tous les modèles ayant le même *profil de réponses aux formules* sont considérés comme un seul modèle.

Quotient de l'ensemble des formules par la relation d'équivalence $\equiv_{\mathcal{M}}$. Symétriquement il se peut que plusieurs formules discriminent de la même manière l'ensemble des modèles. Dans ce cas, expérimenter l'une ou l'autre de ces formules permet de sélectionner les mêmes modèles. Nous nous intéressons donc à l'ensemble quotient $\mathcal{F}/\equiv_{\mathcal{M}}$ des classes d'équiva-lence de $\mathcal{F}$: les formules ayant le même *profil de réponses aux modèles* sont considérées comme une seule formule/expérience.

L'étape de construction des deux ensembles quotients permet de limiter l'espace de recherche pour le retour paillasse. Elle permet de particulariser l'approche en faisant le lien entre les capacités expérimentales et les modèles théoriques d'intérêt car plusieurs outils, méthodologies expérimentales, schémas d'expériences peuvent élaguer l'ensemble de modèles théoriques de la même manière. Le problème de recherche des expériences à implémenter pour arriver à sélectionner le ou les modèles correspondant au système biologique revient à construire un arbre dont les nœuds sont des ensembles de modèles (ceux qui sont cohérents avec les comportements biologiques connus) munis d'une formule à expérimenter pour partager cet ensemble en deux ensembles non vides. Chaque formule ne doit être utilisée qu'une seule fois, ce qui justifie d'étiqueter chaque nœud par l'ensemble des formules pouvant encore être utilisées. Les arbres possibles sont donc formalisés par les $\mathcal{MF}$-arbres.

Définition 10 *Les $\mathcal{MF}$-arbres sont des arbres binaires dont les noeuds sont étiquetés par des triplets $(M, f, F) \in \mathcal{P}(\mathcal{M}) \times (\mathcal{F} \cup \{\bot\}) \times \mathcal{P}(\mathcal{F})$ tels que*
- *la racine est étiquetée par $(\mathcal{M}, f, \mathcal{F} \setminus \{f\})$ avec $f \in \mathcal{F}$,*
- *un nœud (M, φ, F) est une feuille si aucune formule de F n'est discriminante pour l'ensemble des modèles M. Dans ce cas, $\varphi = \bot$,*
- *si le nœud (M, φ, F) n'est pas une feuille, alors son fils droit est étiqueté par $(M(\neg\varphi), \psi', F \setminus \{\psi'\})$ avec $\psi' \in (F \cup \{\bot\})$, et son fils gauche est étiqueté par $(M(\varphi), \psi'', F \setminus \{\psi''\})$ avec $\psi'' \in (F \cup \{\bot\})$.*

La condition aux feuilles regroupe deux cas. Lorsque l'ensemble des modèles est un singleton, aucune formule n'est discriminante. Dans ce cas, on a déterminé le seul modèle cohérent avec un ensemble d'expériences réalisées. Si l'ensemble de modèles comporte plus de 2 éléments, alors, les modèles de cette feuille forment une classe d'équivalence par rapport aux expériences possibles restantes. L'ensemble des possibles ayant été défini, nous définissons plusieurs problèmes de recherche de schémas d'expériences (liste non exhaustive).

Définition 11 (Problèmes d'optimisation) *Soit $\mathcal{M} = \{m_1, m_2, ..., m_m\}$ un ensemble fini de modèles et soit $\mathcal{F} = \{f_1, f_2, ..., f_n\}$ un ensemble fini de formules représentant les expériences envisageables.*

1. *Le problème de recherche du schéma d'expériences* **à horizon 1 à coût fixe** *consiste à trouver une unique formule qui permet de minimiser l'espérance du nombre de modèles restants après mise en œuvre de l'expérience.*
2. *Le problème de recherche* **à horizon** n **à coût fixe***, consiste à trouver l'arbre de profondeur* n *qui minimise l'espérance du nombre de modèles restants après la mise en œuvre des* n *expériences de la branche définie par les réponses successives du système biologique.*
3. *Si de plus chaque expérience a un coût spécifique, le problème de recherche* **à horizon** n **à coût variable***, consiste à trouver l'arbre de profondeur* n *qui minimise l'espérance du coût pour chacune des feuilles de l'arbre.*
4. *Le problème de recherche totale* **à coût fixe** *consiste à trouver le* $\mathcal{MF}$*-arbre qui discrimine tous les modèles, et qui minimise le nombre moyen d'expériences à faire pour discriminer le modèle du système biologique.*
5. *Le problème de recherche totale* **à coût variable** *consiste à trouver l'arbre qui discrimine tous les modèles, et qui minimise le coût moyen pour discriminer le modèle du système biologique.*

4.5.2 Quelques éléments de réponse

Recherche du schéma d'expériences à horizon 1 à coût fixe

Cas où les modèles sont tous équiprobables. Supposons que l'on mette en œuvre la formule $\varphi \in \mathcal{F}$. Notons N_φ le nombre de modèles qui satisfont φ, et de manière similaire $N_{\neg\varphi}$ le nombre de modèles ne satisfaisant pas φ. L'espérance du nombre de modèles restants après la mise en œuvre de la formule s'exprime par :

$$\begin{aligned} E(n(\varphi, M)) &= N_\varphi \times \frac{N_\varphi}{N} + (N - N_\varphi) \times (1 - \frac{N_\varphi}{N}) \;=\; N \times (\frac{N_\varphi}{N})^2 + N \times (1 - \frac{N_\varphi}{N})^2 \\ &= N \times (2(\frac{N_\varphi}{N})^2 - 2(\frac{N_\varphi}{N}) + 1) \end{aligned}$$

Le minimum de la fonction $E(n(\varphi, M))$ est atteint lorsque $\frac{N_\varphi}{N} = 1/2$. Il suffit donc de rechercher la formule qui coupe l'ensemble des modèles M en deux parties de cardinal égal. Comme cette formule n'existe pas forcément, on recherche la formule φ_{opt} qui minimise la différence entre les cardinaux des ensembles N_φ et $N_{\neg\varphi}$:

$$\varphi_{opt} = \underset{\varphi \in \mathcal{F}}{argmin} \quad ||N_\varphi| - |N_{\neg\varphi}||$$

Cas où l'on connaît une loi de probabilité sur les modèles. Nous nous plaçons maintenant dans le cas où une distribution de probabilité est donnée sur l'ensemble M des modèles : chaque modèle m_i est muni d'une probabilité p_i. Dans ce cas, il faut minimiser l'espérance de la mesure de l'ensemble des modèles restants après la mise en œuvre d'une formule :

$$E(\mu(\varphi, M)) = (p_\varphi)^2 + (1 - p_\varphi)^2$$

où $p_\varphi = \sum_{m_i \in M_\varphi} p_i$. De la même manière, l'espérance est minimale pour $p_\varphi = 1/2$. Il faut donc rechercher la formule qui coupe l'ensemble des modèles M en deux parties dont les mesures sont les plus proches :

$$\varphi_{opt} = \underset{\varphi \in \mathcal{F}}{argmin} \quad |(\sum_{m_i \in M_\varphi} p_i) - (\sum_{m_j \in M_{\neg\varphi}} p_j)|$$

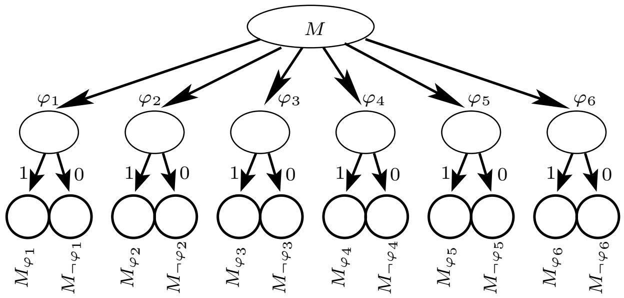


FIG. 4.4 – *Arbre du jeu à profondeur 1. A la racine, l'ensemble des modèles est l'ensemble M, et il est possible de jouer n'importe laquelle des formules φ_1, φ_2, φ_3, φ_4, φ_5, φ_6. Le système donne alors la réponse à l'expérience φ choisie, à savoir* 1 *si le système satisfait la propriété φ, 0 sinon.*

Recherche du schéma d'expériences à horizon n

Pour aborder ce problème, nous nous basons sur un arbre de jeu à deux joueurs. Le joueur 1 choisit une formule à jouer parmi un ensemble fini, alors que le deuxième joueur choisit de répondre par 0 ou 1 (0 signifiant que le système biologique ne valide pas la formule que le joueur 1 vient de jouer). On construit l'arbre de jeu comme dans la figure 4.4 jusqu'à la profondeur n. Chaque feuille est étiquetée par le cardinal de l'ensemble des modèles restants dans la feuille.

Cette structure permet d'utiliser l'algorithme Min-Max de J. Von Neuman. En remontant progressivement des feuilles à la racine, on étiquette les nœuds du joueur 1 par le minimum des étiquettes des fils et les autres nœuds par le maximum des étiquettes des deux fils. Cette stratégie optimise le choix d'une branche de l'arbre de jeu de telle sorte que l'on minimise le pire des cas : on choisit à chaque étape, la formule $\varphi \in \mathcal{F}$ qui minimise la taille de la feuille dans le cas le plus défavorable.

Nous pouvons adapter l'algorithme Min-Max au problème de recherche du schéma d'expériences à horizon n. En remplaçant, lors de la phase d'étiquetage des nœuds, le maximum des étiquettes des fils des nœuds du joueur 2 par leur moyenne, l'algorithme permet de choisir une branche de l'arbre qui minimise l'espérance du nombre de modèles restant après les n expériences.

Recherche du schéma d'expériences total à coût fixe

Le problème de recherche totale à coût fixe est équivalent à la recherche d'un arbre de décision de profondeur minimale. Or ce problème est connu comme étant NP-complet grâce à la réduction au problème 3-DM de L. Hyafil et R.L. Rivest [51]. De plus D. Sieling a ensuite démontré [92] que s'il existait un algorithme d'approximation avec ratio constant polynomial en temps, alors $P = NP$. Les heuristiques peuvent alors être intéressantes. Une heuristique simple consiste à choisir à chaque étape la formule qui partage de manière équilibrée l'ensemble des modèles, mais on peut construire un exemple sur lequel cette heuristique donne une solution aussi éloignée de l'optimum que l'on souhaite.

Nous nous sommes aussi intéressés à la recherche d'un arbre dont le nombre de formules différentes à l'intérieur de l'arbre est minimal. Ce problème est intéressant pour la biologie parce que le temps nécessaire pour mettre au point une expérience est parfois très long. Minimiser le nombre de formules susceptibles d'être mises en œuvre peut éventuellement permettre de commencer à préparer une expérience avant de connaître la réponse du système à l'expérience précédente. Par une réduction à 3-DM, ce problème est aussi NP-complet.

Chapitre 5

Lien entre topologie et dynamique

Résumé

Nous reprenons dans ce chapitre, les deux conjectures proposées par René Thomas sur les liens entre dynamique et topologie des réseaux de régulation génétique : la présence d'un circuit de rétroaction positif (resp. négatif) dans un graphe d'interactions est nécessaire pour qu'il y ait plusieurs états stables dans la dynamique (resp. pour qu'il y ait un comportement cyclique stable). Bien que ces conjectures aient été démontrées dans différents cadres mathématiques, nous y revenons en montrant que les outils théoriques que nous avons proposés (pont le long de cycles, matrice jacobienne définissant le graphe d'interactions local pour les systèmes dynamiques discrets) permettent d'étendre ces résultats à tout systèmes dynamiques discrets.

5.1 Modéliser permet d'énoncer des règles générales

La modélisation donne un cadre théorique dans lequel il est possible de raisonner. L'exemple de la théorie de R. Thomas sur les réseaux de régulation génétique nous a donné l'occasion de confronter la modélisation à des propriétés générales énoncées sur l'ensemble des modèles. Nous avons voulu donner une preuve mathématique des conjectures de R. Thomas portant sur le lien entre le système de transitions représentant la dynamique et la représentation des interactions.

La première conjecture stipule que la présence d'un circuit positif dans le graphe d'interactions d'un système est une condition nécessaire pour que ce système présente plusieurs états stables. Depuis 1995, différentes preuves de cette conjecture ont été obtenues dans des contextes différents : en 1995, E. Plahte, T. Mestl et S.W. Omholt [84] l'ont prouvé dans un contexte différentiel, suivi en 1998 par J.-L. Gouzé [44] et par E.H. Snoussi [99], en 2002 par O. Cinquin et J. Demongeot [21] et enfin dans un contexte général en 2003 par C. Soulé [101]. La conjecture a été récemment prouvée dans le cadre booléen par E. Rémy, P. Ruet et D. Thieffry [89] et C. Soulé [102] a ensuite prouvé par discrétisation que la conjecture restait juste dans le cadre multivalué de R. Thomas. La preuve qui en est donnée ici, issue de la thèse d'Adrien Richard [85], est plus générale car elle s'applique à tout système dynamique discret.

La deuxième conjecture stipule que la présence d'un circuit négatif dans le graphe d'interactions d'un système est une condition nécessaire pour que ce système présente des oscillations dans la dynamique. Elle fut démontrée d'abord dans des formalismes différentiels de plus en plus généraux [84, 44, 99] avant d'être aussi prouvée dans un cadre booléen [89]. Le résultat que nous avons obtenu avec Adrien Richard au cours de sa thèse [85] se place dans un cadre très général de systèmes dynamiques discrets.

La démarche habituelle de l'étude des systèmes dynamiques repose sur l'analyse de la matrice jacobienne qui donne en tout point de l'espace des phases les dérivées partielles du premier ordre d'une fonction vectorielle. Considérons le système dynamique défini par un système d'équations différentielles $d\vec{X}/dt = \vec{F}(X)$ où $\vec{F}$ est une fonction d'un ouvert de $\mathbb{R}^n$ à valeurs dans lui même déterminée par ses n fonctions composantes à valeurs réelles $(f_1(x_1, ..., x_n), ..., f_n(x_1, ..., x_n))$. Les dérivées partielles de ces fonctions en un point M, si elles existent, constituent une matrice $n \times n$, appelée matrice jacobienne de F. Cette matrice code les interactions entre les différentes variables du système. En effet, la modification de la composante i en un point M entraîne la modification du système d'équations différentielles. La composante j en sera modifiée si F_j dépend de i, autrement dit si l'élément i, j de la matrice jacobienne est non nul. De plus, l'influence de la variable i sur la variable j est positive (resp. négative) si la dérivée partielle $\partial F_j/\partial x_j$ est positive (resp. négative). Cette remarque permet de définir un *graphe d'interactions* en chacun des points de l'espace des phases.

Dans le but de prouver les conjectures de R. Thomas dans un cadre discret très général, A. Richard a généralisé au sein de sa thèse la jacobienne booléenne définie pour les systèmes booléens par F. Robert [87, 88], aux systèmes discrets [85]. Cette jacobienne explique comment le graphe d'interactions est construit à partir de la dynamique, elle nous a permis de raisonner sur un ensemble de dynamiques beaucoup plus grand que celui atteignable par des modèles de R. Thomas, en particulier, on peut considérer des dynamiques de $\mathbb{Z}^n$ dans $\mathbb{Z}^n$. La conjecture sur l'homéostasie se prouve ensuite facilement en parcourant le cycle sans issue et en construisant les interactions successives. L'autre conjecture qui porte sur la multi-stationnarité, repose sur le même schéma de preuve que celui utilisé par E. Remy, P. Ruet et D. Thieffry [89].

Ces deux résultats n'ont en fait pas le même statut : dans le cas de la multi-stationnarité, l'existence de la boucle positive est démontrée pour au moins un état dans l'espace des phases, alors que pour le comportement cyclique, l'existence de la boucle négative est pour l'instant prouvée pour une superposition de graphes d'interactions différents dans l'espace des phases, ce qui en fait un résultat plus faible.

5.2 Boucles négatives

5.2.1 Système dynamique discret, domaines et cycles sans issue

On désigne par *système dynamique discret* la donnée d'un ensemble d'états discrets S (un sous ensemble de $\mathbb{Z}^n$ de la forme $\Pi_{i=1}^n[a_i, b_i]$) et d'une fonction de S dans lui-même. La dynamique asynchrone, notée $\mathcal{A}(F)$, est donnée par l'ensemble des désynchronisations (cf. définition 5 du chapitre 3, page 41) des transitions synchrones $(a, F(a))_{a\in S}$. Nous appellerons gène chacune des dimensions de S par analogie avec la théorie de R. Thomas.

Un chemin dans $\mathcal{A}(F)$ est une séquence d'états $(x^1, \dots, x^k)$ telle que $k > 1$ et $(x^i, x^{i+1}) \in \mathcal{A}(F)$ pour $i = 1, \dots, k-1$. Un *cycle* est un chemin $(x^1, \dots, x^k)$ où $(x^k, x_1) \in \mathcal{A}(F)$. Un cycle $(x^1, \dots, x^k)$ est appelé un *trap cycle* ou un *cycle sans issue* si pour tout état x^i du cycle, il n'existe qu'une seule désynchronisation de la transition synchrone $(x^i, F(x^i))$, en d'autres termes s'il n'existe pas de chemin entre un état du cycle et un état hors du cycle. Plus généralement, un *domaine sans issue* ou *trap domains* est un ensemble d'états dont il n'est pas possible de sortir.

5.2.2 Graphe d'interactions le long d'un chemin

Lorsqu'on se déplace le long d'un chemin dans la dynamique asynchrone d'un système dynamique discret, certaines transitions expriment une interaction. En effet, deux transitions consécutives dans des directions différentes reflètent une interaction entre les gènes représentés par ces directions. Soit le chemin $(a, b, c) \to (a, b+1, c) \to (a+1, b+1, c)$. L'augmentation du ni-

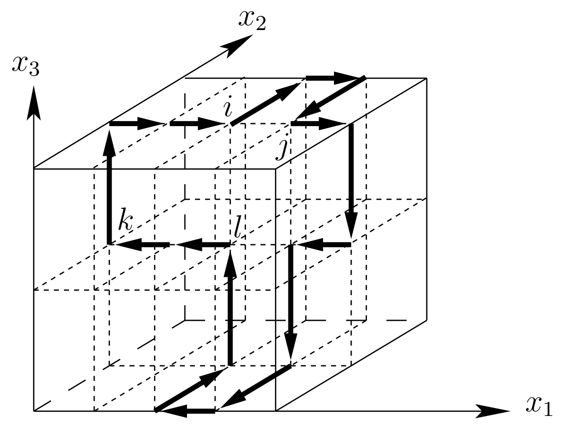


FIG. 5.1 – *Pour ce circuit, seuls deux états ne sont pas des coudes.*

veau de concentration du deuxième gène entraîne une augmentation du niveau de concentration du premier, autrement dit, le deuxième gène agit positivement sur le premier.

Un tel enchaînement de transitions est appelé un *coude*. Dans l'exemple de la figure 5.1, l'état x^i est un coude pour le cycle représenté, et le long de ce cycle il n'y a que deux états qui n'en sont pas : $i-1$ et $k-1$. Si maintenant, on considère que le cycle de la figure 5.1 est un cycle sans issue, de chacun des états, on ne peut aller que dans un seul état suivant, et on déduit de chaque coude du cycle une interaction entre deux gènes. Par exemple, à l'état x^{i-1} de la figure 5.1, seul le gène x_1 évolue. Mais lorsque le gène x_1 évolue, c'est à dire lorsque le système passe dans l'état x^i, le gène x_2 évolue à son tour. En d'autres termes, à l'état x^{i-1}, l'accroissement du gène x_1 mène à un accroissement de la variable x_2 : il y a une action positive de x_1 sur x_2 ($x_1 \xrightarrow{+} x_2$). Pour le coude j, la diminution du gène x_2 mène à un accroissement de x_1, et ainsi il y a une action négative de x_2 sur x_1 ($x_2 \xrightarrow{-} x_1$)... Ces interactions sont représentées par un graphe d'interactions.

Définition 12 *Soit $C = \{x^1, x^2, \ldots, x^r\}$ un cycle sans issue dans la dynamique asynchrone d'un système dynamique discret. Le* graphe d'interactions *du cycle C est le graphe orienté signé défini par :*

- *l'ensemble des sommets est l'ensemble des dimensions du système dynamique,*
- *il existe un arc de k à l s'il existe un coude $i \in [1, \ldots, r]$ tel que le gène qui évolue à l'état x^{i-1} est k et celui qui évolue à l'état x^i est l,*
- *le signe de cet arc est positif si l'évolution dans la direction k et l'évolution dans la direction l sont de même signe, négatif sinon. Formellement,*
  le signe de cet arc est $\begin{cases} +1 & si \quad (x_k^i - x_k^{i-1}) = +(x_l^{i+1} - x_l^i) \\ -1 & si \quad (x_k^i - x_k^{i-1}) = -(x_l^{i+1} - x_l^i) \end{cases}$.

Le signe de l'interaction de k à l est positif si l'évolution du niveau de k à l'état x^{i-1} (transition de x^{i-1} à x^i) entraîne une évolution du niveau de l à l'état x^i (transition de x^i à x^{i+1}) dans le même sens[1].

[1]Pour l'étude des boucles positives, nous introduirons dans la section suivante, des matrices jacobiennes discrètes à partir desquelles seront construits des graphes d'interactions. La condition d'existence d'une interaction entre k et l basée sur les coudes le long d'un circuit sans circuit s'exprime aussi avec la jacobienne.

5.2.3 Condition nécessaire pour un comportement cyclique stable

Pour prouver qu'un circuit négatif dans le graphe d'interactions est nécessaire pour avoir dans la dynamique un cycle sans issue, il suffit de parcourir le cycle sans issue et de construire les interactions successives. On peut montrer que faire le tour complet du cycle sans issue n'est pas nécessaire pour construire un circuit négatif (voir la notion de 0-pont et de 1-pont). La partie du cycle que l'on doit parcourir permet de plus de spécifier quels sont les gènes impliqués dans le circuit négatif.

Soit $C = \{x^1, x^2, \ldots, x^r\}$ un cycle sans issue dans la dynamique asynchrone d'un système dynamique discret. Comme il s'agit d'un cycle sans issue, à chaque état x^i de ce cycle, l'évolution du système ne peut se faire que suivant une direction donnée notée $\varphi(i)$. La définition 12 donne la condition pour avoir une interaction positive (resp. négative) entre deux gènes. En notant $\sigma(\varphi(i-1), \varphi(i))$ le signe de l'interaction de $\varphi(i-1)$ à $\varphi(i)$, on obtient la relation suivante :

$$(x^i_{\varphi(i-1)} - x^{i-1}_{\varphi(i-1)}) \ = \ \sigma(\varphi(i-1), \varphi(i)) \ \times \ (x^{i+1}_{\varphi(i)} - x^i_{\varphi(i)}). \tag{5.1}$$

Considérons maintenant un état $x^i \in C$ qui n'est pas un coude ($\varphi(i-1) = \varphi(i)$). Nous avons naturellement $(x^i_{\varphi(i-1)} - x^{i-1}_{\varphi(i-1)}) \ = \ (x^{i+1}_{\varphi(i)} - x^i_{\varphi(i)})$ puisque $\varphi(i-1) = \varphi(i)$[2]. Pour chaque $k \in [1, \ldots, r]$ et pour chaque $l \in [1, \ldots, k]$, l'équation 5.1 et la remarque précédente nous permettent d'écrire

$$(x^{k+1}_{\varphi(k)} - x^k_{\varphi(k)}) \ = \left(\prod_{\substack{i = k, \\ \varphi(i) \neq \varphi(i+1)}}^{l-1} \sigma(\varphi(i), \varphi(i+1)) \right) \times (x^{l+1}_{\varphi(l)} - x^l_{\varphi(l)}). \tag{5.2}$$

Choisissons maintenant l tel que d'une part $\varphi(l) = \varphi(k)$ et d'autre part que le signe de l'évolution du gène $\varphi(l)$ entre x^l et x^{l+1} est opposé au signe de l'évolution du même gène entre x^k et x^{k+1}. Il est trivial de montrer qu'un tel l existe pour chaque $k \in [1, \ldots, r]$. On appelle un tel couple (k, l), un *0-pont*.

Définition 13 *Soit $C = \{x^1, \ldots, x^r\}$ un cycle sans issue de $\mathcal{A}(F)$ et soit $k, l \in [1, \ldots, r]$. Le couple (k, l) est appelé un 0-pont si*

$$\begin{cases} l \neq k \\ \varphi(l) = \varphi(k) \\ (x^{k+1}_{\varphi(k)} - x^k_{\varphi(k)}) \ = \ -(x^{l+1}_{\varphi(l)} - x^l_{\varphi(l)}). \end{cases} \tag{5.3}$$

L'équation 5.2 reste valide lorsque le couple (k, l) forme un 0-pont. Ce choix de (k, l) implique que $(x^{k+1}_{\varphi(k)} - x^k_{\varphi(k)}) \ = \ -(x^{l+1}_{\varphi(l)} - x^l_{\varphi(l)})$, on en déduit que le produit des signes des interactions le long du 0-pont est négatif. Comme entre deux coudes consécutifs b_1 et b_2, on a pour tout $i \in]b_1, \ldots, b_2[$, $\varphi(i) = \varphi(b_1)$ (i n'est pas un coude), les interactions impliquées dans le produit 5.2 constituent un circuit dans le graphe d'interactions. D'où le théorème suivant.

Théorème 5.2.1 *Pour tout 0-pont (k, l) d'un cycle sans circuit $C = \{x^1, \ldots, x^r\}$, le graphe d'interactions construit le long du 0-pont (k, l) contient un circuit négatif. Les gènes de ce circuit appartiennent à l'ensemble $\{\varphi(i) \mid i \text{ est un coude }, i \in [k, \ldots, l]\}$.*

Considérons le graphe d'interactions entre 3 gènes x, y et z dont la dynamique asynchrone contient le cycle sans circuit $C = \{x^1, x^2, x^3, x^4, x^5, x^6\}$ présenté dans la figure 5.2. Sur ce cycle, il y a six 0-ponts qui peuvent être énumérés selon le premier état du pont.

[2]Nous considérons séparément le cas du cycle sans issue de longueur 2 : $x^1 \longleftrightarrow x^2$.

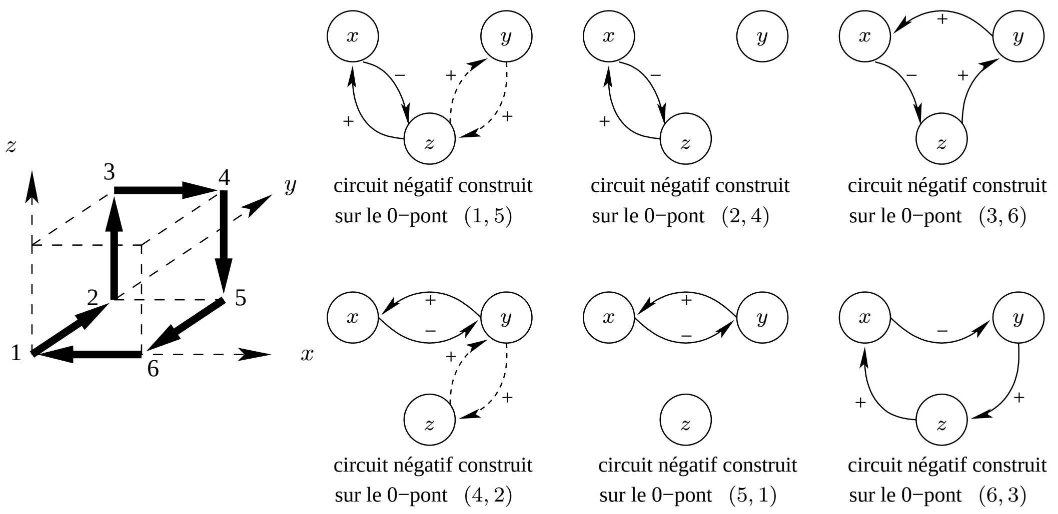


FIG. 5.2 – *Un exemple de cycle sans issue et ses circuits négatifs construits à partir des différents 0-ponts.*

1. $k = 1$, $\varphi(k) = 2$, le 0-pont est $(1, 5)$, les interactions dues aux coudes du 0-pont $(1, 5)$ sont respectivement $\varphi(1) \xrightarrow{+} \varphi(2)$, $\varphi(2) \xrightarrow{+} \varphi(3)$, $\varphi(3) \xrightarrow{-} \varphi(4)$ et $\varphi(4) \xrightarrow{+} \varphi(5)$, *i.e.* $(y \xrightarrow{+} z)$, $(z \xrightarrow{+} x)$, $(x \xrightarrow{-} z)$ et $(z \xrightarrow{+} y)$. Le circuit $(x \xrightarrow{-} z \xrightarrow{+} x)$ est négatif.
2. $k = 2$, $\varphi(k) = 3$, le 0-pont est $(2, 4)$, les interactions dues aux coudes du 0-pont $(2, 4)$ sont $\varphi(2) \xrightarrow{+} \varphi(3)$ et $\varphi(3) \xrightarrow{-} \varphi(4)$, *i.e.* $(z \xrightarrow{+} x)$ et $(x \xrightarrow{-} z)$, ce qui consitue un circuit négatif.
3. $k = 3$, $\varphi(k) = 1$, le 0-pont est $(3, 6)$, les interactions dues aux coudes du 0-pont $(3, 6)$ sont $\varphi(3) \xrightarrow{-} \varphi(4)$, $\varphi(4) \xrightarrow{+} \varphi(5)$ et $\varphi(5) \xrightarrow{+} \varphi(6)$, *i.e.* $(x \xrightarrow{-} z)$, $(z \xrightarrow{+} y)$ et $(y \xrightarrow{+} x)$, ce qui consitue un circuit négatif.
4. $k = 4$, $\varphi(k) = 3$, le 0-pont est $(4, 2)$, les interactions dues aux coudes du 0-pont $(4, 2)$ sont $\varphi(4) \xrightarrow{+} \varphi(5)$, $\varphi(5) \xrightarrow{+} \varphi(6)$, $\varphi(6) \xrightarrow{-} \varphi(1)$ et $\varphi(1) \xrightarrow{+} \varphi(2)$, *i.e.* $(z \xrightarrow{+} y)$, $(y \xrightarrow{+} x)$, $(x \xrightarrow{-} y)$ et $(y \xrightarrow{+} z)$. Le circuit $(y \xrightarrow{+} x \xrightarrow{-} y)$ est negatif.
5. $k = 5$, $\varphi(k) = 2$, le 0-pont est $(5, 1)$, les interactions dues aux coudes du 0-pont $(5, 1)$ sont $\varphi(5) \xrightarrow{+} \varphi(6)$ et $\varphi(6) \xrightarrow{-} \varphi(1)$, *i.e.* $(x \xrightarrow{-} y)$ et $(y \xrightarrow{+} x)$, ce qui constitue un circuit négatif.
6. $k = 6$, $\varphi(k) = 1$, le 0-pont est $(6, 3)$, les interactions dues aux coudes du 0-pont $(6, 3)$ sont $\varphi(6) \xrightarrow{-} \varphi(1)$, $\varphi(1) \xrightarrow{+} \varphi(2)$ et $\varphi(2) \xrightarrow{+} \varphi(3)$, *i.e.* $(x \xrightarrow{-} y)$, $(y \xrightarrow{+} z)$ et $(z \xrightarrow{+} x)$ ce qui constitue un circuit négatif.

Il est clair avec cet exemple que les gènes impliquées dans le circuit négatif dépendent du 0-pont choisi. Nous pouvons montrer qu'il existe des ponts (les 1-ponts) qui mènent à un circuit négatif dont les sommets sont tous les gènes qui évoluent lorsque l'on parcourt le cycle.

Définition 14 *Le 0-pont (k, l) d'un circuit sans issue $C = \{x^1, \ldots, x^r\}$ est un 1-pont s'il est le plus court des 0-ponts commençant en k. L'entier l tel que (k, l) est un 1-pont est noté $\sigma(k)$*[3].

Lemme 5.2.1 *Soit $C = \{x^1, \ldots, x^r\}$ un cycle sans issue de $\mathcal{A}(F)$. Il existe $k \in [1, r]$ tel que $\varphi([k, \sigma(k)[) = \varphi([1, r])$. Un* 1-pont *qui satisfait cette contrainte est appelé un* 1-pont maximal.

[3] L'application σ n'est pas une permutation.

Intuition de la preuve : Par l'absurde, considérons qu'aucun 1-pont n'est maximal et choisissons en un que l'on note $[k, \sigma(k)]$. Alors on peut construire un 1-pont $[k', \sigma(k')]$ qui le contienne strictement : $x^{k'}$ est le plus proche état précédant x^k dans le cycle tel que $\varphi(k') \notin \varphi([k, \sigma(k)])$. Ainsi, il existe une séquence infinie $(k_i)_{i \in \mathbb{N}}$ tous différents telle que $[k_i, \sigma(k_i)[\subsetneq [k_{i+1}, \sigma(k_{i+1})[$. Contradiction car le cycle est fini. □

En utilisant le théorème 5.2.1 pour un 1-pont maximal $[k, \sigma(k)]$, il existe un circuit négatif dont les sommets sont dans $\varphi(C)$. Par le lemme précédent, tous les gènes de $\varphi(C)$ sont dans le circuit.

Théorème 5.2.2 *le graphe d'interactions d'un cycle sans issue $C = \{x^1, \ldots, x^r\}$ contient un circuit négatif qui passe par tous les gènes qui évoluent lorsque l'on parcourt C.*

Il est intéressant de noter que d'autres conditions sont nécessaires pour construire directement un circuit négatif *élémentaire*. Nous définissons pour cela la longueur d'un pont (k, l) comme le nombre de coudes dans (k, l). La longueur d'un 1-pont maximal (k, l) est nécessairement plus grande que le nombre de gènes qui évoluent lorsque l'on parcourt C. Si la longueur d'un 1-pont maximal $(k, \sigma(k))$ est égale à cette borne minimale, alors la construction du circuit mène à un circuit négatif élémentaire. Un tel circuit est appelé un *1-pont principal*.

5.3 Boucles positives

Bien que la conjecture de R. Thomas concernant les circuits positifs ait été démontrée dans plusieurs cadres théoriques, il n'en existe pas de preuve directe dans un cadre discret multivalué. E. Remy, P. Ruet et D. Thieffry [89] proposent une preuve de cette conjecture dans le cas booléen en se fondant sur la définition de la jacobienne booléenne pour les systèmes dynamiques booléens. Nous reprenons cette idée en l'étendant au cas discret.

5.3.1 Jacobienne booléenne pour les systèmes dynamiques discrets

Soit x un état d'un système dynamique, nous noterons $x^{j \lhd \epsilon}$ l'état du système dont seule la coordonnée j diffère de celles de x : $x_j^{j \lhd \epsilon} = x_j + \epsilon$. Pour un système dynamique booléen $F : \{0, 1\}^n \to \{0, 1\}^n$, la matrice jacobienne booléenne $J_F(x)$ évaluée à l'état x a été introduite par F. Robert [87, 88] et s'exprime par :

$$J_F(x)_{ij} = \begin{cases} 1 & \text{si } F_i(x) \neq F_i(x^{j \lhd \epsilon}) \\ 0 & \text{sinon} \end{cases} \quad \text{où } \epsilon = \begin{cases} +1 & \text{if } x_j = 0 \\ -1 & \text{if } x_j = 1 \end{cases}, \qquad \forall i, j \in \{1, \ldots, n\}.$$

De manière intuitive, $J_F(x)_{ij} = 1$ si une commutation (de 0 à 1 ou de 1 à 0) de x_j induit une commutation de $F_i(.)$.

Pour $F : \mathbb{Z}^n \to \mathbb{Z}^n$, il n'y a pas de manière naturelle de définir une commutation, puisque à l'état x, la variation ε de la définition précédente peut prendre la valeur $+1$ ou -1. Ainsi, la matrice jacobienne dépend de x et d'un vecteur de variations $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, +1\}^n$. Naturellement, pour calculer la matrice jacobienne $J_F^{\varepsilon}(x)$, on compare $F_i(x)$ et $F_i(x^{j \lhd \varepsilon_j})$ pour tout $i, j \in \{1, \ldots, n\}$. La condition $F_i(x) \neq F_i(x^{j \lhd \varepsilon_j})$ signifie que la variation ε_j de x_j induit une modification de la coordonnée i du point focal au point x. Cependant, la modification du point focal peut ne pas être visible localement : si les deux points focaux que l'on compare font évoluer le système dans la même direction, ce changement ne sera pas visible localement dans la dynamique asynchrone. On introduit ainsi une condition supplémentaire pour pouvoir conclure à la présence d'une interaction.

Définition 15 *Etant donnés $F : \mathbb{Z}^n \to \mathbb{Z}^n$, $x \in \mathbb{Z}^n$ et $\varepsilon \in \{-1,+1\}^n$, la* matrice jacobienne discrète *évaluée à l'état x avec le vecteur de variation ε, notée $J_F^\varepsilon(x)$, est la matrice booléenne $n \times n$ définie par :*

$$J_F^\varepsilon(x)_{ij} = \begin{cases} 1 & si \quad F_i(x) < x_i + \varepsilon_i/2 < F_i(x^{j \lhd \varepsilon_j}) \quad ou \quad F_i(x^{j \lhd \varepsilon_j}) < x_i + \varepsilon_i/2 < F_i(x) \\ 0 & sinon. \end{cases}$$

Pour avoir $J_F^\varepsilon(x)_{ij} = 1$, il est nécessaire que la variation de j entraîne un déplacement du point focal $F_i(.)$ visible localement : les deux points focaux doivent être de part et d'autre du « seuil » qui sépare x_i et $x_i + \varepsilon_i$. Pour les systèmes booléens, cette définition est équivalente à celle de la jacobienne booléenne. Pour éviter la manipulation de la valeur $\varepsilon_i/2$, nous introduisons la définition équivalente suivante.

Définition 16 *Etant donnés $F : \mathbb{Z}^n \to \mathbb{Z}^n$, $x \in \mathbb{Z}^n$ et $\varepsilon \in \{-1,+1\}^n$, la* matrice jacobienne discrète *$J_F^\varepsilon(x)$ est la matrice booléenne $n \times n$ définie par :*

$$J_F^\varepsilon(x)_{ij} = 1 \iff \begin{cases} (\alpha) \;\; F_i(x) \le x_i < F_i(x^{j \lhd \varepsilon_j}) \quad ou \quad (\beta) \;\; F_i(x^{j \lhd \varepsilon_j}) \le x_i < F_i(x) \quad si\ \varepsilon_i = +1 \\ (\gamma) \;\; F_i(x) < x_i \le F_i(x^{j \lhd \varepsilon_j}) \quad ou \quad (\delta) \;\; F_i(x^{j \lhd \varepsilon_j}) < x_i \le F_i(x) \quad si\ \varepsilon_i = -1 \end{cases}$$

Pour associer un signe à l'interaction $(j \longrightarrow i)$ si $J_F^\varepsilon(x)_{ij} = 1$, remarquons que dans les cas (α) et (δ), j a un effet positif sur i puisqu'une augmentation (resp. diminution) du niveau de j induit une augmentation (resp. diminution) de la coordonnée i du point focal. Dans les cas (β) et (γ) la situation est inversée et j a un effet négatif sur i. Ainsi le signe de l'interaction de j sur i est donné par

$$\varepsilon_j \cdot sign(F_i(x^{j \lhd \varepsilon_j}) - F_i(x)) \qquad \text{où} \qquad sign(a) = +1 \;\text{ si }\; a > 0 \;\text{ et }\; sign(a) = -1 \;\text{ si }\; a < 0.$$

Le graphe d'interactions basé sur la jacobienne $J_F^\varepsilon(x)$ est alors défini comme suit.

Définition 17 *Etant donnés $F : \mathbb{Z}^n \to \mathbb{Z}^n$, $x \in \mathbb{Z}^n$ et $\varepsilon \in \{-1,+1\}^n$, le* graphe d'interactions local *évalué à l'état x avec le vecteur de variations ε, noté $G_F^\varepsilon(x)$, est le graphe orienté signé sur les sommets $\{1,\dots,n\}$ dont l'ensemble des arcs est donné par :*

$$\{(j,\alpha,i) \mid i,j \in \{1,\dots,n\}, J_F^\varepsilon(x)_{ij} = 1 \;\; et \; \alpha = \varepsilon_j \cdot sign(F_i(x^{j \lhd \varepsilon_j}) - F_i(x))\}.$$

En suivant E. Rémy, P. Ruet et D. Thieffry [89], nous définissons le *graphe d'interactions global $G(F)$ d'un système dynamique discret $F : \mathbb{Z}^n \to \mathbb{Z}^n$* comme le graphe sur les sommets $\{1,\dots,n\}$ dont l'ensemble des arcs est l'union des ensembles des arcs de tous les graphes d'interactions locaux.

5.3.2 Théorème sur l'existence d'un chemin de longueur minimale

Nous présentons ici un théorème qui donne une condition suffisante pour la présence d'un chemin de longueur minimale reliant deux états de la dynamique asynchrone d'un système dynamique discret. Ce théorème est fondamental pour la preuve de la conjecture de R. Thomas concernant la multi-stationnarité puisqu'il permet de déduire sous les conditions du théorème l'existence d'un chemin entre deux états.

Rappelons que la distance Manhattan d est définie sur $\mathbb{Z}^n$ par : $\forall x,y \in \mathbb{Z}^n$, $d(x,y) = \sum_{i=1}^n |x_i - y_i|$. Un chemin de x à y dans la dynamique asynchrone est nécessairement de longueur supérieure ou égale à $d(x,y)$. Si cette borne est atteinte, le chemin est dit *chemin de longueur minimale*. Comme les états d'un plus court chemin de x à y appartiennent au pavé $\pi(x,y) = \prod_{i=1}^n \{\min(x_i,y_i),\dots,\max(x_i,y_i)\}$, nous allons nous restreindre à l'étude de ce pavé. Nous dirons

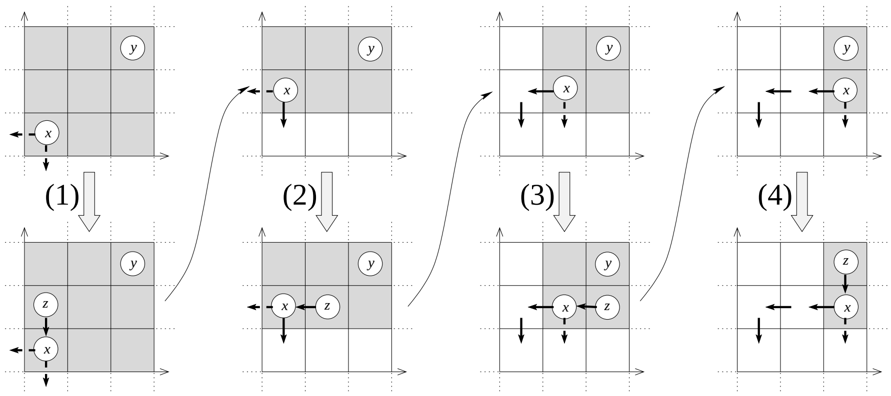


FIG. 5.3 – *Illustration du théorème du chemin de longueur minimale. La taille de* $\pi(x, y)$ *(en gris) est réduite à chacune des étapes* (1), (2), (3) *et* (4)*. Les flèches en pointillé partant de* x *sont rajoutées pour illustrer la* $\pi(x, y)$*-stabilité de* x*.*

que x est $\pi(x, y)$-stable si aucun des successeurs de x dans la dynamique asynchrone n'appartient à $\pi(x, y) \setminus \{x\}$. Si x est un point fixe de F alors x est $\pi(x, y)$-stable pour tout $y \in \mathbb{Z}^n$. Finalement, notons par $G_F(x, y)$ le sous graphe d'interactions de $G_F^{\varepsilon}(x)$, où $\varepsilon_i = sign(x_i, y_i)$ pour tout i tel que $x_i \neq y_i$, ne contenant que les gènes l tels que $x_l \neq y_l$.

Les deux lemmes suivants jouent un rôle crucial dans la suite. Leurs démonstrations sont publiées dans la thèse d'Adrien Richard [85] : celle du premier s'appuie sur les travaux d'E. Rémy, P. Ruet et D. Thieffry [89] et de M.-H. Shi et J.-L. Dong [90] et celle du deuxième découle d'un jeu d'écriture.

Lemme 5.3.1 *Soit* $F : \mathbb{Z}^n \to \mathbb{Z}^n$ *et* $x, y \in \mathbb{Z}^n$ *avec* $x \neq y$*. Si* x *est* $\pi(x, y)$*-stable et si* $G_F(x, y)$ *n'a pas de circuit positif, alors* $G_F(x, y)$ *n'a pas de circuit.*

Lemme 5.3.2 *Soit* $F : \mathbb{Z}^n \to \mathbb{Z}^n$ *et* $x, y \in \mathbb{Z}^n$ *avec* $x \neq y$*. Si* x *est* $\pi(x, y)$*-stable et si* $G_F(x, y)$ *n'a pas de circuit, alors il existe* $z \in \pi(x, y)$ *tel que* $(z, x) \in \mathcal{A}(F)$ *et* z *est* $\pi(z, y)$*-stable.*

Nous pouvons maintenant énoncer le théorème du chemin de longueur minimale.

Théorème 5.3.1 (Théorème du chemin de longueur minimale) *Soit* $F : \mathbb{Z}^n \to \mathbb{Z}^n$ *et* $x, y \in \mathbb{Z}^n$ *avec* $x \neq y$*. Si* x *est* $\pi(x, y)$*-stable et si, pour tout* $z \in \pi(x, y)$*,* $G_F(z, y)$ *n'a pas de circuit positif, alors il y a un chemin de longueur minimale de* y *à* x *dans* $\mathcal{A}(F)$*.*

La preuve repose sur l'itération des deux lemmes précédents. Dans les conditions du théorème, le lemme 5.3.1 permet d'assurer l'absence de circuit dans $G_F(x, y)$ et le lemme 5.3.2 permet d'affirmer qu'il existe $z \in \pi(x, y)$ tel que $(z, x) \in \mathcal{A}(F)$ et z est $\pi(z, y)$-stable. On peut alors itérer l'utilisation de ces deux lemmes à partir de z jusqu'à arriver en y (cf. figure 5.3). Comme cette preuve est constructive, elle donne un algorithme pour trouver un chemin de longueur minimale de y à x. A partir de x, on détermine un nœud j de $G_F(x, y)$ qui n'a pas de successeur. L'état $z = x^{j \lhd sign(x_j, y_j)}$ est le prédécesseur de x dans un chemin de longueur minimale possible. On réitère la procédure à partir de z jusqu'à atteindre y.

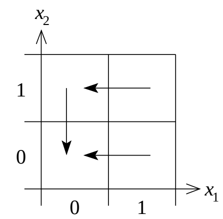
5.3.3 Condition nécessaire pour avoir multi-attractivité

Le théorème du chemin de longueur minimale permet de prouver la première conjecture de R. Thomas dans le cadre des systèmes dynamiques discrets.

(a)

x	$F(x)$
$(0,0)$	$(0,0)$
$(0,1)$	$(0,0)$
$(1,0)$	$(0,0)$
$(1,1)$	$(0,1)$

(b)



(c)

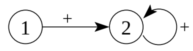


FIG. 5.4 – *(a) Application $F : \{0,1\}^2 \to \{0,1\}^2$. (b) dynamique asynchrone associée dont les domaines sans issus sont : $\{(0,0)\}$, $\{(0,0),(0,1)\}$, $\{(0,0),(1,0)\}$ et $\{(0,0),(0,1),(1,0),(1,1)\}$. (c) Le graphe d'interactions local $G_F((1,1),(0,0))$ contient un circuit positif, sans qu'il existe deux domaines sans issue d'intersection vide.*

Théorème 5.3.2 *Soit $F : \mathbb{Z}^n \to \mathbb{Z}^n$. Supposons que la dynamique asynchrone de F possède deux domaines sans issue disjoints X et Y et soit (x,y) un élément minimal de l'ensemble $X \times Y$ muni de la relation d'ordre partiel $(x,y) \leq (x',y')$ si $\pi(x,y) \subset \pi(x',y')$. Il existe alors $z \in \pi(x,y)$ tel que $G_F(z,y)$ contient un circuit positif.*

Preuve : Prouvons d'abord que sous les conditions du théorème, x est $\pi(x,y)$-stable. Supposons que x ne soit pas $\pi(x,y)$-stable. Alors il existe $z \in \pi(x,y) \setminus \{x\}$ tel que $(x,z) \in \mathcal{A}(F)$ et $z \in X$ (par définition d'un domaine sans issue). Comme $\pi(z,y) \subsetneq \pi(x,y)$ on a $|\pi(x,y)| > |\pi(z,y)|$. Contradiction.

Supposons maintenant que pour tout $z \in \pi(x,y)$, $G_F(z,y)$ ne contient pas de circuit positif. Le théorème du chemin dit qu'il existe un chemin de y à x, ainsi $x \in Y$. Contradiction. □

Ce résultat est particulièrement intéressant puisque la coexistence de deux domaines sans issue disjoints implique la présence d'un circuit positif dans un graphe *local*. De plus, le théorème donne des informations d'une part sur la localisation de ce graphe local, et d'autre part sur les gènes impliqués dans ce circuit (ils appartiennent à l'ensemble des gènes i tels que $x_i \neq y_i$). Ce théorème est une généralisation de la conjecture de R. Thomas : si x et y sont deux points fixes distincts de F alors $\{x\}$ et $\{y\}$ sont deux domaines sans issue disjoints, et la présence d'un circuit positif dans un graphe *local* est assurée. Cette condition nécessaire est plus forte que celle proposée par R. Thomas puisque la présence d'un circuit positif dans un graphe *local* implique la présence d'un circuit positif dans le graphe *global*. Notons cependant que la réciproque de ce théorème n'est pas vraie (figure 5.4).

5.4 Conclusion

La matrice jacobienne que nous avons introduite dépend non seulement de l'état du système x mais aussi d'un vecteur de variations élémentaires ε. Si on place un observateur sur l'état singulier voisin de l'état courant x en direction de $x+\varepsilon$, il est possible de relier la valeur des éléments de la jacobienne à ce que cet observateur peut distinguer. Cet observateur a une capacité d'observation limitée : il compare deux états z_1 et z_2 et ne peut dire qu'ils sont différents que si pour au moins une coordonnée, il est lui-même au milieu de ces deux états. L'élément ij de la jacobienne (voir définition 15) est égal à 1 si d'une part x_j et $x_j + \varepsilon_j$ sont de part et d'autre de l'observateur (évident car l'observateur est positionné entre x_j et $x_j + \varepsilon_j$) et si les $F_i(x)$ et $F_i(x^{j \lhd \varepsilon_j})$ sont aussi de part et d'autre de l'observateur. Cet observateur voit si le changement de la coordonnée j entraîne un changement de la coordonnée i du point focal.

Le travail de modélisation de ce chapitre se concentre essentiellement sur la définition d'un *graphe d'interactions local* associé à un état ou encore associé à un pont d'un cycle sans issue, pour une dynamique asynchrone donnée. La définition des matrices jacobiennes a permis d'expliciter comment une interaction peut être vue à travers la dynamique (voir notion d'observateur dans la dynamique) et surtout comment interpréter une dynamique en termes d'interactions. La notion de coude a permis de se focaliser sur les états où les interactions sont visibles.

Les conséquences de cette définition sont importantes car certains liens entre dynamique et topologie des interactions peuvent être faits. La présence dans la dynamique d'un cycle sans issue (resp. de deux domaines sans issue disjoints) n'est possible que si dans le (resp. un) graphe d'interactions global (resp. local) un circuit négatif (resp. positif) est présent. Le résultat pour la multi-stationnarité est finalement plus fort puisque l'existence du circuit positif est assurée dans un graphe *local* et que le théorème donne une information sur la localisation de ce graphe.

Conclusion

La biologie moléculaire a ouvert la voie à une biologie des systèmes qui repose essentiellement sur des questions de reverse engineering puisqu'il s'agit de comprendre le fonctionnement d'un système donné, de prédire son fonctionnement dans certaines conditions, et de déterminer où et comment modifier le système pour qu'il satisfasse certaines spécifications. Ces questions ne pourront pas être traitées sans la construction de théories informatiques adaptées à la biologie des systèmes. Mon parcours scientifique tend à montrer au travers certaines questions précises comme la comparaison de séquences, l'analyse du transcriptome ou la modélisation des réseaux génétiques, que l'informatique théorique est nécessaire à la biologie des systèmes. La modélisation dans son acception informatique est cruciale pour chacune des questions traitées, laissant penser qu'elle sera aussi nécessaire pour de nombreux autres aspects de la biologie des systèmes. La modélisation constitue ainsi le maître-mot de mon parcours scientifique.

- *Modéliser pour générer des connaissances.* Lors de la comparaison de séquences biologiques, la significativité d'un score d'alignement est fondamentale pour décider s'il y a ou non un lien biologique entre deux séquences. Pour estimer cette significativité le Z-score fait appel à un modèle de séquences aléatoires (séquences indépendantes, identiquement distribuées de même longueur et même composition qu'une séquence réelle). En reprenant la modélisation du score d'alignement de M.S. Waterman et M. Vingron, nous donnons une modélisation du Z-score : quelles que soient les deux séquences à comparer, le Z-score suit une loi de Gumbel de paramètres fixés.
- *Modéliser pour exprimer des liens potentiels entre transcriptome et comportement macroscopique.* Dans l'étude du transcriptome, la modélisation du niveau d'expression des gènes (modèle ANOVA) qui sépare l'effet de la classe des effets individuels permet d'envisager une méthode simple pour trouver les gènes témoins de la différence de classes.
- *Modéliser pour comprendre.* La compréhension des processus biologiques complexes nécessite le développement de modèles qui combinent des informations quantitatives avec des détails moléculaires. Le modèle numérique résultant, est basé sur les données de la littérature lorsqu'elles sont disponibles. Ces modèles peuvent alors être utilisés pour faire des prédictions qui peuvent à leur tour être confrontées à des expériences.
- *Modéliser pour automatiser partiellement la construction d'un modèle.* Lorsque le biologiste construit un « modèle mathématique » d'un réseau de régulation génétique, il construit une abstraction du système d'étude comprenant non seulement les nœuds du réseau, mais aussi l'ensemble des valeurs des paramètres associés permettant de construire de manière non ambiguë la dynamique. La détermination de ces paramètres constitue alors un point clef de l'activité de modélisation. Cette phase doit prendre en compte les connaissances biologiques disponibles sur le système, et elle recherche de manière exhaustive l'ensemble des valeurs de paramètres permettant de construire une dynamique cohérente avec ces connaissances. Cette recherche exhaustive peut être assistée par ordinateur.
- *Modéliser pour énoncer des règles générales.* La modélisation donne un cadre théorique dans lequel il est possible de raisonner. L'exemple de la théorie de R. Thomas sur les

réseaux de régulation génétique nous a donné l'occasion de confronter la modélisation à des propriétés générales énoncées sur l'ensemble des modèles. Nous avons donné une preuve mathématique des conjectures de R. Thomas portant sur le lien entre le système de transitions représentant la dynamique et la représentation des interactions.

Cycle de modélisation en biologie

Pour approcher un modèle valide, les modélisateurs utilisent souvent une technique itérative inspirée du cycle de développement en spiral du génie logiciel. Un cycle de modélisation (voir figure 5.5) commence par la sélection d'un ensemble de connaissances biologiques, comme par exemple les gènes impliqués dans le système, éventuellement l'ordre de certains seuils d'interactions et des propriétés comportementales. A partir de ces connaissances, on construit l'ensemble de tous les modèles cohérents. Pour chacun des modèles, on prédit par simulation leurs comportements et on élimine les modèles qui ne sont pas en accord avec des propriétés biologiques connues. Les modèles qui passent cette étape sont de bons candidats pour une analyse plus précise. A partir des prédictions faites sur l'ensemble des modèles survivants, on peut extraire un schéma d'expériences dont le résultat *in vivo* va départager l'ensemble de ces modèles : ceux qui « réagissent » comme le système vivant (ils possèdent les mêmes propriétés expérimentales) et les autres. Un schéma d'expérience est choisi et mis en œuvre *in vivo*. Les modèles qui survivent à ce cycle sont jugés consistants avec les connaissances courantes, et le cycle peut être rejoué en incorporant de nouvelles connaissances.

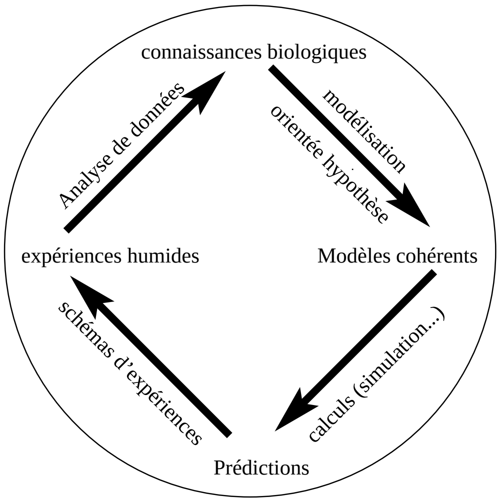


FIG. 5.5 – *Cycle de développement de modèles*

Même s'il s'agit d'un processus idéal, on peut raisonnablement penser que les avancées en modélisation, en méthodes formelles, simulation, technologies de mesures et en génomique, vont transformer graduellement la recherche en biologie pour se rapprocher de ce cycle. La biologie sera alors une science orientée hypothèses où les hypothèses faites à une étape donnée guident fortement le chercheur pour l'étape suivante.

Méthodes formelles et modélisation de systèmes biologiques

Les approches nouvelles en biologie regroupées sous le terme de biologie des systèmes visent à comprendre le fonctionnement des objets biologiques (organelles, cellules, organes, organismes, populations) comme un système. Dans cette quête, l'analyse des séquences participe à la détermination des gènes ainsi qu'à l'inférence par homologie de leurs fonctions hypothétiques, le transcriptome permet de sélectionner des gènes d'intérêt pour l'étude de réponses à des stress particuliers. Mais les études du transcriptome ont montré que les systèmes biologiques étaient des *systèmes complexes* au sens où leur fonctionnement global ne peut être prédit de manière simple par une décomposition raisonnée de leurs parties. Pour les réseaux de régulation génétique, l'ajout d'un gène, et *a fortiori* la fusion de deux réseaux de régulation, entraîne généralement des changements non prévisibles de comportement. L'approche incrémentale pour étudier des grands réseaux ne semble donc pas adaptée, et de par l'explosion combinatoire du nombre de modèles, il est déraisonnable d'envisager une approche de type force brute comme celle décrite au chapitre 4. Ainsi la complexité du vivant provient aussi de l'incapacité à utiliser des méthodes de décomposition hiérarchique. S'agit-il d'un contrôle distribué ? comment pourrait se faire la transmission de ce contrôle ? Quelles seraient les caractéristiques topologiques de ces grands graphes ? Toutes ces questions rendent difficiles les prédictions en biologie de l'effet de la modification de l'un des composants voire des liens entre composants.

Cependant, toutes les études font apparaître des réseaux de régulation présentant des groupes de gènes coexprimés voire qui présentent un niveau de connectivité significativement supérieur à la valeur moyenne. Ces systèmes sont donc modulaires dans un certain sens. En reprenant la terminologie de M.J. Stuart [103], ces groupes sont appelés modules et ces modules sont confirmés lorsque les fonctions associées à ces gènes sont biologiquement rattachées. Le modélisateur est donc poussé à mettre en œuvre une approche « compositionnelle » dans laquelle les sous-réseaux liés à des « fonctions biologiques » identifiables sont en premier lieu modélisés puis « composés » suivant certaines règles inspirées par la biologie. La question principale réside dans la définition de ces règles afin de rendre compte du fonctionnement global du système à partir du fonctionnement de ses entités.

L'acception informatique du mot « module » ne peut pas ici être utilisée parce qu'un gène peut participer à plusieurs fonctions biologiques, et donc plusieurs modules. Cette particularité des systèmes biologiques est fondamentale, c'est le « bricolage » de l'évolution. Ainsi les méthodes de l'informatique, plus particulièrement du génie logiciel à base de réutilisations, d'adaptations, etc. sont des schémas insuffisants dans le contexte de la modélisation en biologie. Pourtant, les questions de sûreté de fonctionnement, optimisations de redondances, etc. en informatique sont des domaines de recherches fertiles, et on peut se demander si les réponses apportées par le génie logiciel ne pourraient pas être des pistes pour la modélisation en biologie.

Nous avons montré dans ce mémoire que ce rapprochement entre méthodes formelles et bio-informatique était porteur pour l'étude des réseaux de régulation génétique et pour la formulation des propriétés dynamiques du système biologique. Ainsi on peut penser que les aspects formels du génie logiciel pourraient être bien adaptés pour faire émerger de nouvelles méthodes pour structurer et analyser ces objets biologiques complexes. Même si chacun des deux domaines se distingue par des éléments de bases bien spécifiques, les notions de propriétés du système sont partagées, qu'elles soient émergeantes ou invariantes. Le génie logiciel s'intéresse à la méthodologie pour assurer un développement de logiciels conformes à leurs spécifications alors que la biologie s'intéresse à la compréhension des propriétés d'un système *a priori* non modifiable.

Richesse du domaine

Le domaine de la formalisation des réseaux de régulation génétique ouvre de nouvelles perspectives de recherche, que cela concerne l'abstraction des trajectoires, la représentation d'information de type coopération ou concurrence, ou encore la recherche exhaustive des modèles qui satisfont une spécification donnée.

Dans la modélisation de R. Thomas, les trajectoires continues sont abstraites par des dynamiques discrètes, qui ne prennent en compte aucun aspect temporel quantitatif, seule la succession des états qualitatifs est représentée. Ainsi, il n'est pas facile de proposer des spécifications du type : *le système présente un comportement cyclique stable de période 24h*. D'un côté, pour enrichir le pouvoir d'expression des spécifications de la sorte, il est nécessaire de raffiner le modèle de R. Thomas en proposant par exemple des trajectoires continues dans l'espace des phases des concentrations. De l'autre côté, il faut choisir une modélisation qui soit compatible avec une procédure de model-checking. Ainsi, nous collaborons avec O. Roux de l'IRCCyN pour modéliser les réseaux de régulation à l'aide d'automates hybrides linéaires. Certaines des variables de ces automates sont discrètes, et d'autres continues. Les états de ces automates représentent les états qualitatifs, et lorsque le système se trouve dans un état particulier, les trajectoires des variables continues sont linéaires. Un premier raffinement consiste donc à approximer les trajectoires du système d'équations linéaires par morceaux par des lignes brisées, sur lesquelles les vitesses sont constantes. L'approximation est peut-être un peu grossière, mais elle permet une vérification automatique de propriétés par le model-checker `HyTech` [4, 49].

Un deuxième raffinement possible consiste à découper les domaines abstraits en sous-domaines en fonction des trajectoires futures. En effet, dans le modèle de R. Thomas, l'abstraction des trajectoires continues introduit un non-déterminisme. Ces trajectoires continues peuvent cependant servir à subdiviser les états qualitatifs afin de proposer un modèle qualitatif raffiné sur lequel les méthodes de model-checking sont exploitables.

L'approche de modélisation itérative, qui nécessite de rechercher exhaustivement tous les modèles d'une certaine classe qui satisfont les spécifications, est sensible à l'explosion combinatoire du nombre de modèles à élaguer. Pour passer à l'échelle, il devient nécessaire de diminuer le nombre de modèles à vérifier ou de passer à des techniques de résolution plus symbolique.

La réduction du nombre de modèles à vérifier peut être faite avant même toute vérification si une description plus précise des interactions entre gènes est disponible. L'introduction de graphes bipartis pourrait permettre de prendre en compte très rapidement dans le processus de modélisation, des informations de type coopération/concurrence, informations codées dans les valeurs des paramètres de la modélisation de R. Thomas. Un type de sommets abstrait les variables du systèmes (gènes), et l'autre (les transitions) les processus régulatoires. Chaque transition peut avoir plusieurs prédécesseurs (les variables participant au même processus régulatoire) et successeurs (variables contrôlées par le processus considéré). Par exemple, l'activation d'un opéron qui comporte les trois gènes a, b et c, peut être modélisée par une transition unique ayant comme prédécesseur le régulateur de l'opéron et trois successeurs (a, b et c). De plus l'information sur le processus régulatoire du type « le processus est effectif uniquement sous certaines conditions », peut être rajoutée dans la transition sous forme logique. La diminution du nombre de modèles (valuations possibles des paramètres) dépend de l'information que l'on code dans les transitions. Ces informations de coopération / concurrence ne supprimeront malheureusement pas la vérification systématique d'un grand ensemble de modèles qui restera le goulot d'étranglement à cause du grand nombre d'appels au model-checker.

Des techniques symboliques pourraient aussi être envisagées car les modèles présentent des caractéristiques communes. Pour cela, on doit représenter l'ensemble de modèles par une seule structure plus riche, et développer un algorithme de model-checking sur cette structure. L'idée générale consiste à étiqueter chaque transition possible par l'ensemble des modèles qui partagent cette transition. La sémantique d'une formule dans un état donné est alors définie comme étant le sous ensemble des modèles qui satisfont à cet état donné la formule temporelle. Ce multi-modèle de Kripke est très général, car il nécessite simplement que les états des différents modèles soient en correspondance et partagent les mêmes étiquettes. Dans le cas des modèles issus de la modélisation de R. Thomas, bien d'autres informations sont disponibles. En particulier, le changement d'une des valeurs des paramètres n'entraîne que des changements locaux dans la dynamique associée. On pourrait donc envisager une classe de *systèmes de transitions symboliques* dans laquelle, les transitions existent sous certaines *contraintes* sur les paramètres du système. L'algorithme de model-checking pourrait alors calculer les états qui satisfont les spécifications sous certaines contraintes sur les paramètres. La procédure de model-checking d'une telle structure symbolique fait donc intervenir des phases de construction de contraintes et des phases de résolution de contraintes.

Vers une biologie formelle

Pour arriver à assister la biologie dans la compréhension des systèmes biologiques, la modélisation en biologie peut être représentée en plusieurs niveaux de modélisation. Chacun permet de se focaliser sur un niveau d'abstraction.

1. Le premier niveau est le *niveau sémantique* dont le but est la représentation de la dynamique du système que l'on cherche à modéliser. Elle se focalise sur les traces dans l'espace des phases que le système peut emprunter.
2. Le *niveau formel* se base sur le niveau sémantique : il permet de représenter le niveau sémantique sous forme syntaxique, ainsi interprétable et manipulable par un ordinateur. Nous avons montré que pour les réseaux de régulation génétique, ce niveau syntaxique était à même de proposer des outils de vérification/validation de propriétés temporelles représentant soit des connaissances biologiques soit des hypothèses biologiques. L'utilisation intensive de cet outil de vérification, voire le développement d'outils symboliques, peut aussi permettre d'aider à préciser les valeurs de certains paramètres.
3. Le *niveau fondamental* se focalise sur les objets biologiques et les propriétés des systèmes. Le passage du niveau formel au niveau fondamental fait le lien entre les objets biologiques, leurs propriétés et leurs représentations formelles. Par exemple, pour les réseaux de régulation génétique, le modèle de R. Thomas appartient au niveau formel, alors que les conjectures de R. Thomas (condition nécessaire de multi-stationnarité/homéostasie) relève du niveau fondamental.

Le travail présenté sur les réseaux de régulation génétique qui a motivé ce découpage en niveaux, incite à penser que la biologie peut gagner beaucoup en s'ouvrant sur les niveaux formel et fondamental.

Annexe scientifique

```
/* Les paramètres                    */
/* kx0  = 0        ky0 = 0           */
/* kxx  = 2        kyx = 1           */
/* kxy  = 1                          */
/* kxxy = 2                          */
/***********************************/
kx([],   0).
kx([x],  2).
kx([y],  1).
kx([x,y],2).

ky([],   0).
ky([x],  1).

/* calcul des ressources pour un état donné */
/*********************************************/
ressources(x,X,Y,[])    :-  X <   2,  Y =:= 1.
ressources(x,X,Y,[x])   :-  X >=  2,  Y =:= 1.
ressources(x,X,Y,[y])   :-  X <   2,  Y =:= 0.
ressources(x,X,Y,[x,y]) :-  X >=  2,  Y =:= 0.
ressources(y,X,Y,[])    :-  X =:= 0.
ressources(y,X,Y,[x])   :-  X >=  1.

suivant((X,Y),(NewX,Y)):- NewX is X+1,
      ressources(x,X,Y,Rx), kx(Rx,Kx), Kx > X.
suivant((X,Y),(NewX,Y)):- NewX is X-1,
      ressources(x,X,Y,Rx), kx(Rx,Kx), Kx < X.
suivant((X,Y),(X,NewY)):- NewY is Y+1,
      ressources(y,X,Y,Ry), ky(Ry,Ky), Ky > Y.
suivant((X,Y),(X,NewY)):- NewY is Y-1,
      ressources(y,X,Y,Ry), ky(Ry,Ky), Ky < Y.
suivant((X,Y),(X,Y)):-
      ressources(x,X,Y,Rx), kx(Rx,Kx), Kx =:= X,
      ressources(y,X,Y,Ry), ky(Ry,Ky), Ky =:= Y.

atteignable((X,Y),(X,Y)).
atteignable((X,Y),(XX,YY)) :-
    suivant((X,Y),(NX,NY)), atteignable((NX,NY),(XX,YY)).

/* formule : x=2 => EX EF(x=2) */
formule1(X,Y) :- X=\=2.
formule1(X,Y) :- X=:=2, suivant((X,Y),(NX,NY)),
                 atteignable((NX,NY),(XX,YY)), XX=:=2.
```

FIG. 1 – *Code prolog décrivant la dynamique d'un réseau de régulation génétique comportant 2 gènes ($x \xrightarrow{+,1} y$, $x \xrightarrow{+,2} x$, $x \xrightarrow{-,1} y$). Le dernier prédicat permet de verifier si un état du système satisfait la formule CTL $(x = 2) \Longrightarrow EXEF(x = 2)$.*

```
/* le graphe est traduit par les ressources */
/*********************************************/
ressources(x,X,Y,[])    :-  X #<   2,  Y #= 1.
ressources(x,X,Y,[x])   :-  X #>=  2,  Y #= 1.
ressources(x,X,Y,[y])   :-  X #<   2,  Y #= 0.
ressources(x,X,Y,[x,y]) :-  X #>=  2,  Y #= 0.
ressources(y,X,Y,[])    :-  X #= 0.
ressources(y,X,Y,[x])   :-  X #>=  1.

kx(L,[]   ,K) :- nth(1, L, K).
kx(L,[x]  ,K) :- nth(2, L, K).
kx(L,[y]  ,K) :- nth(3, L, K).
kx(L,[x,y],K) :- nth(4, L, K).
ky(L,[]   ,K) :- nth(5, L, K).
ky(L,[x]  ,K) :- nth(6, L, K).

/* calcul des successeurs */
/**************************/
suivant(Ks,(X,Y),(NewX,Y)):- NewX #= X+1,
    ressources(x,X,Y,Rx), kx(Ks,Rx,Kx), Kx #> X.
suivant(Ks,(X,Y),(NewX,Y)):- NewX #= X-1,
    ressources(x,X,Y,Rx), kx(Ks,Rx,Kx), Kx #< X.
suivant(Ks,(X,Y),(X,NewY)):- NewY #= Y+1,
    ressources(y,X,Y,Ry), ky(Ks,Ry,Ky), Ky #> Y.
suivant(Ks,(X,Y),(X,NewY)):- NewY #= Y-1,
    ressources(y,X,Y,Ry), ky(Ks,Ry,Ky), Ky #< Y.
suivant(Ks,(X,Y),(X,Y)):-
     ressources(x,X,Y,Rx), kx(Ks,Rx,Kx), Kx #= X,
     ressources(y,X,Y,Ry), ky(Ks,Ry,Ky), Ky #= Y.

lesKs(Ks) :-
    Ks = [Kx0,Kxx,Kxy,Kxxy,Ky0,Kyx],
    fd_domain([Kx0,Kxx,Kxy,Kxxy],0,2),
    fd_domain([Ky0,Kyx],0,1),
    suivant(Ks,(0,0),(1,0)), suivant(Ks,(1,0),(1,1)),
    suivant(Ks,(1,1),(0,1)), suivant(Ks,(0,1),(0,0)),
    suivant(Ks,(2,0),(2,1)), suivant(Ks,(2,1),(2,1)),
    fd_labeling(Ks).

find_all(L,LL) :- findall(Ks,lesKs(Ks), L), length(L,LL).
```

FIG. 2 – *Code Gnu Prolog naïf pour la détermination des paramètres menant à une dynamique qui présente certaines transitions connues.*

Curriculum Vitae

1 Situation

Depuis février 2000, je suis maître de conférences à l'Université d'Evry-Val d'Essonne. J'effectue mes recherches au sein de l'équipe *Réseaux et Méthodes Formelles* dirigée par P. Le Gall au laboratoire IBISC F.R.E 2873 CNRS. Auparavant j'étais dans l'équipe *Bioinfo* dirigée par G. Bernot au LaMI.

Depuis fév. 2000	Maître de conférences à l'Université d'Evry-Val d'Essonne. Laboratoire de rattachement : Informatique, Biologie Intégrative et Systèmes Complexes, FRE CNRS 2873 (IBISC), équipe *Réseaux et Méthodes Formelles.*
2004–2006	Deux ans de *délégation CNRS* du 1er septembre 2004 au 1er septembre 2006. Laboratoire de rattachement : IBISC, FRE CNRS 2873.
1999-2000	Stage post-doctoral en bioinformatique au *Whitehead Institute/MIT Center for Genome Research*, Cambridge (MA) USA. Axes de recherche : Modélisation de l'intensité observée sur les puces à ADN, prise en compte de l'erreur de mesure dans les méthodes de prédiction.
1995-1998	Doctorat d'Informatique à l'Université de Technologie de Compiègne, laboratoire d'accueil : INRIA Rocquencourt. Directeur de thèse : J. Henry. Titre : *Programmation dynamique et alignements de séquences biologiques.*
1994	Ingénieur au CNETI, groupe des Caisses d'Epargne.
1992-1993	Service national en tant que scientifique du contingent au CEA/DAM.
Juin 1992	D.E.A. de *Statistiques, Probabilités et Informatique* à l'université de Rouen. Titre : *Optimisation du choix du paquet d'ondelettes et du filtre de décomposition* sous la direction d'E. Lenglart.
	Diplôme d'ingénieur mathématicien : option *statistiques et recherche opérationnelle.* Institut National des Sciences Appliquées de Rouen.

2 Thématiques de Recherche

Alignements de séquences biologiques.	1995–2001
Exploitation des données d'expression de gènes.	1999–2004
Modélisation des réseaux de régulation.	2002–

3 Encadrement d'étudiants en 3ème cycle

3.1 Thèses soutenues

1. *Modèle formel pour les réseaux de régulation génétique et influence des circuits de rétroaction.* Adrien Richard. Encadrement : G. Bernot (10%) et moi-même (90%). Soutenue le 20/09/2006.

3.2 Thèses en cours

1. *Modélisation des réseaux de régulation génétique : utilisation des "Extended Labelled Transition Systems"*. Daniel Mateus, CEA Saclay. Encadrement : P. Le Gall (50%), moi-même (50%). Thèse commencée en décembre 2004.
2. *Modélisation géométrique de processus biologiques*. Mathieu Poudret. Encadrement : P. Le Gall (20%), A. Arnould (50%) et moi-même (30%). Thèse commencée en septembre 2005.
3. *Réseaux de Petri hybrides temporisés, réseaux de régulation génétique et vérification*. Sylvie Troncale. Encadrement : G. Bernot (50%) et moi-même (50%). Thèse commencée en septembre 2005.

3.3 Etudiants en DEA/Master 2 Recherche

1. *Analyse combinée des réseaux de régulation biologique par contraintes et model-checking*, J. Fromentin (Juin 2006). Master 2 Recherche IMBI (Informatique et Mathématiques pour la Biologie Intégrative). Encadrement : P. Le Gall (33%), O. Roux (33%) et moi-même (33%).
2. *Classification non supervisée de modèles discrets de réseaux génétiques*, C. Damon (septembre 2005). DEA Mathématiques, Vision et Apprentissage, ENS Cachan. Encadrement : F. d'Alché-Buc (50%) et moi-même (50%).
3. *Modélisation des réseaux de régulation génétique par des graphes bipartis*, Dorian Girin (juin 2005). DEA Application des Mathématiques et de l'Informatique à la Biologie (AMIB), Evry. Encadrement à 100%.
4. *Intégration des aspects temporels dans l'analyse des réseaux de régulation*, H. Hanen (Juin 2004). DEA AMIB, Evry. Encadrement : F. Delaplace (50%) et moi-même (50%).
5. *Techniques de Tests pour proposer des plans d'expériences biologiques*, X. Goffin (Juin 2004). DEA AMIB, Evry. Encadrement P. Le Gall (33%), G. Bernot (33%) et moi-même (33%). Travail présenté sous forme d'un poster à IPG'2004.
6. *Outils pour la modélisation du comportement qualitatif des réseaux de régulation génétique*, A. Richard (juin 2003). DEA AMIB, Evry. Encadrement à 100%.
7. *Une nouvelle sémantique pour les réseaux de régulation génétique*, C. Müller (juin 2003). DEA AMIB, Evry. Encadrement : F. Delaplace (50%) et moi-même (50%).
8. *Extension du modèle de R. Thomas pour l'exploitation des données issues des puces à cellules*, C. Auliac (juin 2003). DEA AMIB, Evry. Encadrement : V. Frouin (50%) du CEA/SGF et moi-même (50%).
9. *Un langage formel pour les réseaux de régulation génétique*, S. Pérès (Juin 2002). DEA AMIB, Evry. Encadrement : G. Bernot (25%) et moi-même (75%).
10. *Clustering des données du transcriptome : mise en œuvre de la méthode de Shamir et Sharan*, M. Van de Voorde (juin 2001). DEA AMIB, Evry. Encadrement à 100%.

4 Autres activités

4.1 Examination de thèse

J'ai été examinateur de la thèse d'Olivier Bastien, *Développements théoriques et méthodes numériques pour les analyses comparatives de génomes et protéomes biaisés, Application à la comparaison des génomes et protéomes de Plasmodium falciparum et d'Arabidopsis thaliana*, CEA/Universite Joseph Fourier, Grenoble, soutenue le 21 Avril 2006.

4.2 Réalisation de logiciels / transfert technologique

J'ai participé au développement de `SMBioNet`, une plateforme d'aide à la modélisation des réseaux de régulation génétique en utilisant le cadre de la théorie de R. Thomas. 6400 lignes, écrit en Java, appel au model checker NuSMV, 2 ans de développement (Adrien Richard est le principal développeur), diffusé sur le web (voir `http ://smbionet.lami.univ-evry.fr/`).

J'ai participé pendant ma thèse au code de LASSAP (LArge Scale Sequence compArison Package), logiciel de comparaison de séquences biologiques. Depuis, Gene-IT, société française de bioinformatique, a été créée sur le savoir-faire de LASSAP.

4.3 Participation à l'animation de la recherche

- Coordinateur (avec M. Kaufman, biologiste) d'un numéro spécial de TSI sur le thème : Modélisation et simulation pour la post-génomique, publication prévue en janvier 2007 (appel à contribution, comité de sélection, rédaction de l'éditorial).
- Organisation de deux journées thématiques sur les "Réseaux d'interaction : analyse, modélisation et simulation" (RIAMS 2005 et RIAMS 2006) à Lyon, journées satellites d'IPG. 30 novembre 2005 et 29 novembre 2006. `http ://riams.lami.univ- evry.fr` (mise en place du comité scientifique, recherche de financement, appel à contribution, sélection des articles, édition des actes).
- Edition d'un livre : *Modelling and simulation of biological processes in the context of genomics*, eds P. Amar, J.-P. Comet, F. Képès, V. Norris. Publisher Frontier group, ISBN : 2 84704 0374, juin 2004. Relance des auteurs pour obtenir leurs contributions, mise en page (210 pages).
- Organisation de la venue de R. Thomas (Bruxelles) au printemps 2005 grâce au soutien du « programme épigénomique ». Une semaine, animation des rencontres avec une dizaine de participants.
- Referee pour les revues *Bioinformatics*, *IJMMS*, *Transactions on Computational Biology and Bioinformatics*, pour les conférences internationales *International Colloquium on Theoretical Aspects of Computing (ICTAC 2006)*, *ISMB 2005*, *ISMB 2003*, *CMSB 2006*, *DNA Computing 2004*, pour la revue nationale *TSI*, et diverses conférences nationales.
- Membre du *Program Committee* de *ISMB'2005*, International conference on Intelligent Systems for Molecular Biology.
- Membre du comité local de l'école *Modélisation et simulation des processus biologiques*, génopole Evry, 2004 ; membre du comité local du *international symposium on macromolecular networks*, Institut Pasteur (Paris), 8-12 juillet 2002 ; membre du comité local du *workshop on regulatory proteins interplay and traffic on DNA*, Evry, France, 12-13 juillet 2002.

4.4 Activités d'Enseignement

J'interviens en cours et en TDs dans les filières rattachées au département de biologie : Licence SDV, Master 1 et 2 « Génie Biologique et Informatique », ainsi que dans des filières d'informatique : MIAGE, Licence, Master 2 Recherche IMBI.

Polycopiés de cours (disponible en ligne)

1. J.-P. Comet et A. Richard, Modélisation des réseaux de régulation génétique, DESS CCI, 23 p., fév. 2004.
2. J.-P. Comet. Introduction à l'Algorithmique pour la bioinformatique, IUP GBI Licence, 39 p., oct. 2003.

3. J.-P. Comet. Introduction à la programmation Prolog, Licence Informatique, 34 p., sept. 2003.
4. J.-P. Comet. Cours d'Algorithmique pour la bioinformatique, DESS GBI, 89 p., nov. 2002.

5 Responsabilités Collectives

– Co-animation avec J.-C. Aude du CEA de Saclay du groupe « G.T.Ex »(2000-2001) un groupe de travail sur la problématique de l'analyse des données d'expression.
– Responsable de la licence MIAGE (initiale et par apprentissage) de sept 2003 à sept 2005. 2 groupes de 30 dont un en apprentissage, sélection des étudiants sur dossiers (et entretiens pour l'apprentissage), rencontre des délégués tous les deux mois, discussions avec le CFA, organisation des jurys, président de jury.
– Responsable de l'enseignement d'Informatique de l'IUP « Génie Biologique et Informatique »d'Evry (2000-2003) : gestion de 450h d'enseignements d'informatique (co-gestion des emplois du temps, participation à toutes les soutenances de stages, organisation des réunions pédagogiques, participation à la création d'un DESS).
– Responsable de l'enseignement d'Informatique du DESS « Génie Biologique et Informatique »d'Evry (2001–2003) : participation à sa mise en place – programme des 350 heures équivalent TD d'enseignements d'informatique, gestion de 350h éq. TD d'enseignements d'informatique, jury de sélection (formation initiale et continue), co-gestion des emplois du temps, recherche des enseignants, etc.
– Participation à la rédaction de la plaquette LMD pour le Master IMBI (Informatique et Mathématiques pour la Biologie Intégrative).
– Membre du conseil de laboratoire (de sept 2000 à jan 2004 : responsable des réunions, de l'ordre du jour), membre de la Commission Matériel (2000–2002), membre du conseil de l'UFR Sciences Fondamentales et Appliquées (2002–2005), membre supplé- ant de la Commission de spécialistes (2002–2005).
– Membre de la Commission de spécialistes de l'université de Paris 11 (depuis 2005).

6 Publications et communications

Récapitulatif	**Total**	de 2002 à 2005	2006
Editions	**2**	1	1
Chapitres de livres	**5**	3	1
Revues internationales avec comité de lecture	**9**	6	1
Revues nationales avec comité de lecture	**1**	1	
Conférences internationales avec comité de lecture (avec actes)	**5**	4	1

Editions

1. J.-P. Comet and M. Kaufman, editors. *Modélisation et simulation pour la post-génomique.* TSI, numéro spécial, Janvier 2007.
2. P. Amar, J.-P. Comet, F. Képès, and V. Norris, editors. *Modelling and simulation of biological processes in the context of genomics.* Publisher Frontier group (ISBN : 2-84704-0374), June 2004.

Chapitres de livres

1. A. Richard, J.-P. Comet, and G. Bernot. *Modern Formal Methods and Applications*, chapter Formal Methods for Modeling Biological Regulatory Networks. Springer, ISBN : 1-

4020-4222-1, 2006.

2. J. Mary, G. Mercier, J.-P. Comet, A. Cornuéjols, C. Froideveaux, and M. Dutreix. *Informatique pour l'analyse du transcriptome*, chapter Utilisation d'une méthode de sélection d'attributs pour l'analyse du transcriptome de cellules de levure exposées à de faibles doses de radiation, pages 189–205. Hermes (ISBN : 2-7462-0850-4), 2004.
3. G. Bernot, J. Guespin-Michel, J.-P. Comet, P. Amar, A. Zemirline, F. Delaplace, P. Ballet, and A. Richard. *Proc. of the Dieppe Spring school on Modelling and simulation of biological processes in the context of genomics*, chapter Modelling, observability and experiment : a case study, pages 49–55. Publisher Frontier group (ISBN : 2-84704-036), 2003.
4. J. Mary, G. Mercier, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. *Proc. of the Dieppe Spring school on Modelling and simulation of biological processes in the context of genomics*, chapter Using an attribute estimation technique for the analysis of microarray data, pages 69–77. Publisher Frontier group (ISBN : 2-84704-036), 2003.
5. J.-J. Codani, J.-P. Comet, J.-C. Aude, E. Glémet, A. Wozniak, J.-L. Risler, A. Hénaut, and P.P. Slonimski. *Automation, Genomic and Functional Analyses*, volume 28, chapter Automatic analysis of large scale pairwise alignments of protein sequences, pages 229–244. Academic Press (ISBN : 0-12-521527-4), 1999.

Revues internationales

1. D. Filopon, A. Merieau, G. Bernot, J.-P. Comet, R. Leberre, B. Guery, B. Polack, and J. Guespin-Michel. Epigenetic acquisition of inducibility of type III cytotoxicity in p. aeruginosa. *BMC Bioinformatics*, 7 :272, 2006.
2. A. Richard, J.-P. Comet, and G. Bernot. R. Thomas' modeling of biological regulatory networks : introduction of singular states in the qualitative dynamics. *Fundamenta Informaticae*, 65(4) :373–392, 2005.
3. J. Guespin-Michel, G. Bernot, J.-P. Comet, A. Mérieau, A. Richard, C. Hulen, and B. Polack. Epigenesis and dynamic similarity in two regulatory networks in pseudomonas aeruginosa. *Acta Biotheoretica*, 52(4) :379–390, 2004.
4. G. Bernot, J.-P. Comet, A. Richard, and J. Guespin. Application of formal methods to biological regulatory networks : Extending Thomas' asynchronous logical approach with temporal logic. *Journal of Theoretical Biology*, 229(3) :339–347, 2004.
5. G. Mercier, N. Berthault, J. Mary, J. Peyre, A. Antoniadis, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. Biological detection of low radiation by combining results of two microarray analysis methods. *Nucleic Acids Research*, 32(1) :e12, 2004.
6. J.-P. Comet. Application of max-plus algebra to biological sequence comparisons. *Theoretical Computer Science*, 293 :189–217, 2003.
7. J.-P. Comet and J. Henry. Pairwise sequence alignment using a PROSITE pattern-derived similarity score. *Computers and Chemistry*, 26(5) :421–436, 2002.
8. J.-N. Bacro and J.-P. Comet. Sequence alignment : an approximation law for the Z-value with applications to databank scanning. *Computers and Chemistry*, 25 :401–410, 2000.
9. J.-P. Comet, J.-C. Aude, E. Glémet, A. Hénaut, J.-L. Risler, P.P. Slonimski, and J.-J. Codani. Significance of Z-value statistic of Smith-Waterman scores for protein alignments. *Computers and Chemistry*, 23(3-4) :317–331, 1999.

Revues nationales

1. J. Guespin-Michel, J.-P. Comet, and G. Bernot. Les réseaux de régulation biologique : rencontre entre biologie et informatique. *T.S.I.*, 23(7) :939–945, 2004.

Conférences et workshops internationaux avec comité de lecture et actes

1. Jamil Ahmad, Adrien Richard, Gilles Bernot, Jean-Paul Comet, and Olivier Roux. Delays in biological regulatory networks. In *Proceedings of IWBRA06*, LNCS, 2006, taux d'acceptation : 37%.
2. J.-P. Comet, H. Klaudel, and S. Liauzu. Modeling multi-valued genetic regulatory networks using high-level Petri nets. In *In ICATPN 2005*, volume 3536 of *LNCS*, pages 208–227, 2005.
3. A. Richard, J.-P. Comet, and G. Bernot. Graph-based modeling of biological regulatory networks : Introduction of singular states. In *Proceedings of the second International workshop on Computational Methods in Systems Biology (CMSB)*, volume 3082 of *LNBI*, pages 58–72, 2004.
4. G. Bernot, F. Cassez, J.-P. Comet, F. Delaplace, Müller, O. Roux, and O.H. Roux. Semantics of biological regulatory networks. In *Proceedings of the Workshop on Concurrent Models in Molecular Biology (BioConcur'2003)*, 2003.
5. S. Pérès and J.-P. Comet. Contribution of computational tree logic to biological regulatory networks : example from pseudomonas aeruginosa. In *International workshop on Computational Methods in Systems Biology*, volume 2602 of *LNCS*, pages 47–56, February 24-26, 2003, taux d'acceptation : 28%.

Conférences et workshops internationaux avec comité de lecture sans acte

1. J.-P. Comet, H. Klaudel, and S. Liauzu. Qualitative modelling of genetic networks using high-level Petri nets. IPG'2004, Lyon, France, October 14-15, 2004.
2. G. Bernot, O. Roux, C. Auberger, V. Bassano, F. Cassez, J.-P. Comet, F. Delaplace, A. Richard, O.H. Roux, and F. Tahi. Temporal logics and biological regulatory networks. ECCB Satellite Meeting on Modeling and Simulation of Biological Regulatory Processes, Paris, October 2003.
3. S. Pérès and J.-P. Comet. A formal approach to modelize biological regulatory networks. Workshop theoretical biophysics, Hiddensee, Germany, April 02-05, 2003.
4. J.-C. Aude and J.-P. Comet. Construction of protein sequences families. In *24th Aharon Katzir-Katchalsky conference BIOINFORMATICS-STRUCTURE*, Jerusalem, Israel, November 1996.

Conférences internationales invitées avec comité de lecture

1. G. Bernot, J.-P. Comet, and J. Guespin. Formal methods from computer science to treat temporal properties of biological regulatory networks, applied to mucus production in pseudomonas aeruginosa. Symposium Modeling approaches in Systems Biology, Physiology, and Pathophysiology, Poitiers (Futuroscope), France, October 7–12 2005.
2. M. Dutreix, J.-P. Comet, A. Cornuéjols, and C. Froidevaux. Determination of cellular drug targets : Searching for functional informations in the jungle of microarrays data. International Conference on Current Trends in Drug Discovery Research, Abstract will be

published in a special issue of journal "Medicinal Chemistry Research", India, Feb. 17–20. 2004.

3. P.P. Slonimski, M.-O. Mossé, P. Golik, A. Hénaut, Y. Diaz, J.-L. Risler, J.-P. Comet, J.-C. Aude, A. Wozniak, E. Glémet, and J.-J. Codani. The first laws of genomics. In *Microbial and Comparative Genomics in Microbial Genomes II : Sequencing, Functional Analysis and Comparative Genomics*, volume 46(3), Jan 1998.
4. P.P. Slonimski, J.-L. Risler, M.-O. Mossé, A. Hénaut, Y. Diaz, J.-P. Comet, and J.-C. Aude. What can we learn about proteins from complete sequences of microbial genome. In *Protein Science*, volume 6(1), 1997.

Conférences nationales invitées avec comité de lecture

1. G. Bernot and J.-P. Comet. Logique temporelle et model-checking pour les réseaux de régulation biologiques. 3èmes Journées de Post-Génomique de la Doua (JPGD'03), Lyon (France), 14-16 mai 2003.
2. J. Michel-Guespin, J.-P. Comet, and G. Bernot. Comment valider un modèle mathématique en biologie ? AFADL'03, Rennes (France), 15-17 janvier 2003.
3. G. Bernot, J. Michel-Guespin, A. Zermiline, J.-P. Comet, F. Delaplace, P. Ballet, and P. Amar. Modélisation, observabilité et expérimentation : étude d'un cas, boucle de rétroaction positive dans un réseau de régulation génétique. In *Séminaire d'Autrans « Modélisation et simulation des processus biologiques dans le contexte de la génomique »*, Mars 2002.
4. J.-P. Comet and F. Tahi. Structuration et analyse des données d'expression de gènes. Journées Peri-Bio à l'université d'Evry, Avril 2000.

Conférences nationales avec comité de lecture sans acte

1. J. Guespin, G. Bernot, J.-P. Comet, A. Mérieau, and B. Pollack. Une approche bio-informatique de la dynamique d'un réseau de régulation chez pseudomonas aeruginosa. VIème congrès de la SFM, 10 - 12 Mai 2004.
2. G. Mercier, J. Mary, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. Information mining of microarray data : comparison of results generated by different analysis methods. Ecole de printemps "Modélisation et simulation de processus dans le contexte de la biologie", 11-16 mai 2003, Dieppe (France).
3. J. Mary, G. Mercier, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. Utilisation d'une méthode de sélection d'attributs pour l'analyse du transcriptome de cellules de levure exposées à de faibles doses de radiation. 3èmes Journées de Post-Génomique de la Doua (JPGD'03), Lyon (France), 14-16 mai 2003.
4. J.-P. Comet. Méthodes de prédiction à partir des données transcriptionnelles : Prise en compte des erreurs de mesure. IMPG Marseille, France, Mars 2001.
5. J.-N. Bacro and J.-P. Comet. Comparaison de séquences : Loi asymptotique du Z-score et applications. In *XXXIèmes journées de Statistiques*, Grenoble, France, 17-21 mai 1999.
6. J.-P. Comet. Comparaison de séquences biologiques : Interprétation statistique du score Smith et Waterman, un autre indice de similarité : Le Z-score. XXIXèmes journées de Statistiques, 26-30 Mai 1997.

Posters

1. D. Mateus, J.-P. Gallois, P. Le Gall, and J.-P. Comet. Determining constraints in models of genetic regulatory networks using symbolic analysis. Jobim'2006, Bordeaux, France, July 5–7, 2006.
2. J. Ahmad, A. Richard, G. Bernot, J.-P. Comet, and O. Roux. Delays in biological regulatory networks. IPG'2005, Lyon, France, December 1-2, 2005.
3. J.-P. Comet, P. Le Gall, and G. Bernot. Qualitative modelling of genetic networks : Extracting experiment schema from modelling. IPG'2004, Lyon, France, October 13-15, 2004.
4. J.-P. Comet and A. Richard. SMBioNet : a tool for modeling biological regulatory networks driven by temporal behavior. Poster, ECCB'03, Paris, September 27-30 2003.
5. J.-P. Comet, S. Pérès, and G. Bernot. CTL and biological regulatory networks : example from p. aeriginosa. Poster, Symposium on Macromolecular networks, Paris, France, July 8-12, 2002.
6. J.-P. Comet, S. Pérès, and G. Bernot. Un langage formel pour les réseaux de régulation : un exemple avec multi-stationarité. Poster, JOBIM2002 St-Malo, France, 10-12 Juin 2002.
7. J.-P. Comet. Extension of dynamic programming sequence alignment to pattern. Poster, WABI2001, Arhus, Denmark, August 28-31, 2001.

Thèse & Rapports de Recherche

1. A. Richard and J.-P. Comet. Necessary conditions for multistationarity in discrete dynamical systems. Technical report, LaMI, Université d'Evry-Val-d'Essonne, 2005.
2. J.-P. Comet, H. Klaudel, and Liauzu L. Modeling multi-valued genetic regulatory networks using high-level petri nets. Technical report, LaMI, Université d'Evry-Val-d'Essonne, 2004.
3. J.-P. Comet, Richard A., and G. Bernot. Modeling biological regulatory networks with singular states. Technical report, LaMI, Université d'Evry-Val-d'Essonne, 2004.
4. G. Bernot, J.-P. Comet, A. Richard, and J. Guespin-Michel. A fruitful application of formal methods to biological regulatory networks : extending Thomas' logical approach with temporal logic. Technical report, LaMI, Université d'Evry-Val-d'Essonne, 2003.
5. G. Bernot, J. Guespin-Michel, A. Zemirline, J.-P. Comet, P. Amar, F. Delaplace, and Ballet P. Modelling, observability and experiment : a case study. Technical report, LaMI, Université d'Evry-Val-d'Essonne, 2002.
6. J.-P. Comet and J. Henry. Sequence alignment : Extension of dynamic programming to pattern. Technical report, LaMI, Université d'Evry-Val-d'Essonne, 2002.
7. J.-N. Bacro and J.-P. Comet. Approximated asymptotic law of Z-value and applications. Technical report, LaMI, Université d'Evry-Val-d'Essonne, 2000.
8. J.-P. Comet. *Programmation dynamique et comparaison de séquences biologiques*. PhD thesis, Université de Technologie de Compiègne, Novembre 1998.

Bibliographie choisie

Ce rapport reflète l'ensemble des travaux de recherche que j'ai effectués depuis ma thèse et au travers des encadrements. Pour approfondir les différents aspects techniques, on pourra se rapporter aux articles cités ci-dessous, classés par chapitre. Ils sont disponibles sur ma page web http ://www.ibisc.univ-evry.fr/∼comet/.

Pour le chapitre 1 : Séquences

– J.-N. Bacro and J.-P. Comet. Sequence alignment : an approximation law for the Z-value with applications to databank scanning. *Computers and Chemistry*, 25 :401–410, 2000.

Pour le chapitre 2 : Puces à ADN

– G. Mercier, N. Berthault, J. Mary, J. Peyre, A. Antoniadis, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. Biological detection of low radiation by combining results of two microarray analysis methods. *Nucleic Acids Research*, 32(1) :e12, 2004.

Pour les chapitres 3 et 4 : Fondements sémantiques des RRG / Méthodologie pour une modélisation itérative

– G. Bernot, J.-P. Comet, A. Richard, and J. Guespin. Application of formal methods to biological regulatory networks : Extending Thomas' asynchronous logical approach with temporal logic. *Journal of Theoretical Biology*, 229(3) :339–347, 2004.

– J.-P. Comet, H. Klaudel, and S. Liauzu. Modeling multi-valued genetic regulatory networks using high-level Petri nets. In *In ICATPN 2005*, volume 3536 of *LNCS*, pages 208–227, 2005.

Pour le chapitre 5 : Lien entre topologie et dynamique

– A. Richard and J.-P. Comet. Necessary conditions for multistationarity in discrete dynamical systems. Technical report, LaMI, Université d'Evry-Val-d'Essonne, 2005.

Bibliographie

[1] M. Adélaïde and G. Sutre. Parametric analysis and abstraction of genetic regulatory networks. In *2nd Workshop on Concurrent Models in Molecular Biology (BioCONCUR'04)*, Electronic Notes in Theor. Comp. Sci., London, UK, 2004 [To appear].

[2] J. Ahmad, A. Richard, G. Bernot, J.-P. Comet, and O. Roux. Delays in biological regulatory networks. In *Proceedings of IWBRA06*, LNCS, 2006, taux d'acceptation : 37%.

[3] D. Aldous. *Probability Approximations via the Poisson Clumping Heuristic*. Springer, New York, 1989.

[4] R. Alur, T.A. Henzinger, and P.-H. Ho. Automatic symbolic verification of embedded systems. *IEEE Transactions on Software Engineering*, 22 :181–201, 1996.

[5] R. Apweiler, M. Biswas, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E.V. Kriventseva, V. Mittard, N. Mulder, I. Phan, and E. Zdobnov. Proteome analysis database : online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Res.*, 29(1) :44–48, 2001.

[6] R. Arratia, L. Gordon, and L. Goldstein. Two moments suffice for poisson approximations : the Chen-Stein method. *The Annals of Probability*, 17 :9–25, 1989.

[7] R. Arratia, L. Gordon, and M.S. Waterman. An extreme value theory for sequence matching. *Ann. Stat.*, 14 :971–993, 1986.

[8] R. Arratia, L. Gordon, and M.S. Waterman. The Erdös-Renyi law in distribution for coin tossing and sequence matching. *The Annals of Statistics*, 18(2) :539–570, 1990.

[9] R. Arratia and M.S. Waterman. The Erdös-Renyi strong law for pattern matching with a given proportion of mismatches. *Ann. Prob*, 17 :1152–1169, 1989.

[10] R. Arratia and M.S. Waterman. A phase transition for the score in matching random sequences allowing deletions. *Ann. appl. Prob.*, 4(1) :200–225, 1994.

[11] R. Backofen and A. Dovier, editors. *Proceedings of WCB05 - Workshop on Constraint Based Methods for Bioinformatics*, 2005.

[12] J.-N. Bacro and J.-P. Comet. Sequence alignment : an approximation law for the Z-value with applications to databank scanning. *Computers and Chemistry*, 25 :401–410, 2000.

[13] O. Bastien. *Développements théoriques et méthodes numériques pour les analyses comparatives de génomes et protéomes biaisés. Application à la comparaison des génomes et protéomes de Plasmodium falciparum et d'Arabidopsis thaliana*. PhD thesis, Université Joseph Fournier, Grenoble, 2006.

[14] O. Bastien, J.-C. Aude, S. Roy, and E. Marechal. Fundamentals of massive automatic pairwise alignments of protein sequences : theoretical significance of Z-value statistics. *Bioinformatics*, 20(4) :534–537, 2004.

[15] O. Bastien, S. Lespinats, S. Roy, K. Métayer, B. Fertil, J.-J. Codani, and E. Maréchal. Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference. *Gene*, 336(2) :163–173, 2004.

[16] G. Bernot, F. Cassez, J.-P. Comet, F. Delaplace, C. Müller, O. Roux, and O.H. Roux. Semantics of biological regulatory networks. In *Proceedings of the Workshop on Concurrent Models in Molecular Biology (BioConcur)*, 2003.

[17] G. Bernot, J.-P. Comet, A. Richard, and J. Guespin. Application of formal methods to biological regulatory networks : Extending Thomas' asynchronous logical approach with temporal logic. *Journal of Theoretical Biology*, 229(3) :339–347, 2004.

[18] G. Bernot, J. Guespin-Michel, J.-P. Comet, P. Amar, A. Zemirline, F. Delaplace, P. Ballet, and A. Richard. *Proc. of the Dieppe Spring school on Modelling and simulation of biological processes in the context of genomics*, chapter Modelling, observability and experiment : a case study, pages 49–55. Publisher Frontier group (ISBN : 2-84704-036), 2003.

[19] C. Chaouiya, E. Remy, P. Ruet, and D. Thieffry. Qualitative modelling of genetic networks : From logical regulatory graphs to standard Petri nets. In *ICATPN 2004*, LNCS 3099, pages 137–156. Springer-Verlag, 2004.

[20] A. Cimatti, E. Clarke, F. Giunchiglia, and M. Roveri. NuSMV : a reimplementation of SMV. In *Proceeding of the International Workshop on Software Tools for Technology Transfer (STTT-98)*, BRICS Notes Series, NS-98-4, pages 25–31, 1998.

[21] O. Cinquin and J. Demongeot. Positive and negative feedback : striking a balance between necessary antagonists. *Journal of Theoretical Biology*, 216(2) :229–241, 2002.

[22] J.-P. Comet. *Programmation dynamique et comparaison de séquences biologiques*. PhD thesis, Université de Technologie de Compiègne, Novembre 1998.

[23] J.-P. Comet. Application of max-plus algebra to biological sequence comparisons. *Theoretical Computer Science*, 293 :189–217, 2003.

[24] J.-P. Comet, J.-C. Aude, E. Glémet, A. Hénaut, J.-L. Risler, P.P. Slonimski, and J.-J. Codani. Significance of Z-value statistic of Smith-Waterman scores for protein alignments. *Computers and Chemistry*, 23(3-4) :317–331, 1999.

[25] J.-P. Comet, H. Klaudel, and L. Liauzu. Modeling multi-valued genetic regulatory networks using high-level Petri nets. Technical report, LaMI, Université d'Evry-Val-d'Essonne, 2004.

[26] C. Damon. Classification non supervisée de modèles discrets de réseaux génétiques. Master's thesis, DEA Mathématiques, Vision et Apprentissage, ENS Cachan, septembre, 2005.

[27] H. de Jong. Modeling and simulation of genetic regulatory systems : a literature review. *J. Comput. Biol.*, 9(1) :67–103., 2002.

[28] H. de Jong, J. Geiselmann, C. Hernandez, and M. Page. Genetic network analyzer : qualitative simulation of genetic regulatory networks. *Bioinformatics*, 19(3) :336–44., 2003.

[29] H. de Jong, J.-L. Gouzé, C. Hernandez, M. Page, T. Sari, and H. Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. Technical Report RR-4407, INRIA, France, 2002.

[30] A. Dembo and S. Karlin. Strong limit laws of empirical functionals for large exeedences of partial sums of i.i.d. variables. *Ann. Prob.*, 19 :1737–1755, 1991.

[31] A. Dembo and S. Karlin. Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of Markov variables. *Ann. Prob.*, 19 :1756–1767, 1991.

[32] D. Diaz and P. Codognet. Design and implementation of the GNU Prolog system. *Journal of Functional and Logic Programming (JFLP)*, 2001.

[33] M.B. Elowitz, A.J. Levine, E.D. Siggia, and P.S. Swain. Stochastic gene expression in a single cell. *Science*, 297 :1183–1186, 2002.

[34] E.A. Emerson. *Handbook of theoretical computer science, Volume B : formal models and semantics*, chapter Temporal and modal logic, pages 995–1072. MIT Press, 1990.

[35] F. Fages. *Programmation logique par contraintes*. Ellipses, 1998.

[36] E. Fanchon, F. Corblin, L. Trilling, B. Hermant, and D. Gulino. Modeling the molecular network controlling adhesion between human endothelial cells : Inference and simulation using constraint logic programming. In *Proceedings of the second International workshop on Computational Methods in Systems Biology (CMSB)*, pages 104–118, 2004.

[37] A.F. Filippov. *Differential equations with discontinuous righthand sides*. Kluwer Academic Publishers, 1988.

[38] D. Filopon, A. Merieau, G. Bernot, J.-P. Comet, R. Leberre, B. Guery, B. Polack, and J. Guespin-Michel. Epigenetic acquisition of inducibility of type III cytotoxicity in P. aeruginosa. *BMC Bioinformatics*, 7 :272, 2006.

[39] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10) :906–914, 2001.

[40] O. Gandrillon and R. Houlgatte. *Informatique pour l'analyse du transcriptome*, chapter Le transcriptome : un nouveau monde ? Hermes Science, 2004.

[41] L. Goldstein and M.S. Waterman. Poisson, compound poisson and process approximations for testing statistical significance in sequence comparisons. *Bulletin of Mathematical Biology*, 54(5) :785–812, 1992.

[42] L. Goldstein and M.S. Waterman. Approximations to profile score distributions. *Journal of computational biology*, 1(2) :93–104, 1994.

[43] M. Gondran and M. Minoux. *Graphes et Algorithmes*. Eyrolles, 1995.

[44] J.-L. Gouzé. Positive and negative circuits in dynamical systems. *Journal of Biological Systems*, 6 :11–15, 1998.

[45] J. Guespin-Michel, G. Bernot, J.-P. Comet, A. Mérieau, A. Richard, C. Hulen, and B. Polack. Epigenesis and dynamic similarity in two regulatory networks in Pseudomonas aeruginosa. *Acta Biotheoretica*, 52(4) :379–390, 2004.

[46] J. Guespin-Michel, J.-P. Comet, and G. Bernot. Les réseaux de régulation biologique : rencontre entre biologie et informatique. *T.S.I.*, 23(7) :939–945, 2004.

[47] E.J. Gumbel. *Statistics of extremes*. Columbia University Press, Columbia, 1958.

[48] J. Hasty, D. McMillen, and J.J. Collins. Engineered gene circuits. *Nature*, 420(6912) :224–30, 2002.

[49] T.A. Henzinger, P.-H. Ho, and H. Wong-Toi. HyTech : A model checker for hybrid systems. *Software Tools for Technology Transfer*, 1997.

[50] S. Huang. Genomics, complexity and drug discovery : insights from boolean network models of cellular regulation. *Pharmacogenomics*, 2(3) :203–22, 2001.

[51] L. Hyafil and R.L. Rivest. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1) :15–17, 1976.

[52] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life : Systems biology. *Annu. Rev. Genomics Hum. Genet.*, 2 :343–372, 2001.

[53] N.L. Johnson and S. Kotz. *Distribution in statistics : Continuous univariate distributions - 1*. The Houghton Mifflin Series in statistics. Houghton Mifflin, 1970.

[54] S. Karlin and S.F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.*, 87 :2264–2268, 1990.

[55] S. Karlin and S.F. Atlschul. Application and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci.*, 90 :5873–5877, 1993.

[56] S. Karlin and A. Dembo. Limit distributions of maximal segmental score among Markov-dependent partial sums. *Ann. Appl. Prob.*, 24 :113–140, 1992.

[57] S. Karlin, A. Dembo, and T. Kawabata. Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.*, 18 :571–581, 1990.

[58] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *20th International Conference on Machine Learning*, pages 321–328, 2003.

[59] M. Kaufman and R. Thomas. Model analysis of the bases of multistationarity in the humoral immune response. *J. Theor. Biol.*, 129(2) :141–62, 1987.

[60] M. Kaufman, J. Urbain, and R. Thomas. Towards a logical analysis of the immune response. *J. Theor. Biol.*, 114(4) :527–61, 1985.

[61] K. Kira and L. Rendell. A practical approach to feature selection. In D. Sleeman and P. Edwards, editors, *Proc. International Conf. on Machine Learning*, pages 249–256. Morgan Kaufman, 1992.

[62] H. Kitano. Computational systems biology. *Nature*, 420(6912) :206–10, 2002.

[63] H. Kitano. Looking beyond the details : a rise in system-oriented approaches in genetics and molecular biology. *Curr. Genet.*, 41(1) :1–10, 2002.

[64] I. Kononenko. Estimating attributes : Analysis and extensions of RELIEF. In *Proc. of the European Conference on Machine Learning, ECML*, pages 171–182, 1994.

[65] E.V. Kriventseva, W. Fleischmann, E.M. Zdobnov, and R. Apweiler. CluSTr : a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res.*, 29(1) :33–36, 2001.

[66] E.V. Kriventseva, F. Servant, and R. Apweiler. Improvements to CluSTr : the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Res.*, 31(1) :388–389, 2003.

[67] C. Lefebvre, J.-C. Aude, E. Glémet, and C. Néri. Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates. *Bioinformatics*, 21(8) :1550–1558, 2005.

[68] M.J. Levsky, S.M. Shenoy, R.C. Pezo, and R.H. Singer. Single-cell gene expression profiling. *Science*, 297 :836–840, 2002.

[69] D.J. Lipman, W.J. Wilbur, T.F. Smith, and M.S. Waterman. On the statistical signifiance of nucleic acid similarities. *Nucl. Acids Res.*, 12 :215–226, 1984.

[70] A. Louis, E. Ollivier, J.-C. Aude, and J.-L. Risler. Massive sequence comparisons as a help in annotating genomic sequences. *Genome Research*, 11(7) :1296–1303, 2001.

[71] M. Mäkelä. Maria : Modular reachability analyser for algebraic system nets. In *ICATPN 2002*, number 2360 in LNCS, pages 434–444. Springer-Verlag, 2002.

[72] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F.

Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057) :376–380, 2005.

[73] J. Mary, G. Mercier, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. *Proc. of the Dieppe Spring school on Modelling and simulation of biological processes in the context of genomics*, chapter Using an attribute estimation technique for the analysis of microarray data, pages 69–77. Publisher Frontier group (ISBN : 2-84704-036), 2003.

[74] J. Mary, G. Mercier, J.-P. Comet, A. Cornuéjols, C. Froideveaux, and M. Dutreix. *Informatique pour l'analyse du transcriptome*, chapter Utilisation d'une méthode de sélection d'attributs pour l'analyse du transcriptome de cellules de levure exposées à de faibles doses de radiation, pages 189–205. Hermes (ISBN : 2-7462-0850-4), 2004.

[75] E. Mayr. The evolution of living systems. *PNAS*, 51 :934–941, 1964.

[76] E. Mayr. The objects of selection. *PNAS*, 94 :2091–2094, 1997.

[77] M. Meila and J. Shi. Learning segmentation by random walks. In *NIPS*, 2000.

[78] G. Mercier, N. Berthault, J. Mary, J. Peyre, A. Antoniadis, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. Biological detection of low radiation by combining results of two microarray analysis methods. *Nucleic Acids Research*, 32(1) :e12, 2004.

[79] S. Mohseni-Zadeh, A. Louis, P. Brezellec, and J.-L. Risler. PHYTOPROT : a database of clusters of plant proteins. *Nucleic Acids Research*, 32 :D351–D353, 2004.

[80] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48 :443–453, 1970.

[81] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering : Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, 2001.

[82] J. Park and S.A. Teichmann. DIVCLUS : an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics*, 14 :144–150, 1998.

[83] P. Pavlidis and W. Noble. Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology*, 2(10) :0042.1–0042.15, 2001.

[84] E. Plahte, T. Mestl, and S.W. Omholt. Feedback loops, stability and multistationarity in dynamical systems. *Journal of Biological Systems*, 3 :409–413, 1995.

[85] A. Richard. *Modèle formel pour les réseaux de régulation génétique et Influence des circuits de rétroaction.* PhD thesis, Université d'Evry-Val d'Essonne, Septembre 2006.

[86] A. Richard, J.-P. Comet, and G. Bernot. R. Thomas' modeling of biological regulatory networks : introduction of singular states in the qualitative dynamics. *Fundamenta Informaticae*, 65(4) :373–392, 2005.

[87] F. Robert. *Discrete iterations : a metric study*, volume 6 of *Series in Computational Mathematics*. Springer, 1986.

[88] F. Robert. *Les systèmes dynamiques discrets*, volume 19 of *Mathématiques et Applications*. Springer, 1995.

[89] E. Rémy, P. Ruet, and D. Thieffry. Graphic requirements for multistability and attractive cycles in a boolean dynamical framework. Technical report, IML, France, 2005.

[90] Mau-Hsiang Shih and Jian-Lang Dong. A combinatorial analogue of the jacobian problem in automata networks. *Advances in Applied Mathematics*, 34 :30–46, 2005.

[91] G.R. Shorack and J.A. Wellner. *Empirical processes with applications to statistics*. Wiley, New-York, 1986.

[92] D. Sieling. Minimization of decision trees is hard to approximate. Technical report, Univ. Dortmund, Germany, 2002.

[93] D.K. Slonim, P. Tamayo, J.P. Mesirov, T.R. Golub, and E.S. Lander. Class prediction and discovery using gene expression data. In *the 4th Annual International Conference on Computational Molecular Biology (RECOMB)*, 2000.

[94] T.F. Smith and M.S. Waterman. Comparision of bio-sequences. *Advan. Appl. Math.*, 2 :482–489, 1981.

[95] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147 :195–197, 1981.

[96] T.F. Smith, M.S. Waterman, and W.M. Fitch. Comparative biosequence metrics. *J. Mol. Evol.*, 18 :38–46, 1981.

[97] P. Smolen, D.A. Baxter, and J.H. Byrne. Modeling transcriptional control in gene networks : Methods, recent results, and future directions. *Bulletin of Mathematical Biology*, 62(2) :247–292, 2000.

[98] E.H Snoussi. Qualitative dynamics of a piecewise-linear differential equations : a discrete mapping approach. *Dynamics and stability of Systems*, 4 :189–207, 1989.

[99] E.H. Snoussi. Necessary conditions for multistationarity and stable periodicity. *Journal of Biological Systems*, 6 :3–9, 1998.

[100] E.H. Snoussi and R. Thomas. Logical identification of all steady states : the concept of feedback loop caracteristic states. *Bull. Math. Biol.*, 55(5) :973–991, 1993.

[101] C. Soulé. Graphic requirements for multistationarity. *ComPlexUs*, 1 :123–133, 2003.

[102] C. Soulé. Mathematical approaches to differentiation and gene regulation. Technical report, IHES, 2005.

[103] J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of concerved genetic modules. *Science*, 302 :249–255, 2003.

[104] D. Thieffry and R. Thomas. Dynamical behaviour of biological regulatory networks - II. immunity control in bacteriophage lambda. *Bull. Math. Biol.*, 57(2) :277–97, 1995.

[105] D. Thieffry and R. Thomas. Qualitative analysis of gene networks. In *Proceedings of the Pac. Symp. Biocomput.*, pages 77–88, 1998.

[106] R. Thomas. Logical analysis of systems comprising feedback loops. *J. Theor. Biol.*, 73(4) :631–56, 1978.

[107] R. Thomas. On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations. *Springer Series in Synergies*, 9 :180–193, 1981.

[108] R. Thomas, A.M. Gathoye, and L. Lambert. A complex control circuit. Regulation of immunity in temperate bacteriophages. *Eur. J. Biochem.*, 71(1) :211–27, 1976.

[109] R. Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behavior. *Chaos*, 11 :170–179, 2001.

[110] R. Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. II. Logical analysis of regulatory networks in terms of feedback circuits. *Chaos*, 11 :180–195, 2001.

[111] R. Thomas, D. Thieffry, and M. Kaufman. Dynamical behaviour of biological regulatory networks - I. Biological role of feedback loops an practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.*, 57(2) :247–76, 1995.

[112] V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9) :5116–5121, 2001.

[113] D. Verma and M. Meila. A comparison of spectral clustering algorithms. Technical report, University of Washington, 2003.

[114] M.S. Waterman. Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B*, 344 :383–390, 1994.

[115] M.S. Waterman. The statistical significance of local alignement score. Technical report, University of Southern California, USA, 1994.

[116] M.S. Waterman and M. Vingron. Sequence comparison significance and poisson approximation. *Statistical Science*, 9(3) :367–381, 1994.

[117] D.J. Wilkinson. *Stochastic Modelling for Systems Biology.* Chapman & Hall/CRC, 2006.

[118] O. Wolkenhauer. Systems biology : the reincarnation of systems theory applied in biology ? *Brief Bioinform.*, 2(3) :258–70, 2001.

[119] T.T. Wu, W.M. Fitch, and E. Margoliash. The information content of protein amino acid sequences. *Annu. Rev. Biochem.*, 43 :539–566, 1974.

[120] E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray. In *Proc. of the Int. Conf. on Machine Learning, ICML*, pages 601–608, 2001.

Index

Résumé

La bio-informatique est née de la nécessité d'utiliser des méthodes provenant de l'informatique pour le traitement des informations dans le cadre de la biologie. Elle constitue un large domaine de recherche pluridisciplinaire où la biologie, les mathématiques, l'informatique, la physique et la chimie théoriques apportent des éléments cruciaux pour la recherche. Elle est en pleine mutation depuis sa naissance, suivant ainsi les évolutions du génie biologique. Après la comparaison de séquences et l'analyse des données du transcriptome, la biologie des systèmes a fait rentrer la modélisation mathématique en biologie dans le champs de la bio-informatique.

Dans ce document, nous montrons comment le terme « modélisation » peut être compris dans ses différentes acceptions au travers des différents thèmes abordés que sont la comparaison de séquences, l'analyse du transcriptome et la biologie des systèmes. L'activité de modélisation en biologie permet entre autres de générer des connaissances (inférence de fonction par homologie), d'exprimer des liens potentiels entre transcriptome et comportement macroscopique (modèles statistiques pour les données du transcriptome), de comprendre la dynamique de systèmes (réseaux de régulation génétique), d'automatiser partiellement la construction d'un modèle (en déléguant à l'ordinateur la vérification de propriétés) ou encore d'énoncer des règles générales (lien entre topologie d'un réseau d'interactions et sa dynamique)...

Au travers de la modélisation des réseaux de régulation génétique, nous montrons que le rapprochement entre méthodes formelles et biologie moléculaire est fécond, que d'autres champs de recherche en informatique peuvent aussi apporter à la biologie et enfin que la biologie est devenue et restera sans doute encore longtemps une source de sujets de recherche novateurs pour l'informatique théorique.

Abstract

Bioinformatics stemmed from the need to use modern data-processing methods in the general area of biological research. It now constitutes a broad multidisciplinary field of study in which biology, mathematics, computer science, theoretical physics and chemistry each contribute essential elements to scientific investigation. From the start bioinformatics is in a continuous transformation, thus following the rapid development of advanced biological engineering. After the biological sequence comparison and analysis of transcriptomic data, systems biology introduced mathematical modelling in biology into the world of bioinformatics.

In this document we show how the term "modelling" can be understood in its different meanings according to the different areas chosen for study. These are : sequence comparaison, analysis of transcriptomic data and systems biology. The modelling activity in biology allows, among other things, knowledge generation (functional inference by homology), the definition of potential links between transcriptome and macroscopic behavior (statistical models for transcriptomic data), the clearer understanding of systems dynamics (genetic regulatory networks), the partial automation of model-building (by delegating property verification to the computer), or, again, the determination of general rules (links between the topology of an interaction network and its dynamics)...

By modelling genetic regulatory networks, we show that the use of formal methods in molecular biology is productive, that other areas of research in computer science can also contribute to biology, and, finally, that biology has become and will long remain a source of innovative research projects for theoretical computer science.