Identification de réseaux de régulation génique à partir de données d'expression : une approche basée sur les modèles affines par morceaux.

Samuel Drulhe

N° d'ordre : 0000
N° attribué par la bibliothèque : 0000

UNIVERSITÉ JOSEPH FOURIER - GRENOBLE I

THÈSE
pour l'obtention du titre de
Docteur de l'université Joseph Fourier - Grenoble I
spécialité : Mathématiques Informatique
au titre de l'école doctorale
Mathématiques, Sciences et technologies de l'information, Informatique

présentée et soutenue publiquement le 09/12/2008
par Samuel DRULHE

Identification de réseaux de régulation génique à partir de données d'expression : une approche basée sur les modèles affines par morceaux

---

Composition du jury :

M. Hans GEISELMANN,	Président
Me. Florence D'ALCHER-BUC,	Rapporteur
M. Diego DI BERNARDO,	Rapporteur
M. Oded MALER,	Examinateur
M. Hidde de JONG,	Directeur de thèse
M. Giancarlo FERRARI-TRECATE,	Co-Directeur de thèse

Thèse préparée au sein de l'équipe HELIX à l'INRIA Rhône-Alpes

Table des matières

Liste des figures

Première partie

Contexte

Chapitre 1

Réseaux de régulation

1.1 Biologie moléculaire et réseaux d'interactions

1.1.1 Systèmes biologiques moléculaires et leur identification

Déjà en considérant le vivant à un niveau cellulaire, une grande unité [71] apparait : tous les êtres vivants sont apparentés dans leur composition chimique, leurs fonctions et leur plan. En poussant encore la réduction d'échelle, on arrive à une mécanique moléculaire universelle. Les êtres vivants partagent à ce niveau un grand nombre de caractéristiques fondamentales, ce qui entraine que les découvertes faites pour une espèce donnée s'appliquent le plus souvent directement à beaucoup d'autres.

Une approche classique, en biologie moléculaire, a consisté à explorer les réseaux complexes d'éléments cellulaires (éléments réagissant les uns avec les autres, catalysant des réactions, déplaçant physiquement d'autres éléments dans le milieu cellulaire) en se centrant sur des molécules isolées et leurs réactions [115]. Ces mécanismes précisés, il était possible d'envisager des systèmes d'éléments cellulaires et d'expliquer leur multistationnarité. Mais deux maillons manquaient à cette approche systématique. D'une part la connaissance de ces éléments n'était pas exhaustive. D'autre part, les biologistes expérimentateurs ne disposaient pas d'outil de mesure à grande échelle des processus élémentaires internes aux cellules vivantes.

Ces limites ont été notablement repoussées ces dernières décennies. Le séquençage complet de l'ADN d'organismes variés a conduit à la mise en place de techniques qui permettent de mesurer un large spectre de phénomènes en parallèle, se produisant à l'intérieur des cellules vivantes. Des techniques dites "puces à ADN" autorisent ainsi à estimer pour plusieurs dizaines de milliers de gènes leur expression différentielle au niveau de l'ARN messager : extrêmement courantes, industrialisées, ces techniques sont de moins en moins couteuses ; standardisées, elles fournissent des résultats expérimentaux accessibles et répétables. Il est possible, pour donner un exemple concernant un autre type d'élément cellulaire, de distinguer les protéines produites par spectrométrie de masse, et de tester leurs interactions. Pour tous ces éléments, on voit l'apparition de bases de données qui accumulent des quantités énormes de données concernant différents aspects du développement et du fonctionnement des cellules.

Ces avancées permettent à la communauté scientifique d'envisager l'approche systémique appliquée à la biologie moléculaire avec de plus en plus d'intérêt [64]. En effet, la qualité et la quantité des données expérimentales allant croissantes, il est possible de raffiner ou de préciser les modèles d'interactions développés théoriquement. De sorte que l'approche systémique peut être envisagée dans les développements qu'elle connait dans d'autres domaines. La simulation permet d'accéder à la prédiction. Une nouvelle science a ainsi vu le jour qui permet d'ajouter des fonctions à des êtres vivants qui en étaient naturellement dépourvus, qui se basent sur des systèmes d'interactions d'éléments cellulaires [55, 52, 39, 13].

De nombreuses applications sont pensées en terme de traitement pour des maladies, en terme d'optimisation de bioprocédés. Mais la problématique systémique qui nous intéressera dans ce manuscrit concerne le problème dit inverse de l'inférence de réseau à partir

de données de mesure.

En automatique, cela se nomme l'*identification* [67]. On définit en général cette problématique comme l'opération de détermination des caractéristiques dynamiques d'un procédé (système). Sans cette connaissance, il n'est pas possible de concevoir et de mettre en œuvre un système performant de régulation. En biologie, la connaissance du réseau d'interactions est en fait un objectif en soi. La premier sous-objectif consiste à identifier les éléments interagissants. Le deuxième consiste à déterminer, s'il y a lieu, quel élément (ou groupe d'éléments) agit sur tel autre élément (ou groupe d'éléments), éventuellement lui-même. Le troisième sous-objectif consiste à identifier, au delà de la topologie du réseau d'interactions, la dynamique de ces interactions. En effet, celle-ci présente un enjeu majeur pour l'analyse et la simulation des systèmes : on peut caractériser les réponses des divers éléments du système, et éventuellement les grouper ou les distinguer. Il est ainsi possible de décomposer les interactions en modules qui sont autant de sous-systèmes indépendants qu'il faut considérer à leur échelle dynamique propre. Tout cela est déterminé par les procédés et les modalités de l'expérience dont sont issues les données.

Cela nous conduit à exposer un schéma simplificateur de représentation de la mécanique moléculaire cellulaire qui est généralement admis par la communauté des chercheurs en biologie systémique, de manière à énoncer une nomenclature utile pour distinguer les types de systèmes et leurs problèmes spécifiques.

1.1.2 Mécanique moléculaire et réseaux d'interactions

Le schéma simplificateur de représentation de la mécanique moléculaire cellulaire commence par le fait qu'une sous-partie des gènes sont exprimés (après transcription et traduction, voir Figure 1.1) en protéines, qui interagissent entre elles (soit qu'elles réagissent, soit qu'elles catalysent les réactions d'autres molécules, soit qu'elles modifient la structure ou la position d'autres molécules modifiant ainsi leur fonction) : certaines protéines ont ainsi pour fonction, entre autre, d'influencer les interactions d'autres molécules, et, par exemple, l'expression des gènes. Il apparait ici une première forme de chaine d'interactions (éventuellement parallèles ou bouclées) qu'il sera pratique de représenter sous la forme d'un réseau. Ce mode de représentation est adopté en biochimie. En fait, d'autres types d'interactions à l'intérieur d'une cellule sont aussi envisagés : il s'agit d'abstraire les réactions physico-chimiques élémentaires (i.e. la physique concrète de ces réactions lors de la synthèse et de la dégradation) pour se concentrer sur les influences entre les éléments cellulaires étudiés.

Ces réseaux biochimiques peuvent être construits à plusieurs niveaux et peuvent représenter différents types d'interactions. Dans le cas où l'on s'intéresse à une chaine particulière d'interactions, on emploie le terme de voie plutôt que de réseau.

Les réseaux d'interactions sont généralement classés en trois grandes familles : les réseaux de régulation génique, les réseaux d'interactions protéine à protéine, et les réseaux métaboliques. À vrai dire, une autre catégorie de réseaux concerne aussi la transduction des signaux à travers la membrane cellulaire : cependant, pour simplifier, dans le cadre de l'étude des trois premiers types d'interactions, on considère généralement qu'il s'agit d'un cas spécial des réseaux d'interactions protéine à protéine. Donnons maintenant une brève description de ces trois familles. Les réseaux de régulation génique représentent les relations qui peuvent être établies entre les gènes, lorsqu'on mesure la manière par laquelle l'expression de chacun d'entre eux affectent le niveau d'expression des autres. Les réseaux de protéines traduisent les interactions protéiques telles que la formation de complexe, ou la

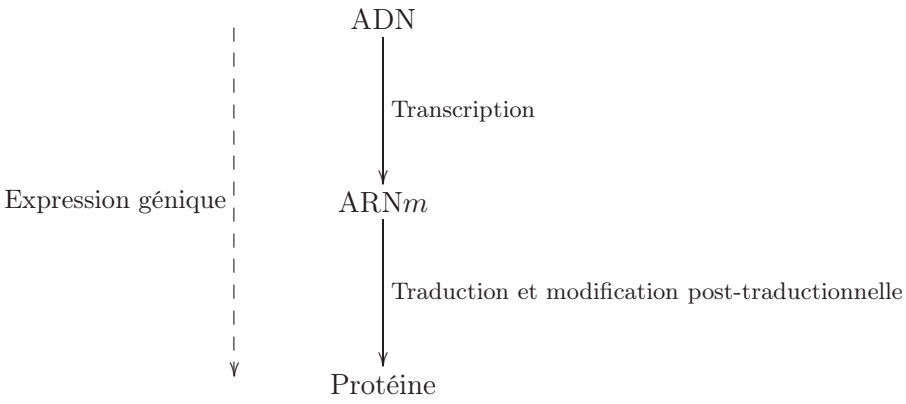


FIGURE 1.1 – Principe fondamental de la biologie moléculaire : un gène identifié sur la séquence d'ADN est transcrit en ARNm qui est ensuite traduit au niveau du complexe ribosomal conduisant à la production d'une chaine d'acides aminés qui, après des modifications post-traductionnelles, constitue la protéine produit.

modification d'une protéine par une enzyme. Enfin, les réseaux de métabolites représentent les transformations enzymatiques entre métabolites (petites molécules de la cellule).

Cette nomenclature laisse faussement penser qu'il y a une distinction entre les processus envisagés par ces trois familles. Dans la mécanique cellulaire, ces réseaux sont en fait totalement intriqués. Chaque famille de réseaux est la simplification du système cellulaire complet. Les gènes, par exemple, lorsque l'on considère la chaine des réactions biochimiques, n'interagissent pas directement : l'activation ou la répression de l'expression d'un gène se produit du fait de l'action spécifique de protéines qui sont les produits de l'expression de gènes. Il existe aussi des cas où les métabolites contribuent à la régulation de l'expression des gènes. Cependant, on distingue et on hiérarchise les trois niveaux de manière à ce que, lorsque l'on se place dans le cadre d'une famille d'interactions données, on puisse faire abstraction de ce qu'il se passe pour les autres niveaux.

Pour la suite de ce manuscrit, nous nous centrerons sur les réseaux de régulation génique. L'expression génique est un phénomène compliqué pendant lequel on distingue différentes étapes aboutissant à la synthèse des protéines : on résume en général le processus comme sur la Figure 1.1. La régulation la mieux étudiée est la régulation de la transcription de l'ADN : en effet le taux de transcription d'un gène est contrôlé en grande partie par la séquence d'ADN nommé site *promoteur* qui le précède, déterminant l'affinité chimique de l'ARN polymérase servant à la transcription, comme illustré par la Figure 1.2. Le contrôle est lié à des protéines dites *facteurs de transcription* qui peuvent se lier au site promoteur, affectant ainsi l'affinité de l'ARNp : selon le cas, un facteur de transcription peut agir comme un activateur (qui augmente le taux de transcription) ou comme répresseur (s'il le réduit). Nous donnons un exemple sur la Figure 1.2. En plus de cette régulation, on observe que l'expression génique peut être contrôlée pendant la production de l'ARNm (et son transport pour les eucaryotes), durant la traduction (affinité avec le ribosome) et les modifications post-traductionnelles qui suivent éventuellement. Enfin, les protéines sont naturellement dégradées grâce à la présence de protéases. L'activité et la stabilité des protéines sont soumises à de multiples influences environnementales et physiologiques.

Tout cela est simplifié au sens où c'est l'action d'un régulateur sur sa cible qui est prise en compte pour la construction du réseau de régulation génique. Non seulement le

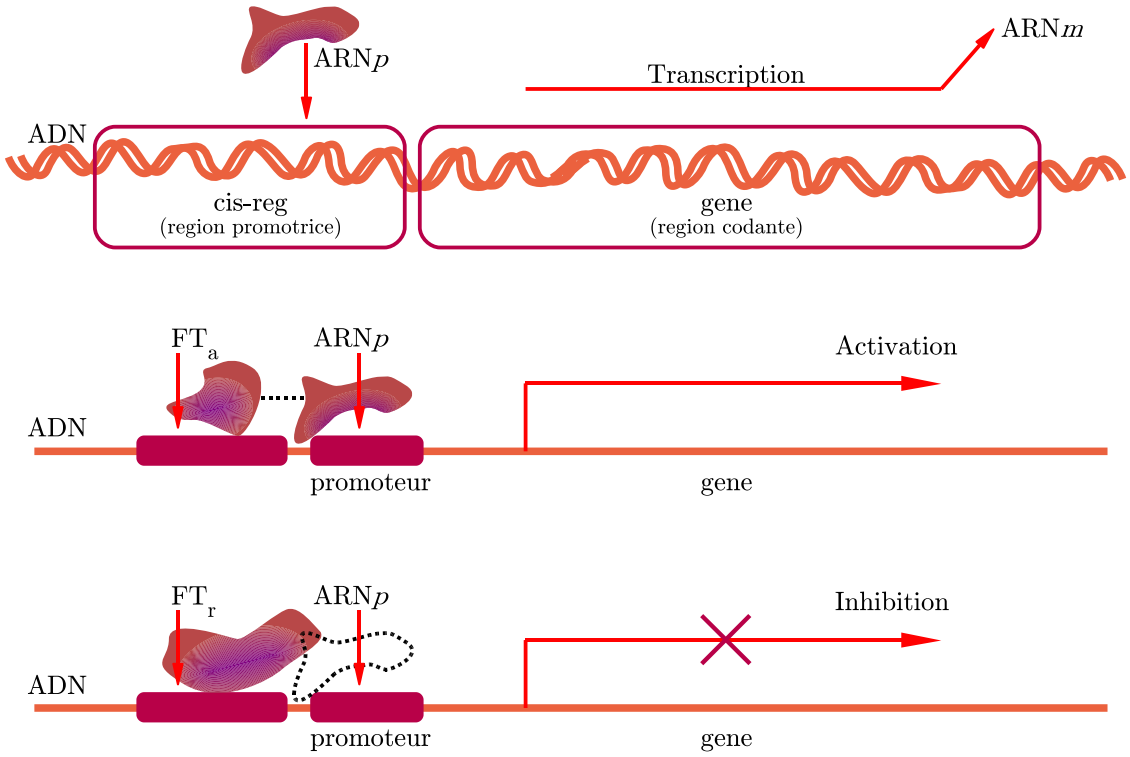


FIGURE 1.2 – Régulation de la transcription. La transcription de la région codante intervient lorsqu'une ARN polymérase s'associe avec la région promotrice qui la précède. Nous illustrons ici deux cas de figure. (Activation) Un facteur de transcription FTa se lie au promoteur et favorise l'affinité du site avec l'ARNp : la transcription est favorisée, c'est un activateur. (Inhibition) Un facteur de transcription FTr se lie au promoteur et diminue l'affinité du site avec l'ARNp : la transcription est inhibée, c'est un répresseur.

mode de régulation est abstrait, mais encore les étapes intermédiaires qui ont conduit à la régulation sont elles masquées. En effet, l'action entre les gènes peut se produire de diverses manières. La plus directe est que la protéine résultant de l'expression d'un premier gène est un régulateur de l'expression d'un second. Or, on peut envisager que la protéine influence en fait la production d'un métabolite qui est le véritable régulateur contrôlant l'expression du second gène. Cela est étudié de la même manière : on dira que le premier gène contrôle l'expression du second. Un exemple est donné sur la Figure 1.3.

1.2 Modéliser les interactions géniques

1.2.1 Complexité et sophistication

Nous avons pu voir que les réseaux de régulation génique traduisent en fait une intrication complexe de processus physicochimiques : la connaissance détaillée acquise à leur sujet depuis plusieurs décennies ne permet toujours pas d'envisager la reconstruction de réseaux à partir de celles-ci. L'analyse du réseau contrôlant la réponse à un manque de nutriments de la part de la bactérie *E. coli*, par exemple, fait partie des phénomènes intensivement étudiés par les biologistes moléculaires depuis des dizaines d'années [102, 54] : si les régulateurs principaux impliqués dans la réponse bactérienne ainsi que leurs cibles sont connus, ce n'est que très récemment que l'étude du fonctionnement du réseau de régulation concerné et celle de son comportement dynamique ont commencé [21, 7, 10, 98, 37].

La reconstruction des réseaux de régulation génique est donc un enjeu majeur pour la compréhension de la machinerie cellulaire et pour l'élaboration d'interventions ayant un intérêt en terme médical ou en terme bio-ingénierique. Selon le détail de la complexité de ces réseaux qu'il est nécessaire d'envisager, différents types de modèle pour abstraire cette complexité ont été proposés. Il existe une gradation dans la sophistication de ces types de modèle : cela va de pair avec le détail de la traduction de la complexité des interactions ; pratiquement, cela correspond surtout aux caractéristiques des mesures traitées. En effet, chaque type de modèle comporte un grand nombre de paramètres auxquels il faut attribuer une valeur numérique soit sur la base d'une connaissance a priori de leur valeur (ce qui est rarement le cas pour la plupart des systèmes biologiques actuellement traités), soit sur la base de données expérimentales. La qualité de l'estimation des paramètres dépendra intrinsèquement de la quantité et de la qualité des données qui sont mises à la disposition de la procédure d'identification des paramètres du modèle. À côté de ces limitations sur l'information disponible, un autre problème surgit dans le pouvoir informatif des données pour une procédure d'identification : les modèles dynamiques représentent des processus intrinsèquement non linéaires et les paramètres du modèle peuvent être fortement corrélés entre eux. La question de l'*identifiabilité* des paramètres du modèle s'avère être une tâche essentielle.

Par ailleurs, dans ce manuscrit, nous n'aborderons pas la question de la pertinence d'un niveau de complexité adopté, c'est-à-dire la question du choix du type de modèle. Comme c'est le cas dans l'étude de tout système, se demander quel est le "bon type de modèle" est crucial : il est en effet nécessaire que le modèle choisi et identifié se comporte effectivement comme le processus qu'il prétend représenter. Cette question de la *validation* du type de modèle est de ce fait un sous-problème de l'*identification* des systèmes : on ne peut pas se contenter d'abstraire des phénomènes complexes sans préciser les domaines d'utilisation des modèles. Quelque soit le degré de sophistication choisi, on ne saurait aborder dans toute sa complexité des systèmes dynamiques fortement non-linéaires de grande dimension : il s'agit

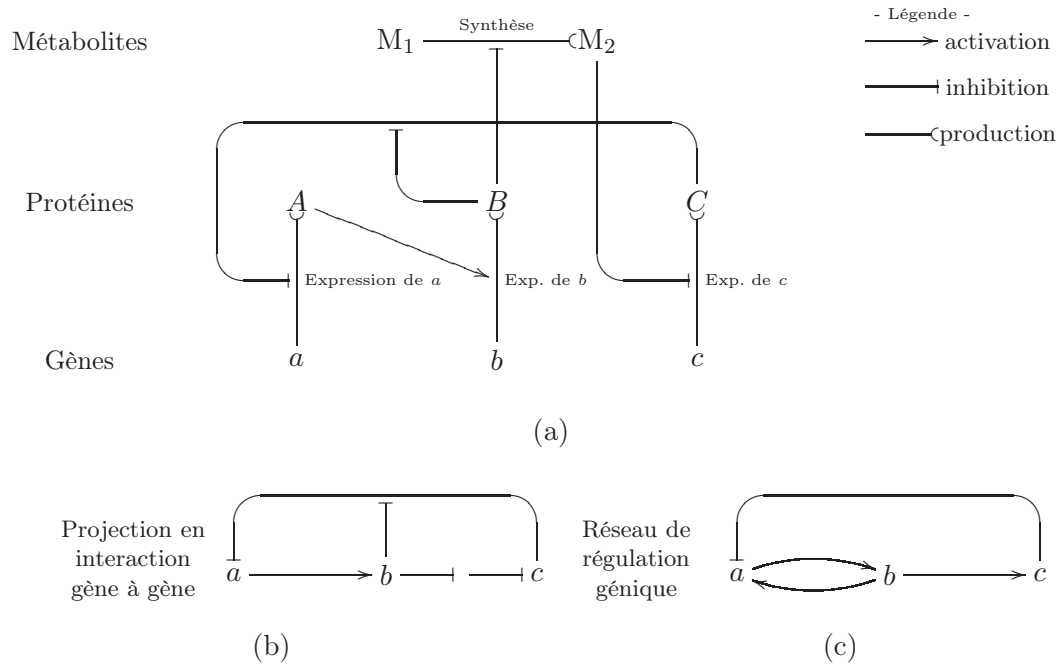


FIGURE 1.3 – Modélisation des interactions géniques (adapté de [28] et [19]). (a) La protéine A active l'expression du gène b. La protéine B réprime la synthèse de la métabolite M_2 qui inhibe l'expression du gène c. La protéine C inhibe l'expression du gène a, mais ceci en concurrence avec la présence de la protéine B qui, quant à elle, n'a aucun effet direct sur l'expression. (b) Par projection de ces diverses actions comme des interactions gène à gène, on déduit le réseau de régulation génique. Entre le gène b et le gène c, il existe deux actions d'inhibition, ce qui équivaut à une activation. (c) Il vient au final que a active b, qui active c, qui inhibe a si l'expression de b est suffisamment faible. Dans une représentation de type graphe où les sommets sont des gènes et les arêtes des actions orientées et valuées (activation ou inhibition) gène-a-gène, b devient un activateur de a : ce type de représentation n'est pas suffisant pour montrer comment les actions se combinent.

alors d'espérer au mieux une analogie de comportement plutôt qu'une identité. Pourtant, un modèle non-validé, plus qu'inutile, peut s'avérer dangereux. Le choix de la complexité du modèle, qui peut être crucial selon le pouvoir prédictif espéré, doit se baser sur la caractérisation précise des modèles : nous mettrons de côté ce problème dans le cadre du travail présenté ici qui se concentrera sur l'inférence des paramètres à partir de données de mesure.

1.2.2 Quelques modèles de réseaux de régulation génique

Il existe plusieurs formalismes mathématiques permettant de modéliser les réseaux de régulation génique [106, 30]. Chaque formalisme permet de représenter avec davantage de précision certains aspects caractéristiques des réseaux. On peut décomposer les formalismes en considérant les quatres axes suivants :

1. selon que les expressions géniques sont soit des niveaux de présence, soit des concentrations, soit des intervalles de concentrations, on les dira discrets, ou continus, ou mélangés ;
2. entre modes dynamiques, ils peuvent être à transition continue, ou discrète (automate synchrone, automate asynchrone, système hybride) ;
3. ils sont à évolution déterministe ou stochastique ;
4. enfin, ils peuvent être à paramétrage qualitatif ou quantitatif.

On peut combiner à peu près tous les éléments de chaque niveau, mais en pratique, les principaux formalismes investis ont été : les réseaux logiques (par exemple booléens [2]), les réseaux de Pétri, les réseaux bayésiens, les équations différentielles ordinaires, les équations aux dérivées partielles, les équations stochastiques. De manière grossière, disons que cette gamme permet de passer d'abstractions très fortes, pour lesquelles il est possible de déduire des propriétés avec un cout de traitement faible, à des approches décrivant plus précisément les mécanismes des interactions biochimiques, qui sont plus gourmandes en terme de connaissances et en terme de traitement, mais qui permettent d'investir des propriétés dynamiques éventuellement déterminantes (voir [30]).

Les modèles d'équations différentielles ordinaires sont probablement le formalisme le plus utilisé pour modéliser les réseaux de régulation génique. Ces modèles représentent la concentration des produits de l'expression des gènes (ARNm puis protéines) par des variables continues qui varient dans le temps. Ainsi, $\mathbf{x} : \mathbb{R} \to \mathbb{R}^{+n}$ est un vecteur de dimension $n \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$ pour lequel chacune de ses composantes est linéairement liée (par idéalisation) à la concentration de n produits d'expression génique observés dans le temps de l'expérience. Aucune concentration n'étant négative, pour tout $i \in \{1, \ldots, n\}, x_i(t) \in \mathbb{R}^+$: le gain de mesure est positif. Le fait que l'on représente une concentration moléculaire signifie deux hypothèses majeures : (1)- il y a suffisamment de molécules dans les cellules observées pour se ramener à une étude macroscopique ; (2)- le milieu cellulaire est homogène. La première condition n'est pas toujours vérifiée : il se peut que le produit de l'expression d'un gène ait un effet régulateur même en présence de très peu de molécules. La deuxième condition n'est pas vraie : cependant, cette hypothèse d'homogénéité est surtout intéressante localement, dans la zone cellulaire où se produit le contrôle, et les phénomènes de transport intracellulaire ne sont pas pris en compte pour ce type de modèles.

Puisqu'il s'agit d'un système auto-régulé, la variation dans le temps de chaque composante $\Delta x_i(t)$ dépend de l'action des autres éléments, c'est-à-dire de leurs concentrations représentées au moyen de $x_j(t), j \in \{1, \ldots, n\}$. Par passage à la limite, on écrit alors que

la dérivée du premier ordre par rapport au temps est une fonction de l'état présent de $x_1, \ldots, x_n$:

$$\dot{x_i} = f_i(\mathbf{x}), i \in \{1, \ldots, n\}. \tag{1.1}$$

La fonction $f_i : \mathbb{R}^{+n} \to \mathbb{R}$ est souvent fortement non-linéaire, mais de très nombreuses approches, exploitant l'approximation des petites variations, travaillent avec des fonctionnelles f_i linéaires pour caractériser les interactions.

FIGURE 1.4 – Exemple de réseau de régulation à deux gènes. Les interactions sont représentées avec des arcs orientés : s'il s'agit d'une activation, l'arc se termine par une flèche, s'il s'agit d'une inhibition, l'arc se termine par une barre.

Exemple 1. *Deux gènes, a et b, forment un réseau de régulation génique et se régulent de la manière suivante : l'expression de a est activée par la présence du produit de l'expression de b et par son propre produit d'expression, quand celle de b est inhibée par le produit de l'expression de a en concurrence avec l'activation due à sa propre expression. Le réseau résultant est représenté Figure 1.4.* $\mathbf{x} = [x_1\ x_2]^T$ *est tel que* x_1 *représente la concentration du produit de l'expression du gène a et* x_2 *celle relative à b.*

Lorsque le passage à la limite est défini, on a $\dot{x_1} = f_1(x_1, x_2)$ *et* $\dot{x_2} = f_2(x_1, x_2)$*, avec* $\frac{\partial f_1}{\partial x_1} > 0$ *et* $\frac{\partial f_1}{\partial x_2} > 0$*, ainsi que* $\frac{\partial f_2}{\partial x_1} < 0$ *et* $\frac{\partial f_2}{\partial x_2} > 0$.

En fait, il existe trois grandes familles de modèles à base d'équations différentielles ordinaires : l'approche où $\mathbf{f} = \{f_1, \ldots, f_n\}^T : \mathbb{R}^{+n} \to \mathbb{R}^n$ est linéaire [41, 22, 34, 119], celle où elle est non-linéaire continue [59, 96, 63], et en dernier lieu celle où elle est linéaire par morceaux [45, 77, 38, 32, 11, 43, 44, 10] (c'est-à-dire non-linéaire du fait de discontinuités). Nous allons très brièvement décrire ces modèles ici.

Dans le cas linéaire, l'écriture mathématique du réseau de régulation génique est :

$$\dot{\mathbf{x}} = A\mathbf{x} + b, \quad A \in \mathcal{M}_n(\mathbb{R}), b \in \mathbb{R}^n. \tag{1.2}$$

A est une matrice[1] carrée de dimension n à composantes dans $\mathbb{R}$. Selon qu'une composante a_{ij} est négative (ou respectivement positive), on dira que l'action du gène j (pour simplifier, on omet de dire "gène correspondant à l'indice j") sur le gène i est une inhibition (ou respectivement une activation), car $\dot{x_i} = \sum_{k \neq j} a_{ik}x_k + a_{ij}x_j + b_i$, ce qui montre que la présence de l'élément j (pour simplifier, on omet de dire "le produit de l'expression du gène correspondant à l'indice j") tend à faire décroitre (respectivement accroitre) la présence de l'élément i. Ce type de modèles est particulièrement bien adapté autour d'un point d'équilibre du système. À l'aide d'un nombre limité d'observations, il est relativement simple de reconstruire les valeurs de A et de b.

Exemple 2. *En reprenant l'Exemple 1, s'il est modélisé par une écriture mathématique telle que celle de l'équation (1.2), on peut écrire*

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

1. $\mathcal{M}_{np}(\mathbb{R})$ est l'ensemble des matrices réelles à n lignes et à p colonnes. $\mathcal{M}_n(\mathbb{R})$ est l'ensemble des matrices carrées à n lignes et colonnes.

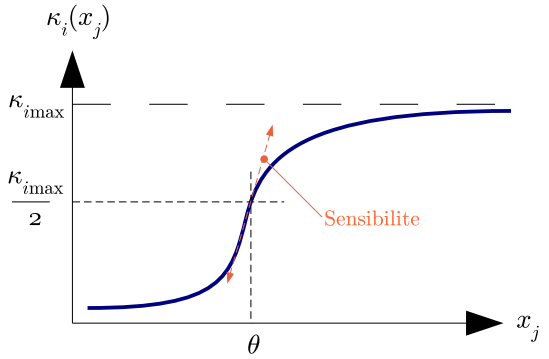
et on a alors $a_{11} > 0$, $a_{12} > 0$, $a_{21} < 0$ *et* $a_{22} > 0$.

Une description plus adéquate des aspects de la dynamique des réseaux de régulation génique nécessite l'introduction de davantage de paramètres et de non-linéarités [91]. Comme mentionné dans [77, 113], les réseaux de régulation génique sont souvent à dominante seuillés, i.e. les gènes sont activés seulement quand les concentrations de certains produits de l'expression génique sont compris entre certaines bornes définies. Les approches non-linéaires et affines-par-morceaux sont nécessaires à envisager dès lors que l'amplitude des phénomènes observés est grande, ou encore si l'on observe des phénomènes transitoires.

On distingue en général la régulation opérée sur l'expression elle-même (la synthèse) de celle opérée sur la dégradation du produit d'expression indépendamment des variations de dilution dues à la croissance de la cellule.

$$\dot{\mathbf{x}} = \overbrace{\kappa(\mathbf{x})}^{\text{terme de synthèse}} - \overbrace{\gamma(\mathbf{x})}^{\text{terme de dégradation}}, \quad \kappa : \mathbb{R}^{+n} \to \mathbb{R}^{+n}, \gamma : \mathbb{R}^{+n} \to \mathbb{R}^{+n}. \tag{1.3}$$

Typiquement, la fonction de synthèse par unité de temps reliant un gène à son régulateur est une fonction monotone bornée (classiquement la fonction de Hill, telle que représentée sur la Figure 1.5) : elle est croissante s'il s'agit d'une activation, décroissante dans le cas contraire. En effet, il est en général vérifié que l'action d'un gène sur un autre sature : ce niveau de saturation $\kappa_{\max}$, ainsi que la sensibilité σ de l'action et le seuil θ autour duquel l'effet régulateur se fait sentir sont autant de paramètres à régler.



$$\kappa_i(x_j) = \frac{\kappa_{i_{\max}} {x_j}^{\sigma}}{\theta^{\sigma} + {x_j}^{\sigma}}, \ \sigma \in \mathbb{N}^*$$

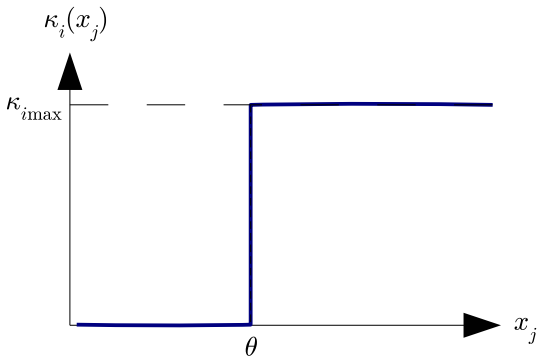
$$\kappa_i(\theta) = \frac{\kappa_{i_{\max}}}{2}$$

FIGURE 1.5 – Fonction de Hill.

Dans ce type de modèle, on a bien une action d'un gène j sur un gène i : en effet, le gène i voit son taux d'expression affecté sensiblement par les variations du taux de présence du produit de l'expression du gène j.

La multiplication des paramètres rend le traitement mathématique de ces modèles très difficiles. Il a donc été envisagé de se concentrer sur une forme légèrement simplifiée où la sensibilité aux actions régulatrices est admise comme étant infinie. On obtient alors des fonctions de régulation constantes par morceaux [40], dites en escalier (dont un exemple est représenté sur la Figure 1.6). On a pu montrer que des propriétés dynamiques essentielles sont ainsi préservées tout en facilitant grandement le traitement mathématique [45, 31].

Remarque : il existe aussi une représentation intermédiaire, dîte logoïde, conduisant à des modèles multi-affines [12, 11], qui introduit une pente au niveau du seuil afin de ne pas avoir de discontinuité (voir [85]).



$$\kappa_i(x_j) = \begin{cases} 0 & \text{si } x_j < \theta \\ \kappa_{i_{\max}} & \text{si } x_j > \theta \end{cases}$$

FIGURE 1.6 – Fonction en escalier.

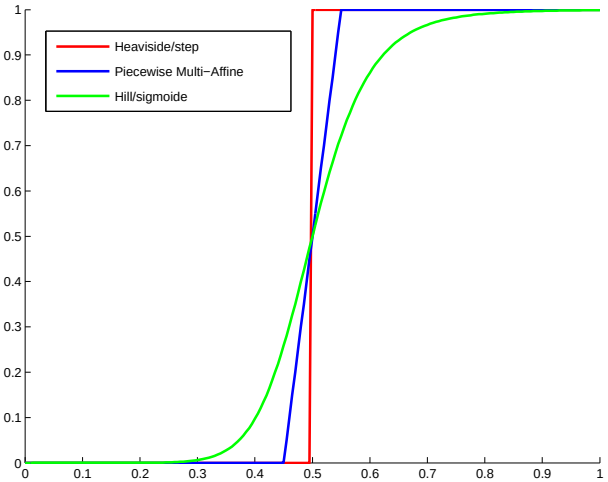


FIGURE 1.7 – Trois formes de fonctionnelles $\breve{\kappa}(x_i, \theta)$ pour le terme de synthèse régulé à partir d'un seuil θ (ici à 0.5 pour illustration) selon une concentration x_i en abscisse.

Exemple 3. *En reprenant l'Exemple 1 du réseau de régulation génique à deux gènes, nous allons illustrer l'effet d'idéaliser la fonction de régulation, nommée ici $\breve{\kappa}$, soit par une sigmoïde, soit par une logoïde, soit par un échelon : cette fonctionnelle est donnée Figure 1.7 avec un même seuil $\theta = 0.5$ pour comparaison.*

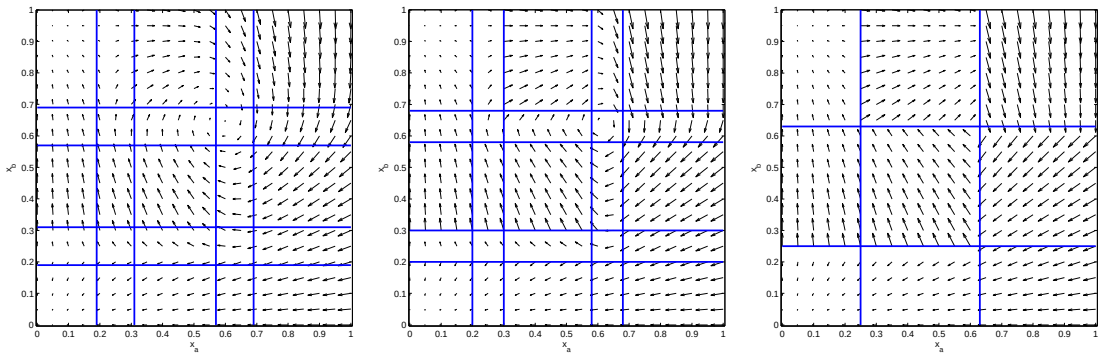
Le système d'équations dynamiques peut être écrit en reprenant la décomposition introduite à l'équation (1.3) :

$$\begin{cases} \dot{x_1} &= \kappa_1 \breve{\kappa}(x_1, \theta_{(1),1}) \breve{\kappa}(x_2, \theta_{(2),1}) - \gamma_1 x_1 \\ \dot{x_2} &= \kappa_2 \breve{\kappa}(x_2, \theta_{(2),2})(1 - \breve{\kappa}(x_1, \theta_{(1),2})) - \gamma_2 x_2 \end{cases}$$

avec les constantes $\kappa_1, \kappa_2, \gamma_1, \gamma_2$ réelles strictement positives, et les seuils sont supposés être tels que $\theta_{(1),1} < \theta_{(1),2}$, $\theta_{(2),2} < \theta_{(2),1}$.

On peut alors dessiner le flot du vecteur $\mathbf{x}$ dans l'espace d'état (x_1, x_2) : il apparait sur la Figure 1.8 selon les trois cas de fonctionnelles proposées Figure 1.7.

Il apparait que l'allure générale du flot est préservé ainsi que les attracteurs (les points d'équilibre).



(a) (b) (c)

FIGURE 1.8 – Flot du vecteur $\mathbf{x}$ dans l'espace des concentrations des produits de l'expression des gènes a et b du réseau de la Figure 1.4. (a) $\breve{\kappa}$ sigmoïdales : les barres bleues sont les mêmes que pour le cas suivant. (b) $\breve{\kappa}$ logoïdales : les barres bleues sont les iso-niveaux extrêmes des $\breve{\kappa}$. (c) $\breve{\kappa}$ constantes par morceaux : les barres bleues correspondent aux seuils $\theta_{(1),1}$ puis $\theta_{(1),2}$ dans la première direction, $\theta_{(2),2}$ puis $\theta_{(2),1}$ dans la deuxième.

Autour du vortex, on peut représenter des trajectoires pour quatre conditions initiales identiques dans les trois cas (Figure 1.9) : la courbure est modifiée, mais elles atteignent les mêmes domaines. Les variations du flot mesurées à l'aide de la fonction $|\dot{\mathbf{x}}|$ ne présentent pas de larges différences (Figure 1.10).

Dans la suite de ce manuscrit, la méthode d'identification proposée se rapportera au type de modèles affines-par-morceaux [77].

1.2.3 Modèles affines-par-morceaux de réseaux de régulation génique

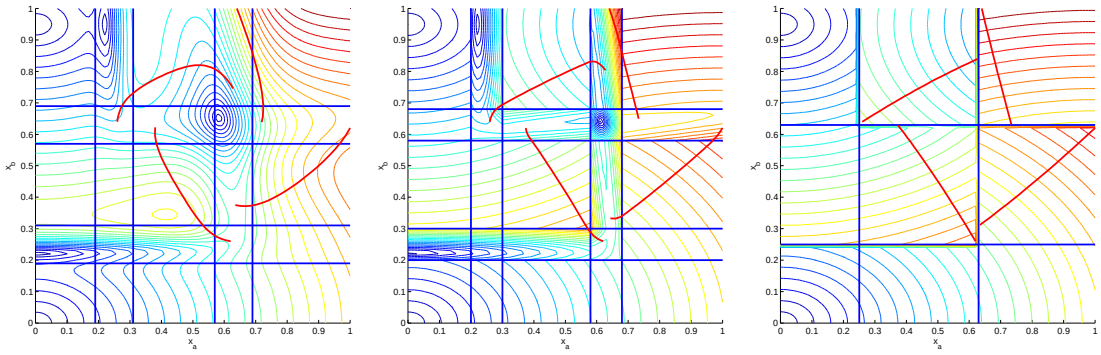
Présentons maintenant plus en détail les modèles affines-par-morceaux (APM) introduits dans la section précédente. L'utilisation de ces modèles pour l'analyse des réseaux de régulation génique a été proposée par [45], et étendue depuis, notamment par [50] qui adjoignirent la notion d'inclusion différentielle pour décrire analytiquement le comportement au niveau des discontinuités de la fonction de régulation.

Soit $\mathbf{x} \in \Omega \subset \mathbb{R}^{+n}$, le vecteur des concentrations. La dynamique de chaque composante peut être décomposée sous la forme :

$$\dot{x_i} = {\kappa_i}^{(j)} - {\gamma_i}^{(j)} x_i \text{ si } \mathbf{x} \in \Delta_j, j \in \{1, \ldots, s\}. \tag{1.4}$$

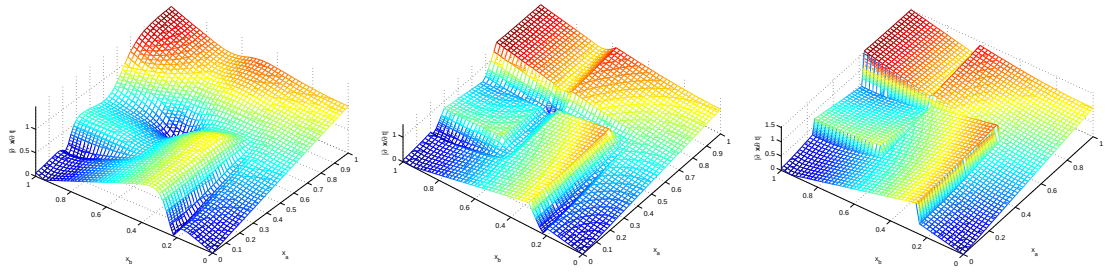
Les Δ_j sont des sous-domaines hyperrectangles de l'espace des concentrations Ω ($\Delta_j \subset \Omega$) qui le partitionnent ($\bigcup_{j=1}^{s} \Delta_j = \Omega$ et $\forall (j, j') \in \{1, \ldots, s\}^2, j \neq j' \Rightarrow \Delta_j \cap \Delta_{j'} = \emptyset$). À l'intérieur de ces sous-domaines, la variation temporelle de chaque composante est donc la somme d'un terme de synthèse représenté par la constante positive ${\kappa_i}^{(j)}$ et d'un terme négatif de dégradation à coefficient réel constant ${\gamma_i}^{(j)}$. La résolution de cette équation du premier ordre à coefficient constant est directe. À l'intérieur des sous-domaines partitionnant Ω, le système présente donc un comportement dynamique d'ordre un qu'il n'est pas compliqué de traiter.

Comment en sommes-nous arrivés à partitionner l'espace des concentrations ? Dans le chapitre 1.2.2 précédant, nous avons vu que la fonction associée à la régulation est une



(a) (b) (c)

FIGURE 1.9 – Dans le même espace des concentrations que dans la Figure 1.8 et pour les mêmes trois cas (a), (b) et (c) que cette même figure, on représente quatre segments de trajectoires avec les mêmes conditions initiales dans les trois cas. Les barres bleues sont les mêmes que celles de la Figure 1.8 et les courbes en dégradé de couleurs sont les iso-niveaux de la surface représentée sur la Figure 1.10.



(a) (b) (c)

FIGURE 1.10 – Dans le même espace des concentrations que dans la Figure 1.8 et pour les mêmes trois cas (a), (b) et (c) que cette même figure, on a représenté la fonction $|\dot{\mathbf{x}}|$, avec un dégradé coloré pour en facilité la lisibilité.

fonction en escalier. Prenons le terme de synthèse $\kappa_i(\mathbf{x}) : \mathbb{R}^{+n} \rightarrow \mathbb{R}^+$ qui décrit comment le taux de synthèse du produit de l'expression du gène i dépend du vecteur des concentrations de l'ensemble des protéines étudiées et supposant former un système fermé. Cette fonction représente l'action de toutes ces concentrations au niveau de l'expression génique : certaines actions sont en fait indépendantes et peuvent se produire en sus des autres, mais beaucoup se combinent. En effet, non seulement nous avons vu au chapitre 1.1.2 que les réseaux de régulation génique projettent sur le plan génique des mécanismes moléculaires complexes s'interconnectant (l'exemple le plus simple pour se représenter cela est la formation d'un complexe de protéines qui a alors une fonction régulatrice), mais aussi il y a la concurrence des effets régulants (par exemple, certains régulateurs ayant davantage d'affinité chimique avec la région promotrice du gène cible auront la priorité). Étant donné que nous faisons l'hypothèse que la sensibilité de la régulation est infinie (l'action régulatrice est active ou ne l'est pas), il est possible d'utiliser la logique combinatoire pour représenter la manière par laquelle les effets se combinent. On écrira ainsi, avec L_i l'ensemble d'entiers naturels éventuellement vide, κ_{il} des constantes réelles positives et $\breve{s}_{il}$ des fonctions de $\mathbb{R}^n \rightarrow \{0,1\}$:

$$\kappa_i(\mathbf{x}) = \sum_{l \in L_i} \kappa_{il} \breve{s}_{il}(\mathbf{x}) \tag{1.5}$$

selon qu'une certaine condition est remplie à l'état $\mathbf{x}$ [86] amenant les $\breve{s}_{il}$ à valoir 0 ou 1. En l'occurrence, la condition concerne le fait que les concentrations dépassent les seuils idéalisant le fait qu'elles sont activatrices, ou qu'elles soient en deçà les seuils idéalisant le fait qu'elles sont inhibitrices, selon le cas. Ces seuils sont dits *seuils de régulation*.

La fonction de base utilisée pour écrire les $\breve{s}_{il}(\mathbf{x})$ est une fonction en escalier de type :

$$s(x_i, \theta) = \begin{cases} 0 & \text{si} \quad x_i < \theta \\ 1 & \text{si} \quad x_i > \theta \end{cases}, \text{ pour } i \in \{1, \ldots, n\}, \text{ composante où intervient le seuil.}$$

La notation $s^+(x_i, \theta) = s(x_i, \theta)$ est aussi utilisée, ainsi que $s^-(x_i, \theta) = 1 - s(x_i, \theta)$. La négation logique NON devient l'inverse d'avec 1. L'opération logique ET (conjonction) se traduit par le produit des fonctions en escalier. L'opération logique OU (disjonction) se traduit par leur somme (à ceci près que $1 + 1 = 1$, et non pas 2 ou 0, ce qui fait que l'on écrit pour lever toute ambigüité la forme duale issue des lois de De Morgan : a ou b = non(non(a) et non(b))= $1 - (1 - a)(1 - b)$). Au final, la combinaison logique des conditions est traduit dans la condition $\breve{s}_{il}(\mathbf{x})$, qui est un polynôme de fonctions en escalier.

L'écriture du terme de dégradation est identique, à ceci près qu'il constitue le terme d'ordre 1 de l'équation différentielle :

$$\gamma_i(\mathbf{x}) = \left(\sum_{l \in L'_i} \gamma_{il} \breve{s}'_{il}(\mathbf{x}) \right) x_i \tag{1.6}$$

où L'_i est un ensemble d'entiers naturels éventuellement vide, γ_{il} des constantes positives réelles et $\breve{s}'_{il}(\mathbf{x})$ des fonctions de $\mathbb{R}^n \rightarrow \{0,1\}$. Sauf si toutes les molécules ont disparu, le processus de dégradation a toujours lieu : $\gamma_i(\mathbf{x}) > 0$.

L'équation du système (1.3) $\dot{\mathbf{x}} = \kappa(\mathbf{x}) - \gamma(\mathbf{x})$ constitue donc un système d'équations différentielles d'ordre 1 couplées : cela est vrai partout sauf éventuellement là où les concentrations ont franchi au moins un seuil idéalisant les actions de régulation. En prenant l'ensemble fini de tous ces seuils, au moyen des hyperplans parallèles aux axes leur correspondant, on peut partitionner l'espace des concentrations $\Omega \subset \mathbb{R}^n$ en s sous-domaines Δ_j,

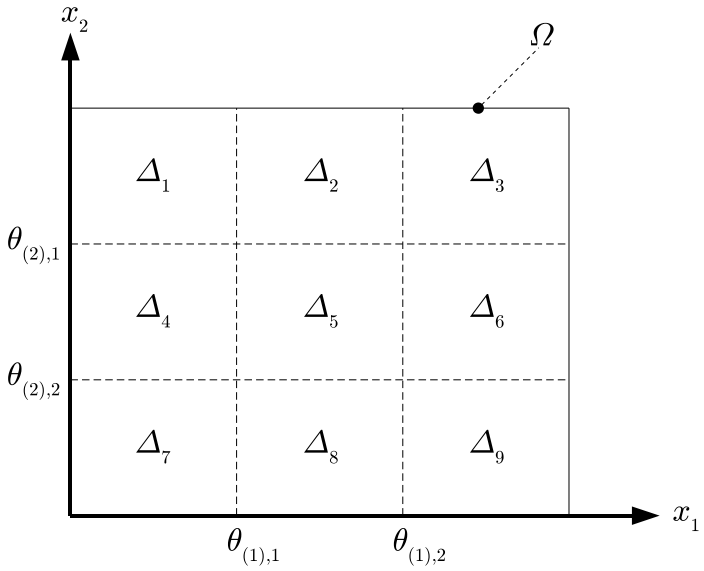


FIGURE 1.11 – L'espace d'état Ω pour l'Exemple 4 est divisé en neuf domaines de régulation rectangles qui le partitionnent.

$j \in \{1, \ldots, s\}$, qui sont par construction hyperrectangles. On appelle ces sous-domaines les *domaines de régulation* [77].

Exemple 4. *En reprenant l'Exemple 1 du réseau de régulation génique à deux gènes, dans la cadre d'un modèle APM, la dynamique de ce réseau peut être écrite sous la forme :*

$$\left\{ \begin{array}{l} \dot{x_1} = \kappa_{11} \breve{s}_{11}(\mathbf{x}) - \gamma_{11} \breve{s}'_{11}(\mathbf{x})\ x_1 \\ \dot{x_2} = \kappa_{21} \breve{s}_{21}(\mathbf{x}) - \gamma_{21} \breve{s}'_{21}(\mathbf{x})\ x_2 \end{array} \right. \quad (\sum)$$

$$\textit{avec} \left\{ \begin{array}{lcl} \breve{s}_{11}(\mathbf{x}) & = & s^+(x_1, \theta_{(1),1}) s^+(x_2, \theta_{(2),1}) \\ \breve{s}'_{11}(\mathbf{x}) & = & 1 \\ \breve{s}_{21}(\mathbf{x}) & = & s^-(x_1, \theta_{(1),2}) s^+(x_2, \theta_{(2),2}) \\ \breve{s}'_{21}(\mathbf{x}) & = & 1 \end{array} \right.$$

dans la mesure où l'on considère que l'expression du gène a est activée quand $\left[\left(x_1 > \theta_{(1),1}\right)\ ET\ \left(x_2 > \theta_{(2),1}\right)\right]$ *(la présence des deux produits d'expression est nécessaire), et l'expression du gène b est activée si* $\left[\left(x_2 > \theta_{(2),2}\right)\ ET\ \left(x_1 < \theta_{(1),2}\right)\right]$ *(l'auto-activation de b est limitée par la présence du produits d'expression de a).*

Expression du gène b	$x_1 < \theta_{(1),2}$	$x_1 > \theta_{(1),2}$
$x_2 < \theta_{(2),2}$	*inhibée*	*inhibée*
$x_2 > \theta_{(2),2}$	*activée*	*inhibée*

L'ensemble des seuils $\{\theta_{(1),1}, \theta_{(1),2}, \theta_{(2),1}, \theta_{(2),2}\}$ *intervenant sur les concentrations du produit de a pour les deux premiers, du produit de b pour les deux derniers, divisent l'espace d'état* Ω *en neuf domaines de régulation* $\{\Delta_j\}_{j=1}^9$*, comme on le voit sur la Figure 1.11.*

Le système dynamique (1.3) décomposé en (1.4) peut être réécrit sous la forme APM matricielle : pour $j \in \{1, \ldots, s\}$,

$$\dot{\mathbf{x}} = b^{(j)} - A^{(j)}\mathbf{x} \text{ si } \mathbf{x} \in \Delta_j \quad (1.7)$$

avec $b^{(j)} = \left[b_1^{(j)} \dots b_n^{(j)}\right]^T$ $(b_i^{(j)} \geq 0)$ et $A^{(j)} = \mathrm{diag}(A_1^{(j)}, \dots, A_n^{(j)})$ $(A_i^{(j)} > 0)$.

$(A^{(j)}, b^{(j)}, \Delta_j)$ permet de décrire chacun des modes dynamiques du réseau de régulation génique.

Exemple 5. *En poursuivant l'Exemple 4 :*

$$\begin{aligned} b^{(2)} &= [\kappa_{11} \;\; \kappa_{21}]^T, \\ b^{(3)} &= [\kappa_{11} \;\; 0]^T, \\ b^{(1)} = b^{(4)} = b^{(5)} &= [0 \;\; \kappa_{21}]^T, \\ b^{(6)} = b^{(7)} = b^{(8)} = b^{(9)} &= [0 \quad 0]^T, \\ \forall j \in \{1, \dots, s\}, \quad A^{(j)} &= \mathrm{diag}(\gamma_{11}, \gamma_{21}). \end{aligned}$$

En reprenant la définition de [46], on dira que le *point focal* associé au mode dynamique $j \in \{1, \dots, s\}$ est :

$$\Phi^{(j)} = {A^{(j)}}^{-1} \cdot b^{(j)} \tag{1.8}$$

($A^{(j)}$ est évidemment inversible dans la mesure où les constantes associées à la dégradation ne peuvent être nulles).

À l'intérieur de chaque domaine de régulation, la trajectoire $\mathbf{x}(t)$ est exponentielle. Il est aisé de montrer qu'elle tend de façon monotone vers son point focal qui est l'unique solution de $\dot{\mathbf{x}} = 0$ (point stationnaire régulier). Si cette trajectoire vient à franchir un des seuils de régulation (c'est-à-dire à sortir du domaine de régulation auquel elle appartient), on ne peut pas décrire trivialement la trajectoire. Dans le cas le plus simple, l'effet du seuil est masqué dans la combinaison des effets régulateurs et les paramètres cinétiques restent inchangés : la trajectoire se poursuit identiquement. Il est possible que le franchissement d'un seuil entraine la modification des paramètres de la dynamique (synthèse ou dégradation). Si la trajectoire se poursuit dans le nouveau domaine de régulation, on observe une discontinuité dynamique qui est la trace d'un franchissement de seuil. Dans le cas contraire, ce n'est pas une unique trajectoire qui est possible, mais éventuellement un ensemble de trajectoires : on arrive sur les limites de la formulation APM, et on dit que la trajectoire est *glissante*.

Sans vouloir donner ici tous les détails concernant les solutions glissantes, nous esquissons dans ce paragraphe seulement le trait général. Le problème lié aux trajectoires glissantes peut être appréhendé grâce à l'approche de [49, 50], qui propose d'utiliser une extension des équations différentielles à des inclusions différentielles définies sur des sous-parties des hyperplans correspondant aux seuils et de leurs intersections, qui seront appelées les *domaines de transition*. À l'intérieur de ces domaines, la dynamique du système peut éventuellement ne pas être déterminée de façon unique (les paramètres affines appartenant à des intervalles, ils peuvent prendre une infinité bornée de valeurs) : pour de tel mode, dit glissant, la notion de point focal est remplacée par celle d'ensemble focal. Si cet ensemble se trouve à l'intérieur du domaine de transition à laquelle la trajectoire appartient, celle-ci convergera monotoniquement vers lui, puis elle en restera prisonnière tout en pouvant fluctuer. Si ce n'est pas le cas, la trajectoire va "glisser" entre les domaines de régulation jusqu'à un nouveau franchissement de seuil, auquel cas, la trajectoire peut éventuellement être à nouveau caractérisée de façon unique, à moins que l'on ne soit de nouveau dans une configuration glissante. Un exemple est donné Figure 1.12.

Le travail qui suit tente d'exploiter les qualités brièvement décrites d'un réseau de régulation génique APM pour en faire la reconstruction à partir de données de mesure.

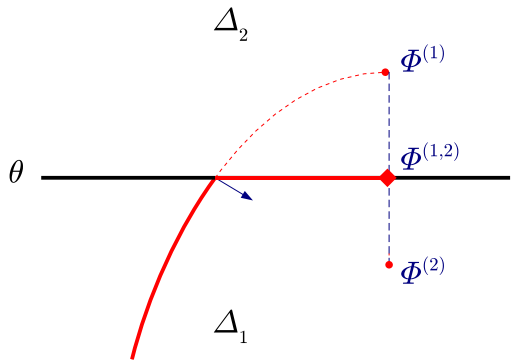


FIGURE 1.12 – Trajectoire glissante - La trajectoire évolue dans le domaine de régulation Δ_1 exponentiellement vers le point focal $\Phi^{(1)}$ se trouvant dans Δ_2. Au moment de franchir le seuil θ séparant Δ_1 de Δ_2, la dynamique est modifiée de manière à ce que le point focal devient $\Phi^{(2)} \in \Delta_1$. La trajectoire va alors glisser sur l'hyperplan de transition jusqu'à l'ensemble focal $\Phi^{(1,2)}$ qui, dans notre exemple, se résume à un point. Un état stable y est atteint.

Chapitre 2

Inférence de réseaux de régulation génique

2.1 Le problème de l'inférence de réseau

L'analyse à grande échelle de profils d'expression génique, puis de motifs d'expression, a été encouragée par le présupposé que l'information concernant un état fonctionnel d'un organisme est grandement déterminé par l'expression de ses gènes. Ce mécanisme peut être abstrait au moyen de boucles de rétroaction sur les gènes, qui forment un réseau de régulation génique : les flux informationnels attachés à des motifs d'expression génique viennent entre autre influencer l'expression des gènes à travers un enchainement de processus intracellulaires voire même intercellulaires. La réponse d'un organisme a un stimulus donné, ses dysfonctionnements de type pathologique, voire sa dégénérescence, peuvent alors être envisagés sous l'angle du comportement du réseau de régulation génique sous-jacent : il s'agit donc d'une approche conceptuelle tout à fait puissante [112].

Une fois les motifs d'expression cernés, le but ultime de l'analyse des données d'expression est l'identification du réseau de régulation, c'est-à-dire l'inférence des interactions causales "directes" entre les gènes et la caractérisation quantitative des ces interactions selon un modèle séquentiel ou dynamique choisi. On dira *direct* si l'on ne peut pas envisager d'interaction génique intermédiaire susceptible de justifier le lien de causalité observé : ainsi, si le gène i régule le gène j lui-même régulant le gène k, il existe bien un mécanisme de régulation justifiant l'adaptation de l'état du niveau d'expression du gène k lorsque celui du gène i se modifie, cependant, l'expression du gène j étant observée, on dira que la régulation du gène i sur le gène k est indirecte.

De manière à effectuer une inférence de réseaux qui ait un sens, il est important que chaque gène soit observé avec diverses conditions : l'observation en série temporelle de réponses à des perturbations est aujourd'hui rendue possible pour des mesures en parallèle à haute résolution. Dans tous les cas, la particularité du défi liant l'inférence du réseau de régulation génique aux méthodes issues du traitement du signal est la quantité et la qualité des données disponibles.

À vrai dire la caractérisation complète du comportement des réseaux n'est pas quelque chose de neuf en biologie moléculaire : ce qui l'est, c'est l'approche systématique, qui tente une reconstruction globale à partir d'une observation globale. Les approches antérieures tentaient de caractériser des interactions gène-à-gène en essayant diverses combinaisons de stress sur les organismes, ou de perturbations irréversibles. Ainsi, on peut manipuler l'activité d'un gène en utilisant diverses techniques issues de la biologie moléculaire. Cela inclut les techniques permettant de sur-exprimer un gène choisi, ainsi que les techniques permettant d'obtenir l'effet inverse, jusqu'à la suppression totale du gène lui-même (on obtient des mutants). Mais ces méthodes supposent des modifications systématiques dont on contrôle mal les effets secondaires. Il a toutefois été envisagé d'optimiser le choix des gènes à perturber [111] de manière à maximiser la quantité d'information extraite à chaque expérience : la procédure est de fait itérative sur la séquence d'expériences à entreprendre.

Cette approche peut ressembler à une identification en boucle ouverte : l'effet de perturbations contrôlées est directement mesuré pour inférer les interactions. Cependant, le

système biologique étudié est intrinsèquement un procédé en boucle fermée : ce sont les propriétés de cette rétroaction qui présentent pour nous un intérêt. Ainsi, il se peut très bien qu'une causalité directe existe entre l'expression d'un gène et celle d'un autre, sans que l'action possible ne survienne jamais dans l'organisme vivant. Ou encore, il se peut qu'une causalité directe ne soit pas mesurable alors que la fermeture de la boucle de régulation rend la notion de causalité pertinente.

La méthode d'identification classique en automatique consiste alors à construire un prédicteur ajustable qui précise l'ensemble des modèles possibles, que l'on restreint de manière itérative de façon à minimiser l'écart avec la sortie bruitée mesurée du procédé physique. Les algorithmes exploitent en général l'erreur d'adaptation. Cependant, dans le contexte actuel de la biologie moléculaire, cette approche n'est pas envisageable. Un système d'organismes vivants ne peut pas être replacé plusieurs fois dans des conditions initiales identiques. Au mieux, on s'efforce de synchroniser ces organismes sur des cycles que l'on peut initialiser de la même manière.

Cela change complètement la perspective. Le type d'expériences qu'il est raisonnable d'envisager sont des expériences de lâché : une modification ponctuelle intervient au début de l'observation, et nous mesurons comment le système réagit de lui-même sur une fenêtre de temps finie. De plus, cette modification ponctuelle n'est pas totalement arbitraire : les organismes vivants vont se placer dans un état stable donné et réagiront à une perturbation dans la limite de ce qu'ils peuvent pour rejoindre un autre état stable.

On peut dire que les différentes méthodes d'inférence proposées se divisent alors en deux familles : celles qui exploitent la différence entre les deux états stables, celles qui exploitent la réponse transitoire du système. Nous allons en décrire quelques unes dans ce qui suit.

2.2 Quelques approches existantes pour l'inférence de réseaux

Il existe une très longue série de méthodes maintenant disponibles dans la littérature (pour quelques revues récentes, voir [42, 118, 23, 72, 126]). Les domaines les plus actifs sont sans nul doute les méthodes à base de réseaux (dynamiques) bayésiens [124] qui offrent une très grande souplesse pour modéliser divers réseaux biologiques, mais qui présentent quelques difficultés pour inférer les boucles de rétroaction, ce qui constituent essentiellement les réseaux de régulation - et les méthodes pour modèles linéaires [4] -particulièrement bien adaptées pour les contextes où peu d'expériences, peu de points de mesure et peu de stimulations sont possibles.

Il faut cependant relever que, si les résultats des ces techniques peuvent être décrits d'une même manière, en l'occurrence un graphe représentant les interactions (souvent orienté, parfois valué), les "interactions" ne caractérisent que rarement le même objet. D'ailleurs, il ne fait aucun sens de comparer sans précaution les résultats de deux méthodes distinctes, même pour un même ensemble de données. De fait, la comparaison des méthodes entre elles n'est pas évidente : depuis quelques années, la compétition DREAM [1] tente de proposer des problèmes autorisant une certaine mise en perspective.

Nous avons choisi de ne présenter ici, dans les grands traits, que quelques méthodes dans la mesure où, d'une part, elles traitent des séries temporelles de données, et, d'autre part, elles se rapprochent du formalisme qui est le nôtre.

1. http ://wiki.c2b2.columbia/edu/dream/index.php/The_DREAM_Project : en 2008, quarante équipes ont participé, preuve supplémentaire de l'intérêt pour le sujet.

2.2.1 Inversion de modèle linéaire

Peeters & Westra

Cette méthode a été introduite dans [83].

Soit $\mathbf{x}$ le vecteur de concentration du produit de l'expression des N gènes observés. La représentation adoptée est la représentation linéaire différentielle classique :

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}. \tag{2.1}$$

Les variations de l'expression d'un gène $\dot{x}_i(t)$ est donc contrôlée par l'expression des autres gènes proportionnellement à leur présence $\mathbf{x}$ ($A \in \mathcal{M}_N(\mathbb{R})$), ainsi que par des grandeurs de contrôle extérieur $\mathbf{u}$ ($B \in \mathcal{M}_m(\mathbb{R})$). On dispose d'observations à $t_1, \ldots, t_M$: on écrira $\mathbf{x}[k] \hat{=} \mathbf{x}(t_k)$. En passant à l'écriture matricielle ($X = (\mathbf{x}[1] \cdots \mathbf{x}[M])$ et $U = (\mathbf{u}[1] \cdots \mathbf{u}[M])$), l'Équation (2.1) devient :

$$\dot{X} = AX + BU \tag{2.2}$$

et par transposition $\dot{X}^T = X^T A^T + U^T B^T$. On peut alors écrire, pour $i \in \{1, \ldots, N\}$:

$$\delta_i = X^T \mathbf{a}_i + U^T \mathbf{b}_i \tag{2.3}$$

avec $\mathbf{a}_i$ la i-ième ligne de A, $\mathbf{b}_i$ celle de B, et δ_i celle de $\dot{X}$.

Il s'agit de N systèmes de M équations linéaires à $M \times (N + m)$ inconnues : chacun de ces systèmes est donc particulièrement sous-déterminé.

Les auteurs se ramènent alors à un problème de minimisation pour la norme ℓ_1 qui peut être résolu par Programmation Linéaire [80].

Une extension a été proposée dans [65] pour des modèles APM. L'Équation (2.2) devient dans ce cas :

$$\dot{X} = A^{(\sigma)} X + B^{(\sigma)} U \tag{2.4}$$

où σ représente l'indice d'appartenance à l'un des modes dynamiques. Les poids $w_{k\sigma}$ sont les fonctions de l'observation en k du système, qui valent 0 si $\{\mathbf{x}[k], \mathbf{u}[k]\}$ n'appartient pas au mode σ, ou prend des valeurs jusqu'à 1 autrement. De plus, $\sum_\sigma w_{k\sigma} = 1$.

S'il est possible de résoudre comme précédemment l'équation $\dot{X} = H_1 X + H_2 U$, obtenue en compilant les Équations (2.4) pour tous les σ, le problème devient alors d'inverser :

$$\begin{array}{rcl} H_1 & = & f_1(W, A^{(1)}, \ldots, B^{(1)}, \ldots), \\ H_2 & = & f_2(W, A^{(1)}, \ldots, B^{(1)}, \ldots). \end{array} \tag{2.5}$$

Les auteurs alternent alors itérativement l'estimation des $A^{(\sigma)}$ et $B^{(\sigma)}$, les poids étant fixés, avec l'estimation de W, les matrices $A^{(\sigma)}$ et $B^{(\sigma)}$ étant fixées : cette estimation se fait en minimisant

$$\sum_{k,\sigma} w_{k\sigma} \left\| A^{(\sigma)} \mathbf{x}[k] + B^{(\sigma)} \mathbf{u}[k] - \dot{\mathbf{x}}[k] \right\|_{\ell_1}$$

(par Programmation Linéaire).

L'approche de la méthode décrite dans [92] est très similaire. Cependant, elle introduit de manière intéressante une réduction de l'espace des gènes par construction optimale de nuées ("fuzzy C-means" modifié). Cela permet de supprimer l'effet lié au fait que certains gènes en régulent énormément d'autres et font apparaitre des motifs de co-expression.

Bansal & di Bernardo

Utilisant le même formalisme, la méthode décrite dans [5] est légèrement différente. Ici, les auteurs proposent de se débarrasser du terme de dérivée du premier ordre dans (2.2) par intégration :

$$X[k] = A\int_0^{t_k} X(t)dt + B\int_0^{t_k} U(t)dt \tag{2.6}$$

(avec $X(0) = 0$), ce qui devient par approximation :

$$X[k] = A\sum_{i=1}^{k} \delta t X[i] + B\sum_{i=1}^{k} \delta t U[i] \tag{2.7}$$

où δt est le pas d'échantillonnage. L'Équation (2.7) est réécrite sous la forme :

$$Y[k] = AH[k] \text{ avec } \begin{cases} Y[k] = X[k] - B\sum_{i=1}^{k} \delta t U[i], \\ H[k] = \sum_{i=1}^{k} \delta t X[i]. \end{cases} \tag{2.8}$$

En compilant les M équations sous forme matricielle, on arrive à la forme compacte :

$$Y = AH. \tag{2.9}$$

Les auteurs proposent de résoudre le système décrit par (2.9) par pseudo-inversion : en prenant A_i la i-ième ligne de A et Y_i celle de Y, alors on se ramène à $A_i = Y_i H^T (HH^T + \lambda_i I)^{-1}$, avec I matrice identité, et λ_i la plus petite valeur singulière de $H^T Y_i^T$. Cette équation peut être résolue si $M \geq N$, ce qui est rarement le cas. Par contre, en faisant l'hypothèse qu'un gène donné n'est régulé que par au maximum $K \leq M$ gènes, la résolution est rendue possible : il faut considérer les combinaisons de K éléments parmi N.

Les auteurs utilisent alors une méthode de régression "forward step-wise" : on considère toutes les N solutions pour $K = 1$, et on les range en fonction de l'erreur au sens des moindres-carrés pour ne garder que les D solutions avec l'erreur la plus faible. Avec ce qu'il reste, on recommence pour $K = 2$, et ainsi de suite jusqu'à $K = M$.

2.2.2 Inversion de modèle non-linéaire

Müller et *al.*

Dans [79], une ambitieuse méthode est proposée afin d'inférer un système de n équations différentielles non-linéaires de m paramètres représentés par un vecteur x :

$$\dot{y}(t) = f(y(t), x) \text{ avec } y(0) = y_0. \tag{2.10}$$

La structure est entièrement connue : seules les valeurs des paramètres sont inconnues.

La résolution du système conduit à l'opérateur de solutions pour $N + 1$ observations

$$\begin{aligned} F : \quad & U_x \to (U_y)^{N+1} \\ & x \mapsto (y(t_0), \ldots, y(t_N)) \end{aligned} \tag{2.11}$$

avec $U_x \subset \mathbb{R}^m$ l'espace des paramètres possibles, $U_y \subset \mathbb{R}^n$ l'espace des solutions envisageables, $t_0, \ldots, t_N$ étant les temps discrets des observations.

Dans le cadre du problème inverse, on cherche à identifier x à partir des séries temporelles bruitées de mesure y^δ, vérifiant $F(x) = y^\delta$ (on obtient $F(x)$ par intégration numérique). En présence de bruit, la résolution, si elle est parfois possible, conduit à des résultats abérants et est extrêmement instable. Les auteurs proposent de se ramener au problème :

$$\text{trouver } x \text{ minimisant } \left\| F(x) - y^\delta \right\|_{\ell_2}^2 \tag{2.12}$$

Cependant, ce problème n'est pas correctement posé au sens de Hadamard [51]. Les auteurs utilisent alors une stratégie de régularisation (dite de Tikhonov), se ramenant au problème :

$$\text{trouver } x \text{ minimisant } \left\| F(x) - y^\delta \right\|_{\ell_2}^2 + \alpha \left\| x - x^* \right\|_{\ell_2}^2 \tag{2.13}$$

où α est le paramètre de régularisation et x^* l'estimation *a priori* de la solution. α doit être le plus petit possible mais pas trop pour ne pas retomber dans les travers du problème non-régularisé.

Pour trouver un bon compromis, l'erreur totale (entre la solution régularisée x_α^δ pour des mesures bruitées et la solution idéale pour le cas sans bruit et $\alpha = 0$) est bornée par la décomposition :

$$\left\| x_\alpha^\delta - x_0^0 \right\| \leq \left\| x_\alpha^\delta - x_\alpha^0 \right\| + \left\| x_\alpha^0 - x_0^0 \right\|. \tag{2.14}$$

$\left\| x_\alpha^\delta - x_\alpha^0 \right\|$ est l'erreur de propagation, liée au bruit sur les données : du fait de l'instabilité dans le cas non-régularisé, cette erreur explose lorsque $\alpha \to 0$. $\left\| x_\alpha^0 - x_0^0 \right\|$ est l'erreur de régularisation (liée à α) : elle tend à s'annuler lorsque $\alpha \to 0$. Pour éviter cet effet ainsi que la sur-estimation qui en est l'origine, une borne inférieure au choix de α est donc appliquée.

Perkins et *al.*

L'approche la plus complémentaire avec la méthode que nous décrirons dans les chapitres suivants est celle décrite dans [84]. Seule la synthèse est régulée. Le produit de l'expression d'un gène est représenté de façon continue ($x_i(t)$ pour l'élément i), qui implique un état logique $X_i(t) \in \{0, 1\}$ (il existe un seul seuil de régulation par gène, qui est fixé), et c'est cet état qui est impliqué dans la régulation de l'expression des gènes :

$$\dot{x_i}(t) = f_i(X_{R_i}(t)) - x_i(t), \tag{2.15}$$

avec $X_{R_i}(t)$ vecteur contenant les états logiques de tous les régulateurs du gène i à l'instant t.

Les auteurs cherchent à déterminer la structure et les fonctions de régulation du système représenté par (2.15), en se basant sur un ensemble de données temporelles non bruitées de concentration. Pour cela, un ensemble de trois règles est défini, qui travaille sur la série d'états logiques observés, en rapport avec les taux de production correspondants. Cette approche permet aussi de reconstruire des résultats même quand des points sont manquants dans les séries de mesures relevées.

Chapitre

3 Identification des réseaux de régulation génique : l'approche APM

Comme nous l'avons vu au Chapitre 1.2.3, la représentation affine-par-morceaux (APM) fait que le système représenté a une dynamique continue jusqu'à l'avènement d'évènements intervenant à des instants discrets que sont les franchissements des seuils idéalisant la régulation.

En automatique, il existe un formalisme plus généraliste qui se concentre sur des systèmes à évolution continue dont le comportement est contrôlé par des phénomènes discrets. Il s'agit des *systèmes hybrides*. Dans un premier temps nous montrerons comment il est possible d'adapter le formalisme des réseaux de régulation génique APM à celui des systèmes hybrides.

De nombreux travaux existent quant à leur identification : les systèmes hybrides connaissent en effet des applications très variées et leur développement a conduit à une théorie formelle aboutie et à des outils puissants. Pour les travaux portant sur l'identification, nous présenterons la problématique des approches existantes exploitable pour résoudre notre problème d'inférence de réseaux. Cependant, nous verrons que leur application au sens strict ne conduit pas à des résultats biologiquement probants. Cela nous permettra d'introduire notre approche qui sera développée dans la partie suivante.

3.1 Approche APM et systèmes hybrides

Un système dynamique sera dans notre cas représenté au moyen d'un vecteur $\mathbf{x} \in \mathbb{R}^n$ de variables dites d'*état* évoluant dans un espace dit d'état, formant une *trajectoire* dont on observe une sous-partie au moyen d'une *observation* représentée par le vecteur $\mathbf{y} \in \mathbb{R}^q$ de variables dites de sortie. L'évolution de $\mathbf{x}$ peut être commandée au moyen de variables d'entrée représentées par un vecteur $\mathbf{u} \in \mathbb{R}^p$. $(n, p, q) \in \mathbb{N}^3$.

Pour un système hybride, la dynamique du système est affine mais des évènements intervenants à des temps discrets peuvent la modifier. La séquence temporelle de ces évènements est représentée au moyen de σ à valeur dans $\mathbb{N}$. On écrit alors :

$$\left\{ \begin{array}{rcl} \dot{\mathbf{x}} & = & A^{(\sigma)}\mathbf{x} + B^{(\sigma)}\mathbf{u} + \mathbf{f}^{(\sigma)} + \mathbf{w} \\ \mathbf{y} & = & C^{(\sigma)}\mathbf{x} + D^{(\sigma)}\mathbf{u} + \mathbf{g}^{(\sigma)} + \mathbf{v} \end{array} \right. \tag{3.1}$$

où $\mathbf{w} \in \mathbb{R}^n$ et $\mathbf{v} \in \mathbb{R}^q$ sont des termes de bruit additionnel. Pour tout $t \in \mathbb{R}^+$, $\sigma(t)$ est l'état discret décrivant dans quelle dynamique affine le système se comporte à l'instant t : on fait en général l'hypothèse que σ peut prendre un nombre fini de valeurs : $\sigma : \mathbb{R}^+ \to \{1, \ldots, s\}$ (on dit alors qu'il existe s sous-modèles affines). En pratique, σ peut être non seulement une fonction du temps t, mais encore de l'état $\mathbf{x}$ ou/et de l'entrée $\mathbf{u}$. Pour les modèles APM, la séquence des transitions est généralement donnée par :

$$\forall t, \sigma(t) = j \text{ si et seulement si } \mathbf{r} = \left[\mathbf{x}^T \mathbf{u}^T\right]^T \in \Omega_j \tag{3.2}$$

avec $\{\Omega_j\}_{j=1}^s$ formant une partition complète de l'espace $\Omega \subseteq \mathbb{R}^{n+p}$ à l'intérieur duquel le vecteur $\mathbf{r}$ est amené à évoluer.

En pratique, l'observation est échantillonnée et on se ramène à un temps échantillonné (discret). Soit $\{t_1, \ldots, t_N\}$ la séquence des temps d'échantillonnage. On écrira k à la place de t_k pour $k \in \{1, \ldots, N\}$. Cela donne la formulation équivalente à temps discret :

$$\left\{ \begin{array}{rcl} \mathbf{x}\,[k+1] & = & A^{(\sigma[k])}\mathbf{x}\,[k] + B^{(\sigma[k])}\mathbf{u}\,[k] + \mathbf{f}^{(\sigma[k])} + \mathbf{w} \\ \mathbf{y}\,[k] & = & C^{(\sigma[k])}\mathbf{x}\,[k] + D^{(\sigma[k])}\mathbf{u}\,[k] + \mathbf{g}^{(\sigma[k])} + \mathbf{v} \end{array} \right. \tag{3.3}$$

(remarque : les paramètres ne sont pas les mêmes en continu et en échantillonné, $A^{(\sigma[k])}$ (et respectivement $B^{(\sigma[k])}$, $C^{(\sigma[k])}$, $D^{(\sigma[k])}$, $f^{(\sigma[k])}$, $g^{(\sigma[k])}$) n'est pas $A^{(\sigma)}$ (et respectivement $B^{(\sigma)}$, $C^{(\sigma)}$, $D^{(\sigma)}$, $f^{(\sigma)}$, $g^{(\sigma)}$)).

Les méthodes existantes utilisent souvent ce formalisme général, mais il existe d'autres formalismes [81] (par exemple à base de regresseurs). Cela n'a pas une importance décisive.

Montrons maintenant que l'équation APM (1.4) des réseaux de régulation génique peut être ramené à ce formalisme :

$$\dot{x_i} = {\kappa_i}^{(j)} - {\gamma_i}^{(j)} x_i \text{ ssi } \mathbf{x} \in \Delta_j, j \in \{1, \ldots, s\}$$

où les Δ_j sont les domaines de régulation hyper-rectangles compris entre les hyperplans définis par les seuils de transition et les bords de l'espace d'état. Dans un domaine Δ_j, la résolution de cette équation différentielle donne :

$$x_i(t) = \left(x_i(0) - \frac{{\kappa_i}^{(j)}}{{\gamma_i}^{(j)}} \right) \exp(-{\gamma_i}^{(j)} t) + \frac{{\kappa_i}^{(j)}}{{\gamma_i}^{(j)}}.$$

On en déduit, en temps échantillonné, la formule de récurrence :

$$x_i[k+1] = x_i[k] \exp(-{\gamma_i}^{(j)}(t_{k+1} - t_k)) + \frac{{\kappa_i}^{(j)}}{{\gamma_i}^{(j)}} \left(1 - \exp(-{\gamma_i}^{(j)}(t_{k+1} - t_k)) \right).$$

On en arrive ainsi à la formulation matricielle :

$$\begin{array}{rcl} \mathbf{x}\,[k+1] & = & \left[\begin{array}{cccc} \exp[-{\gamma_1}^{(j)}(t_{k+1} - t_k)] & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \exp[-{\gamma_n}^{(j)}(t_{k+1} - t_k)] \end{array} \right] \mathbf{x}[k] \\ & & + \left[\begin{array}{c} \frac{{\kappa_1}^{(j)}}{{\gamma_1}^{(j)}} \left(1 - \exp[-{\gamma_1}^{(j)}(t_{k+1} - t_k)] \right) \\ \vdots \\ \vdots \\ \frac{{\kappa_n}^{(j)}}{{\gamma_n}^{(j)}} \left(1 - \exp[-{\gamma_n}^{(j)}(t_{k+1} - t_k)] \right) \end{array} \right]. \end{array} \tag{3.4}$$

On se ramène bien à un formalisme hybride où il n'y a pas d'entrée. $\Omega \subset \mathbb{R}^n$ est l'espace d'état.

Dans le cadre d'une observation avec erreur sur la sortie (par rapport à (3.3), $\mathbf{w}$ est un vecteur nul), on estime que la sortie est bruitée avec une erreur aléatoire (qui sera souvent prise comme étant gaussienne) :

$$\mathbf{y}[k] = \mathbf{x}[k] + \eta[k].$$

Avec certains appareils certifiés de mesure, on pourra aussi imposer une erreur bornée, voire un bruit de quantification à pas connu.

3.2 Identification de systèmes hybrides APM

Nous n'allons pas décrire en détail dans ce chapitre tout ce qui concerne le sujet de l'identification des systèmes hybrides : il existe de bonnes publications qui rendent le sujet accessible [95, 81] de façon plus détaillée. Nous souhaitons juste poser les problèmes de l'identification hybride dans le cadre le plus fréquent et donner en survolant les grands traits de la résolution : cela nous aidera à en dessiner les limites.

Problème 1 (Identification de systèmes hybrides APM)**.** *Étant donnée une collection finie de N paires d'entrée-sortie $\{(u[k], y[k])\}_{k=1}^{N}$, (a) estimer l'ordre n du modèle, (b) le nombre s de sous-modèles, (c) les paramètres $A^{(j)}, B^{(j)}, f^{(j)}, C^{(j)}, D^{(j)}, g^{(j)}$ pour j allant de 1 à s, ainsi que (d) la séquence $\{\sigma[k]\}_{k=1}^{N}$ et (e) les régions $\{\Omega_i\}_{i=1}^{s}$.*

Dans la plupart des méthodes de résolution proposées, certaines quantités sont supposées connues, ou des hypothèses fortes sur le bruit sont faites.

C'est le cas dans le problème d'identification de réseau de régulation génique. En effet, l'état et la mesure sont assimilées (à un bruit près), ce qui rend l'ordre du modèle fixé par l'expérience biologique. Il se peut que trop peu de gènes soient mesurés : cela ne fait pas parti de la problématique abordée ici.

Toutes les méthodes d'identification pour systèmes hybrides disponibles savent résoudre les sous-problèmes (c) et (d) (en supposant au pire n et s connus) du Problème 1. On peut considérer que le sous-problème (e) peut être résolu en déterminant des hyperplans séparateurs obtenus aux moyens des méthodes de reconnaissance de motifs telles que la programmation linéaire robuste multicatégorique (MRLP [14]) ou les classifieurs à vecteur support (SVC [120]). En ce qui concerne les sous-problèmes (b), et, encore plus délicat, le sous-problème (a), quelques méthodes proposent des approches : il s'agit cependant d'une estimation très délicate.

Pour ce qui concerne les réseaux de régulation génique et leur identification, le Problème 1 précédant peut se ramener à la reformulation suivante :

Problème 2 (Identification de réseaux de régulation génique APM)**.** *Étant donnée une collection finie de N mesures sur les produits de l'expression de n gènes $\{\mathbf{x}[k]\}_{k=1}^{N}$, $\mathbf{x} \in \Omega$ (a) estimer le nombre s de sous-modèles traversés, (b) reconstruire la séquence $\{\sigma[k]\}_{k=1}^{N}$, (c) estimer les domaines de régulation hyperrectangulaires $\{\Delta_j\}_{j=1}^{s}$, tels qu'ils ne se superposent pas dans Ω ($\Delta_j \subset \Omega$ et $\Delta_j \cap \Delta_{j'} = \emptyset$), (d) estimer les paramètres dynamiques $\{(\kappa^{(j)}, \gamma^{(j)})\}_{j=1}^{s}$ décrivant les taux de synthèse et de dégradation à l'intérieur de chacun des domaines de régulation.*

Dans notre cadre, qui est celui du Problème 2, l'étape (a) ne peut pas être envisagée comme prérequis. D'autre part, le sous-problème (a) ne doit pas être confondu avec l'estimation du nombre total de domaines de régulation envisageables pour le réseau de régulation génique inféré, de même que le sous-problème (c) ne conduit pas à la reconstruction d'un ensemble de domaines hyper-rectangles formant une partition de Ω (la condition $\sum_{j=1}^{s} \Delta_j = \Omega$ n'est pas exigée) : la trajectoire donnée par les points de mesure ne traverse pas nécessairement tous les domaines de régulation. Il n'est possible d'inférer d'un réseau de régulation génique seulement la partie sollicitée lors d'une expérience : il s'agit là de la question de l'identifiabilité pratique. Toutefois, on pourra envisager de fusionner les sous-problèmes (a) et (b) : la connaissance de la séquence σ donne immédiatement le nombre s.

Les méthodes proposées dans la littérature pour l'identification de systèmes hybrides sont donc applicables dans une certaine mesure pour faire l'inférence de réseau de régulation génique dans le cadre d'une modélisation affine par morceaux. Cependant, il existe plusieurs caractéristiques liées aux modèles APM de réseau de régulation génique qui ne sont pas directement prises en compte par les méthodes "généralistes". Deux d'entre elles nous paraissent particulièrement critiques.

La première d'entre elles consiste à prendre en compte la relative simplicité du modèle dynamique à l'intérieur d'un domaine de régulation. Il est possible de raffiner sérieusement l'inférence de cette dynamique en la contraignant au modèle d'équation différentielle d'ordre un et ainsi améliorer très nettement la reconstruction de la séquence des transitions d'un mode dynamique vers un autre.

La deuxième caractéristique est quant à elle d'ordre géométrique. Les seuils idéalisant la régulation interviennent sur une seule concentration et non pas sur une combinaison de concentrations : dès lors les domaines de régulation ne peuvent être qu'hyperrectangles avec leurs faces normales à un vecteur porteur d'axe de concentration. Or, ni MRLP ni SVC ne permettent d'imposer ce type de contraintes sur les hyperplans séparateurs de classe. De sorte que, même si la séquence de transitions est parfaitement connue, il n'y a aucune garantie que les méthodes existantes produisent des domaines Δ_j contraints comme dans le sous-problème (c) du Problème 2. Cela pourrait conduire à des modèles hybrides qui n'ont aucun sens d'un point de vue biologique, dans la mesure où ils ne conservent pas la notion de seuils franchis de régulation associés aux variables de concentration.

D'autre part, la formulation du Problème 2 ne doit pas faire oublier qu'il s'agit en premier lieu d'inférer la structure du réseau de régulation : les seuils de transition idéalisent une interaction génique qu'il s'agit de reconstruire. Or, cette information doit être extraite des résultats précédents : nous montrerons par la suite qu'il est nécessaire pour ce faire d'exploiter à la fois l'information sur les seuils reconstruits et sur les paramètres dynamiques identifiés.

Enfin, une dernière différence majeure d'avec les méthodes d'identification pour systèmes hybrides concerne l'unicité du résultat. En l'occurrence, toutes les méthodes proposent le meilleur modèle qui optimise un certain nombre de critères et de contraintes. Or nous verrons que pour une observation donnée, plusieurs modèles peuvent être identifiés de manière également pertinente : il n'est pas alors question de privilégier l'un d'entre eux de manière tout à fait arbitraire. Ainsi, on peut, par exemple, imaginer de rajouter des interactions dont l'influence n'est pas solliciter dans l'expérience et les données d'observation qui en découlent. On appliquera en premier lieu un principe de parcimonie permettant d'éliminer tous les modèles consistants avec les données mais dont la cardinalité de leurs interactions n'est pas minimale. Cela ne conduit pourtant pas à l'unicité de la solution. Seules de nouvelles expériences biologiques permettant de connaitre des modes de régulation non observés peuvent amener à discriminer avec plus d'acuité les solutions ayant un sens biologique.

Pour toutes ces raisons, nous proposons un algorithme d'identification de réseau de régulation génique à modèle affine par morceaux décrite dans les chapitres qui suivent.

Deuxième partie

Méthode

Chapitre

4 Description de la chaîne de traitement

Les récentes avancées portant sur les techniques expérimentales en biologie moléculaire ont conduit à la production d'une énorme quantité de données concernant la dynamique de réseaux de régulation génique. Nous présentons une approche pour l'identification des modèles affines par morceaux [30, 53] à partir des données expérimentales. Comme relevé dans les chapitres précédents, ces modèles se fondent sur l'hypothèse que la régulation se fait au niveau de la synthèse et de la dégradation des produits de l'expression des gènes : les paramètres cinétiques sont supposés constants jusqu'à ce que la concentration d'une protéine régulatrice franchisse un seuil de transition.

La méthode que nous présentons se concentre sur les problèmes de la détection des transitions entre les différents modes opératoires dans les données d'expression génique (ce problème a été introduit dans [89]) et de la reconstruction des seuils de transition associés avec les interactions régulatrices (ce problème a été introduit dans [37]). En particulier, notre méthode prend en compte les contraintes géométriques qui sont spécifiques aux modèles affines par morceaux de réseaux de régulation génique. Une telle méthode est prévue pour des systèmes avec erreurs sur la sortie pour lesquels les données d'observation sont des séries temporelles de mesures de concentrations cellulaires au niveau de populations de cellules ou de cellules individuelles.

Les points sont d'abord classifiés selon les modes à l'intérieur desquels le comportement dynamique est sensé être correctement décrit par un système d'équations différentielles du premier ordre. À partir de cette classification, une méthode de reconnaissance de motifs est utilisée pour reconstruire toutes les combinaisons de seuils de transition qui sont cohérentes avec les données mesurées. Pour chaque combinaison de seuils, il est alors possible de proposer un réseau de régulation identifié et les paramètres dynamiques pour chacun des modes opératoires.

Les performances de notre approche ont été analysées en utilisant des données synthétiques simulées à partir d'un modèle simplifié de la réponse à un stress nutritionnel chez la bactérie *Escherichia coli* (qui s'inspire de celui proposé dans [99]). En particulier, nous avons évalué l'influence du niveau de bruit et du pas d'échantillonnage sur les réseaux identifiés. Nos résultats montrent que la méthode, couplée avec des données de mesure temporelle suffisamment précises, lesquelles peuvent être obtenues à l'aide de méthodes utilisant des gènes rapporteurs, nous permet une identification quantitative des modèles affines par morceaux de réseaux de régulation génique.

4.1 Position du problème

Les Problèmes 3 et 4 suivants nous permettent de repositionner notre problématique. Le Problème 5 rappelle que la résolution de ces deux problèmes doit en fait être effectuée en même temps.

Problème 3. *Reconstruire un(des) réseau(x) orienté(s) d'interactions à partir de données d'expression.*

On représente souvent les réseaux d'interaction génique par des graphes où les sommets sont les gènes et les arcs leurs interactions. L'approche affine par morceaux est un

formalisme qui donnent davantage de détails : dans la formulation proposée originellement par Glass et Kauffmann [45], les actions des protéines sur l'expression des gènes sont exprimées via des combinaisons logiques. En effet, les actions de diverses protéines sur le taux d'expression d'un gène peuvent se combiner, de manière plus ou moins complexe, de manière plus ou moins intriquée, et cela se traduit en terme mathématique par l'utilisation d'une combinatoire (se reporter au chapitre 1.2.3). Le lien idéalisé entre les processus physico-chimiques modélisés et leur écriture mathématique voudrait que cette combinatoire traduise la combinaison de ces processus-mêmes. Cela serait peut-être envisageable si nous avions la possibilité d'observer tous les éléments concernés, ce qui n'est pas le cas pour ce travail. De sorte que, dans le cadre de ce manuscrit, il ne sera envisagé de modéliser que les interactions des éléments soumis à l'expérience biologique et mesurés. Assumons, par exemple, que nous disposons de la mesure de la concentration de la protéine résultant de l'expression du gène nommé ici i, et celle pour le gène nommé j. Ces deux concentrations moléculaires mesurées dans leur évolution sont appelées la trajectoire de notre expérience. Pour ce manuscrit, nous nous en tiendrons à déterminer si la concentration de la protéine associée au gène i agit sur l'expression du gène j étant donnée une observation d'une trajectoire, c'est-à-dire une mesure de l'évolution des concentrations des protéines produites par l'expression des gènes i et j.

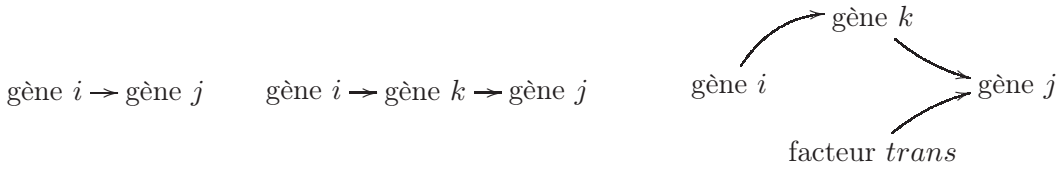


FIGURE 4.1 – Trois situations qui conduiront à la même reconstruction : dans tous les cas, les variations de l'expression des gènes i et j sont mesurées. Dans le premier cas, le résultat de l'expression du gène i agit directement d'un point de vue physico-chimique sur celle du gène j. Dans le deuxième cas, le gène k est un intermédiaire dans la chaine d'actions régulatrices mais n'est pas observé. Dans le troisième cas, le produit de l'expression du gène k se combine avec un autre facteur transcriptionnel présent, mais k n'est pas observé. Si au cours d'une expérience, il est possible de reconstruire une causalité entre l'expression de i et celle de j, nous ne pourrons pas déduire l'influence d'un gène intermédiaire k non observé, et encore moins décrire les conditions éventuelles de cette influence.

À noter que la notion de combinatoire prévue dans le modèle traduit le fait que la notion de réseau d'interaction génique fait l'abstraction des interactions qui se passent, entre autre, au niveau des protéines (se reporter à l'introduction en 1.1.2 et à l'exemple de la Figure 1.3). Cependant, si l'on observe uniquement les produits de l'expression du gène i et du gène j, comme sur la Figure 4.1, il nous semble impossible de prédire qu'en fait l'expression de i ait une influence causale sur l'expression d'un gène k non observée qui aura une influence causale sur l'expression du gène j ; d'autant plus que l'on peut imaginer encore que la protéine associée à i catalyse en fait une réaction sur d'autres molécules qui vont réagir pour produire in fine un facteur de transcription du gène j. Dans ce qui suit, il ne s'agira pas de déduire si la causalité de i sur j est une causalité biologique directe. Les gènes d'un réseau, dont les variations temporelles sont analysées, doivent être déterminés au préalable, en utilisant les informations relevées dans la littérature ou en utilisant les méthodes de motifs d'expression ou en se basant sur des analogies avec d'autres organismes.

Voici le présupposé qui est le nôtre :

1. nous notons que la dynamique de l'expression de j a été affectée à un moment donné.
2. nous supposons que ce changement est dû à l'action régulatrice de l'une des protéines observées.
3. nous assumons, conformément à notre modèle, que cette action régulatrice correspond au franchissement d'un seuil de concentration sur l'une des molécules observées.
4. chaque seuil correspond à un arc dans le réseau d'interactions : le franchissement du seuil intervient pour la protéine régulatrice, et le changement de dynamique intervient sur l'expression du gène régulé (que la causalité soit directe ou pas).
5. parmi tous les seuils trouvés pour la trajectoire mesurée, nous cherchons les seuils qui expliquent le plus de transitions dynamiques possibles, ce qui permet de réduire le nombre d'hypothèses.
6. pour les seuils retenus, une sous-partie d'entre eux suffit souvent à expliquer tous les changements de dynamique. En appliquant le principe de parcimonie, l'ensemble des réseaux reconstruits sont ceux pour lesquels il y a le moins d'arcs possibles : c'est-à-dire, étant donnée la trajectoire observée, le moins de régulations invoquées le mieux.

Il s'agit alors de relever les variations des dynamiques de l'expression des gènes observés, et cela nous conduit au problème suivant.

Problème 4. *Identifier les paramètres de la dynamique du système supposé APM.*

Si nous faisons l'hypothèse que nous connaissons les régulations (donc les seuils), nous pouvons immédiatement déduire les instants sur la trajectoire où la dynamique va changer (éventuellement, car du fait de la combinatoire, il se peut que l'action régulatrice soit "masquée") : en tout cas, la dynamique ne peut pas changer ailleurs qu'à ces instants de franchissement des seuils. Nous pouvons donc segmenter la trajectoire observée, et se concentrer sur chacun des segments (ou portions, ou intervalles de temps) pour inférer les paramètres de la dynamique du premier ordre à coefficient constant (dans le cadre de l'approximation liée à notre modèle). Ce dernier problème est classique (voir [68]).

Or, dans le pire des cas (qui est souvent celui dans lequel on se trouve), nous ne connaissons pas les seuils. À noter que, ici, toute information a priori sur les seuils est bonne à prendre. On peut intégrer de l'information a priori (ce qui permet d'être plus précis), pour peu que celle-ci soit avérée, ou au moins, qu'elle ne rentre pas en contradiction avec l'observation (les trajectoires mesurées). Pour la suite, nous n'intégrons pas d'information a priori, et nous ne traitons pas le problème compagnon qui consiste à vérifier que l'a priori n'est pas falsifié. Puisque nous ne connaissons pas les seuils, il va falloir décrire un moyen de détecter quand la trajectoire change de mode dynamique.

De plus, le fait de connaitre les seuils, et d'inférer les dynamiques, ne donne pas forcément un résultat cohérent au sens du modèle affine-par-morceaux qui sous-tend notre approche. En effet, les paramètres de chaque segment seront, du fait du bruit (causé, par hypothèse, par erreur sur la mesure), différents les uns des autres. Certains de ces paramètres seront "proches", mais cette notion topologique n'est pas triviale à manipuler. Or, pour un gène donné, la dynamique de son expression peut prendre un nombre de valeurs assez limité : pour justifier cela de manière sommaire, on considère en général que les réseaux de régulation génique sont en effet assez peu interconnectés et que les influences régulatrices sont de plus combinées ; en admettant que cet énoncé "général" est au moins parfois vérifié (comme dans bon nombre de réseaux décrits actuellement), il n'y a donc pas de

raison d'avoir autant de valeurs pour les dynamiques que de segments. Il s'agirait donc de s'appuyer sur les interactions pour que les variations des dynamiques inférées conduisent au moins d'hypothèses possibles. Ainsi, l'information sur les seuils est primordiale pour faire une inférence correcte, qui respecte au minimum la division de l'espace des concentrations (ou l'espace d'état) en région hyper-rectangles. Cela nous ramène au Problème 3.

Problème 5. *Nous avons vu que les Problèmes 3 et 4 sont intriqués, il faut donc les résoudre en parallèle.*

Pour ce manuscrit, nous nous en tiendrons là. Le fait de remonter ensuite à partir des seuils reconstruits et à partir des paramètres dynamiques inférés à l'équation dynamique d'ordre 1 affine-par-morceaux en reconstituant les combinaisons logiques des interactions est à ce jour un problème encore ouvert. Ce problème soulève de nouveaux défis. À l'échelle d'une trajectoire (pour une condition initiale donnée), certaines interactions peuvent être non identifiables : en l'occurrence, ce qui est observé ne les rend pas détectables. Il existe aussi des configurations de réseaux où les interactions sont structurellement non identifiables, c'est-à-dire que quelque soit la condition initiale, une seule trajectoire ne permet pas de les observer : il faudrait alors choisir plusieurs conditions initiales distinctes (l'évolution du système permettrait éventuellement de trancher) ; or cela ne correspond pas à la réalité d'un organisme vivant pour lequel les conditions initiales sont uniquement les quelques états stables viables pour la vie, et faire passer d'un état à un autre n'est même pas une évidence pour le biologiste[1].

La méthode que nous présentons suit l'enchainement proposé Figure 4.2. Enfin d'illustrer cette approche, nous allons donner un exemple simple introductif.

Un exemple introductif

Un réseau de régulation génique élémentaire est un réseau à deux gènes. S'il s'agira pour nous, dans ce chapitre, d'en décrire un à titre exemplaire : il est intéressant de constater que ce type de réseau est déjà utilisé en biologie, par exemple pour expliquer des phénomènes caractéristiques de la différenciation cellulaire [113, 90].

Nous reprenons, en donnant maintenant tout le détail, l'exemple du Chapitre 1 avec la Figure 1.4 que nous rappelons sur la Figure 4.3 : les gènes a et b ont pour unique produit de leur expression les protéines A et B respectivement. Ce sont ces protéines qui contrôlent l'expression de ces mêmes gènes : dans notre exemple, le contrôle se fera uniquement au niveau de la production. Plus précisément, lorsque la concentration de la protéine A dépasse un certain seuil, nommé $\theta_{(A),2}$, l'expression du gène b est inhibé : le seuil indique que le taux de présence de la protéine A est en effet suffisant pour que l'on considère qu'il réprime la production de B qui disparaît petit à petit par dégradation. De même, lorsque la concentration de la protéine B dépasse un autre seuil, nommé $\theta_{(A),1}$, l'expression du gène a est activée. De sorte que le gène a s'auto-active, et si l'on ignore l'influence de B, l'expression du gène a est bistable : si A est suffisamment présente dans le milieu, a sera exprimé conduisant à davantage de A (jusqu'à ce que cela s'équilibre avec la dégradation) ; inversement, si A est peu présente, il n'y aura pas d'expression de a et A viendra à disparaître. Mais il ne faut pas négliger l'influence de B qui a la possibilité d'activer a à partir d'un seuil de concentration fixé à $\theta_{(B),2}$, et de s'auto-activer, lui-aussi, à partir de $\theta_{(B),1}$.

1. Il existe toutefois des moyens pour faire varier les conditions initiales, en changeant les conditions du milieu de croissance, voire en perturbant de façon contrôlée l'expression génique.

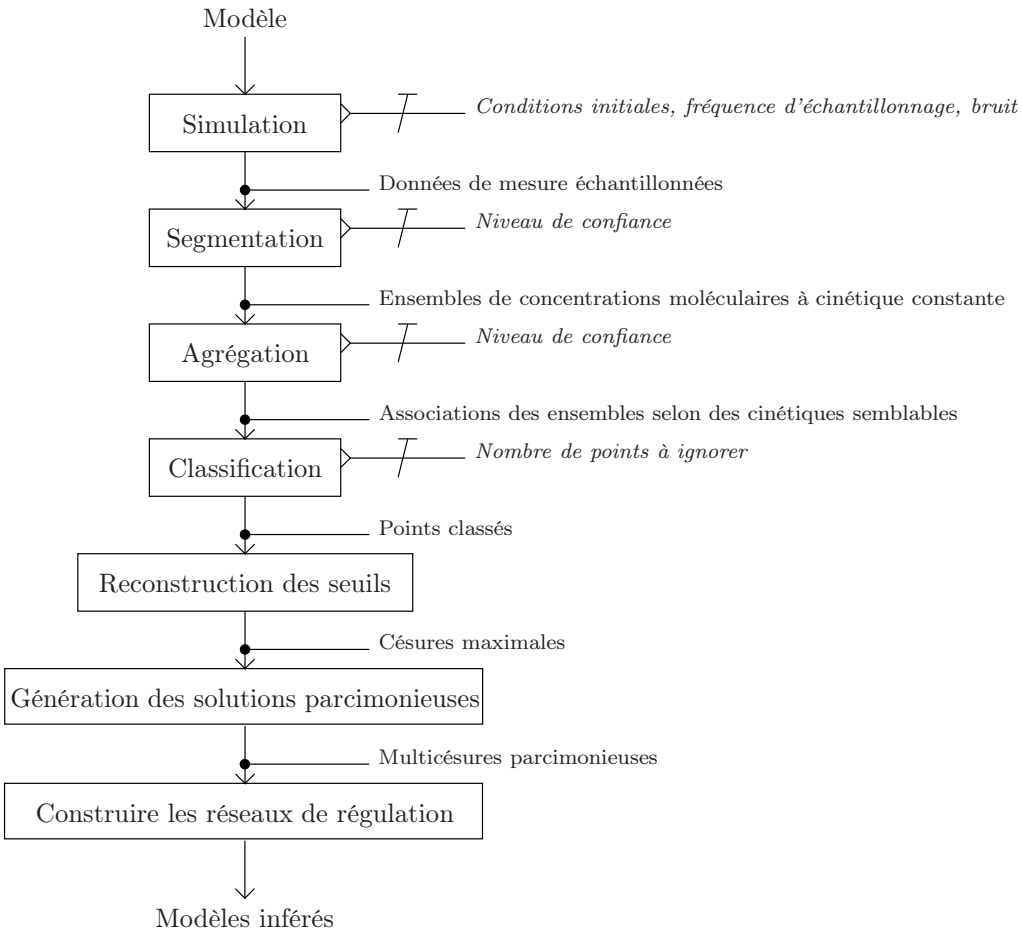


FIGURE 4.2 – Chaine de traitement : les principales étapes sont encadrées. Les paramètres de ces étapes sont donnés en italique.

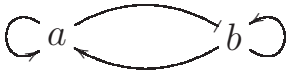


FIGURE 4.3 – Exemple de réseau de régulation à deux gènes. Les gènes a et b codent respectivement pour les protéines A et B qui sont leur seul et unique produit d'expression. La concentration de A est assimilée à la variable x_A et celle de B à x_B. Les interactions sont représentées avec des arcs orientés : s'il s'agit d'une activation, l'arc se termine par une flèche, s'il s'agit d'une inhibition, l'arc se termine par une barre.

Nous assimilerons la concentration de A à la variable x_A et celle de B à x_B. En suivant l'Exemple 3 à la page 25, l'écriture de l'évolution cinétique de la concentration de A dans le modèle APM devient (4.1). La dérivée du premier ordre $\dot{x}_A$ de la variable x_A est donc le taux de variation instantanée de la concentration de A : celui-ci est la somme d'un terme de production où se traduit les interactions géniques, et d'un terme de dégradation proportionnel à la concentration elle-même. Il s'agit donc au final d'une équation différentielle du premier ordre. Les fonctions s^+ et s^- (définies page 27) prennent des valeurs dans $\{0,1\}$, d'où le fait que leur produit prend de même des valeurs dans $\{0,1\}$, et elles passent de l'une à l'autre de leurs valeurs lorsque la variable indiquée franchit le seuil indiqué. Cela fait que le terme de production est constant par morceaux. De même, x_B varie d'après l'équation différentielle APM du premier ordre comme décrit en (4.2). La résolution de l'équation (4.1) nécessite la résolution de (4.2) et inversement : il s'agit d'un système d'équations dynamiques couplées.

$$\dot{x}_A = \kappa_A\, s^+(x_A, \theta_{(A),1}) s^+(x_B, \theta_{(B),2}) - \gamma_A\, x_A \tag{4.1}$$
$$\dot{x}_B = \kappa_B\, s^-(x_A, \theta_{(A),2}) s^+(x_B, \theta_{(B),1}) - \gamma_B\, x_B \tag{4.2}$$

Les paramètres en κ et en γ sont des réels positifs car une concentration ne peut pas prendre de valeur négative. Dans notre cas, les interactions se combinent. Par exemple, expliquons (4.1) : il suffit que A n'active pas a (i.e. $x_A < \theta_{(A),1}$) ou B n'active pas a (i.e. $x_B < \theta_{(B),2}$) pour que le taux de production de A soit nul. Par contraposée, en appliquant la loi de De Morgan, le taux de production de A sera κ_A si $x_A > \theta_{(A),1}$ et si $x_B > \theta_{(B),2}$. Il est possible d'en conclure que, dans notre écriture, $s^+(x,\theta) = 1$ si $x > \theta$, 0 sinon. Nous laissons au lecteur le soin de faire le même raisonnement pour (4.2).

Exemple 6. *Nous utiliserons l'exemple du réseau décrit pour la Figure 4.3 avec les valeurs des paramètres suivantes.* $\theta_{(A),1} = \theta_{(B),1} = 0,25$. $\theta_{(A),2} = \theta_{(B),2} = 0,63$. $\kappa_A = \gamma_A = 1$. $\kappa_B = 1,4$, $\gamma_B = 1,5$. *Les variables* x_A *et* x_B *varient dans l'intervalle* $[0,1]$.

Un exemple de trajectoire échantillonnée obtenue par la résolution du système d'équations est donné Figure 4.4 pour des conditions initiales proches de 1 *(i.e. les concentrations des protéines* A *et* B *sont proches de leur maximum) : nous expliquerons en 4.2 comment nous avons obtenu cette trajectoire. Il est à noter que les seuils* $\theta_{(A),2}$ *et* $\theta_{(B),2}$ *qui apparaissent en magenta sont franchis : ils engendrent une rupture dans la dynamique qui est soulignée par une barre verticale rouge. Nous verrons en 4.3 que la première étape de notre méthode consiste à retrouver ces barres rouges. La trajectoire tend à se rapprocher de ces seuils : indiquons donc ici, sans préciser davantage, qu'une trajectoire ne converge pas nécessairement vers un point obtenu par la résolution directe de l'équation* $\dot{x}_A = \dot{x}_B = 0$ *à l'intérieur des différents domaines de définition, mais un équilibre stable sera atteint (dans la résolution du système d'équations couplées, les taux de variation varient à des instants infiniment rapprochés, ce qui est un phénomène dit de Zénon [62, 60]).*

Il est encore à noter que les seuils $\theta_{(A),1}$ *et* $\theta_{(B),1}$ *qui apparaissent en jaune sur la figure ne sont jamais franchis au cours de cette trajectoire. Puisqu'ils ne sont jamais franchis, il ne sera pas possible de déduire ce que leur éventuel franchissement pourrait induire : il ne sera donc pas possible d'inférer les interactions associées, en l'occurrence, celles d'auto-activation du gène a et du gène b. Cela nous permet d'introduire d'un point de vue intuitif ce que nous appellerons une interaction* pratiquement non-identifiable. *Le lecteur pourrait être amené à penser que cette situation est contingente des données mesurées, c'est-à-dire ici des conditions initiales de notre trajectoire. Cela est vrai en général, et c'est pour cela*

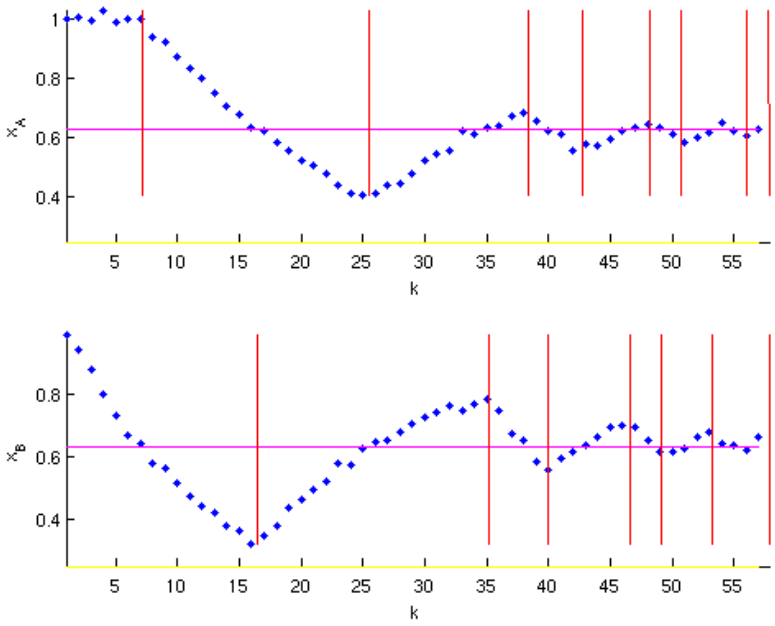


FIGURE 4.4 – Exemple de trajectoire obtenue pour le réseau de régulation à deux gènes décrits sur la Figure 4.3 : un bruit gaussien a été ajouté pour représenté l'erreur sur la mesure.

*que l'on parle d'*identifiabilité pratique. *Il se trouve que dans le cas de l'exemple choisi (et il a d'ailleurs été choisi pour cela), quelque soit la condition initiale, à partir d'une trajectoire, il ne sera jamais possible de reconstruire ces interactions : nous introduirons, toujours de manière intuitive, l'*identifiabilité structurelle. *Pour justifier ce fait, il est nécessaire de considérer le champ des vecteurs* $(\dot{x}_A, \dot{x}_B)$ *représentés Figure 4.5 dans l'espace des concentrations* (x_A, x_B) *(dit aussi* espace d'état *dans la terminologie de l'Automatique). Sur cette figure, nous constatons que les seuils divisent l'espace* (x_A, x_B) *en régions rectangles nommées* domaines de régulation, *et nous savons qu'à l'intérieur de celles-ci les paramètres de la dynamique sont constants. Cependant, il y a neufs rectangles, mais, étant données* (4.1) *et* (4.2), *il n'y a que quatre possibilités pour les paramètres cinétiques : celle où les deux taux de production sont nuls, celle où les deux sont non-nuls, les deux dernières où l'un est nul l'autre pas. Plus précisément, il n'existe que deux domaines de régulation qui ont leur ensemble de paramètres qu'ils sont les seuls à posséder : il s'agit des deux rectangles supérieurs à droite. En nommant les domaines* $\Delta_1, \ldots, \Delta_9$ *par ordre séquentiel croissant suivant le sens de la lecture, les deux rectangles supérieurs à droite sont donc* Δ_2 *et* Δ_3. *On note ensuite que* $\Delta_1, \Delta_4, \Delta_5$ *ont des mêmes paramètres, et* $\Delta_6, \ldots, \Delta_9$ *aussi. Dès lors, examinons l'identifiabilité de* $\theta_{(B),1}$. *Il est possible d'obtenir une trajectoire qui passerait de* Δ_6 *à* Δ_9, *franchissant ainsi* $\theta_{(B),1}$. *Or, dans ce cas,* $x_A > \theta_{(A),2}$ *et A inhibe donc la propension de b à s'auto-activer : il n'y a pas de changement de dynamique observable. À l'interface de* Δ_6 *et de* Δ_9, $\theta_{(B),1}$ *n'est donc pas pratiquement identifiable. Cependant, les champs de vecteurs* $(\dot{x}_A, \dot{x}_B)$ *ont des directions opposées si l'on regarde de part et d'autre de* $\theta_{(B),1}$, *entre* (Δ_4, Δ_5) *et* (Δ_6, Δ_7). *Il n'y aura donc jamais de trajectoire qui pourrait franchir* $\theta_{(B),1}$ *lorsqu'un changement de dynamique est observable. En conclusion,* $\theta_{(B),1}$ *est structurellement non-identifiable. Nous laissons le lecteur faire la même démonstration*

pour $\theta_{(A),1}$.

Le lecteur retiendra alors que seules les interactions associées aux seuils $\theta_{(A),2}$ *et* $\theta_{(B),2}$ *seront identifiables, comme représentées Figure 4.6.*

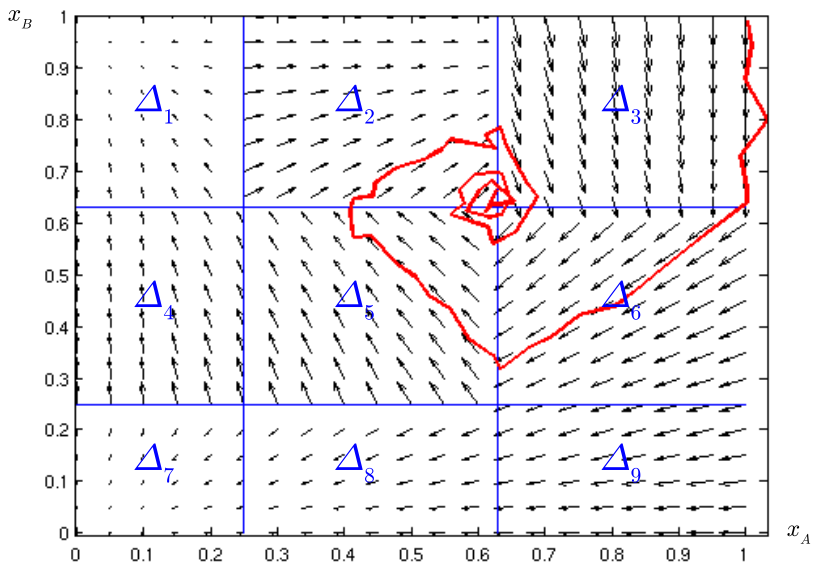


FIGURE 4.5 – Représentation de la trajectoire dans l'espace de phase de la trajectoire de la Figure 4.4 (rouge), avec le champ vectoriel $(\dot{x}_A, \dot{x}_B)$. Les lignes bleues sont les seuils du système. Ils divisent l'espace en domaines de régulation à l'intérieur desquels les paramètres cinétiques de la trajectoire sont constants.

FIGURE 4.6 – Exemple de réseau de régulation à deux gènes de la Figure 4.3. Pour les paramètres choisis dans l'Exemple 6, seules les interactions représentées ici seront pratiquement identifiables d'après la trajectoire "mesurée" présentée sur la Figure 4.4.

4.2 Données synthétiques

Le problème que nous nous proposons de résoudre est un problème inverse. Quand on pose une équation différentielle, le problème direct est d'écrire la trajectoire qui respecte cette équation étant données des conditions initiales ; dans notre cas, on part de l'observation d'une trajectoire et l'on veut retrouver les paramètres de cette équation. Le problème est, comme c'est usuellement le cas, un problème mal-posé, au sens de Hadamard [51]. Il n'est donc pas inutile de décrire le problème direct avant de décrire le problème inverse.

Dans ce paragraphe, nous décrivons le simulateur de trajectoires que nous avons utilisé pour élaborer et évaluer notre procédure d'identification. Il est donc important d'en décrire

le fonctionnement, même si la procédure d'identification doit pouvoir fonctionner indépendamment du simulateur. Ceci est d'ailleurs vérifié dans la mesure où le simulateur décrit ici n'est que la dernière version d'une série de simulateurs qui étaient d'abord très spécialisés, ne savaient générer des données que pour un type de modèle, pour des conditions initiales précisées, sans prendre en compte des interventions externes, et ainsi de suite.

Cette version est donc l'aboutissement d'une réflexion qui tendait à satisfaire plusieurs objectifs :

- être indépendant du modèle de réseaux (nombre de gènes, équations de leur dynamique) : à l'heure actuelle, un fichier externe est chargé, mais une piste abordée et non aboutie était de charger des modèles écrits en SBML ou même XML généré par des outils dédiés tels que Iogma ou plus précisément le Genetic Network Analyzer [33].
- générer des trajectoires affines par morceaux de façon rapide (quelques secondes) en prenant en compte la résolution très simple des équations différentielles du premier ordre avec conditions initiales (critère de Cauchy) et en se concentrant sur la détection des changements de mode dynamique.
- avoir comme paramètre les conditions usuelles, telles que les conditions initiales, la fréquence d'échantillonnage, le bruit pour le cas "output-error".
- stopper automatiquement la simulation, principalement à cause du fait qu'un état stationnaire est atteint. Deux autres situations conduisent encore à un arrêt de la simulation. Le premier est un comportement dit de Zénon, lorsque des évènements discrets (i.e. lorsqu'une modification de la régulation intervient) se produisent à des instants très rapprochés dont l'écart tend à être nul. Le second est un comportement glissant dit de "mur noir", pour lequel les dynamiques de chaque côté d'un seuil oriente la trajectoire chacune de l'autre côté du seuil, si bien que la trajectoire reste "attachée" au seuil et tend à glisser sur lui ([7, 21]). L'évolution de la trajectoire est alors assez difficile à simuler et ne l'est pas dans notre cas.
- prendre en compte aussi des changements externes, qui font que le milieu change, ou plutôt est affecté par une intervention externe (du biologiste expérimentateur par exemple) ou du fait d'un phénomène physique : ces changements interviennent à un temps discret qui peut être soit connu, soit considéré comme étant suffisamment grand pour que le système se soit stabilisé. La conséquence sur la concentration des molécules est soit une translation (par exemple si l'expérimentateur ajoute des éléments) soit un forçage à un niveau donné (cas usuel du signal d'entrée).
- il a aussi semblé utile de générer pendant la simulation le réseau identifiable (défini au Chapitre 8.1), qui dépend du modèle du réseau et aussi de la trajectoire observée.

La construction d'un simulateur tend donc à préciser l'ensemble des caractéristiques des signaux que nous nous proposons de traiter. S'il est une réduction, et en ce ce sens s'il ne reflète que partiellement ce que seront les données de mesure, il permet de traduire algorithmiquement ce que nous nous attendons à observer dans la réalité pour un large spectre d'expériences, notamment celles à base de gènes rapporteurs.

Enfin, le choix des paramètres de simulation et du réseau exemplaire peuvent être expliqués de cette manière :

- l'équipe de Jong du projet Helix à l'INRIA Rhône-Alpes travaille en étroite collaboration avec l'équipe Geiselmann du LAPM au CNRS UMR 5163 (université Joseph Fourier) : l'analyse de la régulation globale de la transcription chez *Escherichia coli*, en particulier la réponse à un stress nutritionnel, a été privilégié (Action ACI IMP-BIO BacAttract, Ministère de la Recherche & ARC INRIA GDyn)
- en exploitant notamment les informations de la littérature, Delphine Ropers (INRIA)

a permis de préciser le réseau de régulation génique associé à la réponse à un manque de nutriment dans le milieu de la bactérie ([99])
- ce réseau a été étudié et précisé au moyen d'outils pour l'analyse qualitative ([10, 98])
- sur la base des études qualitatives effectuées, un sous réseau à quatre gènes a été choisi, préservant le comportement global du système, tout en ayant une faible dimension. Par ailleurs ce sous-réseau avait la propriété de ne pas nous exposer à des comportements dits glissants [50], cas que nous ne voulions pas avoir à traiter dans un premier temps.

4.2.1 Description du modèle de réseaux utilisé par le simulateur de trajectoires

Modèle

On commence tout d'abord par charger le modèle affine par morceaux. Un modèle est composé d'une liste d'éléments (des protéines dans notre cas) $m^* = \{m_1, \ldots, m_n\}$ dont on considère que les concentrations sont observables, représentées au moyen du vecteur $\mathbf{x} = (x_1, \ldots, x_n)^T$. Le principe est que le modèle dynamique biologique non-linéaire de ces éléments est approximé par une série de discontinuités qui surviennent à des seuils sur les concentrations. L'écriture des équations dynamiques n'est pas simple. Pour clarifier la lecture, on souligne que la dynamique est un modèle cinétique d'ordre un dans des régions données de l'espace des concentrations , et on écrit donc :

$$\dot{\mathbf{x}} = \kappa(\mathbf{x}) - \gamma(\mathbf{x}) \tag{4.3}$$

Rappelons rapidement ce qui a été vu en 1.2.3. Dans le cadre des modèles affines par morceaux, à part au niveau des discontinuités, le taux de synthèse $\kappa_i(\mathbf{x})$ et la constante de dégradation $\frac{\gamma_i(\mathbf{x})}{x_i}$ de chaque élément m_i sont constants. L'action des discontinuités peut être combinée (par exemple, biologiquement, deux protéines sont nécessaires pour former un complexe qui aura une action donnée sur l'expression d'un gène à partir d'un certain taux de présence), si bien que $\kappa_i(\mathbf{x}) = \sum_{l \in L_i} \kappa_{il} \breve{s}_{il}(\mathbf{x})$ (où L_i est un ensemble d'entiers éventuellement vide, κ_{il} sont des constantes, et $\breve{s}_{il}(\mathbf{x})$ sont des fonctions booléennes logiques combinant des tests de type $x_i > \theta$ qu'on écrira à l'aide de fonctions en escalier $s(x_i, \theta)$) et de même $\gamma_i(\mathbf{x}) = \sum_{l \in L'_i} \gamma_{il} \breve{s}'_{il}(\mathbf{x}) x_i$.

Le principe adopté pour la simulation est de recalculer la valeur des constantes de synthèse et de dégradation à chaque fois qu'un seuil est franchi. Dans le fichier où est décrit le modèle, chacune de ces constantes (κ_i, γ_i) est décrite par une équation qui compose des fonctions en escalier centrées sur des seuils et de gain donné. Sont donc listés :
- les seuils pour chaque molécule qui en dispose : $\left\{\theta_{(i),l}\right\}_{i \in \{1,\ldots,n\}, l \in T_i}$,
- les gains selon ce qui est nécessaire : $\{\kappa_{il}\}_{i \in \{1,\ldots,n\}, l \in L_i}$ et $\{\gamma_{il}\}_{i \in \{1,\ldots,n\}, l \in L'_i}$,
- les équations écrites de manière normalisée : celles-ci peuvent être évaluées pour n'importe quel point de la trajectoire, afin de calculer les taux de synthèse et de dégradation.

Il est à noter que pour lever certaines ambiguïtés (les fonctions en escalier peuvent renvoyer un intervalle de valeurs au niveau de leur seuil), l'évolution est prise en compte, de telle sorte que 0 est renvoyé au niveau du seuil si l'on va de 0 à 1, et inversement 1 est renvoyé si l'on va de 1 à 0.

L'espace des concentrations Ω est partitionné en s hyperrectangles $\Delta^* = \{\Delta_1, \ldots, \Delta_s\}$

tels que, pour $j \in \{1, \ldots, s\}$, avec les listes $(\delta_{(i),l})_{l \in \{1,\ldots,|T_i|+2\}} = (0, \theta_{(i),1}, \ldots, \theta_{(i),|T_i|}, \max x_i)$,

$$\Delta_j = \left\{ \mathbf{x} : \forall i \in \{1, \ldots, n\}, \exists! l \in \{1, \ldots, |T_i| + 1\}, \left| \begin{array}{ll} \delta_{(i),l} \leq x_i < \delta_{(i),l+1} & \text{si } l = 1, \\ \delta_{(i),l} < x_i \leq \delta_{(i),l+1} & \text{si } l = |T_i| + 1, \\ \delta_{(i),l} < x_i < \delta_{(i),l+1} & \text{sinon.} \end{array} \right. \right\}.$$

Ces régions sont les *domaines de régulation*. On peut alors réécrire la constante de synthèse $\kappa_i(\mathbf{x})$ et celle de dégradation $\frac{\gamma_i(\mathbf{x})}{x_i}$ de chaque élément m_i sur chacun de ces domaines de régulation : pour $\mathbf{x} \in \Delta_j$, $\kappa_i(\mathbf{x}) = \kappa_i^{(j)}$ et $\gamma_i(\mathbf{x}) = \gamma_i^{(j)} x_i$. En prenant l'écriture matricielle $\kappa^{(j)} = \left[\kappa_1^{(j)} \cdots \kappa_n^{(j)}\right]^T$ et $\gamma^{(j)} = \mathrm{diag}(\gamma_1^{(j)}, \ldots, \gamma_n^{(j)})$, l'équation (4.3) est plus généralement réécrite au moyen d'équations affines :

$$\dot{\mathbf{x}} = \left\{ \begin{array}{ll} \kappa^{(1)} - \gamma^{(1)}\mathbf{x} & \text{si } \mathbf{x} \in \Delta_1 \\ \vdots & \\ \kappa^{(s)} - \gamma^{(s)}\mathbf{x} & \text{si } \mathbf{x} \in \Delta_s \end{array} \right. \tag{4.4}$$

Paramètres

On charge ensuite les paramètres décrivant l'expérience :
- les mesures sont supposées être faites à temps discrets avec un pas d'échantillonnage T constant ;
- les conditions initiales : dans le cas biologique usuel, elles sont choisies dans le domaine de régulation qui correspond à un état stable de la cellule (cela sert en fait à synchroniser une population de cellules) ;
- les évènements externes décrivant des altérations de l'évolution des concentrations simulées, soit au bout d'un temps donné, soit lorsque la simulation est interrompue (elle peut l'être pour les raisons suivantes : un état stationnaire est atteint, ou l'évolution d'une molécule a un comportement dit de Zénon, ou un mode de glissement [49] est atteint pour une configuration en "mur noir" que nous expliquerons ultérieurement) ;
- l'écart type σ du bruit gaussien qui sera ajouté à la trajectoire simulée : il s'agit d'une simulation pour une trajectoire avec erreur sur l'observation.

4.2.2 Simulation de la trajectoire échantillonnée

La résolution de l'équation (4.4) dans un domaine de régulation donné (où les taux de synthèse et de dégradation sont constants) conduit à la forme continue suivante : pour tout $i \in \{1, \ldots, n\}$

$$\left\{ \begin{array}{lcll} x_i(t) & = & (x_i(0) - \frac{\kappa_i^{(j)}}{\gamma_i^{(j)}}) \exp(-\gamma_i^{(j)} t) + \frac{\kappa_i^{(j)}}{\gamma_i^{(j)}} & \text{si } \mathbf{x} \in \Delta_j \text{ en tout instant } t \in [0, t_s] \\ y_i(t) & = & x_i(t) + \eta_i(t) & \text{avec } \eta_i \text{ bruit gaussien} \end{array} \right. \tag{4.5}$$

Le principe de la simulation est donc de calculer les instants où l'on bascule dans un nouveau domaine de régulation (c'est-à-dire, par exemple si $\mathbf{x}(0) \in \Delta_j$, l'instant t_s tel que $\forall \epsilon > 0$, $\mathbf{x}(t_s + \epsilon) \notin \Delta_j$). Cela permet de calculer d'abord les points échantillonnés non bruités $x_i(kT)$, $k \in \mathbb{N}$, en fonction de l'équation (4.5), puis on ajoute le bruit $\eta_i(kT)$ afin d'obtenir les données d'observation $y_i(kT)$ à traiter.

L'essentiel est donc de décomposer le temps en intervalles pendant lesquels la trajectoire appartient à un même domaine de régulation. A chaque fois que l'on rentre dans un nouvel intervalle, on recalcule alors les paramètres de l'équation (4.5) qui sont le point d'entrée dans le domaine, le taux de synthèse et celui de dégradation. Il n'est d'ailleurs pas besoin à ce stade de générer explicitement les points $\mathbf{x}(kT)$ ni même $\mathbf{y}(kT)$. La trajectoire est estimée dans sa totalité en travaillant seulement sur les temps pour lesquels au moins une transition intervient. L'intervalle qui se situe entre deux transitions est appelé *intervalle sans transition*. Pour chacun de ces intervalles, les valeurs des taux de synthèse et de dégradation, ainsi que la valeur du point d'entrée dans le nouveau mode sont simplement mis en mémoire.

Précisons le calcul du temps final d'un intervalle sans transition. Le temps d'entrée dans l'intervalle est appelé ici le temps présent, qui n'est autre que le temps de l'expérience réinitialisé à zéro pour se placer dans le formalisme décrit dans l'équation (4.5). La trajectoire tend vers un point, nommé point focal ϕ. Pour chacune des dimensions $i \in \{1, \ldots, n\}$, la trajectoire tend vers la valeur ϕ_i, valeur qui se trouve soit à l'intérieur du domaine de régulation, soit à l'extérieur : dans ce dernier cas, une transition vers un autre domaine de régulation va survenir. Il s'agit de calculer l'ensemble des instants où une transition survient pour un élément donné, et de se concentrer sur le plus petit d'entre eux. Si cet ensemble d'instants est vide, c'est donc que la trajectoire tend vers ϕ qui est à l'intérieur du domaine de régulation présent : la trajectoire tend vers un point de stationnarité qui est stable à moins qu'un évènement extérieur n'intervienne.

Pour déterminer les transitions, calculons la composante du point focal de chaque élément m_i à l'instant présent : celui-ci est le rapport du taux de synthèse présent par le taux de dégradation présent $\phi_i = {\gamma_i}^{-1}\kappa_i$. Un élément m_i engendre une transition si un seuil de l'ensemble $\{\theta_{il}\}_{l \in T_i}$ est compris entre la position x_i^e d'entrée dans l'intervalle et la position focale ϕ_i. Il est possible à ce moment de vérifier que nous ne sommes pas dans le cas de figure d'un glissement dû à un *mur noir* (pour une introduction formelle, se reporter à [49]) : cela est le cas lorsque l'évolution de la concentration avant la dernière transition rentre en contradiction avec la position du nouveau point focal qui tend à faire "rebondir" la concentration sur la transition franchie. Pour chaque élément m_i pour lequel au moins un seuil est compris entre la position d'entrée dans l'intervalle présent et le point focal présent, nous ne désignons ici par simplification que le seuil parmi ceux-là qui est le plus proche de la position d'entrée : il est noté θ^i. Le temps de la prochaine transition est le plus petit des temps de chacune des molécules où une transition est possible. L'expression analytique du temps du transition peut être écrit comme étant :

$$t_s = \min_{\{\theta^i\}_i \text{ t.q. } \theta^i \text{ défini}} \frac{1}{\gamma_i} \ln\left(\frac{\kappa_i - \gamma_i x_i^e}{\kappa_i - \gamma_i \theta^i}\right). \tag{4.6}$$

Un dernier élément à considérer avant d'utiliser ce temps pour limiter l'intervalle sans transition présent est de vérifier si aucun évènement extérieur n'a été programmé auparavant. Dans tous les cas, le nouveau point d'entrée est calculé comme étant le point atteint au temps de fin de l'intervalle sans transition. Si aucun seuil n'intervient entre une position d'entrée et le point focal (i.e $\{\theta^i\}_i$ est vide), le système atteint un régime stationnaire et l'intervalle sans transition est borné lorsque la trajectoire a atteint 95% de son état final par rapport au dernier point d'entrée. Si trois intervaux successifs durent moins que la durée d'un pas d'échantillonnage, on fera l'hypothèse que le système a un comportement de type Zénon. Au final, un état stationnaire, un comportement de Zénon ou une trajectoire glissante sur un mur noir mettent fin à la simulation de la trajectoire, à moins qu'un

évènement externe n'intervienne à ce moment pour la relancer.

La génération des données échantillonnées est alors triviale : pour chaque intervalle sans transition caractérisé par son point d'entrée $\mathbf{x}^e$, les taux de synthèse κ et de dégradation γ, le temps t_s passé dans l'intervalle avant changement de domaine de régulation, ainsi que le temps t_e écoulé depuis l'instant d'entrée jusqu'au prochain plus petit multiple entier du temps d'échantillonnage T, en appelant K ce temps discret du premier point de l'intervalle, pour chaque élément i, on a pour tout k entier naturel tel que $t_e + kT \leq t_s$,

$$x_i(K+k) = (x_i^e - \frac{\kappa_i}{\gamma_i}) \exp(-\gamma_i(t_e + kT)) + \frac{\kappa_i}{\gamma_i}. \tag{4.7}$$

Sans formaliser pour l'instant cette notion, nous appellerons *seuil identifiable* tout seuil franchi lors de la simulation tel que les paramètres κ ou γ soient différents avant et après le franchissement. Il est possible de générer la liste des seuils identifiables, ainsi que les segments pour chaque molécule qui sont l'ensemble des points pour lesquels les paramètres cinétiques sont constants. On déduit aussi le réseau identifiable constitué par l'ensemble des interactions gène-à-gène actives lors de la trajectoire simulée.

Pour finir, la dernière étape consiste à simuler les données bruitées d'observation : on ajoute à la trajectoire non bruitée un bruit gaussien normal avec un gain sigma : $y(k) = x(k) + \sigma\eta(k)$ avec $\eta \sim \mathcal{N}(0,1)$.

Exemple 7. *Poursuivons avec le cas introduit dans l'Exemple 6. Un exemple de simulation est proposé sur la Figure 4.7, avec les mêmes conditions initiales que celle de la trajectoire de la Figure 4.4. Pour pouvoir exploiter la solution introduite en (4.7), il s'agit de détecter les instants où le prochain seuil est franchi : cela se traduit par une barre verticale verte dans la figure. A ce moment, il peut éventuellement intervenir un changement de dynamique sur l'une des molécules observées : ce changement de mode est marqué par une barre verticale rouge. Ce fait rend le seuil franchi identifiable pratiquement, et de jaune, il appairait sous la forme d'une ligne horizontale magenta.*

4.3 Classification des données bruitées

4.3.1 Détection des transitions

Cette étape est aussi appelée la *segmentation* car elle permet, pour chaque molécule observée, de reconstituer les ensembles de points consécutifs pour lesquels les paramètres dynamiques semblent être constants. Cette étape est particulièrement critique pour la chaîne de traitement, car elle intervient très tôt et les caractéristiques des données telles que le niveau de bruit et le pas d'échantillonnage ont une influence essentiellement pour cette segmentation.

Une méthode pour résoudre le problème de segmentation a été publié dans [89].

L'objectif de cette partie consiste à détecter les transitions entre deux modes dynamiques à partir de données d'observation échantillonnées bruitées, sans connaître au préalable les paramètres cinétiques des concentrations mesurées. Précisons ici que cette dernière hypothèse est la plus contraignante : l'ajout de connaissances préalables est tout à fait envisageable à partir de ce travail, et permet de préciser la segmentation, puisque le problème devient alors très simple dans la mesure où il s'agit seulement d'attribuer les données à des modes prédéfinis. Dans notre cas, nous ne connaissons par avance ni le nombre de ces modes, ni la valeur des paramètres cinétiques correspondant.

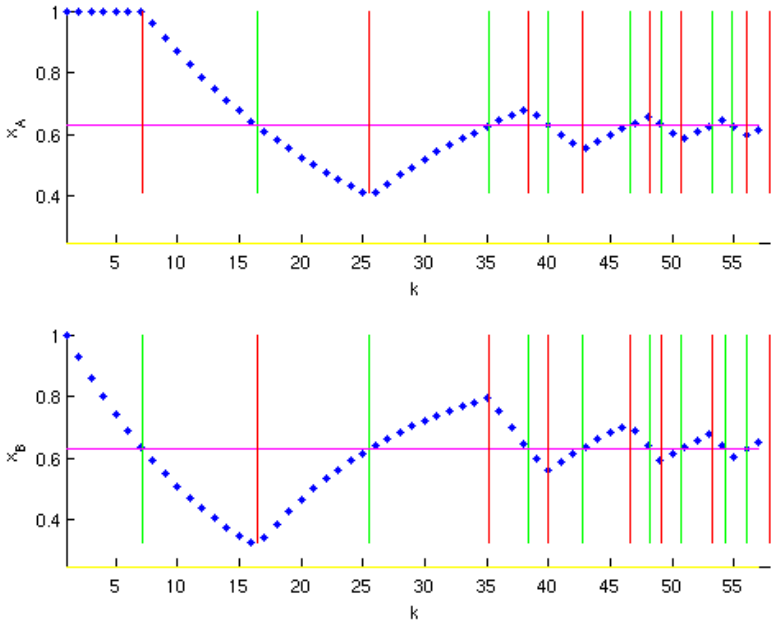


FIGURE 4.7 – Trajectoire simulée pour le réseau de régulation génique à deux gènes proposé Figure 4.3. Les lignes verticales rouges sont les vraies transitions d'un mode à un autre ; les lignes magenta horizontales sont des seuils franchis pendant la trajectoire et qui conduisent à une transition d'un mode à un autre (ils sont donc ce que nous appellerons des seuils identifiables). Les seuils sont franchis au niveau des lignes verticales vertes.

Exemple 8. *En poursuivant l'Exemple 6, il s'agit de retrouver les barres verticales rouges de la Figure 4.4. Pour cela il s'agit de regrouper les points s'ils peuvent avoir été obtenus à partir d'une même trajectoire exponentielle à un bruit gaussien additif près. Il n'est pas possible de transiger parfois quant à l'appartenance à un segment : les points sont alors exclus. Le résultat obtenu par la méthodologie décrite dans ce qui suit est donné Figure 4.8 : on notera que l'aspect le plus délicat consiste à déterminer avec précision l'instant où la trajectoire passe d'un mode dynamique à un autre.*

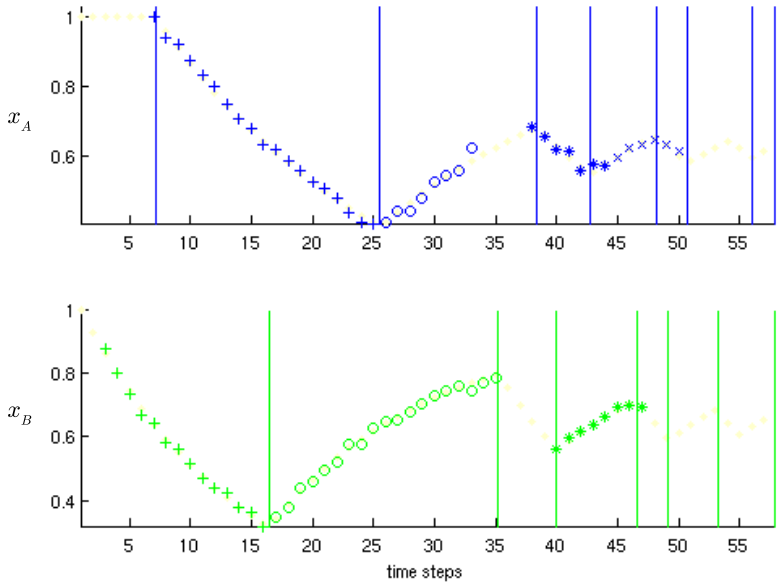


FIGURE 4.8 – Trajectoire segmentée : pour chaque molécule, les modes et leurs paramètres dynamiques sont inférés. Les données qui sont générées par des modes différents sont représentées par des symboles différents. Nous avons de plus représenté au moyen des lignes verticales les transitions véritables : idéalement, aucun segment ne devrait chevaucher ces lignes.

Dans cette partie, le raisonnement porte pour un seul élément m_i de m^*, et chaque élément est donc traité de manière indépendante. De sorte que nous appellerons $y(k)$ au lieu $y_i(k)$ les données de mesure, $x(k)$ au lieu de $x_i(k)$ les concentrations non-bruitées.

La cinétique des points est estimée par une identification itérative non-linéaire, les modes étant distingués par le biais d'un test d'hypothèse.

Régression non-linéaire moindres-carrés

Pour la procédure d'identification de la trajectoire exponentielle, il s'agit d'inférer les paramètres d'une trajectoire exponentielle de type

$$x(k) = (x_0 - \frac{\kappa}{\gamma}) \exp(-\gamma k T) + \frac{\kappa}{\gamma}, \tag{4.8}$$

c'est-à-dire, la condition initiale $x_0 = x(0)$, le taux de synthèse κ et celui de dégradation γ. Pour cela, on utilise une régression moindres-carrés non-linéaire [6, 68].

Le principe général de la régression moindres-carrés non-linéaire est de résoudre un système de p équations à q inconnues ($p > q$) de type

$$\begin{cases} y_1 = f(t_1; \lambda_1, \ldots, \lambda_q) \\ \vdots \\ y_p = f(t_p; \lambda_1, \ldots, \lambda_q) \end{cases} \tag{4.9}$$

tel que les solutions, paramètres de la fonction non-linéaire $f(t)$ choisie, satisfassent de manière optimale le système d'équations (4.9) : on cherche donc à trouver les valeurs des λ_j qui rendent aussi faibles que possible les $d\beta_i(\lambda_1, \ldots, \lambda_q) = y_i - f(t_i; \lambda_1, \ldots, \lambda_q)$ (on parle alors de minimisation des résidus) : $d\beta_i$ mesure l'écart entre l'observation et la valeur théorique "non bruitée". Les algorithmes de minimisation itérative, tels que celui de Gaus-Newton ou celui de Levenberg-Marquardt [66, 73], cherchent à linéariser le modèle en utilisant la matrice jacobienne de f par rapport à ses paramètres. Le principe est le suivant. Le choix des paramètres λ_j se fait de manière itérative. Pour améliorer ce choix, on approxime l'influence des petites variations $d\lambda_j$ sur $d\beta_i$: pour $i \in \{1, \ldots, p\}$

$$d\beta_i = \sum_{j=1}^{q} \frac{\partial f}{\partial \lambda_j} d\lambda_j \Big|_{t_i, \lambda} \tag{4.10}$$

avec $\lambda = [\lambda_1 \cdots \lambda_q]^T$, ce qui est écrit par simplification sous la forme matricielle $d\beta = [d\beta_1 \cdots d\beta_p]^T = Ad\lambda$ où A est une matrice $p \times q$ telle que $A_{ij} = \frac{\partial f}{\partial \lambda_j}\Big|_{t_i, \lambda}$. L'astuce des moindres carrés usuelle consiste à multiplier cette équation par la transposée de A pour obtenir une matrice carrée inversible, ce qui donne : $A^T d\beta = \left(A^T A\right) d\lambda$. Cette équation résolue, on actualise les valeurs λ_j en prenant $\lambda_j + d\lambda_j$. L'itération est stoppée lorsque la nouvelle valeur de la somme des résidus au carré $R^2 = d\beta^T d\beta = \|d\beta\|$ devient inférieure à une valeur très petite fixée. La convergence est d'autant plus rapide que l'initialisation des valeurs des λ_j était proche de la solution, mais celle-ci n'est pas garantie en toute généralité. De plus le problème principal vient du fait qu'un choix erroné pour l'initialisation peut conduire la minimisation itérative à être piégée dans un minimum local. Enfin, il est à noter que, tout comme le méthode des moindres carrés dans le cas linéaire, l'estimation est sensible aux données atypiques.

Dans le cas de la trajectoire exponentielle (4.8), on dispose des points bruités $y(k) = x(k) + \eta(k)$. Prenons pour hypothèse que K^M points consécutifs jusqu'au temps présent $\tilde{k}$ sont considérés comme faisant partie d'un même mode dynamique M, i.e comme appartenant à la même trajectoire exponentielle. Nous allons utiliser les points $y(\tilde{k} - K^M + 1), \ldots, y(\tilde{k})$ pour estimer la valeur des paramètres, soit $\widehat{{x_0}^M}, \widehat{\kappa^M}$ et $\widehat{\gamma^M}$. Pour cela, il nous faut donner les dérivées partielles :

$$\begin{aligned} \frac{\partial x}{\partial \kappa^M}\Big|_{k; {x_0}^M, \kappa^M, \gamma^M} &= \frac{1}{\gamma^M}\left(1 - \exp(-\gamma^M kT)\right) \\ \frac{\partial x}{\partial \gamma^M}\Big|_{k; {x_0}^M, \kappa^M, \gamma^M} &= -\frac{\kappa^M}{{\gamma^M}^2} + \left(\frac{\kappa^M}{{\gamma^M}^2} + (\frac{\kappa^M}{\gamma^M} - {x_0}^M)kT\right)\exp(-\gamma^M kT) \\ \frac{\partial x}{\partial {x_0}^M}\Big|_{k; {x_0}^M, \kappa^M, \gamma^M} &= \exp(-\gamma^M kT) \end{aligned} \tag{4.11}$$

De plus, l'itération est initialisée en prenant en compte les considérations suivantes. La trajectoire part de ${x_0}^M$ et tend vers $\phi^M = \frac{\kappa^M}{\gamma^M}$. La valeur initiale de $\widehat{{x_0}^M}$ est donc

pris comme étant le premier point de la série. De plus, pour des points appartenant au même mode dynamique M, le temps caractéristique d'une trajectoire exponentielle est retrouvé à l'aide d'un rapport sur le taux de variation : $\frac{x(k+1)-x(k)}{x(k)-x(k-1)} = \exp(-\gamma^M T)$. Cependant, l'approximation $\frac{y(k+1)-y(k)}{y(k)-y(k-1)}$ est très sensible au bruit. De manière à limiter l'influence du bruit, il s'agit d'exploiter des points le plus distants possibles, et donc le plus éloignés dans le temps du fait de la monotonicité de la trajectoire exponentielle. Comme de nouveau avec K quelconque tel que les points appartiennent à un même mode dynamique M, $\frac{x(k)-x(k-K+1)}{x(k-1)-x(k-K)} = \exp(-\gamma^M T)$, son approximation par $\frac{y(k)-y(k-K+1)}{y(k-1)-y(k-K)}$ est d'autant plus robuste que K est grand : en effet, pour un rapport de type $\frac{a+\Delta a}{b+\Delta b}$, l'influence de Δa et de Δb est d'autant plus faible que a et b sont grands. On utilisera donc les points extrêmes dont on dispose pour obtenir une approximation de γ^M. Enfin, $\kappa^M = \frac{\gamma^M}{1-\exp(-\gamma^M T)}\left(x(k) - x(k-1)\exp(-\gamma^M T)\right)$: l'espérance empirique préserve le résultat et est ici utilisée pour limiter l'influence du bruit. Il semble donc utile d'initialiser la procédure itérative en utilisant les valeurs suivantes :

$$\begin{aligned}
\widehat{x_0^M}^{(0)} &= y(\tilde{k} - K^M + 1) \\
\widehat{\gamma^M}^{(0)} &= -\tfrac{1}{T}\ln\left(\tfrac{y(\tilde{k})-y(\tilde{k}-K^M+2)}{y(\tilde{k}-1)-y(\tilde{k}-K^M+1)}\right) \\
\widehat{\kappa^M}^{(0)} &= \frac{\widehat{\gamma^M}^{(0)}}{1-\exp(-\widehat{\gamma^M}^{(0)}T)}\frac{1}{K^M-1}\sum_{i=0}^{K^M-2}\left(y(\tilde{k}-i) - y(\tilde{k}-i-1)\exp(-\widehat{\gamma^M}^{(0)}T)\right)
\end{aligned} \tag{4.12}$$

Au moins un certain nombre de points consécutifs sont nécessaires pour estimer la condition initiale, le taux de synthèse et celui de dégradation de la trajectoire exponentielle qui semble être le support des points bruités considérés. On appelle N_c le nombre minimal de points consécutifs utilisés pour commencer un nouveau segment : $N_c > 3$. Il est évident que plus N_c est grand, plus l'estimation sera précise. Mais a contrario, si N_c est trop grand et dépasse le nombre moyen de points consécutifs appartenant à un même mode dynamique, l'estimation n'a plus de sens. Ce compromis est de plus à prendre en compte en envisageant le fait que ce paramètre est aussi à rattacher indirectement au pas d'échantillonnage : en effet, avec une résolution telle que le temps caractéristique (l'inverse du taux de dégradation dans le cadre de notre modèle dynamique) de la variation d'une concentration moléculaire est grand par rapport au temps T d'échantillonnage, les points consécutifs envisagés sont en fait alignés sur la tangente à la trajectoire, au bruit près. L'identification de la trajectoire exponentielle n'a alors pas vraiment de sens. Il faut donc que N_c soit suffisamment grand pour que l'identification itérative non-linéaire puisse renvoyer la condition initiale, le taux de synthèse et celui de dégradation.

Sur la figure 4.9, il apparaît que le paramètre le plus sensible en terme de compromis liant N_c et T est le taux de dégradation (qui correspond au temps caractéristique de la trajectoire exponentielle).

Test d'hypothèse

Afin d'évaluer les transitions entre les divers modes dynamiques, la décision pour mettre fin à un segment de points consécutifs appartenant à un même mode est prise par le biais d'un test d'hypothèse. Le paramètre principal est le niveau de confiance $(1-\alpha)$ choisi pour rejeter l'hypothèse : typiquement $\alpha = 0,05$.

Plus précisément, le principe est le suivant. A l'instant présent $\tilde{k}$, $K^M > N_c$ points consécutifs sont considérés comme faisant parti d'un même mode dynamique M : sur

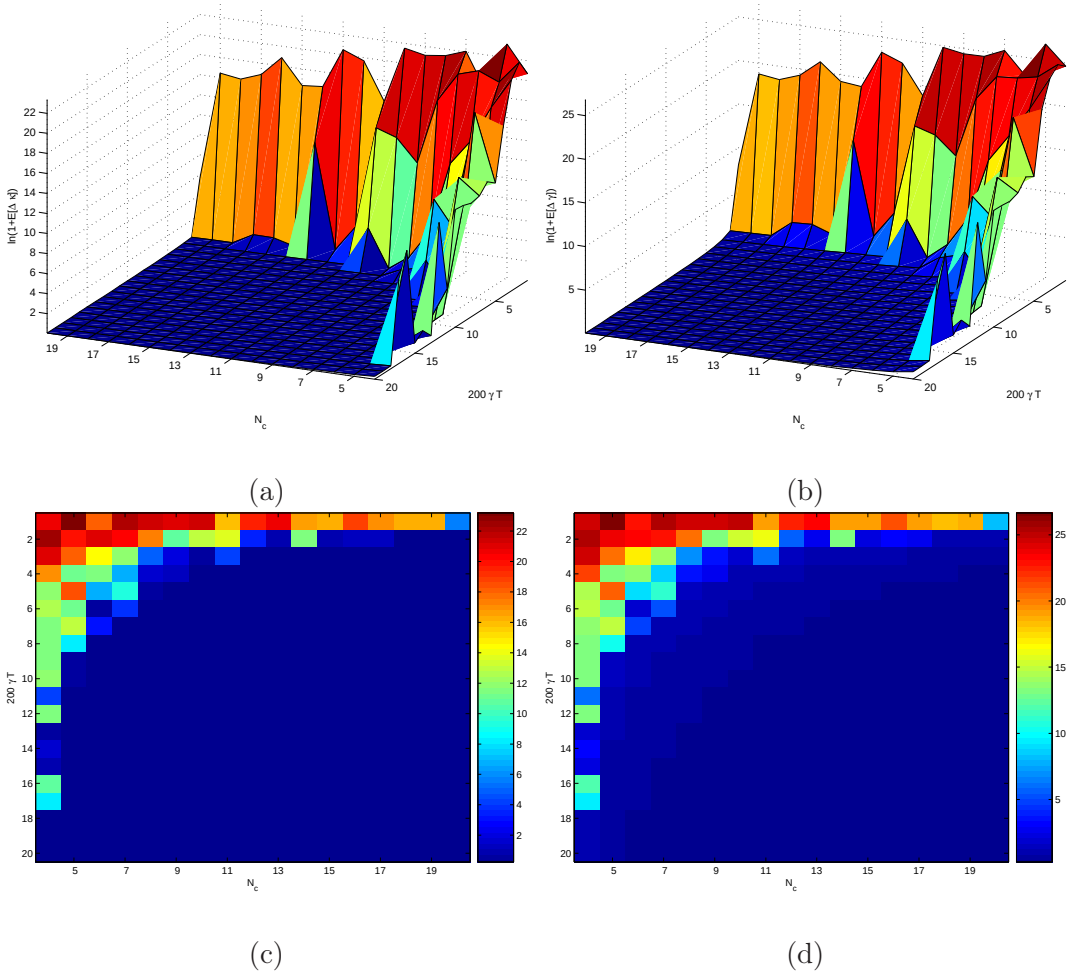


FIGURE 4.9 – Erreurs sur l'évaluation de κ et γ en fonction de T et de N_c : elles deviennent négligeables si T et N_c sont suffisamment grands. $\Delta\kappa = |\kappa - \widehat{\kappa}|$ et $\Delta\gamma = |\gamma - \widehat{\gamma}|$. Les paramètres estimés $\widehat{\kappa}$ et $\widehat{\gamma}$ le sont pour les N_c premiers points échantillonnés selon le pas défini par T. $\kappa/\gamma - x_0 = 1$. Les espérances de $\Delta\kappa$ et $\Delta\gamma$ ont été estimées pour 1000 tirages de bruit différents avec $\sigma = 0,05$. (a) et (c), ainsi que (b) et (d) représentent respectivement en 3D et en niveaux colorés les variations de $\ln(1 + E[\Delta\kappa])$ et de $\ln(1 + E[\Delta\gamma])$ en fonction de T et de N_c.

ces points a été inféré un modèle de trajectoire exponentielle ayant pour paramètres $\widehat{x_0^M}, \widehat{\kappa^M}, \widehat{\gamma^M}$. Le point suivant est alors examiné pour savoir s'il a pu être généré par la même trajectoire. C'est-à-dire, le point $y(\tilde{k}+1)$ peut-il appartenir à la trajectoire exponentielle avec les paramètres décrits étant donné un bruit additif à distribution gaussienne ? Si l'on écrit $\widehat{x}(k) = \left(\widehat{x_0^M} - \frac{\widehat{\kappa^M}}{\widehat{\gamma^M}}\right) \exp(-\widehat{\gamma^M}(k - \tilde{k} + K^M - 1)T) + \frac{\widehat{\kappa^M}}{\widehat{\gamma^M}}$, l'hypothèse nulle est donc :

$$H_0 : y(\tilde{k}+1) - \widehat{x}(\tilde{k}+1) \sim \mathcal{N}(0, \sigma^2). \tag{4.13}$$

La valeur de l'écart type σ est supposée être ici connue. Il est aussi possible de l'estimer empiriquement à partir du calcul de la variance de la liste $(y(k) - \widehat{x}(k))_{k=k-\tilde{K^M}+1}^{\tilde{k}}$.

Étant donné α, puisque la distribution du bruit est supposée normale, l'hypothèse H_0 est acceptée si :

$$y(\tilde{k}+1) - \widehat{x}(\tilde{k}+1) \in \left[-z_{1-\alpha/2}\sigma, z_{1-\alpha/2}\sigma\right] \tag{4.14}$$

où z_q est le q-ième quantile d'une distribution normale standart.

Dans le cas où l'hypothèse est rejetée, c'est-à-dire dans le cas où le point est très peu favorablement issu de la trajectoire estimée avec les points précédents pour un bruit additionnel gaussien, le segment semble être terminé et l'on doit initier un nouveau segment. Dans le cas contraire, le point accepté est alors utilisé pour améliorer l'estimation de la trajectoire qui supporte le segment : on progresse en actualisant $\tilde{k} \leftarrow \tilde{k} + 1$ (et donc $K^M \leftarrow K^M + 1$).

Il se peut que le point traité soit en fait très dégradé par le bruit : pour éviter qu'un tel cas atypique ne soit la cause d'une fragmentation erronée d'un segment, il s'agit de poursuivre le test sur les N_s points suivants. Il est évident que si la trajectoire a changé de dynamique, ces points suivants seront vraisemblablement rejetés par le test d'hypothèse. Inversement, s'ils sont validés, le point présent rejeté est un cas atypique qui appartient au segment : il est possible de ne pas le prendre en compte dans l'estimation des paramètres pour les étapes suivantes. La Figure 4.10 illustre l'influence du choix de N_s sur la segmentation.

Les points critiques d'un segment sont en fait les points extrêmes : en effet, en faisant l'hypothèse que les transitions sont correctement détectées, les points consécutifs à l'intérieur du segment sont très fiables pour le calcul des paramètres de la trajectoire support. Le test d'hypothèse vérifie l'appartenance du point suivant à un segment. Il se base sur l'ensemble des points précédents. Or l'initialisation d'un segment est difficile. De manière à préciser le résultat, on effectue un test d'hypothèse de manière à vérifier que le point à l'instant $\tilde{k} - K^M + 1$ (le premier du segment supposé correspondre au mode dynamique M) appartient bien au segment en se basant sur l'estimation de la trajectoire à partir des $K^M - 1$ points consécutifs qui ne l'incluent pas. Si ce n'est pas le cas, le point est rejeté : $K^M \leftarrow K^M - 1$.

Initialisation d'un segment

Lors de l'arrêt d'un segment, il s'agit d'initialiser un nouveau segment. Pour cela, on prend le point $\tilde{k}$ qui n'a pas été admis comme faisant parti du segment précédant et on envisage les $N_c - 1$ points précédents pour former un nouveau segment, qui ne sera validé qui si le test d'hypothèse sur les N_c points conclut la validité d'un tel segment. Dans ce cas, la procédure décrite auparavant recommence jusqu'à la détection d'une transition. Dans le

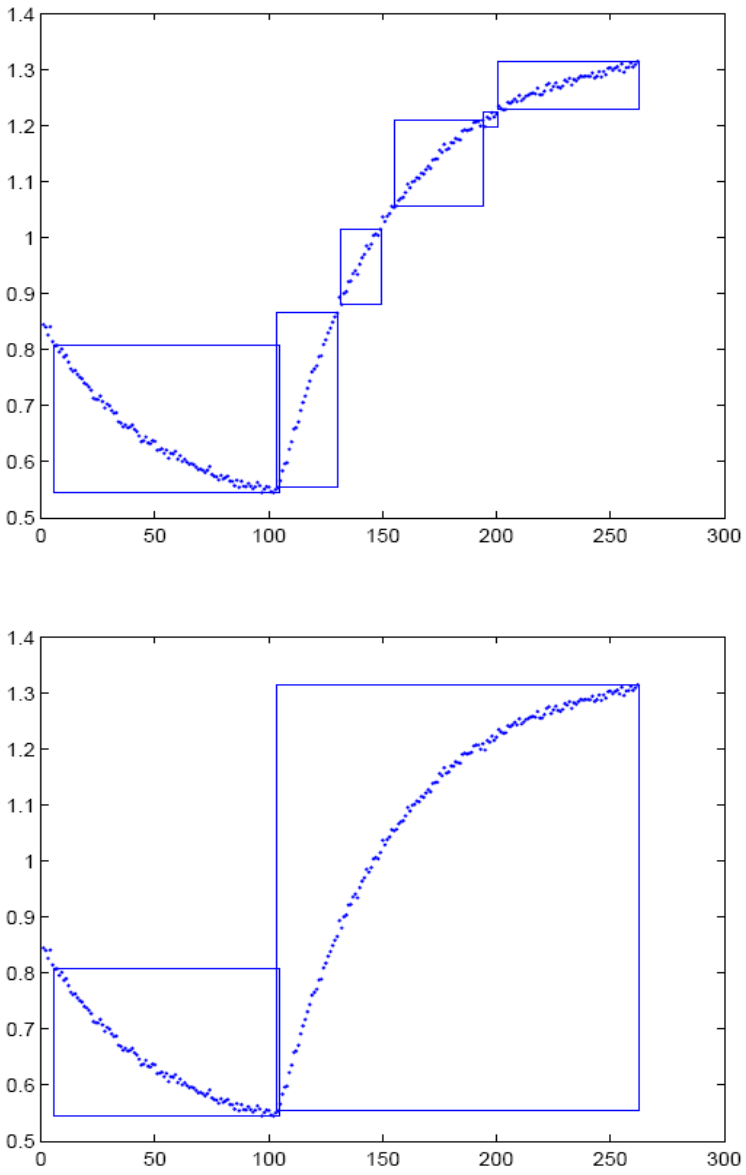


FIGURE 4.10 – Segmentation de la trajectoire : influence du choix de N_s sur le nombre de segments. (en haut) $N_s = 0$; (en bas) $N_s = 1$. Il n'y avait qu'une seule transition pour cet exemple, qui est correctement détectée dans les deux cas.

cas contraire, on incrémente l'instant "présent" ($\tilde{k} \leftarrow \tilde{k}+1$) et l'on cherche de nouveau à initialiser un segment.

Il faut donc souligner ici que l'ensemble des segments générés peuvent éventuellement se superposer. Pour se débarrasser de ce problème d'ambiguïté, les points qui appartiennent à plusieurs segments sont soit éliminés, soit équitablement répartis dans les segments "absorbants" (qui ne partagent que peu de points).

4.3.2 Agrégation des modes

Cette étape permet d'exploiter la segmentation pour classer les points en mode dynamique constant pour toutes les concentrations moléculaires. La version initiale de ce travail a été développée en collaboration avec Riccardo Porreca et Giancarlo Ferrari-Trecate à l'Université de Pavie ; Riccardo Porreca a ensuite optimisé la procédure [88].

Il s'agit de regrouper l'ensemble des points x_i pour un élément donné m_i qui ont un même comportement dynamique : ceux-ci peuvent en effet appartenir à des segments distincts qui correspondent à des domaines de régulation ayant des paramètres dynamiques identiques. Cette étape est justifiée par le fait que c'est l'information liée aux dynamiques qui sera par la suite exploitée. Cela veut dire qu'il faut regrouper les segments qui ont des paramètres semblables : on dira qu'on les agrège. D'un point de vue mathématique, il s'agit de considérer l'ensemble des partitions possibles de la liste des segments et de tester si chaque agrégat résultant peut avoir les mêmes coefficients dynamiques avec une confiance à 99% (ou le niveau choisi). Or, le nombre de partitions possibles étant donnés n éléments est caractérisé par le nombre de Bell B_n (cf. [100]), que l'on peut générer par récurrence avec la formule $B_{n+1} = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) B_k$ (cela donne la suite $1, 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, \ldots$). Il se pose un problème d'optimisation combinatoire dont nous allons montrer qu'une stratégie d'élagage permet d'améliorer son traitement.

La première chose à relever est que si le niveau de confiance est trop élevé, il y a un risque que les agrégats n'agrègent pas correctement les modes dynamiques concernés : dans l'estimation des paramètres faite lors de la segmentation, nous avons vu comment celle-ci est influencée par le bruit ; le problème se pose de nouveau ici. Or, si deux segments qui appartiennent à un même domaine de régulation dans l'espace des concentrations ne sont pas agrégés, notre algorithme va soit faire apparaître des interactions qui n'existent pas, soit aboutir à un cas qu'il ne sait pas inférer. Inversement, si le niveau de confiance est trop bas, des segments vont être agrégés qui n'appartiennent pas au même domaine de régulation : dans ce cas, nous perdons éventuellement l'information sur une interaction et nous ne pourrons donc pas la reconstituer.

Le deuxième point à relever est que, en général, plusieurs résultats sont envisageables. En effet, soit l'élément m_i de m^*, et supposons que nous avons trouvé trois segments S_1^i, S_2^i, S_3^i à l'étape précédente. Admettons que ces trois segments aient des paramètres dynamiques très semblables. Or, pour notre exemple, si l'on regroupe les points de S_1^i et de S_2^i, et on infère les paramètres de la dynamique qui leur correspondent, admettons que ce résultat est accepté étant donné le modèle de bruit. Il se peut que les points de S_3^i ne soient pas compatibles avec le modèle inféré pour l'agrégat de S_1^i et de S_2^i qui n'est pas ni celui inféré pour S_1^i ni celui inféré pour S_2^i. Maintenant, il se peut que l'agrégation de S_1^i et de S_3^i soit possible telle que S_2^i ne soit pas compatible. Nous nous retrouvons avec deux cas, également envisageables : $\{\{S_1^i, S_2^i\}, \{S_3^i\}\}$ et $\{\{S_1^i, S_3^i\}, \{S_2^i\}\}$. Pourtant,

l'agrégation $\{\{S_1^i, S_2^i, S_3^i\}\}$ peut être rejetée. En conclusion, la chaine de traitement peut proposer plusieurs solutions envisageables qu'il n'est pas possible de discriminer.

Rapport de vraisemblance généralisé

Commençons par décrire brièvement la manière par laquelle on décide d'agréger ou pas des segments. Dans cette partie, le raisonnement porte pour un seul élément m_i de m^*, et chaque élément est donc traité de manière indépendante. De sorte que nous appellerons $y(k)$ au lieu $y_i(k)$ les données de mesure, $x(k)$ au lieu de $x_i(k)$ les concentrations non-bruitées.

Appelons[2] $\mathcal{S} = \{S_1, \ldots, S_{|\mathcal{S}|}\}$ l'ensemble des segments obtenus à l'étape précédente. Comme un segment comprend un ensemble de points consécutifs, chaque segment $S \in \mathcal{S}$ est défini par les temps discrets du premier point et du dernier point ; de plus, on a inféré par régression moindres-carrés non-linéaire les paramètres que sont le taux de synthèse $\widehat{\kappa^S}$, le taux de dégradation $\widehat{\gamma^S}$ et la condition initiale $\widehat{x_0{}^S}$. Des segments sont agrégés en une collection $A \subseteq \mathcal{S}$ s'ils ont des taux de dégradation et de synthèse similaires. Pour calculer les paramètres pour A en assumant que le taux de synthèse et de dégradation restent inchangés en tout point quelque soit le segment auquel il appartient, nous utilisons la même méthode qu'en 4.3.1. La seule nuance à apporter par rapport à la partie 4.3.1 est qu'il y a donc moins de paramètres : les taux de synthèse $\widehat{\kappa^A}$ et de dégradation $\widehat{\gamma^A}$ de l'agrégat, ainsi que les conditions initiales spécifiques à chaque segment de l'agrégat $\widehat{x_0{}_1^A}, \ldots, \widehat{x_0{}_{|A|}^A}$. On initialisera l'itération pour les conditions initiales $\{\widehat{x_0{}_j^A}\}_{j=1}^{|A|}$ à l'aide des conditions initiales $\{\widehat{x_0{}^{S_j}}\}_{j:S_j \in A}$, on initialisera le taux de synthèse par la moyenne des taux de synthèse $\frac{1}{|A|}\sum_{j:S_j\in A}\widehat{\kappa^{S_j}}$, et de même pour le taux de dégradation avec $\frac{1}{|A|}\sum_{j:S_j\in A}\widehat{\gamma^{S_j}}$.

Il s'agit ensuite de déterminer si le modèle inféré est consistant avec les données observées. Posons que $\mathcal{S}$ est partitionné en agrégats collectionnés dans P. On utilise le test pour l'hypothèse H_0 défini par

$$\forall A \in P, \forall (j,j') \in \{1,\ldots,|\mathcal{S}|\}^2, (S_j, S_{j'}) \in A^2 \Rightarrow \kappa^{S_j} = \kappa^{S_{j'}} \text{ et } \gamma^{S_j} = \gamma^{S_{j'}} \tag{4.15}$$

contre l'hypothèse où ce n'est pas le cas. Soit $J(P)$ la somme des résidus au carré :

$$J(P) = \sum_{A\in P}\sum_{j:S_j\in A}\sum_{k:y(k)\in S_j}\left[y(k) - \widehat{x(k)}\Big|_{\widehat{\kappa^A},\widehat{\gamma^A},\widehat{x_0{}_j^A}}\right]^2. \tag{4.16}$$

Soit aussi le somme des résidus dans le cas où aucun segment n'est agrégé avec un autre :

$$\bar{J} = \sum_{j=1}^{|\mathcal{S}|}\sum_{k:y(k)\in S_j}\left[y(k) - \widehat{x(k)}\Big|_{\widehat{\kappa^{S_j}},\widehat{\gamma^{S_j}},\widehat{x_0{}^{S_j}}}\right]^2. \tag{4.17}$$

Comme introduit dans [94], et en approximant linéairement $\widehat{x(k)}$ en fonction de κ, γ, x_0, $\frac{v_2}{v_1}(J(P) - \bar{J})/\bar{J}$ a une $F-$distribution avec degré de liberté (v_1, v_2), où v_2 est la différence entre le nombre de données et le nombre de paramètres, et v_1 est le nombre de contraintes liée à H_0 (le nombre minimum d'égalités entre des couples (κ, γ) à écrire). En suivant [94],

2. Par notation, la cardinalité d'un ensemble fini $\mathcal{A}$ sera désignée par $|\mathcal{A}|$. La définition de $\mathcal{S}$ est donc circulaire, et ne sert qu'à préciser explicitement l'écriture des membres de l'ensemble.

le test du *rapport de vraisemblance généralisé* pour rejeter H_0 avec un niveau de confiance $(1-\alpha)$ est si $J(P) \geq (1+\Delta_\alpha(|P|))\bar{J}$, avec

$$\Delta_\alpha(n) = \frac{2(|\mathcal{S}|-n)}{N-3|\mathcal{S}|} F_\alpha(2(|\mathcal{S}|-n), N-3|\mathcal{S}|) \tag{4.18}$$

où $N = |\{y(k) : \exists S \in \mathcal{S}, y(k) \in S\}|$ (ce n'est pas le nombre total de points car certains points ont disparu du fait de la segmentation) et $F_\alpha(v_1, v_2)$ est le $(1-\alpha)$-ième quantile de la F distribution avec les degrés de liberté (v_1, v_2).

Ordre sur les agrégats

Il s'agit maintenant de considérer toutes les partions possibles $\mathcal{P}$ de $\mathcal{S}$ et d'appliquer le test précédant. Cependant, la génération de toutes les partitions possibles est un problème combinatoire. De plus, nous ne désirons pas trouver toutes les partitions non-rejetées par le test d'hypothèse, mais seulement celles qui "agrègent" le plus. Définissons la relation d'ordre pour l'inclusion des agrégats par :

$$\text{soit } (P,Q) \in \mathcal{P}^2, P \leq Q \text{ si et seulement si } \forall A \in P, \exists B \in Q, A \subseteq B. \tag{4.19}$$

Nous recherchons les partitions cohérentes avec les données qui sont maximales pour cette relation (voir Annexe A pour un rappel sur les relations d'ordre). Remarque : cet ordre partiel forme un treillis complet où la borne inférieure est la partition grossière en un seul sous-ensemble et la borne supérieure la partition en singletons.

Un point essentiel pour la suite est la propriété suivante sur $\mathcal{P}$:

Propriété 1. $P \leq Q \Rightarrow J(P) \leq J(Q)$.

Démonstration. $P \leq Q$ implique que l'on étend les contraintes sur les égalités entre les paramètres (κ, γ) en passant de P à Q (les égalités qui étaient vraies pour P le sont pour Q). Or, $J(P)$ est la somme des résidus au carré optimale, et cette optimalité fait que $J(Q)$ ne peut pas être strictement plus petit. ■

On a de plus, si $J(P) \geq (1+\Delta_\alpha(1))\bar{J}$, alors $J(P) \geq (1+\Delta_\alpha(|P|))\bar{J}$ car $\Delta_\alpha(n) \leq \Delta_\alpha(1)$. En utilisant la propriété 1, il vient que, dans le cas où $P \leq Q$, $J(Q) \geq (1+\Delta_\alpha(1))\bar{J}$, ce qui signifie que $J(Q) \geq (1+\Delta_\alpha(|Q|))\bar{J}$. Cela justifie la proposition suivante.

Propriété 2. *Soit $P \in \mathcal{P}$. Si $J(P) \geq (1+\Delta_\alpha(1))\bar{J}$, alors, pour tout $Q \in \mathcal{P}$ tel que $Q \geq P$, l'agrégation correspondant à Q sera rejetée par le test du rapport de vraisemblance généralisé.*

Cette dernière propriété est la base de la méthode qui va permettre de limiter le nombre de partition de $\mathcal{S}$ à tester. Le recherche se produit sur l'arbre des partitions, telle que les enfants d'un noeud P sont toutes les partitions obtenues en agrégeant seulement deux agrégats de P : de sorte qu'en parcourant l'arbre à partir de P, on trouvera toujours des éléments supérieurs. Relevons que l'arbre n'est autre que le diagramme de Hasse (voir Annexe A) pour la relation d'ordre pour l'inclusion des agrégats : l'élément initial est bien-sûr l'élément minimal $\{\{S\}_{S\in\mathcal{S}}\}$. D'après la Propriété 2, si un noeud P vérifie $J(P) \geq (1+\Delta_\alpha(1))\bar{J}$, alors il n'est pas nécessaire de parcourir les noeuds enfants.

Une des difficultés avec cette approche est qu'un noeud donné peut avoir plusieurs parents : pour donner un exemple simple avec trois segments S_1, S_2 et S_3, $\{\{S_1, S_2\}, \{S_3\}\}$

et $\{\{S_1, S_3\}, \{S_2\}\}$ sont deux parents différents de $\{\{S_1, S_2, S_3\}\}$ car $\{\{S_1, S_2\}, \{S_3\}\} \leq \{\{S_1, S_2, S_3\}\}$ et $\{\{S_1, S_3\}, \{S_2\}\} \leq \{\{S_1, S_2, S_3\}\}$ et exactement deux agrégats ont été unifiés dans les deux cas. L'arbre complet pour cet exemple est donné Figure 4.11. Il s'agit alors, à un noeud de l'arbre donné, de ne générer les enfants qui n'ont pas déjà été générés antérieurement (l'approche serait lourdement redondante autrement). Il est possible de rendre cette étape efficace : la description complète de cette étape nécessite l'introduction de notations relatives à la littérature des problèmes combinatoires (i.e. *chaîne de croissance restreinte* [78]), ce qui alourdirait cet exposé ; nous avons préféré ne pas insister sur ce point.

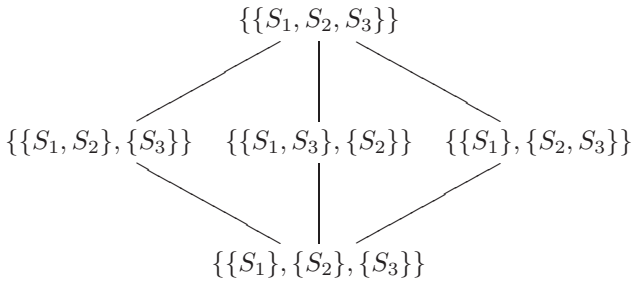


FIGURE 4.11 – Diagramme de Hasse de l'ensemble $\mathcal{P}$ de toutes les partitions possibles de $\mathcal{S} = \{S_1, S_2, S_3\}$, ensemble partiellement ordonné pour la relation d'ordre définie dans la relation (4.19).

Une manière efficace de parcourir l'arbre des partitions s'inspire de l'algorithme *Apriori* [1], méthode classique dans la communauté de l'extraction de connaissances (data mining) pour apprendre des règles d'association (se rapporter à [110]). Cet algorithme inclut la possibilité de limiter la recherche en se basant sur la Propriété 2. L'idée est faire la recherche dans l'arbre en largeur-d'abord. Pour une profondeur donnée, pour tous les noeuds non-rejetés (il n'est donc pas nécessaire de connaître tous les noeuds), on génère les enfants. Pour chacun de ces enfants, on vérifie qu'il n'y a pas au moins un parent Q qui vérifie la condition de rejet $J(Q) \geq (1 + \Delta_\alpha(|Q|))\bar{J}$, auquel cas l'enfant est rejeté. On progresse ainsi jusqu'à la profondeur maximale à moins qu'à une profondeur donnée, tous les enfants soient rejetés. Les solutions sont les noeuds non-rejetés pour la profondeur la plus grande.

4.3.3 Classification

La classification consiste à regrouper les points de mesure $\{\mathbf{y}(k)\}_{k=0}^{N}$ avec $\mathbf{y}(k) = (y_1(k), \ldots, y_{|m^*|}(k))^T$ en une liste de modes, nommée[3] $\mathcal{F}^* = \{\mathcal{F}_1, \ldots, \mathcal{F}_{|\mathcal{F}^*|}\}$, tels que tous les points d'un même mode correspondent à des paramètres dynamiques identiques : souvent, un mode correspond à un domaine de régulation, mais il peut arriver aussi qu'un mode corresponde à plusieurs domaines de régulations voisins ayant des paramètres dynamiques identiques.

Appelons $\mathcal{P}^i$ l'ensemble des partitions obtenues pour l'ensemble des segments $\mathcal{S}^i$ des données de mesure de l'élément $m_i \in m^*$. A tout élément de $\mathcal{P}^1 \times \ldots \times \mathcal{P}^{|m^*|}$ correspond un cas de classification.

Prenons un cas donné par les partitions $P^1 \in \mathcal{P}^1, \ldots, P^{|m^*|} \in \mathcal{P}^{|m^*|}$. Il s'agit de calculer l'intersection entre chacun des agrégats : des points appartiennent à une même classe s'ils appartiennent aux mêmes agrégats pour l'ensemble des molécules. Formellement,$\forall \mathcal{F} \in$

3. voir Note 2.

$\mathcal{F}^*, \forall \mathbf{y}(k) \in \mathcal{F}, \exists!(A_1, \ldots, A_{|m^*|}) \in P^1 \times \ldots \times P^{|m^*|}$ tel que $\forall i \in \{1, \ldots, |m^*|\}, \exists S_i \in A_i$ tel que $y_i(k) \in S_i$.

Exemple 9. *En poursuivant l'Exemple 6 pour un réseau de régulation à deux gènes, il s'agit de combiner les segments de la Figure 4.8 s'ils ont des paramètres dynamiques suffisamment semblables. Un seul cas est obtenu, pour lequel les points sont classés comme indiqué dans la Figure 4.12.*

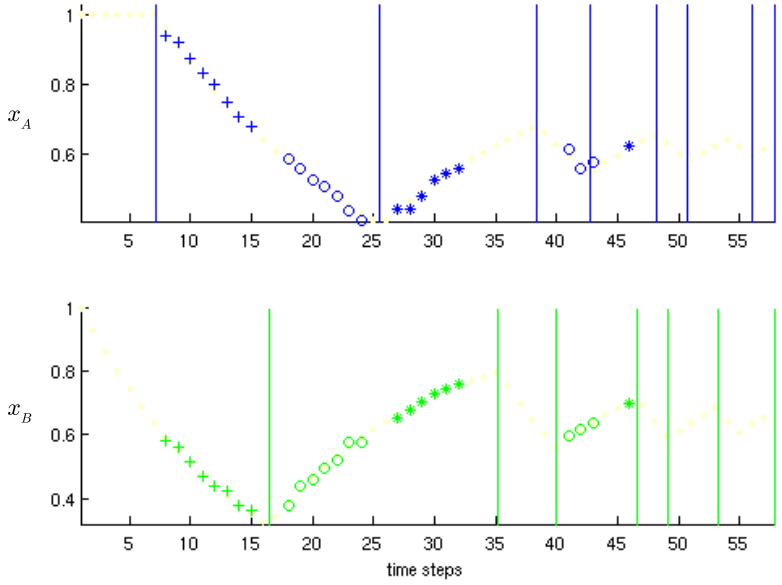


FIGURE 4.12 – Classification des points en modes pour lesquels les paramètres dynamiques sont constants pour toutes les molécules (auquel cas il y aura un même symbole). Il y a trois symboles utilisés, il y aura donc trois classes : $\mathcal{F}_1$ (croix), $\mathcal{F}_2$ (cercle) et $\mathcal{F}_3$ (étoile). Les lignes sont les vraies transitions.

4.4 Énumération des solutions parcimonieuses

4.4.1 Reconstruction des seuils

Il faut raisonner ici pour un cas donné issu de la classification. On dispose donc d'un ensemble de s ensembles de points regroupés par mode, nommé $\mathcal{F}^* = \{\mathcal{F}_1, \ldots, \mathcal{F}_s\}$, qu'il s'agit de séparer de manière à reconstruire les seuils.

Nous avons développé un formalisme, qui sera détaillé dans les chapitres 5 et 6 suivants, qui permet de préciser les interactions entre gènes à partir de la classification que nous venons de proposer. En effet, lorsque l'on passe d'une classe à un autre, l'information ne porte que sur un seul niveau : nous savons qu'un certain nombre d'éléments de m^* ont vu leur cinétique modifiée, et nous pouvons même qualifier ce changement en considérant les écarts sur le taux de synthèse et sur le taux de dégradation d'une classe à l'autre. En effet, si les taux $\kappa_i^{(j)}$ et $\gamma_i^{(j)}$ sont les taux de synthèse et de dégradation de l'élément m_i pour la

classe $\mathcal{F}_j$, alors, si l'on passe de $\mathcal{F}_j$ à $\mathcal{F}_{j'}$, $\kappa_i^{(j')} - \kappa_i^{(j)}$ sera négatif si l'expression de m_i est bridée, positif si elle est encouragée ; de même, $\gamma_i^{(j')} - \gamma_i^{(j)}$ sera positif si la dégradation de m_i est stimulée, négatif si elle est bridée. Qu'ils soient négatifs ou positifs, il existe donc une interaction arrivant vers l'élément m_i.

Cependant, nous ne disposons pas de l'information qui nous permet de préciser l'élément qui est à l'origine de ce changement. Dans le cadre de notre modèle, nous savons qu'il existe une transition sur un élément de m^* qui a été franchie. Le franchissement de ce seuil, s'il est de nouveau franchi, doit aboutir aux mêmes effets. C'est donc cette notion même de franchissement de seuil qui est ici travaillée. En l'occurrence, si un seuil est trouvé pour un premier élément qui explique la variation de dynamique d'un deuxième élément (dinstinct ou non), alors il y a une interaction possible entre le premier et le deuxième.

Un point à souligner à ce niveau est que la classification des points en modes est beaucoup plus informative que la collection des transitions entre ces modes observés pour la trajectoire étudiée. En effet, un seuil peut éventuellement expliquer plusieurs transitions. Mais plus encore, l'action d'un seuil peut ne pas être traduit en transition dans la dynamique à un temps donné (du fait de l'action prioritaire d'une autre élément à ce même moment) : et pourtant, il se peut que son observation à un autre instant permette de le reconstruire.

Pour ce faire, nous exploitons le fait que les points (supposés correctement classés) occupent un sous-espace d'un même domaine de régulation (en approximation du fait qu'il s'agit de l'ensemble des domaines de régulation joints par une face ayant les mêmes paramètres dynamiques). Ces domaines étant hyper-rectangles, il s'agit d'en reconstituer les faces au moyen d'hyperplans qui intersectent les axes au niveau des seuils de transition des facteurs de la régulation. Un hyperplan qui sépare deux classes est nommé césure. L'ensemble des césures est regroupé dans l'ensemble $\mathcal{C}^*$.

Un hyperplan $\mathcal{C}$ de $\mathcal{C}^*$ sépare en général plus que deux classes. L'ensemble des couples de classes qu'il sépare est nommé son pouvoir de séparation. La propriété qui sera intéressante est le fait que certaines césures auront un pouvoir de séparation qui inclut le pouvoir de séparation d'autres césures. En l'occurrence, elles disent la même chose sur la séparation des classes, et elles en disent davantage. Nous nous servirons de ce critère pour éliminer un certain nombre de césures dont le pouvoir explicatif est faible.

Exemple 10. *En poursuivant l'Exemple 6, nous avons obtenu une classification des données Figure 4.12. Lorsque l'on observe ces classes dans l'espace des concentrations (même représentation que sur la Figure 4.5), nous constatons que deux hyperplans (des droites) césures permettent de séparer les classes, comme cela est représenté sur la Figure 4.13.*

4.4.2 Génération des réseaux

Cependant, pour un cas de classification donné, nous disposons toujours d'un ensemble de césures qui donnent des informations en terme de séparabilité des classes qui sont redondantes. En fait, seul un sous-ensemble de ces césures est suffisant pour expliquer complètement la classification. Or, nous avons vu qu'à chaque césure correspond au moins une interaction : à chaque ensemble de césures correspond un réseau de régulation qui est consistant avec les données de mesure (en fait de classification).

Nous appellerons multicésure $\mathcal{M}$ un ensemble de césures qui est suffisant pour expliquer la classification. On ne peut pas préjuger de la cardinalité minimale des multicésures : nous montrerons que si l'ensemble des césures est une multicésure, alors nous pouvons

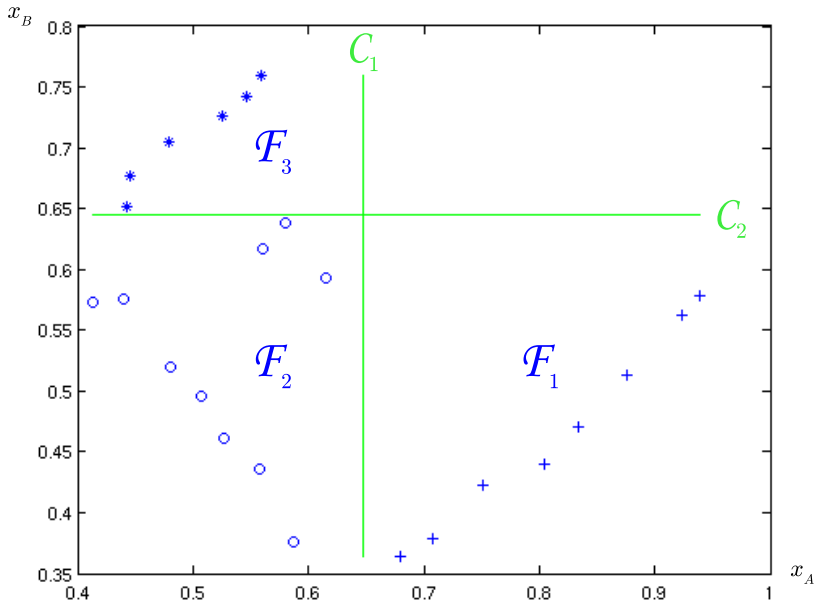


FIGURE 4.13 – Les classes de données $\{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3\}$, issues du traitement de la trajectoire de la Figure 4.4, sont séparées dans l'espace des phases au moyen d'hyperplans nommés césures, $\mathcal{C}_1$ (césure sur x_A) et $\mathcal{C}_2$ (césure sur x_B).

dériver toutes les multicésures de cardinalité inférieure. Chaque multicésure correspond à un modèle de réseau de régulation, et plus il y a de césures, plus il y a de seuils, plus il y a d'interactions.

Il n'y a pas a priori aucune raison de privilégier un modèle de réseau ou un autre. Cependant, il nous est apparu intéressant d'utiliser ici le principe de parcimonie : parmi plusieurs modèles qui ont le même pouvoir explicatif, le modèle le plus simple possible est celui qu'il faut privilégier. En effet, il est toujours possible de complexifier un modèle sans que cela n'affecte son comportement sur une plage d'observations ; il semble donc que le modèle le plus simple suffisant pour expliquer l'observable est aussi celui qui est le plus probablement à l'origine de l'observation, tous les autres ne comprenant que des "ajouts" inutiles étant donné ce qui a été constaté. Nous sommes dès lors davantage intéressés par des modèles qui, tout en restant consistants avec la classification, contiennent le nombre minimal de seuils. Nous verrons dans le chapitre 6 suivant comment formaliser ce problème d'optimisation.

En conclusion, pour un cas de classification donné, on donnera la liste ${\mathcal{M}^*}_{\min}$ de toutes les multicésures de cardinalité minimale et les réseau d'interactions associés. Il ne s'agit donc pas d'une "réponse" unique. Les multicésures de ${\mathcal{M}^*}_{\min}$ sont appelées solutions parcimonieuses.

Exemple 11. *En poursuivant l'exemple du réseau à deux gènes de la Figure 4.3, nous pouvons constater qu'il n'est pas possible de se passer d'aucune des deux césures $\mathcal{C}_1$ et $\mathcal{C}_2$ décrites Figure 4.13 : $\mathcal{C}_1$ est la seule césure à séparer $\mathcal{F}_1$ et $\mathcal{F}_2$, quand $\mathcal{C}_2$ est la seule à séparer $\mathcal{F}_2$ et $\mathcal{F}_3$. La multicésure $\mathcal{M} = \{\mathcal{C}_1, \mathcal{C}_2\}$ est donc de cardinalité minimale. Pour*

notre exemple, $\mathcal{M}^*_{\min} = \{\mathcal{M}\}$. En comparant les Figures 4.14 et 4.7, on constate que ces césures reconstituent bien les seuils franchis (qui étaient représentés en magenta). L'analyse des actions liées aux césures permet bien d'inférer le réseau identifiable décrit Figure 4.6.

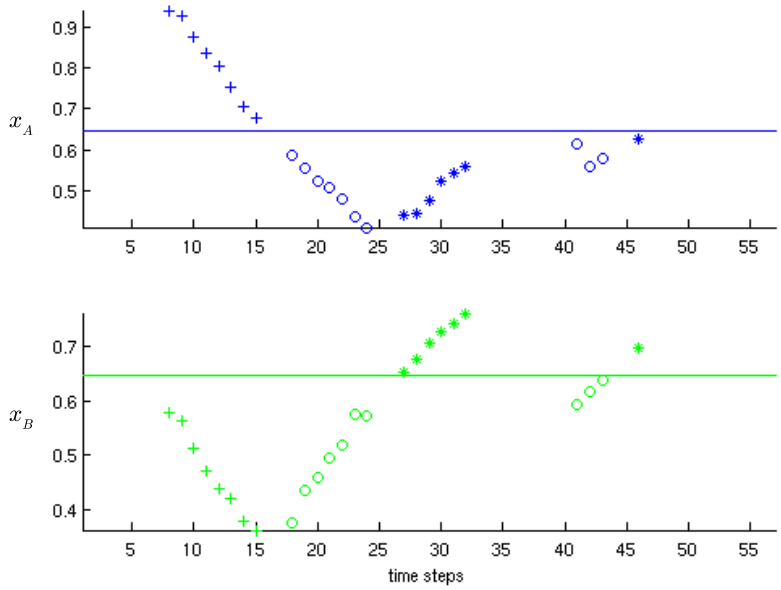


FIGURE 4.14 – Les données classées sont séparées de manière optimale par deux lignes horizontales, une pour chaque molécule, qui identifient correctement les seuils (lignes magenta de la Figure 4.4) et leur action (la protéine A inhibe le gène dont le produit de l'expression est B, et la protéine B active le gène exprimé en A).

4.5 Chaine de traitement

Nous sommes maintenant à même de décrire la chaine de traitement qui est schématisée sur la Figure 4.2 (p.47). À partir du modèle de réseau d'interaction génique tel que décrit dans le chapitre 4.2.1, il est possible de simuler des trajectoires en fonction des conditions initiales sur les concentrations x_i des protéines $m_i \in m^*$ ($i \in \{1, \ldots, |m^*|\}$), d'un pas d'échantillonage T et de l'écart-type σ du bruit gaussien additif. L'idée est bien-sûr de pouvoir remplacer ultérieurement les données simulées par des données d'expérience biologique pour des réseaux à inférer. La simulation permet de préciser exactement les conditions soumises à la chaine de traitement en aval : cela permet de tester intensivement l'effet des divers paramètres.

La partie correspondant à la segmentation (traitement décrit en 4.3.1), à l'agrégation (voir en 4.3.2) puis à la classification (décrite en 4.3.3) permet de regrouper les données d'observation en classes $\mathcal{F}^*$ telles que leurs modèles dynamiques inférés soient constants pour les points d'une même classe mais différents les uns des autres. Si la segmentation est unique pour une trajectoire donnée, l'agrégation conduit quant à elle à des choix de classes élémentaires multiples pour lesquels il n'est pas possible d'en privilégier un seul. L'étape

de classification qui en fait le produit cartésien démultiplie encore cet effet.

L'existence de la classification (qui distingue plusieurs modes dynamiques) est la "trace" laissée par les seuils de transition du modèle identifiable dans le cas de la trajectoire considérée. Cette trace est exploitée pour reconstruire les seuils. La première étape de cette reconstruction consiste à décrire les césures (voir 4.4.1) qui séparent les classes et à déduire les actions correspondantes (arc sur le diagramme du réseau d'interaction identifiable). En supposant que les césures $\mathcal{C}^*$ séparent bien l'ensemble des classes, la dernière étape consiste à énumérer l'ensemble ${\mathcal{M}^*}_{\min}$ des sous-ensembles de $\mathcal{C}^*$ qui vérifient une condition d'optimalité obtenue en appliquant le principe de parcimonie. On les appelle les multicésures parcimonieuses. Chacune d'entre elles représente un modèle de réseau d'interactions acceptable étant donnée la trajectoire.

Dans le cas de l'Exemple 6, nous obtenons une seule solution. Pour des paramètres plus contraignants, il est possible de ne pas avoir de solution ou d'en avoir plusieurs. Pour un modèle possédant plus de deux gènes, on peut s'attendre à rencontrer plus fréquemment plusieurs solutions. S'il ne sera pas possible de privilégier l'une d'entre elles, il sera cependant intéressant de relever les interactions qui reviennent assez régulièrement parmi elles : elles disposent d'un pouvoir informatif plus important.

4.6 Des données aux réseaux de régulation

Nous avons développé en collaboration avec Riccardo Porreca (Université de Pavie) un programme sous Matlab de Mathworks permettant de générer les résultats de la procédure d'identification que nous avons décrite. Matlab est un logiciel de calcul scientifique assez puissant, à compilation à la volée, ce qui permet de développer très simplement des programmes exécutables. Il est possible de visualiser les variables et leur état en cours d'exécution. Des outils d'affichage scientifique ainsi que des boites à outil (pour le traitement statistique, pour l'inférence, pour l'optimisation, par exemple) sont aussi proposés.

Nous n'allons pas ici fournir une documentation détaillée du programme. Nous donnerons seulement un aperçu de la simplicité avec laquelle la chaine de traitement peut être appliquée. En Annexe B, on trouvera un exemple de script en Matlab permettant de procéder à la simulation puis à l'identification d'un réseau de régulation génique à deux gènes.

Dans un premier temps, le modèle est chargé en mémoire : il décrit la collection d'éléments en interaction (i.e. les produits d'expression des gènes), puis donne les équations de la dynamique de la concentration de ces éléments en fonction des concentrations de toutes les éléments. Les valeurs des seuils, des taux de synthèse et de dégradation dans les équations, sont entrées à ce niveau.

Dans la deuxième étape, les points de la trajectoire sont simulés en choisissant les conditions initiales, le pas T, le bruit σ. Comme il s'agit d'un modèle avec erreur sur la sortie, le bruit est ajouté à la fin. Nous sauvegardons ainsi la trajectoire non bruitée de manière à pouvoir tester seulement l'effet du bruit (avec des tirages différents).

Nous procédons ensuite à la segmentation, à l'agrégation puis à la classification. Chacune de ces étapes possède une série de paramètres qu'il est possible de régler. Le plus important est sans doute le niveau de confiance α qui permet de rendre l'inférence plus ou moins rapide/exigeante. Un certain nombre de cas pathologiques sont corrigés de manière automatique : par exemple, les classes contenant trop peu de points sont éliminées.

Viennent ensuite les étapes de reconstruction des seuils et des interactions géniques : le

principe est celui proposé au Chapitre 6. Les performances sont aussi évaluées : ceci sera présenté dans le Chapitre 8.

La structure du programme ainsi que les variables internes majeures sont décrites dans la Section 9.2.2. En effet, il nous est apparu intéressant de pouvoir formaliser celles-ci au travers d'outils plus généraux qui permettent de commencer à faire sortir notre procédure de la sphère Matlab au sein de laquelle elle a été développée et testée.

Chapitre

5 Césures, aspects théoriques

Soit $\mathcal{F}_1, \ldots, \mathcal{F}_s$ une collection d'ensembles disjoints finis de points dans $\mathbb{R}^n_+$ appelés *ensembles de données* avec $s \in \mathbb{N} \setminus \{0\}$ et $n \in \mathbb{N} \setminus \{0\}$. Appelons $\mathcal{F}^* = \{\mathcal{F}_1, \ldots, \mathcal{F}_s\}$ (de manière générale dans ce mémoire, un ensemble étoilé est l'ensemble de tous les ensembles d'un même type).

Dans ce qui suit, nous allons nous concentrer sur le problème de la séparation des ensembles de $\mathcal{F}^*$ par l'intermédiaire d'hyperplans parallèles à une combinaison linéaire de $n-1$ axes (c'est-à-dire, dont le vecteur normal se confond avec le vecteur support d'un des axes) : en effet, dans notre modèle biologique, ce sont les seuils sur les concentrations qui séparent les domaines de régulation, et cela impose donc cette contrainte géométrique.

Pour illustrer les concepts principaux, nous utiliserons l'exemple de la collection $\mathcal{F}^*$ dépeinte Figure 5.1(a) : trois classes sont séparés dans un espace bidimensionnel. Les paires d'ensembles distincts dans $\mathcal{F}^*$ auront souvent pour indice des paires dans l'ensemble $U = \{(p,q) \in \{1, \ldots, s\}^2 : p < q\}$. Enfin, la cardinalité d'un ensemble fini $\mathcal{A}$ sera désignée par $|\mathcal{A}|$.

5.1 Hyperplans séparateurs

Définition 1 (Parax-hyperplan)**.** *Un* (parax-) hyperplan *dans* $\mathbb{R}^n$*, parallèle à* $n-1$ *des axes, avec pour direction l'axe orthogonal* $i \in \{1, \ldots, n\}$*, est un hyperplan solution de l'équation* $x_i = \alpha$*,* $\alpha \in \mathbb{R}$*, ou, de manière équivalente, l'ensemble de niveau zéro de la fonction* $\theta(x) = x_i - \alpha$*.*

Pour préciser la notation, x_i est la i-ième composante du point représenté par le vecteur x. Par abus de notation, θ désignera à la fois un parax-hyperplan et sa fonction associée. La fonction $dir(\theta)$ donne la direction i du parax-hyperplan θ, et la fonction $\mathrm{Z}(\theta)$ donne le niveau zéro α. $dir(\theta)$ et $\mathrm{Z}(\theta)$ caractérisent sans ambigüité θ.

Définition 2 (Séparabilité)**.** *Soient* $\mathcal{F}_p$ *et* $\mathcal{F}_q$ *des ensembles disjoints finis de points appartenant à* $\mathbb{R}^n$*. Un parax-hyperplan* θ *de* $\mathbb{R}^n$ sépare $\mathcal{F}_p$ *et* $\mathcal{F}_q$ *s'il existe* $\delta \in \{+1, -1\}$ *tel que, pour tout* $x \in \mathcal{F}_p \cup \mathcal{F}_q$*, on ait*

$$\begin{cases} \delta\,\theta(x) > 0, & si\ x \in \mathcal{F}_p, \\ \delta\,\theta(x) < 0, & si\ x \in \mathcal{F}_q. \end{cases} \tag{5.1}$$

Dans ce cas, on écrira $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$*.* $\mathcal{F}_p$ *et* $\mathcal{F}_q$ *sont* séparables *s'il existe un parax-hyperplan les séparant.*

Exemple 12. *Sur la Figure 5.1(c),* $\mathcal{F}_1$ *et* $\mathcal{F}_2$ *sont séparables dans la mesure où il existe des parax-hyperplans dans la direction* 1 *(e.g.,* $\theta_{(1),1}$ *ou* $\theta_{(2),1}$ *ou* $\theta_{(3),1}$*), tels que les points de* $\mathcal{F}_1$ *sont positionnés sur un même côté de chaque hyperplan quand tous les points de* $\mathcal{F}_2$ *se trouvent de l'autre côté. On notera encore que les ensembles* $\mathcal{F}_1$ *et* $\mathcal{F}_2$ *ne sont pas séparables dans la direction* 2*.*

Comme il est possible de le vérifier sur la Figure 5.1(c), le parax-hyperplan $\theta_{(1),1}$ *sépare davantage d'ensembles de* $\mathcal{F}^*$ *que le parax-hyperplan* $\theta_{(2),1}$*. L'exemple de* $\theta_{(2),1}$ *illustre le*

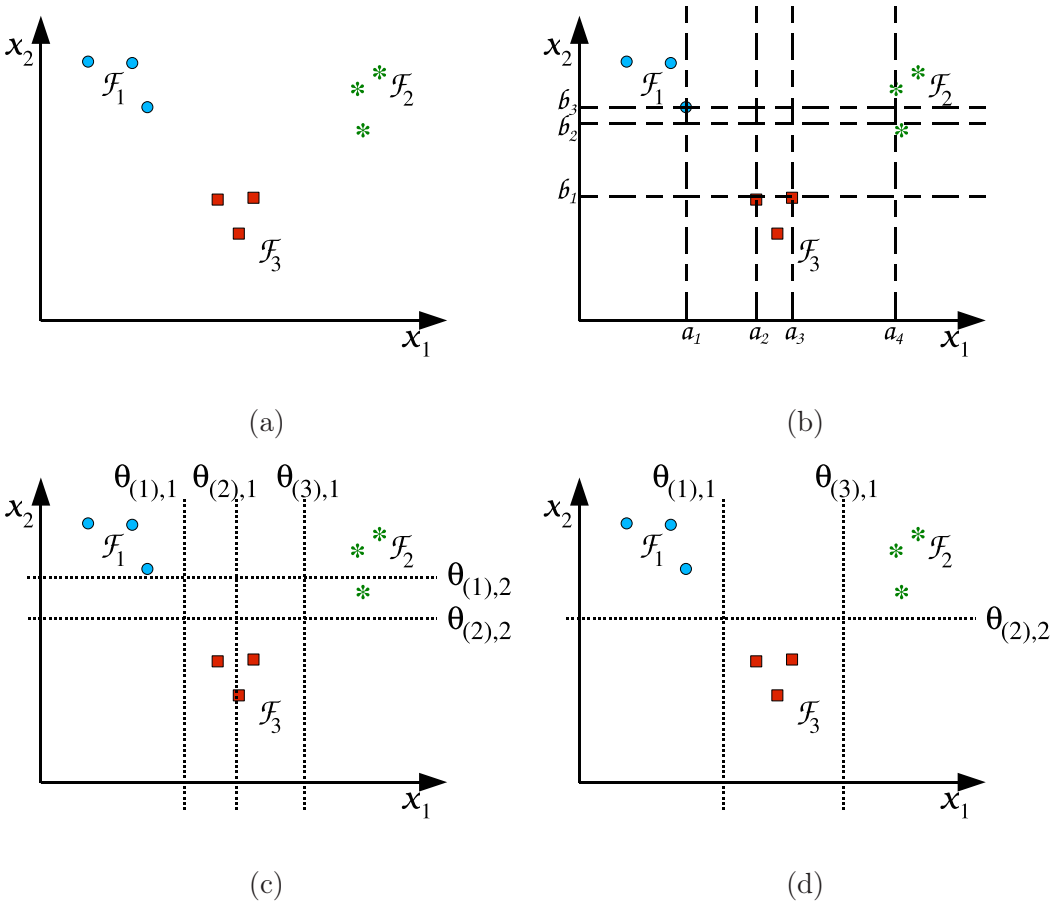


FIGURE 5.1 – Exemple simple. (a) Ensemble de données $\mathcal{F}^*$. (b) Limite des classes d'équivalence. (c) Multicésure $\mathcal{C}^* = \{\theta_{(1),1}, \theta_{(2),1}, \theta_{(3),1}, \theta_{(1),2}, \theta_{(2),2}\}$. (d) Multicésure $Max_{\preceq}\, \mathcal{C}^* = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$.

fait que les ensembles de données peuvent soit se trouver d'un côté d'un parax-hyperplan, soit chevaucher un parax-hyperplan.

Pour indiquer les positions relatives des ensembles de $\mathcal{F}^*$, on introduit les fonctions :

$$\begin{array}{l} \mathcal{I}_0(\theta) = \{j : (\exists x \in \mathcal{F}_j, \theta(x) \leqslant 0) \text{ et } (\exists x \in \mathcal{F}_j, \theta(x) \geqslant 0)\}, \\ \mathcal{I}_-(\theta) = \{j : \forall x \in \mathcal{F}_j, \theta(x) < 0\}, \\ \mathcal{I}_+(\theta) = \{j : \forall x \in \mathcal{F}_j, \theta(x) > 0\}. \end{array} \tag{5.2}$$

Il est intéressant de relever que $\mathcal{I}_0(\theta), \mathcal{I}_-(\theta)$ et $\mathcal{I}_+(\theta)$ partitionnent $\{1, \ldots, s\}$ ($s = |\mathcal{F}^*|$), c'est-à-dire les ensembles sont deux à deux disjoints et

$$\mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta) \cup \mathcal{I}_0(\theta) = \{1, \ldots, s\}. \tag{5.3}$$

De la définition 1 et de (5.2), il vient aussi que :

$$\begin{array}{l} \forall p \in \mathcal{I}_-(\theta), \forall x \in \mathcal{F}_p, x_{dir(\theta)} < \mathrm{Z}(\theta), \\ \forall q \in \mathcal{I}_+(\theta), \forall x \in \mathcal{F}_q, x_{dir(\theta)} > \mathrm{Z}(\theta). \end{array} \tag{5.4}$$

De sorte que $(p, q) \in \mathcal{I}_-(\theta) \times \mathcal{I}_+(\theta)$ si et seulement si $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$.

Exemple 13. *Sur la Figure 5.1, $\theta_{(1),1}$ sépare à la fois $\mathcal{F}_1$ de $\mathcal{F}_2$ et $\mathcal{F}_1$ de $\mathcal{F}_3$. De plus, on a $\mathcal{I}_-(\theta_{(1),1}) = \{1\}$ et $\mathcal{I}_+(\theta_{(1),1}) = \{2, 3\}$. On notera aussi que le parax-hyperplan $\theta_{(2),1}$ sépare seulement $\mathcal{F}_1$ de $\mathcal{F}_2$.*

La différence en terme de pouvoir de séparation des parax-hyperplans peut être définie de manière plus formelle comme suit.

Définition 3 (Pouvoir de séparation)**.** *Le* pouvoir de séparation *d'un parax-hyperplan θ est la fonction*

$$S(\theta) = \{(p, q) \in U : \mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q\}. \tag{5.5}$$

Par la suite, nous nous préoccuperons des parax-hyperplans appartenant à la restriction à l'ensemble $\Theta = \{\theta : S(\theta) \neq \emptyset\}$. Nous ferons aussi l'hypothèse que Θ n'est pas vide.

La comparaison du pouvoir de séparation des parax-hyperplans de Θ ayant une même direction suscite l'introduction des classes d'équivalence de parax-hyperplans.

Définition 4 (Équivalence)**.** *Deux parax-hyperplans $\theta, \theta' \in \Theta$ sont* équivalents *si $dir(\theta) = dir(\theta')$ et $S(\theta) = S(\theta')$. La relation d'équivalence*[1] *est notée $\theta \sim \theta'$.*
*L'*ensemble quotient *est noté $\mathcal{E}^* = \Theta/\sim$. La classe d'équivalence d'un parax-hyperplan $\theta \in \Theta$ est désignée par $[\theta] = \{\theta' : \theta' \sim \theta\}$, que l'on confond avec la surjection canonique de Θ dans l'ensemble quotient $\mathcal{E}^*$.*

Exemple 14. *Comme il peut être vérifié au moyen de la Définition 4, les parax-hyperplans $\theta_{(1),1}$ et $\theta_{(2),1}$ sur la Figure 5.1(c) ne sont pas équivalents.*

Il est alors utile de généraliser les fonctions définies pour un parax-hyperplan de Θ à sa classe d'équivalence dans $\mathcal{E}^*$ en utilisant la notion d'invariance. Rappelons que, étant donnée une relation d'équivalence $\sim$ définie sur un ensemble X et une fonction $f : X \to Y$, f est *invariante* pour $\sim$ si $x \sim y$ implique $f(x) = f(y)$. Il est évident que les fonctions *dir* et S sont invariantes pour la relation d'équivalence $\sim$ introduite dans la Definition 4.

1. Il est aisé de montrer que la relation $\sim$ est réflexive, symétrique et transitive

Propriété 3. *Les fonctions $\mathcal{I}_0$, $\mathcal{I}_-$ et $\mathcal{I}_+$ sont invariantes pour la relation d'équivalence $\sim$ introduite dans la Definition 4.*

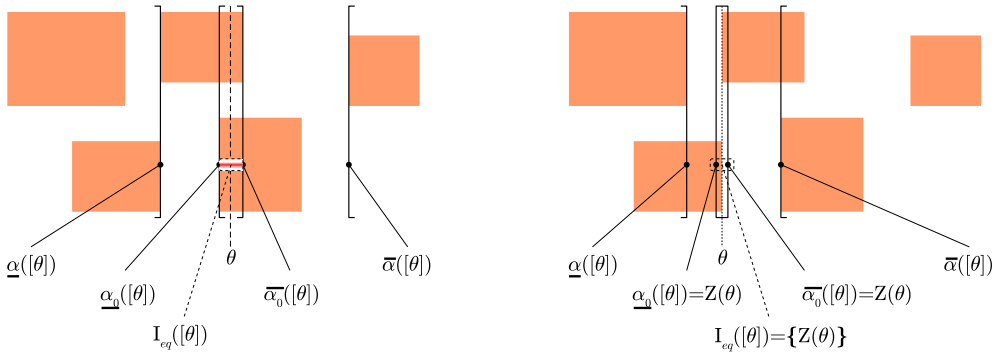
Démonstration. Soient $\mathcal{E} \in \mathcal{E}^*$ et (θ, θ') une paire de parax-hyperplans appartenant à $\mathcal{E}$. Le cas où $\mathrm{Z}(\theta) = \mathrm{Z}(\theta')$ est trivial. Si $\mathrm{Z}(\theta) \neq \mathrm{Z}(\theta')$, puisque $S(\theta) = S(\theta')$ il vient que $\mathcal{I}_-(\theta) = \mathcal{I}_-(\theta')$ et $\mathcal{I}_+(\theta) = \mathcal{I}_+(\theta')$. Aussi $\mathcal{I}_0(\theta) = \{1, \ldots, s\} \setminus (\mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta)) = \{1, \ldots, s\} \setminus (\mathcal{I}_-(\theta') \cup \mathcal{I}_+(\theta')) = \mathcal{I}_0(\theta')$. ■

Soit $\mathcal{E} \in \mathcal{E}^*$. Dans ce qui suit, nous caractérisons les valeurs des niveaux zéro des parax-hyperplans appartenant à une même classe d'équivalence $\mathcal{E}$. Posons pour cela les fonctions :

$$\begin{aligned}
\underline{\alpha}(\mathcal{E}) &= \max_{p \in \mathcal{I}_-(\mathcal{E})} \max_{x \in \mathcal{F}_p} x_{dir(\mathcal{E})};\\
\overline{\alpha}(\mathcal{E}) &= \min_{q \in \mathcal{I}_+(\mathcal{E})} \min_{x \in \mathcal{F}_q} x_{dir(\mathcal{E})};\\
\underline{\alpha_0}(\mathcal{E}) &= \max_{j \in \mathcal{I}_0(\mathcal{E})} \min_{x \in \mathcal{F}_j} x_{dir(\mathcal{E})}, \quad \text{si } \mathcal{I}_0(\mathcal{E}) \neq \emptyset;\\
\overline{\alpha_0}(\mathcal{E}) &= \min_{j \in \mathcal{I}_0(\mathcal{E})} \max_{x \in \mathcal{F}_j} x_{dir(\mathcal{E})}, \quad \text{si } \mathcal{I}_0(\mathcal{E}) \neq \emptyset.
\end{aligned}$$

Il faut ici souligner que, dans la mesure où nous présupposons que $S(\mathcal{E}) \neq \emptyset$, les fonctions $\underline{\alpha}$ and $\overline{\alpha}$ sont toujours définies sur $\mathcal{E}^*$. Cependant, $\underline{\alpha_0}(\mathcal{E})$ et $\overline{\alpha_0}(\mathcal{E})$ n'ont un sens que si $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$.

La Figure 5.2 illustre ces fonctions pour une configuration choisie arbitrairement d'ensembles de données. Étant donné un parax-hyperplan θ (représenté par un trait en pointillé), on retrouve les fonctions $\underline{\alpha}([\theta])$, $\overline{\alpha}([\theta])$, $\underline{\alpha_0}([\theta])$ et $\overline{\alpha_0}([\theta])$.



(a) (b)

FIGURE 5.2 – Les rectangles saumons représentent les plus petits rectangles qui contiennent tous les points des ensembles de données choisis arbitrairement (c'est-à-dire qu'il existe au moins un point appartenant à chaque côté de rectangle). Deux parax-hyperplans sont ici envisagés. L'intervalle $\mathrm{I}_{eq}([\theta])$ contenant les zéros des hyperplans de la classe d'équivalence $[\theta]$ est représenté au moyen d'un rectangle en pointillé. (a) $\mathrm{I}_{eq}([\theta])$ est un intervalle fermé. (b) $\mathrm{I}_{eq}([\theta])$ est un point.

Du fait de la Propriété énoncée en (5.4), les fonctions $\underline{\alpha}(\mathcal{E})$ et $\overline{\alpha}(\mathcal{E})$ vérifient par construction :

$$\begin{aligned}
&\forall \theta \in \mathcal{E}, \mathrm{Z}(\theta) > \underline{\alpha}(\mathcal{E}),\\
&\forall \theta \in \mathcal{E}, \mathrm{Z}(\theta) < \overline{\alpha}(\mathcal{E}).
\end{aligned} \tag{5.6}$$

Il découle de cette Propriété (5.6) que $\forall\theta \in \mathcal{E}$, $\underline{\alpha}(\mathcal{E}) < \mathrm{Z}(\theta) < \overline{\alpha}(\mathcal{E})$, et donc que $\underline{\alpha}(\mathcal{E}) < \overline{\alpha}(\mathcal{E})$. Dans le cas où $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$, on obtient encore $\forall\theta \in \mathcal{E}$, $\forall j \in \mathcal{I}_0(\mathcal{E})$, $\min_{x\in\mathcal{F}_j} x_{dir(\mathcal{E})} \leqslant \mathrm{Z}(\theta) \leqslant \max_{x\in\mathcal{F}_j} x_{dir(\mathcal{E})}$, d'où $\forall(j,j') \in \mathcal{I}_0(\mathcal{E})^2$, $\min_{x\in\mathcal{F}_j} x_{dir(\mathcal{E})} \leqslant \mathrm{Z}(\theta) \leqslant \max_{x\in\mathcal{F}_{j'}} x_{dir(\mathcal{E})}$. Par conséquent, on obtient l'inégalité $\underline{\alpha_0}(\mathcal{E}) \leqslant \overline{\alpha_0}(\mathcal{E})$.

Étant données les inégalités précédentes, il est possible de décrire l'intervalle $\mathrm{I}_{eq}(\mathcal{E})$ couvert par les zéros des hyperplans de la classe d'équivalence $\mathcal{E} \in \mathcal{E}^*$:

$$\mathrm{I}_{eq}(\mathcal{E}) \widehat{=} \begin{cases}]\underline{\alpha}(\mathcal{E}), \overline{\alpha}(\mathcal{E})[\,, \text{ si } \mathcal{I}_0(\mathcal{E}) = \emptyset, \\ ]\underline{\alpha}(\mathcal{E}), \overline{\alpha}(\mathcal{E})[\cap [\underline{\alpha_0}(\mathcal{E}), \overline{\alpha_0}(\mathcal{E})], \text{ autrement.} \end{cases} \tag{5.7}$$

Nous appellerons la fonction I_{eq} définie sur $\mathcal{E}^*$ *intervalle d'équivalence.*

La Figure 5.2 fournit une représentation graphique de cet intervalle s'équivalence $\mathrm{I}_{eq}(\mathcal{E})$. A partir de la Définition (5.7), il est aisé de déduire que $\mathrm{I}_{eq}(\mathcal{E})$ peut être selon le cas soit un intervalle ouvert, soit semi-ouvert, soit fermé (tel que dans l'exemple donné Figure 5.2(a)), soit même un simple point (tel que dans l'exemple donné Figure 5.2(b)).

La propriété suivante permet de caractériser une classe d'équivalence donnée au moyen de l'intervalle d'équivalence I_{eq}.

Propriété 4. *Pour tout $\mathcal{E} \in \mathcal{E}^*$:*

$$\mathcal{E} = \{\theta : dir(\theta) = dir(\mathcal{E}) \text{ et } \mathrm{Z}(\theta) \in \mathrm{I}_{eq}(\mathcal{E})\} \tag{5.8}$$

Démonstration. Pour $\mathcal{E} \in \mathcal{E}^*$ et $\theta \in \Theta$ tels que $dir(\theta) = dir(\mathcal{E})$, nous devons prouver que $\theta \in \mathcal{E}$ si et seulement si $\mathrm{Z}(\theta) \in \mathrm{I}_{eq}(\mathcal{E})$.

($\Rightarrow$) Soit $\theta \in \mathcal{E}$, il vient que $\mathrm{Z}(\theta) \in]\underline{\alpha}(\mathcal{E}), \overline{\alpha}(\mathcal{E})[$ et, si cela a du sens, $\mathrm{Z}(\theta) \in \left[\underline{\alpha_0}(\mathcal{E}), \overline{\alpha_0}(\mathcal{E})\right]$. Par conséquent $\mathrm{Z}(\theta) \in \mathrm{I}_{eq}(\mathcal{E})$, ce qui prouve encore que $\mathrm{I}_{eq}(\mathcal{E}) \neq \emptyset$ pour tout $\mathcal{E} \in \mathcal{E}^*$.

($\Leftarrow$) Soit $\theta \in \Theta$ tel que $dir(\theta) = dir(\mathcal{E})$ et $\mathrm{Z}(\theta) \in \mathrm{I}_{eq}(\mathcal{E})$. Puisque $\mathrm{Z}(\theta) \in]\underline{\alpha}(\mathcal{E}), \overline{\alpha}(\mathcal{E})[$, il apparaît que $\forall p \in \mathcal{I}_-(\mathcal{E}), \forall x \in \mathcal{F}_p, x_{dir(\mathcal{E})} < \mathrm{Z}(\theta)$ et que $\forall q \in \mathcal{I}_+(\mathcal{E}), \forall x \in \mathcal{F}_q, x_{dir(\mathcal{E})} > \mathrm{Z}(\theta)$, si bien que $\mathcal{I}_-(\mathcal{E}) \subseteq \mathcal{I}_-(\theta)$ et $\mathcal{I}_+(\mathcal{E}) \subseteq \mathcal{I}_+(\theta)$. Cela implique que $S(\mathcal{E}) \subseteq S(\theta)$. Nous distinguons maintenant le cas où $\mathcal{I}_0(\mathcal{E}) = \emptyset$ de celui où $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$.

Si $\mathcal{I}_0(\mathcal{E}) = \emptyset$, du fait de (5.3), il vient que $\mathcal{I}_-(\mathcal{E}) \cup \mathcal{I}_+(\mathcal{E}) = \{1, \ldots, s\} \subseteq \mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta)$. Comme on a les inclusions $\{1, \ldots, s\} \subseteq \mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta) \subseteq \{1, \ldots, s\}$, nous pouvons déduire que $\mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta) = \{1, \ldots, s\}$ ce qui implique que $\mathcal{I}_0(\theta) = \emptyset = \mathcal{I}_0(\mathcal{E})$. D'où $\mathcal{I}_-(\mathcal{E}) = \mathcal{I}_-(\theta)$, $\mathcal{I}_+(\mathcal{E}) = \mathcal{I}_+(\theta)$, et donc $S(\mathcal{E}) = S(\theta)$.

Si $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$, sachant que $\mathrm{Z}(\theta) \in [\underline{\alpha_0}(\mathcal{E}), \overline{\alpha_0}(\mathcal{E})]$, nous déduisons que, par construction, tout élément de $\mathcal{I}_0(\mathcal{E})$ est aussi élément de $\mathcal{I}_0(\theta)$, c'est-à-dire $\mathcal{I}_0(\mathcal{E}) \subseteq \mathcal{I}_0(\theta)$. Or, du fait de (5.3), nous savons encore que $\mathcal{I}_-(\mathcal{E}) \cup \mathcal{I}_+(\mathcal{E}) \subseteq \mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta)$, d'où $\mathcal{I}_0(\mathcal{E}) \supseteq \mathcal{I}_0(\theta)$. Comme cela signifie donc que $\mathcal{I}_0(\mathcal{E}) = \mathcal{I}_0(\theta)$, nous arrivons de nouveau à la conclusion $S(\theta) = S(\mathcal{E})$.

De sorte que dans tous les cas, si $\mathrm{Z}(\theta) \in \mathrm{I}_{eq}(\mathcal{E})$, alors $S(\theta) = S(\mathcal{E})$, ce qui revient à dire $\mathcal{E} = [\theta]$. ■

Exemple 15. *Sur la Figure 5.1(b), les lignes en pointillés correspondent aux extrémités des intervalles associés aux différentes classes d'équivalence. Par exemple, $\underline{\alpha}([\theta_{(1),1}]) = a_1$, $\overline{\alpha}([\theta_{(1),1}]) = a_2$, et $\mathcal{I}_0(\theta_{(1),1}) = \emptyset$, de telle sorte que $\mathrm{I}_{eq}([\theta_{(1),1}]) =]a_1, a_2[$. De manière semblable, il apparaît que $\mathrm{I}_{eq}([\theta_{(3),1}]) =]a_3, a_4[$. Un autre exemple est donné pour $\theta_{(2),1}$: $\underline{\alpha}([\theta_{(2),1}]) = a_1$, $\overline{\alpha}([\theta_{(2),1}]) = a_4$, et, comme $\mathcal{I}_0(\theta_{(2),1}) = \{3\}$, on obtient que $\underline{\alpha_0}([\theta_{(2),1}]) = a_2$ et que $\overline{\alpha_0}([\theta_{(2),1}]) = a_3$. Il en découle que $\mathrm{I}_{eq}([\theta_{(2),1}]) = [a_2, a_3]$. Nous venons d'expliciter l'ensemble des classes d'équivalence de la première direction.*

Relevons sans en dire pour l'instant davantage que tous ces intervalles d'équivalence sont concomitants : cela donne $]a_1, a_2[$, puis $[a_2, a_3]$ et enfin $]a_3, a_4[$. Cette remarque sera

exploitée dans la partie algorithmique 6.2.1 pour générer de façon efficace l'ensemble des classes d'équivalence.

La prochaine Propriété justifie que le nombre de classes d'équivalence est fini.

Propriété 5. *La cardinalité de $\mathcal{E}^*$ est finie.*

Démonstration. Puisque S est invariant par $\sim$, on peut l'envisager comme une fonction $S : \mathcal{E}^* \to \mathcal{P}(U)$ où $\mathcal{P}(U)$ dénote l'ensemble des parties de $U = \{(p,q) \in \{1,\ldots,s\}^2 : p < q\}$. Pour deux classes d'équivalence $\mathcal{E}, \mathcal{E}' \in \mathcal{E}^*$ ayant une même direction, $\mathcal{E} \neq \mathcal{E}'$ implique que $S(\mathcal{E}) \neq S(\mathcal{E}')$ (voir la Définition 4). Par suite, le nombre des classes d'équivalence ayant la même direction ne peut pas être plus grand que $|\mathcal{P}(U)|$, qui est borné : $|\mathcal{P}(U)| = 2^{\frac{s(s-1)}{2}}$. Puisque le nombre des directions possibles est n, on obtient la borne supérieure $|\mathcal{E}^*| \leq n|\mathcal{P}(U)|$, ce qui suffit à prouver la finitude de $\mathcal{E}^*$. ■

5.2 Césures

Une classe d'équivalence $\mathcal{E} \in \mathcal{E}^*$ contient l'ensemble des parax-hyperplans qui séparent les mêmes ensembles de données (qui correspondent à des modes dynamiques distincts). Nous pourrions donc raisonner directement sur les classes d'équivalence. Cependant, nous travaillerons avec le parax-hyperplan de chaque classe qui est optimal au sens statistique [120] : il s'agit de celui qui maximise la marge. Cet hyperplan sera nommé *césure*.

Définition 5 (Césure)**.** *Soient $\mathcal{E} \in \mathcal{E}^*$ et $i = dir(\mathcal{E})$. La* césure *associée à $\mathcal{E}$ est le parax-hyperplan $\theta \in \mathcal{E}$ tel que* $\mathrm{Z}(\theta)$ *soit le milieu de* $\mathrm{I}_{eq}(\mathcal{E})$.

Comme $\mathrm{I}_{eq}(\mathcal{E}) \neq \emptyset$ quelque soit $\mathcal{E} \in \mathcal{E}^*$, il existe un isomorphisme entre les césures et les classes d'équivalence.
Nous écrivons $\mathcal{C}^*$ l'ensemble de toutes les césures. Comme $\mathcal{E}^*$ et $\mathcal{C}^*$ sont isomorphes, la cardinalité de $\mathcal{C}^*$ est aussi bornée.

Exemple 16. *Pour les trois ensembles de données de la Figure 5.1(a), $\mathcal{C}^*$ est composé de cinq césures : $\theta_{(1),1}$, $\theta_{(2),1}$, $\theta_{(3),1}$, $\theta_{(1),2}$, et $\theta_{(2),2}$. Elles sont représentées sur la Figure 5.1(c) au moyen de ligne en pointillés.*

En reprenant le cas de cet exemple, nous aurions tendance à trouver intuitivement que la césure $\theta_{(1),1}$ est plus puissante que $\theta_{(2),1}$, dans la mesure où la première sépare aussi bien $\mathcal{F}_1$ et $\mathcal{F}_2$ que $\mathcal{F}_1$ et $\mathcal{F}_3$, tandis que la seconde sépare uniquement $\mathcal{F}_1$ et $\mathcal{F}_2$ (c'est-à-dire, $S(\theta_{(1),1}) = \{(1,2),(1,3)\}$ et $S(\theta_{(2),1}) = \{(1,2)\}$).

Cela motive l'introduction d'une relation d'ordre sur $\mathcal{C}^*$, que l'on notera $\preceq$. Les bases théoriques associées à la notion d'ordre sont rappelées dans l'Annexe A.

Définition 6 (Relation d'ordre sur les césures)**.** *Soient $\theta, \theta' \in \mathcal{C}^*$. Nous définissons la relation binaire $\preceq$ sur $\mathcal{C}^*$ par :*

$$\theta \preceq \theta' \text{ si } S(\theta) \subseteq S(\theta') \text{ et } dir(\theta) = dir(\theta'). \tag{5.9}$$

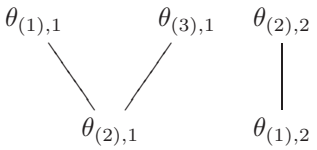
Propriété 6. *La relation $\preceq$ de la Définition 6 est un ordre partiel sur $\mathcal{C}^*$.*

Démonstration. La relation est :

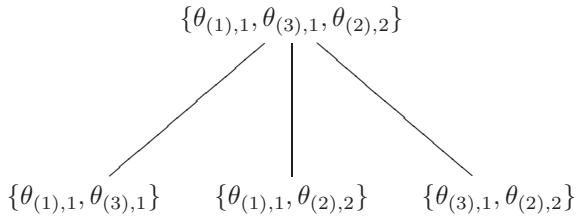
- refléxive : $\forall \theta \in \mathcal{C}^*$, $\theta \preceq \theta$
- antisymétrique : soient θ et $\tilde{\theta}$ tels que $\theta \preceq \tilde{\theta}$ et $\tilde{\theta} \preceq \theta$, cela signifie que $S(\theta) \subseteq S(\tilde{\theta})$ et que $S(\tilde{\theta}) \subseteq S(\theta)$, c'est-à-dire que $S(\theta) = S(\tilde{\theta})$, et comme $dir(\theta) = dir(\tilde{\theta})$, il vient que $\theta = \tilde{\theta}$.
- transitive : soient $(\theta_1, \theta_2, \theta_3) \in \mathcal{C}^{*3}$ tels que $\theta_1 \preceq \theta_2$ et $\theta_2 \preceq \theta_3$, alors $S(\theta_1) \subseteq S(\theta_2)$ et $S(\theta_2) \subseteq S(\theta_3)$, ce qui implique que $S(\theta_1) \subseteq S(\theta_3)$. Or $dir(\theta_1) = dir(\theta_2) = dir(\theta_3)$, donc $\theta_1 \preceq \theta_3$.

L'ordre est de plus évidemment partiel dans la mesure où des césures avec des directions différentes ne sont pas directement comparables. ■

Exemple 17. *Le diagramme de Hasse correspondant à l'exemple de la Figure 5.1 est montré Figure 5.3(a).*



(a)



(b)

FIGURE 5.3 – (a) Diagramme de Hasse de l'ensemble de césures $\mathcal{C}^*$ de la Figure 5.1 ordonné par $\preceq$. Le diagramme montre, par exemple, que $\theta_{(2),1} \preceq \theta_{(1),1}$. (b) Diagramme de Hasse de la fermeture inférieure de $\mathcal{M} = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$, qui est une multicésure des ensembles de données de la Figure 5.1. À noter, $\mathcal{M}$ est égal à $\text{Max}_{\preceq} \mathcal{C}^*$.

Comme tout ensemble partiellement ordonné, $\mathcal{C}^*$ admet des éléments minimaux et des éléments maximaux. L'ensemble des éléments maximaux et celui des éléments minimaux de $\mathcal{C}^*$ sont notés respectivement $\text{Max}_{\preceq} \mathcal{C}^*$ et $\text{Min}_{\preceq} \mathcal{C}^*$.

Exemple 18. *Sur la Figure 5.3(a), $Max_{\preceq} \mathcal{C}^* = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$.*

Comme le suggère la figure de l'Exemple 17, il sera utile de restreindre $\mathcal{C}^*$ à une direction donnée. Nous définissons donc, pour tout $i \in \{1, \ldots, n\}$,

$$\mathcal{C}^{i*} = \{\theta \in \mathcal{C}^* : dir(\theta) = i\}. \tag{5.10}$$

$\mathcal{C}^{i*}$ est un sous-ensemble de $\mathcal{C}^*$ qui hérite de la relation d'ordre partiel définie sur $\mathcal{C}^*$.

Exemple 19. *Pour revenir à notre exemple, la Figure 5.3(a) montre que $Max_{\preceq} \mathcal{C}^{1*} = \{\theta_{(1),1}, \theta_{(3),1}\}$ et que $Max_{\preceq} \mathcal{C}^{2*} = \{\theta_{(2),2}\}$.*

5.3 Ensembles particuliers de césures

5.3.1 Multicésures

De manière générale, plusieurs césures seront nécessaires pour séparer tous les ensembles de données de $\mathcal{F}^*$. Cela nous conduit à introduire la notion de multicésure.

Définition 7 (Multicésure)**.** *Une* multicésure $\mathcal{M}$ *de* $\mathcal{F}^*$ *est un ensemble fini de césures telles que, pour toute paire* $(p,q) \in U$*, il existe une césure* $\theta \in \mathcal{M}$ *telle que* $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$*. La collection* $\mathcal{F}^*$ *est dite* séparable *(on dira aussi* m-séparable *pour éviter toute confusion) s'il existe une multicésure de* $\mathcal{F}^*$*.*

On notera $\mathcal{M}^*$ l'ensemble de toutes les multicésures. Puisque $\mathcal{C}^*$ est une ensemble fini, $\mathcal{M}^*$ l'est aussi (car il est un sous-ensemble de l'ensemble des parties de $\mathcal{C}^*$). Il est à noter que $\mathcal{M}^*$ peut éventuellement être vide, cas dans lequel $\mathcal{F}^*$ n'est pas séparable.

Exemple 20. *Pour l'exemple de la Figure 5.1,* $\mathcal{M} = \{\theta_{(3),1}, \theta_{(1),2}\}$ *est une multicésure car* $S(\theta_{(3),1}) = \{(1,2),(2,3)\}$ *et* $S(\theta_{(1),2}) = \{(1,3)\}$*.*

En ce qui concerne le pouvoir de séparation, nous utiliserons l'extension usuelle de l'image d'un point à l'image d'un ensemble pour définir le pouvoir de séparation d'un ensemble de césures par :

$$\text{soit } \mathcal{C} \text{ un ensemble de césures, } S(\mathcal{C}) = \bigcup_{\theta \in \mathcal{C}} S(\theta). \tag{5.11}$$

L'inverse de la fonction S est défini aussi usuellement comme :

$$S^{-1}(p,q) = \{\theta \in \mathcal{C}^* : (p,q) \in S(\theta)\} \tag{5.12}$$

Les propriétés suivantes seront utiles pour la suite.

Propriété 7. *Un ensemble* $\mathcal{C}$ *de césures est une multicésure si et seulement si l'image de* $\mathcal{C}$ *par* S *est* $U = \{u = (p,q) \in \{1,\dots,s\}^2 : p < q\}$*.*

Démonstration. ($\Rightarrow$) En résonnant par l'absurde, soit $(p,q) \in U$ et $(p,q) \notin \cup_{\theta \in \mathcal{C}} S(\theta)$. $\mathcal{F}^*$ n'est pas séparable, ce qui conduit à la conclusion contradictoire que $\mathcal{C}$ n'est pas une multicésure. Il n'existe donc aucune césure $\theta \in \mathcal{C}$ vérifiant $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$, c'est-à-dire, $(p,q) \in S(\theta)$.
($\Leftarrow$) $\mathcal{C}$ est un ensemble fini de césures tel que $\cup_{\theta \in \mathcal{C}} S(\theta) = U$. Alors, par Définition 7, $\mathcal{C}$ est une multicésure. ■

Propriété 8. $\mathcal{F}^*$ *est séparable si et seulement si* $\mathcal{C}^*$ *est une multicésure.*

Démonstration. ($\Rightarrow$) Soit $(p,q) \in U$. Par hypothèse, $\mathcal{F}_p$ et $\mathcal{F}_q$ sont séparables et il existe donc un parax-hyperplan $\tilde{\theta}$ qui les sépare. Soit θ la césure correspondant à la classe d'équivalence $[\tilde{\theta}]$. Alors $\theta \in \mathcal{C}^*$ et $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$. Dans la mesure où l'on peut répéter cet argument pour l'ensemble des $(p,q) \in U$, il découle que $\mathcal{C}^*$ est une multicésure. ($\Leftarrow$) Par définition. ■

Nous définissons un ordre partiel $(\mathcal{M}^*, \subseteq)$ usuel sur l'ensemble des multicésures de $\mathcal{F}^*$.

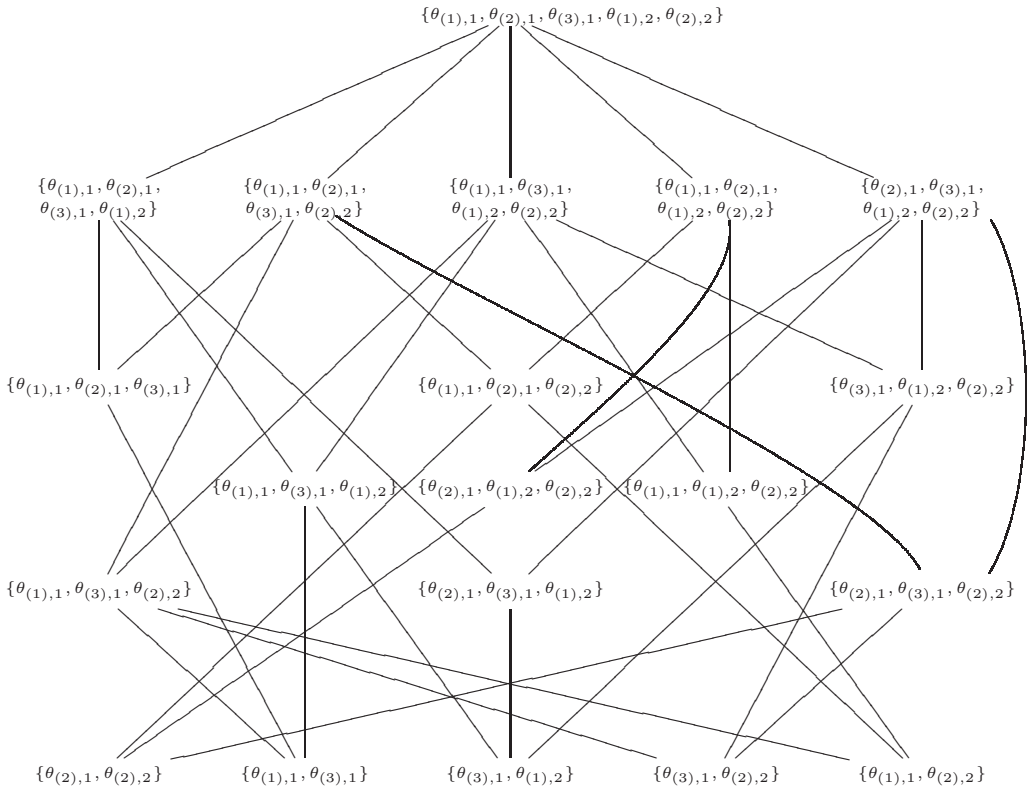


FIGURE 5.4 – Diagramme de Hasse pour l'ensemble partiellement ordonné des multicésures $\mathcal{M}^*$ pour la relation d'inclusion, dans le cas de l'ensemble des données de l'exemple de la Figure 5.1. L'ensemble de toutes les césures $\{\theta_{(1),1}, \theta_{(2),1}, \theta_{(3),1}, \theta_{(1),2}, \theta_{(2),2}\}$ est bien-sûr l'élément maximal.

Exemple 21. *Pour l'exemple de la Figure 5.1, $\mathcal{M}^*$ consiste de vingt multicésures représentées sur le diagramme de la Figure 5.4.*

À n'importe quel sous-ensemble $\mathcal{B}$ de $\mathcal{M}^*$ il est possible d'associer sa fermeture inférieure selon la relation d'inclusion $\subseteq$, qui consiste en l'ensemble des multicésures de $\mathcal{M}^*$ inférieures (c'est-à-dire incluses) à l'une des multicésures de $\mathcal{B}$. Pour des raisons qui apparaîtront claires ultérieurement, nous nous concentrerons sur la fermeture inférieure des singletons $\mathcal{B} = \{\mathcal{M}\}$, avec $\mathcal{M} \in \mathcal{M}^*$.

Définition 8 (Fermeture inférieure d'une multicésure)**.** *La* fermeture inférieure *de $\{\mathcal{M}\}$, $\mathcal{M} \in \mathcal{M}^*$, notée $\downarrow\{\mathcal{M}\}$, est définie par*

$$\downarrow\{\mathcal{M}\} = \{\mathcal{M}' \in \mathcal{M}^* : \mathcal{M}' \subseteq \mathcal{M}\}. \tag{5.13}$$

Dans la mesure où $(\mathcal{O}(\mathcal{M}^*), \subseteq)$ est l'ordre partiel dual de $\mathcal{M}^*$ ($\mathcal{O}(\mathcal{M}^*)$ étant l'ensemble de toutes les fermetures inférieures de $\mathcal{M}^*$ comme indiqué en Annexe A), par induction, $(\downarrow\{\mathcal{M}\}, \subseteq)$ est aussi un ordre partiel.

Exemple 22. *Reprenons pour exemple la multicésure $Max_{\preceq}\mathcal{C}^*$ de la Figure 5.3(a). La fermeture inférieure de $\{Max_{\preceq}\mathcal{C}^*\}$ est $\{\{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}, \{\theta_{(3),1}, \theta_{(2),2}\}, \{\theta_{(1),1}, \theta_{(2),2}\}, \{\theta_{(1),1}, \theta_{(3),1}\}\}$ et est représentée sur la Figure 5.3(b). Le diagramme de Hasse sur la Figure 5.3(b) est en fait une sous-partie du diagramme de la Figure 5.4 : en effet, $Max_{\preceq}\mathcal{C}^*$ est le premier élément à gauche sur la deuxième ligne en partant du bas. Comme il apparaît ici, ne considérer que $Max_{\preceq}\mathcal{C}^*$ (à la place de $\mathcal{C}^*$) réduit foncièrement le nombre de multicésures à étudier.*

Définition 9 (Multicésure parcimonieuse)**.** *Une multicésure $\mathcal{M}$ pour $\mathcal{F}^*$ est dite* parcimonieuse *si*

$$|\mathcal{M}| = \min_{\tilde{\mathcal{M}} \in \downarrow\{Max_{\preceq}\mathcal{C}^*\}} |\tilde{\mathcal{M}}| \tag{5.14}$$

Une multicésure parcimonieuse est une multicésure minimale au sens de la relation $\subseteq$ sur $\mathrm{Max}_{\preceq}\mathcal{C}^*$. L'inverse n'est pas vrai.

5.3.2 Ensemble des césures requises, ensemble des césures candidates

Rappelons tout d'abord que nous nommons $U = \{u = (p, q) \in \{1, \ldots, s\}^2 : p < q\}$.

Définition 10 (Césure requise)**.** *Étant donné un sous-ensemble $\mathcal{C}$ de l'ensemble des césures $\mathcal{C}^*$, nous dirons que la césure $\theta \in \mathcal{C}$ est* requise *si pour $u \in S(\theta)$ quelconque, il n'existe pas une autre césure de $\mathcal{C}$ qui sépare les ensembles de données d'indice dans $u = (p, q)$. Le* sous-ensemble requis *de $\mathcal{C}$ est l'ensemble de toutes les césures requises dans $\mathcal{C}$. On pourra omettre de préciser "dans $\mathcal{C}$" si cela est explicite. On associe la fonction suivante définie sur $\mathcal{C}^*$ au sous-ensemble requis :*

$$Req(\mathcal{C}) = \left\{\theta \in \mathcal{C} : \exists u \in S(\theta), \nexists \theta' \in \mathcal{C} \setminus \{\theta\}, u \in S(\theta')\right\}. \tag{5.15}$$

Si $Req(\mathcal{C}) = \mathcal{C}$, cela signifie que toutes les césures sont requises : il n'existe pas de césure dans $\mathcal{C}$ qui puisse être omise sans perdre la capacité de séparer au moins un couple de classes de $\mathcal{F}^*$. Une propriété intéressante est alors dérivée de la Propriété 7 :

Propriété 9. *Soit $\mathcal{C} \subseteq \mathcal{C}^*$, $\mathcal{C}$ est une multicésure minimale pour l'ordre partiel $(\mathcal{M}^*, \subseteq)$ si et seulement si $Req(\mathcal{C}) = \mathcal{C}$ et $S(\mathcal{C}) = U$.*

Démonstration. D'après la Propriété 7, $\mathcal{C}$ est une multicésure si et seulement si $S(\mathcal{C}) = U$. Il reste à montrer qu'une multicésure est minimale si et seulement si $Req(\mathcal{C}) = \mathcal{C}$.

($\Rightarrow$) $\mathcal{C}$ étant minimale, pour $\theta \in \mathcal{C}$ quelconque, $\mathcal{C} \setminus \{\theta\}$ n'est pas une multicésure : il existe toujours $u \in U$ tel que $u \notin S(\mathcal{C} \setminus \{\theta\})$; un tel élément u n'est donc séparé par aucune césure $\theta' \in \mathcal{C} \setminus \{\theta\}$. Donc θ est requise. Par généralisation, $Req(\mathcal{C}) = \mathcal{C}$.

($\Leftarrow$) Puisque $Req(\mathcal{C}) = \mathcal{C}$, on peut pas avoir un sous-ensemble de $\mathcal{C}$ qui soit une multicésure, d'où la minimalité pour la relation d'inclusion. ■

Définition 11 (Césure superflue, césure candidate)**.** *Étant donné un sous-ensemble $\mathcal{C}$ de l'ensemble des césures $\mathcal{C}^*$ tel que $Req(\mathcal{C}) \subset \mathcal{C}$, une césure θ de $\mathcal{C}$ qui n'est pas requise est appelée :*

– superflue *si $S(\theta) \subset S(Req(\mathcal{C}))$;*

– candidate *si elle n'est pas superflue, c'est-à-dire si $S(\theta) \cap S(Req(\mathcal{C})) \neq S(\theta)$.*

Le sous-ensemble candidat *de $\mathcal{C}$ est l'ensemble de toutes les césures candidates dans $\mathcal{C} \subseteq \mathcal{C}^*$. Le sous-ensemble candidat a la fonction associée sur $\mathcal{C}^*$:*

$$Can(\mathcal{C}) = \{\theta \in \mathcal{C} : S(\theta) \cap S(Req(\mathcal{C})) \neq S(\theta)\}. \tag{5.16}$$

Il est évidant que chacun de ces types de césures est disjoint des deux autres. Une illustration en est donnée Figure 5.5.

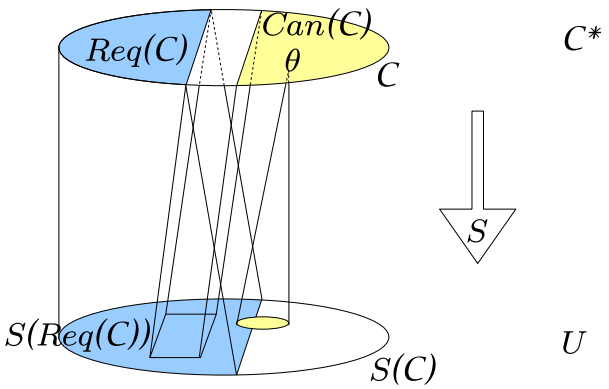


FIGURE 5.5 – Pour un sous-ensemble $\mathcal{C}$ de $\mathcal{C}^*$, il est possible de partitionner les césures en trois ensembles : les césures requises ($Req(\mathcal{C})$, zone bleue en haut), les césures superflues (zone non colorée) et les césures candidates ($Can(\mathcal{C})$, zone jaune en haut). Alors que l'image des césures superflues par S est toujours incluse dans $S(Req(\mathcal{C}))$ (zone bleue en bas), l'image de $\theta \in Can(\mathcal{C})$ quelconque (zone jaune en bas) est telle que $S(\theta) \setminus S(Req(\mathcal{C})) \neq \emptyset$.

Définition 12 (Pouvoir de séparation candidate)**.** *Pour prolonger la notion de pouvoir de séparation d'un ensemble de césures, nous définissons le* pouvoir de séparation candidate

d'une césure θ relativement à un ensemble de césures $\mathcal{C} \in \mathcal{C}^*$ auquel elle appartient par la fonction :

$$S_{\mathcal{C}}(\theta) = S(\theta) \setminus S(Req(\mathcal{C})). \tag{5.17}$$

On notera que cela permet de donner une définition alternative de l'ensemble candidat :

$$Can(\mathcal{C}) = \{\theta \in \mathcal{C} : S_{\mathcal{C}}(\theta) \neq \emptyset\}. \tag{5.18}$$

La fonction inverse de $S_{\mathcal{C}}$ est définie par :

$$S_{\mathcal{C}}{}^{-1}(u) = \{\theta \in \mathcal{C} : u \in S_{\mathcal{C}}(\theta)\}. \tag{5.19}$$

Le domaine de $S_{\mathcal{C}}{}^{-1}$ est bien-sûr $S(\mathcal{C}) \setminus S(Req(\mathcal{C}))$. Il est évident que toute césure de l'image de $S_{\mathcal{C}}{}^{-1}$ appartient à $Can(\mathcal{C})$. Cela se traduit dans la propriété suivante :

Propriété 10. $\forall u \in S(\mathcal{C}), |S_{\mathcal{C}}{}^{-1}(u)| = 0$ *si et seulement si* $u \notin S_{\mathcal{C}}(Can(\mathcal{C}))$.

Démonstration. ($\Rightarrow$) Soit $u \in S(\mathcal{C})$ tel que $|S_{\mathcal{C}}{}^{-1}(u)| = 0$. Pour tout antécédant de u par S dans $\mathcal{C}$, soit $\theta \in \mathcal{C} \cap S^{-1}(u)$ quelconque, on a donc $u \in S(\theta)$ et $u \notin S_{\mathcal{C}}(\theta)$ qui s'écrit encore $u \notin S(\theta) \setminus S(Req(\mathcal{C}))$. De sorte que $u \in S(Req(\mathcal{C}))$, d'où il vient que $u \notin S(\mathcal{C}) \setminus S(Req(\mathcal{C})) = S_{\mathcal{C}}(\mathcal{C}) = S_{\mathcal{C}}(Can(\mathcal{C}))$.

($\Leftarrow$) Soit $u \in S(\mathcal{C})$ tel que $u \notin S_{\mathcal{C}}(Can(\mathcal{C}))$. Comme $S_{\mathcal{C}}(Can(\mathcal{C})) = S_{\mathcal{C}}(\mathcal{C})$, on a $|S_{\mathcal{C}}{}^{-1}(u)| = 0$. ■

La propriété suivante montre que les antécédents par $S_{\mathcal{C}}$ des couples séparés par des césures candidates sont forcément multiples.

Propriété 11. $\forall u \in S_{\mathcal{C}}(Can(\mathcal{C})), |S_{\mathcal{C}}{}^{-1}(u)| > 1$.

Démonstration. Soit $u \in S_{\mathcal{C}}(Can(\mathcal{C}))$: d'après la Propriété 10, $|S_{\mathcal{C}}{}^{-1}(u)| \neq 0$. De plus, raisonnons par l'absurde. Si $|S_{\mathcal{C}}{}^{-1}(u)| = 1$, soit θ la césure telle que $S_{\mathcal{C}}{}^{-1}(u) = \{\theta\}$. Par construction, $u \notin S(Req(\mathcal{C}))$, donc il n'existe pas d'antécédent de u par S dans $\mathcal{C} \setminus Can(\mathcal{C})$. Donc θ est l'unique antécédant de u par S, et cela signifie que θ est une césure requise, ce qui est contradictoire. ■

Nous verrons dans le chapitre suivant l'intérêt de chercher les minima de $(\mathcal{M}^*, \subseteq)$. Pour cela, il vient, d'après la Propriété 9, que nous devons retirer toutes les césures superflues de $\mathcal{C}^*$. La propriété précédente aide à caractériser le fait que de retirer une césure candidate d'un ensemble de césures $\mathcal{C}$ peut conduire des césures candidates restantes à devenir requises, alors que les césures requises restent bien-sûr inchangées. Étant donné une multicésure $\mathcal{M} \in \mathcal{M}^*$, les minima de sa fermeture inférieure sont donc constitués des césures requises pour $\mathcal{M}$, plus des césures candidates pour $\mathcal{M}$. La propriété suivante sert à clarifier cette idée.

Propriété 12. *Étant donné un ensemble $\mathcal{C}$ de césures, pour tout couple de césures candidates $(\theta, \theta') \in \mathcal{C}^2$ tel que ces césures soient les seuls antécédents par $S_{\mathcal{C}}$ de $u \in S(\theta) \cap S(\theta')$, alors $\theta \in Req(\mathcal{C} \setminus \{\theta'\})$ et $\theta' \in Req(\mathcal{C} \setminus \{\theta\})$.*

Démonstration. Soit $u \in S_{\mathcal{C}}(\mathcal{C})$ tel que $S_{\mathcal{C}}{}^{-1}(u) = \{\theta, \theta'\}$. θ est l'unique césure qui sépare u dans l'ensemble $\mathcal{C} \setminus \{\theta'\}$: d'où $\theta \in Req(\mathcal{C} \setminus \{\theta'\})$. De même, θ' est l'unique césure qui sépare u dans l'ensemble $\mathcal{C} \setminus \{\theta\}$: d'où $\theta \in Req(\mathcal{C} \setminus \{\theta\})$. ■

Exemple 23. *Reprenons pour exemple la Figure 5.1(c). Soit* $\mathcal{C} = \mathcal{C}^* = \{\theta_{(1),1}, \theta_{(2),1}, \theta_{(3),1}, \theta_{(1),2}, \theta_{(2),2}\}$. $Req(\mathcal{C}) = \emptyset$ *car toute césure peut être omise sans que l'on ne perde la séparabilité des ensembles de données. Toutes les césures sont donc des césures candidates :* $Can(\mathcal{C}) = \mathcal{C}$.

Alternativement, soit $\mathcal{C} = \{\theta_{(1),1}, \theta_{(2),1}, \theta_{(3),1}\}$ *: on a maintenant* $\theta_{(1),1}$ *qui est une césure requise car elle est la seule à séparer l'ensemble* $\mathcal{F}_1$ *de* $\mathcal{F}_3$*, et* $\theta_{(3),1}$ *qui est de même une césure requise car elle est la seule à séparer l'ensemble* $\mathcal{F}_2$ *de* $\mathcal{F}_3$*. De sorte que* $Req(\mathcal{C}) = \{\theta_{(1),1}, \theta_{(3),1}\}$ *et* $S(Req(\mathcal{C})) = \{(1,2),(1,3),(2,3)\}$*. Vu que* $S(\theta_{(2),1}) = \{(1,2)\}$*,* $S(\theta_{(2),1}) \subset S(Req(\mathcal{C}))$ *implique que* $\theta_{(2),1}$ *est une césure superflue. Elle peut être omise. L'ensemble* $\mathcal{C}' = \{\theta_{(1),1}, \theta_{(3),1}\}$ *est tel que* $S(\mathcal{C}') = U$*, donc c'est une multicésure ; de plus,* $Req(\mathcal{C}') = \mathcal{C}'$*, donc c'est une multicésure minimale pour la relation* $\subseteq$ *sur les multicésures, comme on le vérifie sur le diagramme de la Figure 5.4.*

Envisageons maintenant le cas où $\mathcal{C} = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$*. De nouveau,* $Req(\mathcal{C}) = \emptyset$*. Cependant, pour tout* $\theta \in \mathcal{C}$*,* $Req(\mathcal{C} \setminus \{\theta\}) = \mathcal{C} \setminus \{\theta\}$ *et de plus* $S(\mathcal{C} \setminus \{\theta\}) = U$ *: ce sont donc des multicésures minimales, ce qui est corroboré par la Figure 5.3(b). Enfin, puisque* $Req(\mathcal{C}) = \emptyset$*,* S *et* $S_{\mathcal{C}}$ *donneront le même résultat, ce qui fait que l'on vérifiera aisément que tout* u *de* $U = \{(1,2),(1,3),(2,3)\}$ *admet exactement deux antécédents par* $S_{\mathcal{C}}$.

Chapitre

6 Césures, aspects algorithmiques

Ce chapitre reprend pleinement les notations du chapitre précédent. Nous allons maintenant détailler comment résoudre le problème introduit à la Section 4.4, qui concerne l'énumération des solutions parcimonieuses. La section 6.1 nous permettra de mieux formaliser ce problème. La résolution sera détaillée dans la section 6.2.

6.1 Formulation du problème de reconstruction des seuils

L'introduction des concepts de césures et de multicésures, ainsi que les ordres partiels définis sur eux, nous permet de formuler le problème de reconstruction des seuils de manière plus précise. Chaque césure θ correspond à un seuil de transition $\mathrm{Z}(\theta)$ pour les variables de concentration ayant pour index $dir(\theta)$. Quand la concentration passe ce seuil, les paramètres de la dynamique du système affine par morceaux peuvent changer, correspondant au passage d'un mode de régulation à un autre. Par extension, une multicésure correspond à un ensemble de seuils de transition qui sont suffisants pour expliquer tous les modes de régulation.

En général, les données d'observation sont consistantes avec un grand nombre de multicésures, et donc un grand nombre de modèles affines par morceaux de réseau de régulation génique. A priori, il n'y a aucune raison de préférer l'un de ces modèles plutôt qu'un autre. Cependant, en pratique, nous sommes davantage intéressés par les modèles qui comprennent un nombre minimal de seuils et qui séparent toutes les paires d'ensemble de données de $\mathcal{F}^*$. En faisant l'hypothèse que les ensemble de données sont séparables, $\mathcal{C}^*$ est une multicésure (Propriété 8, page 80). Il existe en général des multicésures inférieures au sens de la relation d'ordre $\subseteq$, et ce sont donc les multicésures minimales de la fermeture inférieure de $\mathcal{C}^*$ pour $\subseteq$ qui nous intéressent.

Cependant, $\mathcal{C}^*$ peut contenir nombre de césures ayant un faible pouvoir de séparation, qui peuvent être éliminées sous certaines conditions. En l'occurence, nous pouvons ignorer $\theta \in \mathcal{C}^*$ s'il existe une autre césure $\theta' \in \mathcal{C}^*$ non équivalente à θ telle que $\theta \preceq \theta'$. Le fait d'éliminer ces césures n'affectent pas la séparabilité de $\mathcal{F}^*$, comme indiqué dans la propriété suivante qui doit être comparée avec la Propriété 8.

Propriété 13. *$\mathcal{F}^*$ est séparable si et seulement si $Max_{\preceq}\,\mathcal{C}^*$ est une multicésure.*

Démonstration. ($\Rightarrow$) Puisque $\mathcal{F}^*$ est séparable, d'après la Propriété 8, $\mathcal{C}^*$ est une multicésure. Pour toute césure $\theta \in \mathcal{C}^*$, il existe une césure $\tilde{\theta} \in \mathrm{Max}_{\preceq}\,\mathcal{C}^*$ telle que $S(\theta) \subseteq S(\tilde{\theta})$. Par conséquent, toute paire d'ensembles de données est séparée par au moins une césure de $\mathrm{Max}_{\preceq}\,\mathcal{C}^*$, ce qui montre que $\mathrm{Max}_{\preceq}\,\mathcal{C}^*$ est une multicésure. ($\Leftarrow$) Évident puisque $\mathrm{Max}_{\preceq}\,\mathcal{C}^* \subseteq \mathcal{C}^*$. ■

Cette réduction nous permet de porter notre attention sur les multicésures minimales de la fermeture inférieure de $\mathrm{Max}_{\preceq}\,\mathcal{C}^*$.

Exemple 24. *En reprenant l'exemple de la Figure 5.1 (p.74), $Max_{\preceq}\,\mathcal{C}^*$ se composent de trois césures, comme il apparaît Figure 5.3(a) (p.79). Nous pouvons ignorer les deux césures ayant un pouvoir de séparation plus faible ($\theta_{(2),1}$ et $\theta_{(1),2}$). La fermeture inférieure*

de l'ensemble constitué par l'ensemble des césures maximales $\{Max_{\preceq}\mathcal{C}^*\}$ *est montrée Figure 5.3(b). Il y a trois éléments minimaux au sens de* $\subseteq$ *:* $\{\theta_{(1),1}, \theta_{(3),1}\}$, $\{\theta_{(1),1}, \theta_{(2),2}\}$, *et* $\{\theta_{(3),1}, \theta_{(2),2}\}$.

Cependant, comme évoqué au chapitre 4.4, pour appliquer le principe de parcimonie, nous voulons les multicésures minimales en terme de cardinalité. Or les éléments minimaux pour l'ordre partiel lié à la relation $\subseteq$ ne sont pas nécessairement minimaux en terme de cardinalité. Pour justifier cela, considérons l'exemple suivant.

Exemple 25. *Prenons le cas de* $\mathcal{F}^*$ *tel que montré sur la Figure 6.1(a) : les deux multicésures* $\mathcal{M}_1$ *et* $\mathcal{M}_2$ *(qui apparaissent sur les Figures 6.1(b) et 6.1(c), respectivement) appartiennent à* $Min_{\subseteq} \downarrow \{Max_{\preceq}\mathcal{C}^*\}$. *Or, puisque* $|\mathcal{M}_1| = 3$, $|\mathcal{M}_2| = 2$ *et comme il n'y a pas de singleton qui soit une multicésure de* $\mathcal{F}^*$, $\mathcal{M}_2$ *est la seule multicésure de cardinalité minimale.*

Au final, nous devons donc obtenir l'ensemble des multicésures parcimonieuses (Définition 9).

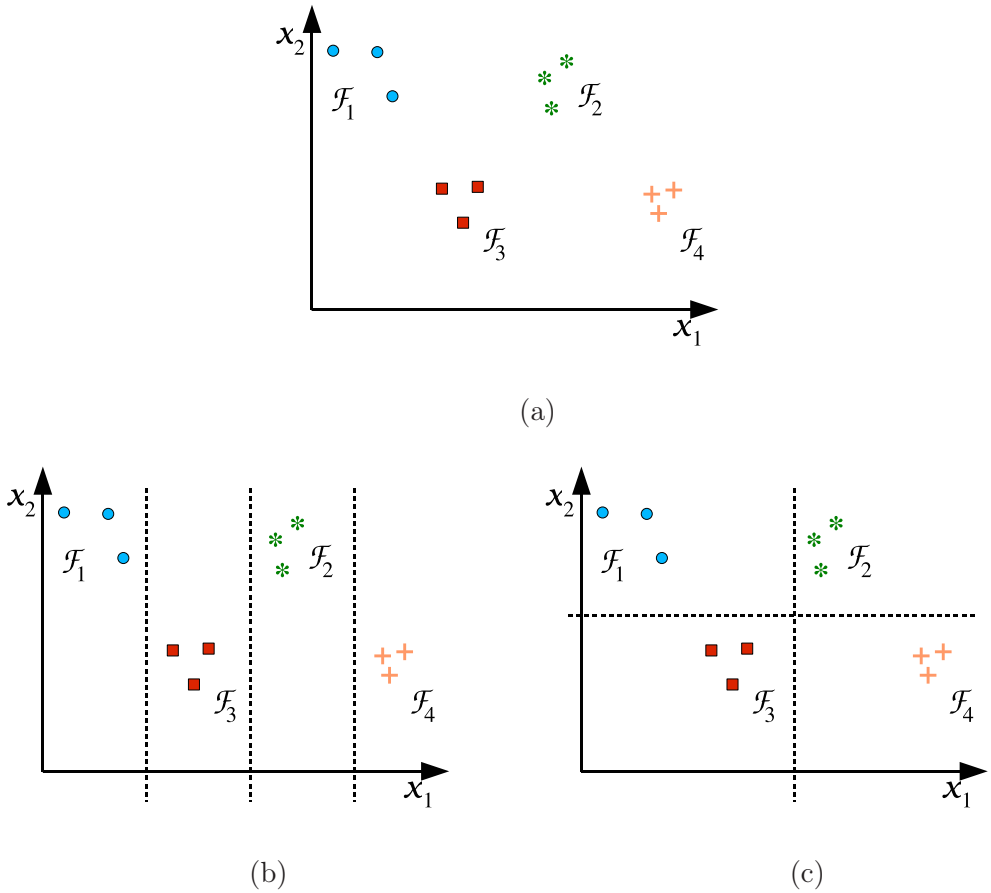


FIGURE 6.1 – Deux multicésures minimales au sens de $\subseteq$ pour $\mathcal{F}^*$. (a) Ensembles de données : $\mathcal{F}^* = \{\mathcal{F}_1, \ldots, \mathcal{F}_4\}$. (b) Multicésure $\mathcal{M}_1$: $|\mathcal{M}_1| = 3$. (c) Multicésure $\mathcal{M}_2$: $|\mathcal{M}_2| = 2$.

6.2 Algorithmes pour reconstruire les solutions parcimonieuses

Nous proposons maintenant une approche permettant de reconstruire l'ensemble des multicésures parcimonieuses pour des ensembles de données classées $\mathcal{F}^*$, d'où nous pourrons déduire l'ensemble des réseaux de régulation génique les plus simples consistants avec ces données.

En particulier, nous introduisons des algorithmes pour générer l'ensemble $\mathcal{C}^*$ des césures (§ 6.2.1), puis permettant d'en extraire les césures maximales au sens de $\preceq$ (§ 6.2.2), puis les multicésures parcimonieuses qui en découlent (§ 6.2.3).

6.2.1 Générer les césures

L'énumération de l'ensemble de toutes les césures est basée sur la Définition 5 et nécessite l'énumération de toutes les classes d'équivalence. Avant de décrire un algorithme permettant de réaliser cette tâche, examinons un exemple basé sur la Figure 5.1(b).

Exemple 26. *Considérons l'ensemble $\mathcal{F}^*$ tel que donné sur la Figure 5.1(a) et concentrons nous sur la première direction. Définissons $a_0 = \min_{x\in\mathcal{F}_1} x_1$ et $a_5 = \max_{x\in\mathcal{F}_2} x_1$. La liste ordonnée $L_1 : a_0 \leq a_1 \leq \ldots \leq a_5$ peut être simplement obtenue en calculant les quantités $\max_{x\in\mathcal{F}_j} x_1$ et $\min_{x\in\mathcal{F}_j} x_1$ pour tous les ensembles $\mathcal{F}_j, j \in \{1,2,3\}$. En utilisant les résultats de l'Exemple 15 (p.77), les intervalles $\mathrm{I}_{eq}(\mathcal{E})$ associés aux classes d'équivalence peuvent être obtenus en retirant les points extrêmes a_0 et a_5 de L_1 et en construisant les intervalles définis par des éléments consécutifs dans L_1 (i.e. $]a_1, a_2[$, $[a_2, a_3]$ et $]a_3, a_4[$). Les césures associées à chaque classe d'équivalence sont calculées à partir du milieu de chacun des intervalles et sont représentés sur la Figure 5.1(c). Soulignons que cette procédure produit* toutes *les césures pour la première direction. De plus, dans la mesure où nous sommes intéressés par la génération des césures, il n'est pas nécessaire de spécifier si les intervalles $\mathrm{I}_{eq}(\mathcal{E})$ incluent ou n'incluent pas les points extrêmes.*

Cette méthode peut être généralisée à des classes pour lesquelles l'intervalle d'équivalence ne consiste qu'en un seul point, comme c'est le cas dans l'exemple de la Figure 5.2(b). La procédure est résumée dans l'Algorithme 1.

Algorithme 1. *Génère l'ensemble de toutes les césures $\mathcal{C}^*$ de $\mathcal{F}^*$*

Soit $\mathcal{C}^* = \emptyset$. Initialisons les listes $m_i = \emptyset$ et $M_i = \emptyset$, pour tout $i = 1, \ldots, n$,
pour** $i = 1, \ldots, n$ **faire
  pour** $j = 1, \ldots, s$ **faire
    $m_i = m_i \cup \{\min_{x\in\mathcal{F}_j} x_i\}$,
    $M_i = M_i \cup \{\max_{x\in\mathcal{F}_j} x_i\}$.
  fin pour
  pour** $j = 1, \ldots, s$ **faire
    si** $m_i(j) < \min(M_i)$ **alors
      $m_i = m_i \setminus \{m_i(j)\}$.
    fin si
    si** $M_i(j) > \max(m_i)$ **alors
      $M_i = M_i \setminus \{M_i(j)\}$.
    fin si
  fin pour
  $L_i = m_i \cup M_i$.
  Trie les éléments de L_i par ordre croissant.

17: ***pour*** $k = 1, \ldots, |L_i| - 1$ ***faire***
18: *Soit θ le parax-hyperplan tel que $dir(\theta) = i$ et $\mathrm{Z}(\theta) = L_i(k) + \frac{L_i(k+1) - L_i(k)}{2}$. Ajoute la césure θ à $\mathcal{C}^*$.*
19: ***si*** $k > 1$ *et* $L_i(k) \in m_i$ *et* $L_i(k) \in M_i$ ***alors***
20: *Soit θ le parax-hyperplan tel que $dir(\theta) = i$ et $\mathrm{Z}(\theta) = L_i(k)$. Ajoute la césure θ à $\mathcal{C}^*$.*
21: ***fin si***
22: ***fin pour***
23: ***fin pour***

Justification

Tâchons maintenant de prouver l'exactitude de cet algorithme. Il est suffisant, pour cela, de démontrer que les lignes 3 à 22 permettent de générer toutes les césures pour une direction i donnée. En effet, si cela est vérifié, la boucle extérieure commençant à la ligne 2 permet ensuite de trouver toutes les césures dans toutes les directions.

Pour simplifier la notation, nous définissons, pour tout $j \in \{1, \ldots, s\}$:

$$\begin{aligned} m_{ij} &= \min_{x \in \mathcal{F}_j} x_i, \\ M_{ij} &= \max_{x \in \mathcal{F}_j} x_i, \\ \tilde{m}_i &= \{m_{ij} : m_{ij} < \min(\{M_{ij}\}_{j=1}^s)\}, \\ \tilde{M}_i &= \{M_{ij} : M_{ij} > \max(\{m_{ij}\}_{j=1}^s)\}. \end{aligned}$$

Pour clarifier la démonstration, nous la diviserons en plusieurs affirmations. Le premier concerne une propriété critique pour l'ensemble L_i généré ligne 3 à 15 de l'algorithme.

Affirmation 1 : Si $\exists \mathcal{E} \in \mathcal{E}^*$, $\mathrm{dir}(\mathcal{E}) = i$, alors les extrémités de $\mathrm{I}_{eq}(\mathcal{E})$ sont des éléments de $L_i = m_i \cup M_i$, où $m_i = \{m_{ij}\}_{j=1}^s \setminus \tilde{m}_i$ et $M_i = \{M_{ij}\}_{j=1}^s \setminus \tilde{M}_i$.
Preuve de l'affirmation 1. Puisque $\mathcal{E} \in \mathcal{E}^*$ on a $S(\mathcal{E}) \neq \emptyset$ et à partir des définitions de $\underline{\alpha}$ et de $\overline{\alpha}$, il vient que $\exists (p, q) \in U : \underline{\alpha}(\mathcal{E}) = M_{ip}$ et $\overline{\alpha}(\mathcal{E}) = m_{iq}$. De plus, à partir de (5.6) il vient que $M_{ip} < m_{iq}$ ce qui implique que $M_{ip} \notin \tilde{M}_i$ et que $m_{iq} \notin \tilde{m}_i$. Si $\mathcal{I}_0(\mathcal{E}) = \emptyset$, à partir de la Propriété 4 et de (5.7) il vient que $\mathrm{I}_{eq}(\mathcal{E}) =]M_{ip}, m_{iq}[$ et l'affirmation est donc démontrée. Si $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$, à partir des définitions de $\underline{\alpha_0}$ et de $\overline{\alpha_0}$ il vient que $\exists (p', q') \in \mathcal{I}_0(\mathcal{E})^2$ tel que $\underline{\alpha_0}(\mathcal{E}) = m_{ip'}$ et que $\overline{\alpha_0}(\mathcal{E}) = M_{iq'}$. Alors, $\mathrm{I}_{eq}(\mathcal{E}) =]M_{ip}, m_{iq}[\cap \left[m_{ip'}, M_{iq'}\right]$. Soit $\alpha_1,\ \alpha_2,\ \alpha_1 \leq \alpha_2$ désignant les extrémités de $\mathrm{I}_{eq}(\mathcal{E})$. Si $m_{ip'} \leq M_{ip}$, alors $\alpha_1 = M_{ip} \notin \tilde{M}_i$. De plus, si $M_{ip} < m_{ip'}$, on a $m_{ip'} \notin \tilde{m}_i$ et aussi $\alpha_1 = m_{ip'}$. De même, si $m_{iq} \leq M_{iq'}$, alors $\alpha_2 = m_{iq} \notin \tilde{m}_i$ et si $m_{iq} > M_{iq'}$, on a $M_{iq'} \notin \tilde{m}_i$ et encore $\alpha_2 = M_{iq'}$.□

Dans ce qui suit, nous justifions le fait que l'algorithme ne génère pas de césure quand L_i est vide (lignes 18 à 21).

Affirmation 2 : L'ensemble L_i est vide si et seulement s'il n'y a aucune classe d'équivalence dans la direction i.
Preuve de l'affirmation 2. Le fait que L_i soit vide signifie qu'il n'existe pas $(p, q) \in \{1, \ldots s\}^2$ tel que $M_{ip} < m_{iq}$ ou, de manière équivalente, qu'aucune paire d'ensembles $\mathcal{F}_p$ et $\mathcal{F}_q$ ne sont séparables le long de la i-ième direction. Réciproquement, si $\nexists \mathcal{E} \in \mathcal{E}^*$ avec $dir(\mathcal{E}) = i$ on a que $\nexists (p, q) \in \{1, \ldots s\}^2$ tel que $\mathcal{F}_p$ et $\mathcal{F}_q$ soient separables le long de la i-ième direction. Ainsi, $\forall (p, q) \in \{1, \ldots s\}^2$ on a toujours $M_{ip} > m_{iq}$ ce qui implique que $M_{ip} \in \tilde{M}_i$ et que $m_{iq} \in \tilde{m}_i$. Il s'en suit que $m_i = M_i = \emptyset$ et donc que $L_i = M_i \cup m_i = \emptyset$. □

La dernière étape consiste à démontrer l'exactitude des lignes 18 à 21 qui génèrent les césures.

Affirmation 3 : Soit $L_i \neq \emptyset$ trié par ordre croissant. Un parax-hyperplan θ est une césure avec pour direction i si et seulement si les conditions suivantes sont vérifiées :

a. $\mathrm{Z}(\theta) = L_i(k) + \frac{L_i(k+1) - L_i(k)}{2}$

b. $k > 1$, $k < |L_i|$, $L_i(k) \in m_i$, $L_i(k) \in M_i$ *et* $\mathrm{Z}(\theta) = L_i(k)$.

Preuve de l'affirmation 3. Associés à un parax-hyperplan, nous introduisons les ensembles suivants, qui nous serons utiles par la suite :

$$\begin{aligned} M_>(\theta) &= \{M_{ij} \in L_i : M_{ij} > \mathrm{Z}(\theta)\}, \\ M_<(\theta) &= \{M_{ij} \in L_i : M_{ij} < \mathrm{Z}(\theta)\}, \\ m_>(\theta) &= \{m_{ij} \in L_i : m_{ij} > \mathrm{Z}(\theta)\}, \\ m_<(\theta) &= \{m_{ij} \in L_i : m_{ij} < \mathrm{Z}(\theta)\}. \end{aligned}$$

($\Rightarrow$) Prenons pour hypothèse que θ est une césure et soit $\mathcal{E} = [\theta]$. Comme montré dans la preuve de l'Affirmation 1, $\underline{\alpha}(\mathcal{E})$ et $\overline{\alpha}(\mathcal{E})$ sont des éléments de L_i et, de même, $\underline{\alpha_0}(\mathcal{E})$ et $\overline{\alpha_0}(\mathcal{E})$ le sont aussi lorsqu'ils sont définis. Nous devons démontrer que les limites de $I_{eq}(\mathcal{E})$ sont des éléments consécutifs ou égaux de L_i. Soit $(\alpha_1, \alpha_2) \in {L_i}^2$ tel que $[\alpha_1, \alpha_2] = cl(I_{eq}(\mathcal{E}))$ (où cl l'opérateur de fermeture sur un ensemble). Alors, $\forall (p, q) \in \mathcal{I}_-(\mathcal{E}) \times \mathcal{I}_+(\mathcal{E})$, $m_{ip} \leqslant M_{ip} \leqslant \alpha_1 \leqslant \alpha_2 \leqslant m_{iq} \leqslant M_{iq}$. Si $\mathcal{I}_0(\mathcal{E}) = \emptyset$, alors $\mathcal{I}_-(\mathcal{E}) \cup \mathcal{I}_+(\mathcal{E}) = \{1, \ldots, s\}$, de sorte qu'il n'y a pas d'élement de L_i entre α_1 et α_2. Si $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$, alors nous pouvons déduire de l'intersection dans (5.7) que la fermeture de $I_{eq}(\mathcal{E})$ peut aussi être écrite comme $[\alpha_1, \alpha_2] = [\max(\underline{\alpha}(\mathcal{E}), \underline{\alpha_0}(\mathcal{E})), \min(\overline{\alpha}(\mathcal{E}), \overline{\alpha_0}(\mathcal{E}))]$. Par conséquent, nous avons encore que $\forall (j, j') \in \mathcal{I}_0(\mathcal{E})^2$, $m_{ij} \leqslant \alpha_1 \leqslant \alpha_2 \leqslant M_{ij'}$. De nouveau, il n'y a pas d'élément de L_i entre α_1 et α_2.

Pour pouvoir conclure la preuve, nous traiterons séparément le cas où $\mathcal{E}$ est un singleton du cas où il ne l'est pas. Si $\mathcal{E}$ est un singleton, de (5.7) il vient que $\mathcal{I}_0(\mathcal{E})$ est un singleton et donc que $\underline{\alpha_0}(\mathcal{E}) = \overline{\alpha_0}(\mathcal{E})$. Puisque, comme montré dans la preuve de l'Affirmation 1, $\exists (p, q) \in \{1, \ldots s\}^2$ tel que $\underline{\alpha_0}(\mathcal{E}) = m_{ip}$ et $\overline{\alpha_0}(\mathcal{E}) = M_{iq}$, à partir de la Definition 5 il vient que $\mathrm{Z}(\theta) = \underline{\alpha_0}(\mathcal{E}) = \overline{\alpha_0}(\mathcal{E})$ et les conditions $L_i(k) \in m_i$, $L_i(k) \in M_i$ et $\mathrm{Z}(\theta) = L_i(k)$ au point (b) sont remplies pour k quelconque. Raisonnons par l'absurde et faisons l'hypothèse que les conditions sont vérifiées pour $k = 1$. Alors, $M_<(\theta) = \emptyset$ et de plus $\mathcal{I}_-(\theta) = \emptyset$ ce qui nous amène à la conclusion, erronée, que θ n'est pas une césure. De manière analogue, si $k = |L_i|$, alors, $m_>(\theta) = \emptyset$ et de plus $\mathcal{I}_+(\theta) = \emptyset$ ce qui implique encore que θ n'est pas une césure.

Si $\mathcal{E}$ n'est pas un singleton, alors $\alpha_1 < \alpha_2$ et de plus il existe k tel que $(\alpha_1, \alpha_2) = (L_i(k), L_i(k+1))$. Dans ce cas, la formule pour $\mathrm{Z}(\theta)$ dans la condition (a) suit la Définition 5.

($\Leftarrow$) Nous montrerons d'abord que si $L_i = \emptyset$, alors $|L_i| \geq 2$. $L_i \neq \emptyset$ implique qu'il existe au moins une classe d'équivalence $\mathcal{E} \in \mathcal{E}^*$ avec pour direction i (voir l'Affirmation 2). De plus, $\underline{\alpha}(\mathcal{E})$ et $\overline{\alpha}(\mathcal{E})$ doivent être tous les deux des points de L_i et comme $\underline{\alpha}(\mathcal{E}) < \overline{\alpha}(\mathcal{E})$, nous avons montré que $|L_i| \geq 2$.

Prenons maintenant pour hypothèse que la condition (a) est vérifiée. Dans ce cas, posons $\alpha_1 = L_i(k) < L_i(k+1) = \alpha_2$ et montrons que α_1, α_2 sont les limites de l'intervalle $I_{eq}(\mathcal{E})$ pour quelque $\mathcal{E} \in \mathcal{E}^*$ (en effet, cela implique que le parax-hyperplan θ avec $\mathrm{Z}(\theta)$ calculé comme dans la condition (a) est une césure). Nous considérons une paire de parax-hyperplans (θ, θ') avec la direction i et faisons l'hypothèse que $Z(\theta) \in]\alpha_1, \alpha_2[$. Il faut

remarquer que les ensembles $M_<(\theta)$ et $m_>(\theta)$ ne sont pas vides, d'où $\exists(p,q) \in U$ tel que $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$. Si $\mathrm{Z}(\theta') \in]\alpha_1, \alpha_2[$, en utilisant le fait que α_1 et α_2 sont des éléments consécutifs de L_i, il vient

$$\begin{array}{ll} M_<(\theta') = M_<(\theta), & M_>(\theta') = M_>(\theta), \\ m_<(\theta') = m_<(\theta), & m_>(\theta') = m_>(\theta) \end{array} \tag{6.1}$$

La première égalité dans (6.1) implique que $\mathcal{I}_-(\theta') = \mathcal{I}_-(\theta)$, la dernière égalité implique que $\mathcal{I}_+(\theta') = \mathcal{I}_+(\theta)$ et les deux autres égalités conduisent à $\mathcal{I}_0(\theta') = \mathcal{I}_0(\theta)$. De sorte que $\theta \sim \theta'$.

A l'opposé, si $\mathrm{Z}(\theta') \notin [\alpha_1, \alpha_2]$, à partir du fait que α_1 et α_2 sont des éléments consécutifs de L_i nous faisons la déduction que au moins une égalité de (6.1) n'est pas vérifiée, de sorte que $\theta \not\sim \theta'$. Cet argument montre que α_1, α_2 sont les limites de $I_{eq}(\mathcal{E})$.

Dans la dernière partie de la preuve, nous faisons l'hypothèse que la condition (b) est vérifiée. Tout d'abord, puisque $k > 1$ et $k < |L_i|$, il vient que $L_i(k) \in (m_i \cap M_i) \setminus \{\min L_i, \max L_i\}$. Cela justifie le fait que les ensembles $M_<(\theta)$ et $m_>(\theta)$ ne sont pas vides, et que donc $\exists(p,q) \in U$ tel que $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$. Montrons maintenant que $[\theta] = \{\theta\}$. Puisque la condition (b) tient, il existe un nouveau couple $(p,q) \in \{1,\ldots,s\}^2$ tel que $L_i(k) = m_{ip} = M_{iq}$, c'est-à-dire que $p \in \mathcal{I}_0([\theta])$ et que $q \in \mathcal{I}_0([\theta])$. Vérifions qu'il n'existe pas de parax-hyperplan équivalent à θ. Si θ' est un parax-hyperplan avec pour direction $dir(\theta) = i$ et tel que $\min L_i < \mathrm{Z}(\theta') < \mathrm{Z}(\theta)$, alors celui-ci appartient à Θ (parce que $\mathrm{Z}(\theta) > \min L_i$) et il vérifie $p \in \mathcal{I}_+(\theta')$. Cela signifie que $\theta' \not\sim \theta$. A l'opposé, si θ' est un parax-hyperplan avec pour direction $dir(\theta') = i$ et tel que $\max L_i > \mathrm{Z}(\theta') > \mathrm{Z}(\theta)$, alors il appartient à Θ (parce que $\mathrm{Z}(\theta) < \max L_i$) et il vérifie $q \in \mathcal{I}_-(\theta')$. Comme dans le cas précédant, on obtient $\theta' \not\sim \theta$. □

Hyper-rectangles contenant les classes

Comme nous avons pu le constater, le raisonnement utilisé pour déduire l'ensemble des césures se base moins sur les points de chaque classe que sur les plus petits hyper-rectangles les contenant. Les classes d'équivalence sont obtenues en utilisant l'espace entre les faces des hyper-rectangles.

D'un point de vue pratique, le bruit de mesure η_i dans (4.5) - page 53 - engendre une dilatation de ces hyper-rectangles. Cette dilatation dépend du bruit sur les points les plus proches des faces. Dans notre cas, nous avons fait l'hypothèse d'un bruit gaussien, donc la dilatation n'a possiblement aucune limite. Cependant, d'un point de vue statistique, on peut considérer que la plupart de l'influence dilatatrice du bruit est cantonnée dans une distance de 2σ où σ est l'écart-type du bruit.

Cela est important à prendre en compte. En effet, pour deux hyper-rectangles ayant des faces très rapprochées, il se peut que la dilatation engendre un chevauchement des hyper-rectangles pour la direction correspondante. Dès lors, il n'existe plus de parax-hyperplans dans cette direction permettant de les séparer. Cela signifie que les classes ne sont plus séparables selon cette direction. Or, ce cas est des plus fréquents dans la mesure où les points sont classés de part et d'autre d'une transition, et plus la densité des points est importante, plus les faces des hyper-rectangles de chaque classe seront proches.

Il est donc nécessaire de contracter (on dit aussi éroder en morphologie mathématique) les hyper-rectangles vers leur barycentre d'une distance de 2σ. D'un point de vue formel, $\min_{x \in \mathcal{F}_j} x_i$ devient $\min_{y \in \mathcal{F}_j} y_i + 2\sigma$ et $\max_{x \in \mathcal{F}_j} x_i$ devient $\max_{y \in \mathcal{F}_j} y_i - 2\sigma$. Bien-sûr, un hyper-rectangle ne peut être contracté au delà de son barycentre, de sorte que si

$\max_{y\in\mathcal{F}_j} y_i - \min_{y\in\mathcal{F}_j} y_i > 4\sigma$, $\min_{x\in\mathcal{F}_j} x_i$ et $\max_{x\in\mathcal{F}_j} x_i$ deviennent $\min_{y\in\mathcal{F}_j} y_i + (\max_{y\in\mathcal{F}_j} y_i - \min_{y\in\mathcal{F}_j} y_i)/2$. A cause de cette dernière remarque, des césures plus proches que σ n'ont pas de sens et doivent donc être supprimées.

6.2.2 Générer les césures maximales

L'énumération de l'ensemble des césures maximales exploite la liste des césures $\mathcal{C}^*$ séparément suivant chacune des n directions, en utilisant la proprité de la restriction d'un ordre partiel introduite en (5.10) au chapitre 5.2. La procédure est résumée dans l'Algorithme 2.

Algorithme 2. *Génère* $Max_{\preceq}\mathcal{C}^*$

Initialise $\bar{\mathcal{C}}^{i*} = \emptyset,\ i = 1,\ldots,n.$
pour $i = 1,\ldots,n$ ***faire***
  Pose $max\_flag = true.$
  pour $k = 1,\ldots,|\mathcal{C}^{i*}|$ ***faire***
    pour $l = \{1,\ldots,|\mathcal{C}^{i*}|\} \setminus \{k\}$ ***faire***
      Soit θ *la* k*-ième césure de* $\mathcal{C}^{i*}$ *et* θ' *la* l*-ième :*
      si $S(\theta) \subset S(\theta')$ ***alors***
        Pose $max\_flag = false.$
      fin si
    fin pour
    si $max\_flag = true$ ***alors***
      Ajoute θ *à* $\bar{\mathcal{C}}^{i*}$.
    fin si
  fin pour
fin pour
Définit $Max_{\preceq}\mathcal{C}^* = \cup_{i=1}^{n}\bar{\mathcal{C}}^{i*}$.

6.2.3 Générer les multicésures parcimonieuses

De manière à énumérer toutes les solutions parcimonieuses, nous pourrions, en principe, générer tous les sous-ensembles de $Max_{\preceq}\mathcal{C}^*$ et vérifier que ce sont bien des multicésures de cardinalité minimale en suivant les Définitions 7 et 9 (pages 80 et 82). Cependant, comme pour le problème des partitions, nous devons faire face à une explosion combinatoire, ce qui rend la résolution extrêmement fastidieuse même pour des exemples fort simples. Il est donc nécessaire d'introduire une méthode plus efficace. Commençons par reformuler le problème.

Problème 6. *Nous voulons déterminer tous les ensembles de césures* $\hat{\mathcal{M}} \in \mathcal{M}^*$ *(ensemble de toutes les multicésures) de telle sorte que*

$$|\hat{\mathcal{M}}| = \min_{\mathcal{M}\in\mathcal{M}^*} |\mathcal{M}|.$$

Dans notre cas, $\mathcal{M}^* = \downarrow \{Max_{\preceq}\mathcal{C}^*\}$.

Le problème 6 est en fait problème d'optimisation combinatoire. Un certain nombre de stratégies existent dans la littérature que nous allons survoler ici.

Commençons par bien spécifier le problème soulevé, et en particulier la contrainte imposée sur l'ensemble $\mathcal{M}$ par $\mathcal{M} \in \downarrow \{Max_{\preceq}\mathcal{C}^*\}$. Peu importe à la limite ce à quoi correspondent les éléments de $\mathcal{M}$, ce qu'il faut retenir, c'est qu'à chacun de ces éléments

correspondent une autre liste d'éléments, qui forment un sous-ensemble d'une liste fixée U. Dans notre cas concret, $U = \{u = (p, q) \in \{1, \ldots, s\}^2 : p < q\}$, et un élément de $\mathcal{M}$ est une césure qui a un pouvoir de séparation dans U. Mais en fait, pour ce qui concerne la résolution du Problème 6, tout cela peut être abstrait, et il faut seulement retenir que, étant donné un ensemble $\mathcal{M}$ d'éléments, nous voulons trouver l'ensemble de tous les sous-ensembles de ses éléments qui vérifient une certaine contrainte et qui ont une cardinalité minimale. Cette contrainte est que l'union des listes dans U correspondantes à chaque élément d'un sous-ensemble de $\mathcal{M}$ est en fait exactement U.

La manière la plus simple de résoudre ce problème est la recherche non-informée ("blind search") : aucune information au sujet de la taille des sous-ensembles à considérer n'est connue. Une manière de résoudre le problème est de considérer un ensemble candidat dont la cardinalité ne peut que grandir (les éléments sont remplacés ou ajoutés), de telle sorte que toutes les solutions candidates atteintes pour une cardinalité donnée vérifiant la contrainte sont, de fait, minimale. L'approche symétrique consiste à partir de l'ensemble de toutes les césures possibles et de retirer un à un les éléments. Dans les deux cas, le processus consiste à construire un arbre de recherche : une branche consiste à ajouter ou à retrancher un élément, ou plusieurs, ou à les remplacer.

Un problème usuel concernant les arbres de recherche est l'explosion combinatoire. Pour illustrer ce fait, prenons pour exemple un labyrinthe que nous parcourons : à chaque intersection, nous faisons des choix, mais si ce choix s'avère mauvais (impasse), il faut repartir en arrière et se souvenir de toutes les combinaisons de couloirs précédemment tentées ; or, plus il y a d'intersections, plus le nombre de possibilités se démultiplient. Si le nombre de cas s'accroît exponentiellement avec la profondeur de la recherche dans l'arbre, nous désirons atteindre les solutions aussi vite que possible, car cela permet de mettre une limite dans la profondeur à explorer, et cela est d'autant plus justifié que, comme c'est le cas pour notre problème, nous désirons obtenir toutes les solutions. De ce fait, une première manière de travailler consiste à examiner les noeuds de l'arbre en augmentant progressivement la profondeur que l'on s'autorise. Il s'agit d'une recherche dite en largeur-d'abord ("breadth-first", voir par exemple [27] au chapitre 22.2, page 531 à 539), où tous les successeurs d'un noeud traité sont pris en considération. Les défauts d'une telle approche sont l'usage de la mémoire qui s'accroît de manière exponentielle en fonction de la profondeur traitée, et le temps de traitement qui s'accroît de la même manière. Une autre stratégie qui demande beaucoup moins de mémoire est la stratégie dite de profondeur d'abord ("depth-first", voir par exemple le chapitre 22.3 de [27], page 540-549), où les noeuds considérés sont toujours ceux avec une profondeur incrémentée. Une telle stratégie permet d'atteindre une solution très rapidement... ou pas, si une branche choisie très tôt ne conduit en fait à aucune solution.

Quand la profondeur de la recherche atteint des valeurs importantes, il devient critique d'utiliser des stratégies de recherche dites informées (voir par exemple [101]), qui tentent d'exploiter l'ensemble des informations obtenues à chaque étape de la recherche. Il s'agira alors de limiter la progression de la recherche sur l'une des branches parcourues si il semble raisonnable de considérer qu'elle ne conduira pas à une des solutions.

Dans la mesure où nous recherchons toutes les solutions, nous proposons maintenant une stratégie informée : il s'agit d'une méthode d'énumération par séparation et évaluation ("branch-and-bound"). Il nous faut alors introduire les concepts essentiels liés à l'évaluation.

Définition 13. *Nous dirons que nous savons comment* évaluer $\mathcal{M} \subseteq \mathcal{M}^*$ *si nous pouvons trouver une borne inférieure à* $\{|\mathcal{C}|\}_{\mathcal{C} \in \downarrow \mathcal{M}}$*, nommée* $eval(\mathcal{M})$*. Nous dirons que nous pouvons*

évaluer exactement $\mathcal{M} \subseteq \mathcal{M}^*$ *si la limite inférieure est un élément minimal :* $eval(\mathcal{M}) = \min_{\mathcal{C} \in \downarrow \mathcal{M}} |\mathcal{C}|$.

Ces définitions nous permettent de poser les propriétés sur lesquelle s'appuie la méthode d'énumération par séparation et évaluation.

Propriété 14 (Principe d'élagage)**.** *Soit $\mathcal{M} \in \mathcal{M}^*$ un ensemble de solutions candidates. S'il existe $\mathcal{M}' \subseteq \mathcal{M}^*$ évalué exactement tel que $eval(\mathcal{M}') < eval(\mathcal{M})$, alors il est possible de ne plus considérer les éléments de $\mathcal{M}$ comme solutions candidates.*

Démonstration. En effet, cela signifie qu'il existe une solution dont la cardinalité est plus petite que n'importe quelle multicésure de $\mathcal{M}$ (ou toute multicésure dérivée à partir d'elle). ■

Propriété 15. *Soit $\mathcal{M} \in \mathcal{M}^*$ une multicésure, alors*

$$\forall \mathcal{M}' \in \downarrow \{\mathcal{M}\}, \begin{cases} Req(\mathcal{M}) \subseteq Req(\mathcal{M}'), \\ Can(\mathcal{M}') \subseteq Can(\mathcal{M}). \end{cases}$$

Démonstration. $\mathcal{M}' \subseteq \mathcal{M}$. Aucune des césures requises de $\mathcal{M}$ ne peut pas être omise par définition : dans le cas contraire, $\mathcal{M}'$ ne serait pas un multicésure. Comme l'ensemble des césures requises ne peut que grandir, celui des césures superflues ou candidates ne peut que réduire. ■

L'Algorithme 3 inspiré de [103] présente la formulation usuelle de la procédure d'énumération par séparation et évaluation. A noter que "*choisir*", "*évaluer*", "*séparer*" sont les trois étapes clés de l'Algorithme 3 : nous allons les expliciter dans ce qui suit. Dans l'algorithme, $\hat{\mathcal{M}}$ est l'ensemble de toutes les solutions possibles à une étape donnée et stocke les multicésures les plus petites trouvées jusque là ; quant à $\mathcal{M}^\dagger$, c'est l'ensemble des solutions candidates à examiner. La condition commençant à la ligne 8 se base sur la Propriété 15. L'algorithme s'arrête lorsque toutes les multicésures "nécessaires et suffisantes" ont été examinées (ligne 4).

Algorithme 3. *Énumération des solutions parcimonieuses par séparation et évaluation.*

1: $\hat{\mathcal{M}} = \emptyset$
2: $best = +\infty$
3: $\mathcal{M}^\dagger = \{Max_{\preceq} \mathcal{C}^*\}$
4: ***tant que*** $\mathcal{M}^\dagger \neq \emptyset$ ***faire***
5: *Choisir $\mathcal{M}$ dans $\mathcal{M}^\dagger$*
6: ***si*** *$\mathcal{M}$ n'a pas été évalué* ***alors***
7: *Évaluer $\mathcal{M}$*
8: ***si*** *$eval(\mathcal{M}) > best$* ***alors***
9: $\mathcal{M}^\dagger = \mathcal{M}^\dagger \setminus \{\mathcal{M}\}$
10: ***autrement***
11: ***si*** *l'évaluation est exacte* ***alors***
12: ***si*** *$eval(\mathcal{M}) = best$* ***alors***
13: $\hat{\mathcal{M}} = \hat{\mathcal{M}} \cup \{\mathcal{M}\}$
14: $\mathcal{M}^\dagger = \mathcal{M}^\dagger \setminus \{\mathcal{M}\}$

15: ***autrement***
16: $best = eval(\mathcal{M})$
17: $\hat{\mathcal{M}} = \{\mathcal{M}\}$
18: $\mathcal{M}^\dagger = \mathcal{M}^\dagger \setminus \{\mathcal{M}\}$
19: ***fin si***
20: ***fin si***
21: ***fin si***
22: ***autrement***
23: *Séparer : remplacer $\mathcal{M}$ pour mettre à jour $\mathcal{M}^\dagger$*
24: ***fin si***
25: ***fin tant que***

Évaluation

Commençons par préciser la Définition 13 : il nous faut borner inférieurement la taille des multicésures de la fermeture inférieure d'une multicésure.
Soit $\mathcal{M} \in \mathcal{M}^*$, $\downarrow\{\mathcal{M}\} \subseteq \mathcal{M}^*$:

- $|Req(\mathcal{M})|$ est une borne inférieure à $\{|\mathcal{M}'|\}_{\mathcal{M}' \in \downarrow\{\mathcal{M}\}}$; l'évaluation est exacte lorsque $|Req(\mathcal{M})| = |\mathcal{M}|$.
- $|\mathcal{M}| - |Can(\mathcal{M})|$ est une borne inférieure plus précise, parce que $|Req(\mathcal{C})| + |Can(\mathcal{C})| \leq |\mathcal{C}|$: l'évaluation exacte est la même puisque, quand $|Req(\mathcal{C})| = |\mathcal{C}|$, $Can(\mathcal{C}) = \emptyset$.

Définition 14. *Pour $\mathcal{M}^\dagger \subseteq \mathcal{M}^*$,*

$$eval(\mathcal{M}^\dagger) = \min_{\mathcal{M} \in \mathcal{M}^\dagger} (|\mathcal{M}| - |Can(\mathcal{M})|).$$

L'évaluation est exacte quand : $\exists \mathcal{M} \in \mathcal{M}^\dagger, |Req(\mathcal{M})| = |\mathcal{M}|$.

Séparation

Nous avons besoin de séparer quand l'évaluation de tout élément de $\mathcal{M}^\dagger$ n'est pas exacte : cela signifie qu'il nous faut remplacer $\mathcal{M} \in \mathcal{M}^\dagger$ (dont l'évaluation n'est pas exacte) par un nouvel ensemble de multicésures qui soient telles que leur pouvoir de séparation soit le même que celui de $\mathcal{M}$.

Si l'évaluation n'est pas exacte, alors : $\nexists \mathcal{M} \in \mathcal{M}^\dagger, |Req(\mathcal{M})| = |\mathcal{M}|$. Pour tout $\mathcal{M} \in \mathcal{M}^\dagger$, deux cas sont alors possibles :

- si $Can(\mathcal{M}) = \emptyset$, alors il reste seulement des césures superflues, que l'on peut donc ignorer : $\mathcal{M}$ sera remplacé par $Req(\mathcal{M})$;
- si $Can(\mathcal{M}) \neq \emptyset$, alors seuls les césures requises et les césures candidates sont intéressantes, et en utilisant la Propriété 11, il apparaît qu'il est possible de retirer au moins une césure candidate sans qu'aucun $u \in U$ ne soit plus séparé : tous les cas sont examinés, de sorte que $\mathcal{M}$ sera remplacé par $\{Req(\mathcal{M}) \cup Can(\mathcal{M}) \setminus \{\theta\}\}_{\theta \in Can(\mathcal{M})}$.

Finalement, si $\mathcal{M}^\dagger$ doit être séparé, alors il sera remplacé par :

$$\{Req(\mathcal{M})\}_{\mathcal{M} \in \mathcal{M}^\dagger : Can(\mathcal{M}) = \emptyset} \cup \left\{ \{Req(\mathcal{M}) \cup Can(\mathcal{M}) \setminus \{\theta\}\}_{\theta \in Can(\mathcal{M})} \right\}_{\mathcal{M} \in \mathcal{M}^\dagger : Can(\mathcal{M}) \neq \emptyset}.$$

Choix

La question du choix est un des aspects les plus critiques pour améliorer l'efficacité d'une optimisation combinatoire basée sur une approche par évaluation et séparation [103]. Cet aspect concerne l'ordre dans lequel il faut traiter les solutions candidates de $\mathcal{M}^\dagger$. Le choix usuel est de prendre en fonction du résultat de l'évaluation, c'est-à-dire que les ensembles non encore évalués sont traités en priorité. Une des stratégie les plus populaires est appelée *retour-sur-trace* ("backtracking") : elle consiste à prendre d'abord les cas générés en dernier (lors de la séparation). En général, il est fait de manière à arriver le plus vite possible à une solution (i.e. une multicésure évaluée exactement) pour pouvoir abaisser la borne *best*. Dès que cela est fait, la stratégie est changée pour la suivante : les ensembles dont l'évaluation est la plus faible sont choisis en premier.

L'idéal pour pouvoir préciser cette question est de pouvoir construire une heuristique sur cette question du choix. Nous ne détaillerons pas plus avant cette question ici.

Troisième partie

Applications

Chapitre 7

Reconstruction d'un réseau simplifié pour *E. coli*

Il est difficile de démontrer en toute généralité la pertinence du type d'algorithme décrit dans les chapitres précédents. Nous effectuerons dans ce chapitre-ci une démonstration par l'exemple en prenant un système déjà bien étudié dans sa formulation affine-par-morceaux : il s'agit de la réponse de la bactérie *Escherichia coli* à un stress nutritionnel [10, 99]. Cela permettra également de montrer la faisabilité de notre approche.

La réponse à un stress nutritionnel chez la bactérie *E. coli* est régulée par un réseau complexe d'interactions géniques, mais aussi protéiques. Cependant, il semble pertinent de se ramener à une machinerie cellulaire centrée sur un réseau de régulation génique incluant une douzaine de protéines régulatrices appelées régulateurs globaux [54, 123]. Ces protéines permettent à la bactérie non seulement de s'adapter aux variations de la présence de nutriments dans leur milieu [102], mais elles jouent également un rôle dans sa capacité à s'adapter aux conditions extérieures sur l'échelle de temps de l'évolution. En effet, les conditions de croissance de ces entérobactéries sont rarement constantes, ne serait-ce que parce que le milieu colonisé possède une quantité finie de ressources qui vont être consommées. Ainsi, tant que les nutriments sont disponibles dans le milieu, les bactéries *E. coli* croissent et se divisent activement pour se multiplier : c'est la phase dite exponentielle (Figure 7.1(a)), dûe au fait qu'une cellule en donne deux, qui deviendront quatre, puis huit, et ainsi de suite. Lorsqu'une source de nutriments (par exemple le carbone) tarit, la population cesse de croitre et entre en phase dite stationnaire (Figure 7.1(b)) car les divisions sont pour ainsi dire arrêtées : on peut relever une modification de la physiologie cellulaire [56]. Le réseau de régulation génique qui contrôle ces modifications peut ainsi s'adapter à un signal de manque de carbone (i.e. l'énergie manque) pour ralentir ou supprimer un certain nombre de voies de biosynthèse, et l'ADN bactérien est aussi mis sous "protection" afin de limiter les dommages qu'un métabolisme ralenti ne saurait compenser [56]. Ce phénomène est évidemment réversible : dès que les nutriments sont à nouveau présents, les bactéries reprennent immédiatement leur phase de croissance exponentielle pour celles qui ont survécu.

Le réseau de régulation de cet organisme concerné par la réponse à un stress nutritionnel est l'objet de nombreuses études depuis des années. Bien que *E. coli* soit un paradigme dans le monde bactérien, on comprend assez mal, même aujourd'hui, comment la réponse

(a) (b)

FIGURE 7.1 – *E. coli* en phase exponentielle (a) et en phase stationnaire (b).

à des conditions de famine se met en place à partir des interactions entre les régulateurs globaux de la bactérie. Dans les travaux précédents [32, 98, 9], l'utilisation d'une méthode qualitative (se basant sur des contraintes d'inégalité entre les paramètres du système modélisant la réponse) a permis de dépasser le manque de données quantitatives (paramètres de la cinétique, variation de concentration d'éléments intracellulaires) : un réseau a pu être proposé [99, 97] à partir des informations disponibles dans les publications issues de la biologie moléculaire expérimentale ou dans les bases de données disponibles. La réponse de ce réseau a pu être qualitativement simulée pour la comparer avec les motifs dans les variations concentrationnaires relevées expérimentalement. Ceci a permis d'identifier des caractéristiques essentielles se produisant lors des transitions entre les phases exponentielles et les phases stationnaires, et de proposer des nouvelles prédictions sur le comportement qualitatif du système à la suite de l'injection de carbone dans le milieu [32, 8] : on a ainsi été amené à considérer que des oscillations amorties de certaines concentrations précèdent l'atteinte d'un nouvel équilibre correspondant à la phase exponentielle, ce qui n'avait encore jamais été proposé dans la littérature.

Par la suite, nous nous sommes concentrés sur la colonne vertébrale du réseau de régulation génique décrit dans [99] : de manière à s'assurer d'une traitabilité numérique raisonnable pour pouvoir se prêter à une batterie intensive de tests faisant varier les paramètres de l'algorithme, nous nous sommes attachés à ne considérer que le réseau constitué par les gènes clés contrôlant la réponse à un stress nutritionnel de *E. coli*.

7.1 Réseau simplifié de la réponse à un stress nutritionnel chez *E. coli*

Le génome de *E. coli* renferme quelques 4500 gènes. Parmi ceux-là, on estime que 150 codent pour des facteurs de transcription impliqués dans la réponse aux stress (température, acidité du milieu, concentration en oxygène, et présence de carbone, entre autres).

Un réseau composé de six des gènes de *E. coli* connus pour jouer un rôle clé dans la réponse au stress nutritionnel a été construit [99]. Il inclut des gènes codant des protéines dont l'activité dépend d'un signal de stress nutritionnel (le régulateur global cAMP - protéine récepteur, CRP, Cya - adémilate cyclase) [47, 48, 74], des gènes impliqués dans le contrôle du métabolisme (le régulateur global Fis) [104], la croissance cellulaire (les gènes rrn codant les ARN stables, considérés comme représentatifs de l'état de croissance) [127, 26], et le superenroulement de l'ADN [114, 109, 75, 76], un modulateur important de l'expression génique (la gyrase GyrAB "enroulant" quand la topoisomérase TopA a l'effet inverse).

Pour résumer, on peut dire que ce réseau est constitué d'une entrée (la voie permettant au signal d'absence en carbone de se propager dans l'univers intracellulaire), une sortie (la concentrations d'ARN stables représentant l'état de croissance de la bactérie), et trois régulateurs globaux (CRP, Fis et la topologie de l'ADN). La Figure 7.2 permet de décrire ce réseau.

Le modèle APM du réseau de la réponse à une absence de carbone est décrit par les équations suivantes, où les variables $x_{\mathrm{Cya}}, x_{\mathrm{CRP}}, x_{\mathrm{Fis}}, x_{\mathrm{GyrAB}}, x_{\mathrm{TopA}}$ et x_{rrn} représentent les concentrations des protéines Cya, CRP, Fis, GyrAB, TopA et des ARN stables respectivement. La variable u_{s} représente la présence ou l'absence du signal de l'absence de carbone dans le milieu, qui vaudra 1 dans le premier cas (famine), 0 dans le second.

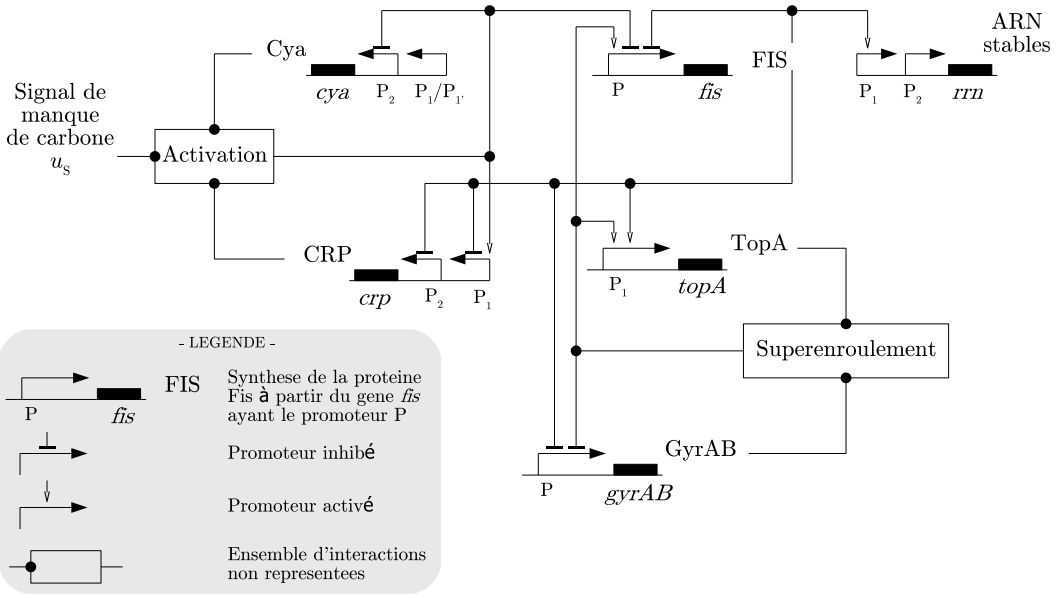


FIGURE 7.2 – Représentation du réseau simplifié de la réponse à un stress nutritionnel chez la bactérie *E. coli*.

$$
\begin{aligned}
\dot{x}_{\mathrm{Cya}} &= \kappa^1_{\mathrm{Cya}} + \kappa^2_{\mathrm{Cya}}\ (1 - s^+(x_{\mathrm{CRP}}, \theta^2_{\mathrm{CRP}})\, s^+(x_{\mathrm{Cya}}, \theta^2_{\mathrm{Cya}})\, s^+(u_{\mathrm{s}}, \theta_{\mathrm{s}})) - \gamma_{\mathrm{Cya}}\ x_{\mathrm{Cya}} \\
\dot{x}_{\mathrm{CRP}} &= \kappa^1_{\mathrm{CRP}} + \kappa^2_{\mathrm{CRP}}\ s^-(x_{\mathrm{Fis}}, \theta^2_{\mathrm{Fis}})\, s^+(x_{\mathrm{CRP}}, \theta^1_{\mathrm{CRP}})\, s^+(x_{\mathrm{Cya}}, \theta^1_{\mathrm{Cya}})\, s^+(u_{\mathrm{s}}, \theta_{\mathrm{s}}) \\
&\quad + \kappa^3_{\mathrm{CRP}}\ s^-(x_{\mathrm{Fis}}, \theta^1_{\mathrm{Fis}}) - \gamma_{\mathrm{CRP}}\ x_{\mathrm{CRP}} \\
\dot{x}_{\mathrm{Fis}} &= \kappa^1_{\mathrm{Fis}}\ s^-(x_{\mathrm{Fis}}, \theta^5_{\mathrm{Fis}})\ (1 - s^+(x_{\mathrm{CRP}}, \theta^1_{\mathrm{CRP}})\, s^+(x_{\mathrm{Cya}}, \theta^1_{\mathrm{Cya}})\, s^+(u_{\mathrm{s}}, \theta_{\mathrm{s}})) \\
&\quad + \kappa^2_{\mathrm{Fis}}\ s^+(x_{\mathrm{GyrAB}}, \theta^1_{\mathrm{GyrAB}})\, s^-(x_{\mathrm{TopA}}, \theta^2_{\mathrm{TopA}})\, s^-(x_{\mathrm{Fis}}, \theta^5_{\mathrm{Fis}}) \\
&\quad \times (1 - s^+(x_{\mathrm{CRP}}, \theta^1_{\mathrm{CRP}})\, s^+(x_{\mathrm{Cya}}, \theta^1_{\mathrm{Cya}})\, s^+(u_{\mathrm{s}}, \theta_{\mathrm{s}})) - \gamma_{\mathrm{Fis}}\ x_{\mathrm{Fis}} \\
\dot{x}_{\mathrm{GyrAB}} &= \kappa_{\mathrm{GyrAB}}\ (1 - s^+(x_{\mathrm{GyrAB}}, \theta^2_{\mathrm{GyrAB}})\, s^-(x_{\mathrm{TopA}}, \theta^1_{\mathrm{TopA}}))\ s^-(x_{\mathrm{Fis}}, \theta^4_{\mathrm{Fis}}) - \gamma_{\mathrm{GyrAB}}\ x_{\mathrm{GyrAB}} \\
\dot{x}_{\mathrm{TopA}} &= \kappa_{\mathrm{TopA}}\ s^+(x_{\mathrm{GyrAB}}, \theta^2_{\mathrm{GyrAB}})\, s^-(x_{\mathrm{TopA}}, \theta^1_{\mathrm{TopA}})\, s^+(x_{\mathrm{Fis}}, \theta^4_{\mathrm{Fis}}) - \gamma_{\mathrm{TopA}}\ x_{\mathrm{TopA}} \\
\dot{x}_{\mathrm{rrn}} &= \kappa^1_{\mathrm{rrn}}\ s^+(x_{\mathrm{Fis}}, \theta^3_{\mathrm{Fis}}) + \kappa^2_{\mathrm{rrn}} - \gamma_{\mathrm{rrn}}\ x_{\mathrm{rrn}}
\end{aligned}
$$

D'après [99], les inégalités suivantes s'appliquent pour la description de notre système :

$$
\begin{gathered}
0 < \theta^1_{\mathrm{Cya}} < \theta^2_{\mathrm{Cya}},\ \theta^1_{\mathrm{Cya}} < \kappa^1_{\mathrm{Cya}}/\gamma_{\mathrm{Cya}} < \theta^2_{\mathrm{Cya}},\ (\kappa^1_{\mathrm{Cya}} + \kappa^2_{\mathrm{Cya}})/\gamma_{\mathrm{Cya}} > \theta^2_{\mathrm{Cya}} \\
0 < \theta^1_{\mathrm{CRP}} < \theta^2_{\mathrm{CRP}},\ \theta^1_{\mathrm{CRP}} < \kappa^1_{\mathrm{CRP}}/\gamma_{\mathrm{Cya}} < \theta^2_{\mathrm{CRP}} \\
\theta^1_{\mathrm{CRP}} < (\kappa^1_{\mathrm{CRP}} + \kappa^2_{\mathrm{CRP}})/\gamma_{\mathrm{CRP}} < \theta^2_{\mathrm{CRP}},\ (\kappa^1_{\mathrm{CRP}} + \kappa^3_{\mathrm{CRP}})/\gamma_{\mathrm{CRP}} > \theta^2_{\mathrm{CRP}} \\
0 < \theta^1_{\mathrm{Fis}} < \theta^2_{\mathrm{Fis}} < \theta^3_{\mathrm{Fis}} < \theta^4_{\mathrm{Fis}} < \theta^5_{\mathrm{Fis}} \\
\theta^1_{\mathrm{Fis}} < \kappa^1_{\mathrm{Fis}}/\gamma_{\mathrm{Fis}} < \theta^2_{\mathrm{Fis}},\ (\kappa^1_{\mathrm{Fis}} + \kappa^2_{\mathrm{Fis}})/\gamma_{\mathrm{Fis}} > \theta^5_{\mathrm{Fis}} \\
0 < \theta^1_{\mathrm{GyrAB}} < \theta^2_{\mathrm{GyrAB}},\ \kappa_{\mathrm{GyrAB}}/\gamma_{\mathrm{GyrAB}} > \theta^2_{\mathrm{GyrAB}} \\
0 < \theta^1_{\mathrm{TopA}} < \theta^2_{\mathrm{TopA}},\ \kappa_{\mathrm{TopA}}/\gamma_{\mathrm{TopA}} > \theta^2_{\mathrm{TopA}} \\
\theta_{\mathrm{rrn}} > 0,\ 0 < \kappa^2_{\mathrm{rrn}}/\gamma_{\mathrm{rrn}} < \theta_{\mathrm{rrn}},\ (\kappa^1_{\mathrm{rrn}} + \kappa^2_{\mathrm{rrn}})/\gamma_{\mathrm{rrn}} > \theta_{\mathrm{rrn}}
\end{gathered}
$$

Proposée par Delphine Ropers, la table qui suit donne les ordres de grandeur des diverses quantités considérées dans ce cas biologique : elles sont choisies pour être physiologiquement réalistes.

Taux de synthèse [M min^{-1}]		Taux de dégradation [min^{-1}]		Valeur des seuils [M]		Conditions initiales [M]	
κ^1_{Cya}	$3.034 \cdot 10^{-12}$	γ_{Cya}	$4.211 \cdot 10^{-2}$	θ^1_{Cya}	$2.748 \cdot 10^{-11}$	x^0_{Cya}	$2.413 \cdot 10^{-10}$
κ^2_{Cya}	$2.317 \cdot 10^{-11}$	γ_{CRP}	$4.327 \cdot 10^{-3}$	θ^2_{Cya}	$2.413 \cdot 10^{-10}$	x^0_{CRP}	$7.101 \cdot 10^{-6}$
κ^1_{CRP}	$1.553 \cdot 10^{-9}$	γ_{Fis}	$4.261 \cdot 10^{-3}$	θ^1_{CRP}	$1.719 \cdot 10^{-7}$	x^0_{Fis}	$5.000 \cdot 10^{-9}$
κ^2_{CRP}	$1.224 \cdot 10^{-9}$	γ_{GyrAB}	$5.188 \cdot 10^{-3}$	θ^2_{CRP}	$7.753 \cdot 10^{-7}$	x^0_{GyrAB}	$1.614 \cdot 10^{-8}$
κ^3_{CRP}	$1.322 \cdot 10^{-8}$	γ_{TopA}	$6.625 \cdot 10^{-3}$	θ^1_{Fis}	$3.991 \cdot 10^{-8}$	x^0_{TopA}	$2.661 \cdot 10^{-9}$
κ^1_{Fis}	$3.404 \cdot 10^{-10}$	γ_{rrn}	$8.468 \cdot 10^{-3}$	θ^2_{Fis}	$1.888 \cdot 10^{-7}$	x^0_{rrn}	$2.999 \cdot 10^{-6}$
κ^2_{Fis}	$8.668 \cdot 10^{-9}$			θ^3_{Fis}	$3.663 \cdot 10^{-7}$		
κ_{GyrAB}	$9.938 \cdot 10^{-10}$			θ^4_{Fis}	$7.472 \cdot 10^{-7}$		
κ_{TopA}	$2.548 \cdot 10^{-9}$			θ^5_{Fis}	$2.020 \cdot 10^{-6}$		
κ^1_{rrn}	$1.488 \cdot 10^{-7}$			$\theta^1_{\mathrm{GyrAB}}$	$3.991 \cdot 10^{-8}$		
κ^2_{rrn}	$1.506 \cdot 10^{-8}$			$\theta^2_{\mathrm{GyrAB}}$	$1.888 \cdot 10^{-7}$		
				θ^1_{TopA}	$1.221 \cdot 10^{-7}$		
				θ^2_{TopA}	$2.491 \cdot 10^{-7}$		
				θ_{rrn}	$8.898 \cdot 10^{-6}$		
				θ_{s}	0.5		

À partir de ce modèle, nous avons simulé des trajectoires x représentant le comportement de la bactérie en réponse à l'absence ou la présence de nutriments. Pour cela des conditions initiales réalistes ont été choisies de manière à se placer dans les conditions biologiques relatives à la phase stationnaire d'une part, et à la phase exponentielle d'autre

part. L'analyse qualitative permet de déterminer les intervalles auxquels doivent appartenir les conditions initiales : cela permet d'avoir une flexibilité sur le choix de celles-ci, conduisant ainsi à des séquences durant la réponse transitoire du système qui ne sont pas similaires (la trajectoire ne passe pas toujours par les mêmes domaines de régulation).

L'entrée en phase exponentielle peut être simulée en imposant les conditions initiales suivantes :

$$\begin{aligned} x^0_{\text{Cya}} &= \theta^2_{\text{Cya}} \\ x^0_{\text{CRP}} &> \theta^2_{\text{CRP}} \\ 0 \leq x^0_{\text{Fis}} &< \theta^1_{\text{Fis}} \\ 0 \leq x^0_{\text{GyrAB}} &< \theta^1_{\text{GyrAB}} \\ 0 \leq x^0_{\text{TopA}} &< \theta^1_{\text{TopA}} \\ 0 \leq x^0_{\text{rrn}} &< \theta_{\text{rrn}} \end{aligned}$$

En ce qui concerne la simulation de la trajectoire, les échelles de temps d'échantillonnage et l'amplitude du bruit de mesure sont choisies pour être cohérentes avec les caractéristiques des données qu'il est possible d'obtenir par une technique dite à base de "gènes rapporteurs". Cette méthode, si elle n'est pas extensive (à l'opposé des méthodes à base de puces à ADN par exemple), permet de suivre finement l'évolution des concentrations des produits de l'expression génique. Nous en donnons une description plus conséquente dans ce qui suit.

7.2 Données de mesure

Plusieurs approches expérimentales existent en ce qui concerne le suivi de l'évolution des concentrations d'éléments intracellulaires dans le temps. Le système le plus utilisé aujourd'hui est la puce à ADN [93, 17, 108]. Cependant, en ce qui concerne l'obtention de séries temporelles quantitatives, RT-PCR [35] et *gènes rapporteurs* [107, 3] sont en général préférés.

Notre approche est très semblable à celle décrite dans [96]. Les systèmes à base de gènes rapporteurs sont l'outil usuel utilisé pour évaluer l'expression d'un gène en fusionnant son site promoteur à une séquence génétique construite appelée gène rapporteur : de cette manière, la machinerie cellulaire synthétise une protéine traçable de la même manière que la protéine que l'on veut observer. Précisément, l'expression du gène rapporteur (dont le produit est facilement mesuré, nous allons le voir) est sensée refléter l'expression du gène d'intérêt.

Dans notre cas, le produit de l'expression des gènes a été choisi pour être visible (par fluorescence ou bioluminescence) : les capteurs optiques, permettant une conversion photon - grandeur électrique - valeur numérique, autorisent alors une mesure en temps réel de cellules vivantes avec un pas d'échantillonnage temporel extrêmement faible, bien suffisant pour pouvoir observer la réponse transitoire des bactéries (de l'ordre de l'heure). De plus, il est possible de mesurer l'expression de plusieurs gènes dans les mêmes cellules en utilisant différents systèmes de gènes rapporteurs : par exemple, l'utilisation de simples filtres optiques permet de discriminer des molécules émettant sur des longueurs d'onde différentes. En pratique, on mesure au plus deux gènes de manière à ce que le flux détourné (ne serait-ce que pour l'énergie consommée) n'affecte pas trop largement le système vivant

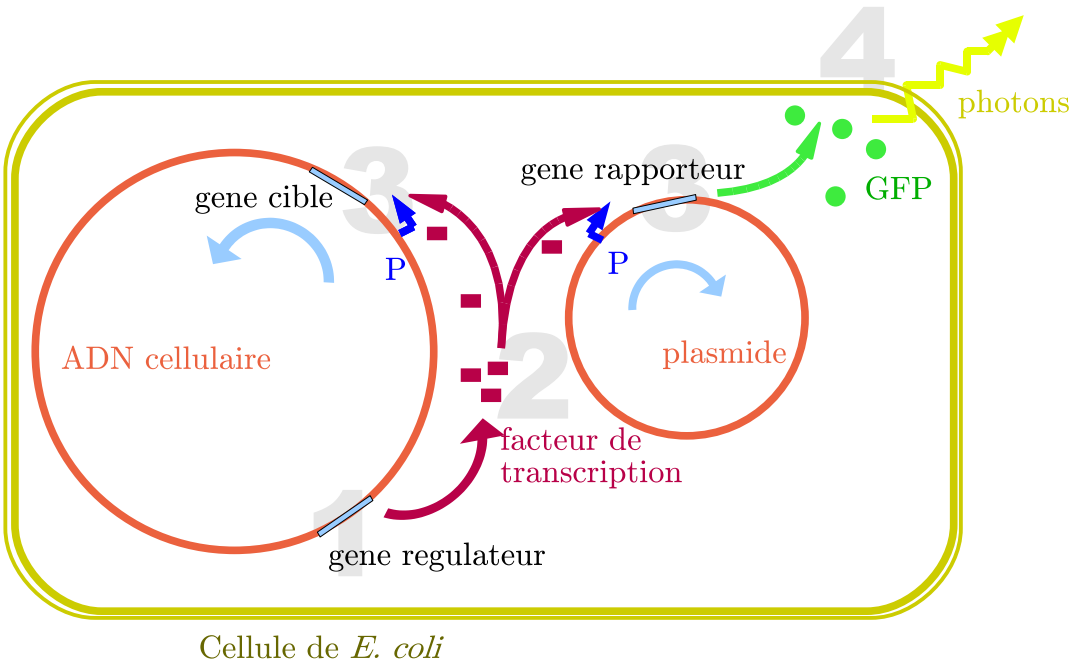


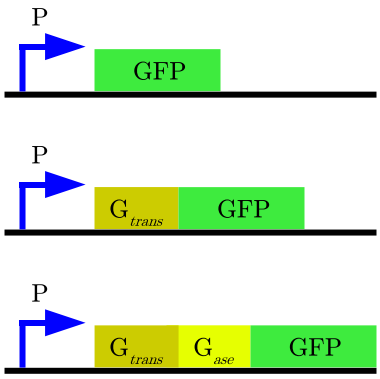
FIGURE 7.3 – Schéma décrivant le principe de la méthode à base de gène rapporteur. L'expression du gène régulateur conduit à la production de facteurs de transcription qui régulent l'expression du gène cible, ainsi que celle du gène rapporteur. Ce dernier code pour des protéines rapportrices (ici GFP) dont le nombre est proportionnel à l'énergie lumineuse mesurable.

étudié, introduisant un biais systématique qui ne permettrait pas de rendre compte de la réalité biologique.

D'un point de vue technique, un système de gènes rapporteurs est obtenu par fusion de la région promotrice (et dans certains cas une partie de la région codante si cela permet d'en préserver certaines propriétés, telles que la régulation ou la reconnaissance pour la dégradation) d'un gène d'intérêt à un gène rapporteur. La fusion génétique est clonée dans un plasmide (ADN circulaire) à faible nombre de copies (afin d'éviter de trop solliciter la machinerie cellulaire sur cet élément introduit), qui est placé dans le milieu intracellulaire de *E. coli*. Ainsi, lorsque les régulateurs du gène cible sont présents, ils régulent "de la même manière" le gène rapporteur : les taux d'expression sont ainsi proportionnels. Pour un gène rapporteur codant par exemple la protéine GFP ("green fluorescent protein"), si le gène cible est exprimé, le gène rapporteur l'est aussi, et la cellule se met alors à avoir une réponse lumineuse dont l'intensité est rapportable au nombre de GFP (la physique de l'émission des photons de ces molécules est bien caractérisée). La Figure 7.3 représente de manière schématique cette méthode d'obtention des données de mesure.

Les bactéries sont mises en croissance dans des microplaques pouvant être lues par un lecteur mesurant aussi bien l'absorbance de la population bactérienne (à rattacher au nombre et à la taille des bactéries) que le signal de fluorescence émis par la GFP.

Il existe plusieurs manières de fusionner les gènes rapporteurs : voir Figure 7.4. Cela



- mesure de la transcription : le promoteur P du gène cible étudié sera reconnu par les facteurs de transcription, ce qui permet la production de la protéine GFP ;

- mesure de la traduction (dans le cas d'une régulation sur la translation) : G_{trans} est la partie codant pour la zone régulant la translation du gène cible ;

- mesure de l'expression avec une régulation sur la dégradation : G_{ase} est la partie servant à la protéine (issue de l'expression du gène cible) pour être reconnue par sa protéase correspondante.

FIGURE 7.4 – Diverses manières d'utiliser la méthode à base de gène rapporteur, fusionnant au gène rapporteur au moins le promoteur du gène cible - et éventuellement des parties codantes supplémentaires - afin de mesurer la régulation à différents stades de l'expression génique. Le code ADN terminal est évidemment le gène rapporteur, codant ici pour la GFP.

permet de mesurer ce qu'il se passe à différents niveaux de régulation (de la transcription, de la translation ou de la dégradation). Cependant, les séquences les plus longues étant les moins stables, il s'agit souvent d'une fusion transcriptionnelle : le seul produit est alors le marqueur (pour notre exemple, la GFP).

En pratique, deux systèmes ont pu être construits pour chaque gène cible envisagé. Le premier est à base de dérivés de la GFP précédemment introduite : cela conduit à une mesure de la fluorescence du système. Le signal obtenu est cependant très faible et, pour un taux d'expression faible, il n'est pas possible de le distinguer du bruit de fond de la chaine de mesure. Un autre désavantage important des GFPs est que ce sont des molécules très stables (demi-vie bien plus grande que le temps d'expérience de quelques heures) : il s'agit donc d'une mesure intégrative. Le deuxième système de gènes rapporteurs utilisé fusionnait l'opéron lux codant pour la luciférase et d'autres enzymes. La luciférase catalyse une réaction qui, en présence d'oxygène, de $FMNH_2$ et ATP, peut émettre spontanément de la lumière visible. L'intensité optique produite est mieux marquée, mais la sollicitation de la machinerie cellulaire est plus importante, ce qui peut limiter la mesure pour des taux d'expression trop important. Cela rend les deux systèmes particulièrement complémentaires.

Si les données obtenues sont d'une qualité inégalée par les autres types de mesure (RT-PCR) en terme de densité et de précision, cette méthode suggère diverses précautions. Soulignons d'abord qu'il est vérifié que les produits de l'expression des gènes rapporteurs ne viennent pas interférer avec le système : cependant, en particulier dans le cas de la luciférase, il se peut que des phénomènes indésirables extérieurs au réseau le perturbe très nettement (cela est observé lorsque l'ATP est surconsommé, privant la cellule d'une source privilégiée d'énergie). D'autre part, comme il apparait très nettement dans la Figure 7.3, un certain nombre de facteurs de transcription sont "détournés" de leur cible : l'idéal serait de pouvoir insérer un copie de gène rapporteur dans le génome-même de la cellule.

La dernière difficulté qui n'est pas des moindres provient du fait que la stabilité des molécules rapportrices n'est en rien asservie à celle du produit de l'expression des gènes

cibles (sauf si la séquence qui sera reconnue par la protéase est insérée dans le rapporteur). Il s'agit alors de procéder à un traitement du signal non-trivial permettant de déduire des mesures effectuées les concentrations des éléments du réseau de régulation observé.

7.3 Reconstruction du réseau de régulation génique

Dans la mesure où nous souhaitons dans un premier temps évaluer la qualité des résultats de la chaine de traitement, nous prendrons des données de mesure simulées. Des résultats ont été présentés dans [36] pour un réseau légèrement plus simple que celui introduit au 7.1. Nous utiliserons donc pour la suite à titre illustratif l'exemple présenté dans [87] obtenu à l'Université de Pavie par Riccardo Porreca.

La Figure 7.5 montre un ensemble de 73 points de mesures obtenu en simulant l'entrée en phase exponentielle après l'introduction de nutriments. On a choisi $T = 10$ minutes et l'écart type du bruit est $\sigma_i = 0,01\,(\max\{x_i\} - \min\{x_i\})$. Les valeurs des σ_i conduisent à un rapport signal-bruit $RSB = 0,01$ pour chaque gène. Signalons que la densité de l'échantillonnage et le niveau de bruit est similaire à ceux des données de mesure obtenues avec des gènes rapporteurs.

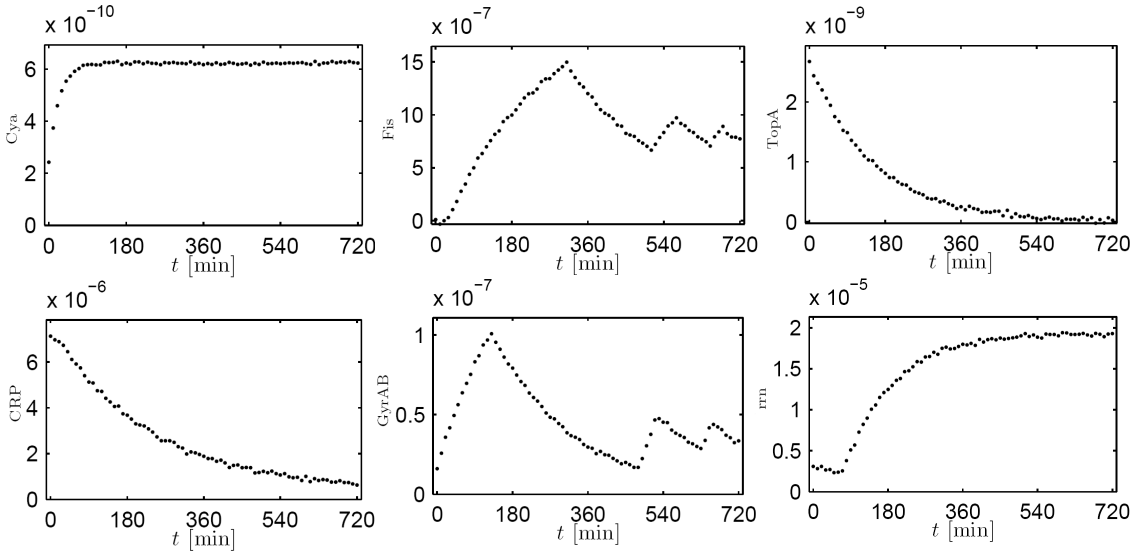


FIGURE 7.5 – Données simulées de la réponse à un manque en carbone du réseau de régulation génique de *E. coli* pour $T = 10$ min et $RSB = 0,01$. Les grandeurs en ordonnées correspondent à des mesures de concentration de Cya, CRP, Fis, GyrAB, TopA, et des ARN stables. A $t = 0$, u_s est forcé à 0 : les nutriments sont mélangés au milieu.

Les résultats des algorithmes effectuant la segmentation puis la classification des données de la Figure 7.5 sont respectivement montrés sur les Figures 7.6 et 7.7.

Les césures maximales reconstruites à partir de ces données de classification sont décrites dans le tableau suivant. Le fait qu'une césure est dite correcte signifie qu'elle correspond bien à un seuil dans le modèle original. Cette correspondance sera explicitée dans le chapitre 8.2 suivant.

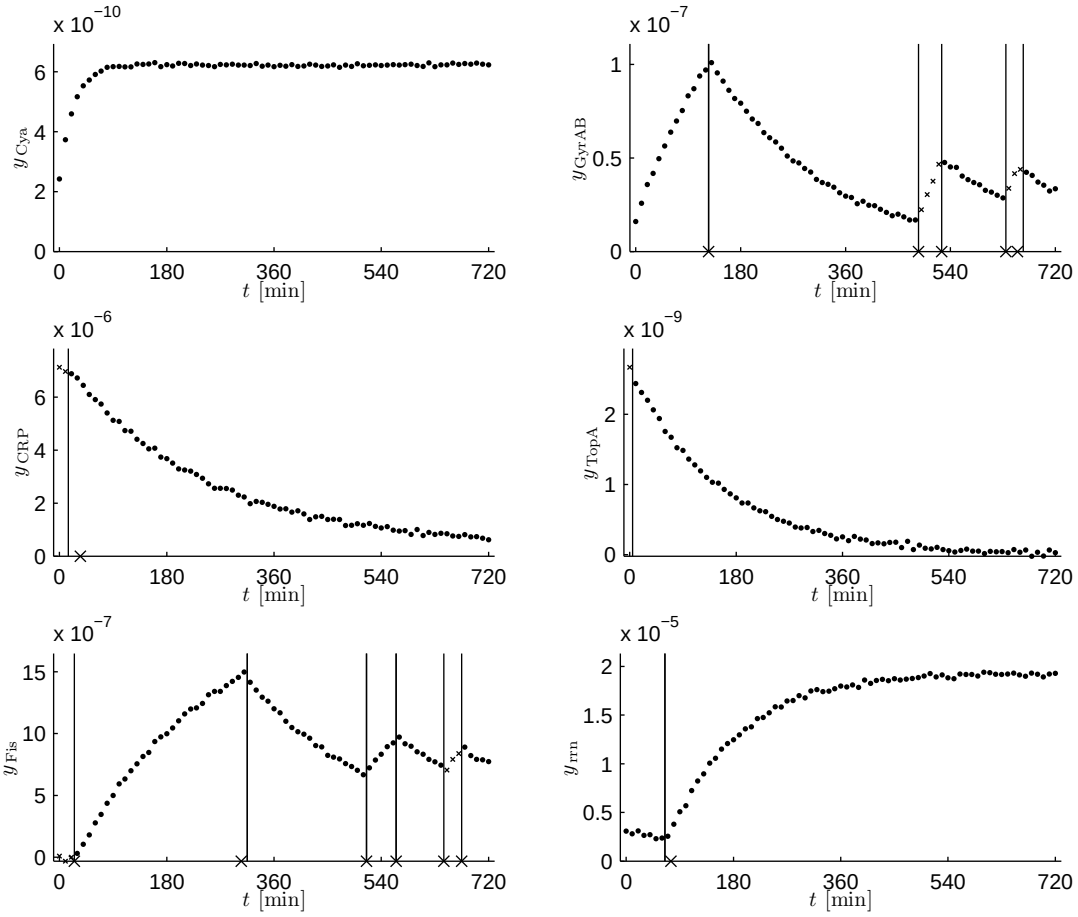


FIGURE 7.6 – Segmentation des données de mesure pour la réponse de *E. coli* à un stress nutritionnel. Les lignes verticales correspondent aux temps de transition détectés ($\alpha = 0,01$, $N_{\mathcal{S}} = 4$) qui définissent les segments. Les croix sur l'axe des abscisses correspondent aux temps de transition réels (connus par la simulation). Les points représentés par une croix n'ont pas pu être attribués à aucun segment.

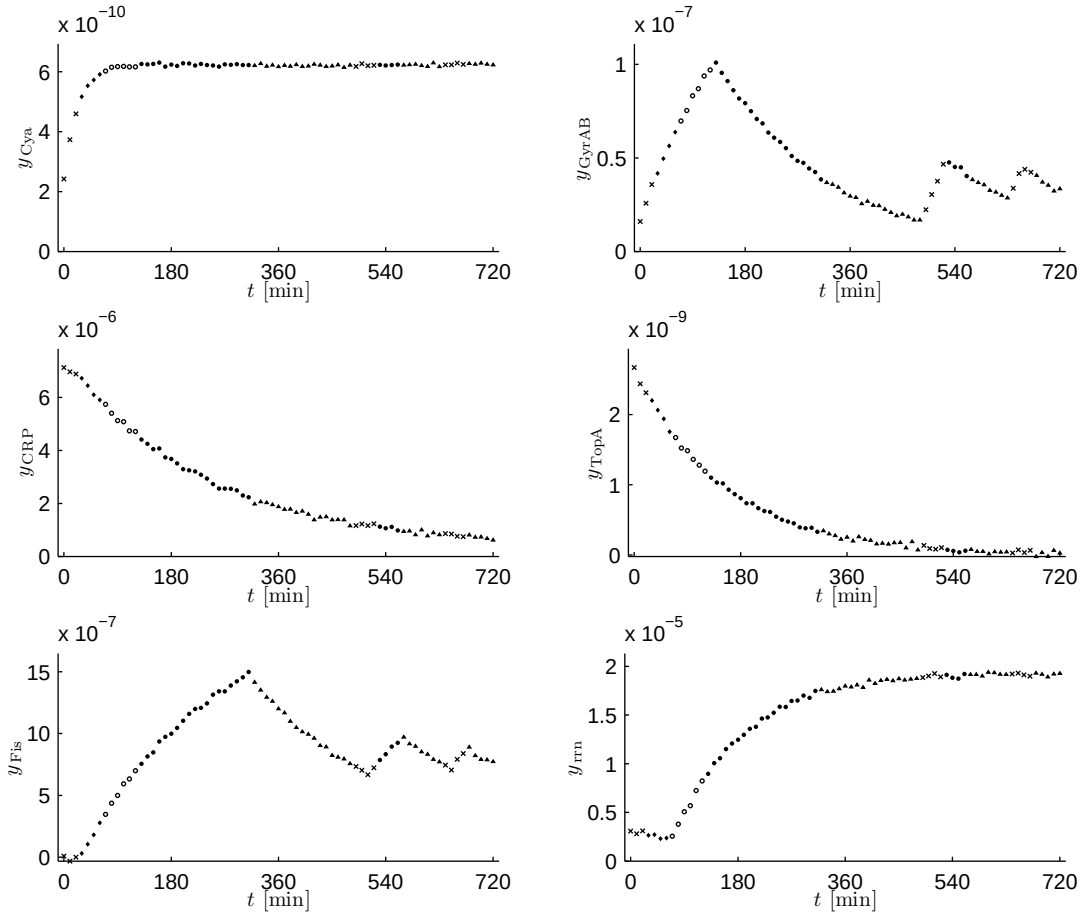


FIGURE 7.7 – Classification avec $\alpha = 0,01$. Les mêmes marqueurs sont attribués aux points d'une même classe.

Césure	Regulateur	Valeur [M]	I_{eq} [M]	Correct
$\hat{\theta}^1_{\text{Cya}}$	Cya	$5.899 \cdot 10^{-10}$	$[5.634 \cdot 10^{-10}, 6.165 \cdot 10^{-10}]$	N
$\hat{\theta}^1_{\text{CRP}}$	CRP	$4.498 \cdot 10^{-6}$	$[4.065 \cdot 10^{-6}, 4.932 \cdot 10^{-6}]$	N
$\hat{\theta}^2_{\text{CRP}}$	CRP	$5.746 \cdot 10^{-6}$	$[5.211 \cdot 10^{-6}, 6.281 \cdot 10^{-6}]$	N
$\hat{\theta}^1_{\text{Fis}}$	Fis	$3.118 \cdot 10^{-7}$	$[1.425 \cdot 10^{-7}, 4.810 \cdot 10^{-7}]$	Y
$\hat{\theta}^2_{\text{Fis}}$	Fis	$7.107 \cdot 10^{-7}$	$[5.892 \cdot 10^{-7}, 8.323 \cdot 10^{-7}]$	Y
$\hat{\theta}^1_{\text{GyrAB}}$	GyrAB	$3.975 \cdot 10^{-8}$	$[3.450 \cdot 10^{-8}, 4.501 \cdot 10^{-8}]$	Y
$\hat{\theta}^2_{\text{GyrAB}}$	GyrAB	$6.585 \cdot 10^{-8}$	$[5.378 \cdot 10^{-8}, 7.792 \cdot 10^{-8}]$	N
$\hat{\theta}^1_{\text{TopA}}$	TopA	$1.154 \cdot 10^{-9}$	$[9.513 \cdot 10^{-10}, 1.358 \cdot 10^{-9}]$	N
$\hat{\theta}^2_{\text{TopA}}$	TopA	$1.723 \cdot 10^{-9}$	$[1.444 \cdot 10^{-9}, 2.002 \cdot 10^{-9}]$	N
$\hat{\theta}^1_{\text{rrn}}$	ARN stables	$3.396 \cdot 10^{-6}$	$[2.507 \cdot 10^{-6}, 4.286 \cdot 10^{-6}]$	N
$\hat{\theta}^2_{\text{rrn}}$	ARN stables	$8.659 \cdot 10^{-6}$	$[6.752 \cdot 10^{-6}, 1.057 \cdot 10^{-5}]$	N

Pour cet exemple, huit césures sur onze reconstruites ne correspondent à rien. Et cela peut même en général être des proportions plus conséquentes. Cela montre bien que la reconstruction seule des césures n'est pas suffisante. Il s'agit de considérer la composition de leur pouvoir de séparation. Cela conduit à examiner les multicésures.

Le tableau suivant donne toutes les multicésures minimales produites par la chaine de traitement.

Multicésure	Césures	Correct
$\mathcal{M}_1$	$\{\hat{\theta}^1_{\text{Cya}}, \hat{\theta}^1_{\text{CRP}}, \hat{\theta}^1_{\text{GyrAB}}\}$	{N,N,Y}
$\mathcal{M}_2$	$\{\hat{\theta}^1_{\text{Cya}}, \hat{\theta}^2_{\text{Fis}}, \hat{\theta}^1_{\text{GyrAB}}\}$	{N,Y,Y}
$\mathcal{M}_3$	$\{\hat{\theta}^1_{\text{Cya}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^1_{\text{TopA}}\}$	{N,Y,N}
$\mathcal{M}_4$	$\{\hat{\theta}^1_{\text{Cya}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{rrn}}\}$	{N,Y,N}
$\mathcal{M}_5$	$\{\hat{\theta}^1_{\text{CRP}}, \hat{\theta}^2_{\text{CRP}}, \hat{\theta}^1_{\text{GyrAB}}\}$	{N,N,Y}
$\mathcal{M}_6$	$\{\hat{\theta}^1_{\text{CRP}}, \hat{\theta}^1_{\text{Fis}}, \hat{\theta}^1_{\text{GyrAB}}\}$	{N,Y,Y}
$\mathcal{M}_7$	$\{\hat{\theta}^1_{\text{CRP}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{GyrAB}}\}$	{N,Y,N}
$\mathcal{M}_8$	$\{\hat{\theta}^1_{\text{CRP}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{TopA}}\}$	{N,Y,N}
$\mathcal{M}_9$	$\{\hat{\theta}^1_{\text{CRP}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^1_{\text{rrn}}\}$	{N,Y,N}
$\mathcal{M}_{10}$	$\{\hat{\theta}^2_{\text{CRP}}, \hat{\theta}^2_{\text{Fis}}, \hat{\theta}^1_{\text{GyrAB}}\}$	{N,Y,Y}
$\mathcal{M}_{11}$	$\{\hat{\theta}^2_{\text{CRP}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^1_{\text{TopA}}\}$	{N,Y,N}
$\mathcal{M}_{12}$	$\{\hat{\theta}^2_{\text{CRP}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{rrn}}\}$	{N,Y,N}
$\mathcal{M}_{13}$	$\{\hat{\theta}^1_{\text{Fis}}, \hat{\theta}^2_{\text{Fis}}, \hat{\theta}^1_{\text{GyrAB}}\}$	{Y,Y,Y}
$\mathcal{M}_{14}$	$\{\hat{\theta}^1_{\text{Fis}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^1_{\text{TopA}}\}$	{Y,Y,N}
$\mathcal{M}_{15}$	$\{\hat{\theta}^1_{\text{Fis}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{rrn}}\}$	{Y,Y,N}
$\mathcal{M}_{16}$	$\{\hat{\theta}^2_{\text{Fis}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{GyrAB}}\}$	{Y,Y,N}
$\mathcal{M}_{17}$	$\{\hat{\theta}^2_{\text{Fis}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{TopA}}\}$	{Y,Y,N}
$\mathcal{M}_{18}$	$\{\hat{\theta}^2_{\text{Fis}}, \hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^1_{\text{rrn}}\}$	{Y,Y,N}
$\mathcal{M}_{19}$	$\{\hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{GyrAB}}, \hat{\theta}^1_{\text{TopA}}\}$	{Y,N,N}
$\mathcal{M}_{20}$	$\{\hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{GyrAB}}, \hat{\theta}^2_{\text{rrn}}\}$	{Y,N,N}
$\mathcal{M}_{21}$	$\{\hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^1_{\text{TopA}}, \hat{\theta}^2_{\text{TopA}}\}$	{Y,N,N}
$\mathcal{M}_{22}$	$\{\hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^1_{\text{TopA}}, \hat{\theta}^1_{\text{rrn}}\}$	{Y,N,N}
$\mathcal{M}_{23}$	$\{\hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^2_{\text{TopA}}, \hat{\theta}^2_{\text{rrn}}\}$	{Y,N,N}
$\mathcal{M}_{24}$	$\{\hat{\theta}^1_{\text{GyrAB}}, \hat{\theta}^1_{\text{rrn}}, \hat{\theta}^2_{\text{rrn}}\}$	{Y,N,N}

Il apparait que trois césures sont suffisantes pour expliquer l'ensemble des classes correspondant aux modes dynamiques inférés. Cette fois-ci, la chaine propose vingt quatre solutions. La treizième reconstitue correctement trois seuils. Mais elle semble un peu perdue au milieu de toutes ces réponses, dont quinze ne contiennent qu'un seuil correctement

reconstitué, huit en contenant deux. Pourtant, une information importante, et pertinente d'un point de vue biologique, est que $\hat{\theta}^1_{\text{GyrAB}}$ intervient dans 100% de ces solutions quand les autres seuils ne sont présents que dans 17% ou 25% des cas : il est donc un seuil qui *très certainement* correspond à une interaction dans le réseau de régulation génique.

Il n'est pas simple de formaliser ce dernier point. Le taux de présence d'un seuil dans les solutions parcimonieuses dépend en effet surtout du fait qu'il est le seul à expliquer la séparation d'au moins une classe. Ici, quelque soit la multicésure, $\hat{\theta}^1_{\text{GyrAB}}$ est une césure requise.

Chapitre 8

Évaluer la qualité des résultats

Dans le chapitre 4.5, nous avons vu que la chaine de traitement propose de résoudre le Problème 5 qui consiste à reconstruire des réseaux orientés d'interactions géniques tout en identifiant le modèle dynamique APM correspondant. Il s'agit ensuite de développer un critère pour l'évaluation des performances de cette procédure d'identification de réseau de régulation génique APM. Le résultat de la chaine de traitement consiste en un ensemble de modèles inférés qui contient toutes les solutions cohérentes avec les données de mesure et qui sont suffisantes pour expliquer le comportement dynamique inféré, dont nous avons une estimation.

Ce critère est inspiré des travaux sur l'évaluation des performances des algorithmes de reconstruction des réseaux en bioinformatique [57, 20, 117]. Nous utiliserons les notions de *rappel* et de *précision* pour caractériser la capacité des algorithmes à retrouver tous les interactions du modèle original et seulement celles-ci.

8.1 Réseau identifiable

Afin de pouvoir évaluer les performances d'un outil d'inférence, il s'agit tout d'abord de pouvoir construire un lien entre le résultat attendu et le résultat obtenu. On peut ensuite comparer et évaluer la fiabilité de l'approche.

Dans ce chapitre, nous allons nous pencher sur le "résultat attendu". Pour ce faire, il nous faut revenir sur la notion d'identifiabilité. De manière générale, une caractéristique est identifiable si, compte tenu des contraintes apportées par les informations disponibles pour la retrouver, il est possible effectivement de la retrouver. Si les données recueillies n'apportent pas d'information se rapportant à la caractéristique recherchée, toute procédure d'identification ne pourra pas mieux faire que de deviner. Parfois, la méthode d'obtention des observations fait que les données sont par nature non informatives : il s'agit d'un problème d'identifiabilité structurelle. Prenons par exemple le cas extrême où toute trajectoire n'est jamais amenée à passer un seuil, quelque soit la condition initiale : si une seule trajectoire est disponible, ce seuil de transition ne pourra jamais être identifié. Parfois les données ne sont pas informatives seulement du fait de la contingence de l'expérience : on parlera alors d'un problème d'identifiabilité pratique. Prenons par exemple le cas où un seuil de transition du modèle "biologique" est franchi, mais cela n'entraine pas de changement dans la dynamique du système : les données collectées autour de ce franchissement ne permettent aucunement de le reconstituer. Dans les deux cas, si l'information n'est pas là, on ne peut pas espérer identifier la caractéristique recherchée, quelque soit la méthode choisie.

Dans notre cas, nous faisons l'hypothèse que nous disposons d'une trajectoire (c'est-à-dire d'une série de mesures issues d'une expérience biologique où le système vivant a été amené d'un état à un autre du fait d'une perturbation, d'un stress). Disons que cette trajectoire correspond à un modèle APM élaboré, incluant toutes les interactions géniques envisageables. Il est possible que le parcours de cette trajectoire passe dans différents domaines de régulation (donc franchisse un certain nombre de seuils), et que plusieurs modes dynamiques soient inférables. Cependant, cette observation ne concerne pas toujours tous les domaines de régulation où la dynamique est différente de celles des autres modes :

certains seuils ne sont pas franchis, et s'ils le sont, leur action régulatrice (c'est-à-dire celle du gène sur lequel intervient le seuil) est peut-être masquée par l'action de l'expression d'un autre gène qui dispose d'une priorité. Dans ce cas, les données n'apportent aucune information sur de tels seuils. Un modèle APM plus simple, n'incluant pas l'action de ces seuils non franchis ou masqués, conduira à la même trajectoire. Le modèle APM identifiable est donc le modèle le plus simple permettant d'obtenir la trajectoire observée dans le cadre de l'expérience ou plutôt, en ce qui concerne l'évaluation des performances de notre méthode, dans le cadre de la simulation.

Définition 15 (Seuil identifiable)**.** *Nous dirons qu'un seuil d'un modèle APM de réseau de régulation génique est* identifiable *si, étant donné une trajectoire, il est traversé au moins une fois, et si, étant traversé, son franchissement modifie le taux de synthèse ou le taux de dégradation d'au moins un élément observé.*

Le but de notre procédure est bien-sûr d'inférer ces seuils identifiables. Il n'est pas difficile d'obtenir la liste des seuils identifiables à partir d'une simulation APM (telle que proposée dans le Chapitre 4.2 page 50). Le principe décrit en 4.2.2 introduisait la notion d'intervalle sans transition, correspondant à l'intervalle temporel pendant lequel la trajectoire reste à l'intérieur d'un domaine de régulation. À la fin de cet intervalle, un seuil au moins est franchi. Si le seuil franchi fait qu'un des taux de synthèse ou de dégradation est modifié (on pourra généraliser en disant que le point focal de part et d'autre est différent), alors, il s'agit d'un seuil identifiable.

Exemple 27. *Pour l'exemple de la Figure 4.7 (p. 56), les seuils identifiables apparaissent en magenta : la cinétique de toutes les concentrations moléculaires sont différentes de part et d'autre de chaque ligne rouge qui représente une vraie transition.*

De cette manière, il est possible de lister uniquement les seuils nécessaires et suffisants pour expliquer la trajectoire. On obtient ainsi le modèle le plus simple que nous pourrons comparer avec les modèles inférés.

Définition 16 (Réseau identifiable)**.** *Le réseau identifiable est le réseau de régulation génique composé par les actions régulatrices qui surviennent au niveau des seuils identifiables, étant donnée une trajectoire dans l'espace des concentrations des éléments observés.*

Nous avons vu que l'action régulatrice d'un gène sur l'expression d'un autre pouvait concerner soit la synthèse (activation ou inhibition de l'expression) soit la dégradation (activation ou inhibition de la disparition du produit de l'expression). Prenons par exemple l'équation suivante, inspirée de l'étude sur *E.coli* décrite au Chapitre 7 : $\dot{x}_{rrn} = 1.12s^{+}(x_{Fis}, \theta_{Fis}) - 1.5x_{rrn}$. Lors du passage d'une phase stationnaire (famine) à une phase exponentielle (croissance de la colonie bactérienne), et inversement, le seuil θ_{Fis} est franchi : le taux de synthèse des ARN stables, produits de l'expression du gène rrn, est modifié. De plus, on peut relever que lorsque la concentration de Fis augmente, le taux de synthèse d'ARN stables augmente aussi, et inversement, la disparition de Fis fait que le taux de synthèse d'ARN stables décroit : de sorte que Fis agit comme un activateur de la synthèse du gène rrn. Nous pouvons donc faire le lien entre la fonctionnelle s^{+} et la modification des paramètres inférés de la dynamique modifiée lors du parcours de la trajectoire.

Pour décrire la procédure basée sur ce principe afin de lister l'action des seuils identifiables, rappelons qu'un segment S_j^i pour une molécule i est un ensemble de points consécutifs $x_i(k)$ pour lesquels les paramètres de la dynamique sont identiques et sont $\kappa^{S_j^i}, \gamma^{S_j^i}$.

L'Algorithme 4 permet de déterminer, au même moment que la simulation, l'ensemble des actions pour les seuils identifiables. On notera cependant que ces actions ne sont pas nécessairement celles qui interviennent biologiquement : il se peut que plusieurs actions de gènes doivent être combinées (par exemple pour la formation d'un complexe) afin que l'action régulatrice se fasse ressentir. On obtient, à ce stade, le fait qu'un gène j semble réguler un gène i.

Algorithme 4. *Liste les actions des seuils identifiables*

pour toute molécule i **faire**
  soit $\mathcal{S}^{i*}$ l'ensemble des segments pour la molécule i
  pour deux segments consécutifs sur cette molécule S_a^i et S_b^i **faire**
    t est l'instant auquel intervient la transition entre S_a^i et S_b^i
    $\Delta\kappa = \kappa^{S_b^i} - \kappa^{S_a^i}$ est l'écart de taux de synthèse entre les segments
    $\Delta\gamma$ est l'écart de taux de dégradation entre les segments pour la molécule i
    pour tout seuil qui est passé à t **faire**
      soit j la direction du seuil
      $trend = \dot{x}_j(t)$ (seul le signe de l'évolution compte)
      $action_\kappa = trend \times \Delta\kappa$
      $action_\gamma = trend \times \Delta\gamma$
      si** $action_\kappa > 0$ **alors
        la molécule j se comporte comme un activateur de la synthèse de la molécule i
      autrement si** $action_\kappa < 0$ **alors
        la molécule j se comporte comme un inhibiteur de la synthèse de la molécule i
      fin si
      si** $action_\gamma > 0$ **alors
        la molécule j se comporte comme un activateur de la dégradation de la molécule i
      autrement si** $action_\gamma < 0$ **alors
        la molécule j se comporte comme un inhibiteur de la dégradation de la molécule i
      fin si
    fin pour
  fin pour
fin pour

Reconstruire les actions correspondant aux césures ne se fait pas tout à fait de la même manière même si le principe est bien-sûr équivalent. Il s'agit d'utiliser à la fois la caractérisation de la césure et les paramètres des modes dynamiques inférés pour la trajectoire bruitée. Même dans le cas où l'effet du bruit peut être négligé, en ne considérant qu'une seule transition détectée pour un seul élément m_i, il est possible d'envisager plusieurs césures : nous nous attendons à ce qu'au moins une des césures corresponde au seuil identifiable qui a causé la modification de la cinétique. Nous expliquerons dans la section suivante cette notion de correspondance.

Quand une trajectoire passe d'une classe (correspondant à un mode dynamique) à une autre (pour lesquelles les paramètres dynamiques sont donc différents), au moins une césure est censée être croisée. Appelons ces classes $\mathcal{F}_a$ et $\mathcal{F}_b$. Souvenons nous que seules les césures maximales sont conservées : cela signifie que la procédure de reconstruction des césures

prend en compte toutes les classes et leur géométrie. Nous ferons de plus l'hypothèse que l'inférence des paramètres dynamiques est correcte. Soit θ la césure séparant les classes : $\mathcal{F}_a \overset{\theta}{\curlyvee} \mathcal{F}_b$. La direction $dir(\theta)$ de la césure est l'index du gène dont le produit de l'expression sera à l'origine de la commande, car le seuil intervient sur le produit de son expression. Comme cela a été introduit en 4.4.1, un élément m_i pour lequel les paramètres de la cinétique varient ($\kappa_i^{(a)} \neq \kappa_i^{(b)}$ ou $\gamma_i^{(a)} \neq \gamma_i^{(b)}$) correspond au gène qui subit l'asservissement, car la synthèse ou la dégradation est affectée par le franchissement du seuil. Il peut y avoir plusieurs gènes commandés simultanément ou plusieurs actions. L'Algorithme 5 permet de déterminer, pour une multicésure minimale, l'ensemble des actions pour les seuils identifiés que révèlent les césures.

Algorithme 5. *Liste les actions des seuils identifiés*

pour *toute césure* θ ***faire***
  pour *deux points classés consécutifs* $(x[k], x[k+1])$ ***faire***
    soient a *et* b *l'index des classes de ces points*
    si $a \neq b$ ***alors***
      si $(a, b) \in S(\theta)$ ***alors***
        $trend = x[k+1]_{dir(\theta)} - x[k]_{dir(\theta)}$
        pour *toute molécule* ***faire***
          soit i *l'index de la molécule*
          $\Delta\kappa = \kappa_i^{(b)} - \kappa_i^{(a)}$
          $\Delta\gamma = \gamma_i^{(b)} - \gamma_i^{(a)}$
          $action_\kappa = trend \times \Delta\kappa$
          $action_\gamma = trend \times \Delta\gamma$
          si $action_\kappa > 0$ ***alors***
            la molécule $dir(\theta)$ *se comporte comme un activateur de la synthèse de la molécule* i
          autrement si $action_\kappa < 0$ ***alors***
            la molécule $dir(\theta)$ *se comporte comme un inhibiteur de la synthèse de la molécule* i
          fin si
          si $action_\gamma > 0$ ***alors***
            la molécule $dir(\theta)$ *se comporte comme un activateur de la dégradation de la molécule* i
          autrement si $action_\gamma < 0$ ***alors***
            la molécule $dir(\theta)$ *se comporte comme un inhibiteur de la dégradation de la molécule* i
          fin si
        fin pour
      fin si
    fin si
  fin pour
fin pour

8.2 Correspondance entre des seuils identifiables et des césures

Nous appelons $\mathcal{T}$ l'ensemble des seuils identifiables étant donnée une trajectoire. $\mathcal{T}$ se compose de tous les seuils identifiables pour chacune des n variables. Un parax-hyperplan θ^a représente un seuil identifiable de $\mathcal{T}$ si sa direction est l'index de l'élément sur lequel intervient le seuil, et si son niveau zéro est égal à la valeur du seuil. La collection de tous ces hyperplans est nommée Θ_{modele}.

$$\Theta_{modele} = \{\theta^a : dir(\theta^a) = i, \mathrm{Z}(\theta^a) = \tau \text{ avec } \tau \in \mathcal{T} \text{ seuil sur } m_i\}. \tag{8.1}$$

Exemple 28. *En reprenant le cas de la bactérie E. coli donné au chapitre 7, en considérant la trajectoire représentée sur la Figure 7.5, on a :*

$$\Theta_{modele} = \{\theta^1_{\mathrm{Fis}}, \theta^3_{\mathrm{Fis}}, \theta^4_{\mathrm{Fis}}, \theta^1_{\mathrm{GyrAB}}\} \ .$$

Seuls quatre seuils parmi les quatorze du modèle initial sont identifiables.

Soit $\mathcal{M}^*_{min}$ l'ensemble des multicésures minimales résultant de la chaine de traitement. Soit θ^e une césure de $\mathcal{M} \in \mathcal{M}^*_{min}$. Le niveau zéro de θ^e est un seuil *identifié*. Nous assimilerons une césure et un seuil identifié si cela ne cause pas d'ambigüité.

Notre objectif est ici d'expliciter la notion de correspondance entre tous les parax-hyperplans $\theta^a \in \Theta_{modele}$ et toutes les césures $\theta^e \in \mathcal{M}$. Il y a deux aspects à prendre en considération. Le premier est la valeur numérique du seuil : or, du fait de l'échantillonnage et du fait du bruit, les valeurs ne coïncident pas. Le deuxième aspect concerne les actions régulatrices : on devrait retrouver les actions d'un seuil identifiable (au moins) parmi les actions d'un seuil identifié correspondant.

Afin de résoudre le premier point, relevons qu'il doit exister une proximité, faute d'être stricte, entre le seuil identifié et le seuil identifiable. Une difficulté supplémentaire est liée au fait que la chaine de traitement peut conduire à un résultat où un seuil n'est pas identifié, ou, au contraire, à des seuils identifiés en surnombre. Or rappelons qu'une césure n'est en fait que le représentant d'une classe d'équivalence de plusieurs parax-hyperplans ayant un même pouvoir de séparation (il va de soi que nous raisonnons dans une même direction). Afin de caractériser la notion de proximité entre seuil identifiable et seuil identifié, nous introduisons la distance définie par :

$$\forall \theta^a \in \Theta_{modele}, \forall \theta^e \in \mathcal{M}^{dir(\theta^a)}, dist(\theta^a, \theta^e) = d(\mathrm{Z}(\theta^a), I_{eq}(\theta^e)) \tag{8.2}$$

avec la distance métrique usuelle $d(p, S) = \min_{p' \in S} \|p - p'\|$.

Définition 17. *Pour $\theta^a \in \Theta_{modele}$, le* seuil identifié le plus proche *de θ^a est*

$$\arg dist(\theta^a, \mathcal{M}^{dir(\theta^a)}) = \arg \min_{\theta^e \in \mathcal{M}^{dir(\theta^a)}} dist(\theta^a, \theta^e).$$

Pour $\theta^e \in \mathcal{M}$, le seuil identifiable le plus proche *de θ^e est*

$$\arg dist(\Theta^{dir(\theta^e)}_{modele}, \theta^e) = \arg \min_{\theta^a \in \Theta^{dir(\theta^e)}_{modele}} dist(\theta^a, \theta^e).$$

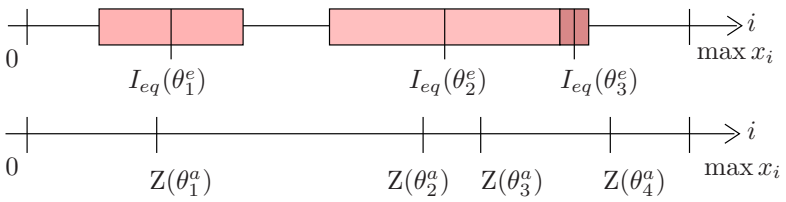


FIGURE 8.1 – Proximité entre seuils identifiables et seuils identifiés. Avec une distance nulle, θ_1^a est le seuil identifiable le plus proche de θ_1^e et, avec une distance non nulle, θ_4^a est le seuil identifiable le plus proche de θ_3^e. Inversement, θ_1^e est le seuil identifié le plus proche de θ_1^a, et il en va de même pour θ_3^e et θ_4^a. De plus, θ_2^e a pour seuils identifiables les plus proches θ_2^a et θ_3^a.

À vrai dire, cette définition est un peu abusive, car il n'existe pas d'unicité : il est possible que la condition de minimalité soit vérifiée par plusieurs éléments. La Figure 8.1 donne un exemple.

En utilisant cette notion de proximité et en comparant les actions issues des Algorithmes 4 et 5, nous pouvons définir la notion de correspondance entre un seuil identifiable et une césure.

Définition 18 (Correspondance)**.** *Un seuil identifiable θ^a et un seuil identifié θ^e ayant une même direction correspondent si $dist(\theta^a, \theta^e) = 0$ et si on retrouve toutes les actions de θ^a parmi celles de θ^e. On notera dans ce cas : $\theta^a \rightleftharpoons \theta^e$.*

Nous allons utiliser cette notion pour évaluer les performances de la procédure d'identification.

8.3 Mesures de performance

De manière à caractériser la qualité d'une multicésure minimale générée par la chaine de traitement étant donnés un modèle initial connu et une trajectoire, nous introduisons la notion de *rappel* et de *précision*. Ces mesures proviennent de la communauté de la Recherche d'informations où l'on désire, par exemple, évaluer des moteurs de recherche au vue des besoins des utilisateurs et de leur requête. Les notions de rappel et de précision ont été introduit dans [24] sous le nom de "rapport de rappel" et "rapport de pertinence". Ils se basent sur la notion de pertinence : étant donné un corpus d'éléments, un moteur de recherche doit répondre à une requête en fournissant la liste de tous les éléments retrouvés, qui pourront être pertinents ou ne pas l'être. Dans les expériences de Cranfield (voir [25]), à la fois le rappel et la précision sont utilisés pour évaluer l'efficacité de la recherche.

Les définitions classiques de rappel et précision sont :

$$\text{Rappel} = \frac{\#\text{ retrouvé pertinent}}{\#\text{ pertinent}}, \tag{8.3}$$

$$\text{Précision} = \frac{\#\text{ pertinent retrouvé}}{\#\text{ retrouvé}}. \tag{8.4}$$

Dans [121], il est souligné que le paradigme de Cranfield se base sur trois hypothèses principales. Premièrement, la pertinence peut être approchée par une similitude de thématique : la pertinence d'un élément est alors indépendant de la pertinence de tout autre

élément. Cela nous conforte à utiliser l'approximation liée à notre notion de correspondance. Deuxièmement, un ensemble d'évaluations de la proximité thématique est valable pour toute la communauté des utilisateurs (ou la représente correctement) : dans notre cas, la notion de correspondance est mathématique. Troisièmement, les listes d'éléments pertinents pour chaque thématique sont complètes : dans la mesure où les seuils identifiables sont connus et fixés pour une expérience donnée (les conditions initiales et le niveau de bruit étant posés), cette hypothèse est remplie.

Dans notre cas, nous remplaçons donc la notion de pertinence par la notion de correspondance. Les éléments retrouvés sont les césures identifiées de $\mathcal{M} \in \mathcal{M}^*_{min}$. La mesure de précision exprime ainsi la proportion de seuils identifiés d'une multicésure minimale qui correspondent à des seuils identifiables du modèle original, tandis que la mesure de rappel indique la proportion de seuils identifiables qui correspondent à des seuils identifiés parmi ces seuils identifiés. Nous définirons donc la mesure ρ de rappel et π de précision de la manière suivante :

$$\rho(\Theta_{modele}, \mathcal{M}) = \frac{card(\{\theta^e \in \mathcal{M} : \exists \theta^a \in \Theta_{modele}, \theta^a \rightleftharpoons \theta^e\})}{|\Theta_{model}|}, \tag{8.5}$$

$$\pi(\Theta_{modele}, \mathcal{M}) = \frac{card(\{\theta^e \in \mathcal{M} : \exists \theta^a \in \Theta_{modele}, \theta^a \rightleftharpoons \theta^e\})}{|\mathcal{M}|}. \tag{8.6}$$

Exemple 29. *Pour l'exemple de la Figure 8.1, en faisant l'hypothèse que les actions sont correctement reconstruites* ($\theta^a_1 \rightleftharpoons \theta^e_1$, $\theta^a_2 \rightleftharpoons \theta^e_2$, $\theta^a_3 \rightleftharpoons \theta^e_2$), $\rho(\Theta^i_{modele}, \mathcal{M}^i) = \frac{2}{4} = 0,5$ *et* $\pi(\Theta^i_{modele}, \mathcal{M}^i) = \frac{2}{3} \approx 0,67$.

Une manière simple d'associer une multicésure à une mesure de performance est de faire la moyenne harmonique du rappel et de la précision, qui résulte en ce qui est appelé la mesure F :

$$F = \frac{2\,\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}. \tag{8.7}$$

La mesure F, comme celle de rappel et de précision, varie de 0 à 1.

Exemple 30. *En reprenant le cas de la bactérie E. coli donné au chapitre 7, en considérant la trajectoire représentée sur la Figure 7.5, la meilleure mesure F obtenue est* $F(\Theta_{modele}, \mathcal{M}_{13}) = \frac{6}{7} \approx 0,86$.

L'utilisation des mesures de rappel et de précision pour évaluer les performances des méthodes concernant l'inférence de réseaux de régulation génique est devenue courante dans la littérature. Deux approches sont utilisées. La définition inspirée de contexte de recherche d'information peut être aussi trouvée dans [82, 20] : pour des réseaux booléens et des réseaux bayésiens dynamiques, les nœuds des graphes sont considérés comme étant pertinents ou retrouvés. Dans une perspective utilisant la terminologie de la classification binaire, une autre définition utilise la notion de vrai positif, faux négatif et faux positif (voir [57, 117] pour des réseaux bayésiens, où les arcs des graphes sont examinés comme pertinents et retrouvés). Dans tous ces cas, les éléments inférés appartiennent à une liste qui est connue depuis le départ, et le but est de déterminer si ces éléments sont bien ceux du modèle original. Dans notre cas, les seuils étant à valeur réelle, ils appartiennent à un espace infini.

Exemple 31. *Nous allons maintenant exploiter le modèle de la bactérie E. coli introduit au chapitre 7 pour illustrer les performances de la chaine de traitement [87]. Plusieurs*

expériences ont été effectuées sur la version du logiciel développée à l'Université de Pavie[1], *en variant le niveau de bruit* $RSB \in \{1e{-}3, 5e{-}3, 1e{-}2\}$ *et le pas d'échantillonnage* $T \in \{5\ min, 10\ min\}$. *Pour chaque combinaison,* 100 *trajectoires ont été simulées puis traitées avec les paramètres* $\alpha = 0,01$ *et* $N_{\mathcal{S}} = 4$, *nous conduisant à des collections de multicésures parcimonieuses différentes. Pour chaque trajectoire, nous avons relevé la meilleure des mesures* F *sur l'ensemble des multicésures parcimonieuses, ainsi que la moyenne de ces mesures* F. *L'histogramme de ces résultats est présenté sur la Figure 8.2.*

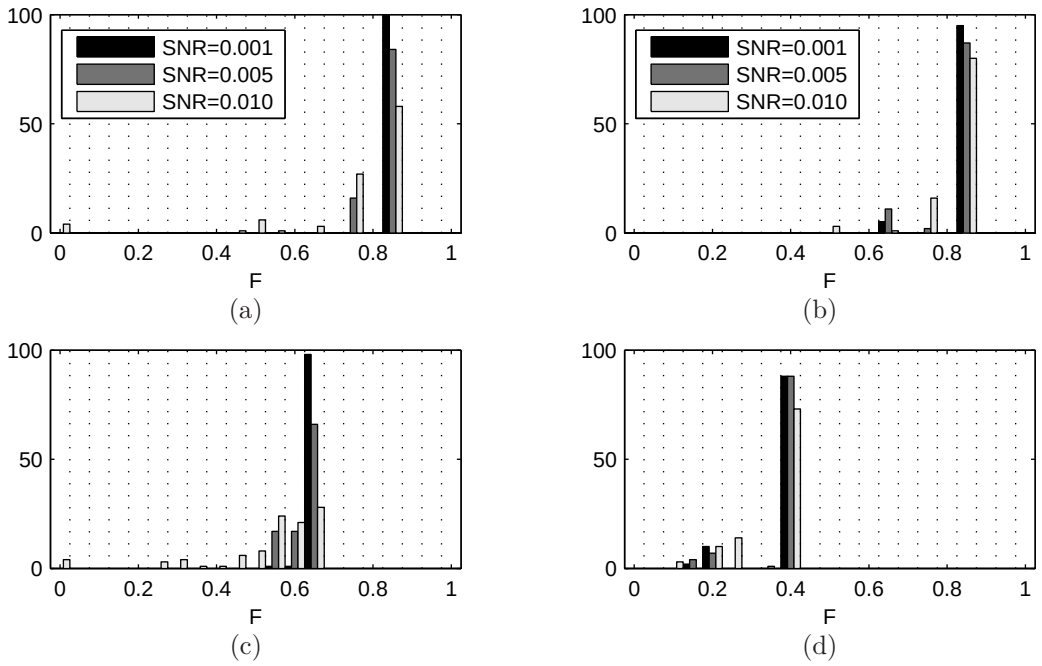


FIGURE 8.2 – Distributions pour divers RSB de la mesure F (a)-(b) maximale et (c)-(d) moyenne pour plusieurs données de mesure simulées avec un pas d'échantillonnage de $T = 5$ min pour (a) et (c), et $T = 10$ min pour (b) et (d).

Ces résultats empiriques montrent que la chaine d'identification se comporte bien même dans le cas le plus défavorable où $T = 10$ *min et* $RSB = 0,01$. *Dans tous les cas, au moins* 85% *des expériences conduisent à une mesure* F *maximale supérieure à* $0,75$. *Ce résultat doit être comparé avec le nombre relativement faible d'hypothèses produites : la distribution du nombre de multicésures parcimonieuses, montrée sur la Figure 8.3, souligne que la chaine de traitement produit, dans la plupart des expériences, moins de* 10 *multicésures parcimonieuses pour* $T = 5$ *min et moins de* 30 *autrement. La présence de cas pathologiques est principalement causée par des trajectoires conduisant à des classifications multiples. La mesure* F *moyenne pour* $T = 10$ *min chute à 60% de ce qui est obtenu pour* $T = 5$ *min : davantage de multicésures parcimonieuses sont générées, dans la mesure où une information plus faible (moins de points) conduit à rendre davantage d'hypothèses compatibles avec les données ; de sorte que des césures éronnées sont introduites qui font baisser les mesures de performance.*

1. Cette version n'utilise pas la notion d'action pour le concept de correspondance.

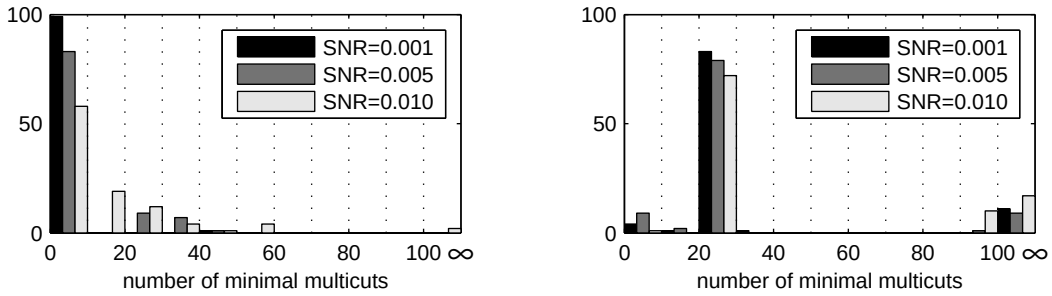


FIGURE 8.3 – Distributions pour divers RSB ("SNR" sur la figure) du nombre de multicésures parcimonieuses pour plusieurs expériences avec un pas d'échantillonnage de (a) $T = 5$ min et (b) $T = 10$ min.

Quatrième partie

Discussion et conclusions

Chapitre 9

Conclusions et perspectives

9.1 Discussion

La modélisation des réseaux de régulation génique, et en particulier, le développement de techniques d'identification permettant d'automatiser cette modélisation, sont en plein développement. Elles arrivent au même moment que les technologies de mesure de l'expression génique à fort débit s'imposent conduisant à une petite révolution en ce qui concerne la compréhension du développement et des maladies. On ne s'intéresse que peu, aujourd'hui, à l'étude d'un seul gène : les fonctions des gènes peuvent seulement être envisagées dans le contexte d'un réseau de gènes, de protéines, et de petites molécules. Ces réseaux sont pris dans un environnement assez confiné, cellulaire (forte compartimentation dans la cellule avec éventuellement des appareils cellulaires, puis membrane plasmique) et extra-cellulaire (matrice cellulaire, puis tissu, puis organisme) vis-à-vis duquel ils s'adaptent en permanence. On ne peut évaluer la fonction d'un gène que dans un environnement fonctionnel de gènes.

Prenons un exemple assez marquant. En se référant à [15, 16, 122, 61] et au travaux de M. J. Bissell (du Lawrence Berkeley National Laboratory) en général, on peut décrire l'expérience suivante. On dispose de cellules malades d'un organisme atteint d'un cancer, et aussi de cellules d'un organisme sain. Pour faire simple, on va dire qu'il est possible de séparer les cellules de leur matrice. Dans un cas, on transfère les cellules saines dans les matrices malades : assez rapidement, ces cellules présentent les caractéristiques des cellules cancéreuses. De sorte que la machinerie cellulaire qui fonctionnait "correctement" se met à dévier de son état usuel souhaitable. Inversement, si l'on met une cellule cancéreuse dans une matrice saine, la cellule se "guérit" et les caractéristiques du cancer disparaissent. Il y a de la réversibilité. Cela montre que la machinerie cellulaire avait bien toutes les bonnes propriétés : il n'y avait pas de "manque" ou de "destruction" (ce n'est pas comme si le moteur cassait). Les fonctions ne sont pas, dans ce cas du moins, exprimées comme il faut ou dans le bon ordre ou dans les bonnes proportions, à moins que de nouvelles fonctions soient révélées qui perturbent la machinerie cellulaire.

La reconstruction de modèles prédictifs décrivant quantitativement les interactions entre les produits d'expression des gènes et les gènes est donc cruciale. Pour procéder à cette identification, plusieurs approches ont été suggérées et utilisées dans des contextes spécifiques. La plupart d'entre elles concernent le cas de l'exploitation de petites perturbations. Les petites perturbations permettent de garder le système étudié proche de son état d'équilibre, en se ramenant à un formalisme linéaire que l'on peut traiter avec des données assez imprécises et en faible quantité. Pour des perturbations plus grandes, conduisant éventuellement à des transitions entre des états stationnaires différents, l'inférence du réseau de régulation doit s'appuyer sur des modèles non-linéaires. Dès lors, la qualité des données ainsi que la densité de l'échantillonnage doivent être notablement accrues. De nombreux outils de traitement du signal existent dans des domaines étrangers à la biologie. Cependant, les conditions expérimentales spécifiques contraignent encore largement ces outils standards et nécessitent leur adaptation. Ainsi, les mesures sont effectuées sur des populations de cellules : le progrès de la microscopie en temps réel, notamment, laisse entrevoir des évolutions prometteuses dans ce sens [116, 69]. D'autre part, une des manières

de compenser les défauts de la mesure est de solliciter de manière variée un système : on améliore sa connaissance, obtenant davantage d'information. Or, l'éventail des perturbations imposables n'est pas en général très grand, sans compter qu'il est très difficile de contraindre les états stables initiaux sans modifier le système lui-même : le panel d'expériences possibles est donc limité. Dans de telles conditions (points de mesure en faible quantité, variabilité de sollicitations du système très réduite), les traitements statistiques deviennent très délicats.

Dans le travail présenté dans ce manuscrit, nous nous sommes concentrés sur ce qui est en notre sens le formalisme (APM) le plus simple permettant de décrire de manière déterministe les évolutions d'un réseau de régulation génique pour des larges perturbations ou des changements d'état. L'approche APM est très intéressante dans la mesure où elle relève que la partie la plus informative de la régulation se trouve dans deux positions, l'une basse (absence de contrôle) et l'autre haute (saturation de l'effet de l'action du régulateur). Notre méthode d'identification s'appuie de manière réaliste sur des données de grande qualité qu'il est possible d'obtenir (pour l'exemple de *E. coli*, ce travail a été réalisé avec succès dans le laboratoire de Hans Geiselmann à l'Université de Grenoble.) Notre méthode d'identification permet de reconstruire de façon rigoureuse des interactions géniques ayant un sens biologique très caractérisé : on relève l'action de l'expression d'un gène sur l'expression d'un autre en terme de synthèse ou de dégradation. De plus, ce sont les dynamiques de ces actions qui sont quantifiées, et cela apporte un pouvoir prédictif jamais égalé. À notre connaissance, le traitement proposé ici est une réelle innovation.

Du point de vue du traitement de données d'expression génique, il existe plusieurs points critiques. Tout d'abord, nous avons appliqué cette méthode sur des systèmes (d'inspiration biologique) d'une dizaine de gènes. Cela est très peu comparé aux dizaines de milliers de gènes que l'on peut mesurer en parallèle sur les meilleures puces à ADN. Il est vrai que notre approche souffre d'explosion combinatoire à divers niveaux. Cependant, il est possible d'envisager de traiter des exemples plus grands : ce qui est déterminant est moins le nombre de gènes que le nombre de transitions entre modes dynamiques, et le lien entre ces grandeurs n'est pas trivial (bien-sûr, on s'attend à ce que s'il y a davantage de gènes, il y aura davantage d'actions régulatrices, donc davantage de transitions). De plus, le parti pris de notre approche est de se concentrer sur des réseaux de faible dimensionnalité en les étudiants autour de fonctions et de caractéristiques biologiques restreintes précises, cela afin de ne pas perdre le pouvoir explicatif biologique ultimement recherché. Enfin, si les limites de traitabilité se voyaient atteintes, il serait assez facilement faisable de construire des approches sous-optimales, considérant un éventail de solutions possibles au lieu de toutes les solutions obtenues par la présente méthode. Une avancée dans ce sens est présentée en 9.2.1.

Un deuxième point est que les données de série temporelle ne sont pas toujours aussi denses et précises que celles obtenues à l'aide de gènes rapporteurs. La technique de mesure la plus populaire est pour le moment la puce à ADN, qui permet d'obtenir pour un coût déjà très fort quelques dizaines de points, et la précision n'est pas bonne. Ces points sont de très grande dimension, mais l'intérêt majeur de cette approche est qu'il est juste nécessaire d'extraire les cellules d'intérêt, quand la technique à base de gènes rapporteurs implique une fenêtre de préparation qui peut difficilement être réduite en dessous de plusieurs mois pour les organismes les mieux étudiés. Ceci est le point faible majeur de notre approche : se basant sur une inférence dynamique par morceaux, l'effet du bruit (ou plus particulièrement l'imprécision de la mesure) combiné au manque de points peut être dramatique. Des librairies de rapporteurs [125] sont toutefois en train d'être mises en place, ce qui permettra

d'accélérer le développement de l'approche expérimentale que nous avons privilégiée.

Un troisième point concerne une hypothèse fondamentale du modèle APM : l'existence de seuils de régulation, impliquant des transitions franches entre les modes dynamiques. Or, pour des données de mesure non synthétiques, cette hypothèse n'est pas toujours strictement vérifiée. En prenant l'approximation un peu meilleure d'une fonction de régulation en forme sigmoïdale, on se rend compte que la transition n'est franche que si la concentration de l'élément régulant passe rapidement la zone intermédiaire autour du seuil. Dans le cas contraire, la dynamique de l'expression du gène régulé va mettre un certain temps (plusieurs points de mesure) avant de redevenir constante. Dans ce cas, on ne peut pas obtenir de segments qui s'enchainent les uns après les autres : il existe des intervalles de temps pour lesquels l'hypothèse d'une dynamique constante par morceaux est mise en défaut. Cela entraine des erreurs durant la segmentation, et ces erreurs de sur-segmentation sont propagées.

Maintenant, en considérant la perspective de l'identification des systèmes, il existe un certain nombre de sujets délicats qu'il s'agira de mieux étudier dans le futur afin d'améliorer l'identification des modèles APM de réseaux de régulation génique. Premièrement, l'incorporation de connaissances préalables doit permettre d'améliorer notablement la reconstruction du modèle. Cet aspect est aussi très important lorsque l'on souhaite mettre en cohérence des résultats obtenus avec des méthodes différentes, ou dans des conditions différentes. En ce qui concerne l'inférence, notre méthode n'a pas été spécifiquement pensée pour incorporer des données *a priori* concernant la dynamique. Il s'agira alors d'associer à l'inférence une étape de décision (par test statistique) permettant de rattacher ou non les points étudiés à une dynamique donnée. Par contre toute la méthode peut très naturellement intégrer une information préalable sur les seuils de transition : la segmentation devient évidente, et l'on peut se ramener à une unique césure, qu'il faut prendre comme requise. Éventuellement, il serait possible de falsifier l'information préalable sur le seuil si son action est connue au préalable et que l'action inférée sur la base des dynamiques des modes quitté et rejoint rentrent en contradiction. La question de la falsification des connaissances préalables n'est cependant pas triviale, surtout si l'on se place dans un compromis entre falsification et validation du (des) résultat(s). On disposera le plus souvent d'information de type "tel gène semble interagir avec tel autre" avec, éventuellement, les conditions dans lesquelles cette interaction a été relevée : on pourrait alors se servir de ces connaissances préalables pour sélectionner les solutions obtenues en ne préservant que celles en cohérence avec ce type d'information, dans les cas où la cohérence stricte peut être établie, ce qui est délicat. Le formalisme qualitatif de vérification présenté dans [10] pourrait être mis à contribution, ainsi que des formalismes tels que celui présenté par [105]. Enfin, et cela est encore plus difficile à exploiter, on dispose parfois d'un niveau de confiance à apporter à une connaissance préalable : il devrait alors être envisagé de donner une mesure de confiance à nos solutions, ce qui permettrait d'une part de les classer, et ce qui permettrait d'autre part de construire des traitements exploitant les informations *a priori* pour raffiner notre ensemble solution.

Deuxièmement, nous avons écarté la question des observations manquantes : les gènes en interaction ne peuvent être que ceux mesurés. Cet aspect est réellement difficile car il souligne la nécessité d'un travail préalable de détermination du système (en utilisant la littérature, les bases de données pour plusieurs organismes, la bioinformatique ou le résultat de méthodes d'identification de réseaux - dans le cadre de faibles perturbations - qui exploitent bien les données de mesure à large débit). En effet, comme nous cherchons à introduire le moins d'hypothèses possibles, nous n'envisageons pas l'influence de gènes

n'appartenant pas au système étudié. Ajouter à cela, les réseaux d'interactions géniques sont une projection sur le plan génétique de nombreux réseaux d'interactions - géniques, protéiques, métaboliques. Notre approche n'apporte aucun élément permettant de dire si un ensemble de gènes doit être inclus dans le système (ni comment les choisir). Le fait de devoir combiner des connaissances et des données de mesure pour reconstruire des réseaux d'interaction ne semble pourtant pas choquant, et c'est d'ailleurs l'approche envisagée par l'analyse [58].

Troisièmement, cette approche ne s'intéresse pas à la modularité du réseau inféré. L'intérêt de cet aspect est qu'il permet de réduire la dimensionnalité du système en se centrant sur un ensemble de sous-systèmes couplés. Le traitement numérique en serait nettement plus simple. Il y a aussi un intérêt biologique : on pourrait comprendre comment des fonctions sont prises en charge par des petits groupes de gènes en interaction. La réduction de système est souvent un préalable pour l'identification des systèmes de grande dimension : on peut notamment se débarrasser, dans notre cas, des gènes qui n'agissent pas ou dont l'influence n'est pas notable. Or, pour les modèles APM, cette réduction est problématique : de faibles variations de concentration peuvent virtuellement avoir un effet régulateur. Nous n'avons jusqu'à présent pas envisagé une manière simple d'aborder ce sujet.

9.2 Perspectives

La méthode d'identification que nous avons présentée dans ce manuscrit est, à notre connaissance, la première approche permettant de reconstruire des réseaux APM de régulation génique à partir de données de mesure de concentration d'éléments cellulaires. La modélisation APM est un bon compromis permettant d'envisager l'étude des réseaux de régulation pour de fortes perturbations. En inférant les caractéristiques dynamiques des réseaux, notre méthode ouvrent des perspectives prédictives nouvelles.

Divers documents relatifs à la méthode d'identification, son application à l'exemple biologique de la réponse à un stress nutritionnel chez *E. coli* et l'évaluation de ses performances ont été publiés [37, 89, 36, 88, 87].

L'application de notre approche à des mesures réelles est en cours de traitement. Quelques difficultés doivent être levées, concernant notamment le problème de la segmentation lorsque les transitions ne sont pas franches.

Le principal défi pour les développements futurs concerne l'utilisation de cette approche avec des données de mesure moins fiables et moins denses que celles obtenues par les gènes rapporteurs. Il sera alors intéressant de proposer à la communauté des chercheurs une application traitant automatiquement des séries temporelles de mesures de concentrations des produits d'expression génique.

Dans cette direction, certains aspects connexes à notre travail et à ses aspects pratiques sont développés dans ce qui suit, qui mériteront une plus ample investigation.

9.2.1 Un manière rapide d'énumérer quelques solutions parcimonieuses.

Il s'agit ici de décrire une approche sous-optimale permettant de résoudre le Problème 6 (p.93) qui puisse être rapide. Cette approche nous a été suggérée par A. Agung Julius de l'Université de Pennsylvanie.

Commençons par introduire quelques nouvelles notations. Rappelons que $\mathcal{C}^*$ est l'ensemble des césures, et $U = \{u = (p, q) \in \{1, \ldots, s\}^2 : p < q\}$ est l'ensemble des paires d'index des classes à séparer : pour la suite, par abus de notation, nous considèrerons ces

ensembles comme étant des listes, et nous noterons $U(i)$ (respectivement $\mathcal{C}^*(i)$) le i-ième élément de la liste U (respectivement $\mathcal{C}^*$). Appelons $\mathbf{v} \in \{0,1\}^{|\mathcal{C}^*|}$ le vecteur de variables de choix : à tout $\mathbf{v}$ correspond un sous-ensemble $\mathcal{C}$ de césures de $\mathcal{C}^*$ dans la mesure où l'on aura :

$$\forall j \in \{1,\dots,|\mathcal{C}^*|\}, v_j = \begin{cases} 1 \text{ si } \theta = \mathcal{C}^*(j) \in \mathcal{C}, \\ 0 \text{ sinon.} \end{cases}$$

Soit encore la matrice $M_S \in \mathcal{M}_{|U||\mathcal{C}^*|}(\{0,1\})$, pour laquelle la composante m_{ij} à la i-ième ligne et à la j-ième colonne est définie par :

$$\forall j \in \{1,\dots,|\mathcal{C}^*|\}, \forall i \in \{1,\dots,|U|\}, m_{ij} = \begin{cases} 1 \text{ si } u = U(i) \in S(\mathcal{C}^*(j)), \\ 0 \text{ sinon.} \end{cases}$$

Cette matrice résume le pouvoir de séparation de toutes les césures : nous l'appellerons *matrice de séparation* associée à $\mathcal{C}^*$. On notera aussi $\mathbf{1}$ le vecteur unitaire de dimension $|U| = s(s-1)/2$. On utilisera aussi le symbole $\succeq$ pour l'inégalité terme à terme.

En utilisant cette notation, on peut avantageusement réécrire la Propriété 7 (p.80) sous la forme suivante.

Propriété 16. *Un ensemble $\mathcal{C}$ de césures, auquel correspond un vecteur de choix $\mathbf{v} \in \{0,1\}^{|\mathcal{C}^*|}$, est une multicésure si et seulement si $M_S\mathbf{v} \succeq \mathbf{1}$.*

Démonstration. Pour la ligne $i \in \{1,\dots,|U|\}$ de l'inégalité, on a $\sum_{j\in\{1,\dots,|\mathcal{C}^*|\}} v_j m_{ij} \geq 1$, ce qui signifie qu'il existe $\theta \in \mathcal{C}$ tel que $u = U(i) \in S(\theta)$. ■

En se basant sur le Problème 6 (p.93), nous pouvons formuler le problème d'optimisation suivant.

Problème 7. *Avec la matrice de séparation M_S associée à l'ensemble des césures maximales $Max_{\preceq}\mathcal{C}^*$,*

$$\begin{array}{c} \textit{minimiser } \|\mathbf{v}\|_{\ell_1} \\ \textit{sujet à } M_S\mathbf{v} \succeq \mathbf{1}. \end{array}$$

La résolution de ce problème conduit à une solution parcimonieuse, dont la cardinalité peut être utilisée pour initialiser l'Algorithme 3 (p.95) avec d'entrée de jeu la bonne valeur pour la variable *best*.

Ce type de problème d'optimisation, pour $\mathbf{v} \in \mathbb{R}$, est appelé *Programme Linéaire* [29, 70]. Il existe toute une gamme de méthodes très efficaces pour en faire la résolution, incluant la méthode simplex de Dantzig ou les méthodes du point-intérieur [18]. Des problèmes avec des centaines de variables et des milliers de contraintes peuvent être résolus en quelques secondes sur un ordinateur personnel.

Dans notre cas, nous avons en plus la contrainte que $\mathbf{v} \in \{0,1\}^{|\mathcal{C}^*|}$. Cela revient alors à un problème dit de *cardinalité convexe* [18] (dans ce cas, la cardinalité d'un vecteur réel est le nombre de ses composantes non nulles, qui est quasi-concave sur $\mathbb{R}^+$: $\text{card}(\mathbf{x}+\mathbf{y}) \geq \min\{\text{card}(\mathbf{x}), \text{card}(\mathbf{y})\}$ pour $\mathbf{x}, \mathbf{y} \succeq 0$). Ce problème est NP-difficile. Une heuristique simple, liée à l'utilisation de la norme ℓ_1, peut être utilisée. Elle est décrite comme une relaxation convexe : au lieu de $v_i \in \{0,1\}$, on prend $v_i \in [0,1]$. Alors, la valeur optimale obtenue par la résolution de ce problème est une borne inférieure pour le problème original.

Cette approche permet d'obtenir une solution sous-optimale instantanément, qui peut éventuellement servir à initialiser l'approche exhaustive décrite dans ce manuscrit (voir § 6.2.3) en accélérant ainsi notablement son traitement.

9.2.2 Stockage et échange des résultats

La méthode d'identification proposée est une chaine de traitement décrite au Chapitre 4. Chacun des modules de cette chaine (Figure 4.2) requiert des entrées que sont les résultats des modules précédents et les paramètres pour régler le processus de traitement, et proposent un certain nombre de résultats. De sorte que des arguments (des résultats antécédents ou des paramètres) sont passés d'un module à l'autre. Il est en général recommandé d'utiliser un fichier extérieur pour mémoriser les variables clés de certains modules ainsi que leurs résultats : de cette manière, il est possible d'avoir une trace complète du traitement qui a été effectué.

Dès que l'on commence à vouloir importer/exporter des données, pour les traiter ultérieurement, pour les partager, pour les sauvegarder, il est déterminant de se pencher sur la question du format. Un des aspects de "l'écriture" des données est de les garder lisibles : le format choisi doit être suffisamment précis pour que l'écriture reste cohérente, et qu'il soit éventuellement possible de contraindre la structure. De plus, les données "écrites" doivent être lisibles quelque soit le contexte : une feuille de données ne devrait pas dépendre d'un programme en particulier qui permettrait de retrouver ces données, ou de les générer. XML [1], un langage standardisé W3C, vise à être une méthode générale pour structurer les données : il s'agit d'un ensemble de règles qui facilitent la construction de fichiers de données telles que ces fichiers ne soient pas ambigus, généralisables, standards, indépendants de la plateforme et du système d'exploitation. Pour toutes ces raisons, XML a été largement utilisé pour l'échange de données par la communauté scientifique, et il est même l'un des formats les plus populaires.

XML est l'acronyme anglais de "*eXtensible Markup Language*" : les données étant explicitées (marquées), elles sont rendues indépendantes de leur contexte, mais leur structure, c'est-à-dire la manière dont elles s'articulent, s'encapsulent les unes dans les autres, est choisie par avance. De sorte que même dans un contexte particulier, écrire des données en un format XML aide à construire une structure qui ne soit pas ambigüe mais contrainte. Pour ce faire, un Schéma XML [2] (XSD), écrit en syntaxe XML, nous permet de spécifier la structure du format d'échange et de préciser le type des variables. Tout fichier XML doit être correctement formaté (respectant la syntaxe XML) et valide (respectant les règles imposées par le XSD).

Le schéma que nous avons adopté est décrit en Annexe C. Il a été construit pour suivre de manière intuitive la chaine de traitement proposée.

Un grand avantage d'un fichier XML correctement formaté et dont le XSD élimine toute ambigüité, est qu'il est très facilement convertible en une autre structure XML (respectant un autre XSD). Cela permet de faire des conversions ou des extractions faciles pour d'autres logiciels, de traitement, de comparaison, ou d'intégration. Le langage XSLT (W3C "eXtensible Style Language for Transformation"), ou XQuery, sont des langages de transformation qui permettent de faire ceci. Or, on notera que beaucoup de réseaux biologiques sont sauvegardés et mis à disposition en réseau au travers un langage XML standard pour la biologie des systèmes, SBML. Il est ainsi possible d'imaginer des ponts avec d'autres applications, soit fournissant des données à identifier, soit exploitant le résultat de notre méthode d'identification de réseau de régulation génique.

1. http ://www.w3.org/XML/
2. http ://www.w3.org/XML/Schema

Cinquième partie

Annexes

Annexe

A Ensembles ordonnés : rappels

Une relation d'ordre dans un ensemble E est une relation binaire $\mathcal{R}$ dans cet ensemble qui permet de comparer ses éléments entre eux de manière cohérente. Un ensemble muni d'une relation d'ordre est un ensemble ordonné ou tout simplement un ordre.

A.1 Relation d'ordre

Une relation d'ordre sur un ensemble E est une relation binaire sur E réflexive, transitive et antisymétrique :

– réflexive : $\mathcal{R}$ met tout élément en relation avec lui-même, c'est-à-dire $\forall x \in E,\ x\mathcal{R}x$;

– antisymétrique : les éléments distincts ne sont jamais en relation mutuelle, c'est-à-dire $\forall x \in E, \forall y \in E,\ [(x\mathcal{R}y) \wedge (y\mathcal{R}x)] \Rightarrow [x = y]$;

– transitive : deux éléments sont mis en relation dès qu'on peut "transiter" par un troisième, c'est-à-dire $\forall x \in E, \forall y \in E, \forall z \in E,\ [(x\mathcal{R}y) \wedge (y\mathcal{R}z)] \Rightarrow [x\mathcal{R}z]$.

On peut tout de suite remarquer que, de par la forme même de ces axiomes, ils sont vérifiés par la relation inverse ou réciproque $\mathcal{R}^{-1}$, qui est définie par : $x\mathcal{R}^{-1}y$ si et seulement si $y\mathcal{R}x$. À toute relation d'ordre est donc associé un ordre réciproque (plus petit/plus grand, inférieur/supérieur etc.).

A.2 Relation d'ordre strict

À une relation d'ordre on associe naturellement la relation obtenue en restreignant celles-ci aux couples d'éléments distincts. On parle alors d'*ordre strict*. Si la relation d'ordre est notée $\leq$, on définit donc la relation d'ordre strict associée, notée $<$ par : $x < y$ si et seulement si $x \leq y$ et $x \neq y$.

On peut alors préciser *relation d'ordre large* quand on veut distinguer la notion de relation d'ordre de celle d'ordre strict.

Il est tout à fait possible d'axiomatiser directement la notion d'ordre strict. Cela peut même s'avérer plus naturel dans certains cas.

Une relation $\mathcal{R}$ d'ordre strict définie sur un ensemble E est une relation binaire irréflexive, et transitive :

– irréflexive : aucun élement de E n'est en relation avec lui-même par $\mathcal{R}$, c'est-à-dire $\forall x \in E, x\ \not\mathcal{R}x$;

– transitive : deux éléments sont mis en relation dès qu'on peut "transiter" par un troisième, c'est-à-dire $\forall x \in E, \forall y \in E, \forall z \in E,\ [(x\mathcal{R}y) \wedge (y\mathcal{R}z)] \Rightarrow [x\mathcal{R}z]$.

On déduit immédiatement de ces deux propriétés qu'une relation d'ordre strict est antisymétrique. À dire vrai une relation d'ordre strict est antisymétrique en un sens plus fort qu'une relation d'ordre large, c'est-à-dire que si x est en relation avec y par $\mathcal{R}$ alors y n'est pas en relation avec x par cette relation. C'est pourquoi on qualifie parfois cette propriété d'antisymétrie forte.

La relation $\mathcal{R}$ définie sur E est *fortement antisymétrique* quand pour tous éléments x et y de E : si $x\mathcal{R}y$ alors $y\ \not\mathcal{R}x$. Cependant pour une relation irréflexive, comme les ordres

stricts, cette propriété est équivalente à la propriété d'antisymétrie définie pour les ordres larges. Il n'y a donc pas d'inconvénient à parler d'antisymétrie dans les deux cas.

À une relation d'ordre strict, notons la $<$, on associe naturellement une relation d'ordre large, notons la $\leq$, définie par : $:x \leq y$ si et seulement si $x < y$ ou $x = y$.

Choisir l'une ou l'autre des axiomatisations n'a pas d'importance en soi. Dans les deux cas on a défini un ordre large et un ordre strict associés. En effet on vérifie facilement, en utilisant les propriétés de l'égalité, que, d'un côté la relation d'ordre strict associée à une relation d'ordre large (transitive, réflexive et antisymétrique) vérifie bien les axiomes d'ordre stricts (elle est transitive et irréflexive), et de l'autre, la relation d'ordre large associée à une relation d 'ordre strict (transitive et irréflexive) vérifie bien les axiomes d'ordre large (elle est transitive, réflexive et antisymétrique).

A.3 Ordre total, ordre partiel

Une relation d'ordre large est *totale* si pour tous x, y dans E, on a $x \leq y$ ou $y \leq x$:

$$\forall (x, y) \in E^2, x \leq y \text{ ou } y \leq x. \tag{A.1}$$

L'ensemble E est alors dit *totalement ordonné*. On dit aussi que la relation d'ordre $<$ est *totale* ou que $(E, \leq)$ est un *ordre total*.

Une relation d'ordre est *partielle* si elle n'est pas totale, c'est-à-dire s'il existe deux éléments que l'on ne peut mettre en relation, ni dans un sens ni dans l'autre :

$$\exists (x, y) \in E^2, x \not\leq y \text{ et } y \not\leq x. \tag{A.2}$$

L'ensemble E est alors dit *partiellement ordonné*.

Deux éléments x et y tels que $x \leq y$ ou $y \leq x$ sont dits *comparables*. Un ordre total est un ordre dont tous les éléments sont deux à deux comparables.

A.3.1 Représentation

Quand on travaille sur un ordre fini, il peut être agréable de disposer d'une représentation visuelle de celui-ci. On peut en proposer une qui est similaire à la représentation habituelle d'un graphe sur papier : il s'agit du diagramme de Hasse.

Pour dessiner un diagramme de Hasse, on représente les éléments de l'ordre par des points de telle sorte que si un élément x est plus grand qu'un autre élément y selon la relation $\leq$, on place la représentation de x plus haut que celle de y ; le fait que deux éléments sont en relation est représenté par un segment entre ces deux points. Du fait de la disposition des points, il n'y a nul besoin d'orienter ces segments avec une flèche. De plus, pour ne pas charger le schéma, on ne représente pas toute la relation d'ordre, mais seulement sa réduction réflexive transitive : d'une part si $x \leq y$, mais qu'il existe z différent de x et de y tel que $(x \leq z) \wedge (z \leq y)$, alors on ne trace pas le segment entre x et y ; d'autre part on ne représente pas les boucles d'un élément vers lui-même.

En cas d'ordre infini, on peut néanmoins aussi utiliser le diagramme de Hasse pour représenter une restriction finie de l'ordre.

A.3.2 Éléments spécifiques usuels d'un ordre partiel

1. – Un ordre partiel $(E, \leq)$ admet un élement *inférieur* $x_\perp$ s'il vérifie : $\forall x \in E, x_\perp \leq x$.

   - De même E admet un élément *supérieur* $x_\top$ s'il vérifie : $\forall x \in E, x_\top \geq x$.

2. - Soit E_S un sous-ensemble de E. Alors $x \in E_S$ est une élément *maximal* de E_S si $\forall y \in E_S$, $y \geq x$ implique que $y = x$. On désigne l'ensemble des éléments maximaux de E_S par Max E_S.
   - De même $x \in E_S$ est un élément *minimal* de E_S si $\forall y \in E_S$, $y \leq x$ implique que $y = x$. On désigne l'ensemble des éléments minimaux de E_S par Min E_S.

3. - Un sous-ensemble E_S est dit *totalement ordonné* si pour tout (x, y) dans E_S on ait $y \geq x$ ou $y \leq x$. Dans un pareil cas, les éléments peuvent être triés de manière à former une *chaîne*. Un tel ensemble E_S admet une *limite supérieure* x_s dans E si $x \leq x_s$ pour tout $x \in E_S$. L'ensemble de toutes les limites supérieures de E_S est désigné par ${E_S}^s$.
   - De même E_S admet une *limite inférieure* x_i dans E si $x \geq x_i$ pour tout $x \in E_S$. L'ensemble de toutes les limites inférieures de E_S est désigné par ${E_S}^i$.

4. - Soient un ordre partiel $(E, \leq)$ et un sous-ensemble E_S de E. La *fermeture supérieure* de E_S est l'ensemble $\{y \in E | \exists x \in E_S, x \leq y\}$, noté $\uparrow E_S$.
   - $\uparrow$ est un opérateur unaire sur l'ensemble des parties de E. Il a les propriétés suivantes :
     - $\uparrow \emptyset = \emptyset$,
     - $\uparrow E_S \subseteq E_S$,
     - $\uparrow\uparrow E_S = \uparrow E_S$,
     - si $E_S \subseteq E'_S$, alors $\uparrow E_S \subseteq \uparrow E'_S$.

     De sorte que $\uparrow$ est un opérateur de fermeture. Un sous-ensemble de E tel que sa fermeture soit lui-même (*i.e.* $\uparrow E_S = E_S$) est appelé *ensemble supérieur*.
   - Par dualité, la *fermeture inférieure* de E_S est l'ensemble $\{y \in E | \exists x \in E_S, x \geq y\}$, noté $\downarrow E_S$. L'ensemble de toutes les fermetures inférieures de E est noté $\mathcal{O}(E)$. $(\mathcal{O}(E), \subseteq)$ est un ordre partiel, nommé *ordre partiel dual* de E.

Il est souvent utile de se souvenir de la proposition suivante :

Lemme 1 (Kuratowski-Zorn lemma)**.** *Tout ensemble partiellement ordonné pour lequel toute chaîne (i.e. tout sous-ensemble totalement ordonné) admet une limite supérieure contient au moins un élément maximal.*

Annexe

B Exemple de script Matlab

```
%% A- Chargement d'un modèle de réseau de régulation génique
twoGenes_model ;
%% B- Simulation
%Conditions initiales
initialcondition(1) = 1 ; initialcondition(2) = 1 ;
%Temps d'énchantillonage
T=0.05 ;
%Evènements externes
externalevents=[] ;
%Bruit additif gaussien
sigma=10e-3 ;
simulation=PWAGRN_Simulator(model,initialcondition,T,externalevents,sigma) ;
%% C- Segmentation
% Niveau de confiance 1 − α
gamma_conf = 0.99 ;
% Nombre initial de points dans un nouveau segment N_C
min_window = 4 ;
% Nombre de point à tester pour détecter une transition N_S
N_test = 2 ;
Y_all=[] ;
for imol=1 :length(model.molecule)
for inpt=1 :length(simulation.noisy.molecule(imol).npoint)
Y_all(imol,inpt)=simulation.noisy.molecule(imol).npoint(inpt) ;
end
end
overlapped_segmentations = segment_all(Y_all, simulation.noiseless.ts,...
    simulation.noisy.sigma, gamma_conf, min_window, N_test) ;
segmentations = solve_segment_overlaps(overlapped_segmentations, 'assign', 5) ;
%% D- Aggregation
% Niveau de confiance
alpha_conf = 0.01 ;
% Classification par molécule
[partitions, all_aggregation_details] = classify_single_all_zeroinit(Y_all,...
    simulation.noiseless.ts, segmentations, alpha_conf) ;
%% E- Classification
whole_classifications = classify_whole_all(partitions) ;
% Nombre minimum de points consécutifs à conserver dans chaque classe
min_consec = 2 ;
% Nombre minimum de points dans chaque classe
min_class_card = 4 ;
% Reduction des classifications selon min_consec et min_class_card :
cases = reduce_whole_classifications(whole_classifications, min_consec, min_class_card) ;
classification = struct('case', repmat(struct('pointclassid',[],...
    'class',struct('molecule',struct('hhsrate',[],'hhdrate',[]))), size(cases))) ;
for icase = 1 :length(cases)
% Retire les petites classes
indClassRm=[] ;
for iClass=1 :length(cases(icase).classes)
if size(cases(icase).classes(iClass).y,1)<3
indClassRm(iClass)=1 ;
else
indClassRm(iClass)=0 ;
end
end
cases(icase).rmClasses=cases(icase).classes(find(indClassRm==1)) ;
cases(icase).classes=cases(icase).classes(find(indClassRm==0)) ;
end
%% F- Reconstruction des seuils et des interactions
maxmc=[] ;
minmc=[] ;
for icase=1 :length(cases)
```

```
F=cell2struct(cases(icase).classes( :).y,'pt')' ;
%Césures maximales
maxmc.case(icase).cut = Cuts2Net (MaxCuts (AllCuts (F, 2*simulation.noisy.sigma)), ...
    classification.case(icase) ) ;
%Multicésures parcimonieuses et mesures de performance
if  isempty(maxmc.case(icase).cut)
minmcut=MinMCuts(maxmc.case(icase).cut) ;
if  isempty(minmcut)
[minmc.case(icase).multicut,minmc.case(icase).Fmeasure] = EvalF_RP(modelTh, minmcut) ;
else
disp(['Case #',num2str(icase),' is not m-separable.'])
end
else
disp(['Case #',num2str(icase),' has just one class.'])
end
end
```

Annexe C

Description du ML pour le stockage et le partage des données d'expérience

C.1 Schéma XML

Rappelons d'abord qu'une déclaration XML ressemble à cela :

```
<élément attributs="..."> contenu (mélangeant éventuellement divers éléments dits enfants) </élément>
```

Un schéma XML (XSD) est utilisé pour exprimer un ensemble de règles auxquelles un document XML doit se conformer. Si cela n'est pas le cas, on dira que le document n'est pas *valide*. Il peut être représenté sous la forme d'une arborescence qui montre comment les éléments sont encapsulés. Le schéma permet notamment de spécifier le nom des éléments et leur ordre, s'ils sont optionnels, s'ils peuvent être répétés (et combien de fois). Le nom des attributs, et s'ils sont requis, sont de même spécifiés. En ce qui concerne le contenu, le type est contraint.

Lorsque l'on ajoute les éléments, pour un même niveau, leur ordre est pris en compte. Au cas où ce n'est pas suffisant, il est aussi possible de préciser qu'un attribut, ou un contenu, est une "clé" : il ne peut alors pas être vide, et il ne peut pas avoir deux fois la même valeur.

Nous pouvons décrire notre structure de données en se basant sur l'enchainement décrit au Chapitre 4. À chaque module doit correspondre un élément dans notre schéma XML. Dans la mesure où, dès l'agrégation, plusieurs cas sont envisageables, certains éléments de premier niveau existent déjà en plusieurs propositions. Cependant, cela n'apparait pas dans le schéma qui précise uniquement la multiplicité éventuelle d'un élément donné.

L'arborescence XSD que nous avons composé en suivant de façon intuitive la chaine de traitement est représentée sur la Figure C.1. Pour comprendre la représentation choisie (moins aride qu'un script XML), nous allons expliquer globalement les symboles. Les quatre points liés signifient que l'on a une séquence : tous les éléments enfants apparaissent en respectant cet ordre. Un élément est donné dans un rectangle. Les nombres se trouvant sous les symboles représentent leur cardinalité : "$a..b$" signifie que l'on peut avoir une répétition de a à b fois. Un rectangle en pointillé dont le nom commence par "@" est un attribut. On précise d'un élément ou d'un attribut si c'est un clé au niveau où l'existence et l'unicité sont vérifiées au moyen d'un symbole "clé". Le chemin permettant de remonter à l'élément parent est fourni. Par exemple, tout attribut d'identification "id" devrait être une clé, car il permet d'identifier un élément de manière unique (cela est utile si l'ordre d'apparition d'un élément ne représente rien).

Sur la Figure C.2, il apparait que seuls "model" et "simulation" ne sont pas optionnels : en effet, une expérience donnée est totalement caractérisée par ces deux éléments, et la chaine de traitement conduira toujours au même résultat.

Les éléments permettant de décrire les résultats de chaque module de la chaine de traitement sont donnés dans les figures suivantes : la description du modèle est donnée par la Figure C.3, celle de la simulation (les données) sur la Figure C.4, celle correspondant à la segmentation C.5, l'agrégation C.6, la classification C.7, la reconstruction des césures maximales C.8 puis celle des multicésures parcimonieuses C.9. Si les résultats d'un module

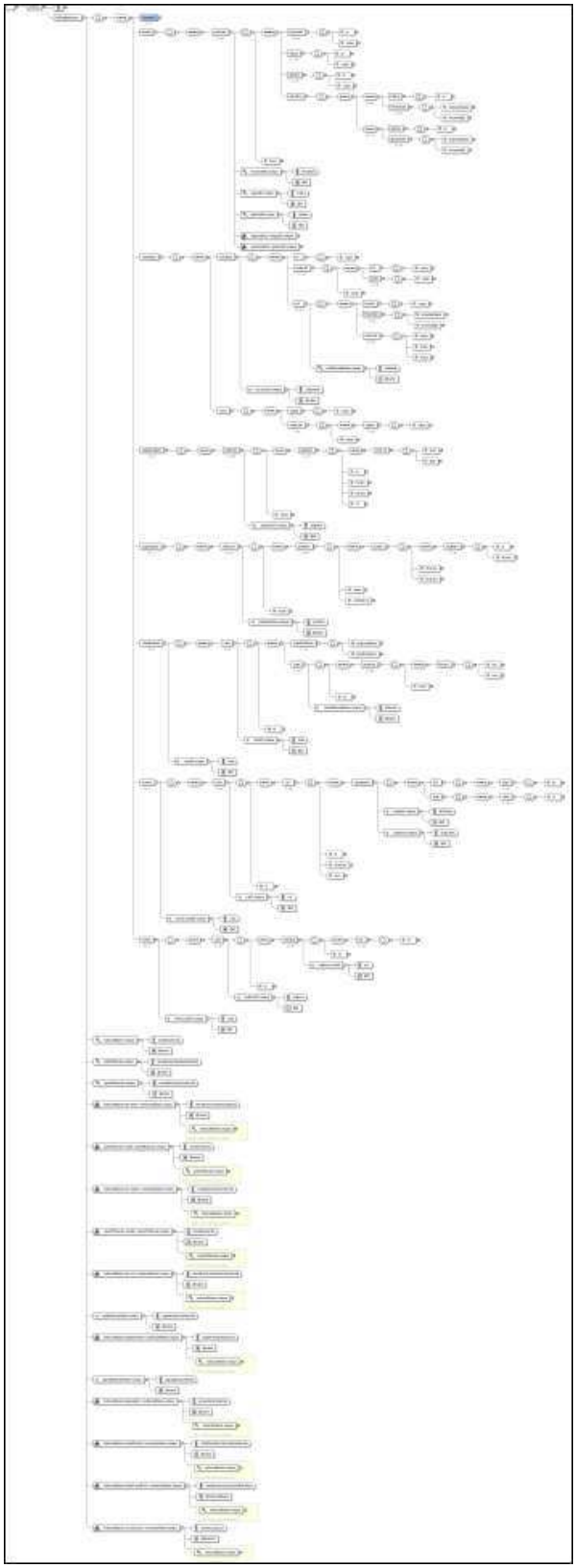


FIGURE C.1 – Arborescence du schéma XML PWAGeRegNet.xsd

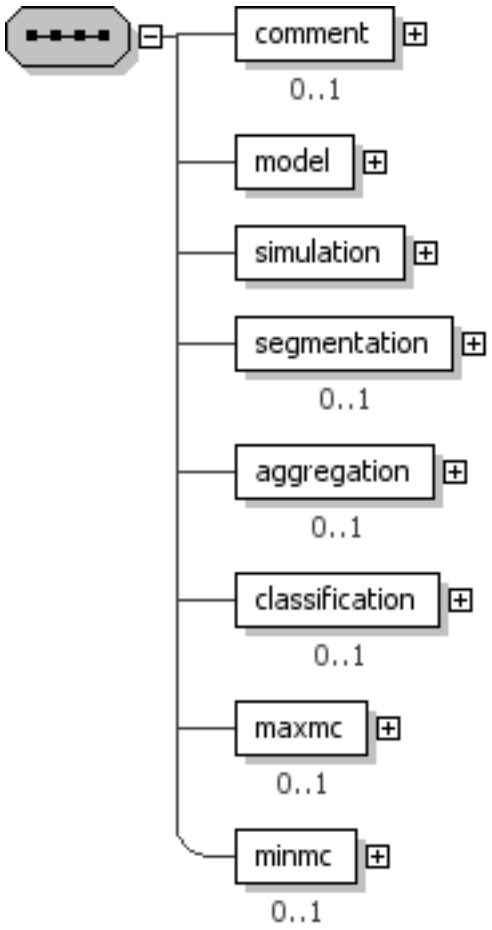


FIGURE C.2 – Arborescence du schéma XML : modules de premier niveau

optionnel de premier niveau sont perdus, il est possible de les retrouver en utilisant les résultats du module juste précédant, sans avoir à reprendre toute la chaine de traitement.

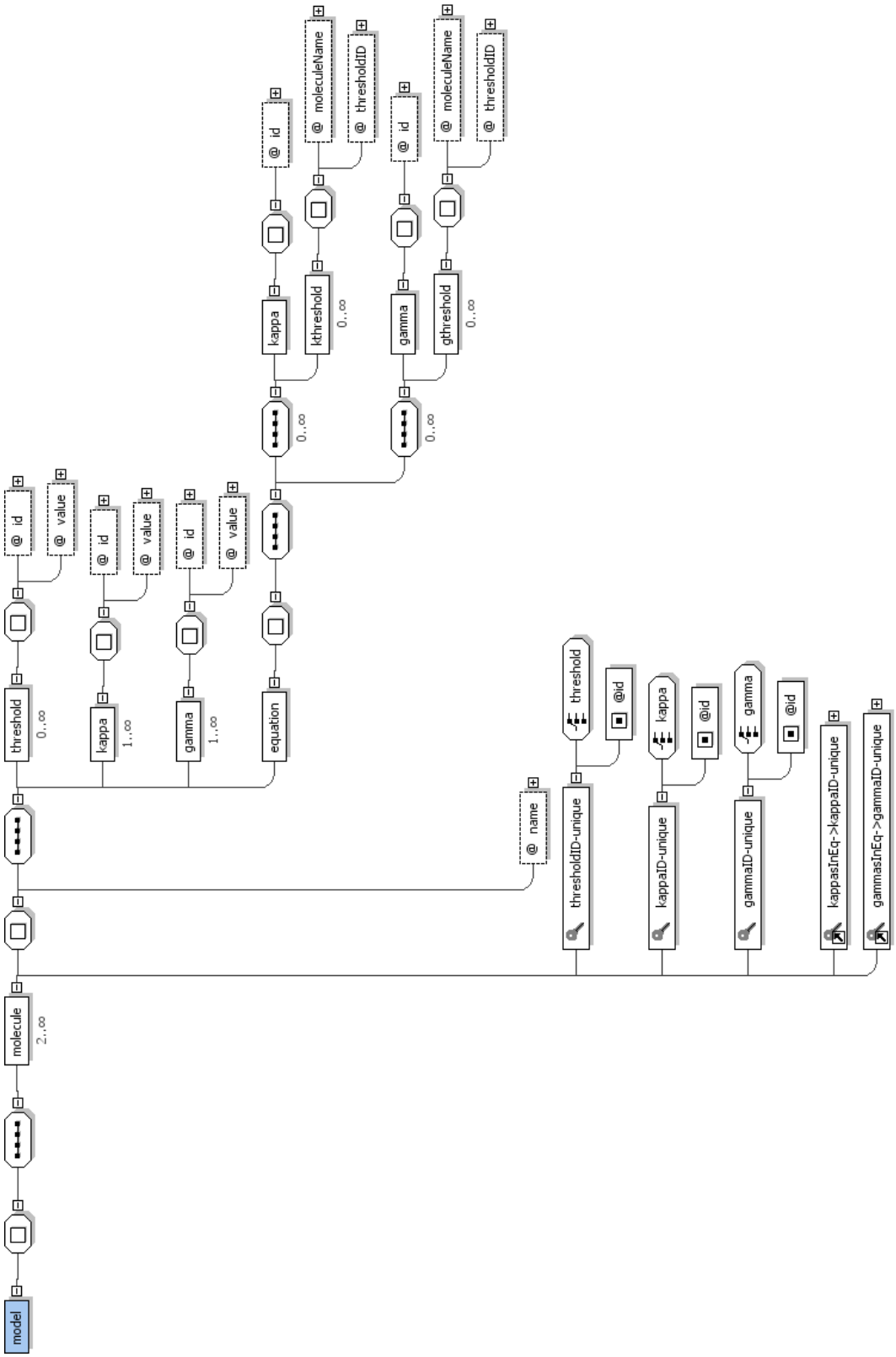


FIGURE C.3 – Arborescence du schéma XML : "model".

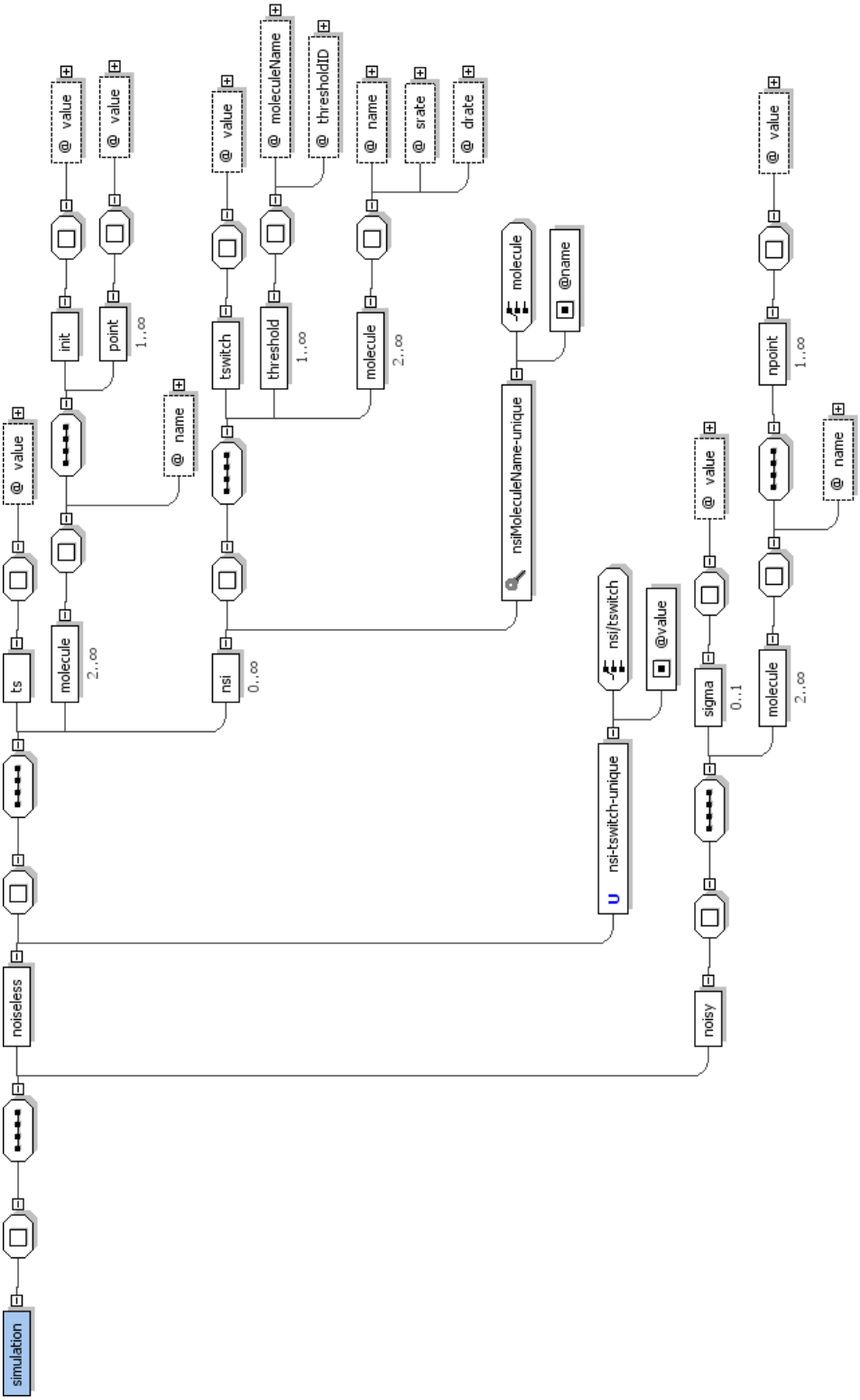


FIGURE C.4 – Arborescence du schéma XML : "simulation".

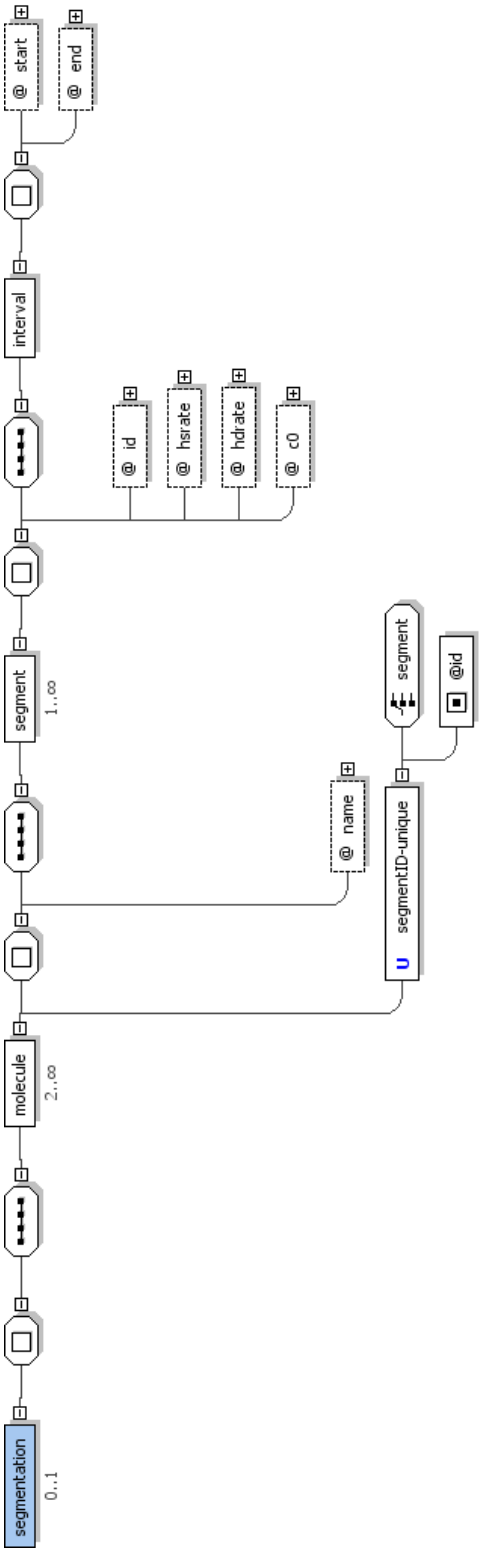


FIGURE C.5 – Arborescence du schéma XML : "segmentation".

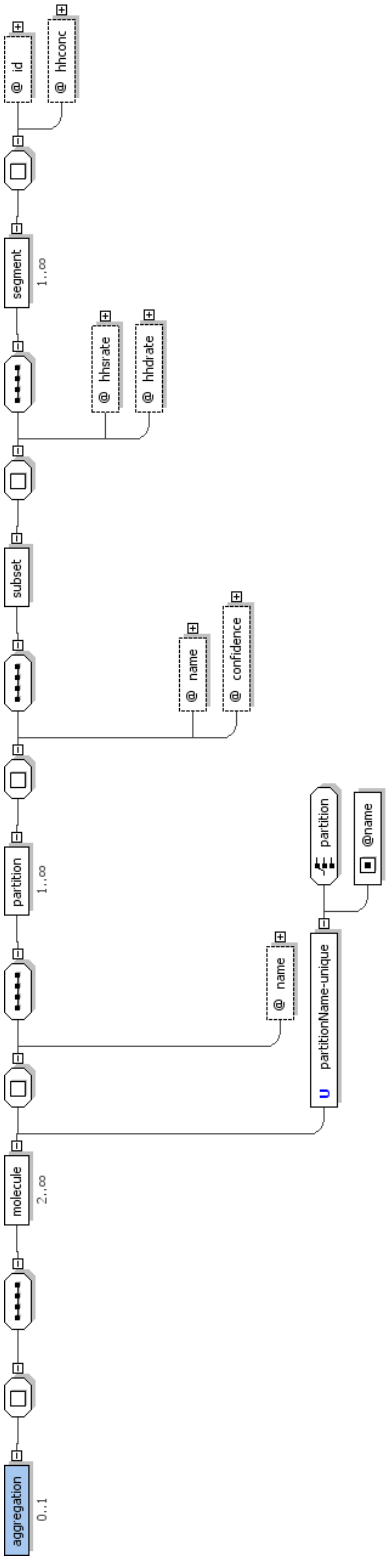


FIGURE C.6 – Arborescence du schéma XML : "aggregation".

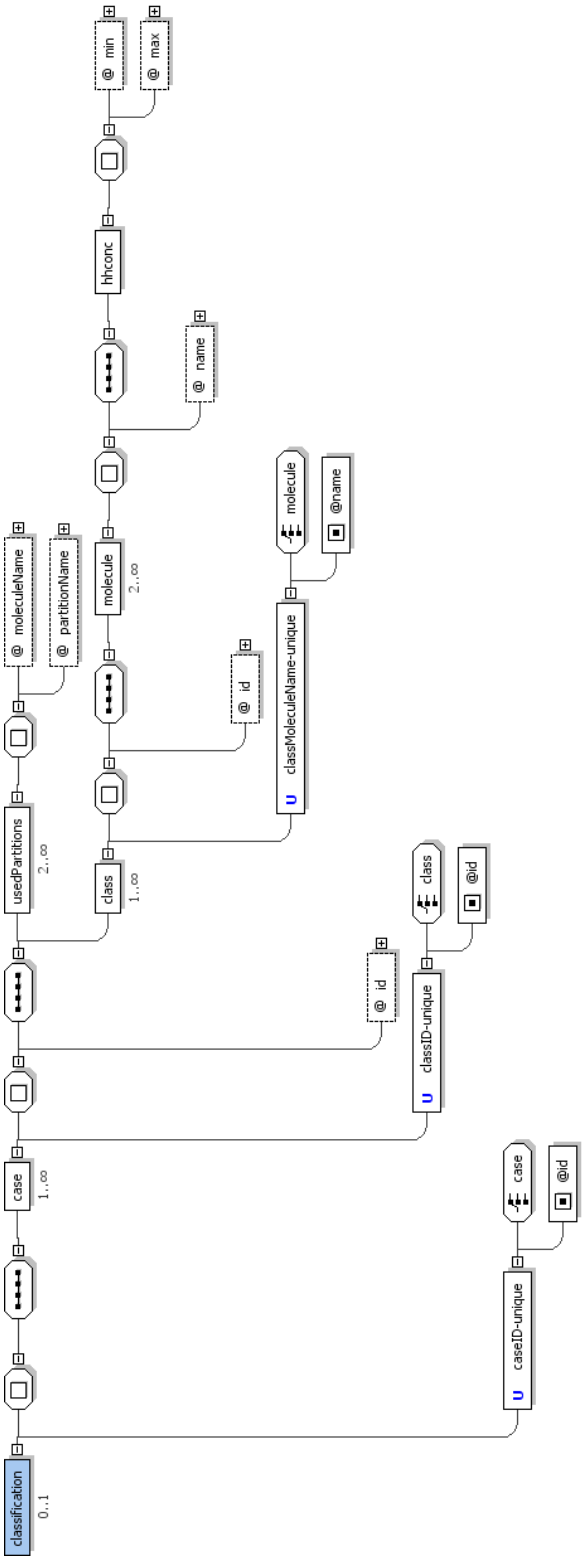


FIGURE C.7 – Arborescence du schéma XML : "classification".

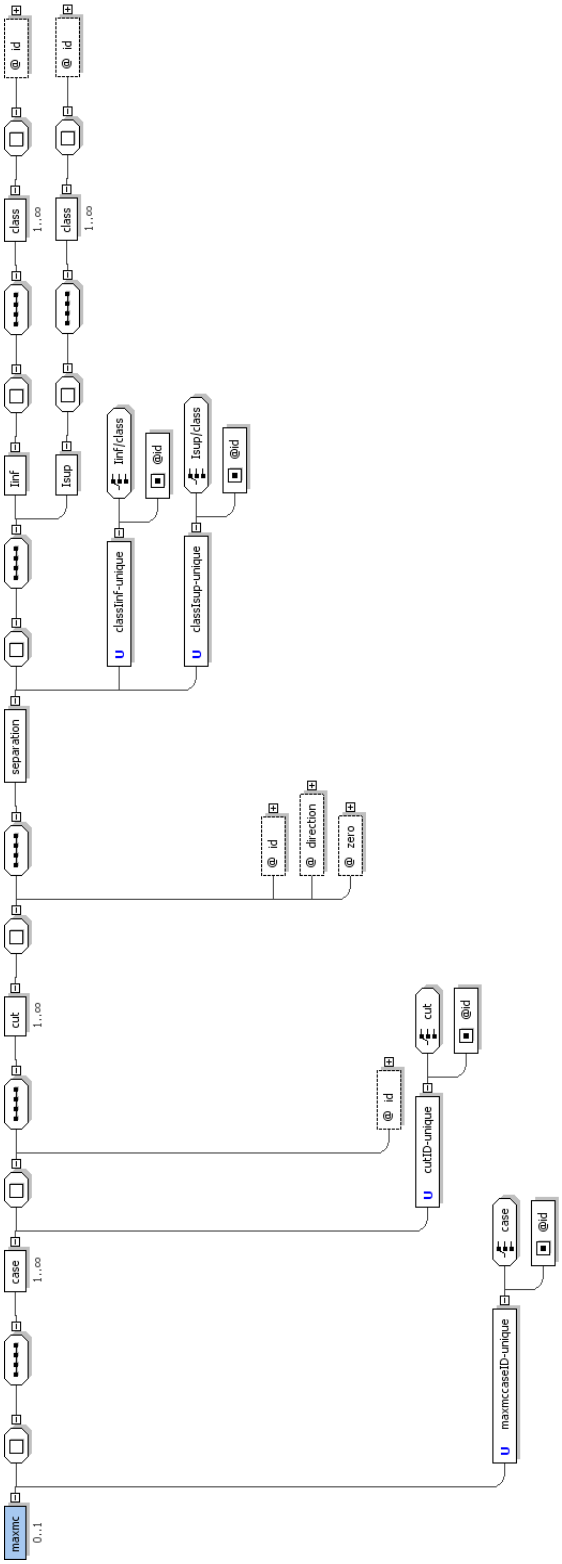


FIGURE C.8 – Arborescence du schéma XML : "maxmc".

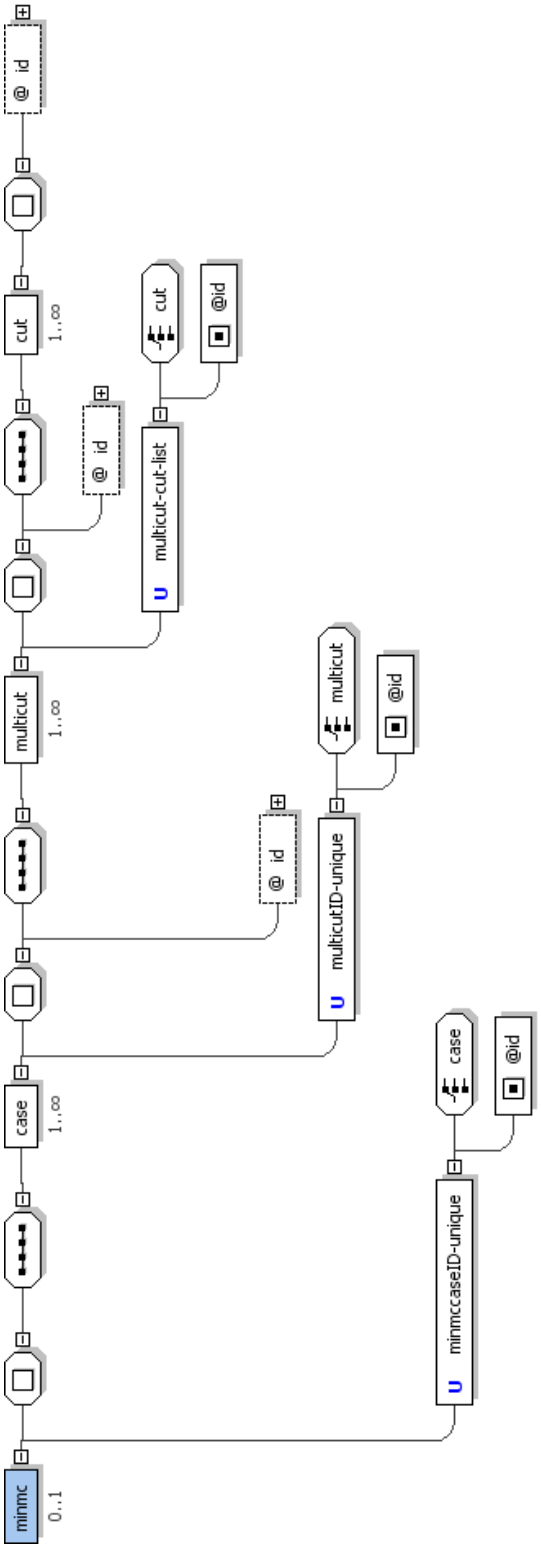


FIGURE C.9 – Arborescence du schéma XML : "minmc".

Bibliographie

Références

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994. 4.3.2

[2] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano. Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model. *Theor. Comput. Sci.*, 298(1) :235–251, 2003. 1.2.2

[3] D. S. Anson. *Reporter genes : a practical guide (Methods in molecular biology).* Humana Press, 2007. 7.2

[4] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, 3 :78, 2007. 2.2

[5] M. Bansal and D. di Bernardo. Inference of gene networks from temporal gene expression profiles. *IET Syst. Biol.*, 1(5) :306–312, 2007. 2.2.1

[6] D. M. Bates and D. G. Watts. *Nonlinear Regression and Its Applications.* John Wiley, NY, 1988. 4.3.1

[7] G. Batt, R. Casey, H. de Jong, J. Geiselmann, J.-L. Gouzé, M. Page, D. Ropers, T. Sari, and D. Schneider. Qualitative analysis of the dynamics of genetic regulatory networks using piecewise-linear models. In E. Pecou, S. Martinez, and A. Maass, editors, *Mathematical and Computational Methods in Biology.* Hermann, 2005. 1.2.1, 4.2

[8] G. Batt, H. de Jong, J. Geiselmann, J.-L. Gouzé, M. Page, D. Ropers, T. Sari, and D. Schneider. Analyse qualitative de la dynamique de réseaux de régulation génique par des modèles linéaires par morceaux. *Technique et Science Informatique*, 26(1-2) :11–45, 2007. 7

[9] G. Batt, H. De Jong, M. Page, and J. Geiselmann. Symbolic reachability analysis of genetic regulatory networks using qualitative abstractions. *Automatica*, 44(4) :982–989, 2007. 7

[10] G. Batt, D. Ropers, H. de Jong, J. Geiselmann, M. Page, and D. Schneider. Qualitative analysis and verification of hybrid models of genetic regulatory networks : Nutritional stress response in *Escherichia coli.* In M. Morari and L. Thiele, editors, *Eighth International Workshop on Hybrid Systems : Computation and Control, HSCC'05*, volume 3414 of *Lecture Notes in Computer Science*, pages 134–150. Springer, 2005. 1.2.1, 1.2.2, 4.2, 7, 9.1

[11] C. Belta, P. Finin, L.C.G.J.M. Habets, A.M. Halász, M. Imieliǹski, V. Kumar, and H. Rubin. Understanding the bacterial stringent response using reachability analysis of hybrid systems. In R. Alur and G.J. Pappas, editors, *Seventh International Workshop on Hybrid Systems : Computation and Control, HSCC'04*, volume 2993 of *Lecture Notes in Computer Science*, pages 111–125. Springer, 2004. 1.2.2, 1.2.2

[12] C. Belta, L. C. G. J. M. Habets, and V. Kumar. Control of multi-affine systems on rectangles with applications to hybrid biomolecular networks. In *Proceedings of the 41st IEEE Conference on Decision and Control, CDC02*, Las Vegas, USA, 2002. 1.2.2

[13] S. A. Benner and A.M. Sismour. Synthetic biology. *Nat. Rev. Genet.*, 6 :535–543, 2005. 1.1.1

[14] K.P. Bennett and O.L. Mangasarian. Multicategory discrimination via linear programming. *Optimization Methods and Software*, 3 :27–39, 1993. 3.2

[15] M. J. Bissell, H. G. Hall, and G. Parry. How does the extracellular matrix direct gene expression ? *J Theor Biol*, 99(1) :31–68, 1982. 9.1

[16] N. Boudreau, C. J. Sympson, Z. Werb, and M. J. Bissell. Suppression of ICE and apoptosis in mammary epithelial cells by extracellular matrix. *Science*, 267(5199) :891–3, 1995. 9.1

[17] D. Bowtell. *DNA microarrays : a molecular cloning manual*. Cold Spring Harbor Laboratory Press, 2002. 7.2

[18] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 9.2.1

[19] P. Brazhnik, A. de la Fuente, and P. Mendes. Gene networks : how to put the function in genomics. *Trends Biotechnol*, 20(11) :467–72, 2002. (document), 1.3

[20] B. Di Camillo, F. Sanchez-Cabo, G. Toffolo, S. K. Nair, Z. Trajanoski, and C. Cobelli. A quantization method based on threshold optimization for microarray short time series. *BMC Bioinformatics*, 6 (Suppl 4) :S11, 2005. doi :10.1186/1471-2105-6-S4-S11. 8, 8.3

[21] R. Casey, H. de Jong, and J.-L. Gouzé. Piecewise-linear models of genetic regulatory networks : Equilibria and their stability. Technical Report RR-5353, INRIA Sophia-Antipolis, 2004. 1.2.1, 4.2

[22] T. Chen, H.L. He, and G.M. Church. Modeling gene expression with differential equations. In R.B. Altman, K. Lauderdale, A.K. Dunker, L. Hunter, and T.E. Klein, editors, *Pac. Symp. Biocomput., PSB'99*, volume 4, pages 29–40, 1999. 1.2.2

[23] K.-H. Cho, S.-M. Choo, S.H. Jung, J.-R. Kim, H.-S. Choi, and J. Kim. Reverse engineering of gene regulatory networks. *IET Syst. Biol.*, 1(3) :149–163, 2007. 2.2

[24] C. W. Cleverdon. Aslib cranfield research project : report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, Cranfield University, 1962. 8.3

[25] C. W. Cleverdon. The significance of the cranfield tests on index languages. In *SIGIR '91 : Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, New York, NY, USA, 1991. ACM Press. ISBN 0-89791-448-1. 8.3

[26] C. Condon, D. Liveris, C. Squires, I. Schwartz, and C.L. Squires. rRNA operon multiplicity in *Escherichia coli* and the physiological implications of *rrn* inactivation. *Journal of Bacteriology*, 177(14) :4152–4156, 1995. 7.1

[27] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press and McGraw-Hill, second edition edition, 2001. 6.2.3

[28] E. J. Crampin. System identification challenges from systems biology. In *Proc. 14th IFAC Symposium on System Identification, Newcastle, Australia*, pages 81–93, 2006. (document), 1.3

[29] G. B. Dantzig. *Linear Programming and Extensions.* Princeton University Press, 1963. 9.2.1

[30] H. de Jong. Modeling and simulation of genetic regulatory systems : a literature review. *J Comput Biol*, 9(1) :67–103, 2002. 1.2.2, 1.2.2, 4

[31] H. de Jong and J. Geiselmann. Modélisation et simulation de réseaux de régulation génique par des équations différentielles ordinaires. In J.-F. Boulicaut and O. Gandrillon, editors, *Informatique pour l'analyse du transcriptome*, pages 143–185. Hermès, 2004. 1.2.2

[32] H. de Jong, J. Geiselmann, G. Batt, C. Hernandez, and M. Page. Qualitative simulation of the initiation of sporulation in *Bacillus subtilis*. *Bulletin of Mathematical Biology*, 66(2) :261–299, 2004. 1.2.2, 7

[33] H. de Jong, J. Geiselmann, C. Hernandez, and M. Page. Genetic Network Analyzer : Qualitative simulation of genetic regulatory networks. *Bioinformatics*, 19(3) :336–344, 2003. 4.2

[34] P. D'Haeseleer, S. Liang, and R. Somogyi. Genetic network inference : From co-expression clustering to reverse engineering. *Bioinformatics*, 16(8) :707–726, 2000. 1.2.2

[35] T. Dorak. *Real Time PCR (BIOS Advanced Methods).* Taylor & Francis, 2006. 7.2

[36] S. Drulhe, G. Ferrari-Trecate, and H. de Jong. Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks. *IEEE Trans. Automat. Control*, 53(1) :153–165, 2008. 7.3, 9.2

[37] S. Drulhe, G. Ferrari-Trecate, H. De Jong, and A. Viari. Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks. In João P. Hespanha and Ashish Tiwari, editors, *HSCC*, volume 3927 of *Lecture Notes in Computer Science*, pages 184–199. Springer, 2006. ISBN 3-540-33170-0. 1.2.1, 4, 9.2

[38] R. Edwards, H.T. Siegelmann, K. Aziza, and L. Glass. Symbolic dynamics and computation in model gene networks. *Chaos*, 11(1) :160–169, 2001. 1.2.2

[39] D. Endy. Foundations of engineering biology. *Nature*, 438 :449–453, 2005. 1.1.1

[40] A.F. Filippov. *Differential Equations with Discontinuous Righthand Sides.* Kluwer Academic Publishers, 1988. 1.2.2

[41] T.S. Gardner, D. di Bernardo, D. Lorenz, and J.J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301 :102–105, 2003. 1.2.2

[42] T.S. Gardner and J.J. Faith. Reverse-engineering transcription control networks. *Phys. Life Rev.*, 2(1) :65–88, 2005. 2.2

[43] R. Ghosh and C.J. Tomlin. Lateral inhibition through Delta-Notch signaling : A piecewise affine hybrid model. In M.D. Di Benedetto and A. Sangiovanni-Vincentelli, editors, *Hybrid Systems : Computation and Control, HSCC'01*, volume 2034 of *Lecture Notes in Computer Science*, pages 232–246. Springer, Berlin, 2001. 1.2.2

[44] R. Ghosh and C.J. Tomlin. Symbolic reachable set computation of piecewise affine hybrid automata and its application to biological modelling : Delta-Notch protein signalling. *Systems Biology*, 1(1) :170–183, 2004. 1.2.2

[45] L. Glass and S. A. Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol*, 39(1) :103–29, 1973. 1.2.2, 1.2.2, 1.2.3, 4.1

[46] L. Glass and J.S. Pasternack. Stable oscillations in mathematical models of biological control systems. *J Theor Biol*, 6 :207–223, 1978. 1.2.3

[47] J.M. Gomez-Gomez, F. Baquero, and J. Blazquez. Cyclic AMP receptor protein positively controls *gyrA* transcription and alters DNA topology after nutritional upshift in *Escherichia coli*. *Journal of Bacteriology*, 178(11) :3331–3334, 1996. 7.1

[48] G. Gonzalez-Gil, R. Kahmann, and G. Muskhelishvili. Regulation of *crp* transcription by oscillation between distinct nucleoprotein complexes. *EMBO Journal*, 17(10) :2877–2885, 1998. 7.1

[49] J.-L. Gouzé and T. Sari. A class of piecewise linear differential equations arising in biological models. Technical Report RR-4207, INRIA Sophia-Antipolis, 2001. 1.2.3, 4.2.1, 4.2.2

[50] J.-L. Gouzé and T. Sari. A class of piecewise linear differential equations arising in biological models. *Dynam. Syst.*, 17(4) :299–316, 2002. 1.2.3, 1.2.3, 4.2

[51] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902. 2.2.2, 4.2

[52] J. Hasty, D. McMillen, and J.J. Collins. Engineered gene circuits. *Nature*, 420(6912) :224–230, 2002. 1.1.1

[53] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins. Computational studies of gene regulatory networks : in numero molecular biology. *Nat. Rev. Genet.*, 2(4) :268–79, 2001. 4

[54] R. Hengge-Aronis. The general stress response in *Escherichia coli*. In G. Storz and R. Hengge-Aronis, editors, *Bacterial stress responses*, pages 161–177. ASM Press, 2000. 1.2.1, 7

[55] B. Hoborn. Gene surgery : on the threshold of synthetic biology. *Med. Klin.*, 75 :834–841, 1980. 1.1.1

[56] G.W. Huisman, D.A. Siegele, M.M. Zambrano, and R. Kolter. Morphological and physiological changes during stationary phase. In F.C. Neidhardt, R. Curtiss III, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, M. Schaechter, and H.E. Umbarger, editors, *Escherichia coli and Salmonella : Cellular and Molecular Biology*, pages 1672–1682. ASM Press, 1996. 7

[57] D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19 :2271–2282, 2003. 8, 8.3

[58] Herrgård M. J., Covert M. W., and B. O. Palsson. Reconstruction of microbial transcriptional regulatory networks. *Current opinion in Biotechnology*, 15 :70–77, 2004. 9.1

[59] J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K.N. Kozlov, Manu, E. Myasnikova, C.E. Vanario-Alonso, M. Samsonova, D.H. Sharp, and J. Reinitz. Dynamic control of positional information in the early *Drosophila* embryo. *Nature*, 430(6997) :368–371, 2004. 1.2.2

[60] K. H. Johansson, S. Sastry, J. Zhang, and J. Lygeros. Zeno hybrid systems. *International Journal of Robust & Nonlinear Control*, 11 :435–451, 2001. 6

[61] P. L. Jones. Extracellular matrix and tenascin-C in pathogenesis of breast cancer. *Lancet*, 357(9273) :1992–4, 2001. 9.1

[62] J.Zhang, K. H. Johansson, J. Lygeros, and S. Sastry. Dynamical systems revisited : Hybrid systems with zeno executions. In *Hybrid Systems : Computation and Control, Lecture Notes in Computer Science 1790*, pages 451–464. Springer-Verlag, 2000. 6

[63] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, 19(5) :643–650, 2003. 1.2.2

[64] H. Kitano. Systems biology : A brief overview. *Science*, 295(5560) :1662–1664, 2002. 1.1.1

[65] Westra R. L. and Peeters R. L. M. State-space modeling and robust identification of piecewise linear gene regulatory networks. *Bull. Math. Biol.*, 2005. 2.2.1

[66] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.*, 2 :164–168, 1944. 4.3.1

[67] L. Ljung and T. Glad. *Modeling of Dynamic Systems*. Prentice Hall PTR, 1994. 1.1.1

[68] Lennart Ljung. *System identification : theory for the user*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2nd edition, 1999. 4.1, 4.3.1

[69] A. Y. Louie, M. M. Huber, E. T. Ahrens, U. Rothbacher, R. Moats, R. E. Jacobs, S. E. Fraser, and T. J. Meade. In vivo visualization of gene expression using magnetic resonance imaging. *Nat Biotechnol*, 18(3) :321–5, 2000. 9.1

[70] D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, second edition, 1984. 9.2.1

[71] A. Lwoff. *L'Ordre biologique*. Robert Laffont, 1969. 1.1.1

[72] F. Markowetz and R. Spang. Inferring cellular networks : A review. *BMC Bioinform.*, 28(Suppl. 6) :S5, 2007. 2.2

[73] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11 :431–441, 1963. 4.3.1

[74] A. Martinez-Antonio and J. Collado-Vides. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, 6(5) :482–489, 2003. 7.1

[75] R. Menzel and M. Gellert. Regulation of the genes for *Escherichia coli* DNA gyrase : homeostatic control of DNA supercoiling. *Cell*, 34(1) :105–113, 1983. 7.1

[76] R. Menzel and M. Gellert. Modulation of transcription by DNA supercoiling : a deletion analysis of the *Escherichia coli gyrA* and *gyrB* promoters. *Proceedings of the National Academy of Sciences of the USA*, 84(12) :4185–4189, 1987. 7.1

[77] T. Mestl, E. Plahte, and S.W. Omholt. A mathematical framework for describing and analysing gene regulatory networks. *J Theor Biol*, 176(2) :291–300, 1995. 1.2.2, 1.2.2, 1.2.2, 1.2.3

[78] S.C. Milne. Restricted growth functions, rank row matchings of partition lattices, and q-stirling numbers. *Advances in Mathematics*, 43 :173–196, 1982. 4.3.2

[79] S. Müller, J. Lu, P. Kügler, and H.W. Engl. Parameter identification in systems biology : solving ill-posed inverse problems using regularization. Technical Report 2008-25, Johann Radon Institute for Computational and Applied Mathematics, Astrian Academy of Sciences (ÖAW), 2008. 2.2.2

[80] Bloomfield P. and Steiger W.L. *Least absolute deviations : Theory, applications and algorithms*. Birkhaüser, 1983. 2.2.1

[81] S. Paoletti, A.Lj. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems : a tutorial. *European Journal of Control*, 513(2-3) :242–260, 2007. 3.1, 3.2

[82] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(S1) :215–224, 2001. 8.3

[83] R.L.M. Peeters and R. L. Westra. On the identification of sparse gene regulatory networks. In *Proc. of the 16th Intern Symp. on Mathematical Theory of Networks and Systems (MTNS 2004)*, 2004. 2.2.1

[84] T. J. Perkins, M. Hallett, and L. Glass. Inferring models of gene expression dynamics. *J. Theor. Biol.*, 230(3) :289–299, 2004. 2.2.2

[85] E. Plahte, T. Mestl, and S.W. Omholt. Global analysis of steady points for systems of differential equations with sigmoid interactions. *Dynamics and Stability of Systems*, 9(4) :275–291, 1994. 1.2.2

[86] E. Plahte, T. Mestl, and S.W. Omholt. A methodological basis for description and analysis of systems with complex switch-like interactions. *Journal of Mathematical Biology*, 36(4) :321–348, 1998. 1.2.3

[87] R. Porreca, S. Drulhe, G. Ferrari-Trecate, and H. de Jong. Structural identification of piecewise-linear models of genetic regulatory networks. *J. Comput. Biol.*, 2008 (in press). 7.3, 31, 9.2

[88] R. Porreca and G. Ferrari-Trecate. Identification of piecewise affine models of genetic regulatory networks : the data classification problem. In *17th IFAC World Congress on Automatic Control*, 2008. 4.3.2, 9.2

[89] R. Porreca, G. Ferrari-Trecate, D. Chieppi, L. Magni, and O. Bernard. Switch detection in genetic regulatory networks. In A. Bemporad, A. Bicchi, and G. C. Buttazzo, editors, *HSCC*, volume 4416 of *Lecture Notes in Computer Science*, pages 754–757. Springer, 2007. ISBN 978-3-540-71492-7. 4, 4.3.1, 9.2

[90] M. Ptashne. *A Genetic Switch : Phage λ and Higher Organisms*. Cell Press and Blackwell Science, Cambridge, MA, 2nd edition, 1992. 4.1

[91] M. Ptashne. *A Genetic Switch : Phage λ Revisited*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 3rd edition, 2004. 1.2.2

[92] Guthke R., Möller U., Hoffman M., Thies F., and Töpfer S. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21(8) :1626–1634, 2005. 2.2.1

[93] J. B. Rampal. *DNA arrays : methods and protocols (Methods in molecular biology)*. Humana Press, 2001. 7.2

[94] V.K. Rohatgi and A.K.M Saleh. *An Introduction to Probability and Statistics*. John Wiley, NY, 2nd edition, 2000. 4.3.2

[95] Jacob Roll. *Local and piecewise affine approaches to system identification*. PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Suède, 2003. 3.2

[96] M. Ronen, R. Rosenberg, B.I. Shraiman, and U. Alon. Assigning numbers to the arrows : Parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences of the USA*, 99(16) :10555–10560, 2002. 1.2.2, 7.2

[97] D. Ropers, H. de Jong, and J. Geiselmann. Mathematical modeling of genetic regulatory networks : Stress responses in *Escherichia coli*. In P. Fu, M. Latterich, and S. Panke, editors, *Systems and Synthetic Biology*. John Wiley & Sons, 2008. 7

[98] D. Ropers, H. de Jong, J.-L. Gouzé, M. Page, D. Schneider, and J. Geiselmann. Piecewise-linear models of genetic regulatory networks : Analysis of the carbon starvation response in *Escherichia coli*. In *Proceedings of the European Conference on Mathematical and Theoretical Biology (ECMTB)*, Modeling and Simulation in Science, Engineering and Technology. Birkhaüser Boston and Basel, 2005. 1.2.1, 4.2, 7

[99] D. Ropers, H. de Jong, M. Page, D. Schneider, and J. Geiselmann. Qualitative simulation of the carbon starvation response in Escherichia coli. *Biosystems*, 84(2) :124–52, 2006. 4, 4.2, 7, 7, 7.1, 7.1

[100] G.-C. Rota. The number of partitions of a set. *The American Mathematical Monthly*, 71(5) :498–504, May 1964. 4.3.2

[101] Stuart Russel and Peter Norvig. *Artificial intelligence, a modern approach*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 1995. 6.2.3

[102] M.H. Saier, T.M. Ramseier, and J. Reizer. Regulation of carbon utilization. In F.C. Neidhardt, R. Curtiss III, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, M. Schaechter, and H.E. Umbarger, editors, *Escherichia coli and Salmonella : Cellular and Molecular Biology*, pages 1325–1343. ASM Press, 1996. 1.2.1, 7

[103] Michel Sakarovitch. *Optimisation combinatoire*. Hermann, 1984. 6.2.3, 6.2.3

[104] R. Schneider, A. Travers, and G. Muskhelishvili. The expression of the *Escherichia coli fis* gene is strongly dependent on the superhelical density of dna. *Molecular Microbiology*, 38(1) :167–175, 2000. 7.1

[105] A. Siegel, O. Radulescu, M. Le Borgne, P. Veber, J. Ouy, and S. Lagarrigue. Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks. *BioSystems*, 84, 2006. 9.1

[106] 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. 1.2.2

[107] M. R. Soboleski, J. Oaks, and W. P. Halford. Green fluorescent protein is a quantitative reporter of gene expression in individual eukaryotic cells. *FASEB J*, 19(3) :440–2, 2005. 7.2

[108] T. Speed. *Statistical analysis of gene expression microarray data*. Chapman & Hall/CRC, 2003. 7.2

[109] R. Sternglanz, S. DiNardo, K.A. Voelkel, Y. Nishimura, Y. Hirota, K. Becherer, L. Zumstein, and J.C. Wang. Mutations in the gene coding for *Escherichia coli* DNA topoisomerase I affect transcription and transposition. *Proceedings of the National Academy of Sciences of the USA*, 78(5) :2747–2751, 1981. 7.1

[110] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005. 4.3.2

[111] J. Tegner, M. K. S. Yeung, J. Hasty, and J. J. Collins. Reverse engineering gene networks : integrating genetic perturbations with dynamical modeling. *Proc Natl Acad Sci U S A*, 100(10) :5944–9, 2003. 2.1

[112] D. Thieffry and H. de Jong. Modélisation, analyse et simulation des réseaux génétiques. *Médécine/Sciences*, 18(4) :492–502, 2002. 2.1

[113] R. Thomas and R. d'Ari. *Biological Feedback*. CRC Press, 1990. 1.2.2, 4.1

[114] A. Travers, R. Schneider, and G. Muskhelishvili. DNA supercoiling and transcription in *Escherichia coli* : the Fis connection. *Biochimie*, 83(2) :213–217, 2001. 7.1

[115] A. M. Turing. The chemical basis for morphegenisis. *Philos. Trans. R. Soc. Lond.*, B 237 :37–72, 1952. 1.1.1

[116] S. Ura, H. Ueda, J. Kazami, G. Kawano, and T. Nagamune. Single cell reporter assay using cell surface displayed Vargula luciferase. *J Biosci Bioeng*, 92(6) :575–9, 2001. 9.1

[117] T. Van den Bulcke, K. Van Leemput, P. van Remortel, B. Naudts, B. De Moor, and K. Marchal. Benchmarking gene network inference algorithms using synthetic gene expression data. In K. Tuyls, R. Westra, Y. Saeys, and A. Nowé, editors, *KDECB workshop 2006*, volume 4366 of *Lecture Notes in Computer Science*. Springer, 2006. ISBN 978-3-540-71036-3. 8, 8.3

[118] N.A.W. van Riel. Dynamic modelling and analysis of biochemical networks : Mechanism-based models and model-based experiments. *Brief. Bioinform.*, 7(4) :364–374, 2006. 2.2

[119] E.P. van Someren, L.F.A. Wessels, and M.J.T. Reinders. Linear modeling of genetic networks from experimental data. In R. Altman and et *al.*, editors, *Proc. Eight Int. Conf. Intell. Syst. Mol. Biol., ISMB 2000*, pages 355–366, Menlo Park, CA, 2000. AAAI Press. 1.2.2

[120] V. Vapnik. *Statistical Learning Theory*. John Wiley, NY, 1998. 3.2, 5.2

[121] E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF '01 : Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, London, UK, 2002. Springer-Verlag. ISBN 3-540-44042-9. 8.3

[122] V. M. Weaver, O. W. Petersen, F. Wang, C. A. Larabell, P. Briand, C. Damsky, and M. J. Bissell. Reversion of the malignant phenotype of human breast cells in three-dimensional culture and in vivo by integrin blocking antibodies. *J Cell Biol*, 137(1) :231–45, 1997. 9.1

[123] L.M. Wick and T. Egli. Molecular components of physiological stress responses in *Escherichia coli*. *Advances in Biochemical Engineering/Biotechnology*, 89 :1–45, 2004. 7

[124] J. Yu, V. A. Smith, P.P. Wang, A.J. Hartemink, and E. D. Jarvis. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18) :3594–3603, 2004. 2.2

[125] A. Zaslaver, A. B., M. Ronen, S. Itzkovitz, I. Kikoin, S. Shavit, W. Liebermeister, M. G. Surette, and U. Alon. A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. *Nat Methods*, 3(8) :623–8, 2006. 9.1

[126] N. Zavbanos, A. Julius, S. Boyd, and G. Pappas. Identification of stable genetic networks using convex programming. In *Proceedings American Control Conference*, pages 2755–2760, juin 2008. 2.2

[127] J.M. Zengel and L. Lindahl. Transcription of ribosomal genes during a nutritional shift-up of *Escherichia coli*. *Journal of Bacteriology*, 167(3) :1095–1097, 1986. 7.1

Résumé

Les progrès récents des techniques expérimentales biologiques ont conduit à la production d'une énorme quantité de données sur le comportement dynamique des réseaux de régulation génique (RRG). Nous présentons une approche pour l'identification des modèles affines-par-morceaux (APM) de RRGs à partir de données expérimentales. Ces modèles reposent sur l'hypothèse que la régulation survient au niveau de la synthèse et de la dégradation des produits de l'expression des gènes : les paramètres cinétiques sont supposés être constants jusqu'à ce que la concentration d'une protéine régulatrice franchisse un seuil de transition.

La méthode que nous présentons se concentrent sur le problème de la détection des transitions entre les différents modes dynamiques à partir des données d'expression génique et sur la reconstruction des seuils de transition associés avec les interactions régulatrices. En particulier, notre méthode prend en considération les contraintes géométriques spécifiques aux modèles APM de RRGs. Une telle méthode d'identification est conçue pour des systèmes à erreur sur la sortie où les observations sont des séries temporelles de mesures bruitées de niveaux de concentration à l'intérieur d'une cellule.

Les données sont d'abord classées en modes dans lesquels le comportement dynamique est considéré comme étant complètement décrit par une équation différentielle linéaire. À partir de la classification résultante, une technique de reconnaissance de forme est utilisée pour reconstruire toutes les combinaisons de seuils de transition qui sont cohérentes avec les données mesurées. Pour chaque combinaison de seuils, il est alors possible de fournir un réseau de régulation et les paramètres dynamiques de chaque mode.

Les performances de notre approche ont été analysées en utilisant des données artificielles simulées pour un modèle simplifié de la réponse à un manque de carbone pour le bactérie *Escherichia coli*. En particulier, nous avons évalué l'influence du niveau du bruit et du pas d'échantillonnage sur les systèmes identifiés. Nos résultats montrent que la méthode, en association avec des séries temporelles de mesures suffisamment précises, lesquelles peuvent être obtenues avec des systèmes à gène rapporteur, permettent une identification quantitative de modèles APM de RRGs.

Abstract

Recent advances of experimental techniques in biology have led to the production of enormous amounts of data on the dynamics of genetic regulatory networks (GRN). We present an approach for the identification of piecewise affine (PWA) models of GRNs from experimental data. These models rely on the assumption that regulation happens at the level of synthesis and degradation of gene expression products : kinetic parameters are considered as being constant until a regulating protein concentration crosses a switching threshold.

The method we present focuses on the problem of detecting switches among different modes of operation in gene expression data and on the reconstruction of switching thresholds associated with regulatory interactions. In particular, our method takes into account geometric constraints specific to PWA models of GRNs. Such an identification method is designed for output-error systems where the observations are noisy time-series measurements of concentration levels inside a cell.

Data points are first classified into modes on which the dynamical behavior is assumed to be fully described by a linear differential equation. From the resulting classification, a pattern recognition technique is used to reconstruct all combinations of switching thresholds that are consistent with measured data. For each combination of thresholds, it is then possible to provide an identified regulatory network and the dynamical parameters of each mode of operation.

The performance of our approach has been analyzed using synthetic data simulated from a simplified model of the carbon starvation response in the bacterium *Escherichia coli*. In particular, we evaluated the impact of the noise level and sampling time on the identified systems. Our results show that the method, coupled with sufficiently precise time-series data, which can be obtained from gene reporter systems, enables a quantitative identification of piecewise affine models of genetic regulatory networks.