Modélisation statistique et formelle de la régulation de l'épissage alternatif

Damien Eveillard

UFR Sciences et Techniques Biologiques
Ecole doctorale BioSE

Modélisation statistique et formelle de la régulation de l'épissage alternatif

THÈSE

présentée et soutenue publiquement le 14 mai 2004

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité Analyse et Modélisation des Systèmes Biologiques)

par

Damien Eveillard

Composition du jury

Président :	M. Jacques NICOLAS	Chargé de Recherche IRISA
Rapporteurs :	M. Claude THERMES	Chargé de Recherche CNRS
	M. Alain VIARI	Directeur de Recherche INRIA Rhône-Alpes
Examinateurs :	M. Alexander BOCKMAYR	Professeur LORIA-UHP (Directeur de thèse)
	Mme Christiane BRANLANT	Directrice de Recherche CNRS (Directrice de thèse)

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503
& Laboratoire de Maturation des ARN et Enzymologie Moléculaire — UMR 7567

Mis en page avec la classe thloria.

Remerciements

Ce travail pluridisciplinaire n'est pas l'œuvre d'un seul étudiant en thèse. Son accomplissement n'a été possible que dans le cadre de collaborations qui ont permis de créer un cadre de travail unique. C'est notamment pour cette raison que je tiens à remercier les personnes qui y ont contribué :

- Alexander Bockmayr pour m'avoir acceuilli au sein de son équipe de recherche encore balbutiante. Il m'a initié à l'informatique avec patience et pédagogie tout en subissant mes égarements intellectuels. Il a également su catalyser autour de moi les compétences qui étaient nécessaires pour avancer dans un domaine de recherche pluridisciplinaire. Sa vision et sa conception de la recherche constitueront pour moi une référence lors de mes futurs travaux.
- Christiane Branlant pour m'avoir guidé dans le monde de la Biologie Moléculaire et de m'avoir sensibilisé à sa complexité. Elle m'a permis de dégager une ligne directrice dans mes travaux.
- Alain Viari pour avoir jugé mes travaux et pour ses indications informatiques qui donneront surement lieu à d'autres perspectives. Je le remercie de prendre part à ce jury.
- Claude Thermes qui a bien voulu rapporter ce manuscrit. Je lui suis très reconnaissant de faire partie de ce jury pour rendre compte des aspects pluridisciplinaires que peut prendre la biologie.
- Jacques Nicolas de faire partie de ce jury pour se prononcer sur mes travaux à l'interface de deux domaines scientifiques.
- Delphine Ropers pour sa pédagogie et nos discussions biologiques qui ont largement inspiré mes travaux. Merci d'avoir supporté mes errances de modélisation et d'avoir recadré biologiquement nos modèles.
- Hidde de Jong pour avoir contribué à me mettre sur les rails de la modélisation des processus moléculaires. Son ouverture scientifique contribue largement au plaisir que j'ai à travailler avec lui.
- Olivier Bernard et Antoine Sciandra qui m'ont initié à la modélisation des systèmes biologiques et développé chez moi le goût de la conceptualisation du vivant. Ils m'ont appris à toujours être critique sur mes travaux et éventuelles découvertes.
- Yann Guermeur pour m'avoir initié à la "vraie" statistique et d'avoir bien voulu faire des bornes ensemble malgré sa double vie d'agent secret.
- Myriam, Sandrine pour les précieuses corrections finales qu'elles ont apportées à ce manuscrit. Abdelhalim pour nos discussions et pour son thé marocain magique. Les autres membres de MODBIO (dans le désordre : Stéphanie, Yasmin, Arnaud, Eric, Nicolaï, Manu, Yannick) pour l'ambiance de travail particulièrement chaleureuse. Sophie pour son soutien logistique malgré mon desordre caractéristique et pour m'avoir appris à ranger un classeur.
- Tous ce qui a remplit les presque trois ans et demi de thèse avec de manière non exhaustive : les rondades de Marco ; le jambon de Miguel ; les crampes de Martial ; le couscous de Walid qui marque une gorge à vie ; les sessions de poudre avec Nico ; les apnées bleues de Sophie ; les crevettes d'océano ; le sourire de Clémence.

Enfin je tiens à remercier mon père et Marielle sans qui je n'aurais jamais repoussé aussi loin ma curiosité et ma quête de connaissances concernant les systèmes vivants qui nous entourent. C'est une grande chance que j'ai eu d'avoir leurs soutiens inconditionnels.

Et le reste de ma famille, soutien de tout les instants, où qu'elle soit...

The greatest challenge today, not just in cell biology and ecology but in all of science, is the accurate and complete description of complex systems. Scientist have broken down any kinds of systems. They think they know most of the element and forces. The next task is to reassemble them, at least in mathematical models that capture the key properties of the entire ensembles.

E. O. Wilson 1998

Table des matières

Avant propos : Une approche informatique pour un problème biologique xv

Partie I Etat des domaines scientifiques en présence 1

Chapitre 1 Épissage alternatif : une régulation post-transcriptionnelle 3

1.1 Généralités concernant l'épissage des ARN prémessagers nucléaires . . 3
1.1.1 La réaction d'épissage 4
1.1.2 Les signaux nécessaires à l'épissage 7
1.2 Généralités sur l'épissage alternatif 8
1.3 Acteurs de la régulation de l'épissage des ARN prémessagers nucléaires 10
1.3.1 Les protéines SR 10
1.3.2 Les protéines hnRNP 12
1.4 Fonctionnalités des sites de fixation des protéines régulatrices 13
1.5 Conclusions 14

Chapitre 2 L'épissage de l'ARN du virus HIV-1 17

2.1 Généralités concernant le virus HIV-1 17
2.2 Cycle de vie du virus HIV-1 18
2.3 Contrôle du cycle du virus par les protéines virales 18
2.4 Régulation de l'épissage du virus 20
2.5 Conclusions 22

Chapitre 3 Modélisation de systèmes biologiques 25

3.1 Démarche de modélisation 26
3.1.1 Détermination du référentiel 26
3.1.2 Hypothèses biologiques 27

3.1.3 Conceptualisation 28
3.1.4 Formalisation 28
3.1.5 Validation qualitative 28
3.1.6 Identification 29
3.1.7 Simulation 30
3.1.8 Validation quantitative 31
3.1.9 Conclusion 31

Chapitre 4 Formalismes en modélisation biologique **33**
4.1 Formalismes discrets 35
4.1.1 Les réseaux booléens 35
4.1.2 Le π-calcul 37
4.2 Formalisme continu 38
4.2.1 Fondements mathématiques 39
4.2.2 Applications en modélisation biologique 40
4.3 Raisonnement qualitatif et automates hybrides 41
4.4 Contraintes en modélisation biologique 43
4.4.1 Programmation concurrente par contraintes hybrides 43
4.4.2 Contraintes d'intervalles et dynamique continue 46
4.4.3 Composition parallèle 48
4.4.4 Utilisation de conditions et de changements discrets 49
4.4.5 Comportement par défaut 50

Chapitre 5 Apprentissage statistique **53**
5.1 Philosophie de l'apprentissage statistique 54
5.1.1 Fondements théoriques de l'apprentissage statistique 57
5.1.2 Mise en pratique de l'apprentissage statistique 57
5.2 Les SVM bi-classes 58
5.2.1 La famille de fonctions 59
5.2.2 Le choix de l'hyperplan optimal 59
5.2.3 Exemple d'apprentissage : interprétation géométrique de l'algorithme des SVM 63
5.2.4 L'approche par noyau 63
5.3 Les SVM multi-classes 66
5.3.1 La famille de fonctions $\mathcal{H}$ 66

5.3.2 Choisir un jeu optimal d'hyperplans 66
5.3.3 Interprétation géométrique et avantage du noyau 67
5.4 Conclusions 68

Partie II Analyse des expériences SELEX : Identification des motifs de régulation d'épissage 69

Chapitre 6 Approches théoriques des expériences SELEX 71
6.1 Protocole expérimental SELEX 71
6.2 Modélisation statistique des expériences SELEX 74
6.2.1 Modélisations mathématiques 74
6.2.2 Apports du modèle 75
6.3 Analyses standards des résultats de SELEX 77
6.3.1 Méthodes bioinformatiques 78
6.3.2 Incohérences des analyses standards 79
6.4 Conclusion 83

Chapitre 7 Classification et analyses des résultats de SELEX 85
7.1 Approche méthodologique 85
7.2 Résultats de la classification des données SELEX 86
7.3 Interprétation de la répartition des séquences 94
7.3.1 Critère de la structure secondaire des acides nucléiques 94
7.3.2 Importance des séquences de faibles affinités 96
7.4 Conclusions 97

Partie III Recherche de motifs : Application à la recherche de motifs de régulation d'épissage 99

Chapitre 8 Approches actuelles des outils de recherche de motifs 103
8.1 Méthodes algorithmiques 103
8.1.1 Approches existantes 104
8.1.2 `grappe` une approche discrète pour la biologie 104
8.2 Méthodes d'apprentissage statistique 105

8.2.1 Utilisation des modèles de Markov cachés 107
8.2.2 Utilisation de SVM 108
8.3 Conclusions 109

Chapitre 9 Localiser les sites de régulation : Développement et application de KOALAB **111**
9.1 Une approche intégrée pour rechercher les motifs de régulation 111
9.2 Une interface graphique dédiée aux problèmes biologiques 123
9.2.1 Contrôle des approches standards de recherche de motifs 123
9.2.2 Intégration des résultats 125
9.3 Application de KOALAB aux motifs de régulation d'épissage sur HIV-1 125
9.3.1 Comparaison des motifs discrets et statistiques 127
9.3.2 Exploitation de KOALAB : vers une cartographie fonctionnelle 128
9.4 Avantages de l'approche *in silico* 133
9.4.1 Validation biologique partielle 134
9.4.2 Émergence d'hypothèses nouvelles 134
9.5 Conclusions 135

Partie IV Modélisation formelle : Analyse de la fonctionnalité des motifs de régulation d'épissage **137**

Chapitre 10 Modéliser la régulation d'un site d'épissage par les protéines SR **141**
10.1 Formalisation du modèle de régulation du site A3 142
10.1.1 Connaissances expérimentales concernant le site A3 142
10.1.2 Hypothèses biologiques 144
10.1.3 Modèle mathématique 146
10.2 Etude du comportement qualitatif 149
10.2.1 Variables d'observations : indicateurs du comportement qualitatif150
10.2.2 Caractéristiques qualitatives du système mathématique 153
10.3 Conclusions 158

Chapitre 11 Modélisation intégrative des régulations de l'épissage alternatif **159**

11.1 Utilisation des contraintes hybrides pour modéliser sur plusieurs échelles 160
11.2 Construction et analyse de modèle multi-échelles 183
11.2.1 Hypothèses et conceptualisation de la régulation multi-site . . . 183
11.2.2 Analyse qualitative formelle de modèles intégrés 200
11.3 Influence de la régulation de l'épissage alternatif dans le cycle de vie de HIV-1 200
11.4 Conclusions 204

Partie V Discussion **205**

Chapitre 12 Vers une méthodologie biologique *in silico* **207**
12.1 Approche méthodologique transversale 207
12.1.1 Approche informatique intégrée 207
12.1.2 Une nouvelle méthodologie biologique 209
12.1.3 Critiques de l'approche bioinformatique 209
12.2 Perspectives 210
12.2.1 Outils d'aide à la décision pour les expériences SELEX 210
12.2.2 Exploitation de classification par la discrimination automatique 211
12.2.3 Détection statistique et fonctionnelle de motifs 211
12.2.4 Modélisation générique par des systèmes de contraintes 212
12.2.5 Analyse automatique des propriétés d'un système biologiques . 212

Bibliographie **215**

Table des figures

1.1 Dogme du transfert d'information de la biologie moléculaire 4
1.2 Relation entre la transcription et la maturation des ARN prémessagers . . 5
1.3 Schématisation des principaux éléments qui composent le spliceosome . . . 6
1.4 Représentation de l'épissage sur une région intronique d'un ARN prémessager 7
1.5 Séquences exoniques et introniques essentielles à la réaction d'épissage . . . 8
1.6 Représentation des différents modes d'épissage alternatif 9
1.7 La fixation des protéines SR sur l'élément ESE favorise le choix des sites 5' ou 3' d'épissage adjacents 14
1.8 Fonction antagoniste des protéines de régulation par sites chevauchant . . . 14

2.1 Cycle réplicatif du virus HIV-1 19
2.2 Expression des gènes du virus HIV-1 au cours du cycle de vie 21
2.3 Organisation du génome de HIV-1 22
2.4 Structure secondaire de la région de l'ARN viral entourant le site d'épissage A3 23

3.1 Protocole de modélisation d'un système biologique 26
3.2 Positionnement d'un modèle dans le contexte biologique et de modélisation 27

4.1 Répartition des formalismes de modélisation suivant le référentiel biologique 34
4.2 Exemple de réseau booléen 36
4.3 Exemple de fonctions de régulation de croissance 40
4.4 Interaction des contraintes avec le solveur de contraintes 43
4.5 Domaine d'atteignabilité d'une espèce moléculaire à cinétique linéaire . . . 47
4.6 Domaine d'atteignabilité de deux éléments moléculaires à cinétique michaelienne 49
4.7 Représentation d'un *switch* entre deux comportements par une condition . 51
4.8 Représentation d'un *switch* entre deux comportements avec un comportement par défaut 52

5.1 Représentation d'un vecteur dans $\mathbb{R}^3$ 54
5.2 Représentation de la discrimination linéaire 55
5.3 Combinaison de protocoles expérimental et statistique pour inférer des fonctions biologiques 56
5.4 Illustration d'un effet de sur-apprentissage 58

5.5 Sous-ensembles imbriqués de fonctions ordonnées en fonction de leurs dimensions VC croissante 61
5.6 Principe de la minimisation structurelle du risque 61
5.7 Représentation de la marge d'après l'algorithme des SVM 63
5.8 Représentation du pré-traitement sur un jeu d'échantillons non séparables linéairement 64
5.9 Représentation de la marge dans le cas multi-classe 67

6.1 Schématisation du protocole expérimental SELEX 73
6.2 Évolution de la distribution de l'affinité au cours des cycles SELEX 76
6.3 Estimation de l'évolution de l'information biologique au cours des cycles SELEX 78
6.4 Analyse standard des données SELEX 79
6.5 Protocole de test de la stabilité des analyses SELEX 82

7.1 Structures secondaires des ARN déterminées expérimentalement 95
7.2 Résultat des données SELEX pour la protéine L7Ae avec les structures secondaires associées 95
7.3 Résultat de classification des données SELEX pour la protéine L7Ae 96

8.1 Protocole de la mise en œuvre de l'apprentissage 106
8.2 Exemple d'automate HMM 107

9.1 Vérification graphique de l'apprentissage statistique par KOALAB 124
9.2 Représentation des scores de la M-SVM par KOALAB 124
9.3 Représentation de l'intégration des résultats de `grappe` et de M-SVM par KOALAB 126
9.4 Scores bruts de sorties de M-SVM 129
9.5 Scores normalisés de sorties de M-SVM 130
9.6 Agrandissement de la région 5350-5430 du graphique des données normalisées de la M-SVM 130
9.7 Agrandissement de la région 5530-5600 du graphique des données normalisées de la M-SVM. 131
9.8 Graphique d'autocovariance des sorties de M-SVM 132
9.9 Graphique des données normalisées de la M-SVM analysé avec les moyennes mobiles 132
9.10 Séquence en cis du HIV-1 associée à la présence des motifs mis en évidence par la M-SVM 133

10.1 Régulation de l'épissage au site A3 143
10.2 Représentation des acteurs de la régulation du site A3 144
10.3 Schéma conceptuel de la régulation du site A3 145
10.4 Représentation de la cinétique de Michaelis-Menten 147
10.5 Comportement qualitatif de l'efficacité d'épissage au site A3 152
10.6 Graphe de transition partiel pour un système à deux variables 155
10.7 Graphe de transition du comportement qualitatif du modèle 156

10.8 Comportement qualitatif du modèle de régulation au site A3 157

11.1 Composition en exons et abondance relative des ARNm du virus HIV-1 . . 184
11.2 Schéma conceptuel du modèle multi-échelles du cycle de vie de HIV-1 . . . 201
11.3 Comportement dynamique du modèle multi-échelle 202
11.4 Effet de l'augmentation de la quantité de protéine ASF/SF2 sur le cycle viral203

Avant propos : Une approche informatique pour un problème biologique

La fin du siècle dernier a très nettement marqué la Biologie. Les aspects moléculaires de cette science ont notamment profité des avancées technologiques comme le séquençage automatique de génomes. Dans ce nouveau contexte, les biologistes ont du faire face à un nombre croissant de données. L'Informatique dont le but est de gérer l'information, permet de stocker et de gérer les données qui restent complexe. L'outil informatique est ainsi devenu rapidement indispensable aux progrès biologiques. Mais au delà de la puissance de calcul et de stockage, l'Informatique permet aussi de raisonner sur l'information. Ce dernier aspect est l'élément essentiel qui laisse envisager que l'emploi de méthodes informatiques est prometteur pour analyser et comprendre la problématique biologique. L'approche consiste pour cela à modéliser le vivant par différentes approches. Nous utiliserons pour cela des domaines formelles variés tels que la statistique ou la programmation par contraintes. Nous les appliquerons au cours de cette thèse sur un problème biologique complexe : l'épissage alternatif.

C'est donc dans un contexte scientifique interdisciplinaire que cette thèse s'inscrit qui se manifeste par une collaboration entre deux laboratoires distincts : le laboratoire de Christiane BRANLANT dont l'équipe de recherche s'intéresse à l'épissage alternatif et le projet de recherche d'Alexander BOCKMAYR au sein du LORIA, qui développe des outils d'analyse et des raisonnements informatiques originaux appliqués à la Biologie. A l'interface de ces deux mondes, cette thèse propose une approche transversale des différentes approches informatiques pour comprendre un processus biologique complexe. Le développement de modèles statistiques permet de paramétrer les différentes hypothèses du processus biologique pour ensuite proposer une méthode qui permet de généraliser la connaissance acquise. Les connaissances peuvent se représenter sous la forme d'hypothèses de fonctionnement du système biologique. Une modélisation formelle permettra ensuite d'analyser et de raisonner sur ces hypothèses de régulation de l'épissage alternatif. Cette démarche transversale en informatique permet donc de raisonner théoriquement sur le système pour proposer ensuite des méthodes dédiées. Cette nouvelle méthodologie aborde donc le problème biologique de l'épissage alternatif dans sa globalité, ce qui représente à notre point de vue un des enjeux actuels de la biologie computationnelle et une évolution naturelle des concepts en Biologie.

Première partie

Etat des domaines scientifiques en présence

Chapitre 1

Épissage alternatif : une régulation post-transcriptionnelle

A la fin des années 70, les travaux de [Berget & Sharp, 1977] mettent en évidence la présence de gènes morcelés dans le génome d'un adénovirus. Ces travaux caractérisent alors la présence de séquences non codantes dans des gènes codants des protéines. Suite à ces travaux, la présence de séquences non codantes dans les gènes nucléaires s'est généralisée aux eucaryotes. Ce type de segments non codants se retrouve dans les gènes d'ARN stables (ARN de transfert et ARN ribosomiaux), dans certains gènes d'archaee et bactériens. Les mécanismes de régulation de ces gènes semblent différents en fonction du cadre biologique. Néanmoins, dans tout les cas, les séquences non codantes sont éliminées. Ce processus biologique d'élimination est appelé **épissage**. On définit ces fragments non codants comme introns[1] par opposition aux exons[2] qui correspondent aux séquences codantes. Les séquences codantes sont reliées entre elles pour former l'ARN messager qui sera exporté dans le cytoplasme pour être traduit en protéines. Ainsi, hormis pour certains gènes nucléaires, on peut considérer de manière générale que tous les génomes des eucaryotes possèdent des introns. Cela fait de l'épissage un phénomène majeur chez les eucaryotes supérieurs. Chez les mammifères, les introns sont de taille importante par rapport aux insectes et levures. La taille moyenne d'un exon chez l'homme est de 75 à 150 nucléotides (nts) contre 3500 nts pour un intron moyen [Deutsch & Long, 1999] qui peut atteindre un maximum de 500 000 nts [Rowen *et al.*, 2002]. Toutes ces observations biologiques renforcent l'intérêt de la communauté scientifique à comprendre le processus biologique qu'est l'épissage des ARN prémessagers nucléaires.

1.1 Généralités concernant l'épissage des ARN prémessagers nucléaires

L'épissage est une étape du processus de maturation des ARN. Ce processus s'effectue juste après la transcription au cours de laquelle sont synthétisés les ARN. L'épissage fait

[1] pour **intr**agenic reg**ions**
[2] pour **ex**pressed reg**ions**

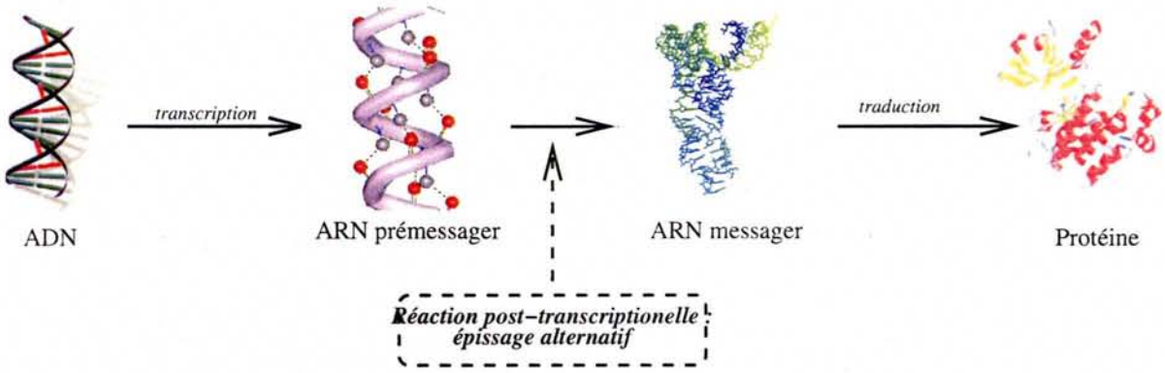

FIG. 1.1 – **Dogme du transfert d'information de la biologie moléculaire.** On représente ici le flux d'information génétique qui utilise différentes formes moléculaires. L'information est initialement contenue dans les gènes (génome) sous forme d'ADN. Une extraction de l'information parmi cette banque s'effectue par la **transcription**. L'information pertinente est alors convertie sous forme d'ARN. Ces molécules après **maturation** peuvent être exportées dans le cytoplasme pour être traduite en protéines grâce à un processus de **traduction**.

partie de la maturation des ARN et possède ainsi une place centrale dans le grand dogme du transfert d'information de la biologie moléculaire schématisé sur la Figure 1.1.

Les ARN prémessagers qui sont issus de la transcription doivent subir certaines modifications afin de pouvoir être traduits en protéines : la maturation. Cette phase de régulation post-transcriptionnelle est liée au processus de transcription, comme le montre la Figure 1.2 (pour revue voir [Maniatis & Reed, 2002]). Au fur et à mesure que l'ARN polymérase *II* synthétise les ARN prémessagers (la **transcription**), ceux-ci sont coiffés d'une coiffe en 5' (la **formation de la coiffe**). Cette coiffe correspond à une liaison phosphate terminale qui va protéger l'ARN nouvellement formé de l'action d'une autre catégorie d'enzyme. Les exonucléases qui hydrolysent l'ARN à partir de l'extrémité 5' de l'ARN vers l'extrémité 3'. Les ARN sont ensuite épissés par un processus que nous détaillerons par la suite (**l'épissage**) pour être ensuite polyadénylés à leur extrémité 3' (**polyadénylation**). Cette dernière étape permettra l'exportation des ARN prémessagers nucléaires devenus messagers vers le cytoplasme où ils seront traduits en protéines. C'est au cours de la maturation des ARN qu'a lieu la régulation post-transcriptionnelle dans laquelle l'épissage joue un rôle central.

1.1.1 La réaction d'épissage

Le processus d'épissage consiste à extraire de la séquence ARN des sous-séquences qui sont appelées les introns par opposition aux exons qui sont conservés. L'élimination des introns se fait au sein d'un macro-complexe moléculaire : le spliceosome (pour revue voir [Will & Lührmann, 2001]). C'est une machinerie qui est constituée de molécules ribonucléiques et protéiques. De manière générale, il est constitué de 5 UsnRNP et de nombreux facteurs protéiques. Les UsnRNP spliceosomales sont des particules constituées de petites molécules d'ARN nucléaires, les snRNA U1, U2, U4, U5 et U6 qui sont associés à des

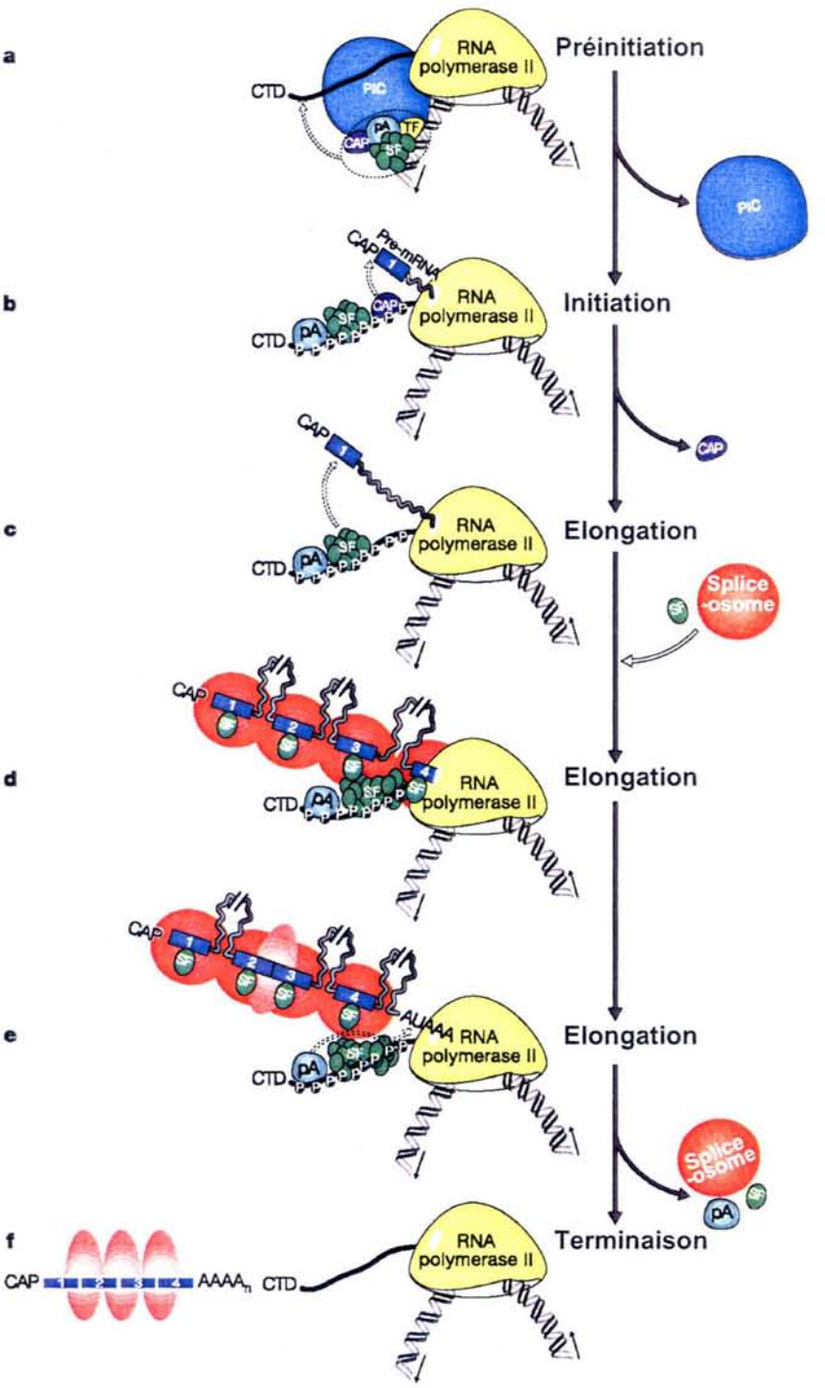

FIG. 1.2 – **Relation entre la transcription et la maturation des ARN prémessagers** d'après [Maniatis & Reed, 2002]. Les machineries de la transcription et de la formation de la coiffe, de l'épissage et de la polyadénylation sont représentées schématiquement. *PIC* complexe de pré-initiation de la transcription ; *TF* facteurs de transcription ; *CAP* facteurs de la coiffe ; *SF* facteurs d'épissage ; *pA* facteurs de polyadénylation ; *CTD* correspond au domaine phosphorylé. Tous ces facteurs agissent de concert afin de maturer les ARN juste après la transcription.

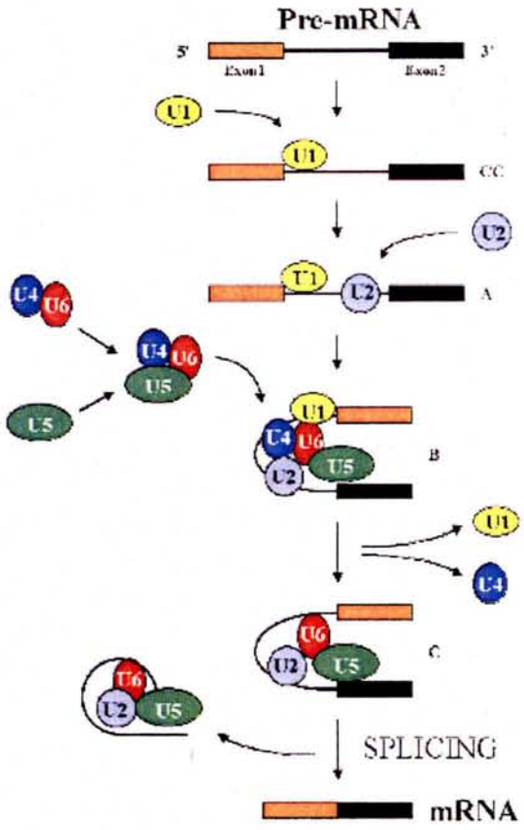

FIG. 1.3 – **Schématisation des principaux éléments qui composent le spliceosome : Macro-complexe ribonucléoprotéique** d'après [Will & Lührmann, 2001]. Le spliceosome est constitué de 5 UsnRNP et de nombreux facteurs protéiques. L'assemblage du macro-complexe modifie la conformation de l'ARN prémessager. Grâce au spliceosome peut avoir lieu la réaction d'épissage.

protéines (voir Figure 1.3 pour illustration). Ces particules nucléaires s'assemblent sur les jonctions intron-exon.

Les introns sont caractérisés par des séquences en 5' qui sont l'objet de l'attaque du complexe spliceosomal. Ce site est considéré comme le site donneur (SD). Une autre séquence en 3' de l'intron est l'objet d'une autre attaque du complexe. Ce site est considéré comme le site accepteur (SA). Ces deux séquences sont deux signaux caractéristiques nécessaires à l'épissage. Un troisième signal est la boîte de branchement (BP) qui se situe sur la séquence intronique. L'élimination des introns se fait par une double réaction de transestérification (représentée Figure 1.4). Le spliceosome va effectuer une première attaque nucléophile par la première réaction de transestérification. Cette réaction va relier l'adénosine de la boîte de branchement au site 5'. Cette réaction libère alors le premier exon. Cette première réaction produit également deux intermédiaires de réaction : l'exon 1, avec une extrémité 3' OH libre et un intermédiaire qui contient l'intron et l'exon en

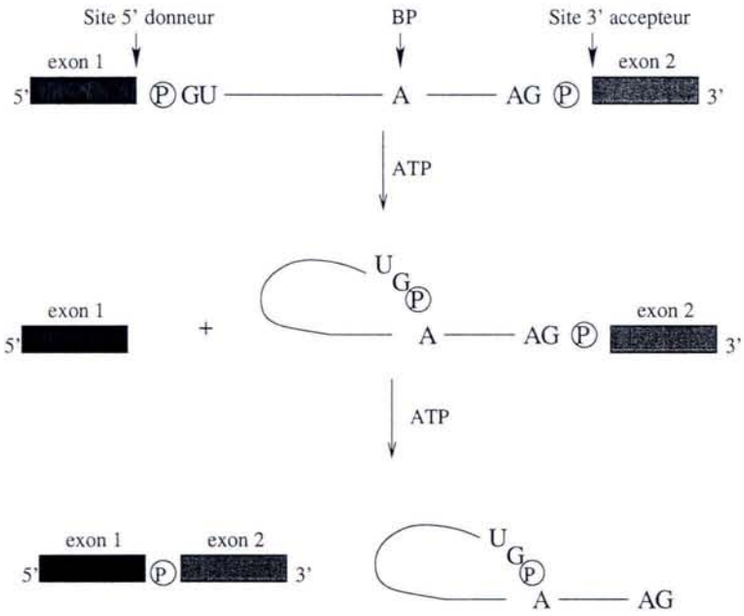

FIG. 1.4 – **Représentation de l'épissage sur une région intronique d'un ARN prémessager.** L'épissage consiste en deux réactions de transestérification. La première réaction relie l'adénosine de la boîte de branchement (BP) au site 5' (site donneur SD). La deuxième réaction relie les deux exons par le site 3' de l'intron (site accepteur SA) avec l'extrémité 3' de l'exon 1.

structure de lasso. A ce stade, on a effectué la première réaction d'épissage.

La deuxième réaction d'épissage est aussi une transestérification. La deuxième attaque a lieu sur le site 3' de l'intron. L'attaque nucléophile de l'extrémité 3'OH de l'exon 1 sur le phosphate en 3' de l'intron conduit à la ligation des deux exons et à la libération de l'intron sous forme de lasso. Après cette deuxième réaction d'épissage, on obtient les produits finaux de réaction d'épissage (voir [Moore *et al.*, 1993] pour revue).

1.1.2 Les signaux nécessaires à l'épissage

Les signaux d'épissage sont d'une importance cruciale dans le processus biologique. Il est ainsi rapidement apparu comme essentiel de les caractériser au mieux. Une approche consiste à effectuer un alignement multiple afin d'extraire des motifs consensus de ces sites. La séquence consensus établie par comparaison des sites 5' d'épissage des introns majeurs des vertébrés est `AG|GURAGU` [3] dans laquelle le dinucléotide `GU` est le plus conservé ([Moore *et al.*, 1993]). Les sites 3' sont quant à eux caractérisés par la séquence `YAG|G`[4]. Chez les vertébrés, ce motif est précédé par une séquence de 10 à 20 pyrimidines. Cette même opération de comparaison par un algorithme d'alignement sur le point de branchement (BP) permet de mettre en évidence la séquence `YNYURAC`[5]. Cette séquence est

[3] R = purine, | = jonction exon-intron
[4] Y = pyrimidine
[5] N = nucléotide quelconque

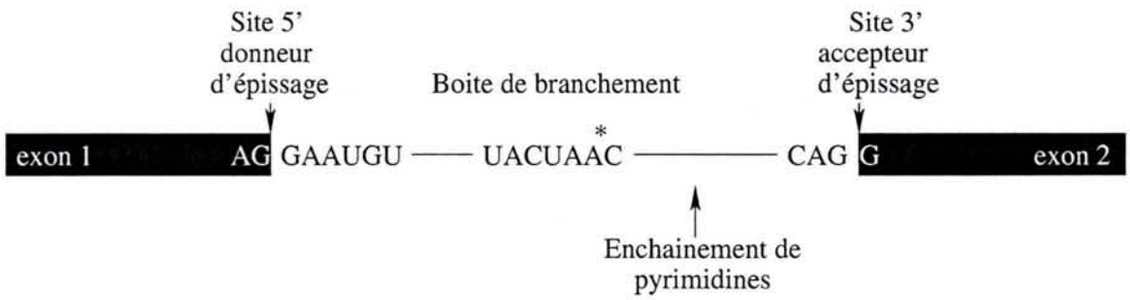

FIG. 1.5 – **Séquences exoniques et introniques essentielles à la réaction d'épissage.** L'astérisque désigne la localisation du point de branchement.

située de 18 à 40 nucléotides en amont du site d'épissage 3'. L'adénosine (A) est utilisée comme site de branchement. C'est en effet avec ce résidu que la réaction d'épissage possède la meilleure efficacité ([Hornig *et al.*, 1986, Query *et al.*, 1995, Query *et al.*, 1996]). L'ensemble de ces motifs est représenté dans la Figure 1.5.

De la capacité de la machinerie biologique à reconnaître ces signaux va dépendre le résultat de l'épissage. C'est une des raisons pour lesquelles les processus de régulation de l'épissage reposent sur l'efficacité de ces signaux d'épissage et sur l'existence de signaux correspondant à des sites de fixation de facteurs de régulation. Les processus de régulation qui vont permettre de choisir un site plutôt qu'un autre dépendent de divers acteurs que nous aborderons dans la Partie 1.3. D'un point de vue méthodologique, les sites 5' d'épissage sont relativement bien caractérisés, ce qui a permis le développement de méthodes informatiques efficaces pour les retrouver dans les génomes contrairement aux sites 3' qui sont dégénérés et par conséquent plus difficile à identifier.

1.2 Généralités sur l'épissage alternatif

Les eucaryotes supérieurs et les virus utilisent une variation du mécanisme d'épissage : l'épissage alternatif. Dans ce type d'épissage les exons ne sont pas définis de manière absolue. Ainsi la séquence qui est assimilée à un exon dans un contexte cellulaire peut ainsi devenir un intron dans un autre contexte. Le choix des introns et des exons est donc relatif. Il n'existe donc plus un seul mode d'épissage comme dans l'épissage constitutif exposé précédemment mais un grand nombre de choix alternatifs d'introns et d'exons. Ce processus est donc très répandu dans le règne vivant et possède divers avantages. Il permet notamment d'accroître la capacité codante des génomes. Chez la drosophile, l'épissage de l'ARN prémessager *Dscam*[6] peut ainsi produire 38 016 ARN messagers différents [Black, 2000]. L'homme utilise l'épissage alternatif comme un processus biologique majeur avec 35 à 60% des 30 000 gènes concernés ([McPherson *et al.*, 2001] [Venter *et al.*, 2001] [Modrek & Lee, 2002]). Une altération de ce processus peut provoquer des pathologies graves. L'épissage alternatif est un moyen supplémentaire pour les organismes de contrôler l'expression des gènes, mais il demande une régulation fine.

[6]pour *Down syndrome cell adhesion molecule*

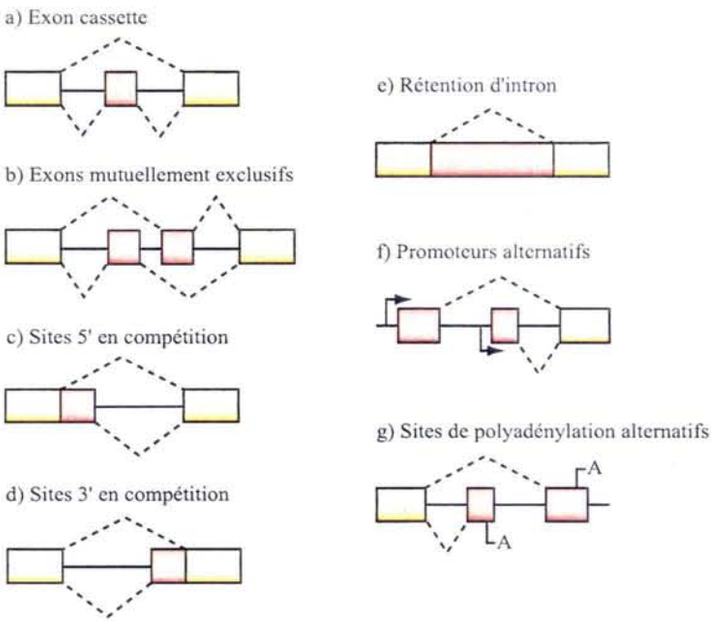

FIG. 1.6 – **Représentation des différents modes d'épissage alternatif (d'après [Roberts & Smith, 2002])** Les exons épissés constitutivement sont représentés par un rectangle orange, ceux épissés alternativement sont représentés par des rectangles roses. Les flèches désignent les promoteurs alternatifs et A les signaux de polyadénylation.

Les exons peuvent être classés selon différentes catégories en fonction du type d'épissage alternatif auquels ils sont soumis.

- **Les exons cassettes :** Ces types d'exons sont soit inclus soit exclus de l'ARN messager mature (voir Figure 1.6 a). Ils sont de petite taille et spécifiques à un tissu.
- **Les exons mutuellement exclusifs :** Ces exons sont présents en succession dans l'ARN prémessager. Ils possèdent des sites 5' et 3' fonctionnels, mais ne sont jamais présents ensemble dans les ARN matures (Figure 1.6 (b)).
- **Les exons variables du fait de la présence de sites donneurs ou accepteurs alternatifs :** Des sites d'épissage alternatifs peuvent être présents dans une séquence exonique et la sélection d'un des sites d'épissage alternatif entraîne l'exclusion d'une partie de l'exon dans l'ARN messager (Figure 1.6 (c) et (d)). Ce système est souvent utilisé par les virus. Cela permet de produire de nombreux ARN messagers différents. L'ARN prémessager du virus HIV-1 contient par exemple des sites d'épissage en compétition. Ainsi 4 sites donneurs et 8 sites accepteurs permettent la production d'une quarantaine d'ARN messagers différents ([?, PurcMart93].
- **Les exons avec la rétention d'un intron :** En fonction des conditions cellulaires, l'intron peut être inclus dans l'ARN messager. Dans ce cas, l'épissage est souvent tissu-spécifique (Figure 1.6 (e)).
- **Les exons dûs à la présence de promoteurs et de sites de polyadénylation :** Le choix du promoteur ou du site de polyadénylation influence l'identité des exons

terminaux (Figure 1.6 (f) et (g)).

Les différents modes d'épissage mentionnés permettent d'aboutir à des ARN messagers différents. Leur production est régulée selon la nature des sites d'épissage et la présence de séquences régulatrices. L'ensemble est fonction du type cellulaire, de l'état de différenciation de la cellule, de l'activation de voies de signalisation ou du génotype sexuel. Tous les ARN prémessagers qui sont régulés par ce type de processus biologique possèdent des sites d'épissage considérés comme faibles. Cette caractéristique rend compte de la faible affinité de ces sites pour la machinerie de l'épissage. La présence d'acteurs moléculaires supplémentaire est donc essentielle.

Les sites d'épissage constitutifs sont quant à eux très fortement conservés. Un alignement permet de mettre facilement en évidence le motif consensus. Nous avons mentionné précédemment l'efficacité des recherches informatiques de ces sites. Néanmoins, les sites d'épissage alternatif sont moins conservés et sont par conséquents plus divergents. Ils sont donc moins efficaces dans leur interaction avec le complexe spliceosomal. Cette propriété explique sans doute la difficulté à les mettre en évidence par des approches *in silico* efficaces.

1.3 Acteurs de la régulation de l'épissage des ARN prémessagers nucléaires

L'élimination des introns majeurs de gènes nucléaires de vertébrés met en oeuvre le spliceosome présenté précédemment. Il est composé de particules ribonucléiques et aussi des facteurs protéiques, comme les protéines SR. Les snRNP et les protéines associées reconnaissent des séquences qui sont nécessaires à l'épissage. On assiste alors à la formation d'interaction ARN-ARN et ARN-protéines de manière ordonnée et transitoire autour des sites d'épissage. Toutes ces interactions contribuent à former le macrocomplexe qu'est le spliceosome. La régulation s'effectue par le choix de ces séquences spécifiques. Le processus d'assemblage est une étape clef et souvent limitante de l'épissage. Cette étape est d'autant plus critique chez les eucaryotes supérieurs qui possèdent une grande taille d'introns (de 3 500 nts en moyenne avec un maximum à 10^5 d'après [Hawkins, 1988, Sterner *et al.*, 1996, Deutsch & Long, 1999] par rapport à la petite taille des exons (100 à 150 nts d'après [Berget, 1995]). Cette différence relative de taille rend difficile la recherche des sites d'épissage par le spliceosome. Pour combler cette incompétence, les premiers facteurs protéiques du spliceosome reconnaissent les exons. Une fois les sites 5' et 3' définis, les autres facteurs s'assemblent autour de l'ARN prémessager.

1.3.1 Les protéines SR

Ce sont les premières protéines qui interviennent dans l'assemblage du spliceosome chez les vertébrés. Nous présenterons ici leur structure moléculaire qui leur donne des fonctions essentielles lors du processus d'épissage.

Structure et fonction associées aux protéines SR

Les **protéines SR** possèdent un domaine C-terminal qui est riche en enchaînement de Serine (S) et Arginine (R), ce explique leur dénomination de protéines SR (pour revue, [Manley & Tacke, 1996, Graveley, 2000]). Ces protéines possèdent des caractéristiques communes [Zahler *et al.*, 1992]. On est alors en mesure de définir la famille des protéines SR par des critères expérimentaux et structuraux [Zahler *et al.*, 1993*a*] :

- Chaque protéine SR peut induire l'épissage si on l'ajoute à un extrait cytoplasmique S100 qui ne permet pas d'épisser à lui seul.
- L'organisation structurale des protéines est conservée quelle que soit la protéine. Elles se composent d'un domaine N-terminal contenant un ou plusieurs motifs RRM, et d'un domaine C-terminal qui possède un enchaînement de dipeptides arginine / serine que l'on appelle le domaine RS.
- Les protéines SR sont conservées en taille et en séquence dans tous les organismes chez lesquelles elles existent.
- D'un point de vue expérimental, les protéines SR sont solubles en présence de sulfate d'ammonium à 65 % et précipitent en présence de $MgCl_2$ à 20 μM.

On compte actuellement dix protéines chez l'homme qui correspondent à ces critères : SRp20, ASF/SF2, SC35, SRp30c, SRp40, SRp55, 9G8 et la protéine p54 qui est plus éloignée des autres en terme de séquence [Zahler *et al.*, 1992] [Ge & Manley, 1990][Krainer *et al.*, 1990] [Ge *et al.*, 1991] [Krainer *et al.*, 1991] [Fu & Maniatis, 1992] [Screaton *et al.*, 1995] [Zahler *et al.*, 1993*b*] [Cavaloc *et al.*, 1994] [Chaudhary *et al.*, 1991] [Zhang & Wu, 1996] [Soret *et al.*, 1998]. Les protéines SR se divisent ensuite en deux familles en fonction du nombre de RRM qu'elles possèdent. La famille qui contient les protéines SC35 et ASF/SF2 possède un domaine de fixation à l'ARN avec un ou deux éléments RRM. L'autre famille qui contient la protéine 9G8 possède dans son domaine de fixation à l'ARN un motif RRM et un motif dit à doigt de zinc. Ce motif particulier correspond à 4 résidus localisés sur un brin β et une hélice α arrangés autour de l'ion métallique zinc (pour revue [Mackay & Crossley, 1998]). Cette structure permet alors de définir une reconnaissance spécifique de l'ARN par la protéine 9G8 ([Cavaloc *et al.*, 1994, Cavaloc *et al.*, 1999*b*]).

Les domaines RS au niveau du domaine C-terminal des protéines permettent d'établir des interactions avec les domaines RS des autres facteurs protéiques. Ces interactions ioniques font alors interagir les arginines qui sont chargées positivement et les serines qui peuvent être phosphorylées. Cette phosphorylation joue un rôle important car c'est elle qui va pouvoir réguler les interactions entre protéines. Les protéines phosphorylées sont en effet nécessaires à l'assemblage des composants spliceosomaux ([Roscigno & Garcia-Blanco, 1995, Xiao & Manley, 1997]). Les protéines SR vont donc subir des cycles de phosphorylation-déphosphorylation de leur domaine RS lors de l'épissage des ARN prémessagers [Xiao & Manley, 1998].

Rôle des protéines SR dans l'épissage des ARN prémessagers nucléaires

Les protéines SR interviennent à différentes étapes de la formation du macro-complexe spliceosomal. Elles jouent donc à ce niveau un rôle majeur (pour revue [Graveley, 2000]).

Au cours de l'épissage constitutif, elles contribuent à définir les exons par la reconnaissance des signaux 5' et 3' et elles favorisent l'assemblage du spliceosome. Toutes ces fonctions dépendent de la structure des protéines qui permet de multiples interactions ARN-Protéine et Protéine-Protéine comme nous l'avons vu dans la Section 1.3.1. Elles constituent ainsi un *ciment* spliceosomal. Les protéines n'ont à ce niveau pas de spécificité propre. Elles étaient d'ailleurs initialement considérées comme redondantes entre elles [Manley & Tacke, 1996]. Les premières expériences ne permettaient pas en effet de différencier les fonctionnalités de chaque protéine SR.

Les protéines jouent donc un rôle important dans le mécanisme de base de l'épissage des ARN prémessagers nucléaires, mais elles jouent un rôle encore plus important dans le processus d'épissage alternatif. Elles influencent le choix d'un site d'épissage en compétition avec les autres. Des tests ont ainsi été effectués pour déterminer la capacité de chaque protéine SR à influer sur le choix d'un site d'épissage. Ainsi, malgré une apparente redondance dans l'épissage constitutif, les protéines SR possèdent une capacité d'activation spécifique de certains sites d'épissage [Manley & Tacke, 1996] [Graveley, 2000] [Smith & Valcárcel, 2000]. Par ailleurs, le profil d'épissage des ARN prémessagers dans les différents tissus peut être expliquée par les ratio des concentrations des différentes protéines qui varient en fonction des tissus [Zahler *et al.*, 1993*a*]. Chaque tissu possède un ratio différent qui permet d'exprimer un profil d'épissage des ARN prémessagers qui est différent.

1.3.2 Les protéines hnRNP

Dès la transcription, les ARN prémessagers sont pris en charge par un grand nombre de protéines dont les protéines hnRNP. Il existe une vingtaine de ce type de protéines dont la taille varie de 34 à 120 kDa. Elles sont présentes en quantité importante dans le noyau allant jusqu'à 10 μM pour la protéine hnRNP A1 [Dreyfuss *et al.*, 2002]. Elles sont localisées majoritairement dans le nucléoplasme, mais certaines d'entre elles sont considérées comme des protéines navettes, pouvant transiter entre le noyau et le cytoplasme. Par cette transition, elles suivent les ARN prémessagers qui deviennent les ARN messagers après maturation. Suivant leurs sites de fixation sur les ARN, les protéines vont interférer avec l'assemblage du spliceosome. Ces protéines jouent des rôles variés dans les étapes de synthèse, maturation, transport, traduction et dégradation des ARN messagers. Mais nous décrirons ici uniquement le rôle de ces protéines dans le processus d'épissage.

La protéine hnRNP A1 est la protéine de la famille qui est la plus étudiée. Le rôle de cette protéine a donc été bien documenté. Elle apparaît globalement comme une protéine inhibitrice d'épissage.

- Elle inhibe les sites 3' d'épissage. On observe notamment cette régulation dans la production de l'ARN messager *tat* du virus HIV-1. Au niveau de ces ARN, la protéine hnRNP A1 se fixe sur plusieurs séquences inhibitrices exoniques qui répriment les sites 3' A3 et A7 en amont. Cette fixation empêche l'assemblage des complexes spliceosomaux autour de ces sites d'épissage [Caputi *et al.*, 1999*b*, Marchand *et al.*, 2002].
- Elle inhibe les sites 5' d'épissage. Elle empêcherait la formation des complexes spliceosomaux autour des sites 5'. La protéine possède la capacité de polymériser. Lors-

qu'elle est présente en quantité importante, la protéine peut alors se multimériser sur la longueur de l'ARN prémessager. Cette fixation massive empêche la fixation d'autres facteurs de régulation comme la protéine ASF/SF2 [Eperon *et al.*, 2000]. Or ces protéines SR favorisent la fixation des facteurs spliceosomaux comme snRNP U1 au site 5' d'épissage. La compétition entre les deux protéines module le taux d'assemblage des facteurs du spliceosome. Nous illustrerons cet antagonisme entre les protéines régulatrices dans la section suivante.

1.4 Fonctionnalités des sites de fixation des protéines régulatrices

Les protéines SR sont les principaux activateurs de l'épissage des ARN prémessagers nucléaires. Ces protéines agissent le plus souvent en se fixant à des séquences exoniques. On retrouve cependant souvent des cas d'inhibition lorsque ces protéines se fixent sur une séquence intronique. Les sites de fixations sont déterminés expérimentalement par l'utilisation d'expériences SELEX que nous développerons dans le Chapitre 6. Cette approche expérimentale donne une représentation des séquences nucléiques sur lesquelles peuvent se fixer les protéines SR. Ces expériences ont mis ainsi en évidence que ce type de protéine est capable de reconnaître un grand nombre de séquences comme site de fixation. Cette hétérogénéité des sites reconnus permet de générer des fonctionnalités différentes.

Fixation des protéines SR sur les éléments ESE

Le plus souvent, les sites de fixation des protéines SR sont des éléments exoniques appelés **ESE**[7] (pour revue [Graveley, 2000]). Actuellement, on considère que les protéines SR permettent l'assemblage du spliceosome.

Les protéines SR fixées à un ESE facilitent l'épissage en interagissant avec les particules snRNP U1 ou le facteur U2AF[35] qui interviennent dans l'assemblage du spliceosome. Les protéines SR stabilisent ainsi ces facteurs protéiques sur l'ARN prémessager favorisant ainsi l'épissage (pour illustration voir la Figure 1.8). La localisation des ESE par rapport aux sites d'épissage est donc essentielle. Un autre critère important est l'affinité plus ou moins forte d'une protéine SR pour un site ESE donné [Graveley *et al.*, 1998]. Plus l'affinité sera forte et plus le site ESE jouera un rôle important dans l'épissage.

La faible affinité d'un élément ESE pour une protéine SR pourra être compensée par la proximité d'éléments semblables. En effet la présence de multiples éléments ESE augmentent ainsi la probabilité d'activer l'épissage.

Rôle antagoniste des protéines SR et hnRNP A/B

Les premières études ont rapidement mis en évidence le rôle antagoniste des protéines SR et hnRNP A/B (pour revue [Smith & Valcarcel, 2000]). Leurs sites de fixation sont généralement juxtaposés générant un gène stérique, empêchant la fixation des deux protéines de manière simultanée. Les protéines SR et hnRNP A/B sont alors en compétition

[7] pour *Exonic Splicing Enhancer*

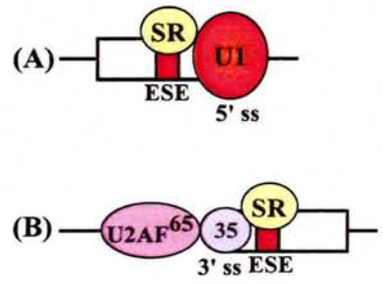

FIG. 1.7 – **La fixation des protéines SR sur l'élément ESE favorise le choix des sites 5' ou 3' d'épissage adjacents** (pour revue [Graveley, 2000]). **(A)** La fixation d'une protéine SR au voisinage du site 5' stabilise la fixation de la snRNP U1. **(B)** La fixation de la protéine SR sur l'ESE favorise la fixation de $U2AF^{35}$ sur la jonction intron-exon, ce qui permet de stabiliser la liaison de $U2AF^{35}$ sur la séquence polypyrimidine.

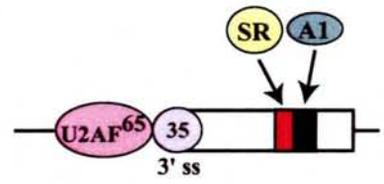

FIG. 1.8 – **Fonction antagoniste des protéines de régulation par sites chevauchant** (pour revue [Graveley, 2000]). Les protéines SR (SR) et hnRNP A/B (A1) sont en compétition par encombrement stérique. L'élément ESE est en rouge et le site de fixation de la protéine hnRNP A/B est en noir.

pour se fixer sur l'ARN prémessager. La fixation de la protéine SR à ESE active un site d'épissage en empêchant la multimérisation de la protéine hnRNP A/B qui bloquerait le même site d'épissage [Marchand *et al.*, 2002, Eperon *et al.*, 2000]. Le rapport entre les concentrations intracellulaires de ces deux protéines va donc permettre de moduler le choix d'un site d'épissage.

1.5 Conclusions

Les protéines SR ont donc un rôle de régulation marqué dans un processus biologique central. L'épissage alternatif peut avoir des conséquences énormes sur le cycle de vie d'une cellule. Connaître le rôle exact de ces protéines régulatrices est donc un enjeu majeur de la biologie actuelle. Nous avons mentionné seulement quelques interactions moléculaires mais qui dans un contexte expérimental s'avèrent générer une complexité combinatoire à la régulation de l'épissage. Le processus biologique apparaît alors naturellement difficile à étudier. Il est en effet difficile de comprendre la mécanique de régulation. Cette difficulté est accentuée par le fait que les sites de fixation des protéines régulatrices sont d'affinité variable et possèdent par conséquent des séquences mal conservées. Dans ce contexte expérimental difficile, l'approche *in silico* peut être un recours précieux pour faciliter les

démarches expérimentales. Néanmoins, les outils informatiques actuels ne suffisent pas à permettre des progrès dans la recherche des motifs de régulation. Il est donc important de proposer une nouvelle méthodologie. Pour être efficace, elle doit tenir compte du contexte biologique afin de se concentrer sur les propositions expérimentales facilitant les recherches sur un thème biologique précis. C'est pour cette raison que nous nous focaliserons dans ces travaux de thèse sur les sites de régulation d'épissage contrôlé par les protéines SR sur le génome du virus HIV-1. Ce contexte génomique permet d'utiliser l'expertise biologique et expérimentale à disposition pour mettre en oeuvre une nouvelle méthodologie qui sera validée sur un cas concret. Les modèles informatiques devront alors permettre d'intégrer l'hétérogénéité des sites de régulation qui jusqu'alors était ignorée.

Chapitre 2

L'épissage de l'ARN du virus HIV-1

L'étude de l'épissage alternatif est très complexe. Développer une méthode informatique pertinente nécessite de se placer dans un contexte particulier. La régulation de l'épissage est un processus biologique central dans le cycle de vie du virus HIV et sa compréhension peut permettre de développer de nouvelles thérapies. Une partie des recherches menées au MAEM concerne cette problématique. C'est dans ce contexte biologique précis avec des résultats et une expertise expérimentale forte que nous nous sommes placés afin de proposer une méthodologie informatique. Notre approche était donc initialement centrée sur un problème spécifique mais pourra être généralisée le cas échéant.

2.1 Généralités concernant le virus HIV-1

Le virus de l'immunodéficience humaine de type 1 (HIV-1) a été isolé en 1983 par les professeurs Luc Montagnié et Robert Gallo. Il est responsable de la pandémie mondiale qui touche 42 millions d'individus (ONUSIDA/OMS 2002). Le virus se manifeste par une immunodépression qui expose le patient à des infections opportunistes qui lui seront fatales. Le virus appartient à la famille des lentivirus qui sont caractérisés par une infection lente. Transmissible par voies sanguines ou sexuelles, le virus converge vers les ganglions lymphatiques où il infecte les lymphocytes T et les macrophages. L'ADN du virus est ainsi intégré dans le génome de l'hôte, où il ne peut être détruit sauf avec destruction de la cellule hôte. L'infection des macrophages par le virus n'est pas cytopathique, ce qui a pour effet de les transformer en centre de production de nouveaux virions.

Lors de la primo-infection, durant 2 à 6 semaines après contamination, le virus se multiplie et se dissémine dans l'organisme. Le système immunitaire réagit au bout de 6 à 8 semaines avant de produire une défense immunitaire afin de réduire la charge virale. Le stade suivant de l'infection par ce virus est la phase de latence pouvant s'étendre sur une dizaine d'années. C'est la phase de séropositivité qui est asymptomatique durant laquelle le système immunitaire est détruit progressivement. Lorsque le nombre de lymphocytes n'est plus suffisant, le système immunitaire succombe. Les maladies opportunistes se manisfestent à ce stade entrainant une phase symptomatique létale. C'est la phase clinique du SIDA.

Le virus est une particule de 100 nm de diamètre délimité par une enveloppe formée

d'une bicouche lipidique qui est un dérivé de la membrane de la cellule hôte. On retrouve à l'intérieur du virus deux molécules d'ARN, des protéines structurales et des protéines à activité enzymatique : l'intégrase, la transcriptase inverse et une protéase.

2.2 Cycle de vie du virus HIV-1

Le virus infecte les cellules du système immunitaire qui possèdent à leur surface les récepteurs CD4 et les co-récepteurs CCR5 et CXCR4 (pour revue [Tang *et al.*, 1999]). Une fois la reconnaissance faite, la membrane du virus va fusionner avec la membrane plasmique de la cellule hôte. Cette fusion des deux couches lipidiques permet à la capside qui renferme l'ARN du virus de pénétrer dans la cellule (voir Figure 2.1 pour illustration). La capside va ensuite se disloquer libérant son contenu dans le cytoplasme. L'ARN du virus est alors converti en ADN complémentaire (ADNc) double brin en utilisant la transcriptase inverse. Une fois l'ADNc obtenu, les protéines virales s'assemblent autour pour former un complexe de pré-intégration. Ce complexe est transporté dans le noyau de la cellule hôte. Une fois en place, l'ADNc est intégré dans le génome de la cellule grâce à l'intégrase. Cette intégration se fait au niveau des sites actifs de transcription.

A partir de cette étape c'est la machinerie de la cellule hôte qui va poursuivre le processus moléculaire. La transcription de l'ADN proviral va conduire à la production d'un unique transcrit qui aura deux rôles. Il subira toutes les étapes de maturation des ARN prémessagers. L'étape d'épissage permettra à cet unique transcrit de produire les différentes protéines virales. Le transcrit aura également une deuxième fonction, pour laquelle il devra rester intact. En effet, le transcrit joue aussi le rôle d'ARN génomique pour les nouveaux virions.

Les nouvelles protéines virales produites après l'épissage s'assemblent autour de l'ARN génomique. Cet assemblage permet la formation d'un nouveau virion qui pourra sortir de la cellule par bourgeonnement. Les nouvelles particules virales ainsi créées sont prêtes à infecter de nouvelles cellules.

2.3 Contrôle du cycle du virus par les protéines virales

Ce cycle de vie est contrôlé par les protéines virales qui sont produites à partir d'un unique transcrit. Le transcrit après régulation de l'épissage alternatif va produire une succession chronologique de protéines (voir Figure 2.2 pour illustration). Dans une première temps phase précoce de l'infection cellulaire, le transcrit est fortement épissé et les ARNm sont de taille avoisinant les 2 kb. Dans cette phase, ils produisent après traduction les protéines virales Nef, Tat et Rev.

- **La protéine Nef** [8], de taille 27 kDa, est la première protéine détectable après infection. Elle protège la cellule infectée de l'apoptose. Nef protège également la cellule infectée des attaques des lymphocytes en réduisant le nombre de récepteurs

[8] pour *Negative Factor*

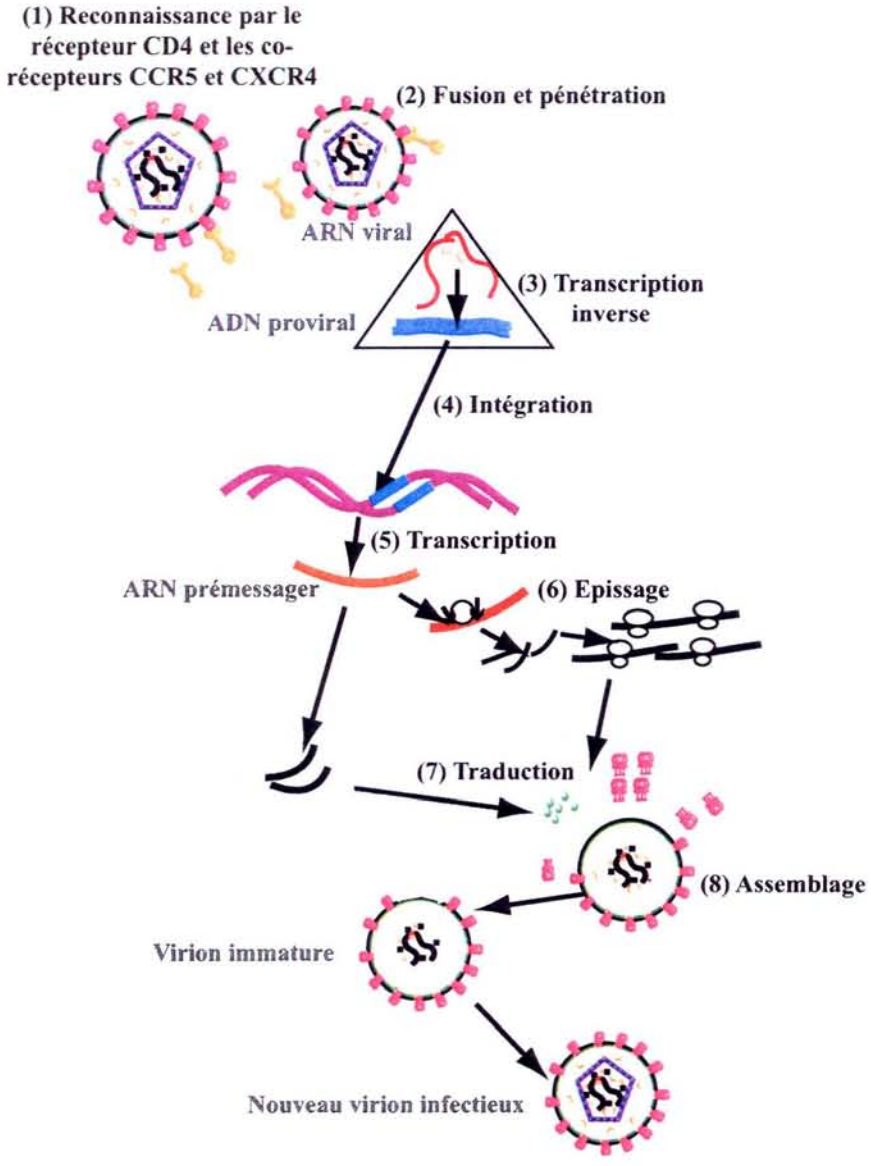

FIG. 2.1 – **Cycle réplicatif du virus HIV-1** d'après [Pavlakis & Felber, 1990]. Le cycle du virus se compose de 6 étapes. (1) fixation du virus sur les récepteurs cellulaires cibles et pénétration dans le cytoplasme, (2) transcription inverse de l'ARN viral en ADN double-brin, (3) intégration dans le génome de la cellule hôte, (4) expression des différents ARN messagers et transport vers le cytoplasme, (5) production des protéines virales et (6) assemblage des différents constituants sous la membrane cellulaire, ce qui va conduire à un bourgeonnement et la maturation de nouvelles particules virales.

$CD4^+$ à la surface de la cellule. C'est une protéine qui contribue à la pathologie du SIDA.

- **La protéine Tat**[9] est un activateur de la transcription. Elle active d'un facteur environ 100 la transcription de l'ARN viral. La production de cette protéine entraine donc un processus d'amplification de la production des protéines virales. Tat possède également la propriété de pénétrer dans les cellules voisines et conduit à leur apoptose, d'où sa contribution à la pathologie SIDA.
- **La protéine Rev** [10] joue un rôle primordiale dans l'expression des gènes du virus HIV-1, en permettant le passage de la phase précoce vers la phase tardive. Pour cela la protéine Rev agit en se fixant à un élément RRE[11] (pour revue [Hope, 1999]). Cette protéine, une fois fixée aux ARN, favorise leur transport dans le cytoplasme. De manière naturelle, les ARN prémessagers contenant des introns encore non épissés sont maintenus dans le noyau par interaction avec des facteurs d'épissage, et ce jusqu'à épissage complet ou dégradation. Dans les cellules infectées par HIV-1, les sites d'épissage ne sont pas de forte affinité pour les facteurs d'épissage. L'assemblage du spliceosome est ralenti. La protéine Rev peut donc se fixer sur l'élément RRE des ARN messagers qu'ils soient épissés ou non, pour les transporter du noyau vers le cytoplasme. La production de protéine Rev va donc permettre la production de protéines virales exprimées à partir des ARN faiblement épissées. Cette caractéristique est représentative des protéines de la phase tardive. Rev permet donc de passer de la phase précoce à la phase tardive en permettant l'exportation des ARN messagers même s'il subsiste des introns.

Les ARN messagers produits durant la phase tardive sont donc moins épissés que ceux de la phase précoce du fait de l'exportation dans le cytoplasme par la protéine Rev. Les ARN messagers sont de taille 4 kb ou 9 kb. Les protéines produites sont destinées à devenir des protéines virales auxiliaires (Vif, Vpu, Vpr) ou des protéines structurales (gag, env) qui serviront à la formation de la capside, et des protéines à activité enzymatique comme la transcriptase inverse, la protéase et l'intégrase.

Le cycle de vie du virus HIV-1 est donc principalement contrôlé par la régulation de l'épissage alternatif qui possède dans ce contexte des conséquences importantes. Il apparaît donc primordial de comprendre la régulation de l'épissage afin de mieux comprendre le cycle du virus et le cas échéant le contrôler.

2.4 Régulation de l'épissage du virus

Toutes les protéines du virus HIV-1 sont produites à partir du même transcrit. Le taux relatif d'expression et l'ordre chronologique d'expression au cours du cycle dépend d'étapes post-transcriptionnelles comme celle de l'épissage. Cette étape est d'une complexité combinatoire. L'ARN viral met en jeu 4 sites donneurs (D1, D2, D3, D4) et 8 sites accepteurs (A1, A2, A3, A4c, A4a, A4b, A5 et A7). Leur utilisation est combinée pour produire une quarantaine d'ARN messagers différents. Les ARN messagers possibles en

[9] pour *Transactivator*
[10] pour *REgulator of Virion expression*
[11] pour *Rev Response Element*

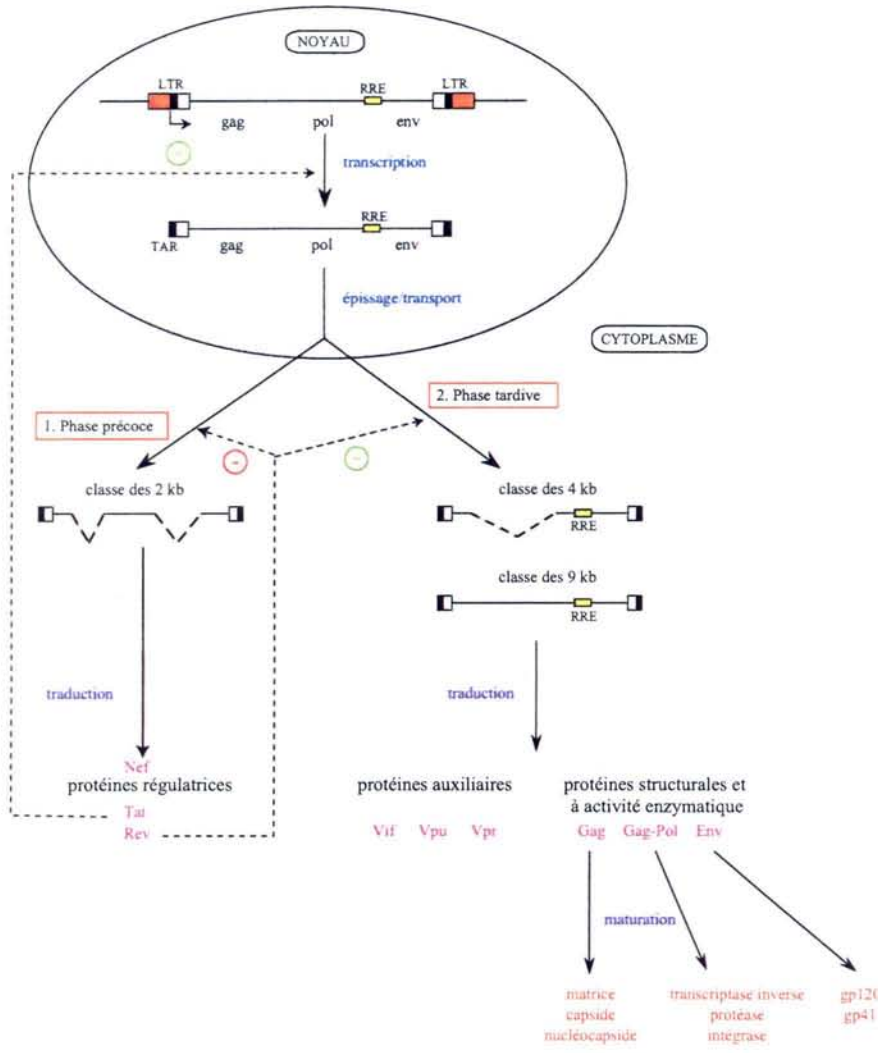

FIG. 2.2 – **Expression des gènes du virus HIV-1 au cours du cycle de vie** d'après [Ropers, 2003]. Le virus possède deux phases distinctes. La phase précoce permet la production des protéines régulatrices. Pour une quantité suffisante de protéine Rev, le virus passe en phase tardive. Au cours de cette phase, le virus produit les protéines structurales et celles à forte activité enzymatique.

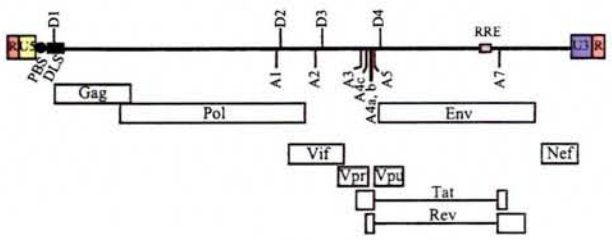

FIG. 2.3 – **Organisation du génome de HIV-1**. On représente les différents sites donneurs (**D**) et les sites accepteurs (**A**). On retrouve ainsi la composition en exons du génome et les phases codantes génératrices des ARN messagers pour les protéines de structure et de régulation du virus.

fonction des sites d'épissage utilisés sont schématisés sur la Figure 2.3.

Nous avons mentionné dans le chapitre précédant les protéines de régulation qui peuvent contrôler l'épissage alternatif. Les travaux expérimentaux développés au MAEM visent à identifier les éléments ESE et ESS qui permettent la fixation des protéines SR et hnRNP A/B sur l'ARN prémessager viral. En fonction de la combinaison des protéines fixées, l'épissage a lieu à un site accepteur ou à un autre. Pour identifier les sites, il faut réaliser des constructions moléculaires qui permettent de produire des ARN adoptant la structure secondaire de l'ARN entier. Nous mentionnerons ici les résultats obtenus par le laboratoire MAEM sur la région de l'ARN viral entourant le site A3 [Jacquenet *et al.*, 2001]. Ces travaux en tenant compte de la structure secondaire de l'ARN viral ont permis d'identifier les sites d'interaction des protéines nucléaires sur l'ARN viral (voir Figure 2.4 pour illustration).

Ces travaux du MAEM ont également permis de mettre en évidence des interactions fonctionnelles entre les éléments de régulations et d'émettre des hypothèses de fonctionnement comme nous le verrons dans le Chapitre 10.

2.5 Conclusions

L'infection par le virus HIV-1 est une pathologie touchant le système immunitaire dans laquelle le rôle de la régulation de l'épissage alternatif est prépondérant. Dans ce contexte biologique, on dénombre de nombreuses contraintes moléculaires qui compliquent l'approche expérimentale. Notre objectif sera donc de développer une méthode de modélisation permettant de faciliter la tâche des expérimentateurs étudiant la régulation des sites d'épissage. En effet, l'identification des motifs de régulation est laborieuse et tester les fonctionnalités de ces sites possède une combinatoire importante. Une approche informatique basée sur une modélisation peut être ici un atout pour les expériences du futur. Ce domaine biologique a été suffisamment travaillé expérimentalement pour que des travaux de modélisation puissent être abordés. Nous sommes donc ainsi en possession d'atouts favorables à une modélisation statistique et formelle d'un processus biologique aussi complexe que l'épissage alternatif.

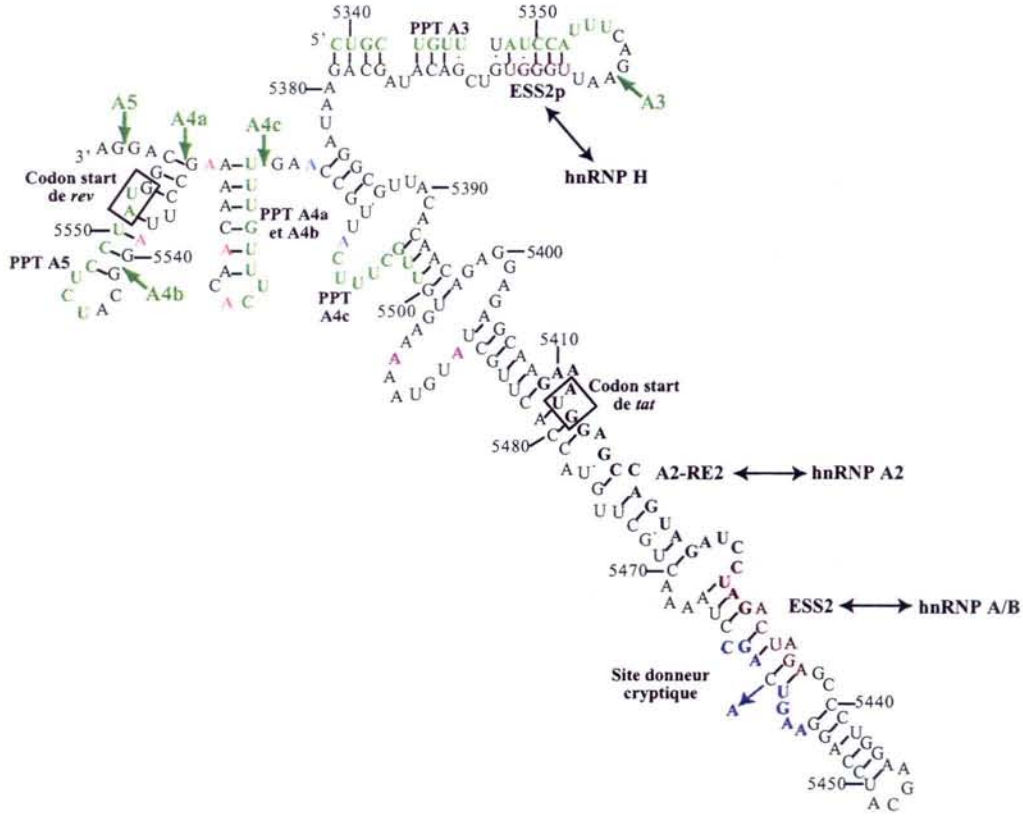

FIG. 2.4 – **Structure secondaire de la région de l'ARN viral entourant le site d'épissage A3** d'après [Jacquenet *et al.*, 2001]. On retrouve les sites accepteurs d'épissage de A3 à A5 ainsi que les boîtes de branchement associées. Les éléments régulateurs sont encadrés en rose ainsi que les protéines de régulation associées.

Chapitre 3

Modélisation de systèmes biologiques

Un des objectifs de la modélisation est d'organiser les concepts pour en comprendre les différents effets lorsque ceux-ci sont juxtaposés. La modélisation en biologie permet d'étudier les conséquences des nombreuses hypothèses biologiques agencées entre elles. De manière historique, les expériences biologiques permettent d'observer de manière systématique et efficace les systèmes vivants. Dans le cas où les observations ne seraient pas directes, la démarche consiste à analyser les résultats expérimentaux avec des outils statistiques ; c'est la **modélisation statistique**. Ces protocoles appartiennent à l'*approche naturaliste* qui optimise les observations du vivant pour comprendre son fonctionnement. Une autre alternative est possible avec la **modélisation formelle** des connaissances biologiques. Elle permet d'une part, une démarche de synthèse qui finalise les hypothèses qui ont été extraites directement des expériences ou après analyse. D'autre part, la modélisation formelle intervient en amont des expériences. Proche de la démarche acquise en physique, elle consiste cette fois-ci à élaborer et à modéliser les hypothèses pour les vérifier ensuite par les expériences. Il existe ainsi une grande diversité d'approches de modélisation formelle pour décrire le comportement d'un système ou issue de la statistique pour décrire un système. Face à cette diversité de modélisation possible, il apparaît important de rappeler la démarche méthodologique à suivre.

La modélisation de systèmes biologiques est récente et appartient historiquement aux domaines des mathématiques et de la biométrie. La notion de modèle en biologie est jeune et implique ainsi bons nombres de limitations. L'approche que nous allons développer dans ce chapitre se doit donc de rester modeste quant aux conclusions qu'elle peut apporter à la biologie compte tenu des approches déjà existantes. Les biologistes ont initialement utilisé les formalismes mathématiques, contribuant à l'émergence du concept de modèle en biologie [Legay, 1973]. D'où la notion actuelle de modèle telle qu'elle est définie dans le dictionnaire :

Modèle...II.1 : Structure formalisée utilisée pour rendre compte d'un ensemble de phénomènes qui possèdent entre eux certaines relations. modèle mathématique : Représentation mathématique d'un phénomène physique, économique, humain etc... réalisée afin de pouvoir mieux étudier celui-ci. (Petit Larousse 1998)

La modélisation apparait donc comme un outil formel qui permet d'étudier les propriétés d'un système vivant. La mise en place de cet outil repose sur une méthodologie générique.

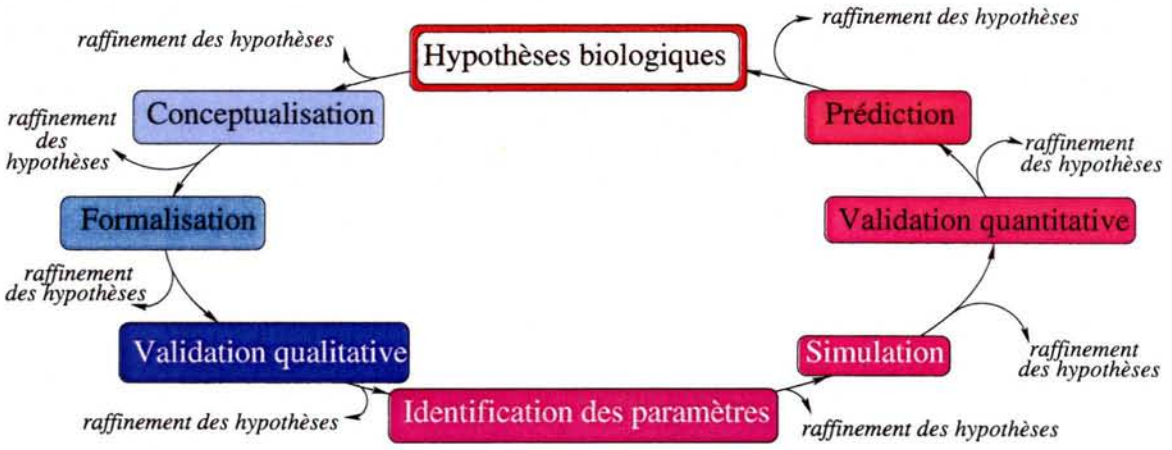

FIG. 3.1 – **Protocole de modélisation d'un système biologique**. Chaque étape est importante pour passer à la suivante. La non-satisfaction d'une étape nécessite la reformulation des hypothèses biologiques formulées au départ. Ce raffinement est la source d'interactions fructueuses entre les biologistes et les modélisateurs

3.1 Démarche de modélisation

La démarche rigoureuse de modélisation se décompose en différentes étapes (voir Figure. 3.1) qui doivent toutes être validées pour certifier un processus de modélisation efficace. Il est important de noter avant de décrire cette démarche la nécessaire collaboration qui doit exister entre les domaines mathématico-informatique et biologique. En effet certaines étapes nécessitent des expertises différentes qu'il faut faire converger pour valider une étape et passer ainsi à la suivante.

3.1.1 Détermination du référentiel

La première étape consiste à déterminer le référentiel biologique dans lequel le phénomène biologique se situe. On remarquera tout au long de la description, le parallèle qu'offre la modélisation biologique avec les protocoles de physique. De la même manière qu'un physicien qui détermine le référentiel de la dynamique qu'il étudie (galiléen, géocentrique etc...), le biologiste doit se restreindre dans une échelle de temps et d'espace donnée. Ainsi, on ne peut pas considérer la dynamique moléculaire tels que les mouvements browniens si l'on s'intéresse au développement cellulaire. Cette restriction ne revient pas à considérer cette dynamique comme inutile mais d'une complexité trop importante pour le problème biologique posé. Le référentiel est alors avant tout déterminé par la question biologique initiale. C'est ensuite au biologiste de déterminer en accord avec les données expérimentales à sa disposition de la nécessité d'étendre le modèle à des échelles d'espace et de temps qui sont accord avec les processus internes et externes du problème biologique considéré. Un des avantages de la modélisation du vivant par des systèmes formels tels que nous allons les présenter, est que l'on peut s'affranchir d'une échelle imposée par les expériences. Il est en effet possible d'intégrer dans un même modèle des données de diverses natures qui représentent le phénomène sur différentes

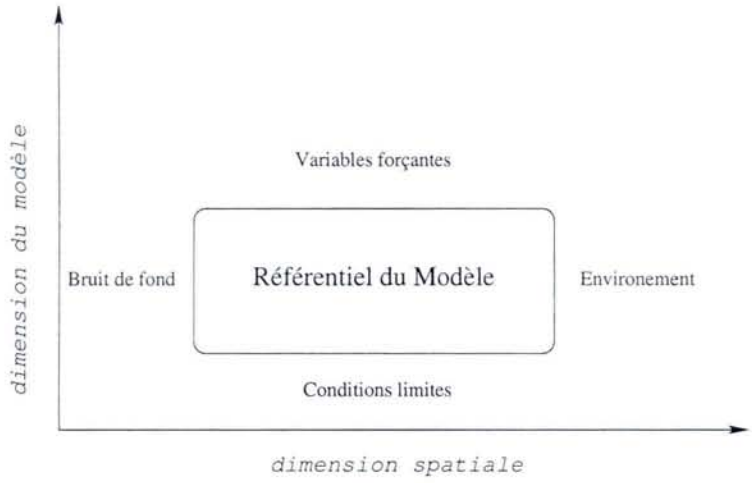

FIG. 3.2 – **Positionnement d'un modèle dans le contexte biologique et de modélisation.**

échelles. On peut ainsi appréhender un phénomène biologique de manière plus globale que par la simple juxtaposition expérimentale de données.

Lors de la détermination du référentiel, nous allons également considérer les processus extérieurs au problème biologique par des variables forçantes ou des conditions limites telles qu'elles sont représentées dans la Figure 3.2. Les conditions limites sont les processus qui existent dans une échelle suffisamment petite pour les considérer comme négligeable au cours de l'étude. Les variables forçantes représentent l'ensemble des processus appartenant à une très grande échelle par rapport à l'objet de notre étude. On considère communément que ces grands processus ont une incidence sur le modèle sans que le comportement du modèle résultant influe sur ce processus. Ils forcent ainsi le système biologique étudié sans interactions en retour. De manière similaire, on peut décrire le référentiel dans une échelle d'espace. La structure biologique trop petite pour être assimilée par le modèle sera considérée comme un bruit de fond et les structures trop grandes seront considérées comme l'environnement.

3.1.2 Hypothèses biologiques

Déterminer le référentiel inclut le fait de déterminer des hypothèses biologiques plus ou moins fortes pour représenter le système biologique dans son contexte. Cette étape délicate doit s'attacher à décrire les hypothèses biologiques qui négligent ou prennent en compte différents processus qui peuvent influer sur le modèle. Lors de la formulation des hypothèses, la composante biologique de la démarche de modélisation est la plus forte. En effet, une hypothèse biologique trop forte peut contraindre le système de manière trop importante. Inversement, négliger une hypothèse peut trop simplifier le modèle entraînant l'impossibilité de raisonner sur le modèle terminé. Une fois de plus cette étape doit être en accord avec le référentiel pour permettre de raisonner et ainsi d'apporter des questions sur les connaissances biologiques. A ce niveau, le biologiste doit pouvoir fournir les hy-

pothèses tant expérimentales qu'empiriques qui doivent être en accord avec les données à disposition. Lors de cette étape que les discussions sont souvent prolifiques. Les diverses discussions sont en effet en mesure de poser des questions concernant des raffinements d'hypothèses souvent insoupçonnées par les biologistes.

3.1.3 Conceptualisation

Une fois les hypothèses énoncées, il est possible de représenter le modèle sous forme d'un schéma conceptuel. Ce mode de représentation synthétise les hypothèses formulées au cours de l'étape précédente.

De manière générale, on représente ici la nature du système et les composantes relatives aux processus biologiques que nous cherchons à modéliser. On spécifie alors la nature du système que l'on cherche à modéliser, en relation avec le choix du référentiel que l'on a déterminé plus haut (voir la section 3.1.1) : le système peut être isolé s'il n'échange aucune interaction avec des processus extérieurs. Il peut être clos, s'il n'échange que de l'énergie. Il peut être enfin ouvert si le système échange de l'énergie et de la matière.

On spécifie également les *variables d'état* qui décrivent l'état du système biologique et qui peuvent être de différentes natures. Les variables d'état décrivent l'état du système. Les *variables d'action* peuvent modifier l'état du système par des actions externes. On considère ainsi des actions telles que les injections de produits de réactions externes au système biologique que l'on étudie dans le but de comprendre diverses stimulations externes (ex : expériences pharmaceutiques). On considère en dernier lieu des *variables d'observations ou observateurs* qui rendent compte de l'état d'un système. Ces variables peuvent être directement liées aux variables d'état.

Une fois les variables définies, on représente les contraintes entre ces différentes variables. On décrit ainsi les processus inhérents au système. L'analyse de systèmes biologiques expérimentaux classiques se restreint souvent à cette seule étape. Néanmoins la complexité de certains systèmes biologiques nécessite une étape de formalisation afin de permettre un raisonnement sur des systèmes devenus trop complexes.

3.1.4 Formalisation

Cette étape sera développée plus profondément lors de la section 4. Néanmoins, il est ici important de rappeler le fondement du choix d'un formalisme plutôt qu'un autre. Le formalisme représente le schéma conceptuel biologique contenant les diverses hypothèses dans un langage qui permet d'extraire des propriétés du système et éventuellement de raisonner sur le modèle biologique [Thieffry & Jong, 2002]. Le choix d'un formalisme plutôt qu'un autre est déterminé en fonction des objectifs à atteindre mais aussi des connaissances à notre disposition. Cette étape est également l'occasion d'une réflexion théorique sur les propriétés d'un langage formel et de son application à un système biologique donné.

3.1.5 Validation qualitative

La validation qualitative est une étape clef dans l'élaboration d'un modèle. C'est à ce moment que l'on peut commencer à valider l'approche. A cette occasion, nous chercherons

à comparer le comportement qualitatif d'un système avec le comportement qualitatif décrit par les expériences. Au cours de cette étape, nous raisonnerons sur le comportement qualitatif du modèle. Les techniques de validation et de raisonnement qualitatifs sont dépendantes des formalismes qui seront choisis afin de représenter le système biologique.

Globalement les techniques de validation qualitative reposent sur une formalisation partielle du modèle. On peut ainsi envisager plusieurs niveaux de description en fonction des hypothèses que l'on veut tester ou du comportement que l'on veut décrire. Une première approche naturelle sur un système continu consiste à analyser le comportement du système à l'équilibre et de comparer celui-ci aux données expérimentales. L'équilibre correspond à la phase stable du système, où les différents processus s'équilibrent telle des réactions stochiométriques. Néanmoins, les données biologiques à l'équilibre ne sont pas toujours celles qui prédominent. En effet, on observe plus volontiers expérimentalement les phases de transition entre les différentes phases d'équilibre. Ces phases sont alors une source importante d'informations à ne pas négliger lors d'une validation qualitative.

Une analyse plus fine sur un système continu consiste à décrire un système par des graphes de transition [Bernard & Gouzé, 1995*a*, Bernard & Gouzé, 1995*b*]. Cette approche décrit le comportement d'un système avec la description qualitative de la variation des variables au cours du temps. Les graphes que l'on obtient ne dépendent alors pas de la valeur des paramètres mais de la structure mathématique du modèle. Les hypothèses qui permettent d'écrire ces graphes à partir du système restent simples mais elles permettent de mettre en évidence trois applications importantes. D'un point de vue mathématique, c'est un outil d'analyse de modèles non linéaires [Bernard & Gouzé, 1998, Bernard & Gouzé, 1999, Bernard & Gouzé, 2002]. D'un point de vue plus pratique, cette méthode compare la dynamique qualitative d'un modèle aux différentes expériences. Si les transitions expérimentales ne correspondent pas à celles qui sont autorisées par le modèle, celui-ci n'est pas valide. Cela se manifeste par une transition qui est illicitement franchie.

Différents langages s'attachent à obtenir les mêmes avantages. On peut notamment citer les travaux de [Thomas *et al.*, 1995, Thieffry & Thomas, 1998] qui extraient un comportement qualitatif d'un système biologique en utilisant un formalisme *logique*. Parallèlement, [Jong & Page, 2000] recherchent des propriétés similaires par une discrétisation en différents états des variations du système par des seuils. L'espace de variation des variables d'état est alors défini comme étant une succession de domaines d'atteignabilité. Les passages des variables parmi ces domaines caractérisent un comportement qualitatif interprétable par le biologiste. Ces deux formalismes seront, ainsi que d'autres formalismes dit hybrides seront développés dans le Chapitre 4.

3.1.6 Identification

Le système une fois validé qualitativement, doit être calibré afin de répondre aux exigences des conditions biologiques. En effet, les coefficients que l'on utilise sont dépendants du référentiel. Il est à ce moment nécessaire de faire un tableau récapitulatif des coefficients et des unités que nous sommes amenés à utiliser. Les coefficients correspondent à des valeurs cinétiques de réaction ou des conditions de transition entre des états booléens. Les unités qui correspondent aux coefficients, doivent être homogènes, d'une part pour

respecter l'homogénéité du système mais aussi pour respecter une correspondance avec les données expérimentales dans le cas d'une modélisation en formalisme continu. On doit retrouver à ce niveau les valeurs de saturation des mécanismes biologistes ou les seuils de fonctionnement identifiés expérimentalement. Face à la diversité des valeurs de ces paramètres issues des expériences, une approche consiste à appliquer grossièrement une valeur moyenne. Cette approche loin d'être parfaite possède le mérite de simplifier le paramètre par incorporation des phénomènes sous-jacents.

Une autre approche que nous privilégierons au cours de cette thèse repose davantage sur les statistiques et l'optimisation. Nous avons en effet à notre disposition des données expérimentales. Il est alors possible par divers algorithmes tels que ceux de Newton, ou des moindres carré, d'optimiser le système construit par rapport aux données à notre disposition. Cette démarche est possible seulement après une validation qualitative de notre modèle. Une optimisation sur un système non validé aurait pour seul effet de reproduire les données sans pour autant réellement comprendre le comportement du système. Il serait alors impossible de vérifier la cohérence des connaissances que nous avons intégrées dans le modèle et encore moins de prédire les divers effets sur le système, ce qui est dans les deux cas un des enjeux majeurs de la modélisation biologique.

3.1.7 Simulation

Cette étape peut être considérée comme une contribution importante de l'informatique à la modélisation. En général, les systèmes mathématiques qui sont générés au cours de la modélisation biologique ne peuvent pas être résolus de manière formelle. Mais il reste néanmoins possible de simuler l'évolution des variables d'états en fonction des mécanismes de régulation. Les méthodes de simulation diffèrent en fonction du formalisme qui est utilisé. Dans un formalisme discret, l'estimation de la variable discrète s'effectue à chaque itération du calcul qui s'apparente alors à un pas de temps. Dans le cas de modèles continus, on utilise des méthodes d'intégration numériques. Cette approche décompose le temps en petites variations linéaires qui correspondra au pas de temps. A partir de conditions initiales, on calcule la variation linéaire pour un pas de temps. Plus le pas de temps sera petit et plus faible sera l'erreur dans l'estimation de la valeur. Cette erreur due au pas d'intégration est une erreur de troncature.

La simulation est un enjeu essentiel dans le cas de formalismes discrets que nous détaillerons par la suite. Il existe des simulations dites de Monte-Carlo qui mettent en oeuvre un nombre important de simulations afin d'obtenir un résultat appréciable par les biologistes.

Dans le cadre d'une formalisation continue, une autre démarche durant la simulation consiste à analyser la sensibilité du modèle aux valeurs des paramètres cinétiques. Par ce protocole, on estime et on apprécie l'action et l'importance de chaque coefficient sur les courbes simulées. Ainsi pour une variation δC_i d'un coefficient C_i, on estime le rapport $\frac{\delta C_i}{C_i}$. Pour une variation δC_i du coefficient, la variable V_i va aussi varier. Pour quantifier cette variation, on doit alors estimer le rapport $\frac{\delta V_i}{V_i}$ de la variable d'état. Cette valeur n'est pas quantifiable et on l'estime alors par le rapport $\frac{\Delta V_i}{V_i}$. On est alors en mesure de

quantifier la sensibilité avec le ratio :

$$sensibility = \frac{\frac{\Delta V_i}{V_i}}{\frac{\delta C_i}{C_i}}$$

Plus le coefficient de sensibilité sera fort et plus la variation imposée par le coefficient sera forte. Cette méthode permet de trier les coefficients qui sont importants. Une fois de plus, il est possible de faire le parallèle avec les connaissances biologiques du phénomène.

3.1.8 Validation quantitative

Au cours de cette dernière étape, l'objectif est de comparer cette fois-ci les valeurs numériques des résultats du modèle avec celles des données biologiques à disposition. Ce point de la modélisation est de nature statistique. Par des outils de multi-régression linéaire à non linéaire, il est possible de quantifier la disparité entre les résultats simulés et les résultats observés expérimentalement. Une approche à envisager à ce niveau est l'emploi de méthodes issues de l'apprentissage statistique telle que les *kernel CCA* [Saigo *et al.*, 2004] qui permettent une régression non linéaire. On est alors en mesure de valider quantitativement les modèles. Lors de la validation quantitative, il est essentiel d'utiliser des données qui n'ont pas été utilisées lors de l'identification des paramètres. Dans le cas contraire, une validation quantitative équivaudrait à seulement vérifier la qualité de l'optimisation des paramètres. Cette étape est alors délicate d'un point de vue expérimentale car elle nécessite d'avoir des données biologiques en nombre suffisant. La démarche sera encore plus performante si de nouvelles données expérimentales sont produites *a posteriori* de la démarche de modélisation.

3.1.9 Conclusion

La méthodologie exposée précédemment simplifie énormément l'approche de modélisation biologique. La méthodologie reste complexe et elle nécessite de nombreuses itérations qui permettent de raffiner le modèle. Cela permet de nombreuses interactions car elle fait appel à de nombreux domaines scientifiques tels que les outils formels parmi lesquels les mathématiques mais aussi des notions indispensables de biologie. Néanmoins la modélisation n'est pas un but en elle-même. Il faut en effet garder à l'esprit que la biologie doit toujours rester en amont du développement méthodologique de la modélisation. Une modélisation déconnectée d'un problème biologique concret ne permettra pas d'obtenir des résultats biologiques ce qui reste un des enjeux majeurs du modèle. D'autre part, il ne faut pas exploiter de manière excessive le modèle. Les propriétés émergentes qui sont déterminé formellement comme étant le comportement du modèle, doivent toujours être soumises à l'épreuve des expériences. D'après notre approche, la modélisation doit aider à élaborer théoriquement des concepts biologiques. Elle représente alors un outil qui doit être adapté aux problèmes du vivant. L'adaptation de la méthodologie de modélisation est dépendante du choix du formalisme qui conditionne les résultats de la démarche.

Chapitre 4

Formalismes en modélisation biologique

Formaliser permet de poser explicitement dans une théorie déductive les règles d'inférence selon lesquelles on raisonne. Dans le cas d'un système vivant, les règles d'inférence sont les règles qui décrivent les processus biologiques qui composent le système. Le choix du formalisme à utiliser pour modéliser un système est un élément crucial au cours de l'élaboration d'un modèle. Le formalisme conditionne les exploitations d'un modèle. Il existe un nombre relativement important de formalismes pour modéliser les processus biologiques. De manière grossière on peut les classer selon deux critères : continus ou discrets. Outre la culture du modélisateur, le choix d'un formalisme est motivé par le référentiel biologique et par les questions auxquelles le modèle doit répondre. La Figure 4.1 représente notre perception actuelle du lien qui existe entre un formalisme et le référentiel biologique associé. Une échelle biologique est définie notamment par la finesse de description d'un phénomène biologique.

Un système vivant est une coordination de processus biologiques. Un processus peut être décomposé en diverses interactions entre les éléments biologiques. Le formalisme discret de modélisation est bien adapté à cette dernière échelle puisqu'il permet de formaliser les interactions entre les éléments biologiques. La modélisation consiste alors à agencer des règles pour représenter au mieux un système. L'agencement de ces règles permet d'obtenir une représentation d'un processus biologique. Cette modélisation par formalisme discret se focalise ainsi à une échelle biologique plus grande. La complexité du modèle est dans ce cas plus importante. Des outils de **model-checking** sont alors adaptés pour interroger un système de cette complexité. L'extension du protocole de modélisation vers les grandes échelles biologiques peut ensuite se poursuivre avec notamment l'analyse de plusieurs processus.

Un processus biologique peut être modélisé avec un formalisme antagoniste : une approche continue. Le formalisme continu est historiquement le plus utilisé dans la modélisation de systèmes biologiques. Les premiers modèles biologiques se basaient sur des systèmes d'équations différentielles (**ODE**). Mais cette approche est limitée dans le type et le nombre d'équations que l'on peut gérer. En effet, les équations non-linéaires qui sont souvent bien représentatives d'un comportement biologique ne sont pas facilement analysables. Une démarche de discrétisation des équations différentielles compense ce biais

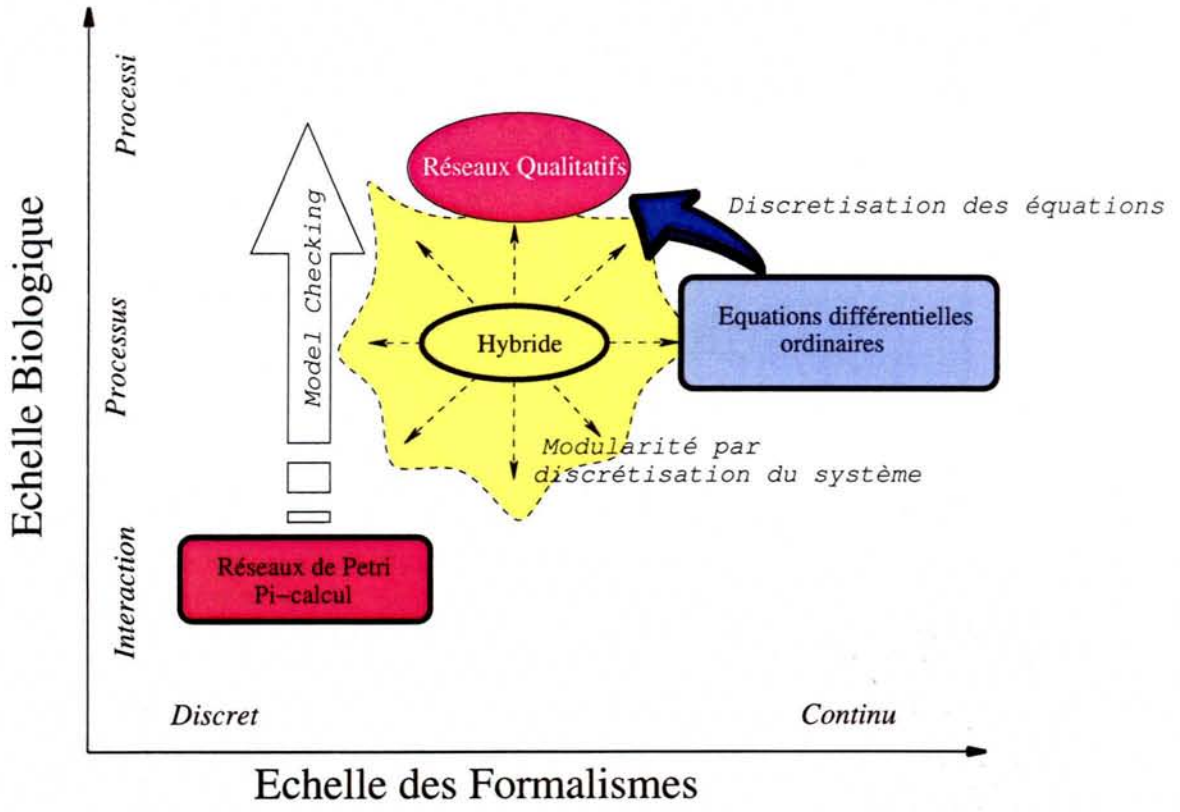

FIG. 4.1 – **Répartition des formalismes de modélisation suivant le référentiel biologique**. On peut représenter la diversité des formalismes sur un plan formé par les échelles biologiques et les formalismes. L'échelle biologique couvre les phénomènes biologiques des interactions aux ensembles de processus. Les formalismes sont représentés suivant une échelle des formalismes discrets à continus. Les formalismes discrets tels que les réseaux de Pétri et le π-calcul-us décrivent des systèmes biologiques au niveau des interactions simples entre objets biologiques. Ce type de formalisme est étendu aux échelles biologiques de plus grande taille par des techniques dites exploratoires de *model checking* représenté par en flèche. Les formalismes continus décrivent des ensembles d'interactions moléculaires avec des équations différentielles ordinaires (*ODE ou EDO*) ou partielles (*PDE ou EDP*). On décrit ainsi un processus biologique. Ces équations peuvent être discrétisées (flèche bleue) permettant ainsi de décrire le système qualitativement avec des seuils. Cette approche analyse ainsi des systèmes biologiques de plus grande taille tels que des ensembles de processus par des *réseaux qualitatifs*. Ces réseaux sont proches des formalismes hybrides qui décrivent des systèmes par des formalismes continus et discrets. Cette dernière famille de formalismes très générique possède l'avantage d'une grande modularité qui lui confère des propriétés intéressantes pour comprendre des systèmes de tailles restreintes à plus importantes.

en générant des **réseaux qualitatifs** qui appréhendent des ensembles de processus. Ces modèles qualitatifs sont considérés proches des formalismes hybrides à la frontière des formalismes continus et discrets.

Les formalismes hybrides, à défaut de discrétiser les équations mathématiques, discrétisent les processus qui composent les systèmes biologiques. Ils proposent pour cela de choisir de façon discrète entre plusieurs processus continus de manière à reproduire le comportement du système vivant. Cette démarche offre une grande adaptabilité du formalisme aux problèmes biologiques. Le formalisme hybride est également très modulaire. Il offre ainsi les avantages des autres formalismes strictement discrets ou continus.

Il existe donc beaucoup de formalismes qui possèdent des intérêts variés pour la modélisation des systèmes biologiques ([Jong, 2002] pour revue). Nous proposons de nous restreindre dans ce Chapitre aux principaux formalismes discrets (Section 4.1) et continus (Section 4.2) pouvant être utilisés dans notre modélisation. Nous détaillerons également d'autres familles de formalismes apparus récemment que l'on peut référencer comme hybride (Section 4.3) car ils associent l'approche discrète et continue. Parmi ceux-ci, nous nous attarderons sur l'utilisation des contraintes hybrides pour modéliser les systèmes vivants (Section 4.4).

4.1 Formalismes discrets

4.1.1 Les réseaux booléens

Dans certaines abstractions des systèmes génétiques, l'état d'un gène ou d'un élément biologique peut être assimilé à une variable booléenne. Cette variable s'exprime comme active associé à la valeur 1 ou inactive et associée à la valeur 0. L'interprétation biologique est alors la production active de l'élément biologique x pour $x = 1$ et l'absence de production pour $x = 0$. On peut alors représenter les interactions entre les variables booléennes correspondant à des éléments par des fonctions booléennes. L'ensemble de ces fonctions représente donc l'ensemble des interactions ce qui caractérise le réseau booléen.

Si l'on considère un vecteur $\hat{x}$ de longueur n représentant l'état d'un système biologique de n variables. Chaque $\hat{x_i}$ de $\hat{x}$ prend la valeur 1 ou 0 ce qui permet de représenter 2^n états possibles pour les variables du système. L'état de $\hat{x_i}$ au temps $t+1$ est calculé au moyen des fonctions booléennes (ou *règles*) $\hat{b_i}$ sur les n états des n éléments au temps t. La variable $\hat{x_i}$ est ainsi une sortie du système booléen décrit par un ensemble de fonctions booléennes et les n variables sont considérées comme les entrées. La dynamique d'un réseau booléen est donc dirigée par :

$$\hat{x_i}(t+1) = \hat{b_i}(\hat{x}(t)), \textit{avec } 1 \leq i \leq n \tag{4.1}$$

Avec $\hat{b_i}$ qui assigne aux n variables, les valeurs des sorties possibles à (t+1).

La structure d'un réseau booléen peut être représentée dans un diagramme comme sur la Figure 4.2. Le diagramme met en avant les transitions possibles entre les états. Ainsi dans l'exemple de la Figure 4.2, un vecteur [000] à t=0 deviendra un vecteur [011] à t=1.

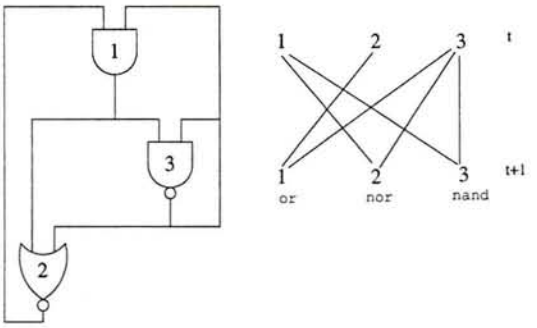

FIG. 4.2 – **Exemple de réseau booléen d'après [Jong, 2002]**. (a) représente un exemple de réseau booléen. (b) est le diagramme des transitions correspondantes. Les équations sont les règles booléennes associées au système.

Cette transformation est déterminée par les fonctions booléennes suivantes :

$$\begin{aligned} x_1(t+1) &= x_2(t) \quad \texttt{or} \quad x_3(t) \\ x_2(t+1) &= x_1(t) \quad \texttt{nor} \quad x_3(t) \\ x_2(t+1) &= x_1(t) \quad \texttt{nand} \quad x_3(t) \end{aligned}$$

Une modélisation avec ce formalisme est dite *déterministe* car une seule possibilité de résultat du modèle est possible pour une entrée donnée. Le système ainsi formalisé est également dit *synchrone* car toutes les sorties $\hat{x}_i$ du modèle change d'état en même temps. Lorsque les sorties varient indépendamment les unes des autres, le système est dit *asynchrone.*

Connaissant les transitions, il est possible de prévoir le comportement d'un système. Pour cela, on est en mesure de déterminer les différents états que le système va suivre au cours du temps. Les transitions entre les états pour des conditions initiales données sont appelées les trajectoires du système. Le devenir du système est alors prévisible si l'on analyse les trajectoires. Ainsi, en fonction des états que l'on atteindra, on peut caractériser un système qui va converger vers un état d'équilibre stable ou d'oscillations. Un état d'équilibre stable correspond à un état final unique alors que un état stable d'oscillations est caractérisé par plusieurs états finaux qui s'enchaînent de manière cyclique.

Ce formalisme est couramment utilisé pour la modélisation de grands réseaux issus de base de données (pour exemples [Akutsu *et al.*, 1999, Karp *et al.*, 1999]). Néanmoins dans ce type de formalisme, l'élément biologique est considéré comme *éteint* ou *allumé.* Dans le cadre du réseau booléen de base, les expressions intermédiaires sont alors ignorées. De plus, les transitions entre les changements d'état des variables sont considérées comme synchrones. Ce type de formalisme ne peut donc pas simuler des systèmes biologiques avec des transitions asynchrones qui sont courantes en biologie. Ces propriétés montrent que le formalisme booléen tel qu'il est ici présenté, n'est pas adapté à tous les types de modélisation biologique. Il existe cependant des extensions de ce formalisme qui permettent de raisonner sur un nombre fini de valeurs pour les variables : la forme multivaluée. Cette

extension permet de mieux représenter les systèmes biologiques avec notamment différents niveaux d'expression pour les variables biologiques.

Une illustration de cette extension des réseaux booléens est fournie par les travaux de [Thomas *et al.*, 1995] qui illustrent la volonté de raisonner qualitativement. Ces travaux représentent les concentrations des éléments biologiques par des variables booléennes (ou *logiques* d'après les auteurs) dans un système asynchrone. Ces travaux à l'interface des modèles continus que nous verrons par la suite, et les modèles booléens sont une forme de modèles hybrides qui rendent compte du comportement qualitatif d'un système vivant. Ainsi, le formalisme permet de raisonner sur les interactions entre éléments biologiques pour inférer des propriétés qui peuvent ensuite être validées ou invalidées par les expériences. Suivant une idée de [Monod & Jacob, 1961], [Thomas *et al.*, 1995] mettent en évidence la relation entre la structure des boucles de rétro-contrôle et les phénomènes biologiques comme l'homeostasis et la différenciation de gènes. Ils développent ainsi un réseau logique à michemin entre les réseaux booléens et les analyses du comportement qualitatif de systèmes biologiques par les mathématiques. Ce formalisme a fait ces preuves sur de nombreux systèmes biologiques complexes comme l'infection par le λ-phage [Thomas *et al.*, 1995] ou le développement dorso-ventral de la Drosophile [Sànchez & Thieffry, 2001]. Il est difficile de placer ce formalisme dans une classification des méthodes de modélisation. Bien que discret et inspiré des réseaux booléens, ce type de formalisme, par ces propriétés d'analyse, possède sa place dans les formalismes hybrides que nous allons développer par la suite. Par ailleurs, les avantages de cette méthode et son ancrage dans les mathématiques la justifie également dans la section dédiée aux formalismes continus.

4.1.2 Le π-calcul

Ce formalisme, comme les réseaux booléens, est un formalisme discret [Milner, 1993]. De la même manière que les réseaux booléens, le π-calcul utilise des règles pour modéliser les interactions des entités biologiques. C'est un formalisme qui modélise des systèmes dont les ressources et les structures varient au cours du temps. Il permet un calcul formel sur les processus concurrents qui se base sur des canaux. Les processus que l'on modélise interagissent en échangeant des noms de canaux.

Pour faire interagir les différents processus entre eux, le formalisme permet certaines règles. Ces règles définissent les changements d'état du système. Ces règles correspondent aux interactions entre les éléments biologiques. Néanmoins, le π-calcul nécessite certaines extensions pour être à même de modéliser un ensemble de processus biologiques qui composent le fonctionnement d'un système biologique.

Une des extensions possibles du π-calcul est stochastique. La différence avec le langage natif réside dans la gestion des canaux. Dans la version native, l'exécution des canaux se fait de manière non déterministe. L'extension stochastique les exécute en fonction de probabilités affectées aux processus biologiques.

Actuellement, différents travaux de modélisation biologique utilisent cette extension. C'est notamment le cas des modèles de cascades de réactions, où seules les interactions protéine-protéine sont prises en compte [Regev *et al.*, 2001]. Ce phénomène biologique est naturellement discret. De plus la perception des objets mis en relation dans ces interactions moléculaires est bien définie et très nettement documentée par les travaux de biologie

moléculaire et de génétique. L'inconvénient de ce type de formalisation reste néanmoins le besoin d'informations très fines concernant le phénomène biologique. C'est sans doute une des raisons majeures qui restreignent le nombre de modèles qui utilisent ce formalisme.

Un autre inconvénient de ce type de formalisme est le caractère descriptif des modèles que formalise ce langage. Les informations nécessaires à ce type de formalisation, sont souvent très complexes à obtenir expérimentalement. Par ailleurs, il sera difficile par du π-calcul seul d'inférer une information biologique nouvelle concernant un système vivant. Ce formalisme discret permet néanmoins d'explorer des modèles complexes par un outil de **model-checking**. Le model-checking est une méthode automatique pour la vérification initialement dédiée à l'analyse de circuits ou de programmes. Cette méthode vérifie des propriétés exprimées par le système dans une logique temporelle. Elle permet concrètement de déterminer les transitions permises par un modèle. Les fondements de la méthode sont actuellement utilisés en modélisation biologique pour valider un modèle [Chabrier & Fages, 2003]. Pour cela, elle vérifie des propriétés biologiques sur un système pour pouvoir les comparer à la réalité expérimentale le cas échéant. Cette démarche permet d'effectuer un raisonnement sur un système biologique descriptif. Il est ainsi possible de déterminer si un événement biologique est possible après un autre. Cette approche est applicable sur des systèmes de taille importante et permet de tirer des conclusions concernant un système complexe. Le modèle discret avec une analyse par model-checking possède alors une connotation qualitative qui rejoint ce que peuvent apporter les réseaux qualitatifs que nous verrons par la suite. Des domaines fonctionnels qui correspondent à l'ensemble des états possibles sont ainsi mis en évidence. Une convergence des réseaux qualitatifs et des méthodes de model-checking est ainsi prometteuse [Batt *et al.*, 2003]. Cette approche est applicable à de nombreux modèles

Un autre aspect intéressant du π-calcul est la possibilité d'appliquer des outils de réécriture. [Danos & Laneve, 2003] élabore dans cette optique un nouveau langage graphique (*graphic κ-calcul*) pour modéliser les interactions protéiques au niveau des différents domaines. Ce nouveau langage issu du π-calcul permet de décrire les différentes interactions possibles telles que l'activation, la complexation et d'autres formes d'interactions par des règles de réécriture de graphes.

La modélisation discrète des systèmes biologiques est un domaine plein de perspectives. C'est en effet un domaine d'applications privilégiées pour tout un domaine de l'informatique formelle. En effet, les systèmes biologiques par leurs complexités sont une source de motivation pour nombre d'outils informatiques théoriques à même de gérer cette complexité. Nous sommes dans ce cas encore face à l'exemple d'une modélisation biologique qui est source de progrès pour l'informatique

4.2 Formalisme continu

L'importance du formalisme continu est historiquement prouvée [Pavé, 1994]. La plupart des modèles dynamiques de systèmes en sciences ou en ingénierie utilisent des équations différentielles ordinaires (EDO ou ODE). Il est donc assez naturel d'utiliser ce formalisme en biologie [Cornish-Bowden, 1995, Heinrich & Schuster, 1996, Voit, 2000].

4.2.1 Fondements mathématiques

Ce formalisme considère les concentrations des éléments biologiques (ARN, protéines...) comme des variables du temps et de paramètres cinétiques constants. De manière plus précise, les régulations des ces variables sont modélisées par des *équations de taux* qui expriment la production de cette variable en fonction des concentrations des autres composants du système biologique. Les équations de taux ont donc la forme suivante :

$$\frac{dx_i}{dt} = f_i(X), 1 \leq i \leq n \tag{4.2}$$

Dans cette équation, $X = [x_1, ..., x_n] \geq 0$ est un vecteur de concentrations des variables du système et $f_i : \mathbb{R}^n \to \mathbb{R}$. Le taux de synthèse de x_i est dépendant des concentrations de X, de la fonction f_i et de la concentration d'éléments d'entrée du système $u \geq 0$. u sont souvent des nutriments dans un système écologique ou des protéines de régulation dans un système moléculaire. Le paramètre u qui correspond aux conditions initiales, est mathématiquement important rendant le problème valeurs initiales dépendant.

La représentation de toutes ces dépendances dans le temps se fait sous la forme d'un système d'ODE ([Voit, 2000] pour compléments). Dans ce système, la fonction f_i représente les différentes réactions biologiques comme la transcription, la traduction ou la diffusion des composants biologiques. Dans le cas d'un système en boucle qui est souvent associé à un réseau de régulation de gènes, les équations sont de la forme :

$$\begin{aligned} \frac{dx_1}{dt} &= \kappa_{1n} x_n - \gamma_1 x_1 \\ \frac{dx_i}{dt} &= \kappa_{i,i-1} x_{i-1} - \gamma_i x_i \quad 1 < i \leq n \end{aligned}$$

Les paramètres $\kappa_{1n}, \kappa_{21}, ..., \kappa_{n,n-1} \geq 0$ correspondent aux constantes de production et $\gamma_1, ..., \gamma_n \geq 0$ aux constantes de dégradation. Ce système exprime l'équilibre qu'il existe entre le nombre d'éléments biologiques qui sont produits et ceux qui disparaissent par unités de temps dans un système biologique.

Les constantes de production κ ne sont pas toujours des constantes. Ce taux peut dépendre de la concentration d'un autre composant du système lors d'interactions entre molécules. Il existe bon nombre d'expressions mathématiques pour raffiner l'expression de ce taux. Il est notamment possible de formaliser ce taux par une expression de la croissance de Hill :

$$h^+(x_j, \theta_j, m) = \frac{x_j^m}{x_j^m + \theta_j^m} \tag{4.3}$$

θ_j est le seuil de l'influence de régulation de x_j et $m \geq 0$ est le paramètre d'escarpement ou *steepness*. Cette fonction donne un résultat compris entre 0 et 1 et augmente pour $x_j \to \infty$. Ainsi, une augmentation de x_j aura tendance à augmenter le taux de croissance dans une ODE. On est alors face à une **activation**. Dans la mesure où l'on veut rendre compte d'une diminution du taux de croissance de l'ODE pour une augmentation de x_j, c'est à dire une **inhibition**, la fonction de régulation $h^+(x_j, \theta_j, m)$ est remplacée par

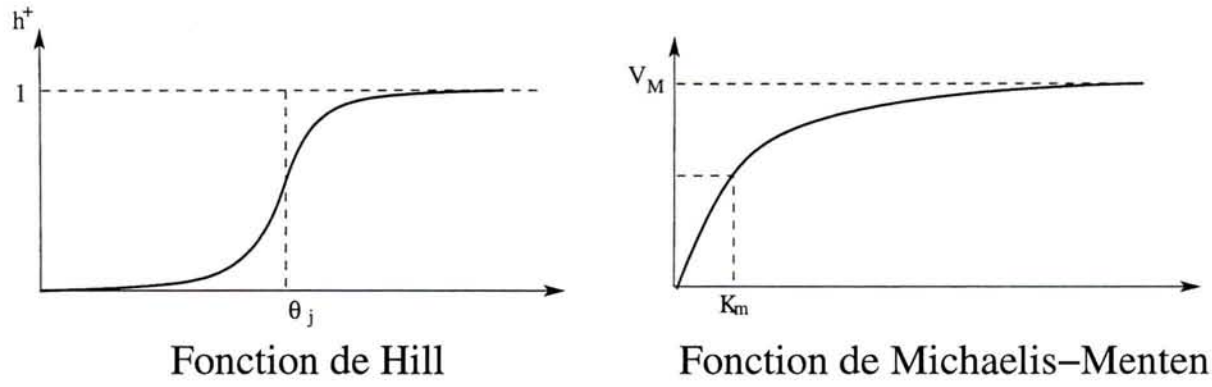

FIG. 4.3 – **Exemple de fonctions de régulation de croissance.**

$h^-(x_j, \theta_j, m) = 1 - h^+(x_j, \theta_j, m)$. Les fonctions de Hill ont une forme sigmoïde comme le représente la Figure 4.3.

Une autre fonction couramment utilisée pour rendre compte de l'interaction de différentes molécules sur le taux de croissance est la fonction de Michaelis-Menten empruntée aux biochimistes qui s'exprime selon la forme :

$$m^+(x_j, V_M, K_m) = \frac{V_M x_j}{K_m + x_j} \tag{4.4}$$

V_M correspond au taux de croissance maximal et K_m à la demi-constante de saturation. Les formes de ces fonctions de Michaëlis-Menten et Hill sont représentées sur la Figure 4.3.

4.2.2 Applications en modélisation biologique

Les expressions de f_i que nous venons de voir sont non-linéaires. Les paramètres sont difficiles à identifier pour rendre compte du comportement biologique. Dès lors, une résolution d'un système de ces équations comme le système 4.2 ne peut pas être analytique. Par ailleurs, dans certains cas spéciaux, il est possible de déterminer les propriétés qualitatives des solutions du système, comme le nombre et la stabilité des états d'équilibre. Ces états correspondent biologiquement aux équilibres stables identifiés avec le formalisme booléen. Les états d'équilibre sont déterminés à partir des propriétés du système mathématique. Il est notamment possible de faire le lien entre ces comportements qualitatifs avec les boucles d'interactions dans un système biologique comme le rétro-contrôle de certains processus métaboliques (pour revue voir [Smolen *et al.*, 2000]).

Dans le cas d'une inhibition d'un élément dans le réseau de régulation par un autre élément, le système possède une boucle de rétro-contrôle négative. Cette boucle est la condition pour que le système converge vers un état d'équilibre : **état stable**. La présence d'une activation dans le réseau de régulation engendre un rétro-contrôle positif. Le système tend alors à converger vers plusieurs états d'équilibre. Ainsi une boucle de rétro-contrôle négative est nécessaire pour une stabilité périodique et un rétro-contrôle positif nécessaire à la multistationnarité des états stables qui permet plusieurs états d'équilibre stables [Gouzé, 1998].

La simulation numérique avec le formalisme continu est largement utilisée pour modéliser des systèmes biologiques qui sont aujourd'hui des références dans la modélisation biologique. Une liste non exhaustive des modèles en fonction des problématiques est la suivante :

- le système du λ-phage par [McAdams & Shapiro, 1995]
- Opéron lactose de *Escherichiae coli* par [Wong *et al.*, 1997]
- l'expression du cycle de vie du virus HIV-1 par [Hammond, 1993]
- le rythme circadien de *Drosophila melanogaster* par [Golbeter, 1995, Leloup & Golbeter, 1998]
- contrôle de la mitose chez l'oocyte de *Xenopus* par [Tyson *et al.*, 1996] et [Tyson, 1999]

Le formalisme continu s'intéresse au travers de ces modèles à la dynamique d'un système dans le temps. Les modèles biologiques continus peuvent également intégrer une dimension spatiale qui peut s'avérer importante dans la compréhension du vivant. Pour cela, il est possible d'intégrer les équations de diffusion des physiciens dans les modèles biologiques pour y injecter une dimension spatiale [Murray, 2003]. Cette perspective de modélisation accentue un des avantages sous-jacents de la modélisation continue. Les systèmes d'équations différentielles partielles sont en effet utilisés par les physiciens pour décrire les comportements dans l'espace. Modéliser avec ce type d'approche permet d'utiliser un formalisme commun avec les sciences physiques ce qui renforce encore les liens entre les différents scientifiques communiquant ainsi sur un formalisme commun. C'est le cas notamment en océanographie où les modèles biologiques sont injectés dans les modèles physiques de circulation océanique pour déterminer le devenir d'une population circumplanétaire.

Malgré la profusion de modèles continus et les avantages d'une telle modélisation, ce formalisme nécessite de nombreux paramètres pour une modélisation biologiquement efficace. Les modèles de ce type sont alors limités par les expériences *in vivo* et *in vitro*. Il existe un réel problème au niveau de l'identification des paramètres. Une démarche alternative consiste alors à privilégier l'analyse du comportement qualitatif qui est indépendant de la valeur des paramètres.

4.3 Raisonnement qualitatif et automates hybrides

Malgré les connaissances expérimentales actuelles, il reste difficile d'évaluer les valeurs des paramètres des modèles. Les expériences *in vitro* ne représentent pas réellement le comportement du système et les expériences *in vivo* sont souvent trop complexes pour extraire des valeurs pertinentes pour identifier les paramètres cinétiques des modèles. De plus une imprécision sur de nombreux paramètres d'un système complexe continu peut présenter des effets multiplicatifs qui ne permettront plus d'appréhender correctement le comportement biologique [Thieffry & Jong, 2002]. C'est dans le but de répondre à ces difficultés que des travaux sont engagés sur une modélisation qui privilégie le caractère qualitatif des modèles.

Des méthodes de formalisation privilégient alors des formalisations qualitatives pour modéliser des systèmes malgré le manque de données à disposition. C'est le cas du for-

malisme logique que nous avons mentionné mais également le cas pour d'autres méthodes utilisant une approche qualitative sur des modèles dynamiques dans le temps. La tendance de ces méthodes est de discrétiser les modèles continus par ODE en utilisant un autre formalisme mathématique : les équations différentielles partielles (PDE pour *Partial Differential Equation*) [Edwards *et al.*, 2001]. L'idée est ici similaire à la méthode logique mentionnée précédemment. [Snoussi, 1989] a notamment démontré que le formalisme logique de [Thomas *et al.*, 1995] est une abstraction d'un cas spécial des équations linéaires par morceaux.

L'idée générale d'un modèle qualitatif est donc de faire une description semi-discrète d'un modèle continu afin de simplifier le comportement des équations continues. Il sera alors possible d'extraire des conclusions concernant la dynamique du système par une discrétisation de celui-ci. Il est pour cela possible d'utiliser une extension des ODE avec les **équations différentielles qualitatives (EDQ ou QDE)** (pour illustration voir [Jong, 2002]).

$$\frac{dx_i}{dt} = f_i(\mathbf{x}), 1 \leq i \leq n \tag{4.5}$$

f_i est une fonction de $\mathbb{R}^n \rightarrow \mathbb{R}$ qui peut être linéaire ou non linéaire. La description du système par les EDQ est formellement la même qu'avec les ODE. Néanmoins, les variables $\mathbf{x}$ prennent cette fois-ci des valeurs qualitatives composées de la *magnitude* qualitative et de la *direction*. La magnitude qualitative d'une variable x_i de $\mathbf{x}$ est une abstraction discrète de sa valeur réelle, tandis que la direction est le signe de sa dérivée. La fonction f_i est quant à elle, une abstraction des *contraintes qualitatives* qui restreignent les valeurs possibles des variables $\mathbf{x}$ [Kuipers, 1994]. Ainsi, à partir du système de QDE, il est possible de déterminer grâce à un algorithme QSIM et à partir d'*états qualitatifs* possibles, un arbre des comportements qualitatifs possibles. Chaque comportement de l'arbre décrit alors les transitions qualitatives possibles à partir de conditions initiales données. L'intérêt de ce formalisme est donc très fort dans le cas de systèmes complexes où seul le comportement qualitatif est connu. Cette méthode est actuellement utilisée pour modéliser la sporulation de *Bascillus subtilis* [Jong *et al.*, 2001].

Les formalismes qui permettent ainsi de raisonner qualitativement gèrent des formalismes discrets et formalismes continus. L'approche actuelle consiste à discrétiser les modèles continus pour les analyser. Tous ces formalismes tendent ainsi à décrire des automates qui peuvent gérer les deux types de formalismes. Ce sont des *automates hybrides*. Ces automates se justifient d'autant plus que certains phénomènes biologiques sont perçus naturellement comme discrets et d'autres comme continus. C'est le cas notamment comme l'état discret physico-chimique d'une molécule qui peut être phosphorylée ou non. D'autres phénomènes qui s'apparentent à des réactions chimiques sont quant à eux modélisés plus naturellement par un formalisme continu. Ainsi, en plus des propriétés des automates hybrides qui permettent de raisonner qualitativement sur des systèmes complexes, une modélisation hybride semble naturelle aux biologistes. Il existe néanmoins d'autres formalismes connus de la communauté informatique qui formalisent des automates hybrides. L'un d'entre eux est un formalisme qui utilise la programmation par contraintes.

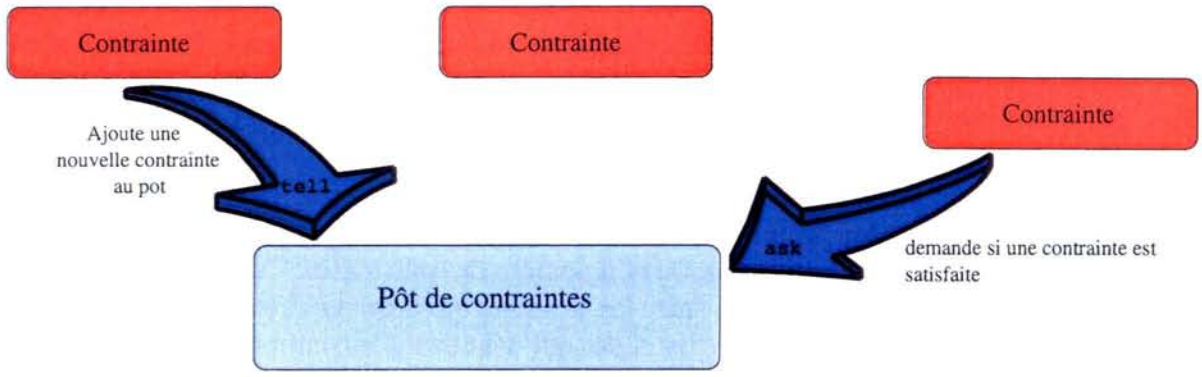

FIG. 4.4 – **Interaction des contraintes avec le solveur de contraintes.**

4.4 Contraintes en modélisation biologique

Certains formalismes hybrides comme ceux mentionnés précédemment sont actuellement utilisés avec succès pour modéliser les systèmes biologiques. Par ailleurs, de récents travaux de biologie [Palsson, 2000, Palsson, 2002] expriment le concept de contraintes biologiques :

Because biological information is incomplete, it is necessary to take into account the fact that cells are subject to certain constraints that limit their possible behaviors. By imposing these constraints in a model, one can then determine what is possible and what is not, and determine how a cell is likely to behave, but never predict its behavior precisely.

Ces travaux biologiques reflètent sans le vouloir le paradigme de la programmation par contraintes. Parmi le domaine de la programmation concurrente par contraintes, il existe un domaine informatique qui s'intéresse entre autres, aux contraintes hybrides qui permettent de formaliser des automates hybrides. Un automate est un dispositif formel qui est défini par les états et par les règles de transitions entre ces états. Un automate hybride possède des règles de transition hybride entre les états qui seront des contraintes continues et/ou des contraintes discrètes. L'automate permet alors de représenter un système biologique par un enchaînement de contraintes qui représenteront les processus inhérents au système.

4.4.1 Programmation concurrente par contraintes hybrides

Le fondement de la programmation par contraintes (*constraint programming*) est que chaque utilisateur spécifie des contraintes sur le comportement du système. Chaque contrainte exprime une information partielle sur l'état du système. Le **solveur de contraintes** vérifie leurs consistances et infère de nouvelles contraintes à partir de celles déjà assimilées. La **programmation concurrente par contraintes hybrides** ou **hybrid cc** permet l'emploi de contraintes qui sont des équations continues et/ou discrètes.

Hybrid cc [Saraswat *et al.*, 1996, Saraswat *et al.*, 1994, Saraswat, 1993,

Gupta *et al.*, 1995, Gupta *et al.*, 1998] est le fruit de multiples évolutions du langage initial de programmation par contraintes **cc** (*concurrent constraint*), auquel on a ajouté de nouvelles caractéristiques. La compréhension de ce type de programmation passe par la connaissance de son évolution de `cc` à `hybrid cc` que l'on utilise aujourd'hui.

Classe cc

Le paradigme de programmation concurrente par contraintes `cc` est un paradigme déclaratif, dans lequel on énumère un certain nombre de contraintes. Ce sont des formules logiques qui définissent des relations entre les variables qui décrivent l'état du système. En biologie, les contraintes seront toutes les règles qui définissent le comportement biologique en fonction des hypothèses du modèle. Toutes les contraintes sont stockées dans un **pot à contraintes** (voir illustration sur Figure 4.4) ce qui permet de restreindre les valeurs que prennent les variables au cours de l'exécution du programme. Un **solveur de contraintes** permet ensuite d'exécuter le programme.

Dans une programmation concurrente par contraintes (cc), chaque contrainte lance un processus et plusieurs processus peuvent être exécutés simultanément, et de manière concurrentielle. Il existe néanmoins une interaction entre les contraintes *via* le pot de contraintes. Les différentes actions possibles que le pot effectue sont soit des requêtes (**ask**), soit un ajout au pot de nouvelles contraintes (**tell**) [Saraswat, 1993].

- `tell` est un processus qui ajoute une nouvelle contrainte au pot. On augmente ainsi la quantité d'informations dans le pot concernant le système.
- `ask` est un processus qui permet d'interroger le pot pour savoir si une contrainte est impliquée par l'information (ensemble des contraintes) déjà présente dans le pot. Dans ce cas, des actions peuvent être exécutées.

Ce langage de programmation est dit monotone, ce qui correspond à augmenter l'information et on ne l'enlève jamais. Il possède néanmoins l'inconvénient majeur de ne détecter que la présence d'une information et non son absence.

Timed cc

Pour palier ce manque, [Saraswat *et al.*, 1994] propose d'ajouter au paradigme `cc` une phase d'exécution dans le temps pour transformer `cc` en `timed cc`. Ce langage permet ainsi de représenter le temps de manière discrète. A chaque temps discret, le programme `cc` est exécuté. A la fin de l'exécution, l'absence d'information est détectée et utilisée pour la phase suivante. Le résultat est cette fois-ci un langage de programmation synchrone et réactif.

Ainsi on est en mesure de détecter l'absence d'informations au bout d'un certain laps de temps qui correspond à un intervalle entre deux temps discrets. Ceci représente des systèmes contrôlés par des stimuli qui réagissent à des événements survenant au cours de l'exécution du modèle.

Default cc

`timed cc` ne permet pas néanmoins de la détection d'informations négatives de manière immédiate. Cette contrainte impose des modifications du langage. Ainsi, on ajoute

le combinateur négatif et `ask` pour définir une nouvelle requête négative :
`if` c `else` A
à la grammaire de `cc` pour obtenir le langage Default cc. Cette requête impose la contrainte (ou l'ensemble de contraintes) A à moins que la contrainte c soit vérifée. L'action A est ainsi perçue comme l'action par défaut du système.

Cette requête est utilisée dans le cas de l'indécidabilité d'une action, notamment lorsqu'on ne sait pas si une condition est juste ou fausse. Dans ce cas, on exécutera alors l'action définie par `else`.

L'exécution de `default cc` est similaire à celle de `cc`. Ce langage émet cependant une hypothèse sur le résultat final avant l'exécution de l'ensemble des contraintes. L'introduction de ce langage dans `timed cc` aboutit à `timed Default cc` [Saraswat *et al.*, 1996]. Ce dernier langage nécessite en effet que de l'ajout de `hence` A qui permet d'exécuter la contrainte A à chaque phase de l'exécution du programme. Il est ainsi possible de simuler un comportement continu. D'autres types de contraintes peuvent être définis grâce à ces contraintes de base. Par exemple `always` A correspond à A, `hence` A qui décrit que A existe maintenant et à partir de maintenant, donc toujours.

Hybrid cc

Toutes les évolutions des langages par contraintes permettent d'aboutir à `hybrid cc` [Gupta *et al.*, 1998, Gupta *et al.*, 1995] qui gère les contraintes comme précédemment avec des contraintes continues et/ou discrètes. `hybrid cc` est une extension de `default cc` sur un temps continu.

Il permet premièrement de prendre en compte des contraintes continues. Les contraintes sont alors des ODE qui s'expriment en fonction des conditions initiales. Deuxièmement, l'opérateur `hence` est interprété sur du temps continu. Cela impose d'appliquer la contrainte A sur chaque pas de temps. L'évolution d'un système de contraintes au cours du temps est continue par morceaux avec un enchaînement de points de calcul et d'intervalles.

Le déroulement d'une exécution s'opère par phases, discrètes ou continues. Au cours de l'exécution on résout les différentes équations. On détermine alors les éventuelles incohérences survenues après l'émission d'une hypothèse, ou alors on ajoute une nouvelle assertion au pot si le programme s'est déroulé correctement. Le langage effectue donc automatiquement un raisonnement sur le système que l'on modélise. Tous les changements discrets sont effectués à chaque changement de phase comme lorsqu'un simple programme en langage `default cc` est exécuté. Dans une phase continue, la résolution des contraintes s'effectue tout au long de l'évolution du temps. Pour une contrainte qui définie un intervalle dont la taille est estimée au point précédent. Cet intervalle est à nouveau exprimé dès que les conditions changent [Gupta *et al.*, 1995]. Le Tableau 4.1 résume les différents combinateurs possibles dans `hybrid cc` pour modéliser le vivant.

Ainsi qu'il est expliqué dans [Bockmayr & Courtois, 2001] et [Bockmayr & Courtois, 2002], le langage `Hybrid cc` est relativement bien approprié à la modélisation de systèmes dynamiques biologiques. Il permet de simuler des systèmes complexes mais surtout de raisonner sur un modèle comme les formalismes qualitatifs. Un autre avantage est de pouvoir gérer le manque d'informations ou l'incertitude des

TAB. 4.1 – Combinateurs de `Hybrid cc`

Agents	Propositions
c	c holds now
`if` c `then` A	si c est vrai, alors A est vraie
`if` c `else` A	si c n'est pas vrai, alors A est vraie
`new` X `in` A	there is an instance $A[T/X]$ that holds now
(A, B)	les deux contraintes A et B sont vraies
`hence` A	A est vraie à chaque instant à partir de maintenant
`always` A	est la même chose que $(A,$ `hence` $A)$
`unless`(c) A `else` B	est la même chose que (`if` c `then` B, `if` c `else` A)

paramètres biologiques en raisonnant sur des intervalles au lieu de valeurs finies. D'après une discussion générale dans [Bockmayr & Courtois, 2002], nous proposons d'illustrer ceci par de petits exemples biologiques. Ils permettent d'observer la combinaison des combinateurs de `Hybrid cc` qui modélisent naturellement les systèmes biologiques. Nous testerons alors les différents avantages qu'offre la programmation par contraintes dans la modélisation d'un système biologique.

4.4.2 Contraintes d'intervalles et dynamique continue

Le langage `Hybrid cc` que nous utiliserons est basé sur les contraintes d'intervalles [Carlson & Gupta, 1998]. Cela signifie que les variables peuvent être définies sur un intervalle de nombres réels et que les calculs sont arithmétiques sur l'intervalle. Cette propriété du langage est très utile en biologie où les paramètres et variables caractéristiques ne sont pas connus de manière déterministe. Cela permet alors une certaine fluctuation qui représente la variabilité expérimentale et/ou la variabilité biologique. Cette caractéristique est illustrée par un exemple simpliste de système biologique en `Hybrid cc` avec une unique contrainte associée à une variable sur un intervalle x (voir Figure 4.5). Afin de raisonner sur le système dynamique, nous utiliserons le combinateur `always` A, qui exprime que la contrainte A est vraie à chaque instant. Le code suivant permet de simuler la variation d'une variable sur un intervalle. Un des atouts biologiques du langage est dans un premier temps son ergonomie. Il est en effet relativement aisé de comprendre la programmation et de programmer un système biologique avec `Hybrid cc` comme le montre le code suivant :

```
interval x;
x = [9.5,10];
always { x' = -(2*x)/(15+x);
       }
sample(x);
```

Cette démarche possède un fort potentiel dans la possibilité de représenter les comportements extrêmes d'une variable biologique. Les expériences biologiques sont en effet basées

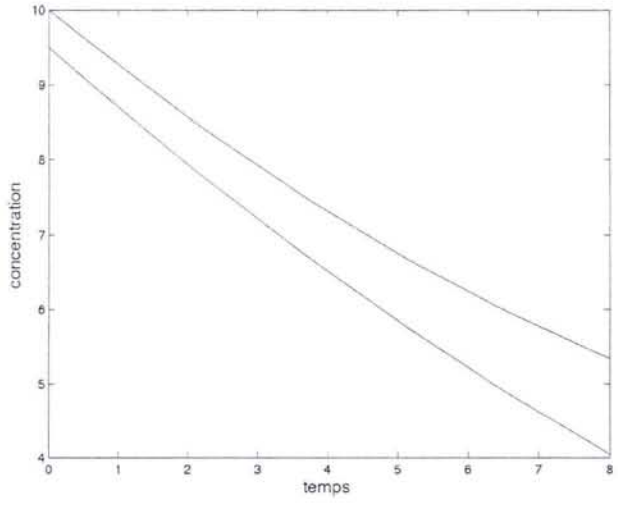

FIG. 4.5 – **Domaine d'atteignabilité d'une espèce moléculaire à cinétique linéaire**. On représente la concentration d'une molécule qui décroît de manière linéaire. L'incertitude concernant la concentration initiale se propage au cours de la simulation.

sur la comparaison de systèmes à l'état dit *naturel* ou *blanc*, avec des systèmes stressés par des mutations ou des conditions expérimentales différentes. Par les différences qualitatives ou quantitatives entre les données aux différentes conditions, il est possible d'inférer sur un comportement biologique. La biologie expérimentale est alors confrontée à un problème dual. La variabilité peut être interprétée de différente manière.

La première est étroitement associée à une interprétation statistique des données. Les fluctuations sont associées à des variabilités expérimentales. Il est alors possible d'extraire des données, une moyenne et un écart-type qui permettent une meilleure représentation des données. Un des avantages des contraintes est de pouvoir observer la propagation de l'intervalle défini par l'écart type. On est alors face à un outil qui peut s'avérer efficace pour une validation quantitative d'un système biologique. On peut en effet simuler le comportement du système avec une variation statistique correspondant aux fluctuations expérimentales.

Une autre interprétation est de considérer la variabilité comme étant dépendante directement du système biologique. Une hypothèse biologique est de considérer que l'ensemble des processus biologiques se situe au sein des processus extrêmes. La modélisation de ces processus permettra alors d'englober le comportement naturel. Cette hypothèse est décrite plus spécifiquement par [Palsson, 2002] sans qu'il utilise spécifiquement la programmation par contraintes pour modéliser les systèmes biologiques.

Il est néanmoins clair qu'une réponse à ce problème dual de l'expérimentation se situe entre ces deux hypothèses extrêmes. Malgré l'explication que l'on donne à la fluctuation, expérimentale ou naturelle, la programmation par contraintes intègre les deux de manière équivalente au sein d'un même langage de programmation laissant présager de très grandes capacités de modélisation encore peu exploitées.

Néanmoins la programmation par contraintes possède des limites assez marquées dans une modélisation sur les intervalles. En effet, pour un système biologique non linéaire, le solveur de contraintes est rapidement dans l'incapacité numérique de résoudre le système.

4.4.3 Composition parallèle

`Hybrid cc` permet l'utilisation de contraintes en parallèle. (A, B) impose ainsi le contraintes pour A et B. D'un point de vue opérationnel, le programme (A, B) se comporte de la même façon que l'exécution simultanée de A et B. Les contraintes A et B peuvent aussi agir sur les mêmes variables et ainsi communiquer par l'intermédiaire du pot des contraintes. Nous proposons d'illustrer ces propos par un petit exemple en `Hybrid cc` qui spécifie une cinétique de Michaelis-Menten entre deux éléments. On considère ainsi deux espèces moléculaires X et Y de concentrations x et y. X est transformé en Y. La concentration initiale est comprise dans l'intervalle $[14, 14.5]$. Le taux de production de Y dépend de la concentration de X suivant la formule michaelienne suivante :

$$y' = (A_{max} \cdot x)/(k_s + x)$$

pour les constantes A_{max} et k_s. La concentration de X diminue de manière symétrique avec la même cinétique. On représente ainsi le passage de la molécule X vers Y.

Afin de permettre la modélisation du système biologique sur un intervalle, il est préférable d'apporter une contrainte supplémentaire avec la conservation de la matière. Il est en effet quelque fois possible de supposer que la matière biologique se conserve au long de la réaction. Cette contrainte supplémentaire permet de gérer la complexité du traitement des contraintes par le solveur. Ainsi nous ajoutons la contrainte $x, y \geq 0$ pour représenter que les concentrations en matériel biologique ne peuvent être négatives et $s = x+y, s' = 0$ pour exprimer la conservation de la matière. Le solveur de contraintes permet alors de définir les domaines de concentrations possibles pour les éléments biologiques (voir Figure 4.6). On peut observer plus particulièrement qu'en fin d'expérience que la concentration de l'élément y et plus importante que celle de l'élément x. L'utilisation des intervalles peut être également très utile au cours de l'analyse de sensibilité du modèle. Au cours de cette étude, l'objectif est de tester l'influence des paramètres sur le comportement du modèle. On peut ainsi observer la propagation de la fluctuation d'un paramètre sur un intervalle. Le code suivant permet de simuler le comportement de la Figure 4.6.

```
#define ks 1.5
#define Amax  2
interval x,y,s;
x = [14,14.5];              /* Initialization */
y = 0;
always {
   x' = -(Amax*x)/(km+x); /* Michaelis-Menten kinetics */
   y' =  (Amax*x)/(km+x);
   x >= 0;                /* Non-negative concentrations */
   y >= 0;
   s = x + y;             /* Conservation of matter */
```

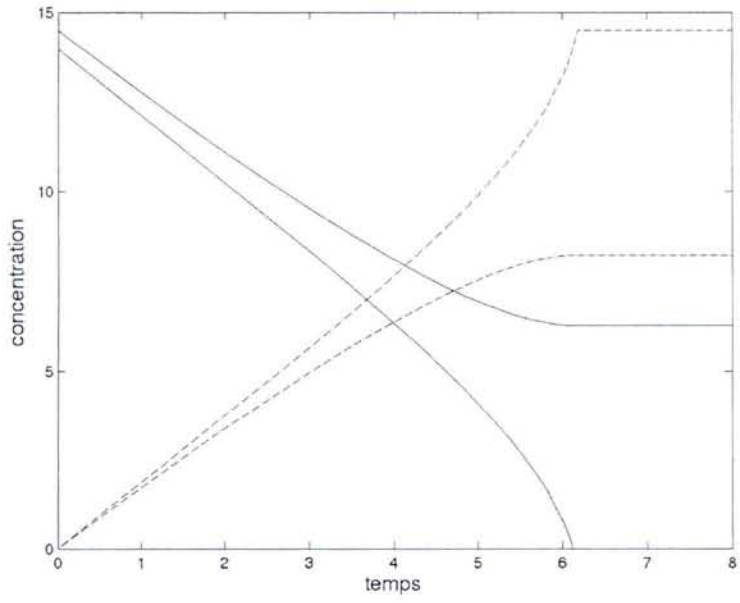

FIG. 4.6 – **Domaine d'atteignabilité d'une espèce moléculaire à cinétique michaelienne**. On représente la concentration d'une molécule qui décroît de manière linéaire. Les lignes pleines représentent le domaine qui borne la concentration de l'élément x. Les lignes pointillées représentent le domaine dans lequel se situe la concentration de y, transformé de x par une cinétique michaelienne. L'incertitude concernant la concentration initiale se propage au cours de la simulation.

```
    s' = 0;
}
sample(x, y);
```

4.4.4 Utilisation de conditions et de changements discrets

De manière générale la dynamique des systèmes biologiques dépend de conditions. Dans `Hybrid cc`, il est possible d'utiliser le combinateur `if` c `then` A qui exprime que si c est vrai, alors la contrainte A sera appliquée. Cette nouvelle contrainte permet d'effectuer des changements discrets par un *switch* d'un état dynamique à un autre. Le programme suivant illustre ces propos en modélisant la transformation de X en Y qui sera effective pour une concentration de protéine p qui peut atteindre un seuil. Le résultat de ce programme est représenté dans la Figure 4.7 en haut.

```
interval x, y, p;
x=[14,14.5];
y=0;
p=0.75;
```

```
always {
   p' = 1.5;
   if (p >= 3)
      { x' = -(Amax*x)/(ks+x);
        y' =  (Amax*x)/(ks+x);
   }
   if (p < 3)
      { x' = 0;
        y' = 0;
   }
}
sample(p, x ,y);
```

Nous considérons donc ici qu'il n'y a pas d'incertitude sur la valeur de p. Sans cette hypothèse, le solveur de contraintes ne peut décider pour des conditions $p \geq 3$ et $p < 3$. Pour combler ce manque, nous proposons l'utilisation du raisonnement par défaut, chose que nous allons décrire par la suite.

4.4.5 Comportement par défaut

Le combinateur par défaut `if` c `else` A (ou `unless`(c) A) exprime le fait que pour l'agent A sera satisfait si c ne l'est pas. D'un point de vue opérationnelle, cela signifie que pour c faux ou c invérifiable, A sera satisfait. On remarquera que ce combinateur par défaut `unless`(c) A n'est pas équivalent à `if` $\neg c$ `then` A. Dans ce dernier cas, A sera exécuté seulement si on peut vérifier que c n'est pas satisfaite. On ne prend ainsi pas en compte l'incertitude du comportement par défaut. Si A est exécuté cela peut être pour deux raisons :

- Le pot de contraintes entraîne $\neg c$ (dans ce cas `unless`(c) A se comporte comme si `if` $\neg c$ `then` A), ou alors
- Le pot de contraintes n'entraîne pas c ni $\neg c$, ce qui correspond à dire que si rien n'est connu, dans ce cas A est exécuté par défaut.

L'agent A n'est pas exécuté si le pot entraîne c. Nous proposons d'illustrer le comportement par défaut avec le même exemple que précédemment. La seule différence réside dans la variable p qui représente la concentration protéique qui sera initialisée cette fois ci sur l'intervalle $[0, 1]$. Comme le montre la Figure 4.8, la modification de réaction sera effectuée quand la borne inférieure de l'intervalle de p atteindra le seuil qui conditionne le changement de cinétique.

```
interval x, y, p;
x=[14,14.5];
y=0;
p=[0,1];
always {
   p' = 1.5;
   if ( p >= 3 )
      { x' = -(Amax*x)/(ks+x);
```

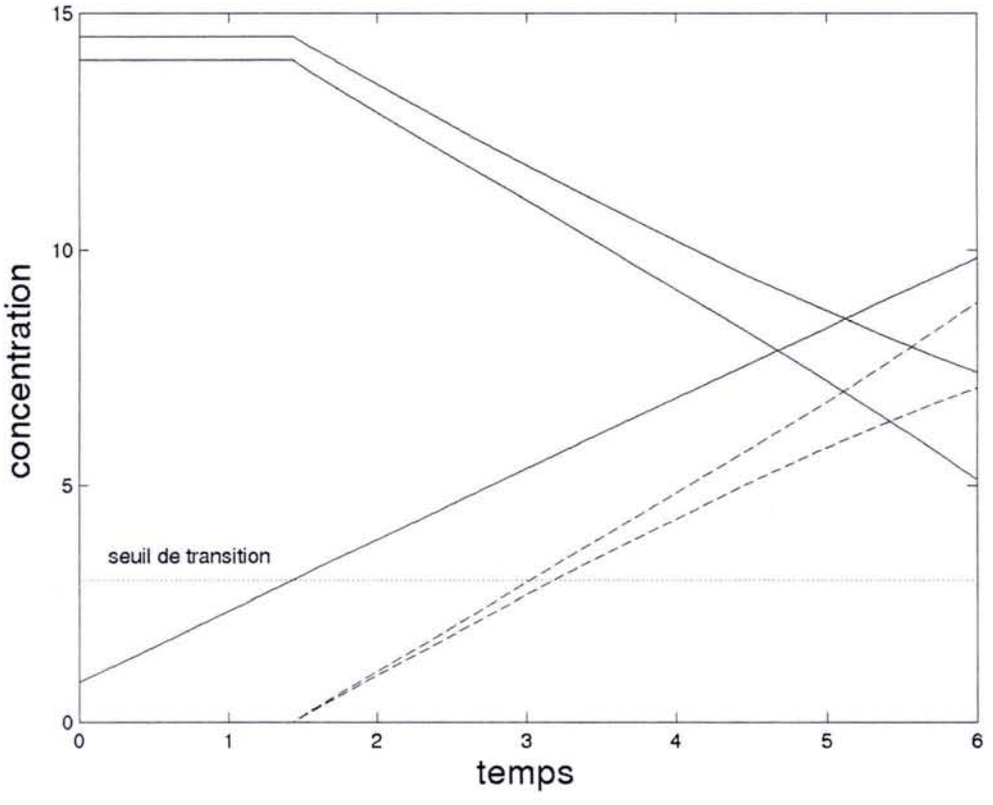

FIG. 4.7 – **Représentation d'un *switch* entre deux comportements par une condition**. Le graphique représente le comportement pour une modélisation avec des conditions sur les contraintes. Le passage d'un comportement à l'autre n'est valable que s'il est prouvé que la concentration en protéines atteint un seuil. Le graphique du bas représente le comportement pour modélisation avec le combinateur par défaut, *défaut*. Le changement de comportement est cette fois-ci effectif si, soit il est prouvé que la concentration protéique est supérieure à un seuil ou s'il n'est pas possible de prouver quelque chose. Dans ce dernier cas, c'est le comportement par défaut qui sera pris en compte.

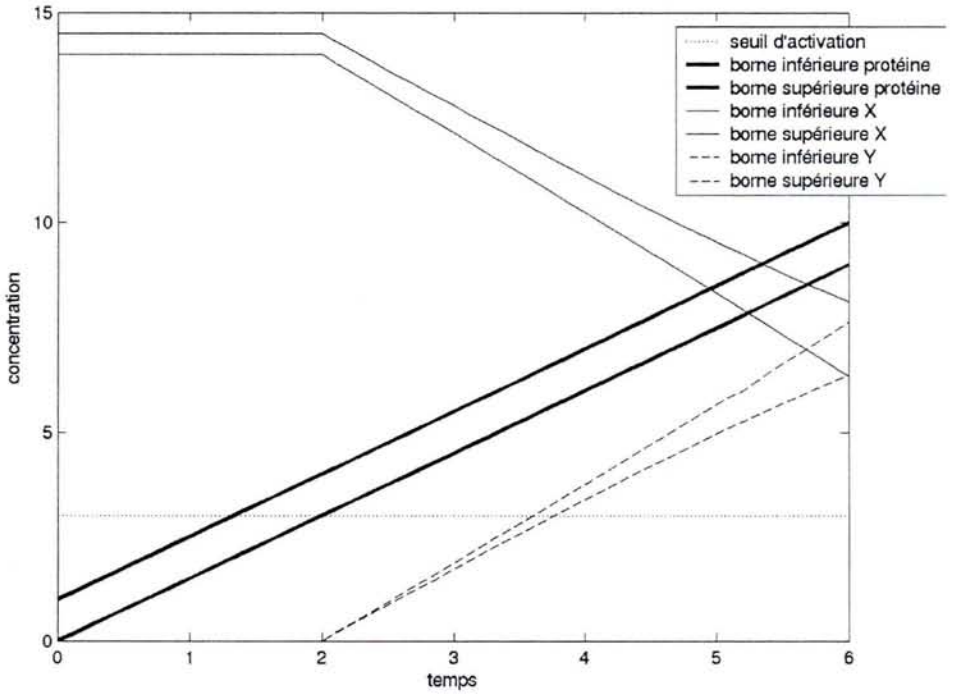

FIG. 4.8 – **Représentation d'un *switch* entre deux comportements avec un comportement par défaut.** Le graphique représente le comportement pour modélisation avec le combinateur par défaut, *default*. Le changement de comportement est effectif si soit il est prouvé que la concentration protéique est supérieure à un seuil. Si cela n'est pas prouvable ou s'il est prouvable que la concentration est inférieure au seuil, aucune modification de cinétique n'est effectuée. Dans ce dernier cas, c'est le comportement par défaut qui sera pris en compte.

```
        y' =  (Amax*x)/(ks+x);
    }
    unless (p >= 3)
       { x' = 0;
         y' = 0;
    }
}
sample(p, x ,y);
```

Le comportement par défaut est ainsi très appréciable pour modéliser les processus dont la connaissance est incomplète. Il est en effet possible de combler les lacunes de connaissance par des comportements par défaut. Nous utiliserons en particulier cette possibilité de modélisation dans la modélisation multi-site qui sera décrite dans le Chapitre 11.

Chapitre 5

Apprentissage statistique

La modélisation formelle que nous avons abordée dans le chapitre précédent permet de formaliser l'agencement des hypothèses concernant le fonctionnement des systèmes biologiques. Les hypothèses peuvent être issues directement de l'observation du vivant ou extraites des expériences qui sont une représentation du système sous des conditions données. Il est possible de proposer un autre type de modélisation de la connaissance qui est sous-jacente aux données biologiques, par une **modélisation statistique**. Nous proposons dans ce chapitre d'exposer ce type de modélisation différente de celle mentionnée précédemment et que nous serons amenés à utiliser dans le cadre de l'étude de la régulation de l'épissage alternatif.

Historiquement les analyses statistiques sont les premiers modèles statistiques qui ont permis une analyse discriminante des résultats biologiques de manière à mieux les comprendre (voir pour l'historique de l'utilisation de la statistique en biologie [Legendre & Legendre, 2000]). Les statisticiens utilisaient alors des méthodes d'inférence pour extrapoler les connaissances biologiques déjà existantes sur des données inconnues. Cette étape qui permet d'extraire la connaissance des hypothèses est aussi appelée **paramétrisation** dans certains domaines biologiques comme l'écologie ou la physiologie. Parallèlement à cette première association entre la biologie et la statistique, les bases théoriques ont énormément évolué.

Il existe différents domaines d'apprentissage statistique qui trouvent leurs applications en biologie. Dans ce chapitre, nous considérerons de manière privilégiée l'analyse discriminante, dont relève plus particulièrement les problèmes biologiques qui nous intérressent.

De nombreux problèmes de bioinformatique peuvent être formalisés en des termes de reconnaissance des formes. Ceci est spécialement vrai dans le cas de l'analyse de séquences biologiques telles que les séquences d'acides nucléiques et les séquences d'acides aminés qui composent les protéines. De nombreux bioinformaticiens considèrent alors le problème d'identification associée à la fonction d'une séquence comme un problème de discrimination entre des séquences possédant de l'information biologique et d'autres qui n'en possèdent pas.

Face à ces questions biologiques, de nombreux modèles statistiques ont été adaptés pour répondre aux besoins de la génomique, comme les réseaux de neurones ou les modèles de chaînes de Markov cachées (*HMM*). On peut citer pour revue le livre de [Durbin *et al.*, 1998] qui expose une liste non exhaustive de modèles discriminants pour

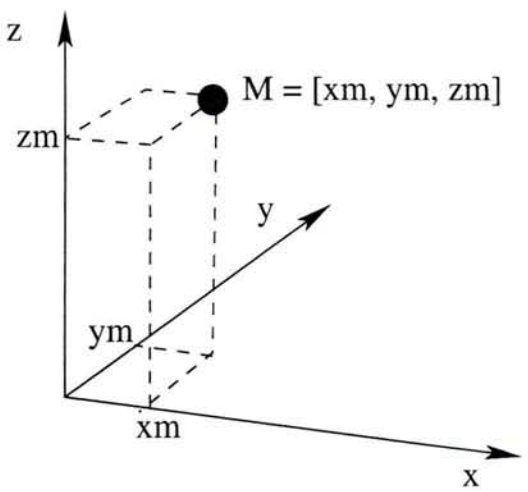

FIG. 5.1 – **Représentation d'un vecteur dans $\mathbb{R}^3$.** Un point M est représenté dans l'espace par 3 coordonnées (x_m, y_m, z_m). L'ensemble de ces trois coordonnées constitue un vecteur $[x_m, y_m, z_m]$ qui caractérise M. Les séquences biologiques peuvent être considérées comme des vecteurs dans un espace de dimension n. Toute l'information biologique associée à la séquence sera contenue dans ce seul vecteur. En appliquant les modèles statistiques de discrimination sur les vecteurs qui représentent les séquences biologiques, on est en mesure de discriminer les séquences.

l'analyse de séquences biologiques. Il est important de mentionner que l'optimisation de ces modèles pour la biologie fédère une communauté de théoriciens de la statistique particulièrement active.

Parmi ces méthodes de discrimination, il existe des machines qui appartiennent au domaine de l'apprentissage statistique qui semblent particulièrement appropriées à notre problème biologique. Ces méthodes se fondent sur des principes fondamentaux qu'il est essentiel d'aborder avant d'adapter l'apprentissage statistique à notre problème biologique. Ainsi, après une description du problème général de l'apprentissage statistique dans la Partie 5.1, nous étudierons les modèles statistiques à base de noyaux dans la Partie 5.2 ainsi que leurs extensions dans le domaine de la discrimination multi-classe qui est particulièrement utile en biologie, dans la Partie 5.3.

5.1 Philosophie de l'apprentissage statistique

En biologie, les principaux problèmes d'apprentissage relèvent de la discrimination entre les objets biologiques qui possèdent l'information voulue et les autres qui ne la possèdent pas. Un des avantages de la discrimination statistique est de permettre son application sur des vecteurs (voir pour illustration la Figure 5.1). Un vecteur peut être plus descriptif qu'une séquence biologique en permettant de décrire des informations autre que de nature séquentielle, comme les conditions expérimentales associées à une séquence ou un critère de structure de la séquence.

L'apprentissage statistique est un domaine de la statistique qui permet notamment

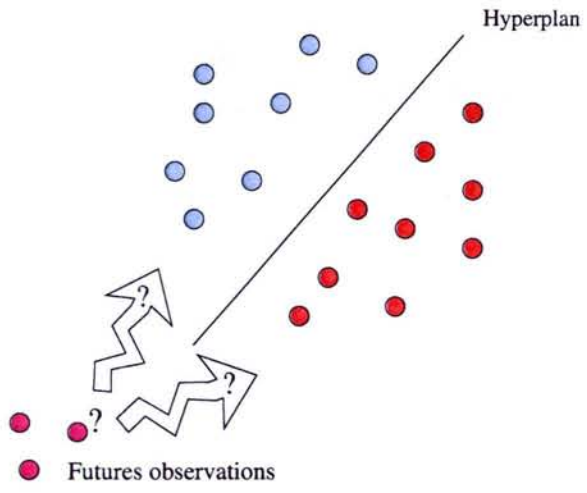

FIG. 5.2 – **Représentation de la discrimination linéaire.** Les données biologiques caractérisées (étiquetées) sont représentées dans $\mathbb{R}^2$. Les nouvelles données non étiquetées doivent être classées. Le choix de la catégorie dans laquel les placer s'effectue par la définition d'un hyperplan (dans notre exemple une droite) qui discrimine les données. Une vision simpliste de l'apprentissage revient à déterminer l'hyperplan qui discrimine au mieux les données de la base d'apprentissage dans leurs catégories correspondantes.

de discriminer les données biologiques. Cette opération permet de *"rassembler"* dans un même ensemble les vecteurs ou séquences biologiques selon une caractéristique commune qui correspondra à une homologie que nous allons mentionner par la suite. Par exemple, nous rassemblerons toutes les séquences nucléotidiques qui fixent une même protéine. L'apprentissage de la machine va alors consister à *"extraire"* les caractéristiques statistiques des séquences qui fixent la protéine et qui font ainsi partie du même ensemble. Après cet apprentissage sur des exemples dont le contenu informationnel est clairement identifié, il est possible de généraliser l'approche sur des exemples dont l'étiquette est cette fois-ci inconnue. De manière concrète, il sera alors possible de prédire si une séquence nucléotidique inconnue se fixe à une protéine donnée.

Ce problème de discrimination peut être illustré de manière informelle par l'exemple de points répartis dans un espace à deux dimensions (voir Figure 5.2). Certaines données possèdent la caractéristique rouge et d'autres la bleue. L'hypothèse sous-jacente de la discrimination est qu'il existe un phénomène artificiel ou naturel correspondant à une loi jointe qui discrimine les points de la sorte. Dans le but de comprendre la discrimination, il est intéressant de comprendre le mécanisme sous-jacent afin de prédire les étiquettes de nouveaux points. C'est le principal objectif des modèles statistiques que nous présentons.

Une **base d'apprentissage** est construite à partir de l'ensemble des exemples biologiques caractérisés biologiquement auxquels on associe une étiquette correspondant à leur caractéristique biologique. Les exemples de la base déterminent les dépendances entre la localisation géométrique d'un exemple et son étiquette. L'apprentissage statistique met alors en relation la position spatiale d'un échantillon avec l'étiquette ou la classe qui lui est associée. En pratique, les étiquettes correspondent à une fonction biologique que l'on associe à un exemple de la base. L'apprentissage va donc localiser un espace associé à la

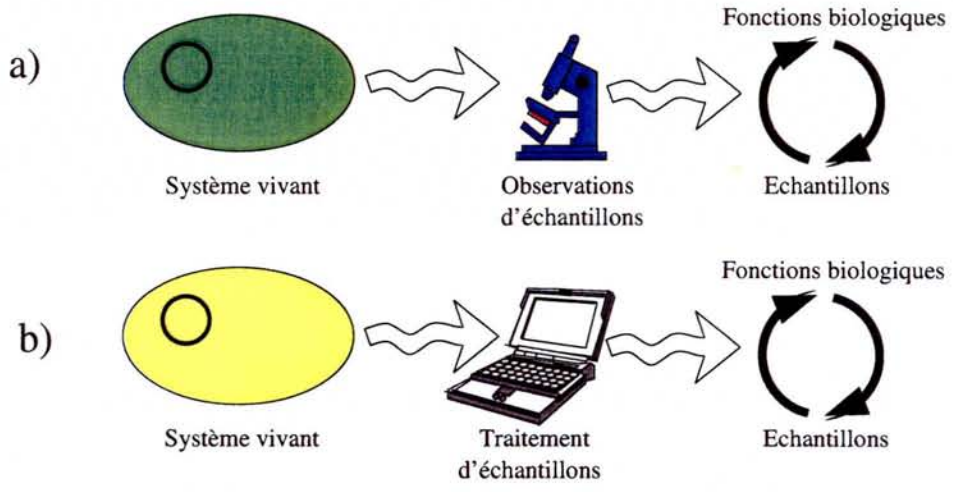

FIG. 5.3 – **Combinaison de protocoles expérimental et statistique pour inférer des fonctions biologiques. a)** représente le protocole expérimental d'extraction de la fonction biologique d'un système vivant. A partir d'un système vivant, un protocole expérimental permet d'obtenir des échantillons qui possèdent une fonction biologique déterminée par une analyse. **b)** est le protocole statistique. Après a) il est possible d'associer statistiquement une fonction biologique à des échantillons qui ont été analysés. Le traitement statistique va permettre de généraliser cette information concernant la fonction sur des échantillons qui ne sont identifiés. L'apprentissage est dans ce cas une forme d'automatisation de l'inférence biologique.

fonction biologique selon un critère statistique. Par ailleurs, on sélectionne une fonction f qui à chaque élément du plan $\mathbb{R}^2$ associe une couleur rouge ou bleue. Cette fonction f doit être déterminée en fonction des différents exemples qui constituent la base d'apprentissage.

Cette détermination de l'espace ainsi que la fonction correspondant aux exemples de l'apprentissage sont les fondements de l'apprentissage statistique. L'ambition de cette approche statistique est d'apporter une démarche efficace à la prédiction d'étiquettes de nouvelles données permettant une compréhension fine du phénomène sous-jacent à la discrimination. D'un point de vue biologique, l'apprentissage permet comme nous allons le voir dans le Chapitre 9, de généraliser les connaissances acquises notamment par les expériences. Ce nouvel outil permet ainsi d'estimer les fonctions biologiques sur de nouveaux éléments qui sont inconnus d'un point de vue expérimental. Cette approche statistique permet alors d'envisager un nouveau protocole qui combine l'approche expérimentale avec l'approche *in silico* (voir Figure 5.3 pour illustration).

Néanmoins, si cette démarche ne permet pas de comprendre l'origine de la fonction, elle permet d'automatiser une inférence de la fonction biologique jusqu'alors réservée à la seule démarche expérimentale. Ce dernier point est d'autant plus avantageux sur les données biologiques de l'ère post-génomique. Elles sont en effet de plus en plus abondantes et de plus en plus difficiles à analyser *à la main*. Une démarche expérimentale préalable reste primordiale puisqu'elle conditionne un apprentissage efficace. C'est dans ce contexte que cette démarche automatique apparaît comme la suite naturelle des expériences biologiques.

5.1.1 Fondements théoriques de l'apprentissage statistique

L'approche la plus didactique pour appréhender les fondements théoriques de l'apprentissage statistique est d'illustrer nos propos par la formalisation originelle (pour revue [Vapnik, 1998]).

Pour une **base d'apprentissage** $s_m = \{(\mathbf{x_i}, c_i)\}_{i=1...m}$ de données étiquetées (c_i est l'*étiquette* de l'exemple $\mathbf{x_i}$), on suppose qu'il existe une distribution de probabilité fixe mais inconnue $P(\mathbf{x}, c)$, qui représente les données $(\mathbf{x_i}, c_i)$. P caractérise alors le phénomène sous-jacent et associe une fonction biologique à l'exemple. Le modèle statistique correspond à rechercher dans un ensemble de fonctions $\mathcal{H}$, une fonction $f_0 : \mathbf{x} \mapsto c$ qui assigne une étiquette c à chaque nouvel exemple $\mathbf{x}$ et ceci en effectuant le moins d'erreurs possibles.

La minimisation des erreurs revient à ce que pour un couple $(\mathbf{x}, c)$, on cherche à minimiser la probabilité $Prob_{(\mathbf{x},c)}\{f_0(\mathbf{x}) \neq c\}$.

La nature des échantillons ($\mathbf{x}$ et c) peut permettre certaines hypothèses qui simplifient l'apprentissage. Dès à présent, on considère que l'on raisonne dans un espace Euclidien E, ce qui permet certaines simplifications.

On peut en effet considérer qu'il existe un jeu fini $\mathcal{C} = \{1, \ldots, Q\}$ de Q catégories. Suivant cette dernière notation, on caractérise différemment la précédente probabilité d'erreur. Le taux d'erreur de la fonction $f : E \to \mathcal{C}$ est aussi appellé **erreur en généralisation** de f ou **risque attendu** de f. Elle peut s'écrire de la façon suivante :

$$R(f) = Prob_{(\mathbf{x},c)}\{f(\mathbf{x}) \neq c\} = \int_{(\mathbf{x},c)\in E\times\mathcal{C}} \mathbf{I}_{\{f(\mathbf{x})\neq c\}} dP(\mathbf{x}, c) \tag{5.1}$$

où $\mathbf{I}$ est la fonction indicatrice, i.e. $\mathbf{I}_{\{f(\mathbf{x})\neq c\}} = 1$ pour $f(\mathbf{x}) \neq c$, et 0 dans les autres cas. Afin de minimiser le risque d'erreur, il faut donc sélectionner dans une famille de fonctions $\mathcal{H}$ donnée, une fonction f_0 qui minimise $R(f)$. Cette sélection de modèle est appelée l'**apprentissage (supervisé)**, ou communément considéré comme la phase d'entraînement de la machine. Cette procédure n'est pas toujours facile à mettre en œuvre et dépend des échantillons qui sont à disposition.

5.1.2 Mise en pratique de l'apprentissage statistique

L'apprentissage consiste donc à tenter minimiser une fonctionnelle R qui va permettre de généraliser correctement l'information présente uniquement dans la base d'apprentissage (qui est la seule source de connaissances biologiques). L'intérêt de l'apprentissage sur les données biologiques va être d'utiliser de manière optimale les échantillons à disposition pour apprendre et généraliser l'information sous-jacente. Or, l'information sous-jacente est statistiquement inconnue (la loi jointe P qui associe la catégorie c à l'exemple $\mathbf{x}$). Il faut alors utiliser la meilleure stratégie possible pour minimiser le risque R de se tromper. Une approche naturelle est de choisir une fonction f_0 qui va minimiser le **risque empirique** qui s'exprime sous la forme :

$$R_{emp}(f) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{I}_{\{f(\mathbf{x_i})\neq c_i\}} \tag{5.2}$$

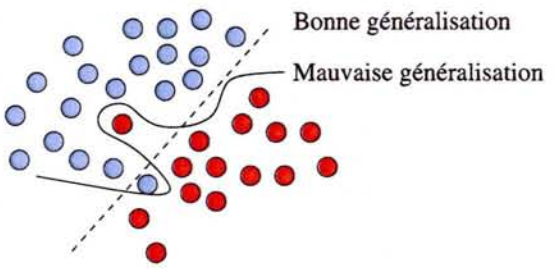

FIG. 5.4 – **Illustration d'un effet de sur-apprentissage.** La fonction f en trait continu ne fait aucune erreur d'apprentissage mais ne généralise pas bien par rapport à la fonction f en trait pointillé qui fait deux erreurs mais qui permet de mieux généraliser.

Le risque empirique $R_{emp}(f)$ représente une estimation du risque théorique R ou erreur de généralisation qui ne peut être calculé directement puisque l'on ne connaît pas la loi jointe P qui associe c à $\mathbf{x}$.

De manière naïve, on pourrait penser que, pour que la fonction f_0 effectue le moins d'erreurs possibles, elle doit faire le moins d'erreurs sur la base d'apprentissage. Cependant, un problème de **sur-apprentissage** peut se présenter. Il est en effet possible de déterminer une fonction f qui ne fait aucune erreur sur la base d'apprentissage et qui par conséquent généralise très mal les échantillons (voir Figure 5.4). Dans ce cas, la fonction f est trop spécifique par rapport aux exemples. Une analogie peut être faite avec un écolier qui apprendrait par cœur ses leçons sans généraliser son savoir. L'apprentissage scolaire est réussi mais l'élève peut s'avérer incapable d'appliquer les connaissances acquises sur un exemple qu'il n'a jamais abordé auparavant. Le problème du sur-apprentissage est particulièrement présent dans l'application de ces modèles statistiques en biologie où les expériences sont soumises à de fortes fluctuations ce qui empêche une discrimination franche des données. Dans ce cas, le modèle apprend parfaitement sur chaque exemple et se retrouve dans l'incapacité de généraliser. C'est le sur-apprentissage. Une généralisation correcte sans sur-apprentissage étant primordiale pour une bonne prédiction de nouveaux échantillons, différentes approches théoriques ont été développées pour augmenter les performances des machines d'apprentissage.

La théorie d'apprentissage a été développée par [Vapnik, 1982], qui propose une nouvelle machine d'apprentissage : **les machines à vecteurs support (SVM)**[Boser *et al.*, 1992, Cortes & Vapnik, 1995] qui proposent une alternative pour gérer le problème de sur-apprentissage. D'un point de vue chronologique, les SVM bi-classes ont été mises en place avant les SVM multi-classes. Dans un but pédagogique, nous suivrons cet ordre pour décrire cette approche statistique que nous appliquerons ultérieurement aux données biologiques dans la Partie III.

5.2 Les SVM bi-classes

Les SVM bi-classes sont des machines qui permettent la discrimination entre deux classes qui est appelée calcul de dichotomie. Nous nous focaliserons donc sur des données étiquetées $(\mathbf{x}, c)$ dans $E \times \{-1, 1\}$. La notation standard des étiquettes pour la discrimi-

nation est souvent -1 et 1 plutôt que 0 et 1 ou 1 et 2, ceci pour des raisons purement techniques. Comme nous l'avons mentionné précédemment, l'algorithme d'apprentissage supervisé est caractérisé par deux éléments prépondérants : la famille $\mathcal{H}$ et la fonction objectif f à optimiser.

5.2.1 La famille de fonctions

Il existe différentes familles de fonctions qui peuvent être associées à l'algorithme des SVM. Dans cette partie, nous allons illustrer nos propos avec la famille $\mathcal{H}$. Elle est composée d'un ensemble de séparateurs linéaires, comme dans le cas d'un perceptron. $\mathcal{H}$ est donc un ensemble d'hyperplans formalisé de la façon suivante :

$$\mathcal{H} = \{h(w,b) : \mathbf{x} \mapsto signe(\mathbf{w}.\mathbf{x} + b) \ / \ \mathbf{w} \in E,\ b \in \mathbb{R}\} \tag{5.3}$$

Dans ce cas, $signe(a) = 1$ si a est positif ou nul, et $signe(a) = -1$ dans le cas contraire. L'algorithme de la SVM revient alors à calculer un hyperplan $(\mathbf{w}, b)$. On dit que le séparateur est linéaire car $\mathbf{w}.\mathbf{x} + b$ caractérise un hyperplan (une droite dans $\mathbb{R}^2$).

La partie qui contient les exemples $\mathbf{x}$ tels que $\mathbf{w}.\mathbf{x}+b \geq 0$ sont associés à l'étiquette 1, l'autre partie qui contient les exemples $\mathbf{x}$ tels que $\mathbf{w}.\mathbf{x}+b < 0$ seront associés à l'étiquette -1. On recherche alors de manière itérative un plan optimal qui va au mieux séparer les échantillons de la base d'apprentissage. C'est l'hyperplan optimal.

5.2.2 Le choix de l'hyperplan optimal

La combinaison des deux propositions suivantes justifie le choix d'un algorithme de la SVM.

Proposition 1 (Borne sur le risque) *Pour une base d'apprentissage définie par* $s_m = \{(\mathbf{x}_i, c_i)\}_{i=1...m}$ *et une famille de fonctions* $\mathcal{H}$*, pour chaque* η *tel que* $0 \leq \eta \leq 1$*, et pour une fonction* f *de* $\mathcal{H}$*, la borne s'exprime de la façon suivante avec la probabilité* $1 - \eta$ *:*

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h(\log(\frac{2m}{h}) + 1) - \log(\frac{\eta}{4})}{m}} \tag{5.4}$$

En d'autres termes, [Cortes & Vapnik, 1995] expriment mathématiquement que le risque possède une borne supérieure. Ils considèrent cette expression du risque comme le **risque garanti** qui est la partie droite de l'inégalité. Ainsi il est "garanti" que le risque ne sera jamais supérieur à l'expression mathématique 5.4 (avec une probabilité $1 - \eta$). On retrouve dans cette expression que le risque sera toujours inférieur à la somme du risque empirique $R_{emp}(f)$ et de l'intervalle de confiance qui dépend d'une probabilité η, du nombre d'échantillons m et de h.

h est un entier positif ou nul appelé la *dimension de Vapnik-Chervonenkis (VC)*. Il constitue une mesure de la capacité de la famille de fonctions $\mathcal{H}$. Cela correspond à la taille maximale d'un ensemble que l'on peut *pulvériser*. La pulvérisation d'un ensemble étant la capacité à calculer toutes les fonctions possibles sur cet ensemble. Mais paradoxalement, la dimension VC ne doit pas être trop grande. En effet, si les points sont trop bien

pulvérisés, il existera un nombre très important de fonctions f qui pourront discriminer les échantillons pour toute dichotomie sur un ensemble de points. En pratique, la dimension VC doit être assez grande pour proposer suffisamment de fonctions pour pulvériser les points, mais pas trop pour permettre une bonne généralisation de l'apprentissage pour un nouvel échantillon. Il faut donc rechercher un compromis. La proposition statistique au travers les méthodes SVM est d'utiliser le principe de *minimisation structurelle du risque (SRM)*[Vapnik, 1982].

Le principe se base sur la structuration de la famille de fonctions. En effet, les fonctions de $\mathcal{H}$ sont structurées en sous-ensembles de fonctions imbriqués (voir Figure 5.5 pour illustration). Elles sont rangées en fonction de la dimension VC $h_1 < h_2 < \ldots$. Dans chaque sous-ensemble de cette structure, une fonction qui minimise le risque empirique est recherchée. Ainsi, en pratique, parmi toutes les fonctions, le processus d'apprentissage recherche une fonction qui minimise le risque garanti. Il est donc pertinent d'avoir une "faible" dimension VC appropriée au problème de discrimination. L'utilisation du principe de SRM est particulièrement avantageuse dans le cas ou l'échantillon de la base d'apprentissage est petit. L'intervalle de confiance n'est alors plus négligeable devant le risque empirique (voir 1).

La définition formelle de la dimension VC peut être les suivantes :

Définition 1 (dimension VC) *Soit $\mathcal{H}$ une famille de fonctions de l'espace E dans $\{-1, 1\}$. Soit s_l un jeu de l exemples de E. Il existe alors 2^l étiquetages possibles pour ces exemples. S'il existe, pour chaque étiquetage possible Y, une fonction f_Y dans $\mathcal{H}$ qui associe correctement s_l à ces étiquettes, s_l est considéré comme pulvérisé par $\mathcal{H}$. La dimension VC de $\mathcal{H}$ est le nombre maximum de points qui peuvent être pulvérisés par $\mathcal{H}$. Le maximum n'existe pas si cette dimension est $+\infty$.*

Proposition 2 ([Vapnik, 1982]) *Soit $\mathcal{H}_\Lambda$ un ensemble de séparateurs linéaires $h(\mathbf{w}, b)$ tel que $\|w\| \leq \Lambda$ et R le rayon de la plus petite sphère centrée sur l'origine contenant les échantillons. Alors, la dimension VC h_Λ de $\mathcal{H}_\Lambda$ est majorée par :*

$$h_\Lambda \leq \min(\lfloor \Lambda^2 R^2 \rfloor, dim(E)) + 1 \tag{5.5}$$

Pour minimiser le risque $R(f)$, il est alors possible d'utiliser la Proposition 1. Il faut pour cela utiliser l'inégalité 5.4 qui formalise le risque garanti. L'intervalle de confiance est alors une fonction croissante de h_Λ qui minimise le risque garanti pour trouver un compromis entre le risque empirique de f et la dimension VC de la famille $\mathcal{H}$ dans laquelle la fonction f a été sélectionnée.

Nous allons à présent illustrer la mise en œuvre des principes théoriques précédents sur un cas linéairement séparable. Suivant la Proposition 2, on définit alors la famille $\mathcal{H}$ qui englobe toutes les fonctions qui sont des séparateurs linéaires. Ils peuvent être structurés de la façon suivante :

$$\begin{aligned} \mathcal{H} &= \mathcal{H}_{\Lambda_1} \cup \mathcal{H}_{\Lambda_2} \cup \ldots \\ &\mathcal{H}_{\Lambda_1} \subset \mathcal{H}_{\Lambda_2} \subset \ldots \end{aligned} \tag{5.6}$$

avec $\Lambda_1 < \Lambda_2 < \ldots$

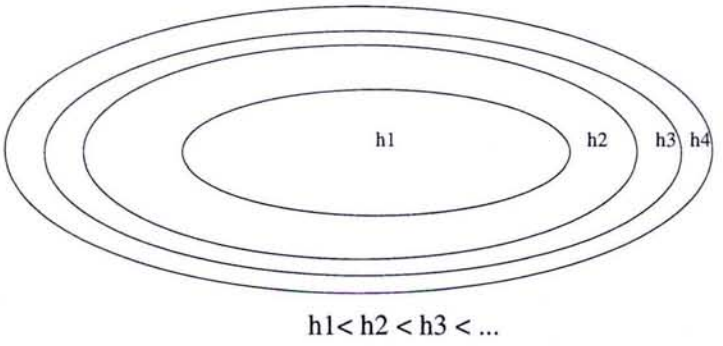

FIG. 5.5 – **Sous-ensembles imbriqués de fonctions ordonnées en fonction de leurs dimensions VC croissante.** Une fonction f optimale pour l'apprentissage sera recherchée dans chaque sous-ensemble. Le protocole de recherche suivra le gradient des dimensions VC.

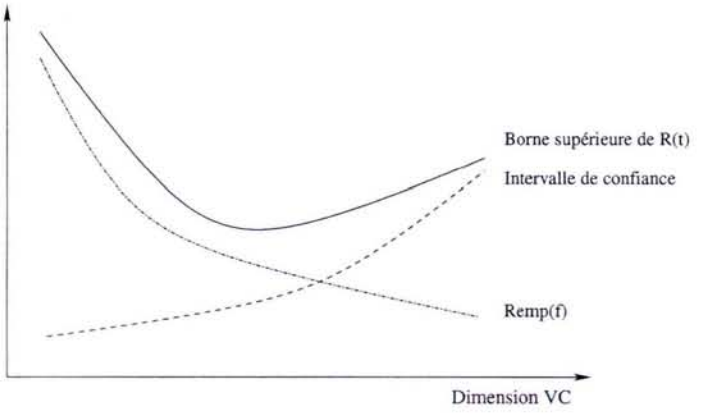

FIG. 5.6 – **Principe de minimisation structurelle du risque.** Le risque empirique et l'intervalle de confiance varient de manière opposée en fonction de la dimension VC.

Ces différents aspects théoriques permettent de représenter ce problème d'apprentissage sous la forme d'un problème d'optimisation. L'optimisation qui nous intéresse consiste à minimiser une fonction quadratique sur un domaine convexe. On est face à un problème classique de programmation quadratique. L'apprentissage consiste à résoudre :

Problème 1

$$min_{\mathbf{w},\xi} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m} \xi_i \right\}$$

avec

$$\forall\, i = 1 \ldots m,\ c_i(\mathbf{w}.\mathbf{x_i} + b) \geq 1 - \xi_i \quad \textit{et les variables d'écart} \quad \xi_i \geq 0$$

C correspond à la constante de *marge douce* qui est une représentation de l'équilibre entre le risque empirique et l'intervalle de confiance. La fonction objectif du Problème 1 est directement reliée au risque garanti : $\frac{1}{2}\|\mathbf{w}\|^2$ est lié à l'intervalle de confiance et $C\sum_{i=1}^{m}\xi_i$ au risque empirique.

Pour des raisons pratiques, le problème de programmation quadratique est résolu sous sa forme duale donnée par :

Problème 2

$$max \left\{ \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j c_i c_j \mathbf{x}_i.\mathbf{x}_j \right\}$$

avec

$$\forall\, i = 1 \ldots m,\ \textit{les variables duales } \alpha_i \textit{ satisfesant} \quad 0 \leq \alpha_i \leq C$$

$$\sum_{i=1}^{m} \alpha_i c_i = 0$$

Cette dernière formulation représente l'expression analytique de la fonction discriminante. Ces fondements théoriques que nous avons partiellement mentionnés permettent d'aboutir à cet algorithme qui est utilisé par les SVM.. Une fois les valeurs optimales (α_i^*) des variables duales connues, les valeurs optimales des paramètres $\mathbf{w}$ et b sont obtenues par application des conditions optimales de Kuhn-Tucker :

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i^* c_i \mathbf{x}_i \tag{5.7}$$

et pour chaque i tel que $0 < \alpha_i < C$,

$$b = c_i - \mathbf{w}.\mathbf{x}_i = c_i - \sum_{j=1}^{m} \alpha_j^* c_j \mathbf{x}_j.\mathbf{x}_i \tag{5.8}$$

La fonction discriminante s'exprime de la manière suivante :

$$f(x) = sign(\mathbf{w}.\mathbf{x} + b) = sign(\sum_{i=1}^{m} \alpha_i^* c_i \mathbf{x}_i.\mathbf{x} + b) \tag{5.9}$$

Cette dernière formulation représente l'expression analytique de la fonction discriminante. Les fondements théoriques que nous avons partiellement mentionnés permettent d'obtenir un algorithme des SVM.

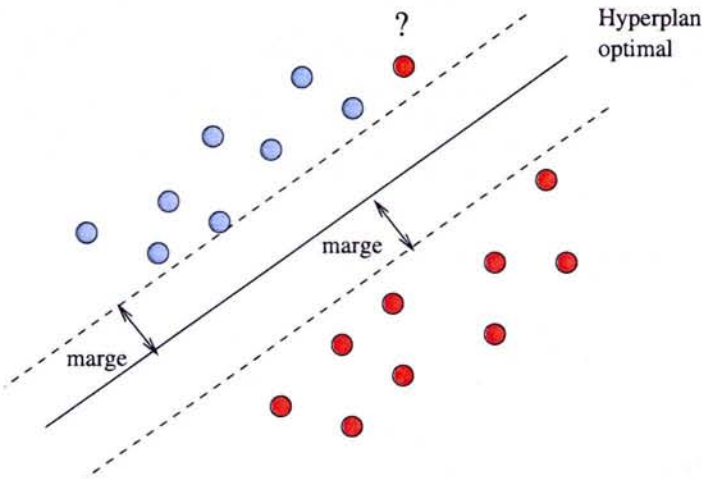

FIG. 5.7 – **Représentation de la marge d'après l'algorithme des SVM dans le cas linéairement séparable.** La marge représente la distance entre l'hyperplan et le point le plus proche de l'hyperplan.

5.2.3 Exemple d'apprentissage : interprétation géométrique de l'algorithme des SVM

Lorsque les échantillons sont linéairement séparables, l'utilisation de l'algorithme SVM avec $C = +\infty$ permet simplement de retrouver l'hyperplan de marge maximale (voir la Figure 5.7 pour illustration).

Définition 2 (marge) *Soit s_m un ensemble d'échantillons (couple ou ensemble de couples) linéairement séparables. Soit $H = (\mathbf{w}, b)$ un hyperplan séparateur. La marge de H est la distance entre l'hyperplan et le point de s_m le plus proche.*

La structure repose alors sur $\|\mathbf{w}\|$ puisque la marge de l'hyperplan calculé par la SVM est égale à $1/\|\mathbf{w}\|$. Ainsi, l'algorithme de la SVM est justifié *a posteriori* dans le cas d'échantillons séparables.

Les points $\mathbf{x}_i$ qui sont proches de l'hyperplan et qui satisfont l'égalité $\mathbf{w}.\mathbf{x}_i + b = c_i$ sont nommés vecteurs support. Dans ce cas linéaire, on peut donner une expression analytique de la marge (grâce à l'expression 5.7).

Dans le cas d'échantillons non linéairement séparables, les calculs ne peuvent pas se simplifier aussi bien. La marge ne peut donc pas être exprimée analytiquement comme précédemment. On considère alors une marge douce ou *soft margin* (ou marge molle dans la littérature du domaine) (voir [Cortes & Vapnik, 1995]). La valeur de C, constante de marge douce, définit alors le compromis entre les erreurs qui peuvent être tolérées et la valeur de l'intervalle de confiance.

5.2.4 L'approche par noyau

Tel que nous avons présenté la SVM, nous considérons la machine comme permettant une séparation linéaire des données d'apprentissage. Nous allons considérer à présent une

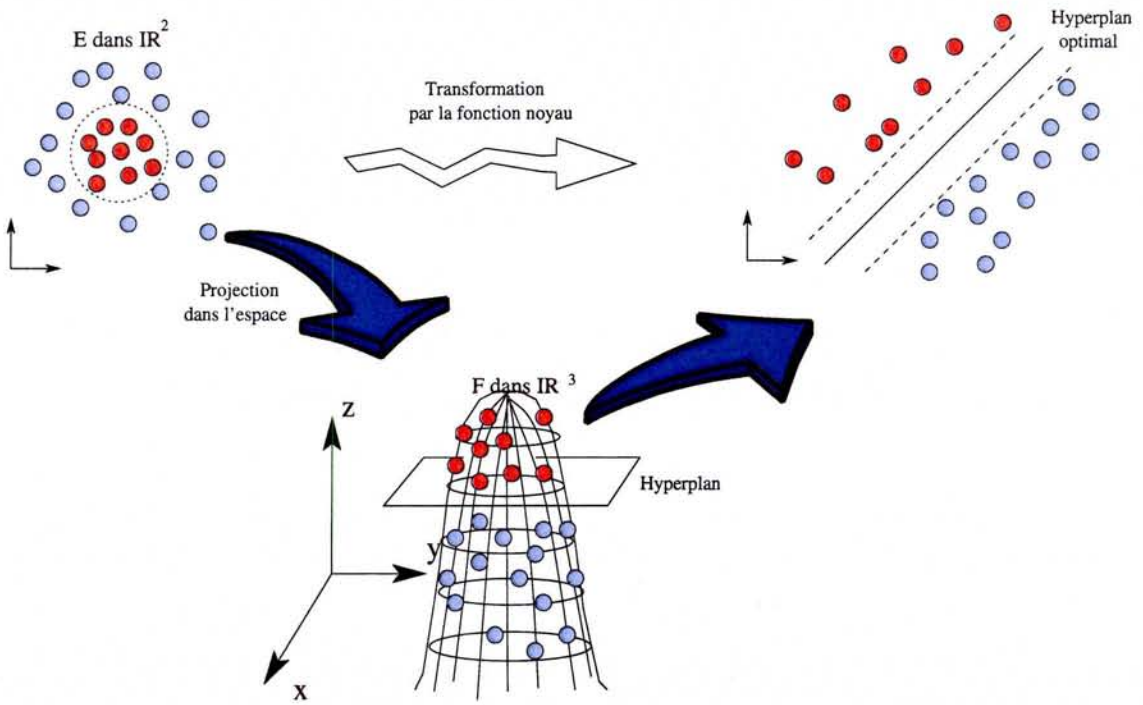

FIG. 5.8 – **Représentation du pré-traitement sur un jeu d'échantillons non séparables linéairement.** Les échantillons ne sont pas linéairement séparables. Le pré traitement consiste à projeter les données dans un nouvel espace de représentation grâce à la fonction (le noyau) qui *pré-traite* les données. Dans ce nouvel espace de représentation, le *feature space*, les données projetées sont linéairement séparables.

machine **non linéaire** (illustrée par la Figure 5.8). En effet, il est rare que les données soient linéairement séparables (et spécialement en biologie). Ce cas est d'autant plus pertinent lorsque l'analyse met en évidence un processus sous-jacent non linéaire (c'est notamment le cas lorsque des structures secondaires interviennent dans les processus biologiques ce qui modifie la fonction des séquences). Dans ce cas, on procède à un pré-traitement des données avant de les séparer avec l'hyperplan. Les données sont alors projetées d'un espace dans un autre (pour illustration de $\mathbb{R}^2$ dans $\mathbb{R}^3$ dans le cas de la Figure 5.8). Dans ce nouvel espace, les données deviennent linéairement séparables.

De manière générale, lorsque les échantillons de la base de données ne sont pas linéairement séparables, on applique aux données une transformation par la fonction ϕ qui projette les exemples dans un espace F de plus grande dimension que l'on appelle l'espace des représentations ou *feature space* dans lequel les exemples sont linéairement séparables.

En effet si les échantillons sont pré-traités par la fonction ϕ, les produits scalaires $\phi(\mathbf{x_i}).\phi(\mathbf{x_j})$ et $\phi(\mathbf{x_i}).\phi(\mathbf{x})$ seront utilisés par la SVM pour apprendre et prédire les étiquettes de $\mathbf{x}$. En raisonnant sur ce produit scalaire, les principes de discrimination s'appliquent dans l'espace de représentation. La plus grande difficulté des SVM consiste donc à déterminer la bonne transformation ϕ. Afin de la déterminer correctement, il convient d'utilisé un *noyau* défini comme suit :

Définition 3 (Définition du noyau) *Pour k une fonction de $E \times E$ dans $\mathbb{R}$, k est un noyau s'il existe une transformation ϕ de E dans un espace des représentations F tel que pour chaque $\mathbf{x}, \mathbf{x}'$ dans E, $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}).\phi(\mathbf{x}')$.*

Cette définition théorique d'un noyau est très utile pour adapter un modèle statistique à de nouvelles données telles que des données biologiques. Les noyaux adaptés aux séquences biologiques sont encore mal identifiés et ils constituent une part importante de la recherche fondamentale statistique. On est ainsi face à un problème de choix de noyau important dont les conditions de validation ont été introduites par [Aizerman *et al.*, 1964].

Proposition 3 *Soit k une fonction symétrique de $E \times E$ dans $\mathbb{R}$. k est un noyau valide* si et seulement si *pour chaque ensemble de nombre fini d'exemple $s_l = \{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$, la matrice de Gram (ou matrice du noyau) $(k(\mathbf{x}_i, \mathbf{x}_j))_{i,j}$ est semi-définie positive.*

Une fonction noyau, k est implicitement liée par une transformation ϕ tel que $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}).\phi(\mathbf{x}')$. L'intérêt des méthodes à noyau est de raisonner directement sur les produits scalaires afin de discriminer dans l'espace de représentation. C'est le *kernel trick*. Dans ce contexte de méthodes à noyau, on peut formuler le problème de programmation quadratique 2 de la façon suivante :

Problème 3

$$max \left\{ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \right\}$$

sous les contraintes

$$\forall\, i = 1 \ldots m,\ 0 \leq \alpha_i \leq C, \quad \alpha_i \in \mathbb{R}$$

$$\sum_{i=1}^{m} \alpha_i c_i = 0$$

Pour calculer la valeur optimale de $\mathbf{w}$ et le biais b, il suffit de remplacer les produits $\mathbf{x}.\mathbf{x}'$ par $k(\mathbf{x}, \mathbf{x}')$ dans les équations 5.7 et 5.8. La fonction pour prédire l'étiquette d'un nouvel échantillon prend alors la forme suivante :

$$f(x) = sign(\mathbf{w}.\phi(\mathbf{x}) + b) = sign(\sum_{i=1}^{m} \alpha_i c_i k(\mathbf{x}_i, \mathbf{x}) + b) \tag{5.10}$$

L'avantage d'utiliser un noyau pour un pré-traitement des données est de traiter les données qui correspondent au produit précédent. Un des avantages du *kernel trick*, mis à part la projection dans un espace de représentation adéquat, est de permettre la gestion de données de très grandes dimensions et ceci avec un faible coût de calcul.

On retrouve une expression similaire à l'équation 5.9 mais qui est cette fois-ci non linéaire avec l'ajout d'une fonction noyau qui correspond au pré-traitement. Il est alors possible d'expérimenter des pré-traitements relativement complexes. Pour illustration, $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\sigma^2)$ est un noyau valide suivant les termes exprimés précédemment. C'est le noyau Gaussien, qui est associé à un espace de représentation de dimension infinie.

5.3 Les SVM multi-classes

Les informations biologiques nécessitent souvent plusieurs classes pour représenter les fonctions associées à des exemples expérimentaux. Il est possible d'effectuer des discriminations deux à deux, mais la démarche n'est pas théoriquement satisfaisante. L'un des avantages des SVM réside dans la possibilité d'extension des SVM bi-classes aux cas multi-classe. On considère alors un nombre d'étiquettes supérieur à deux.

Pour un nombre de classes supérieur à 2, l'idée reste la même. La SVM multi-classe (M-SVM) est une extension proposée du cas bi-classe dont les variantes sont mentionnées par [Weston & Watkins, 1998] et [Vapnik, 1998]. Les fondements théoriques sont détaillés dans les travaux de [Elisseef *et al.*, 1999] et [Guermeur *et al.*, 2002].

5.3.1 La famille de fonctions $\mathcal{H}$

La M-SVM est un modèle affine multivarié :

$$\mathcal{H} = \{\mathbf{x} \mapsto argmax_{i\in\{1\ldots Q\}}(\mathbf{w}_i.\mathbf{x} + b_i)/\mathbf{w}_1, \ldots, \mathbf{w}_Q \in E,\ b_1, \ldots, b_Q \in \mathbb{R}\}$$

On remarquera que les Q classes sont codées de $1, \ldots, Q$.

5.3.2 Choisir un jeu optimal d'hyperplans

Le choix des paramètres $\mathbf{w}_i$ et b_i repose sur la mise en œuvre du principe SRM mentionné dans le cas bi-classe. On se base sur le même résultat de convergence uniforme du risque empirique (voir pour détails [Elisseef *et al.*, 1999]). Ainsi, pour une fonction f déterminée appartenant à une famille $\mathcal{H}$, l'expression est la suivante :

$$R(f) \leq R_{emp}(f) + \mathcal{G}(\mathcal{N}, m, \eta) \tag{5.11}$$

Cette borne est similaire dans le cas multi-classe à la borne 5.4. On retrouve en effet une représentation de la borne du risque $R(f)$. Cette borne correspond au risque garanti qui est la somme du risque empirique $R_{emp}(f)$ et de l'intervalle de confiance $\mathcal{G}(\mathcal{N}, m, \eta)$.

Le risque garanti est obtenu avec la probabilité $1 - \eta$, avec $\mathcal{N}$ le nombre de couverture de la famille $\mathcal{H}$ (voir pour détails [Bartlett, 1998, Guermeur *et al.*, 2002]) et $\mathcal{G}$ l'intervalle de confiance qui est une fonction croissante de $\mathcal{N}$. Dans le cas multi-classe, un nombre de couverture peut être borné supérieurement par une fonction croissante de $\sum_{i,j=1}^{Q} \|\mathbf{w}_i - \mathbf{w}_j\|^2$. En combinant ces deux résultats comme dans le cas bi-classe, l'algorithme de la M-SVM revient encore une fois à résoudre un problème de programmation quadratique :

Problème 4 (primal)

$$min_{\mathbf{w},\xi} \left\{ \frac{1}{2} \sum_{k<l} \|\mathbf{w}_k - \mathbf{w}_l\|^2 + C \sum_{i=1}^{m} \sum_{k=1, k\neq c_i}^{Q} \xi_{ik} \right\}$$

avec

$$\forall i = 1, \ldots, m, k = 1, \ldots, Q,\ k \neq c_i,\ (\mathbf{w}_{c_i} - \mathbf{w}_k).\mathbf{x_i} + b_{c_i} - b_k \geq 1 - \xi_{ik}$$
$$\forall i = 1, \ldots, m,\ k = 1 \ldots Q,\ k \neq c_i,\ \xi_{ik} \geq 0$$

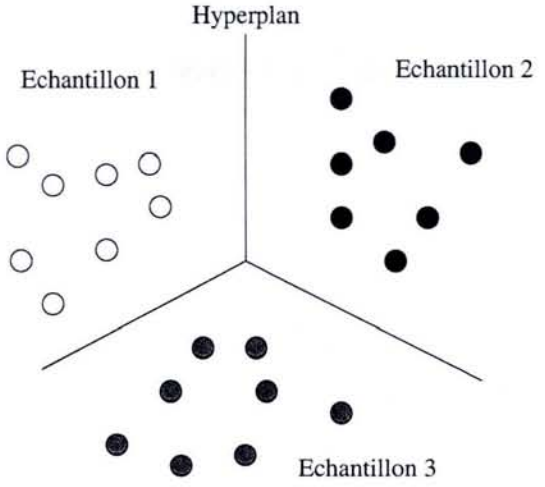

FIG. 5.9 – **Représentation de la marge dans le cas multi-classe.**

Pour les mêmes raisons que dans le cas bi-classe, le problème est résolu sous la forme de son dual de Wolfe. On obtient ainsi :

Problème 5 (dual)

$$max_\alpha \left\{ \mathbf{1}^T \alpha - \frac{1}{2} \alpha^T H \alpha \right\}$$

avec

$$\forall\, k = 1 \ldots (Q-1),\ \sum_{c_i=k} \sum_{l=1}^{Q} \alpha_{il} - \sum_{i=1}^{m} \alpha_{ik} = 0$$

$$\forall\, i = 1 \ldots m,\ k = 1 \ldots Q,\ k \neq c_i,\ 0 \leq \alpha_{ik} \leq C \quad \alpha_i c_i = 0 \quad \forall\, i \in \{1, \ldots, m\}$$

Dans ce cas, α est le vecteur $(\alpha_{1,1}, \alpha_{1,2} \ldots, \alpha_{m,Q})^T$, $\mathbf{1}$ est le vecteur de dimension mQ dont les composantes sont toutes 1 et les composantes du Hessien H de la fonction objective sont des multiples simples des composantes de la matrice du noyau. Elle est définie semi-positive si un noyau valide est utilisé.

5.3.3 Interprétation géométrique et avantage du noyau

Par l'algorithme de M-SVM, la paire de classes (p, q) est séparée par l'hyperplan $(\mathbf{w}_p - \mathbf{w}_q, b_p - b_q)$ et le jeu d'hyperplans $\{(\mathbf{w}_1, b_1), \ldots, (\mathbf{w}_Q, b_Q)\}$ est choisi de façon à minimiser $\sum_{i,j=1}^{Q} \|\mathbf{w}_i - \mathbf{w}_j\|^2$.

L'interprétation de la marge dans le cas bi-classe ne peut pas être étendue naturellement au cas multi-classe. Une illustration des frontières calculées par la M-SVM est représentée dans la Figure 5.9 pour le cas séparable.

L'avantage du noyau expliqué dans le cas bi-classe reste le même dans le cas multi-classe. L'aspect théorique des méthodes à noyau fait l'objet d'études fondamentales d'actualité. Cette approche permet d'améliorer l'apprentissage et la prédiction des étiquettes sur des données multi-classes. Ces travaux théoriques sont principalement motivés par la

biologie qui est source de problèmes statistiquement intéressants. Ainsi, le dé veloppement de SVM multi-classes dédiées aux séquences biologiques est un domaine d'application privilégié pour de nombreux travaux concernant le choix du noyau.

5.4 Conclusions

Pour beaucoup d'expérimentateurs biologistes, la mise en œuvre des méthodes à noyau reste mal appréhendée d'un point de vue pratique. Pourtant la démarche reste accessible sans pré-requis de statistique. Les équations présentées précédemment se résolvent grâce à des algorithmes de programmation quadratique ou linéaire qu'il est possible de télécharger sur le web[12]. Il existe ainsi de nombreuses SVM à disposition des biologistes qui sont soit génériques soit dédiées à des problèmes spécifiques.

Les études théoriques confèrent un très grand potentiel à ces méthodes statistiques notamment dans l'application au domaine biologique. L'approche de l'épissage alternatif et la recherche spécifique de motifs de fixation de protéines de régulation dans un cadre statistique sera discuté plus tard mais il semble être un problème adéquat à la mise en œuvre d'une machine à noyau. C'est pourquoi nous développerons la mise en œuvre d'une SVM multi-classe ainsi que tous les aspects pratiques d'un point de vue biologique dans le Chapitre 8.

[12]notamment sur le site des *kernel machine* : `http ://www.kernel-machines.org`

Deuxième partie

Analyse des expériences SELEX : Identification des motifs de régulation d'épissage

Chapitre 6

Approches théoriques des expériences SELEX

La technologie énormément a fait progresser la connaissance des systèmes biologiques. L'utilisation de techniques de pointes permettant de traiter automatiquement de grandes masses de données issues de l'expérimentation ont notamment contribué à l'essor de la biologie moléculaire. On peut illustrer ceci par les techniques de séquençage de génomes. L'une de ces techniques appelée méthode SELEX permet d'expérimenter la fonction de certaines séquences biologiques. Elle caractérise les molécules ligands qui se fixent par affinité sur une molécule cible donnée. Cette méthode est quasiment automatique dans sa démarche et permet ainsi d'inférer une fonction biologique de manière impartiale. Cette dernière caractéristique confère au biologiste un résultat objectif de son problème, et ceci à condition d'exploiter correctement les données SELEX.

Face au potentiel de cette méthode, une démarche informatique adéquate est nécessaire et nous allons tenter de la caractériser dans ce chapitre. De nombreux paramètres doivent alors être pris en compte afin d'optimiser le traitement des résultats expérimentaux. Ainsi, après un rappel du protocole expérimental dans la Section 6.1, nous dégagerons dans la Section 6.2 les modèles mathématiques qui représentent au mieux les expériences SELEX afin d'en dégager les paramètres importants pour la méthodologie et les fondements théoriques associés. Nous serons alors en possession des fondamentaux qui nous permettront de juger des techniques usuelles d'exploitation des données SELEX dans la Section 6.3 afin de mieux en expertiser les biais éventuels dans la Section 6.3.2

6.1 Protocole expérimental SELEX

La technique expérimentale du SELEX a été développée par [Tuerk & Gold, 1990] et signifie *Systematic Evolution of Ligands by EXponential enrichment*. Cette technique issue de la chimie combinatoire met en évidence les molécules ligands qui vont se fixer par affinité sur une autre molécule cible donnée. La nature des ligands que les expériences SELEX caractérisent, résulte des affinités et des concentrations en molécules cibles. Le protocole SELEX est itératif et se déroule suivant les 4 étapes suivantes (illustrées par la Figure 6.1) :

1. **Génération d'une banque de ligands potentiels :** On procède pour cela à une synthèse aléatoire de molécules. Dans notre cas, ce sont des acides nucléiques de séquences aléatoires qui sont générés. Tous ces acides sont des ligands potentiels.
2. **Accrochage des molécules cibles :** Les ligands potentiels sont mis en contact avec des molécules cibles. Dans notre cas, les molécules cibles seront des protéines de régulation d'épissage. Ce contact permettra la formation de complexes entre les acides nucléiques et les protéines qui peuvent se fixer sur les ARN.
3. **Partitionnement des ligands fixés de ceux non fixés :** On effectue ensuite une filtration des complexes et des molécules encore libres. On récupère ainsi les ligands qui possèdent une affinité pour les molécules cibles. C'est à ce niveau que s'effectue la sélection des ligands.
4. **Amplification des éléments fixés :** Les complexes sélectionnés, on récupère les ligands par un protocole de *reverse transcription*. Les ligands ont subi à ce stade un cycle SELEX de sélection. La base de ligands que l'on récupère est donc enrichie en ligands ayant une bonne affinité avec les molécules cibles. Il est nécessaire de faire plusieurs cycles de sélection pour obtenir un résultat expérimental qui puisse rendre compte de tous les ligands pour une molécule cible. L'expérimentateur décide de mettre fin à la succession des cycles lorsqu'il apparaît que les données obtenues correspondent aux critères biologiques fixés initialement par le biologiste.

Un des grands avantages de cette approche expérimentale est de pouvoir générer de manière automatique et sans *a priori* des molécules qui possèdent toutes le point commun d'avoir une affinité pour la molécule cible. Toutes ces molécules sont donc des ligands potentiels qui n'auraient peut être pas été mis en évidence avec d'autres protocoles expérimentaux qui tiennent compte du contexte biologique. Prendre en compte le contexte biologique consiste à caractériser des molécules qui possèdent une valeur biologique intrinsèque en étant par exemple issues d'un milieu cellulaire *in situ*. Le contexte biologique permet ainsi de trier les molécules biologiques des molécules *artificielles*. Face à ce problème d'identification et d'analyse des résultats expérimentaux, une exploitation efficace est nécessaire.

La méthode expérimentale SELEX génère des ligands pour une molécule cible. Après la sélection expérimentale, on est en mesure de caractériser les molécules qui possèdent une fonction biologique définie. Au cours de l'épissage alternatif, la formation des complexes de transcription se fait principalement par affinité. Cette même affinité est la source de fixation des molécules de régulation de l'épissage alternatif comme les protéines SR ou hnRNP sur l'ARN immature. Ces mêmes protéines de régulation peuvent alors être considérées comme des molécules cibles qui vont, par affinité, pouvoir fixer des ligands qui sont alors des morceaux d'ARN immatures : les motifs de fixation des protéines de régulation. Les expériences de SELEX appliquées à des molécules telles que les protéines SR permettent ainsi de caractériser de manière automatique des motifs de régulation spécifiques. La méthode SELEX représente alors une grande opportunité pour l'annotation fonctionnelle de génomes.

Les données SELEX en notre possession caractérisent ce type d'information pour des protéines de régulation d'épissage alternatif. Ces données correspondent aux résultats ex-

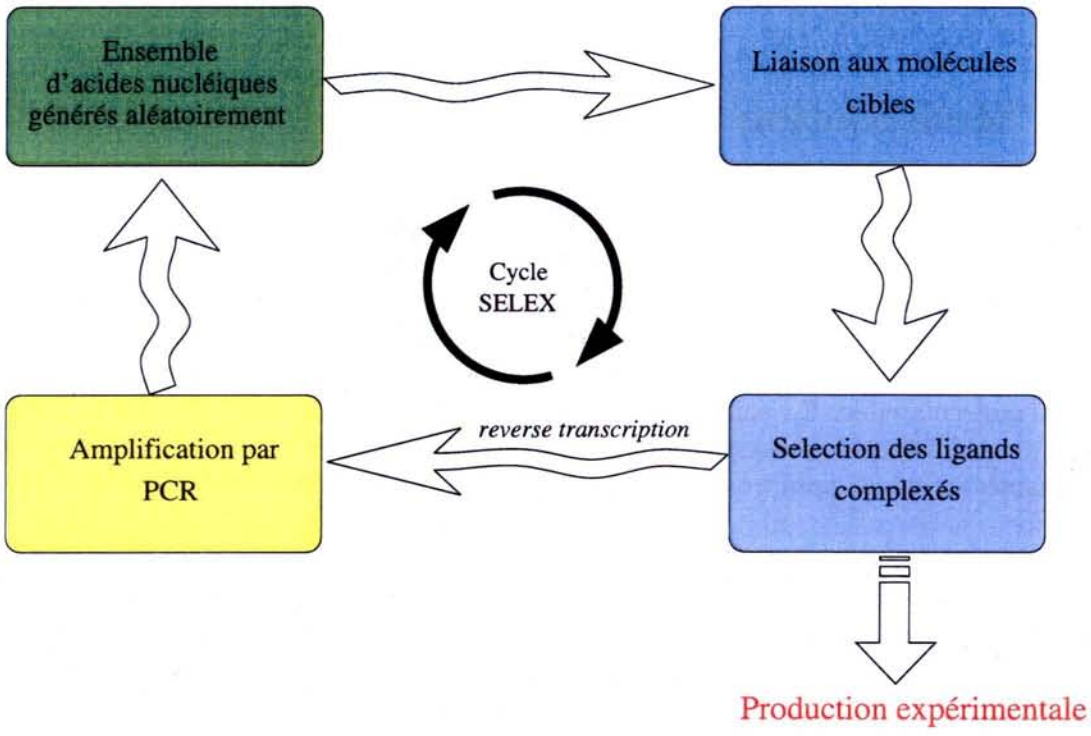

FIG. 6.1 – **Schématisation du protocole expérimental SELEX** (*Systematic Evolution of Ligands by EXponential enrichment*)**.** Le protocole repose sur la synthèse aléatoire d'acides nucléiques. Ils sont ensuite mis en contact avec la molécule cible (une protéine dans notre cas d'étude). Par affinité les acides nucléiques et les molécules cibles forment des complexes qui sont ensuite sélectionnés. Par un protocole de *reverse transcription* il est possible de déterminer les acides nucléiques qui possèdent une affinité pour les molécules cibles. Les acides nucléiques ont alors subi un processus de sélection SELEX (appelé aussi un cycle SELEX). La base des molécules sélectionnées peut être à nouveau enrichie par un nouveau cycle qui purifie les ligands possédant la plus forte affinité.

périmentaux de James Stevenin[13] qui isolent les motifs de fixation des protéines SC35 et 9G8 après différents cycles de sélection. Les données se présentent sous la forme de séquences nucléiques. Les données sont relativement disparates en fonction du nombre de cycles de sélection SELEX. Cette hétérogénéité est associée à l'absence de contexte biologique dans les expériences SELEX. Cela justifie donc une analyse théorique des expériences afin de mieux les comprendre. Ce constat est d'autant plus renforcé que le SELEX est issue de la chimie combinatoire, domaine qui se prête particulièrement bien à la modélisation statistique.

6.2 Modélisation statistique des expériences SELEX

La motivation de modéliser le comportement des expériences SELEX est initialement économique. Les industries pharmaceutiques pratiquent de manière soutenue ce protocole expérimental afin de caractériser des ligands qui auront un fort pouvoir thérapeutique et par conséquent économique sur le marché du médicament. Un des principaux problèmes du SELEX au niveau industriel est de savoir à quel moment arrêter les cycles de sélection pour avoir suffisamment de molécules intéressantes, sans conserver de ligands de faible affinité pour la molécule cible. En considérant que le SELEX peut générer différents ligands en fonction des différentes concentrations et des différentes affinités des molécules cibles, tous ces paramètres sont à prendre en compte dans une modélisation mathématique.

6.2.1 Modélisations mathématiques

Différents travaux ont analysé l'approche expérimentale SELEX d'un point de vue théorique. Les premières études théoriques des expériences SELEX [Sun *et al.*, 1996] se sont focalisées sur les études de SELEX *in vitro*. Parallèlement les travaux de [Burke *et al.*, 1996] étudient les applications de cette technique expérimentale lors de l'étude des inhibiteurs de reverse transcriptase sur l'ARN du virus HIV-1. Ces différentes études aboutissent sur un modèle mathématique qui fait apparaître la nécessité d'une analyse plus statistique développée dans [Vant-Hull *et al.*, 1998]. Nous proposons ici de synthétiser les divers résultats obtenus.

Le modèle mathématique repose sur différentes hypothèses majeures. La première considère comme d'autres études, que l'amplification par PCR est un processus linéaire. Le modèle décrit le processus de sélection SELEX après que les processus atteignent un équilibre. On représente alors la distribution des n ligands L_j qui possèdent une affinité K_j pour une protéine P, avec $j = 1, 2, \ldots, n$. Les équations du modèle mathématique se basent alors sur le principe de la conservation de la matière et l'équilibre du complexe protéine-ligand.

$$[L_j] = \frac{L_j^{tot}}{1 + K_j [P]} \qquad et \qquad [P] = \frac{P^{tot}}{1 + \sum_j K_j [L_j]} \tag{6.1}$$

avec

[13] du laboratoire IGBMC à Illkirch

- $[L_j]$: la concentration en ligands L_j libres
- $[P]$: la concentration en protéine P libre
- P^{tot} : concentration totale en protéine
- $\left[L_j^{tot}\right]$: concentration totale des ligands L_j liés ou non
- K_j : l'affinité ou constante d'association du ligand L_j pour la protéine P définie par $K_j = [L_jP] / [L_j][P]$, $[L_jP]$ étant alors la concentration du complexe ligand-protéine.

Cette équation est la base du modèle qui représente la variation des concentrations en ligands. La somme des ligands qui sont retenus après un cycle de sélection est alors $m_j = \text{eff}\,[L_jP] + bg\,[L_j]$, ce qui correspond à la quantité de complexes réduits par l'effet de partitionnement (eff) de la protéine auxquelles on ajoute une quantité de base (bg) de ligands libres aussi sélectionnés. Compte tenu de ces paramètres inhérents aux expériences et de l'amplification de l'ensemble des acides nucléiques au cours des cycles SELEX, il est plus commode de représenter la quantité des ligands pour différentes affinités par une fréquence de ligands définit comme :

$$[f_j] = \frac{m_j}{\sum_k m_k} = \frac{(\text{eff}K_j\,[P] + bg)\,[L_j]}{\sum_k (\text{eff}K_k\,[P] + bg)\,[L_k]} \tag{6.2}$$

avec :

- m_j total des ligands sélectionnés par le cycle de SELEX
- f_i fréquence de concentration relative de ligands résistant à la sélection
- eff efficacité de partitionnement : fraction de complexes protein-ligand retenue par l'expérience
- bg fraction de ligands libres retenue par l'expérience
- $[L_k]$ concentration en ligands issus de la précédante sélection

Cette équation représente le partitionnement qui a lieu à chaque cycle SELEX en fonction de divers paramètres. Pour analyser le modèle, il est généralement considéré que les affinités possèdent une distribution *log-normale* [Gold *et al.*, 1997]. Cette distribution est la conséquence de la corrélation qu'il existe entre l'énergie de liaison et l'information contenue dans les acides nucléiques. Les motifs qui possèdent le plus d'affinité pour la molécule cible se situeront alors en queue de distribution. A partir de ce postulat, il est possible d'extraire diverses réflexions sur les expériences SELEX.

6.2.2 Apports du modèle

Mise en évidence des limites expérimentales

La simulation des expériences SELEX par ce modèle mathématique est de complexité exponentielle du nombre de cycles de sélection. Néanmoins, elles sont en mesure de décrire le comportement théorique des expériences pour évaluer les conséquences de variations de paramètres divers comme l'affinité pour chaque ligand ou le nombre de cycles. Le modèle présenté par [Vant-Hull *et al.*, 1998] pose un certain nombre de problématiques comme les limites théoriques des expériences. Ainsi pour une séquence de longueur L, il existe 4^L séquences d'acides nucléiques différentes. On considère également une variabilité initiale en molécules cibles de l'ordre de 10^{15} molécules d'ARN. Suivant ces deux hypothèses, la

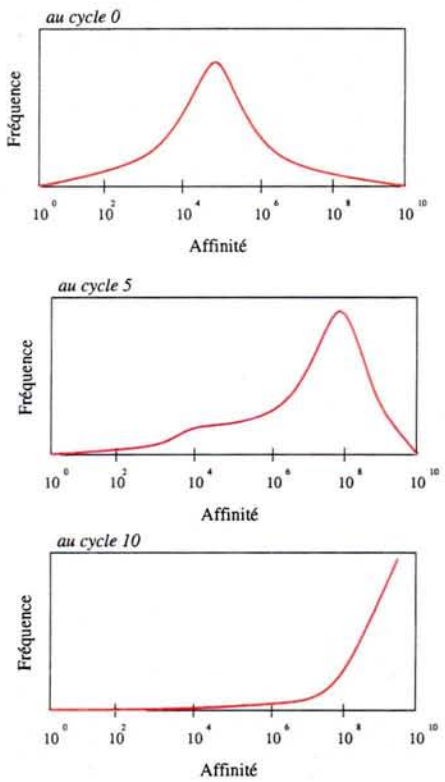

FIG. 6.2 – **Évolution de la distribution de l'affinité au cours des cycles SELEX d'après [Vant-Hull *et al.*, 1998].** Ces graphiques représentent les résultats sous des conditions standards. Pour 0 cycle de sélection, la distribution des affinités dans l'ensemble de ligands initiaux est considérée comme log-normale. Après 5 cycles de sélection, la distribution change pour donner des ligands de fortes affinités après 6 cycles. Après ce nombre de sélections, les ligands sont tous de fortes affinités comme le montre la distribution à 10 cycles de sélection.

fenêtre de variation de la longueur des séquences qui rend compte d'un résultat significatif de sélection est de $15 \leq L \leq 25$. Ainsi, pour une variation extérieure à ces seuils de 15 et 25 nucléotides, les expériences SELEX vont **générer une séquence unique qui ne possédera pas forcément de justification statistique**.

Variations de l'affinité des ligands

Le modèle met également en exergue une variation de l'affinité globale des ligands pour les protéines cibles au cours des cycles SELEX. On considère suivant les hypothèses du modèle que la distribution des affinités globales qui est initialement de forme log-normale va évoluer au cours du temps pour devenir exponentielle après 10 cycles de sélection (voir Figure 6.2). Cette observation corrobore le fait que l'affinité globale augmente au cours du nombre de cycles SELEX. Cette modification s'explique par la sélection exponentielle que subissent les ligands.

Face à cette évolution, on peut s'interroger sur les liens qui unissent les différents

événements de sélection d'une expérience SELEX. Le mode de sélection est stochastique. Cela signifie que les différents événements sont interdépendants dans le temps. Concrètement une séquence sélectionnée au cours d'un cycle modifiera la sélection des autres ligands lors des cycles suivants. Cette interdépendance sera d'autant plus importante que les séquences seront similaires entre elles.

Ces différentes caractéristiques statistiques démontrent que les expériences SELEX ne sont pas reproductibles. Les séquences à forte affinité sont relativement proches. Il existe alors en début d'expérience une très grande incertitude quant à la sélection d'une séquence par rapport à une autre. Ainsi, une séquence de forte affinité peut être éliminée dès les premiers cycles du fait de sa compétition avec d'autres ligands de forte affinité et ce malgré un fort potentiel de fixation.

Fluctuation de l'information au cours des cycles SELEX

Les expériences SELEX rendent compte d'une compétition entre différents ligands pour une protéine cible. Les séquences nucléiques sont donc soumises à un mécanisme de sélection courant en biologie. Néanmoins, les conditions *in situ* ou *in vivo* sont différentes de celles imposées chimiquement par les expériences SELEX. Si l'on considère cet aspect ainsi que le caractère stochastique et non réplicatif du SELEX, on est en droit de s'interroger sur l'information que génère les expériences.

Le protocole SELEX est souvent simplifié à une répétition de cycles de sélection jusqu'à obtention de banques de données suffisamment purifiées. L'objectif de l'expérimentateur est ainsi de simplifier les données expérimentales en une séquence consensus qui va représenter explicitement le ligand qui sera sélectionné par le protocole SELEX. Le raccourci utilisé est donc d'effectuer un grand nombre de cycles afin d'obtenir une séquence qui sera le représentant légitime de la base de ligands.

Selon la variation de l'affinité des ligands, cette démarche possède un risque certain. Il est en effet possible que les expériences convergent vers un unique consensus qui possède une valeur expérimentale uniquement dans un contexte chimique. Si l'on fait l'hypothèse que les ligands importants dans un contexte biologique ne sont pas ceux possédant la plus grande affinité, les expériences SELEX à grands nombres de cycles de sélection perdent de l'information biologique pertinente. Dans ce nouveau contexte, le modèle mathématique des expériences SELEX permet de caractériser la notion de nombre optimal de cycles SELEX (voir Figure 6.3). Un nombre optimal de cycles maximisera le nombre de ligands dans le contexte chimique. On supposera ensuite que les ligands potentiels du contexte biologique font parti de l'ensemble ainsi sélectionné.

6.3 Analyses standards des résultats de SELEX

La méthode SELEX génère un grand nombre de séquences biologiques qui sont toutes potentiellement des ligands pour une protéine cible. La diversité que produisent les expériences nécessite d'appliquer des outils bioinformatiques qui permettent d'extraire des bases de ligands l'information qui intéresse les biologistes.

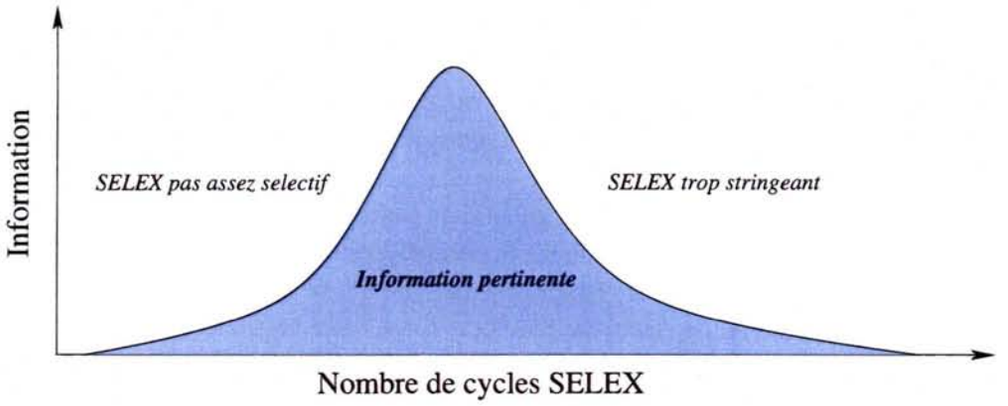

FIG. 6.3 – **Estimation de l'évolution de l'information biologique au cours des cycles SELEX** . L'information biologique pertinente varie au cours des cycles de sélection. Pour un faible nombre de cycles, les ligands ne sont pas assez purifiés et l'information pertinente est contenue dans peu de ligands sur l'ensemble des ligands sélectionnés. Les cycles de SELEX vont enrichir l'ensemble de ligands d'informations biologiques pertinentes. Pour un nombre trop important de cycles, seul un type de ligands de forte affinité sera sélectionné sans avoir forcément de valeur biologique. Le SELEX devient alors stringent envers un unique ligand de forte affinité sans retenir les autres ligands qui possèdent aussi une information pertinente. Un nombre trop important de cycles de sélection diminue l'information que génère le SELEX. Il existe donc un nombre de cycles optimal qu'il faut expérimentalement atteindre.

6.3.1 Méthodes bioinformatiques

Les résultats de SELEX sont constitués d'un ensemble de séquences biologiques qui sont tous potentiellement des ligands pour une molécule cible. Dans les données à notre disposition, les séquences sont des acides nucléiques qui sont toutes potentiellement des motifs de fixation pour une protéine SR qui est la molécule cible. La démarche pourra analyser ces résultats et les synthétiser en une ou plusieurs séquences consensus qui seront les représentants des ligands sélectionnés par les expériences SELEX.

Les méthodes qui permettent de générer un consensus alignent les séquences issues des expériences SELEX. Il est possible d'aligner les séquences dans leur totalité. On procède alors à un alignement global [Needleman & Wunsch, 1970]. Mais il est également possible d'aligner les séquences en considérant les morceaux de séquences. On procède alors à un alignement local [Smith & Waterman, 1981] ([Durbin *et al.*, 1998] pour revue). Ces deux procédés reposent sur un algorithme de programmation dynamique qui permet de retrouver le motif commun aux résultats expérimentaux. Le motif est un consensus qui possède les propriétés communes des séquences biologiques après alignement. Le motif consensus est alors une moyenne ou une tendance des motifs intrinsèques aux échantillons. Une tendance ou une moyenne auront l'avantage de caractériser une expression d'un motif mais pourront paradoxalement éliminer de l'information d'un autre motif sous-jacent. Un consensus unique ne pourra pas ainsi rendre compte de manière efficace de deux motifs biologiques dans les échantillons.

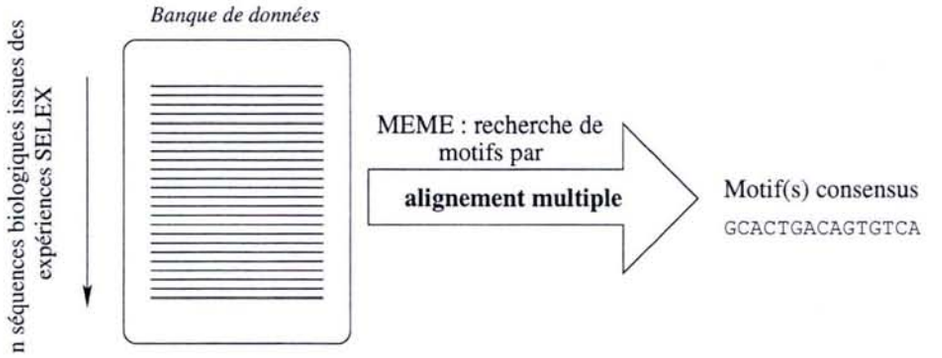

FIG. 6.4 – **Analyse standard des données SELEX.** De manière usuelle MEME [Bailey & Gribskov, 1998] ou ClustalW [Higgins *et al.*, 1994] permettent d'extraire les motifs. Il ressort de cette analyse une ou plusieurs séquences consensus qui possèdent la plus forte affinité pour une protéine motrice de sélection. Le motif devient le représentant biologique des séquences nucléiques obtenues expérimentalement. Par raccourci, on considère les motifs comme étant le représentant des sites de fixation pour la protéine donnée.

Dans le but de contrer ce biais, une alternative consiste à aligner les échantillons par une méthode d'alignement multiple de type local (voir [Duret & Abdeddaim, 1999] pour revue). MEME [Bailey & Gribskov, 1998, Bailey & Elkan, 1994] qui est une méthode d'extraction de motifs utilise l'une de ces méthodes, et notamment l'algorithme bayesien MM qui repose sur le principe inductif de maximum de vraisemblance (utilisé pour la première fois en biologie par [Lawrence & Reilly, 1990]). Il permet de mettre rapidement en évidence le ou les motifs communs aux séquences nucléiques d'une même base d'échantillons qui ne sont pas forcément de la même taille. Une représentation synthétique des motifs est constituée par l'ensemble des séquences consensus qui est retourné par l'algorithme d'alignement..

6.3.2 Incohérences des analyses standards

Différents aspects mettent en évidence des incohérences de cette approche sur les données SELEX appliquée aux protéines SR. Ils sont tout d'abord d'ordre biologique puisque les motifs de fixation des protéines SR identifiés dans la littérature ne sont pas consistants entre eux. Cette observation permet dans un deuxième temps d'appréhender les incohérences d'un point de vue statistique.

Caractérisation biologique des incohérences

Les protéines SR comme nous l'avons vu dans le Chapitre 1 jouent un rôle particulièrement important, puisqu'elles permettent une régulation des sites d'épissage constitutif et alternatif des systèmes biologiques hautement différenciés. Elles sont très conservées en taille et en séquence dans tous les organismes chez lesquels elles existent. Leurs rôles ne se limitent pas à la réaction d'épissage, puisqu'on les retrouve sur les ARN après matura-

Protéine SR	Séquences consensus	Auteurs
ASF/SF2	AGGACAGAGC RGAAGAAC	[Tacke & Manley, 1995*a*]
	(G/C)R(G/C)A(G/C)GA	[Liu *et al.*, 1998]
SC35	GUUCGAGUA	[Tacke & Manley, 1995*a*]
	GGGUAUGCUG	[Cavaloc *et al.*, 1999*a*]
	UGCNGYY	[Schaal & Maniatis, 1999*b*]
SRp40	GAGCRGUYRGCUC	[Tacke *et al.*, 1997]
	AC(A/G/U)G(G/C)	[Liu *et al.*, 1998]
9G8	AGAC(G/U)ACGAY ACGAGAGAY	[Cavaloc *et al.*, 1999*a*]
	AGAC(G/U)ACGA(C/U)	[Lejeune *et al.*, 2001]

TAB. 6.1 – Ensembles de séquences consensus obtenues pour différentes protéines SR. Les parenthèses signalent un choix multiple à une position donnée de la séquence. Les symboles sont standards (R : purine, Y : pyrimidine, N : un des 4 nucléotides).

tion. Globalement, elles renforceraient la stabilité des multiples interactions ARN/ARN et ARN/protéines au sein de complexes spliceosomaux. Elles possèdent des séquences préférentielles de fixation à l'ARN qui leur permettent de moduler l'utilisation de sites d'épissage alternatif. La recherche des motifs nucléiques est donc au centre de l'étude de leurs fonctions, ce qui en fait de manière duale des candidates idéales pour l'évaluation de l'exploitation des expériences SELEX.

De nombreuses études se sont ainsi appliquées à mettre en évidence les motifs susceptibles de fixer ces protéines. Elles ont pu donner lieu, pour une même protéine, à des résultats entièrement différents, comme on pourra le constater dans le Tableau 6.1.

Les motifs identifiés dans la littérature ne sont pas consistants. Diverses hypothèses permettent d'expliquer ces divergences. Une des premières consiste à considérer qu'un seul des motifs décrit est représentatif des sites de fixation d'une protéine SR donnée. Cependant, la qualité des travaux et des données expérimentales laisse à penser à une autre hypothèse qui suppose la présence de plusieurs motifs pour une même protéine SR. Dès lors, les motifs consensus générés par les méthodes standards d'alignement ne sont pas suffisamment représentatifs de cette variabilité de motifs. Le biais pourrait alors provenir des méthodes qui ne sont pas assez précises ou mal employées avec une telle diversité de motifs à décrire avec peu de motifs consensus.

Dans ce contexte d'incertitude, il est important de tester les résultats des méthodes d'alignement sur les données à notre disposition afin de conclure quant à la présence de plusieurs motifs ou non dans les échantillons de SELEX à notre disposition. Dans ce but, nous proposons une démarche statistique qui analyse la consistance des résultats d'alignements des résultats de SELEX.

Caractérisation statistique des incohérences

D'un point de vue statistique, les résultats d'alignements ne doivent pas varier pour une base d'échantillons donnés. Il en découle ainsi que l'ensemble des consensus produits par la méthode doit être un invariant en bijection avec l'ensemble des motifs recherchés. Les résultats ne doivent pas être sensibles aux fluctuations expérimentales, comme les erreurs conduisant à l'élimination d'une ou plusieurs séquences nucléiques. Nous proposons d'évaluer cette propriété sur une des protéines SR du tableau 6.1, la protéine 9G8.

Nous disposons pour cela de données SELEX d'affinité pour la protéine 9G8 obtenues après 11 cycles de sélection [Lejeune *et al.*, 2001]. Les 56 séquences nucléiques produites possèdent une longueur variant entre 18 et 22 bases nucléiques. L'application de MM sur l'ensemble de ces données produit un seul consensus significatif au sens du critère utilisé (issu de la théorie de l'information), `AGAC(U/A)ACG`. Ce consensus, que nous qualifierons par la suite de total, est pratiquement identique au consensus commun aux études de [Lejeune *et al.*, 2001] et [Cavaloc *et al.*, 1999*a*] (voir le tableau 6.1).

Pour tester la robustesse de la méthode de SELEX, on s'intéresse à la variabilité du consensus obtenu lorsque la base de données est amputée d'une à deux séquences. Les sous-ensembles de la base à 55 et 54 éléments étant trop nombreux pour être tous considérés, on extrait de cet ensemble un sous-ensemble représentatif de petite taille. Le critère utilisé pour caractériser la représentativité est la similarité de la ou des séquences consensus produites à la séquence consensus totale. Cette mesure est donnée par le score renvoyé par l'algorithme d'alignement global utilisé ici, celui de [Needleman & Wunsch, 1970]. Plus précisément, on cherche à rendre compte du spectre des similarités obtenues, en assurant en particulier la présence de représentants parmi les sous-ensembles correspondant aux consensus divergeant le plus fortement. La sélection résulte d'un tirage aléatoire sans remise selon une distribution uniforme. Un raisonnement simple permet d'établir qu'en tirant ainsi 59 sous-bases, la probabilité qu'aucun consensus résultant ne se trouve parmi les 5% les plus divergents est inférieure à 5%. Ce protocole expérimental est résumé sur la Figure 6.5.

Une sélection des consensus *partiels* obtenus est présentée dans le tableau 6.2. Elle est caractéristique des différents niveaux de similarité entre ces consensus et le consensus total.

Les scores de l'alignement global [Needleman & Wunsch, 1970] des deux derniers consensus partiels avec le consensus total sont respectivement égaux à -4 et -6. L'analyse biologique, réalisée par Fabrice Leclerc du laboratoire MAEM, établit que ces deux consensus se distinguent significativement des autres. La présence de cette paire constitue un contre exemple démontrant que les résultats d'analyse SELEX pour la protéine 9G8 ne sont pas consistants.

L'algorithme MM peut donc produire des consensus très différents pour des conditions expérimentales très proches. Ainsi que nous l'avons indiqué, une cause de cette instabilité peut être la présence de plusieurs motifs au sein des données SELEX, motifs associés à des familles de séquences nucléiques différentes. Il nous appartient alors pour bien analyser les données SELEX à notre disposition de rechercher les sous ensembles. Il existe dans ce cas plusieurs alternatives qui se basent toutes sur une démarche de classification. Il nous sera alors possible de caractériser des ensembles biologiquement homogènes sur lesquels

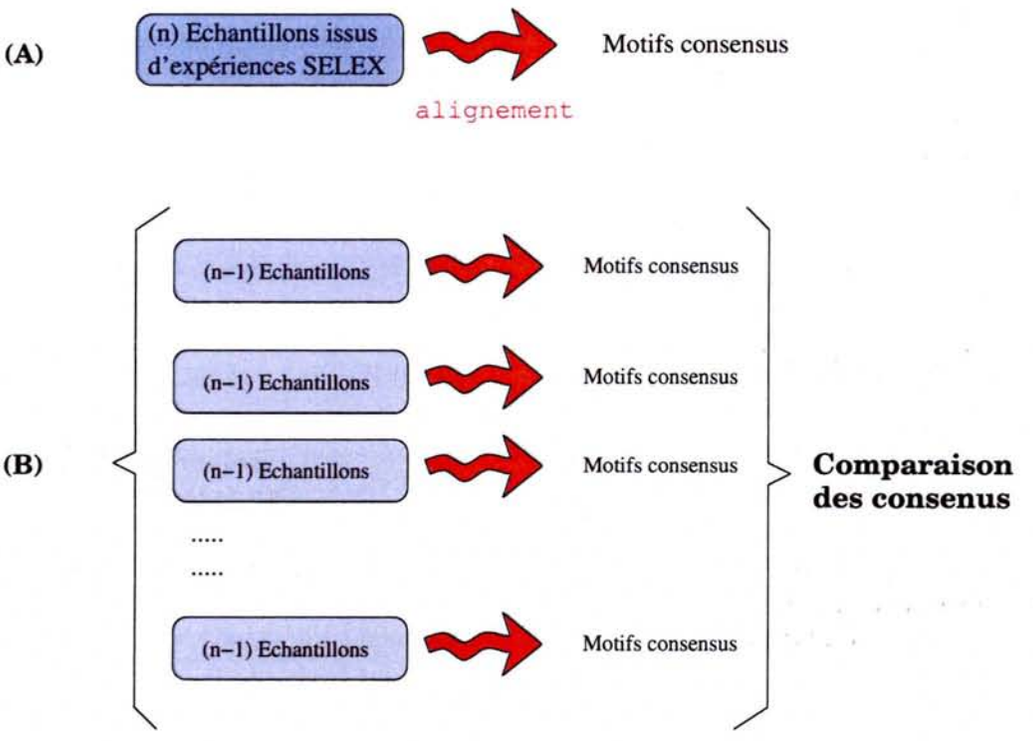

FIG. 6.5 – **Protocole de test de la stabilité des analyses SELEX. (A)** représente une analyse standard d'analyse de résultats SELEX. Un alignement permet de synthétiser la base d'échantillons par un ou plusieurs motifs consensus. **(B)** représente le test de l'alignement sur la base d'échantillons. La démarche consiste à extraire aléatoirement une séquence nucléique de la base. On itére cette opération 59 fois pour avoir une représentation statistiquement significative. On compare ensuite les consensus générés. Pour des consensus différents, l'alignement est invalide.

Consensus pour la protéine 9G8
AGAC(U/A)ACG (total)
...
ACGACGAU
...
ACGACGAU
...
CUACGCUA
GACGAGAGAUU

TAB. 6.2 – **Consensus total et consensus partiels obtenus après ré-échantillonnage**. Ces derniers sont classés par ordre de similitude décroissante avec le consensus total.

nous pourrons affecter une inférence biologique.

6.4 Conclusion

Les expériences SELEX sont intéressantes à différents points de vue. Premièrement, ce sont des expériences quasi-automatiques, dans le sens ou l'inférence du biologiste n'intervient qu'à la fin du protocole expérimental. De nombreux travaux théoriques, illustrés en début de chapitre, se sont alors penchés sur le caractère stochastique particulier de ces expériences. Les SELEX apparaissent alors comme non reproductibles.

D'un point de vue biologique, les données SELEX fournissent des échantillons d'une réalité biologique qu'il faut inférer. Les théories d'échantillonnage qui sont bien connues, nous ont appris l'importance des fluctuations des données. Il est donc essentiel de les prendre en compte. Une méthodologie adéquate apparaît alors comme naturelle. Avant d'inférer et de généraliser les données SELEX, comme peuvent le faire les motifs consensus, il est important de classer les échantillons pour vérifier si les données sont assez bien discriminées entre elles. Cette étape de validation théorique nous permettra ensuite d'inférer les hypothèses biologiques sous-jacentes aux données SELEX.

Chapitre 7

Classification et analyses des résultats de SELEX

Le chapitre précédent nous a démontré que les expériences SELEX restent théoriquement complexes. Cette opinion est confirmée par la mise en évidence statistique des incohérences des approches standard d'analyses des données. C'est un facteur d'incertitude qui est peut être accentué dans le cas des protéines SR qui sont reconnues comme particulièrement hétérogènes. Par ailleurs, l'état actuel des connaissances expérimentales concernant ces protéines confirme que les séquences reconnues par les protéines SR peuvent être extrêmement divergentes [Jacquenet, 2001].

Dans ce contexte, il apparaît important de proposer une nouvelle méthode pour analyser les données SELEX qui puisse tenir compte des hétérogénéités apportées par les protéines SR tout en étant applicable aux autres types de protéines cibles. Une démarche naturelle avant de synthétiser les données en un motif consensus, est de classer les données afin de mettre en évidence des ensembles homogènes qui pourront seulement ensuite être synthétisés. Notre approche exposée dans ce chapitre repose sur ce principe. Nous utiliserons pour cela différents principes théoriques que nous exposerons brièvement dans la section 7.1 pour en présenter ensuite les résultats dans la section 7.2. Nous expliciterons en dernier lieu les avantages de notre démarche qui nous permet de comprendre la signification des ensembles que notre approche classificatoire caractérise.

7.1 Approche méthodologique

La démarche de classification que nous allons mettre en oeuvre pour analyser les données des expériences SELEX nécessite différents outils. Nous présenterons alors succinctement la méthode de pré-traitement que nous appliquerons aux données SELEX, ce qui nous permettra ensuite d'appliquer les modèles standard de classification.

Les outils de classification standard jouent sur des distances euclidiennes entre les différents échantillons. Afin de donner un mode de représentation spatiale aux séquences nucléiques, nous proposons d'utiliser les algorithmes d'alignement. L'approche consiste à comparer deux à deux les séquences par un algorithme. Il retourne de cette comparaison de séquences un score qui est un indice de la similarité entre les deux séquences alignées.

Les scores sont ensuite stockés dans une matrice de similitude. Ainsi, pour n séquences, la comparaison de celles-ci deux à deux donne une matrice quadratique de dimension $n \times n$. Cet algorithme utilise la programmation dynamique. Plus les séquences seront similaires et plus le score sera élevé. Nous proposons dans notre approche d'appliquer les outils de classification sur cette nouvelle représentation de la séquence en vecteur de similarité par rapport aux autres séquences.

Dans l'optique d'obtenir la matrice de similitude, nous avons travaillé avec différents algorithmes d'alignement [Duret & Abdeddaim, 1999]. L'alignement peut être global. On compare alors les séquences dans leurs totalités, c'est l'algorithme de [Needleman & Wunsch, 1970]. L'alignement peut reposer sur un algorithme dit local. On favorise alors la comparaison de morceaux de sous-séquences entre eux. C'est l'algorithme de [Smith & Waterman, 1981]. Ces deux algorithmes testés sur les données SELEX à notre disposition, convergent vers les mêmes résultats que ceux que nous présenterons par la suite. Cela s'explique par la taille restreinte des séquences que nous comparons. Elles sont en effet de 18 nucléotides en moyenne, ce qui ne permet pas de distinguer de différences significatives entre les deux approches de comparaison.

Les outils de classification que nous utiliserons dans notre démarche sont standard. Nous utiliserons d'une part les K-means qui permettent de distribuer les données en clusters [Hartigan & Wong, 1979] et les analyses en composantes principales hiérarchiques (ACP) ([Legendre & Legendre, 2000] pour revue) qui permettent de classer les données sans donner de nombres *a priori* d'ensembles, à la différence des K-means.

7.2 Résultats de la classification des données SELEX

Notre démarche pour analyser les données SELEX consiste à représenter les données dans un nouvel espace issue de l'alignement des données entre elles. Dans cet espace, nous sommes alors dans la capacité d'appliquer deux approches de classification que sont les K-means et l'ACP hiérarchique. Une synthèse de la méthodologie et des résultats obtenus après application a été présentée dans [Eveillard & Guermeur, 2002*a*] et dans [Eveillard & Guermeur, 2002*b*].

Traitement Statistique des Résultats de SELEX

Damien EVEILLARD[†‡] Yann GUERMEUR[†]

[†] LORIA (UMR 7503) – Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex
[‡] MAEM (UMR 7567) – Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex

Courriel : {Damien.Eveillard,Yann.Guermeur}@loria.fr

Résumé

Les expériences de SELEX constituent une source d'informations particulièrement précieuse en biologie moléculaire. Cependant, les données produites, qui peuvent présenter une variabilité importante, ne sont pas d'une exploitation aisée. Après avoir mis en lumière le fait que l'approche standard de cette exploitation peut conduire à des résultats incomplets, voire erronés, cet article propose de pallier les insuffisances rencontrées en introduisant une étape de traitement supplémentaire, de nature statistique. Cette étape, qui a pour but d'identifier le nombre de motifs à déterminer, ainsi que les ensembles de ligands qui leur sont associés, repose sur l'analyse de données et la classification. Elle précède la détermination des séquences consensus. Notre approche est validée sur un cas réel, pour lequel elle produit des résultats permettant de formuler des hypothèses confirmées a posteriori *par une étude biologique indépendante de la nôtre.*

Mots clés : *Données SELEX, classification, protéines SR, alignements multiples.*

1 Introduction

L'essor actuel de la biologie moléculaire appelle l'utilisation de techniques de pointe permettant de traiter automatiquement les grandes masses de données issues de l'expérimentation. La méthode du SELEX, destinée à mettre en évidence les molécules ligands se fixant par affinité sur une molécule cible donnée, entre dans cette catégorie. Son emploi est actuellement largement répandu. Si les données qu'elle produit sont riches d'information, leur exploitation est délicate. Nous soulignons ici les limites de l'approche standard, consistant à dériver une ou plusieurs séquences consensus à partir de l'ensemble des ligands obtenus. Cette approche peut avoir pour conséquence, suivant le cas, soit de produire des séquences consensus sans signification biologique (ne correspondant pas à un motif), soit de négliger des motifs. Afin de pallier ces insuffisances, nous proposons un prétraitement statistique des données SELEX aboutissant à un partitionnement des ligands. Les consensus sont alors calculés indépendamment pour chaque classe. Cette approche est illustrée sur un cas réel, celui de la protéine SR 9G8. Dans ce cas, les ligands obtenus expérimentalement ne possèdent pas de structures secondaires caractéristiques. Ainsi, nous traitons la situation la plus délicate, dans laquelle aucune connaissance biologique n'est disponible a priori pour orienter la recherche. Des arguments purement statistiques permettent dans ce cas de formuler une conjecture de nature biologique, l'existence de motifs multiples, qui se trouve validée par les observations résultant d'expériences effectuées indépendamment par une autre équipe.

L'organisation de cet article est la suivante. La section 2 contient un bref rappel sur le principe et la mise en œuvre de la technique du SELEX. La section 3 décrit l'exploitation standard des données SELEX et met en évidence ses imperfections. La classification des ligands et son influence sur la détermination des séquences consensus sont étudiées dans la section 4. La section 5 est dédiée à la validation biologique de notre principale conjecture.

2 Principe des expériences de SELEX

La technique expérimentale de *Systematic Evolution of Ligands by EXponential enrichment* (SELEX) [19] est une méthode du domaine de la chimie combinatoire qui permet d'extraire des ligands potentiels d'une banque d'oligonucléotides composée initialement de séquences engendrées aléatoirement. La nature des ligands obtenus résulte conjointement des affinités et des concentrations des molécules cibles. Le SELEX est une méthode itérative constituée de quatre étapes schématisées sur la figure 1.

Initialement, un ensemble d'acides nucléiques est engendré aléatoirement. La mise en présence de ces acides nucléiques avec une molécule cible produit des liaisons biochimiques dont l'intensité est fonction du couple considéré. Parmi les complexes constitués, on sélectionne ceux dont la liaison est suffisamment forte. Les ligands contenus dans les complexes sont ensuite identifiés puis amplifiés par *Polymerase Chain Reaction* (PCR). Un nouvel ensemble d'acides nucléiques est ainsi produit, sur lequel il est possible d'itérer le processus de sélection. L'expérimentateur décide de mettre fin à la succession des cycles lorsqu'il lui apparaît que les données obtenues correspondent aux critères biologiques fixés (forte affinité pour la molécule cible).

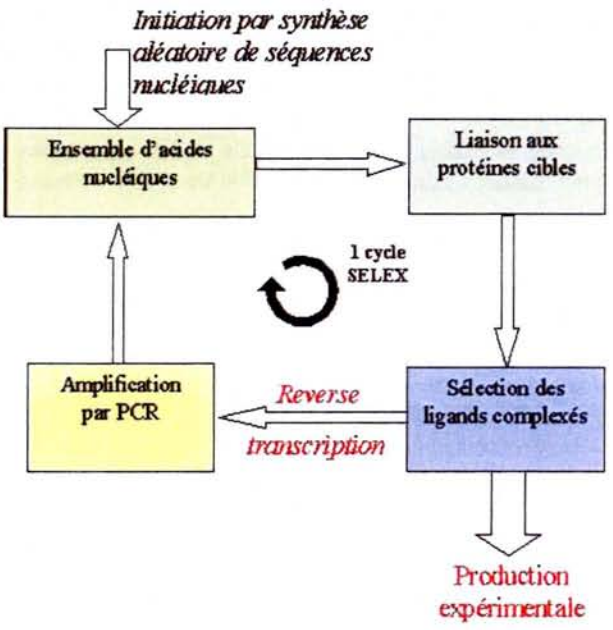

FIG. 1 – Etapes de la méthode SELEX

Dans ce qui suit, dans un but de simplification, l'exposé est restreint au cadre dans lequel se place notre étude, celui des interactions ARN-protéines. La méthode permet alors d'identifier les fragments d'ARN possédant un motif reconnu par la protéine cible. Sous l'hypothèse que ce motif est unique, il sera le représentant biologique du critère de sélection et sera contenu dans chaque séquence issue de l'expérience.
Du fait de son utilité et de sa performance, la méthode SELEX a suscité plusieurs analyses théoriques (voir par exemple [20, 16]). Elles établissent en particulier les faits suivants. La phase d'amplification des oligonucléotides conserve les proportions des différents ligands. Le partitionnement constitue l'étape expérimentale la plus sensible. L'affinité globale des ligands augmente de manière exponentielle avec le nombre de cycles SELEX. Du fait de la nature itérative de la méthode, les grandeurs d'intérêt (concentration et affinité des différents ligands), se comportent comme des processus stochastiques indicés par le numéro de cycle. Il est possible de déterminer, pour un problème donné, un nombre de cycles optimum. Tandis qu'un nombre trop faible de cycles ne permet pas de filtrer suffisamment les oligonucléotides, un nombre trop élevé réduit trop fortement la diversité des ligands obtenus.

3 Limites de l'exploitation standard des données du SELEX

Dans cette section, nous nous appuyons sur un exemple particulier d'expériences SELEX, représentant un enjeu biologique important, pour mettre en évidence les limites de l'approche standard du traitement des données produites.

3.1 Description de l'analyse standard

Chaque ligand issu d'une expérience de SELEX possède une instance d'au moins un motif satisfaisant le critère de sélection. La question se pose de produire un représentant synthétique d'un tel motif. L'approche la plus judicieuse consiste à effectuer un alignement multiple de type local [8]. L'algorithme MM [2, 1], qui repose sur le principe inductif de maximum de vraisemblance, est dans ce cadre expérimental le plus utilisé. Il permet de mettre rapidement en évidence le ou les motifs communs aux séquences biologiques qui lui sont fournies, que ces séquences possèdent ou non la même taille. La représentation synthétique des motifs est constituée par l'ensemble des séquences consensus qui sont retournées.

3.2 Nature des insuffisances rencontrées

Les protéines SR (voir par exemple [11]) jouent un rôle particulièrement important, puisqu'elles permettent une régulation des sites d'épissage constitutif et alternatif des sytèmes biologiques hautement différenciés. Elles sont très conservées en taille et en séquence dans tous les organismes chez lesquels elles existent. Leur rôle ne se limite pas à la réaction d'épissage, puisqu'on les retrouve sur les ARN après maturation. Globalement, elles renforceraient

Protéine SR	Séquences consensus	Auteurs
ASF/SF2	AGGACAGAGC RGAAGAAC	Tacke et Manley [18]
	(G/C)R(G/C)A(G/C)GA	Liu *et al.* [12]
SC35	GUUCGAGUA	Tacke et Manley [18]
	GGGUAUGCUG	Cavaloc *et al.* [4]
	UGCNGYY	Schaal et Maniatis [14]
SRp40	GAGCRGUYRGCUC	Tacke *et al.* [17]
	AC(A/G/U)G(G/C)	Liu *et al.* [12]
9G8	AGAC(G/U)ACGAY ACGAGAGAY	Cavaloc *et al.* [4]
	AGAC(G/U)ACGA(C/U)	Lejeune *et al.* [10]

TAB. 1 – Ensembles de séquences consensus obtenues pour différentes protéines SR. Les parenthèses signalent un choix multiple à une position donnée de la séquence. Les symboles sont standard (R : purine, Y : pyrimidine, N : un des 4 nucléotides).

la stabilité des multiples intéractions ARN/ARN et ARN/protéines au sein de complexes spliceosomaux. Elles possèdent des séquences préférentielles de fixation à l'ARN qui leur permettent de moduler l'utilisation de sites d'épissage alternatif. La recherche des motifs nucléiques est donc au centre de l'étude de leurs fonctions, ce qui en fait de manière duale des candidates idéales pour l'évaluation de l'exploitation des expériences SELEX.
De nombreuses études se sont ainsi appliquées à mettre en évidence les motifs susceptibles de fixer ces protéines. Elles ont pu donner lieu, pour une même protéine, à des résultats entièrement différents, comme on pourra le constater dans le tableau 1.
Il est possible d'avancer deux hypothèses pour expliquer ces variations. D'une part, les consensus peuvent correspondre à autant de motifs (possédant la même structure secondaire). Dans ce cas, le fait que la méthode ne les mette pas systématiquement tous en évidence constitue une faiblesse non négligeable. Un défaut plus important encore apparaît dans le cas où certains consensus ne correspondent pas à des motifs. Cette situation peut survenir lorsque différents motifs sont répartis sur des groupes de séquences différents que l'alignement ne distingue pas. Les consensus se presentent alors comme des résultats d'interférences qui n'ont pas eux-mêmes de sens biologique. Ces deux hypothèses appellent la réalisation d'études supplémentaires en préambule à l'alignement. C'est dans cette optique que nous proposons d'effectuer une analyse statistique des données issues du SELEX.

3.3 Caractérisation statistique des insuffisances

Il découle de ce qui précède que l'ensemble des consensus produits par la méthode doit être un invariant en bijection avec l'ensemble des motifs recherchés. Les résultats ne doivent pas être sensibles aux fluctuations expérimentales, comme les erreurs conduisant à l'élimination d'une ou plusieurs séquences nucléiques. Nous proposons d'évaluer cette propriété sur une des protéines SR du tableau 1, la protéine 9G8.
Nous disposons pour cela de données SELEX d'affinité pour la protéine 9G8 obtenues après 11 cycles de sélection [10]. Dans le cadre de notre étude, cette protéine possède l'"avantage" de ne pas fixer de structure secondaire particulière. Elle apparaît donc comme une candidate idéale pour une analyse reposant uniquement sur l'information de séquence. Les 56 séquences nucléiques produites possèdent une longueur variant entre 18 et 22 bases nucléiques. L'application de MM sur l'ensemble de ces données produit un seul consensus significatif au sens du critère utilisé (issu de la théorie de l'information), `AGAC(U/A)ACG`. Ce consensus, que nous qualifierons par la suite de total, est pratiquement identique au consensus commun aux études de Cavaloc et Lejeune (voir le tableau 1). Pour tester la robustesse de la méthode de SELEX, on s'intéresse à la variabilité du consensus obtenu lorsque la base de données est amputée d'une à deux séquences. Les sous-ensemble de la base à 55 et 54 élements étant trop nombreux pour être tous considérés, on extrait de cet ensemble un sous-ensemble représentatif de petite taille. Le critère utilisé pour caractériser la représentativité est la similarité de la ou des séquences consensus produites à la séquence consensus totale. Cette mesure est donnée par le score renvoyé par l'algorithme d'alignement global utilisé ici, celui de Needleman et Wunsch [13]. Plus précisément, on cherche à rendre compte du spectre des similarités obtenues, en assurant en particulier la présence de représentants parmi les sous-ensembles correspondant aux consensus divergeant le plus fortement. La sélection résulte d'un tirage aléatoire sans remise selon une distribution uniforme. Un raisonnement simple permet d'établir qu'en tirant ainsi 59 sous-bases, la probabilité qu'aucun consensus résultant ne se trouve parmi les 5% les plus divergeants est inférieure à 5%. Ce protocole expérimental

est résumé sur la figure 2.

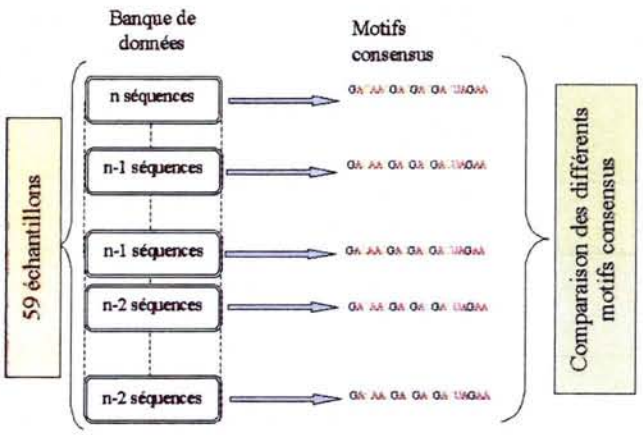

FIG. 2 – Protocole d'étude de la variabilité du consensus associé aux ligands issus d'une expérience de SELEX

Une sélection des consensus "partiels" obtenus est présentée dans le tableau 2. Elle est caractéristique des différents niveaux de similarité entre ces consensus et le consensus total.

Consensus pour la protéine 9G8
AGAC(U/A)ACG (total)
...
ACGACGAU
...
ACGACGAU
...
CUACGCUA
GACGAGAGAUU

TAB. 2 – Consensus total et consensus partiels obtenus après rééchantillonnage. Ces derniers sont classés par ordre de similitude décroissante avec le consensus total.

Les scores de l'alignement des deux derniers consensus partiels avec le consensus total sont respectivement égaux à -4 et -6. L'analyse biologique, réalisée par Fabrice Leclerc, établit que ces deux consensus se distinguent significativement des autres. La présence de cette paire constitue un contre exemple démontrant que les résultats d'analyse SELEX pour la protéine 9G8 ne sont pas consistants. L'algorithme peut produire des consensus très différents pour des conditions expérimentales très proches. Ainsi que nous l'avons indiqué, une cause de cette instabilité peut être la présence de plusieurs motifs au sein des données SELEX, motif associés à des familles de séquences nucléiques différentes. Un moyen de prendre en compte cette possibilité consiste à rechercher les ensembles caractérisés par un motif, en effectuant une classification.

4 Classification sur les séquences produites par la méthode SELEX

4.1 Choix des algorithmes

La classification de séquences fondée sur les scores d'alignement a souvent permis d'obtenir de bonnes performances en génomique. On pourra par exemple consulter sur le sujet [6, 5]. Nous reprenons ici cette stratégie. Deux méthodes standard de classification sont mises en œuvre sur les consensus, la méthode des k-means (voir entre autres [7]) et l'Analyse en Composantes Principales (ACP) hiérarchique, décrite dans [9]. Les scores issus de l'algorithme de Needleman et Wunsch sont de nouveau employés comme mesure de similarité (la "dissimilarité" entre deux séquence étant égale à l'inverse du score de la programmation dynamique). Notons cependant que des expériences supplémentaires, effectuées avec l'algorithme d'alignement local de Smith et Waterman [15], ont conduit aux mêmes résultats qualitatifs (même structure classificatoire).

4.2 Résultats de la classification

Tandis que la méthode des k-means impose de choisir *a priori* le nombre de catégories, l'ACP hiérarchique produit automatiquement une hiérarchie à partir de laquelle on peut construire une structure classificatoire plus ou moins fine, en fonction du seuil retenu. On observe ici une forte concordance des résultats des deux méthodes, ceci pour différents choix du nombre de clusters. Nous illustrons le phénomène dans le cas le plus pertinent, celui de deux classes. La figure 3 localise ainsi les clusters produits par les k-means sur la hiérarchie issu de l'ACP.

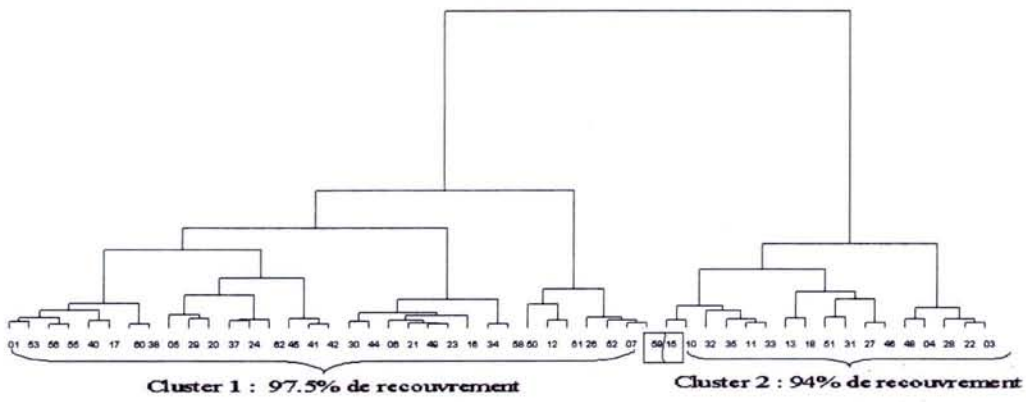

FIG. 3 – Superposition des résultats des deux classifications

Les deux classifications, dont le taux de recouvrement est de 96,4%, répartissent les données entre les deux clusters dans un rapport 3/4 - 1/4. (voir également le tableau 3 contenant la "matrice de confusion" correspondante).

	cluster 1 (k-means)	cluster 2 (k-means)
cluster 1 (ACP)	40	1
cluster 2 (ACP)	1	14

TAB. 3 – Matrice de répartition des séquences entre les clusters issus des deux classifications

La figure 4 propose une représentation alternative, dans laquelle les clusters produits par les k-means sont cette fois matérialisés dans le premier plan principal de l'ACP.

4.3 Analyse des consensus issus de la classification

L'analyse de cette section utilise la classification en deux catégories produite par la méthode des k-means. Les résultats exposés ne varient cependant pas lorsque l'autre classification est retenue. Les clusters identifiés, MM est à nouveau utilisé pour déterminer les consensus correspondants. On retrouve ainsi, avec la séquence `AGAC(U/A)ACG` le motif déjà identifié avant classification. L'autre séquence, `GAC(G/U)A(C/G)(G/A)A`, correspond au second motif recensée dans la littérature [4]. La stabilité des résultats est évaluée suivant le protocole décrit dans la section 3. Une sélection des consensus obtenus, 59 par cluster, apparaît dans le tableau 4.

Cluster 1	Cluster 2
AGAC(U/A)ACG (total)	**GAC(G/U)A(C/G)(G/A)A** (total)
...	...
AGACUACGC	GACGA(G/C)AGAU
...	...
UGACACCG	GACAAGAU
UACGACG	GAC(G/C)AGAG

TAB. 4 – Illustration des séquences consensus obtenues après classification

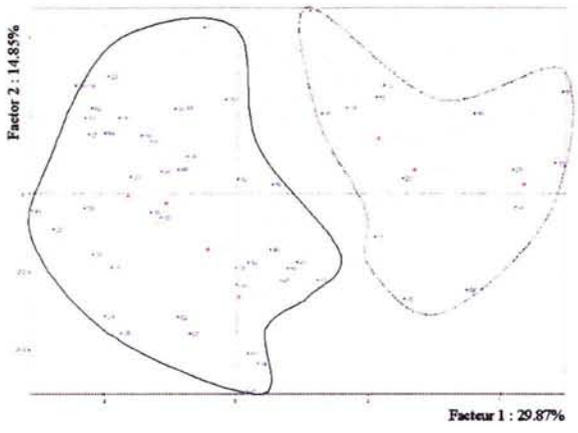

FIG. 4 – Représentation de la classification par k-means dans le premier plan principal de l'ACP

La similitude minimale entre le consensus total et l'un quelconque des consensus partiels est à présent de 1, ceci pour chacun des clusters. Cette valeur est nettement plus élevée que celle obtenue sans classification, -6 (voir la section 3.3). Du fait que les séquences consensus ont toutes approximativement la même taille, la comparaison de ces scores est valide et permet donc de déduire que les données contenues dans les clusters sont plus homogènes. Ici encore, l'expertise biologique vient confirmer les observations numériques, dans la mesure où les séquences issues d'un même cluster apparaissent fonctionnellement similaires. Ceci nous conduit à formuler l'hypothèse que la structure classificatoire déterminée possède une signification biologique et que les deux consensus totaux obtenus sont en fait les représentants de deux motifs différents. Sous cette hypothèse, en analysant les données SELEX suivant la procédure standard, on court le risque de négliger le motif le moins représenté (associé au cluster 2 de notre étude). Ce motif est en quelque sorte masqué par l'autre lorsque l'alignement multiple est effectué.

5 Validation des hypothèses effectuées

La conjecture exprimée dans la section précedente, la présence d'au moins deux motifs, est originale dans le domaine de la biologie fonctionnelle. Jusqu'alors, l'obtention de séquences consensus différentes était seulement imputée aux fluctuations expérimentales. L'hypothèse d'existence de motifs multiples sur des séquences différentes n'était pas retenue. Récemment, l'étude décrite dans [10] a permis d'établir que la protéine 9G8 possède deux domaines. Il s'agit de domaines N-terminaux permettant la fixation aux molécules d'ARN. Ces molécules d'ARN possèdent donc au moins deux types de motifs différents, les interactions multiples permettant à la protéine de moduler l'épissage alternatif et constitutif. Cette observation de nature structurale valide donc qualitativement notre conjecture. Celle-ci est encore corroborée par des observations de nature fonctionnelle. En effet, parmi les protéines SR, la protéine 9G8 est de celles qui contribuent à la modulation de l'épissage tissu-spécifique. Cette fonction multimodale suppose de même la présence de plusieurs motifs. De manière plus générale, la diversité des motifs nous apparaît comme une caractéristique biologique d'une importance centrale. La combinaison SELEX / classification se présente alors comme une solution naturelle permettant de préserver toute l'information biologique pertinente.

6 Conclusions et perspectives

Nous avons exposé une étude critique de la méthode usuelle d'exploitation des données issues de la technique expérimentale du SELEX. Cette méthode, qui soumet sans distinction les données à un alignement multiple, ne permet pas de prendre en compte certains facteurs biologiques importantes, comme la possibilité de l'existence de motifs multiples qui ne seraient pas communs à toutes les séquences. Pour résoudre ce problème, nous proposons d'effectuer sur les séquences produites une classification, et de déterminer ensuite les séquences consensus pour chaque cluster. Dans le cas de la protéine 9G8, cette stratégie nous a permis de mettre en évidence deux motifs différents, ce qui a été confirmé par une étude biologique conduite indépendamment par une autre équipe.
Ces résultats, pour encourageants qu'ils soient, appellent une confirmation qui doit venir d'expériences supplémentaires. Il semble ainsi nécessaire d'appliquer notre approche pour différents nombres de cycles, afin d'étudier l'influence de ce facteur sur le nombre de motifs identifiés. Cette étude est en cours sur un ensemble de protéines SR. La question du choix de la ou des méthodes de classification les plus adaptées est encore pour l'instant lar-

gement ouverte. Un paramètre important doit ici être pris en compte, l'incorporation d'informations biologiques supplémentaires. Ainsi, dans le cas qui nous intéresse plus particulièrement, les relations ARN-protéines peuvent se caractériser par des contraintes très fortes concernant la structure secondaire des ARN. Prendre en compte ces contraintes devrait grandement faciliter la classification. Là encore, des expériences sont en cours. Elles s'appuient sur l'emploi de techniques adaptées aux descripteurs de types différents. La solution que nous privilégions actuellement est le développement de machines à noyaux dédiées, pour mettre en œuvre des algorithmes tels que le *Support Vector Clustering* (SVC) [3].

Remerciements
Les auteurs souhaitent remercier James Stevenin pour la mise à disposition des données expérimentales. Nous exprimons également notre reconnaissance à Fabrice Leclerc pour son analyse des séquences consensus.

Références

[1] BAILEY (T.) et ELKAN (C.), « Fitting a mixture model by expectation maximization to discover motifs in biopolymer », dans PRESS (A.), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 1994, p. 28–36.

[2] BAILEY (T.) et GRIBSKOV (M.), « Methods and statistics for combining motif match scores », *Journal of Computational Biology*, *5*, 1998, p. 211–221.

[3] BEN-HUR (A.), HORN (D.), SIEGELMANN (H.) et VAPNIK (V.), « Support vector clustering », *Journal of Machine Learning Research*, *2*, 2001, p. 125–137.

[4] CAVALOC (Y.), BOURGEOIS (C. F.), KISTER (L.) et STEVENIN (J.), « The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers », *RNA*, *5*, 1999, p. 468–483.

[5] CLÉMENT (E.) et CODANI (J.-J.), « Lassap, a large scale sequence comparison package », *CABIOS*, *2*, 1997, p. 137–147.

[6] FENG (D. F.) et DOOLITLE (R. F.), « Progressive sequence alignment as a prerequisite to correct phylogenetic trees », *Journal of Molecular Evolution*, *25*, 1987, p. 351–360.

[7] HARTIGAN (J.), *Clustering Algorithms*, Wiley, N.Y., 1975.

[8] L.DURET et ABDEDDAIM (S.), « Multiple alignements for structural, functional, or phylogenetic analyses of homologous sequences », dans HIGGINS (D.) et TAYLOR (W.), *Bioinformatics : Sequence, structure and databanks*, Oxford Univ. Press, 2000, p. 51–76.

[9] LEGENDRE (P.) et LEGENDRE (L.), *Numerical Ecology, second english edition*, Elsevier, 1998.

[10] LEJEUNE (F.), CAVALOC (Y.) et STEVENIN (J.), « Alternative splicing of intron 3 of the serine/arginine-rich protein 9g8 », *Gene*, *276*, n° 11, 2001, p. 7850–7858.

[11] LEWIN (B.), *Gene VII*, Oxford University Press, 2000.

[12] LIU (H.-X.), ZHANG (M.) et KRAINER (A. R.), « Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins », *genes & development*, *12*, 1998, p. 1998–2012.

[13] NEEDLEMAN (S.) et WUNSCH (C.), « A general method applicable to the search for similarities in the amino acid sequence of two proteins », *Journal of Molecular Biology*, *48*, 1970, p. 443–453.

[14] SCHALL (T.) et MANIATIS (T.), « Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA », *Molecular and Cellular Biology*, *19*, 1999, p. 261–273.

[15] SMITH (T.) et WATERMAN (M.), « Identification of common molecular subsequences », *Journal of Molecular Biology*, *47*, n° 1, 1981, p. 195–197.

[16] SUN (F.), GALAS (D.) et WATERMAN (M.), « A mathematical analysis of in vitro molecular selection-amplification », *Journal of Molecular Biology*, *258*, 1996, p. 650–660.

[17] TACKE (R.), CHEN (Y.) et MANLEY (J.), « Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation : Creation of an SRp40-specific splicing enhancer », *Proc. Natl. Acad. Sci. USA*, *94*, 1997, p. 1148–1153.

[18] TACKE (R.) et MANLEY (J. L.), « The Human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities », *EMBO*, *14*, 1995, p. 3540–3551.

[19] TUERK (C.) et GOLD (L.), « Systematic evolution of ligands by exponential enrichment : RNA ligands to bacteriophage T4 DNA polymerase », *Science*, *249*, 1990, p. 505–510.

[20] VANT-HULL (B.), PAYANO-BAEZ (A.), DAVIS (R.) et GOLD (L.), « The mathematics of SELEX against complex targets », *Journal of Molecular Biology*, *278*, 1998, p. 579–597.

7.3 Interprétation de la répartition des séquences

L'avantage de notre approche est de permettre une analyse des résultats dédiés aux données SELEX à notre disposition. Cela se manifeste concrètement par la possibilité d'inférer sur les critères biologiques qui justifient une telle répartition des séquences en sous-ensembles qui sont homogènes.

7.3.1 Critère de la structure secondaire des acides nucléiques

La sélection des SELEX se fait par affinité pour certaines protéines. La détermination des groupes homogènes s'effectue sur des critères de distance entre les séquences nucléiques. Il est actuellement possible de déterminer expérimentalement la structure secondaire d'un acide nucléique à partir de sa séquence (structure primaire). On peut donc naturellement supposer que la structure secondaire dépend de la structure primaire. Ainsi si l'on apparente les structures primaires dans des ensembles homogènes, il peut être intéressant d'analyser si les structures secondaires sont-elles aussi homogènes dans ces mêmes ensembles.

Cette remarque est le point de départ d'une réflexion en concertation avec Antoine Cléry et Christiane Branlant du laboratoire MAEM. Nous avons, dans le but de vérifier cette propriété, analysé les données de SELEX pour les protéines L7Ae et Snu13p. Ces protéines interviennent également dans le processus d'épissage. Le laboratoire biologique à mis en oeuvre son expertise afin de déterminer la structure secondaire de chaque acide nucléique sélectionné par les expériences SELEX. Deux structures particulières apparaissent distinctement. Une structure en tige boucle de type (A) et deux autres de type (B) et (C) (voir Figure 7.1). En parallèle, nous avons appliqué notre méthodologie d'analyse sur ces mêmes données SELEX dont la structure secondaire a été identifiée. La classification divise les données SELEX en deux sous-ensembles. La répartition des structures secondaires correspondantes est représentée dans la Figure 7.2. Chacun des sous-ensembles est ainsi caractérisé par une structure spécifique dans les plans significatifs de l'ACP.

On est alors en mesure de comparer les classifications fournies par notre approche et celles déterminées expérimentalement suivant le critère de structure secondaire. Le cluster (A) est le plus significatif, puisque notre méthodologie retrouve 73% de ce cluster avec un indice de confiance de 97,5%. Les structures (B) et (C) se retrouvent de manière moins évidente si l'on considère une classification sur 3 ensembles. L'approche sur 2 ensembles semble donc correspondre à une réalité expérimentale basée sur la structure secondaire des ligands. La classification en deux ensembles peut alors être interprétée comme une discrimination de la structure (A) par rapport aux autres.

Les structures secondaires sont des caractéristiques importantes qui peuvent être un critère de sélection des SELEX. Or pour une structure donnée, il existe beaucoup de séquences nucléiques possible. Cette relation pourrait expliquer les ensembles de séquences hétérogènes pour certaines protéines dont les critères d'affinité reposent principalement sur la structure. Cette hypothèse peut être valider expérimentalement si l'on détermine les structures des séquences ayant une affinité pour les protéines SR. Une perspective expérimentale des SELEX serait alors de justifier l'arrêt des sélections lorsque toutes les séquences sélectionnées possèdent des structures secondaires identiques. On pourra alors

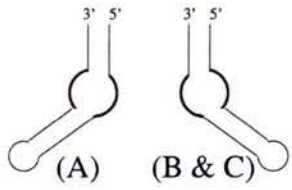

FIG. 7.1 – **Structures secondaires des ARN déterminées expérimentalement**. Ces structures sont celles sélectionnées par les expériences SELEX pour se fixer à la protéine L7Ae. On distinguera deux types de structures secondaires distincts (A) et (B & C) déterminés expérimentalement par Antoine Cléry et Christiane Branlant du laboratoire MAEM.

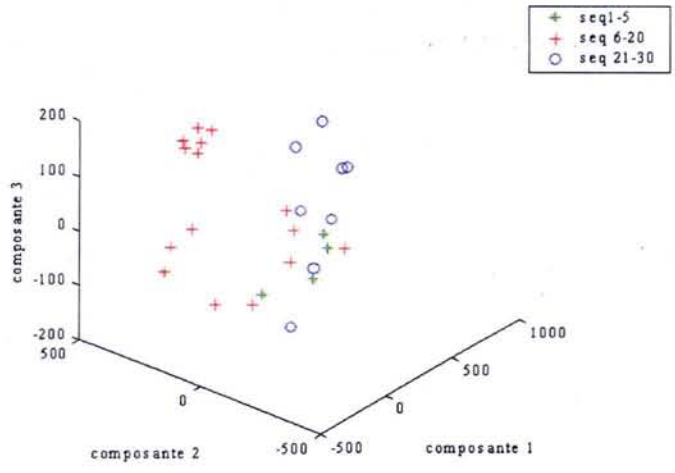

FIG. 7.2 – **Résultat des données SELEX pour la protéine L7Ae avec les structures secondaires associées**. On représente les séquences dans l'espace des 3 premières composantes de l'ACP avec les structures secondaires associées (A, B ou C).

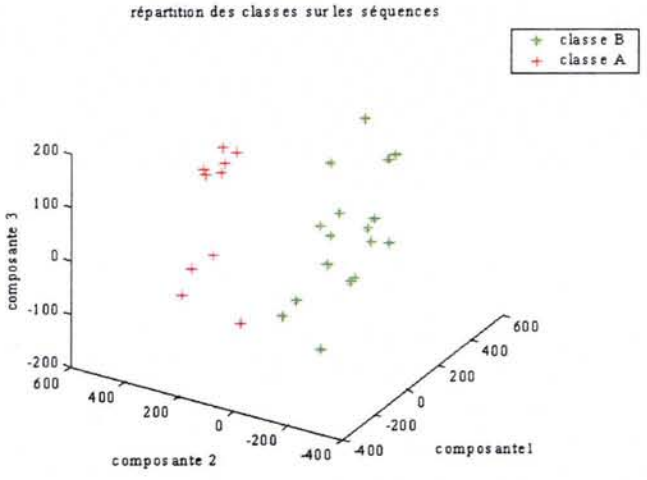

FIG. 7.3 – **Résultat de classification des données SELEX pour la protéine L7Ae.** On effectue la démarche de classification pour 2 ensembles. On représente les séquences associées à leurs ensembles dans les 3 premières composantes de l'ACP.

considérer dans ce cas que le critère d'affinité sélectionné par le SELEX est optimal malgré des séquences divergentes.

7.3.2 Importance des séquences de faibles affinités

Les expériences de SELEX sélectionnent les ARN en fonction de leurs affinités absolues. Seul les plus fortes affinités sont ainsi conservées. Néanmoins, comme nous l'avons vu dans le chapitre précédant, la répartition des affinités est de type *log-normale*. Cela signifie qu'il existe un nombre majoritaire de séquences dans les résultats SELEX qui possèdent une affinité *moyenne*. Ces séquences doivent logiquement être éliminées au cours des cycles de SELEX, tout comme les séquences de faibles affinités pour la protéine cible dans les premiers cycles de sélection.

Il est néanmoins possible que ces séquences passent les sélections si elles sont en nombre suffisamment important. Elles peuvent en effet saturer les sites des protéines et ainsi favoriser leur sélection. On peut donc retrouver après les expériences SELEX, des séquences regroupées en groupes homogènes qui résultent de ce processus de sélection. Cette remarque méthodologique n'est pas forcément à interpréter comme un biais expérimental dans notre contexte biologique. Il apparaît en effet que les séquences de forte affinité ne sont pas forcément les plus efficaces *in vivo*. Dans un système biologique, l'affinité n'est en effet pas le critère prédominant face à la quantité de séquences nucléiques en présence. On peut en effet illustrer ces propos avec le génome de HIV-1 qui possède beaucoup de sites potentiels de fixation de protéines régulatrices d'épissage proche dans l'espace des sites d'épissage. Il est alors possible que des séquences de faible affinité soient importantes si elles sont bien entourées. Le SELEX peut donc être, si les expériences n'utilisent pas

trop de cycles de sélection, une approche unique pour appréhender ces petites séquences potentiellement essentielles à la mécanique de la régulation de l'épissage.

7.4 Conclusions

Une approche par classification permet d'isoler différents sous-ensembles homogènes sur lesquels les méthodes standard sont cette fois-ci efficaces. Notre démarche peut donc être assimilée à un protocole de pré-alignement afin d'en certifier statistiquement le résultat. Au-delà de la méthodologie, notre approche permet de raisonner sur les données pour mieux comprendre l'inférence biologique qui est derrière les résultats de SELEX. Les classes que nous avons mises en évidence après classification possèdent donc une inférence biologique qu'il reste encore à valider. La structure secondaire des acides nucléiques est dans ce cas une piste intéressante. Rechercher ce critère de structure permettra de mieux caractériser les sites de fixation des protéines SR. Néanmoins, les outils actuels de prédiction de structure secondaire des ARN ne sont pas suffisamment performants pour permettre à un outil de recherche de motifs biologiques de se baser uniquement sur ces résultats sans biaiser leurs recherches [Zuker, 2003].

Identifier les propriétés biologiques que la classification met en évidence, peut se baser sur les modèles de discrimination. En effet, la classification étant probante, les données SELEX se composent de classes que l'on peut maintenant rechercher efficacement. Le modèle de discrimination devra alors être en mesure de pouvoir intégrer des données de nature différentes afin de pouvoir à long terme, intégrer les résultats et les hypothèses biologiques que notre analyse classificatoire met en exergue.

Troisième partie

Recherche de motifs : Application à la recherche de motifs de régulation d'épissage

Si la classification est nécessaire pour mieux appréhender un système biologique et les facteurs qui le contrôlent, elle n'est néanmoins pas suffisante. En effet, l'information que l'on obtient est trop souvent restreinte aux données à notre disposition. C'est pour cette raison qu'après identification et analyse des facteurs biologiques, il est important de pouvoir les retrouver dans un génome pour en généraliser leurs propriétés. La recherche des motifs que nous avons précédemment isolés apparaît ainsi comme une démarche naturelle.

La recherche de motifs est un axe central en bioinformatique. Le séquençage automatique des génomes a entraîné une augmentation sans précédent de la quantité de données expérimentales disponibles. Si, dans un premier temps, l'utilisation de méthodes informatiques fut nécessaire à l'agencement des nouvelles données au sein de bases d'informations structurées, les biologistes furent confrontés dans un deuxième temps à la nécessité d'automatiser des protocoles de recherches qui étaient jusqu'alors considérés comme empiriques, tel que la recherche de motifs biologiques tenant compte des substitutions.

Dans notre étude, la recherche consiste à isoler les motifs de régulation de l'épissage alternatif dans un génome intéressant comme peut l'être celui du virus HIV-1. Précédemment, nous avons mentionné le contrôle de la synthèse des protéines virales de HIV-1 par des facteurs de régulation de l'épissage. La recherche des motifs isolés par classification est à nos yeux une justification suffisante pour nous immerger dans un domaine d'application bioinformatique extrêmement bien documenté car historiquement pionnier. Parmi la diversité des méthodologies existantes, nous nous contenterons dans cette partie de thèse d'isoler les approches nécessaires à notre problème. Un premier Chapitre 8 résumera donc les démarches méthodologiques possibles qui seront ensuite appliquées à notre problème dans le Chapitre 9. C'est dans ce dernier chapitre que nous développerons et appliquerons une démarche originale qui nous semble actuellement pertinente pour rechercher les motifs de régulation de l'épissage alternatif. Cette démarche sera conceptualisée par le développement d'un logiciel destiné aux biologistes.

Chapitre 8

Approches actuelles des outils de recherche de motifs

L'inférence biologique est directement issue des expériences biologiques. Elle représente la perception ponctuelle du scientifique d'un système vivant. Un des principaux enjeux de la Biologie réside dans la capacité à généraliser cette inférence. La synthèse de cette généralisation réside dans la formalisation d'un motif biologique qui prendra des formes différentes comme une séquence pondérée, une matrice de poids ou un score statistique. La recherche de ces motifs permettrait d'étendre les connaissances acquises localement sur un système biologique générique à un génome dans sa totalité. La bioinformatique propose différents modèles statistiques ou algorithmiques pour automatiser cette recherche. Dans cette optique de modélisation, il faut donc des outils informatiques spécifiques différents de ceux de classification et d'analyse présentés dans le chapitre précédent.

La recherche de motifs biologiques apparaît comme un problème fédérateur de nombreux domaines informatiques producteurs de diverses méthodologies. Parmi ces domaines, on peut en distinguer deux qui se sont chacun focalisés sur un type de motif particulier à rechercher. La recherche de formes de motifs hétérogènes a été facilement appropriée par les statisticiens par opposition à la recherche de motifs conservés associée aux approches algorithmiques. Ces deux domaines informatiques sont vastes et nous nous restreindrons dans ce chapitre aux méthodes de recherche qui nous semblent cohérentes avec le problème de recherche de sites de fixation de protéines régulatrices d'épissage.

8.1 Méthodes algorithmiques

Certaines protéines doivent gérer des contraintes par leurs interactions avec les acides nucléiques, de sorte que leurs motifs de fixation sont particulièrement bien conservés. Ces protéines se fixent donc systématiquement sur des motifs relativement bien identifiés expérimentalement. Pour illustration, on peut mentionner parmi ces protéines, celles possédant une forte affinité pour une structure secondaire spécifique. Les autres motifs ne satisfaisant pas ces conditions de structure donc de moindre affinité, seront par conséquent évincées de la liste des motifs potentiels. Dans ce cas, on peut citer les protéines SRp40 et ASF/SF2 qui sont relativement bien documentées par la littérature car se fixant cha-

cune sur un motif conservé qui sera plus accessible expérimentalement. Dans ce contexte biologique, la recherche de motifs dans un génome correspond à rechercher un mot plus ou moins exact dans une phrase. Sous cette hypothèse, les méthodes algorithmiques de recherche de mots s'appliquent bien à la recherche de motifs discrets. Les outils d'analyse de texte (voir pour revue [Baeza-Yates & Navarro, 2004]) sont par conséquent une source de méthodologies intéressantes.

8.1.1 Approches existantes

De nombreuses méthodes de recherche de mots sont appliquées à la recherche de motifs biologiques. Elles se basent sur des fondements théoriques différents. Les méthodes généralement utilisées en bioinformatique considèrent une recherche s'effectuant sur un texte non indexé (sans balises particulières). Dans ce contexte algorithmique, une approche standard est le *Backward DAWG matching* (ou BDM) [Crochemore *et al.*, 1994]. Cet algorithme est encore aujourd'hui considéré comme un des plus efficaces pour rechercher des motifs de longueur $m \approx 100$ dans des bases d'ADN. Néanmoins cette approche très rapide ne permet pas de recherches plus complexes comme celles de motifs qui sont variants. Pour combler ce biais, il existe deux approches qui sont à l'origine des principaux progrès du domaine.

La première approche utilise un automate qui permet une recherche déterministe des motifs qui ne le sont pas forcément (les substitutions de nucléotides non déterminées dans un motif rendent la recherche non déterministe). On utilise pour cela des automates tantôt non déterministe comme NFA pour *Non deterministic Finite Automaton* ou déterministes comme DFA pour *Deterministic Finite Automaton* qui ont montré en pratique leur efficacité.

La deuxième approche utilise un procédé de filtration. En effet, les approches par automates sont gourmandes en temps et en espace de calcul. Ils inspectent chaque caractère du texte ce qui peut être peu efficace pour un grand motif. Par économie de temps de calcul, la recherche par filtration n'inspecte pas tous les nucléotides mais filtre le texte à analyser comme le préconise l'approche de [Boyer & Moore, 1977] ou algorithme de BM. Pour cela, la base est découpée par une fenêtre glissante dans laquelle est recherché le motif. On associe un score à chaque fenêtre qui représente le nombre de motifs présents dans celle-ci. Par cette approche, l'algorithme filtre les zones de la base de recherche qui sont intéressantes.

Certains algorithmes actuels utilisent une combinaison des deux précédentes approches. Ainsi, la filtration est effectuée par un automate. Il également possible d'utiliser des automates plus complexes qui gèrent la recherche de motifs multiples ou complexes avec des inconnus. Cette dernière possibilité est particulièrement intéressante en biologie lorsque les motifs sont mal définis.

8.1.2 `grappe` une approche discrète pour la biologie

`grappe` est une approche algorithmique de *pattern matching* proche de l'algorithme BDM mentionné précédemment. Cet outil, développé par [Kucherov & Rusinowitch, 1997], permet de rechercher dans un texte un motif ou

pattern dont certains symboles possèdent des incertitudes qui seront définies. On utilise alors la notion de *wildcard* ou de joker. La longueur du motif peut être alors illimitée. Le motif n'est donc pas borné, ce qui différencie `grappe` de la recherchede motifs par expressions régulières. Par ces attributs, `grappe` est particulièrement efficace pour rechercher un grand nombre de motifs, chacun contenant des incertitudes comme des substitutions ou des motifs en deux parties. Par cet aspect, `grappe` permet la recherche des motifs biologiques plus hétérogènes que ceux habituellement recherchés par le *pattern matching* classique comme la recherche par expressions régulières. Malgré cette complexité dans la description du motif, `grappe` reste aussi performant que les autres outils tels que `egrep` et `agrep` sur la recherche de petits motifs. Mais le plus grand avantage de ce logiciel est de pouvoir définir ergonomiquement des substitutions sur certains motifs comme peuvent l'illustrer les options suivantes :

- `ATCG` ; recherche le motif `ATCG`
- `AT#AT` ; recherche un motif commençant par `AT` suivie d'une distance arbitraire qui le sépare de la deuxième partie du motif `AT`
- `AT(1,5)AT` ; recherche un motif contenant l'occurrence `AT` suivie dans une distance comprise entre 1 et 5 nucléotides de l'occurrence `AT`
- `A[TC]G` ; recherche un motif `ATG` ou `ACG`
- `ATAA|ATCGC` ; recherche les motifs `ATAA` ou `ATCGC`

En collaboration avec Abdelhalim Larhlimi, nous avons modifié cet algorithme pour qu'il donne dans un fichier de sorties toutes les positions d'un génome pour lesquels il existe un motif. Par ces diverses options, `grappe` est un outil efficace pour rechercher des motifs homogènes et ce malgré des incertitudes biologiques concernant des substitutions ou des motifs en deux parties homogènes. L'outil est donc plus flexible que les méthodes standard de *pattern matching* existantes et devient donc un bon candidat pour rechercher les motifs biologiques de fixation des protéines SR.

8.2 Méthodes d'apprentissage statistique

Les méthodes d'apprentissage statistique diffèrent fondamentalement de l'approche algorithmique. Les données biologiques à disposition ne sont alors pas les mêmes que celles nécessaires au bon déroulement d'une approche de *pattern matching*. Comme nous l'avons déjà mentionné dans le Chapitre 5, l'apprentissage statistique ne permet pas l'utilisation de motifs conservés. La méthodologie est en effet performante s'il subsiste une certaine incertitude dans la définition des motifs biologiques à rechercher. Il convient dans ce cadre d'appliquer les méthodologies statistiques sur des motifs hétérogènes qui constitueront une base d'apprentissage. Quelle que soit la méthode statistique que nous serons amenés à utiliser, le protocole expérimental sera similaire à celui représenté dans la Figure 8.1. L'apprentissage statistique, comme dans l'emploi d'une SVM, consiste à identifier les paramètres d'une machine à vecteurs support en fonction des données de la base d'apprentissage. Une fois ceux-ci identifiés, la machine prédira statistiquement un résultat en fonction de l'information biologique sur laquelle l'apprentissage statistique a été effectué. Par cette démarche d'apprentissage, on inocule donc une inférence biologique plus ou moins précise par le biais de la base d'apprentissage. On suppose alors que pour

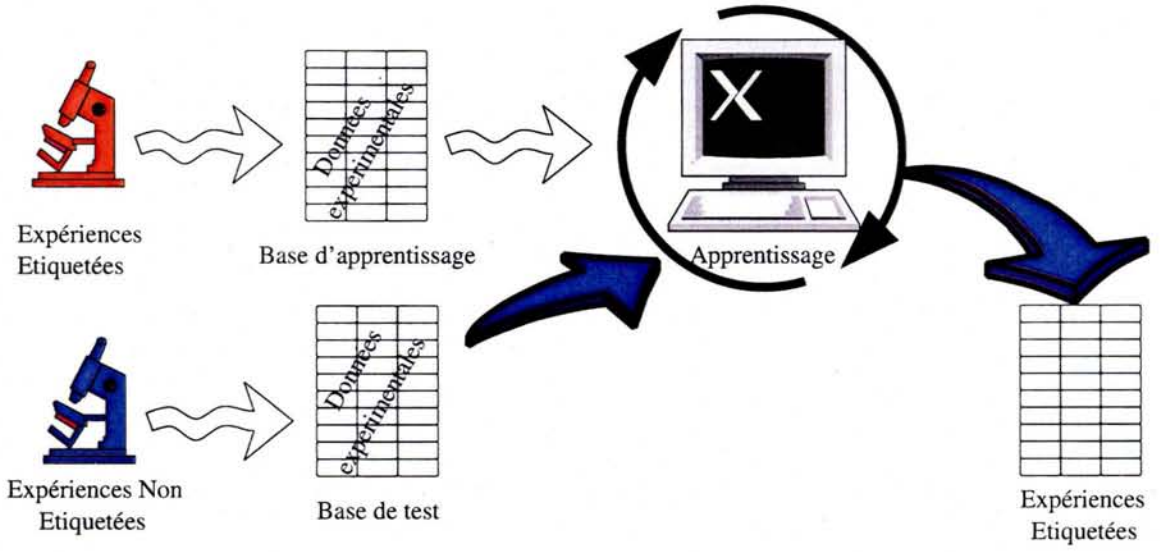

FIG. 8.1 – **Protocole de la mise en œuvre de l'apprentissage** L'apprentissage nécessite des résultats biologiques (rouges) qui doivent être classés ou étiquetés et représentés dans une matrice. Ces résultats forment la base de connaissances qui servira dans cette méthodologie comme une base d'apprentissage. C'est l'information qui est dans cette base que la machine va généraliser. Une itération de calcul permet une optimisation des paramètres de la machine. Au fur et à mesure des itérations, la machine est spécifiée pour le problème posé par les échantillons rouges. Une fois les paramètres identifiés et fiables statistiquement, il est possible de classer de nouveaux échantillons (bleus) qui ne sont pas étiquetés au préalable. Avec un certain seuil statistique, les nouvelles données bleues seront discriminées grâce à l'inférence biologique issue des résultats des échantillons rouges.

un apprentissage optimal, l'inférence biologique inoculée sera généralisée automatiquement sur des données expérimentales indéterminées. Il convient alors de tester chaques paramètres qui permettront de donner des résultats de simulations concordants avec les données d'apprentissage grâce à de nombreuses itérations pour certifier l'apprentissage optimal.

Les diverses méthodes d'apprentissage reposent sur des fondamentaux théoriques différents. Certaines considèrent un automate dont les paramètres de transition permettent de reproduire des objets probabilistes comme des séquences nucléiques. C'est l'approche probabiliste. D'autres méthodes reposent sur un support d'information vectoriel pour discriminer l'information dans un espace. C'est l'approche statistique issue du domaine connexionniste que nous avons précédemment cité dans le Chapitre 5. Dans ces deux cadres théoriques différents, les théoriciens manipulent des outils considérés comme appartenant à la reconnaissance des formes ou (*pattern recognition*) par opposition à l'approche algorithmique qui effectue du *pattern matching*. La différence se situe dans la connaissance du motif que l'on recherche (motif bien identifié dans le *pattern matching* et moins bien, voir pas identifié dans le *pattern recognition*). En fonction des méthodes employées, les paramètres à identifier seront différents ainsi que leur importance dans la méthodologie statistique. Certaines méthodes d'apprentissage statistiques ont montré leurs avantages

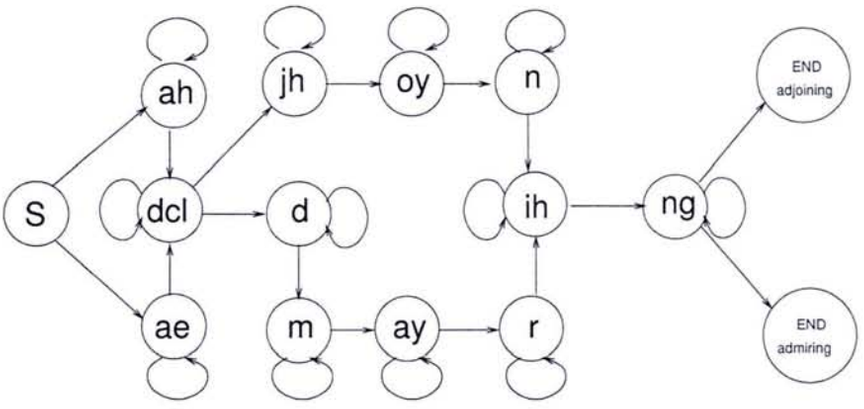

FIG. 8.2 – **Exemple d'automate HMM.** L'automate HMM est emprunté au traitement de la parole. Il permet de reconnaître des phonèmes qui sont les objets élémentaires de la phonétique. Les états de l'automate se représentent par des cercles et les transitions par des flèches entre les états. Cet automate reconnaît pour une entrée phonétique dans l'état initial *S*, les sorties possibles *adjoining* et *admiring* avec une certaine probabilité pour chaque état. Le processus de reconnaissance nécessite de passer parmi les différents états de l'automate avec une certaine probabilité avant de sortir par les deux états finaux. L'entraînement du HMM permet de déterminer les probabilités de transition entre les états. Ainsi pour une parole donnée, l'automate donne le mot correspondant avec une probabilité.

dans la recherche de motifs biologiques. Par conséquent, elles sont des candidates intéressantes pour rechercher les motifs de régulation d'épissage.

8.2.1 Utilisation des modèles de Markov cachés

Les modèles de Markov cachés ou HMM pour *Hidden Markov Model* ont révolutionné les méthodologies de recherche de motifs biologiques [Baldi & Brunak, 2001]. Les HMM et leurs dérivées sont initialement issues du traitement de la parole. Le principe de cette méthodologie est résumé par la Figure 8.2. Les paramètres majeurs de la machine sont les probabilités de transition entre les états de l'automate HMM. Ceux-ci nécessitent d'être initialisés correctement pour être ensuite raffinés avec l'apprentissage.

Les états des HMM dédiés à la biologie sont généralement des nucléotides ou acides aminés (ou également des paires / triplets). Les paramètres sont alors des probabilités d'avoir un nucléotide après un autre dans une séquence biologique. Ces transitions probabilistes permettent de décrire un motif biologique hétérogène. Pour illustration, il est possible dans un HMM d'accentuer la présence d'une adénosine `A` après une tyrosine `T` avec une probabilité forte de 0.80 pour la transition `T-A`. Ce paramètre permettra notamment de dédier une machine à la reconnaissance d'une boite `TATA`. Par ces probabilités, les HMM désignent un motif par des critères statistiques qui composent la structure séquentielle du motif. Elles permettent ainsi une recherche de motifs hétérogènes contrairement au *pattern matching* standard.

Initialiser et identifier les paramètres des HMM est difficile, ce qui rend l'utilisation

de ces automates délicate sur des problèmes biologiques nécessitant de la précision. Néanmoins les travaux de [Roulet *et al.*, 2002] qui dédient un HMM à la recherche de sites de fixation de facteurs de transcription sont encourageants. Ils paramétrisent pour cela un HMM avec des résultats expérimentaux dédiés pour définir les séquences d'ADN spécifiques aux facteurs de transcription. Ils génèrent ainsi un modèle statistique quantitatif où les paramètres d'un HMM sont estimés avec la sélection *in vitro* des expériences SELEX. Les données SELEX sont dans ce cas considérées comme une base d'apprentissage. Cela permet au HMM d'utiliser des transitions suffisamment précises pour appréhender les sites de transcription qui possèdent des taux de fixation protéique de faibles, moyennes, et fortes affinités. Ce sont ainsi les expériences SELEX spécifiques qui dédient le HMM à un problème précis. Les expériences SELEX qui jusqu'alors était considérées comme une simple source de base d'alignement deviennent ici un protocole *pré-computationnel* pour orienter une méthodologie *in silico*. Les progrès de cette méthodologie hybride illustrent alors l'efficacité d'une concertation pluridisciplinaire pour résoudre un problème biologique. L'utilisation d'un HMM semble donc se prêter à la recherche de motifs biologiques complexes comme ceux qui fixent les protéines SR. [Cartegni *et al.*, 2003] ont notamment utilisé cette approche dans *ESEfinder* afin de retrouver les sites qui fixent les protéines activatrices d'épissage.

Néanmoins, l'emploi de ce type de HMM restreint cette méthodologie à la gestion d'une information uniquement localisée sur la séquence biologique qui est un vecteur spécifique. Dans ce contexte, il est également nécessaire de connaître a priori la composition du motif biologique que l'on recherche. Cette composition va influer sur le type d'automate et ses conditions initiales. Or cette information n'est pas toujours facilement identifiable sans a priori. Par ailleurs, l'information n'est pas toujours uniquement séquence dépendante notamment avec une structure secondaire des acides nucléiques caractéristique. Une méthodologie sans a priori et permettant une incorporation de la structure secondaire ou un autre type d'information apparaît donc comme nécessaire pour pallier ce biais, et ce notamment pour rechercher les motifs de régulation d'épissage.

8.2.2 Utilisation de SVM

Pour combler les lacunes des HMM qui nécessitent une initialisation a priori, une alternative réside dans les machines à noyaux dont les fondamentaux théoriques ont déjà été introduits dans le Chapitre 5. De récents travaux ont adapté les méthodes à noyaux aux séquences biologiques. Les adaptations des méthodes génériques à la biologie nécessitent le développement de fonctions noyaux ϕ spécifiques aux séquences protéiques comme le *convolution kernel* de [Haussler, 1999] ou le *Fisher kernel* de [Jaakkola *et al.*, 1999]. Les méthodes standard sont également rapidement optimisées comme la SVM qui est étendue au cas multi-classe [Guermeur, 2002]. Cette extension fait notamment l'objet d'applications à la prédiction de la structure secondaire des protéines dont l'efficacité est inégalée jusqu'à présent. Les SVM sont ainsi compétitives dans de nombreux domaines de la bioinformatique jusqu'à permettre des alignements locaux plus pertinents biologiquement que l'alignement local référence de Smith-Waterman [Saigo *et al.*, 2004].

Dans ce contexte de réussite, les SVM dédiées à la recherche de motifs biologiques apparaissent comme une approche naturelle. [Zhang *et al.*, 2003] utilisent notamment des

SVM dans le but de bioanalyser les sites d'épissage. Ils dissocient ainsi les vrai exons des pseudo-exons en décryptant l'information contenue dans la séquence. Une analyse a *posteriori* des résultats de la machine d'apprentissage permet d'isoler les informations biologiques utiles à la définition d'un vrai site d'épissage. Ainsi, la présence de boites de branchement, des motifs riches en `C` et en `TG` dans la région de 50 nucléotides en amont d'un site et de séquences riches en `C` ainsi que des triplets de `G` dans une région de 80 nucléotides en aval sont nécessaires à la définition. Ces travaux illustrent le potentiel d'analyse des SVM pour mieux comprendre un système biologique. Ce type d'analyse peut être également élargi si l'on considère que le support de l'information est vectoriel au sens large contrairement aux HMM qui se concentrent sur les séquences biologiques. La notion de vecteur permet d'incorporer des informations supplémentaires autre que la séquence seule qui sera alors considérée comme un sous-vecteur. [Sun *et al.*, 2003] utilisent notamment cette propriété pour introduire des données de structures secondaires des ARN afin de mieux différentier les faux positifs des vrais sites d'épissage. Si cette démarche ne permet pas à la machine d'être plus efficace dans la prédiction, elle introduit l'utilisation d'informations biologiques hétérogènes pour spécifier un problème biologique. La machine est alors moins stringente par rapport à prédiction sur la séquence seule.

8.3 Conclusions

Si l'efficacité des outils de *pattern matching* est historiquement établie, les outils de *pattern recognition* sont prometteurs. Ils sont adaptés à la recherche de motifs inconnus ou mal identifiés. Les machines à noyaux ouvrent également des perspectives de bioanalyses qui peuvent aboutir à une meilleure définition des facteurs biologiques qui sont nécessaires à la discrimination naturelle. Ils permettent notamment l'incorporation d'informations autres que la séquence comme la structure secondaire ou des critères d'affinité pour une protéine donnée. Les méthodes statistiques de recherche de motifs peuvent être ainsi configurées spécifiquement à un problème donné notamment avec les expériences SELEX que nous avons à disposition. Ces outils de *pattern recognition* dédiés à la recherche d'un motif biologique sont des approches non négligeables pour identifier les régions fonctionnelles d'un génome.

Néanmoins, les outils, qu'ils appartiennent au *pattern matching* ou au *pattern recognition*, possèdent chacun leurs spécificités d'application dans la recherche de motifs biologiques, en fonction de la définition plus ou moins précise du motif. Dans un cadre biologique, identifier la méthode adéquate pour résoudre un problème reste difficile si celui-ci est mal spécifié. Or la recherche de motifs de régulation d'épissage n'est pas un problème que l'on peut spécifier *a priori*. Certains motifs sont en effet hétérogènes ou mal identifiés, alors que d'autres sont considérés comme fortement homogènes. Face à la diversité méthodologique, une approche intégrative semble naturelle. Elle permettrait idéalement d'utiliser les approches algorithmiques pour certains motifs et une démarche statistique pour des motifs moins bien spécifiés ou hétérogènes. L'objectif d'une telle démarche est de pouvoir aborder globalement la recherche des motifs de régulation d'épissage avec un outil unificateur des méthodologies tout en restant accessible aux biologistes.

Chapitre 9

Localiser les sites de régulation : Développement et application de KOALAB

Rechercher les motifs de régulation reste un problème ouvert en bioinformatique. Le Chapitre 8 a identifié la diversité d'approches possibles pour rechercher les différentes formalisation de motifs biologiques. Il transparaît du chapitre précédant que ces méthodes sont très spécifiques à un type de motifs donné. C'est une des principales sources de motivation pour initier une démarche *in silico* propre à la recherche des motifs de fixation des protéines SR en se basant sur des résultats théoriques existants. Nous avons préalablement étudié les motifs de ces protéines régulant l'épissage dans le Chapitre 7 et ce à partir des données SELEX à notre disposition. Les motifs apparaissent soit comme hétérogènes étant alors mieux formalisé par un motif statistique, soit conservés étant mieux formalisé par un motif lexical et ce en fonction des protéines de régulation considérées. Cette ambivalence justifie une méthode originale qui intègre des méthodologies bioinformatiques existantes de *pattern matching* et de *pattern recognition*. Nous développerons dans ce chapitre notre approche d'un point de vue méthodologique dans un premier temps, et d'un point de vue plus pratique avec le logiciel KOALAB dans un deuxième temps. En dernier lieu, les résultats obtenus seront commentés et discutés.

9.1 Une approche intégrée pour rechercher les motifs de régulation

Face à la complexité de la recherche des motifs qui sont plus ou moins conservés, une première approche pour rechercher des motifs biologiques est de s'inspirer des méthodes déjà existantes sur des problèmes similaires. La méthode qui s'apparente le plus à notre problème est celle décrite dans les travaux de [Roulet *et al.*, 2002]. Elle consiste à utiliser une méthode d'apprentissage dédiée à un problème biologique précis. Pour cela, les auteurs entraînent un modèle de Markov spécifique à la reconnaissance des sites de fixation des facteurs de transcription grâce à des données SELEX suivant l'hypothèse que le motif biologique est présent dans la base expérimentale. Néanmoins, cette méthode qui ne

nécessite pas d'une définition précise de motif probabiliste, ne permet pas de filtrer les faux positifs des vrais résultats. Dans le cadre de la recherche des motifs de fixation des protéines SR, il est envisageable de mettre en place une approche basée également sur un apprentissage statistique mais qui serait également validée par d'autres approches de recherche de motifs comme les méthodes algorithmiques de *pattern matching*.

Notre démarche de recherche des motifs doit donc intégrer différentes méthodes. Ce point méthodologique est corroboré par les connaissances que l'on possède des motifs des protéines que l'on recherche. En effet, certaines protéines SR sont extrêmement bien étudiées comme les protéines SC35 ou ASF/SF2 [Tacke & Manley, 1995*a*] avec un motif lexical bien défini, alors que d'autres sont à un stade de connaissances moins génériques, comme dans le cas des protéines 9G8 [Cavaloc *et al.*, 1999*a*] pour ne citer qu'elles. Dans ce contexte, les approches algorithmiques standards sont très performantes. C'est le cas lors de la recherche de motifs relativement bien connus et documentés comme peuvent l'être certains motifs consensus tels que ceux qui fixent les protéines SC35 et ASF/SF2, ou dans un cadre plus général les motifs qui délimitent les sites d'épissage. Par ailleurs les approches statistiques qui couvrent un autre domaine théorique, sont également performantes dans le contexte exploratoire de la recherche de motifs statistiques.

Après concertation avec Christiane Branlant du laboratoire MAEM, les motifs reconnus par les protéines SR peuvent être de type lexical ou statistique en fonction des protéines considérées. Pour rechercher les motifs de régulation d'épissage, une démarche naturelle consiste alors à intégrer les deux approches : algorithmiques d'une part et d'apprentissage statistique d'autre part. En effet en couplant ces deux approches théoriques, elles pourront se compléter ou se compenser le cas échéant. Néanmoins intégrer des outils issus de domaines théoriques différents reste difficile pour les biologistes non initiés aux méthodes informatiques. C'est une des raisons pour lesquelles il est important de développer un outil qui intègre les deux approches, et ce par le biais d'une interface graphique permettant de piloter l'outil de manière efficace sans expertise préalable. Cette démarche est celle utilisée dans l'outil graphique KOALAB signifiant *KOupled Algorithmic and Learning Approach for Biologists*. Ce logiciel est développé en collaboration avec Abdelhalim Larhlimi et déposé à l'APP sous licence GPL. Il utilise notamment les technologies web pour proposer une certaine ergonomie procurée à l'utilisateur afin d'exécuter les outils des recherches algorithmiques de motifs et des apprentissages statistiques efficaces sur un problème donné. Cette technologie permet également l'exécution de KOALAB sur un serveur distant quel que soit le système d'exploitation du biologiste utilisateur. Le logiciel d'approche intégrée a fait l'objet d'une soumission à la conférence *JOBIM'04* [Eveillard *et al.*, *a*]. Le communiqué met en avant les méthodes de *pattern matching* et de *pattern recognition* qui sont intégrées dans KOALAB. Les résultats du logiciel sont ensuite illustrés sur le problème de recherche de motifs de fixation de protéines SC35 et 9G8 sur le génome de HIV-1.

KOALAB: A new method for regulatory motif search. Illustration on alternative splicing regulation in HIV-1

Damien Eveillard*‡, Abdelhalim Larhlimi*, Delphine Ropers° ,
Stéphanie Billaut*, Sandrine Peyrefitte*‡

* LORIA, Université Henri Poincaré, BP 239, 54506 Vandœuvre-lès-Nancy, France

‡ Laboratoire de Maturation des ARN et Enzymologie Moléculaire, UMR 7567 CNRS-UHP, BP 239, 54506 Vandœuvre-lès-Nancy, France

° INRIA-Rhone-Alpes, 655 avenue de l'Europe, 38334 Montbonnot, France

Abstract

Discovering heterogeneous regulatory motifs remains a difficult problem in biological sequence analysis. In this context, statistical learning or pattern search techniques on their own have shown some limitations. However, significant benefits can be taken from their complementarity. We selected two state-of-the-art methods: a multi-class support vector machine (M-SVM) from the statistical learning domain associated with a performant discrete pattern matching algorithm `grappe`, and integrated them into a web technology based graphical software: KOALAB (KOupled Algorithmic and Learning Approach for Biology)[1]. We applied our method on motif discovery within nucleic acid sequences using experimental SELEX results as training database for the M-SVM. An application dealing with the search for splicing regulatory protein binding sites in HIV-1 genome shows the potential of such an approach.

Keywords: motif discovery, multi-class SVM, SELEX, alternative splicing regulation

1 Introduction

Motif discovery within biological sequences is a key area of bioinformatics. Several biological questions have to deal with short sequences mediating interactions between macromolecules such as nucleic acids and/or proteins. Those interactions have been shown to play crucial roles in numerous phenomena. Because of their rather small size, they are difficult to deal with using standard alignment or `Blast` approaches. Pattern search algorithms are remarkably efficient for conserved, fixed motifs. Unfortunately, a growing number of motifs a biologist has to look for do not correspond to that requirement, notably because of intrinsic heterogeneity of the sequences or technical constraints he has to deal with. Indeed, experimental results for biological motif definition will often come as a collection of potential motifs. A common strategy is hence to apply different alignments to extract a global consensus motif from the collection. From a theoretical point

[1] KOALAB 1.0 is freely available at `http://www.loria.fr/equipes/modbio/KOALAB.html`

of view, very often, this type of approach is not satisfactory because of the heterogeneity of the collection so that the deduced consensus does not bear biological reality anymore. Improvements in the efficiency of the methods available and/or alternative computational methods are hence needed to perform the task.

An interesting strategy to define nucleic acid-protein interaction motifs comes from combinatorial chemistry and is named SELEX for Systematic Evolution of Ligands by Exponential Enrichment [21]. The process consists in the selection, among nucleic acid sequences generated randomly, of sequences that bind by affinity to a protein ligand. The result of a SELEX experiment is a collection of less than 100 sequences from which it is, at best, difficult to retrieve any global consensus. In order to take into account the sequence heterogeneity among such a motif collection of biological significance, Roulet et al [19] proposed to use the entire collection as a training database to define Hidden Markov Models (HMM) parameters. A limiting step of this method could be that the model-building process requires sequence alignments as an intermediate step, which can be problematic with small sequences. From a theoretical point of view it means that, before starting the learning process, one has to perform some prior treatment on the data (the alignment) and this is typically not always satisfactory, notably when no biological knowledge can be used to improve the information from the collection.

We propose here to make the most out of a SELEX collection respecting their heterogeneity by using a statistical learning tool: Multiclass Support Vector Machines (M-SVM). These machine learning tools are very useful for discriminant analyses. The multiclass extension of the original binary classifier offers the opportunity, in our context, to take into account, in a single experiment, data corresponding to several proteins. Even if showing some limitations, global consensus interaction sequences have already been independently defined for numerous biological questions, it hence appears interesting to be able to confront them with the results of our method. Furthermore, some other interaction motifs are very conserved and show no need to use statistical methods. In this case, an algorithmic pattern search is the most suitable.

Statistical learning as well as algorithmic methods appear as different techniques that will be more appropriate for different problems, depending on the heterogeneity of the researched signal. In order to be able to deal with any situation and in the way of taking the best benefits from the complementarity of both methods, we propose in this paper an integrated tool that combines them. In order to make it available to the community as a user friendly tool, we thus designed a web technology based graphical software: KOALAB (KOupled Algorithmic and Learning Approach for Biology).

The goal of the present paper is to introduce a new integrated tool for motif research, KOALAB, and to validate it on the complex problem of alternative splicing regulation in HIV-1. The organization is as follows : we briefly describe our integrated approach with the statistical and algorithmic components present in our software (Sect.2). One can submit a consensus motif and/or a motif collection and train the M-SVM, respectively, and retrieve in the same graphical interface the results of both approaches. We illustrate our method (Sect.3) on the search for alternative splicing regulatory sites in HIV-1 using SELEX results for two SR proteins SC35 and 9G8, components of the spliceosome, that will bind such regulatory sites. Our results, confirmed by independant experimental results show the strength of such an integrated approach.

2 KOALAB

2.1 Statistical learning and algorithmic approaches

In order to discover heterogeneous motifs, statistical learning is an efficient approach because it is specified to learn a generalization rule from such type of data. In [20], this method has been used to discover biological motifs such as splicing sites using pattern recognition support vector machine (SVM) [22]. Note, however, that such an approach typically requires several binary classifications for different kinds of motifs. In KOALAB, we propose to use a multi-class support vector machine (M-SVM) [9] to discover several binding motifs simultaneously. In contrast to HMM based methods, we do not need neither a priori knowledge nor prior treatment of the data. The SVM method is a canonical machine learning tool [10, 5] that was proven as a powerful method for pattern recognition in different problems such as handwritten digit recognition, object recognition, voice identification, and text categorization, etc (see for instance [6]). In such areas, the performance of the SVM was equivalent or higher than that of classical non-linear regression models such as neural networks (multilayer perceptrons, MLP). In biology, SVMs have been applied in different fields among which DNA microarrays gene expression data classification [18], protein function classification [2], help on breast cancer diagnosis [14], identification of splicing sites in eukaryotic sequences [20].

Building upon the uniform strong law of large numbers, a new family of multi-class SVMs (M-SVMs) has been specifed in [9]. This specificity allows the scientist to perform a multi-class classification in a single step instead of applying any decomposition method to several binary classification results. In biology, this is of particular interest in order to avoid a too high decomposition level of a complex system.

Conversely, if one has to look for motifs that are more homogeneous i.e., conserved and sufficiently documented, there is no need to use statistical learning methods. This type of motifs may be identified by a discrete pattern matching approach. Hence, we include in our tool an efficient algorithm for pattern search: grappe [13]. Compared to other discrete pattern matching software, grappe offers additional flexibility in pattern description, such as presence of don't care symbols (wildcards) of unbounded or bounded length. Patterns with substitution errors can also be taken into account. The number of errors and their occurrences in a pattern can be specified by the user. A version of `grappe` devoted to nucleic acids sequences treatment (able to deal with specific substitution codes) is used in our tool.

2.2 Integrated software

Although being useful for biological problems, combining the M-SVM and grappe methods remains difficult for a non-specialist. We overcome this problem by developing a software named KOALAB (KOupled Algorithmic and Learning Approach for Biology), which provides a user-friendly interface to M-SVM technology and grappe. KOALAB has been designed to guide the biologist in the discovery of biological motifs in a genome. Using web technology, the user can apply the latest version of the M-SVM software without having to be concerned about technical details. KOALAB can be installed on a local or remote web server. Only a web browser is needed to use it.

The M-SVM interface is designed to reflect the three stages of the process.

- First, the user is invited to provide the learning database containing the collection of motifs together with their respective tags. The learning process can be started.

- Second, the user can perform an evaluation for the progression of the learning process, providing the validation database which is different from the learning one. The evaluation is made on the ability the SVM shows to retrieve the correct labels for these new objects. Even if the theory of such a tool assumes the learning process has to be carried out thoroughly, one can stop it before the strict optimality criteria are satisfied.

- The user can finally proceed with the exploitation phase for which he has to provide the sequence in which to search for motifs. KOALAB provides a practical graphical interface to handle the M-SVM output. The interface incorporates a variety of tools, including data analysis techniques such as signal filtering. Combining them with an M-SVM is an efficient way to detect both known and unknown biological motifs.

When known biological motifs are available, the users can confront the two methods by searching via grappe a pattern corresponding to the motif of interest. KOALAB integrates the statistical and pattern finding results in a graphical representation and summarizes both of them, highlighting the positions along the sequence for which both results are congruent. Because of the high CPU cost of the M-SVM process and given the type of biological questions this tool has been designed for, the users are not supposed to submit entire genomes but rather regions of interest. By combining an algorithmic and a statistical learning method, KOALAB is particularly designed to explore genomes from a new point of view, considering rather small regions in which interactions with regulatory elements lie.

3 Application to regulatory binding sites discovery

As stated before, discovering nucleic acid/protein interaction motifs remains an open problem. We applied our tool to the search for alternative splicing regulatory binding sites in the genome of HIV-1. This phenomenon is crucial for the virus to generate the full protein repertoire during its life cycle (for review, see [11]). As the virus uses the host cell's spliceosomal machinery to allow the splicing of its primary transcript, the splicing regulation mechanisms are similar to that of the cellular pre-messenger RNAs. Among the 9kb long genome sequence, four donor and eight acceptor splice sites allow the virus to produce about 40 different messenger RNAs from which the whole protein repertoire will be obtained. The use of the different sites, i.e. the alternative splicing, is highly regulated in order to produce the right protein at the right time. The regulatory proteins Nef, Tat and Rev are produced in the early phase whereas auxiliary Vip, Vpu, Vpr and structural-enzymatic Gag, Gag-Pol, Env proteins are specific of the late phase. A sharp control of the alternative splicing will drive the progression in the life cycle of the virus. SR (Serine aRginine rich) cellular proteins take part in the splicing reaction [8] consisting in removing introns from the pre-messenger RNA to release the mature messenger RNA. They allow the spliceosome to assemble around a splicing site close to the protein binding sites, this step being often limiting for the whole splicing reaction. Playing such a central part in the process, they are involved in constitutive as well as alternative splicing and have already shown their crucial connection with the HIV-1 virus physiology. We hence searched for SR proteins binding sites KOALAB to this task.

3.1 Training data: ligands from SELEX experiments

The experimental technique of Systematic Evolution of Ligands by EXponential enrichment (SELEX) [21] is a method from combinatorial chemistry devoted to the identification of ligand molecules that bind by affinity on a given target molecule. This experimental approach generates a large number of potential ligands from a nucleic acid pool. The automatically extracted data are safe and without *a priori* assumption, but their interpretation remains difficult especially for the case of heterogeneous data. As stated before, the standard approach, consisting in deriving a consensus motif using different types of alignment methods is often inappropriately applied because of too high heterogeneity in the results. This approach has two major drawbacks: on the one hand, it is possible to produce consensus sequences that have no biological significance (do not correspond to any motif). On the other hand, some significant motifs may be missed. In [19], SELEX experiments are used as a training set to set HMM parameters. In our case, the knowledge on regulatory motifs is not sufficient to develop any probabilistic model which needs some initial conditions that are not always possible to determine. Therefore, a statistical learning approach seems appropriate. To train KOALAB, we used SELEX results for two different SR proteins: SC35 and 9G8. They belong to different subfamilies of SR proteins. SC35 only contains an RRM domain (for RNA Recognition Motif) in its RNA interaction domain whereas 9G8 contains an RRM and a Zn-finger (for Zinc-Finger) motifs [8]. Those data allow to assume that the RNA binding properties of the two proteins will be slightly different and that they won't bind to the same sequences. Prior studies by global consensus determination have confirmed this statement. However, some functional studies suggest that these two proteins might compete one another directly via a competition for some sequence target. That point was particularly interesting to check if our method was able to give some evidence for this situation. The training data were obtained from SELEX experiments with 11 cycles for SC35 and 12 cycles for 9G8 respectively. For each protein, a collection of about 60 sequences (18 nucleotides-long each) was used for training. Compared to the global consensus sequences obtained by other studies for SR proteins, the size of 18nt correspond to the largest motif available, the average being around 10-12 nucleotides. We hence ensure that our windows would encompass the whole interaction region.

3.2 Results

We submitted the whole HIV-1 genome (isolate BRU, genebank accession K02013) to the analysis (it is rather small, about 9kb). We here concentrate on regions of interest more particularly studied because of their importance in the virus alternative splicing regulation. Several splicing regulatory sequences have been identified on the viral RNA downstream of the acceptor A2, A3, A6, A7 sites and upstream from the donor D4 site. We illustrate our results for A3 and A7 regions which are located in structured parts of the viral RNA. The A3 site is exclusively used to produce the tat mRNA and A7 required for the tat and rev mRNA that will give proteins used in the early phase. They are intensively studied because the target of very complex and sharp regulations. As a general comment, we encountered some difficulties with consensus search by `grappe`. Indeed, the consensus available that have been published (for examples [4, 15]) do not even match experimentally proved regions such as the site between ESS3a and ESS3b in the A7 region [17]. The reasons for such a result can be that the consensus are generally longer than the real sequences bound by the proteins or that, as mentioned before, derived

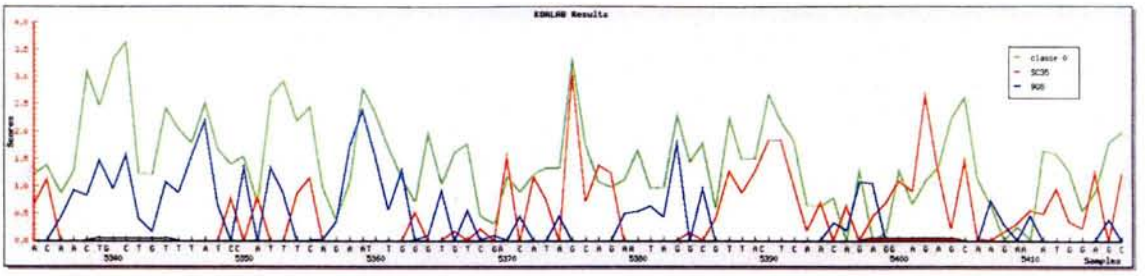

Figure 1: **Areas of alternative splicing regulatory motifs in a part of the HIV-1 genome.** The plots represent the M-SVM results for 2 SR proteins (blue and red) and the non binding segments (green). The bars represent the **grappe** results corresponding to documented SR proteins binding site consensus. The color bar switch to the same SR protein color code when the results for the two methods are consistent.

from a collection of too heterogeneous motifs, the consensus do not correspond to any real sequence anymore. In order to be able to retrieve some matches, one can use central sub-sequences from consensus but soon, the opposite backside effect is observed with the explosion of the system providing a huge quantity of matches. It is anyway of interest to be able to do this confrontation because consensus research was the first technique applied in such type of questions.

An illustration of correlation between grappe and M-SVM result is given in Fig. 1. The graphical interface has been designed to show the grappe results along the sequence as a bar the colour of which corresponds to the M-SVM plots when both results are congruent. In this example covering a region around the A3 site, a SC35 consensus as well as an M-SVM signal for this protein are retrieved around position 5400) where SC35 has effectively been shown to bind (unpublished data). Another consensus grappe hit is found around position 5340 with no M-SVM signal in a region where SC35 is not supposed to bind. Because of the global difficulty to interpret grappe results for SR proteins, we will concentrate here on the presentation of M-SVM results and their confrontation to facts already available on alternative splicing regulation for A3 and A7.

3.2.1 M-SVM results for A3

The A3 site is part from the A3 to A5 sites that are mutually exclusive. A3 is used to produce the tat mRNA whereas A4a, b and c are used for rev mRNA and A5 will take part into env and nef mRNA. We found some sites for SC35 and 9G8 in this region (see Fig. 3.2.2), among which some are experimentally confirmed by a study on SC35 interaction with the viral RNA (unpublished data). An interesting observation is that we find a 9G8 site in a region were SC35 has been proved to bind. This could lead to a competition situation between SC35 and 9G8 that are not equivalent for their effects on the splicing efficiency at the A3 site.

The A3 site activity is also under the control of two Exonic Splicing Silencers, ESS2p and ESS2 that respectively bind hnRNP H and hnRNP A/B repressing proteins. A potential competition between SC35 and hnRNP A/B at the ESS2 site has already been suggested as SC35 has experimentally been shown to bind to ESS2. Our results suggest a potential equivalent situation at the ESS2p site between hnRNP H and 9G8.

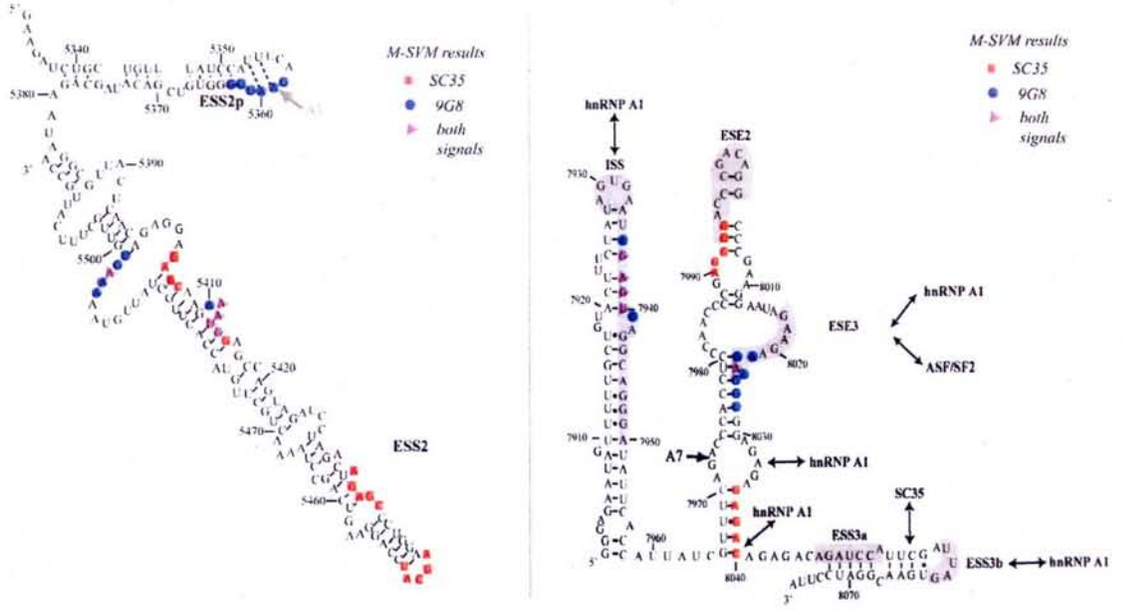

Figure 2: **M-SVM results for HIV-1 splicing acceptor A3 and A7 sites.** The structured regions of the HIV-1 RNA around A3 and A7 sites are represented ([12] [16]). The grey bars show the Exonic Splicing Silencer (ESS) and Exonic Splicing Enhancer (ESE). Red squares and blue circles mark window center positions of M-SVM signals for SC35 and 9G8, respectively. Purple triangles mark positions where both signals are found.

3.2.2 Results for the A7 site

This site is used to generate the tat and rev mRNA, producing the respective early phase regulatory proteins. For this site, we find two potential 9G8 binding sites in regions where hnRNPA1 has been shown to bind, the Intronic Splicing Silencer ISS and the Janus element also called ESE3 (for Exonic Splicing Enhancer 3, [16]). The ESE3 site being known to bind ASF/SF2 and hnRNPA1, there hence might be a three partners balance involved into the effect of this site reaching a level of complexity that begins to be difficult to deal with. It should be noticed that we were unable to retrieve a binding site for SC35 between ESS3a and b where it had been shown in [17]. Our short study of this site confirms the high level of complexity used to regulate its use. Considering that we did not study the binding properties of all the proteins known to be involved into alternative splicing, the authors would like to emphasize the interest there would be to try a formal modelling for these very complex phenomena to acquire a better undestanding of what is really going on during the virus life cycle.

A general comment about the M-SVM results is the global difficulty one has to get a sharp information about the boundaries of the potential interaction sequence given. Indeed, windows used in the training dataset are 18 nucleotides-long. When a 5 nucleotides-long hit is found this doesn't mean that those nucleotides are the target but that they are in the window of the target. This observation highlights the extreme care with which one has to take the results and confirms the need for experimental validation.

4 Discussion and perspectives

The quality of the results we obtained on the highly documented HIV-1 virus genome based on SELEX/M-SVM analysis are very encouraging when compared to independent experimental studies. The `grappe` tool is very useful to make an integrated analysis. To date, KOALAB software offers the opportunity to see and compare the results of the algorithmic and statistical learning methods in a same graphical output but it is also of interest to consider the possibility of real integration of the results from the two methods. An intermediary way of searching nucleic acid-protein interaction sequences is to use probabilistic models such as HMM that have been developed in the ESE finder tool. It would be at least of interest to confront all those techniques in a single study to find out whether an integration is of use or not.

Moreover, some sites were found by the M-SVM inside known splicing activating regions for which no further information about the involved proteins was available so far. This may lead to new hypothesis hence orienting the research of the biologist. The method could be further improved by adding SELEX data for other regulatory proteins. Our results clearly highlight the complementarity between the two methods employed and validates KOALAB method as a posteriori experimental results came and confirmed some of our hypothesis.

KOALAB has been introduced as a tool that helps experimental biologists to discover motifs. It now integrates an algorithmic and a statistical learning methods that can be confronted or used separately, depending on the heterogeneity of the motifs to be searched.

A classical intermediate method corresponds to the use of probabilistic models such as HMMs (Hidden Markov Models). In the case of alternative splicing regulation, HMM models have been developed for some SR proteins from SELEX data as well into a software called ESE finder (for Exonic Splicing Enhancer finder [3]). A possible extension of KOALAB would be to add such a method to be able to cover the full range of available methods for motif discovery. The user could make the choice of using only one or two best fitted methods according to the data and the biological purpose or, if possible, to use the whole repertoire of methods in order to confront their results throughout a single graphical interface.

After interpretation, one is able to identify parts of the genome that are potentially interesting in understanding alternative splicing regulation or any other problem, but still have to be validated experimentally. Furthermore, the resulting data can be used to formulate hypotheses on alternative splicing regulation which can be integrated into a formal model [7] that could lead to new hypotheses regarding global splicing regulation which could be experimentally tested.

Moreover, since the input of a SVM is not restricted to a real-valued vector, this machine could be extended to handle more complex data such as secondary structure information. Our software combines multi-class support vector machines with a discrete method to discover and validate motifs. In future work, KOALAB will integrate a variety of methods for biologists ranging from an algorithmic to a statistical learning approach including probabilistic automata. Following the idea that a computational method should be dedicated to a specific biological problem, KOALAB is designed for motif discovery fitting the biological purpose. Combining experimental analysis and computational tools is the key for providing an efficient motif discovery method.

Acknowledgement: The authors thank Y. Guermeur for M-SVM development, Y. Guermeur, C. Branlant and A. Bockmayr for their helpful comments during the writing of this paper and J. Stevenin for providing SELEX data.

References

[1] Bilodeau, P. S., Domsic, J. K., Mayeda, A., Krainer, A. R., Stoltzfus, C. M. (2001) RNA splicing at human immunodeficiency virus type 1 3' splice site A2 is regulated by binding of hnRNP A/B proteins to an exonic splicing silencer element. *J Virol.*, **75**(18), 8487–8497.

[2] Cai, C. Z., Wang, W. L., Sun, L. Z., Chen, Y. Z. (2003) Protein function classification via support vector machine approach. *Math Biosci*, **185**(2),111–122.

[3] Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q., Krainer, A. R. (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**(13), 3568–3571.

[4] Cavaloc, Y., Bourgeois, C. F., Kister, L., Stevenin, J. (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA*, **5**(3), 468–483.

[5] Cortes, C., Vapnik, V. (1995) Support-Vector Networks *Machine Learning*,**20**(3), 273–297.

[6] Cristianini, N., Shawe-Taylor, J. (2000) An Introduction to Support Vector Machines and other kernel-based learning methods. *Cambridge University Press*

[7] Eveillard, D., Ropers, D., de Jong, H., Branlant, C., Bockmayr, A. (2003) Multiscale modeling of alternative splicing regulation. In *Computational Methods in Systems Biology (CMSB'03)*, Springer LNCS **2602**, 75–87.

[8] Graveley, B. R. (2000) Sorting out the complexity of SR protein functions. *RNA*, **6** (9), 1197–1211.

[9] Guermeur, Y., Elisseeff, A., Paugam-Moisy, H. (2000) A new multi-class SVM based on a uniform convergence result. *IJCNN'00*, **IV**, 183–188.

[10] Guyon, I., Boser, B. E., Vapnik, V. (1992) Automatic Capacity Tuning of Very Large VC-Dimension Classifiers. *NIPS*, 147–155. *Theoretical Computer Science*, **178**, 129–154.

[11] Hope, T. J. (1999) The ins and outs of HIV Rev. *Arch Biochem Biophys*, **365** (2), 186–191.

[12] Jacquenet, S., Ropers, D., Bilodeau, P., S., Damier, L., Mougin, A., Stoltzfus, C.,M., Branlant, C. (2001) Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing. *Nucleic Acids Res.*, **29**(2), 464–478.

[13] Kucherov, G. & Rusinowitch, M. (1997) Matching a set of strings with variable length don't cares. *Theoretical Computer Science*, **178**, 129–154.

[14] Liu, H. X., Zhang, R. S., Luan, F., Yao, X. J., Liu, M. C., Hu, Z. D., Fan, B. T. (2003) Diagnosing breast cancer based on support vector machines. *J Chem Inf Comput Sci.*, **43**(3), 900–907.

[15] Liu, H. X., Chew, S. L., Cartegni, L., Zhang, M.Q., Krainer, A. R. (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol.*, **20**(3), 1063–1071.

[16] Marchand, V., Mereau, A., Jacquenet, S., Thomas, D., Mougin, A., Gattoni, R., Stevenin, J., Branlant, C. (2002) A Janus splicing regulatory element modulates HIV-1 tat and rev mRNA production by coordination of hnRNP A1 cooperative binding. *J Mol Biol.*, **323**(4), 629–652.

[17] Mayeda, A., Badolato, J., Kobayashi, R., Zhang, M. Q., Gardiner, E. M., Krainer, A. R. (1999) Purification and characterization of human RNPS1: a general activator of prem-RNA splicing. *EMBO J*, **18** (16), 4560–4570.

[18] Qian, J., Lin, J., Luscombe ,N. M., Yu, H., Gerstein, M. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*,**19**(15), 1917–1926.

[19] Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J., Mermod, N. & Bucher, P. (2002) Higth-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nature Biotechnology*, **20**, 831–835.

[20] Sun, Y. F., Fan, X. D., Li, Y. D. (2003) Identifying splicing sites in eukaryotic RNA: support vector machine approach. *Comput. Biol. Med.*, **33**(1), 17–29.

[21] Tuerk, C. & Gold, L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.

[22] Vapnik, V. (1992) Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems 4*, 831–838.

9.2 Une interface graphique dédiée aux problèmes biologiques

Les travaux que nous venons de présenter, justifient l'utilité expérimentale que le logiciel KOALAB peut apporter aux biologistes dans la recherche des motifs de régulation d'épissage. Nous proposons dans cette section d'être plus technique concernant les propriétés de KOALAB. Nous illustrerons nos propos avec la recherche des motifs de fixation dans la région du site accepteur d'épissage A3.

9.2.1 Contrôle des approches standards de recherche de motifs

Avant d'intégrer les résultats des deux approches théoriques, KOALAB permet de piloter une M-SVM et `grappe`. L'intérêt de KOALAB est alors de proposer un outil graphique pour contrôler des approches théoriques de pointe. Il est pour cela nécessaire de contrôler le bon déroulement de chaque méthode.

Contrôle de l'apprentissage statistique : piloter M-SVM

KOALAB propose une interface graphique qui exécute la phase d'apprentissage de la M-SVM. Il suffit pour cela à l'utilisateur de sélectionner la base d'apprentissage avec un *browser* ou navigateur. Cette base se compose de vecteurs auxquels sont associées des classes d'appartenance ou étiquettes. La base une fois sélectionnée, KOALAB va générer automatiquement les vecteurs qui seront la base nulle d'apprentissage. Cette base contiendra des échantillons qui ne seront pas reconnus comme possédant les classes définies positives.

KOALAB propose alors différents noyaux (correspondant à la fonction k de la Section 5.2.4) pour exécuter l'apprentissage. Cet apprentissage peut être plus ou moins long en fonction de la taille de la base d'apprentissage et de la taille des vecteurs supports d'information. La complexité est quadratique pour la taille de la base d'apprentissage et linéaire pour la taille du vecteur. Néanmoins la complexité de l'algorithme d'optimisation qui est présent dans la M-SVM est dépendante du type de noyau que l'on utilise. Les critères d'optimisation de l'apprentissage semblent facilement confus pour le non-spécialiste. Un critère est cependant facile à vérifier. Il suffit de comparer les valeurs de la fonction objective du problème dual et du problème primal. Pour rappel, on connaît la valeur du dual et l'apprentissage permet d'estimer la valeur du primal. Dans ces conditions, si les valeurs convergent, on est droit d'estimer que l'apprentissage est terminé. Le ratio de ces deux valeurs est représenté par KOALAB dans la Figure 9.1. Il est alors suffisant pour estimer l'apprentissage. Les valeurs du problème primal et dual doivent converger amenant leur ratio à une valeur de 1. Le suivi et la manière dont converge ce rapport permettent d'estimer le moment où l'apprentissage est optimal pour les données à disposition.

Une fois que l'apprentissage statistique est probant, il est possible d'appliquer la généralisation aux données à traiter, comme ici au génome de HIV-1. Il suffit alors sous KOALAB de charger la séquence du génome qui sera automatiquement segmentée en vecteurs de taille correspondante à celle des vecteurs de la base d'apprentissage. KOALAB

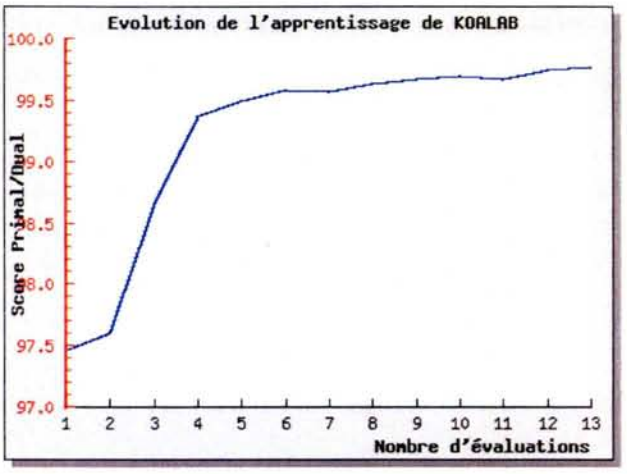

FIG. 9.1 – **Vérification graphique de l'apprentissage statistique par KOALAB.** On représente ainsi le pourcentage de la variation du ratio des valeurs dual sur primal qui permettent de rendre compte de l'apprentissage statistique. Au cours de l'apprentissage, le pourcentage d'apprentissage converge vers la valeur 1 qui correspond à un apprentissage complet. Il est alors intéressant d'analyser la manière dont converge le ratio.

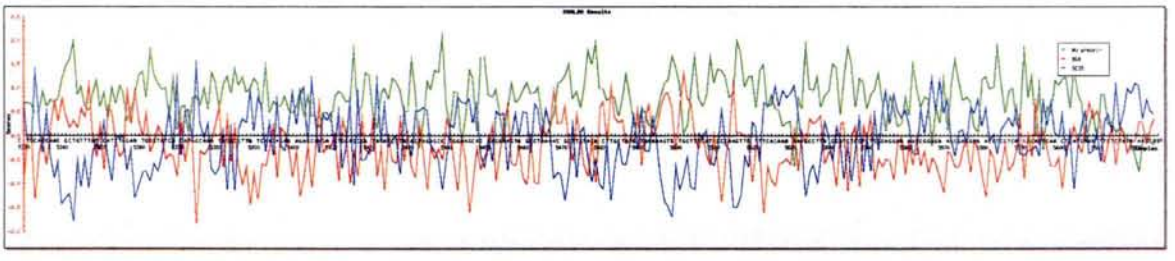

FIG. 9.2 – **Représentation des scores de la M-SVM par KOALAB.** Chaque ligne correspond aux fluctuations du score d'une protéine pour une position donnée du génome. A une position donnée, le score le plus important désignera l'étiquette assignée à cette position.

assignera ensuite une étiquette aux vecteurs à identifier. La généralisation de l'apprentissage donnera des scores statistiques pour chaque étiquette à un vecteur donné. Le score le plus important sera celui de l'étiquette qui sera associée au vecteur, qui correspondra à une position nucléotidique sur le génome de HIV-1. KOALAB projette ensuite les scores sur la séquence d'exploitation de façon à analyser les zones pour lesquelles les scores importants indiquent un site potentiel de fixation de protéine, comme l'indique la Figure 9.2.

Contrôle de la recherche discrète : utilisation de `grappe`

KOALAB permet également de piloter l'outil de recherche discrète `grappe`. Pour cela, il suffit de charger les motifs discrets à rechercher avec la syntaxe de `grappe` pour permettre des fluctuations (des *jokers*). Une adaptation de `grappe` pour KOALAB permet d'utiliser un algorithme qui donne toutes les positions des motifs sur un génome donné. Le résultat

des recherches se représente sous forme de graphique qui assigne une barre sur le génome aux positions des motifs identifiés. Par rapport à l'utilisation standard de `grappe`, une étape supplémentaire intervient dans KOALAB puisqu'il est possible d'associer aux motifs discrets les étiquettes chargées au cours de l'apprentissage statistique.

9.2.2 Intégration des résultats

KOALAB tel qu'il est décrit jusqu'à présent ne permet que de faire une simulation concurrente de deux approches de recherche de motifs. Néanmoins, si l'on considère les outils comme étant complémentaires car aux extrémités d'un gradient de méthodologies de recherche de motifs, il faut les intégrer. Cette intégration s'effectue par le biais d'un algorithme de comparaison entre les sorties statistiques et discrètes. Dans cet algorithme, `p` correspond à la position d'un motif identifié par `grappe` et `l` sa taille en nucléotides. Après assignation des classes aux motifs discrets, on se place à la position identifiée par `grappe` afin de comparer position à position si sur la longueur des motifs donnés par `grappe` correspond aux prédictions après apprentissage.

```
 #assignation de la classe au motif discret
 $index_utilisateur = classe ($motif)
for i = p, i++;
          {#recherche le maximum des scores de M-SVM
           $index_classe=max(score(1,i),score(2,i),score(3,i));
           #compare classe assignée à celle prédite par M-SVM
           if $index_classe = $index_utilisateur
                    {$n = true
                    };
          }
```

Ce simple algorithme compare automatiquement les étiquettes issues de l'apprentissage statistique (`$index_classe`) et celles issues du *pattern matching* (`$index_utilisateur`) et vérifie si elles sont concordantes. Le cas échéant, la variable booléenne `$n` devient `true` et les barres de résultats de `grappe` se colorent de la même couleur que les étiquettes associées à l'apprentissage. Il ainsi facile de repérer les résultats confirmés par les deux approches comme l'illustre la Figure 9.3.

9.3 Application de KOALAB aux motifs de régulation d'épissage sur HIV-1

Le développement de KOALAB est intéressant du point de vue implémentation. Mais au-delà de l'association des outils de recherche discrets et statistiques, le logiciel doit surtout faire la preuve de son efficacité auprès des biologistes. Il est alors important de confronter notre démarche sur le cas concret de la recherche des motifs de régulation d'épissage. Étant en possession des données expérimentales SELEX pour les protéines 9G8 et SC35, nous nous focaliserons sur leurs exploitations par KOALAB.

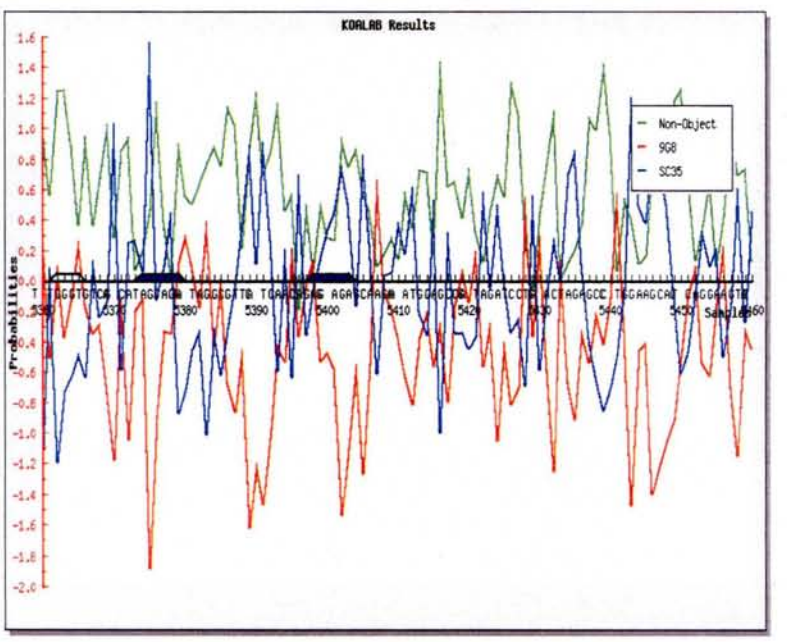

FIG. 9.3 – **Représentation de l'intégration des résultats de grappe et de M-SVM par KOALAB.** Les résultats cohérents avec les deux approches se distinguent par une couleur des barres similaires à celle associée à la classe d'apprentissage. Dans cette figure, les zones de fixation de SC35 confirmées par `grappe` et M-SVM sont bleues. Par ailleurs, on retrouve les scores de M-SVM en bleu pour SC35 ainsi que les motifs de SC35 avec `grappe` non concordants avec la M-SVM sans coloration spécifique.

TAB. 9.1 – **Motifs de protéines régulatrices d'épissage issus de la littérature.** Ces motifs discrets seront identifiés sur le génome de HIV-1 par la méthode `grappe` intégrée à KOALAB. Les parenthèses signalent un choix multiple à une position donnée de la séquence. Les symboles sont standard (R : purine, Y : pyrimidine, N : un des 4 nucléotides)

Protéines de régulation	Motifs consensus	Références
SC35	`GUUCGAGUA`	[Tacke & Manley, 1995*b*]
	`GGGUAUGCUG`	[Cavaloc *et al.*, 1999*a*]
	`UGCNGYY`	[Schaal & Maniatis, 1999*b*]
9G8	`AGAC(G/U)ACGAY`	[Cavaloc *et al.*, 1999*a*]
	`ACGAGAGAY`	[Cavaloc *et al.*, 1999*a*]
	`AGAC(G/U)ACGA(C/U)`	[Lejeune *et al.*, 2001]

9.3.1 Comparaison des motifs discrets et statistiques

KOALAB est dans notre cas un outil de recherche exploratoire de motifs de régulation sur le génome HIV-1. Nous avons à notre disposition des bases de données SELEX pour deux types de protéines SR qui ont déjà été mentionnées dans le Chapitre 7. KOALAB va donc être entraîné sur cette base d'apprentissage. Par ailleurs, il existe dans la littérature des travaux concernant ces mêmes protéines et qui fournissent des motifs consensus. KOALAB va rechercher donc ces motifs discrets par l'intermédiaire de `grappe`. Les motifs que nous rechercherons correspondront alors à ceux mentionnés dans le Tableau 9.1.

En possession de ces données, on effectue alors un entraînement de l'apprentissage sur les données SELEX pour SC35 et 9G8. La qualité de l'apprentissage est celle qui est présentée dans le Figure 9.1. L'exploitation de KOALAB s'effectue sur le génome de HIV-1 mentionné comme `>gi|326417|gb|K02013.1|HIVBRUCG` dans les bases de données correspondant au génome complet du virus d'immunodéficience de type 1 de l'homme isolé BRU (LAV-1) de 9229 nucléotides [Wain-Hobson *et al.*, 1985].

Néanmoins, le procédé intégratif tel qu'il est présenté ici, permet seulement de représenter les résultats communs entre les deux approches. Les approches algorithmiques ont montré leurs limites avec la quantité de faux positifs qu'elles génèrent. Le fait de considérer uniquement les motifs pour lesquels les deux approches sont cohérentes ne permet pas de considérer que les méthodes se compensent. Une approche consiste alors à filtrer les faux positifs en accentuant la compensation. Elle se base sur les technologies de traitement du signal appliqué à la biologie (pour détails voir [Legendre & Legendre, 2000]). On considère alors que la séquence biologique d'ARN du génome de HIV-1 est un signal. Le signal est filtré par l'apprentissage statistique qui retourne un score pour une position donnée. Le signal génomique est donc décomposé en trois signaux de score pour chaque classe qui ainsi possible d'analyser. Les signaux sont composés de bruits autour d'une tendance. Après analyse d'autocovariance, on est en mesure de quantifier le bruit, pour ensuite le filtrer.

9.3.2 Exploitation de KOALAB : vers une cartographie fonctionnelle

La séquence sur laquelle nous allons appliquer KOALAB est constituée de 290 nucléotides situées entre la position 5330 à 5620 du génome de HIV-1 mentionné précédemment. Ce morceau de génome renferme 5 sites accepteurs 3'(A3, A4a, A4b, A4c, A5) ainsi que diverses séquences exoniques activatrices (ESE) et des séquences exoniques inhibitrices (ESS). Cette sous-séquence est représentée sur la Figure 9.10. L'analyse est restreinte à cette sous-séquence dans le but de faire correspondre notre étude à un domaine du génome connu expérimentalement et dont la structure fonctionnelle est particulièrement bien établie par diverses expériences [14], le but étant de calibrer notre approche pré-expérimentale.

Recherche algorithmique des consensus sur la zone isolée

Un premier constat est que les motifs consensus discrets des protéines SC35 et 9G8 recensés dans la littérature et mentionnés dans le Tableau 9.1 ne sont pas retrouvés dans cette partie du génome. La démarche consiste alors à considérer certaines fluctuations dans le motif comme le fait de permettre des substitutions. Dès lors, le nombre de positions explose. Cette observation peut être justifiée par le peu de confiance que l'on peut attribuer aux motifs consensus de la littérature. En effet, les travaux [Eveillard & Guermeur, 2002*a*] mettent en évidence l'hétérogénéité potentielle des motifs de fixation des protéines SC35 et 9G8. Les motifs consensus déterminés par les approches standards ne sont alors pas consistants. Ils nécessitent un traitement statistique au préalable. C'est notamment suivant cette hypothèse, que la recherche des motifs de fixation des protéines SC35 et 9G8 est plus appropriée par une approche statistique comme la M-SVM de KOALAB.

Protocole de recherche statistique sur la zone isolée

Nous privilégierons alors les résultats de la composante statistique de KOALAB pour tester la recherche des motifs. Le protocole se base sur les scores des différentes étiquettes en fonction de leurs positions dans la séquence du HIV-1. Les scores indiquent la possibilité de trouver le motif à la position considérée. Seule la comparaison des différents scores permet de déduire les résultats des prédictions.

Les données de sortie de la M-SVM : Le graphique des données de sortie de la M-SVM (voir Figure 9.4) est difficilement exploitable en raison des nombreuses fluctuations dues aux scores très éparses de la M-SVM. Il convient donc d'envisager un traitement des données afin d'exploiter les résultats. Notre approche pour exploiter les résultats de M-SVM repose sur les techniques d'analyse de données. Pour cela on va s'intéresser au signal que forme l'enchaînement des scores tout au long de la séquence nucléique. Le protocole que nous utiliserons est un protocole d'analyse de séries de données [Legendre & Legendre, 2000]. On suppose alors que les données sont interdépendantes. Cette hypothèse se justifie par le fait qu'un motif peut être présent sur plusieurs fenêtres qui sont chevauchantes. Les scores de la M-SVM des fenêtres sont donc interdépendants.

[14] données issues du laboratoire MAEM

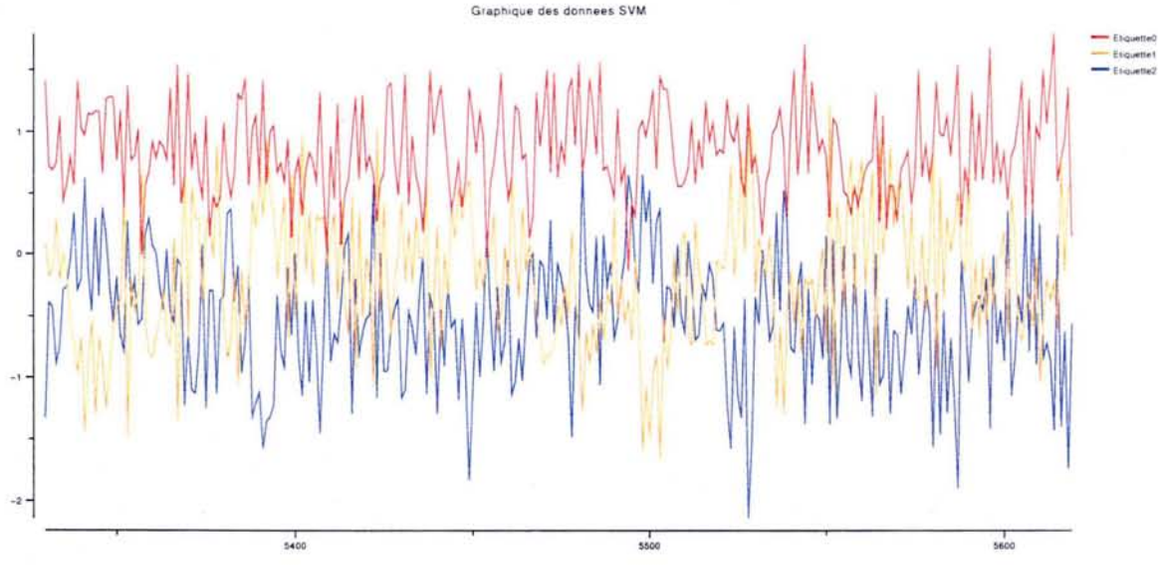

FIG. 9.4 – **Scores bruts de sorties de M-SVM.**

On réalise dans un premier temps une normalisation, qui consiste à niveler les scores. On associe pour cela le score 0 à l'étiquette au score le plus faible. Les autres scores seront estimés en conséquence par soustraction du score minimal. Les scores normalisés présentent alors une lisibilité plus grande sans changer le résultat final.

Les données normalisées de la M-SVM : La Figure 9.5 représente le graphique des données normalisées de la M-SVM. On constate le long de la séquence, un score pour l'étiquette 0 qui est en général supérieur aux deux autres. On considère alors que le motif d'aucune protéine n'est présent à part pour quelques exceptions. Le protocole expérimental mis en place est ainsi confirmé. On constate pour les protéines SC35 et 9G8, la présence de pics supérieurs à celle de l'étiquette 0. Ces pics supérieurs au seuil mis en place par l'absence de motif de protéines, reflètent la présence potentielle en cette position, du motif de la protéine considérée. Néanmoins, la présence de faux positif est possible. Il faut donc émettre des réserves quant à la lecture brute de ces données.

On trouve pour les protéines SC35 et 9G8 respectivement 11 et 1 pics nettement supérieurs à l'étiquette 0. Les pics relevés pour la protéine SC35 sont situés aux positions suivantes : 5356, 5378, 5392, 5401, 5422, 5528, 5550, 5556, 5559, 5564 et 5568. Les Figures 9.6 et 9.7 mettent en évidence l'ensemble de ces pics. Un pic relevé pour la protéine 9G8 est situé à 5493.

Extraction du bruit de fond : Dans le but de minimiser les faux positifs, nous procédons à un lissage des données. La Figure 9.8 représente un variogramme. Il est établi à partir de la fonction d'autocovariance sur les données de sortie de la M-SVM. Le variogramme permet de quantifier le bruit de fond d'une série de données mais plus précisément d'estimer la capacité de lissage que l'on peut appliquer aux données sans perdre de l'information. On constate que les valeurs des trois étiquettes fluctuent autour de la valeur 0. On en déduit qu'il y a un bruit blanc et donc une totale indépendance entre les scores à toutes les positions dans la séquence. Il n'y a donc pas de bruit significatif. On

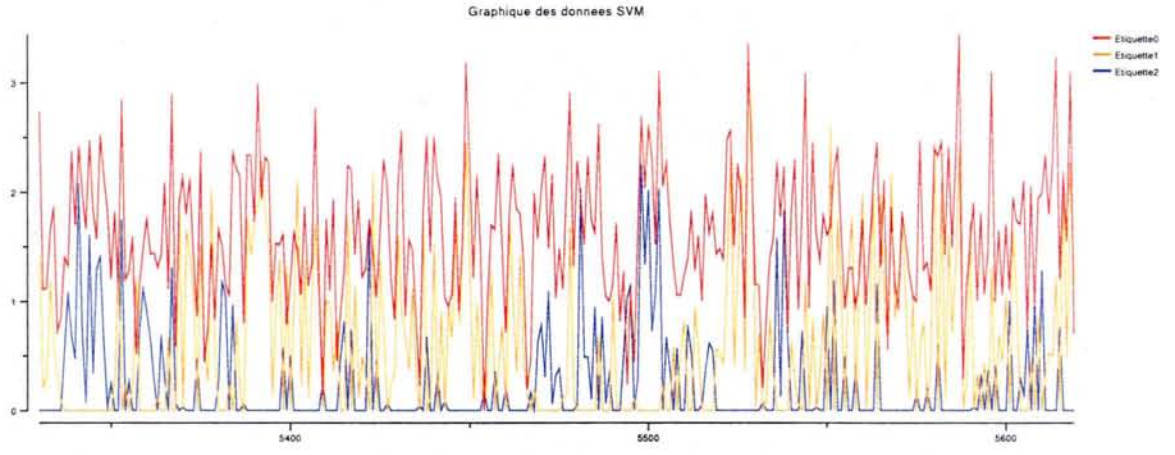

FIG. 9.5 – **Scores normalisés de sorties de M-SVM.**

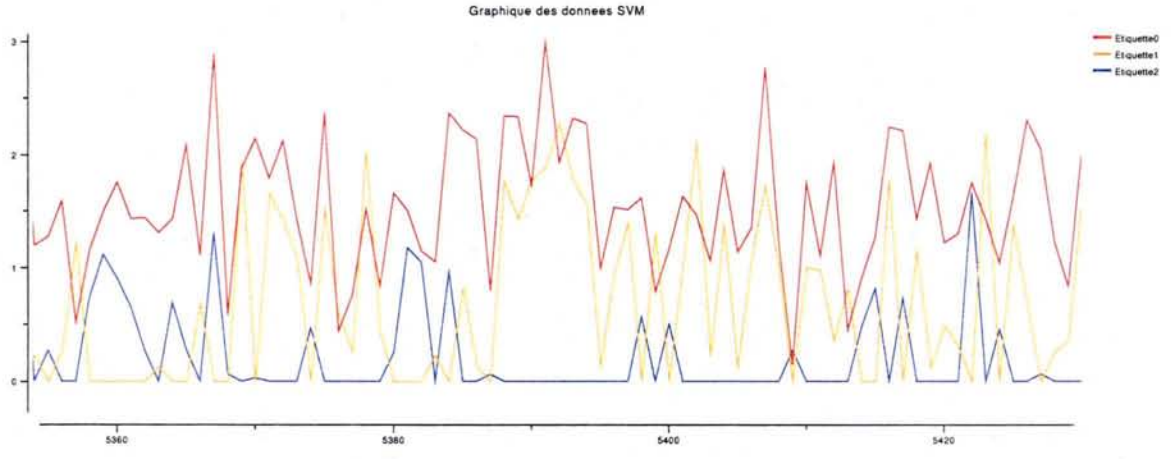

FIG. 9.6 – **Agrandissement de la région 5350-5430 du graphique des données normalisées de la M-SVM.** L'agrandissement est réalisé grâce au zoom de l'interface graphique. On constate que l'axe des abscisses se met automatiquement à l'échelle.

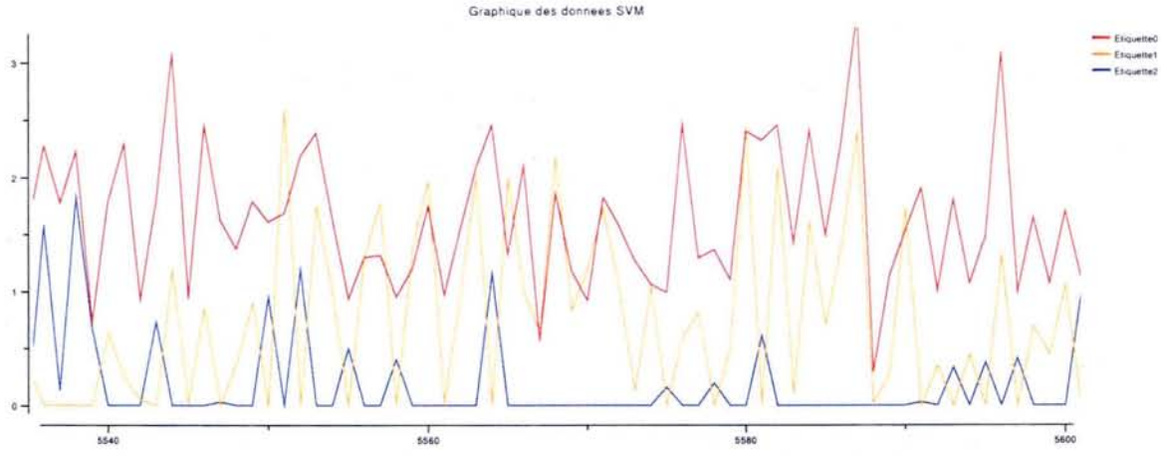

FIG. 9.7 – **Agrandissement de la région 5530-5600 du graphique des données normalisées de la M-SVM.** L'agrandissement est réalisé grâce au zoom de l'interface graphique. On constate que l'axe des abscisses se met automatiquement à l'échelle.

peut alors extraire la tendance de ces données et l'analyser sans contresens, ce qui permet de filtrer les faux positifs.

Extraction de la tendance : Nous utiliserons un protocole de lissage des séries qui repose sur le principe simple des moyennes mobiles. Le principe consiste à faire la moyenne sur une fenêtre de taille n autour d'une position et d'appliquer la valeur moyenne des scores à la position. On obtient alors une nouvelle série de moyennes qui lisse la série initiale d'un indice n. Dans notre cas, les fenêtre de reconnaissance sont de 18 nucléotides. Après confirmation du variogramme, la taille maximale de lissage correspond alors à 18. Au-delà de cette limite, le lissage fait perdre de l'information. La Figure 9.9 représente le graphique des données normalisées analysées avec les moyennes mobiles. Les moyennes mobiles ont été calculés sur une fenêtre de onze positions. On observe alors la tendance des scores des deux protéines 9G8 et SC35.

La protéine 9G8 présente deux domaines de reconnaissance sur la séquence HIV-1 étudiée. Ces deux domaines sont situés respectivement vers 5350 et 5500. Il faut à présent confronter ces résultats à ceux obtenus avec les données de sortie de la M-SVM afin de confirmer ou d'infirmer cette tendance. Si l'on considère les données de sortie de la M-SVM, ces dernières ne mettaient en évidence qu'un pic à la position 5493. Ce pic correspondrait au deuxième domaine retrouvé sur ce graphique alors présent à la position 5493.

La protéine SC35 présente 4 domaines de reconnaissance sur la séquence situés vers 5400, 5430, 5510 et 5550. Ces 4 domaines correspondent tous à la présence de pics dans le graphique des données de sortie de la M-SVM. Les pics considérés sont les suivants : 5401, 5422, 5528 et 5550. On peut ainsi considérer que le motif de la protéine SC35 serait présent à quatre reprises sur la séquence du HIV-1 aux positions 5401, 5422, 5528 et 5550.

Il convient à présent, de mettre en relation les résultats obtenus par l'intermédiaire de l'interface graphique et les connaissances biologiques acquises précédemment. La séquence du HIV-1 étudiée, présente plusieurs sites biologiques : des sites d'épissage 3' et 5' et des séquences exoniques activatrices et inhibitrices (ESE, ESS). Les sites biologiques se situent

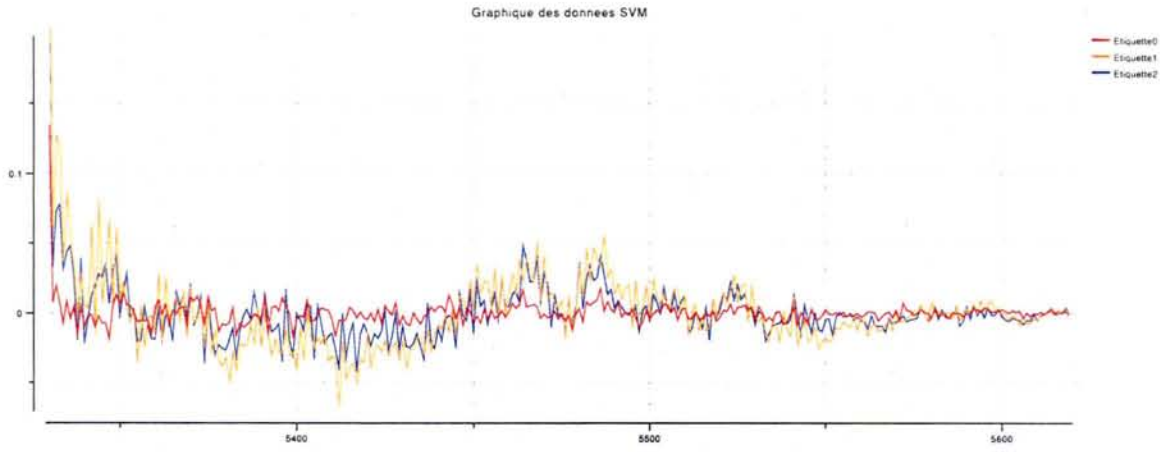

FIG. 9.8 – **Graphique d'autocovariance des sorties de M-SVM.**

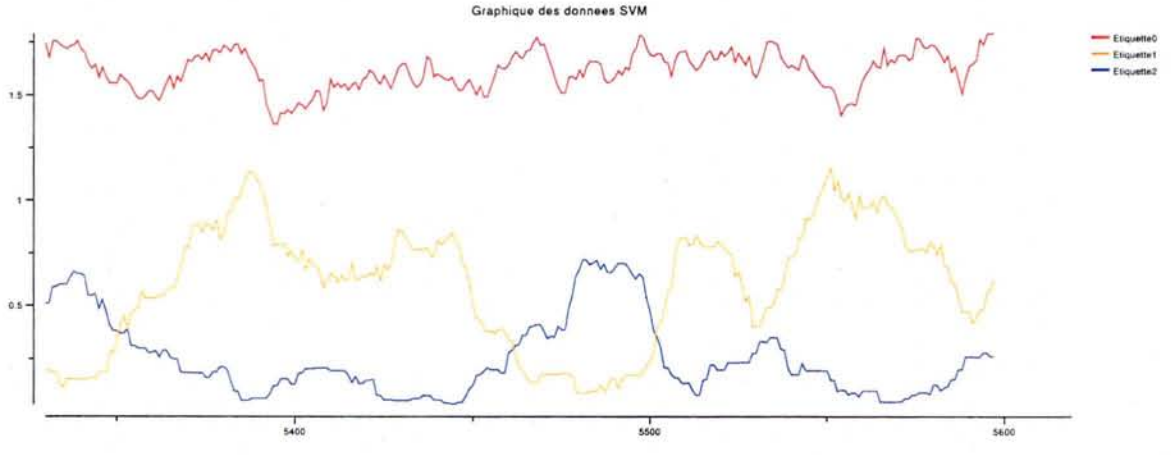

FIG. 9.9 – **Graphique des données normalisées de la M-SVM analysé avec les moyennes mobiles.** Les moyennes mobiles ont été calculé sur une fenêtre de 11 positions.

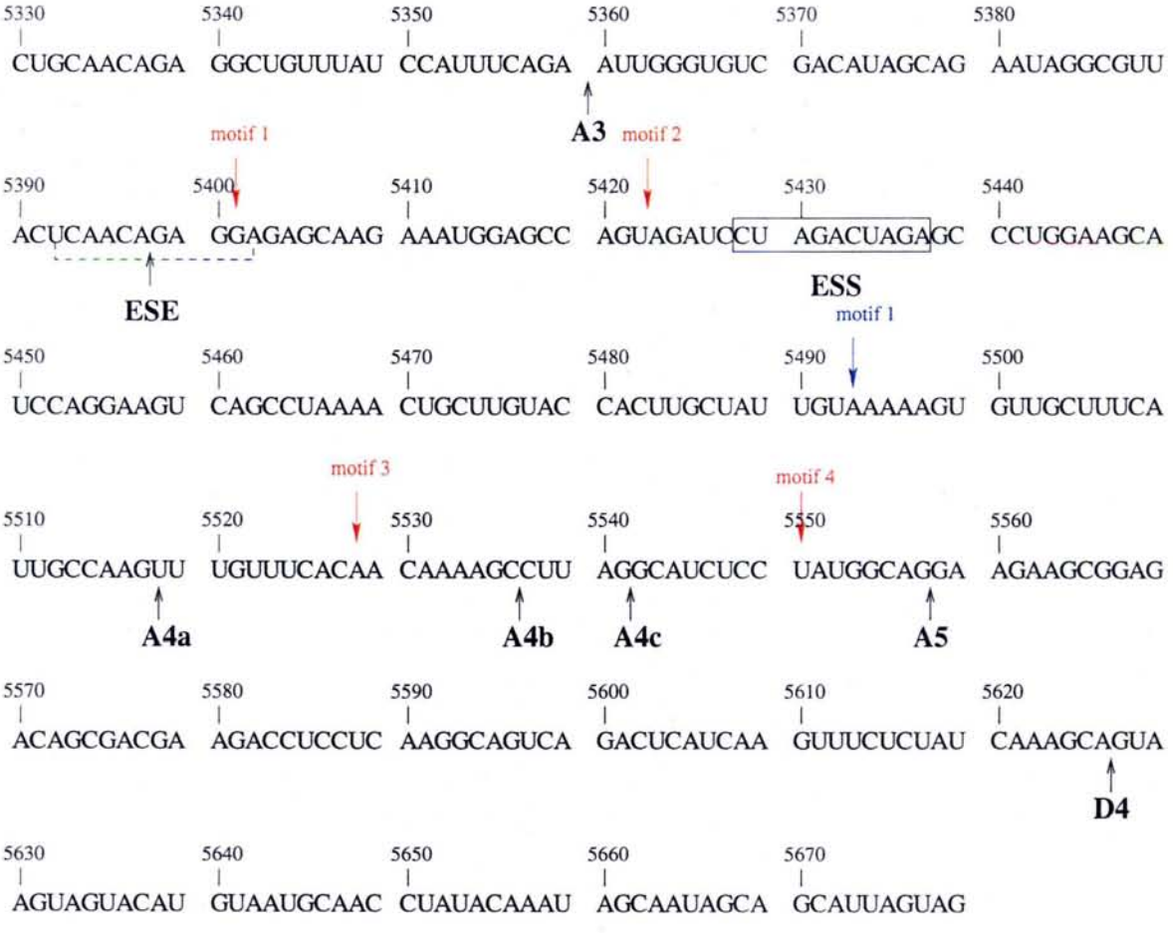

FIG. 9.10 – **Séquence en cis du HIV-1 associée à la présence des motifs mis en évidence par la M-SVM.** Les pics de la protéine SC35 sont représentés en rouge, le pic de la protéine 9G8 est représenté en bleu.

aux positions 5360, 5398, 5428, 5518, 5537, 5541 et 5557 et correspondent respectivement au site A3, à l'ESE, à l'ESS et aux sites A4a, A4b, A4c et A5 (voir Figure 9.10).

Le motif de la protéine 9G8 mis en évidence sur la séquence par la M-SVM, n'a pas de correspondance biologique. En effet, aucun site biologique n'a été expérimentalement mis en évidence à la position 5493. Il s'agit alors, soit d'un faux positif, soit d'un site biologique non encore identifié qui nécessitera l'attention future des biologistes. Les quatre motifs de la protéine SC35 mis en évidence sur la séquence du HIV-1, sembleraient avoir eux tous des correspondances avec des sites biologiquement identifiés. Les pics correspondraient à l'ESE, l'ESS, au site A4a/b et au site A5 (voir Figure 9.10).

9.4 Avantages de l'approche *in silico*

L'approche développée dans KOALAB nécessite d'être étendue à d'autres problèmes pour être validée. Néanmoins, certains résultats sont encourageants.

9.4.1 Validation biologique partielle

D'après les résultats obtenus précédemment, le motif de la protéine 9G8 serait présent sur la séquence du HIV, à la position 5493. Le motif de la protéine SC35 serait lui présent à quatre reprises sur la séquence aux positions 5401, 5422, 5528 et 5550. L'interprétation de ces résultats nécessite de se tourner vers la biologie afin de faire un parallèle entre les connaissances bibliographiques acquises sur la séquence du HIV-1 et sur les protéines SR et les résultats obtenus par expérience *in silico*.

La séquence du HIV-1 étudiée, présente plusieurs sites biologiques. Ces sites biologiques sont les sites d'épissage 3' : A3, A4a, A4b, A4c et A5 et les séquences exoniques activatrices (ESE) et inhibitrices (ESS). Ils sont situés respectivement aux positions 5360, 5518, 5537, 5541, 5557, 5398 et 5428.

On constate que le motif de la protéine 9G8 n'a pas de correspondance biologique. En effet, aucun site n'a été expérimentalement mis en évidence à la position 5493. Il s'agit alors soit d'un faux positif, soit d'un site biologique non encore identifié expérimentalement.

D'après les connaissances bibliographiques actuelles (voir Chapitre 1), les protéines SR possèdent des séquences préférentielles de fixation à l'ARN qui leur permettent de moduler l'utilisation de sites d'épissage alternatif. Elles se fixent aux séquences exoniques activatrices (ESE) et inhibitrices (ESS), et activent ou répriment l'utilisation des sites d'épissage faibles 5' et 3' flanquant. La mise en évidence de motifs de la protéine SC35 au niveau de l'ESE et de l'ESS, nous permet de valider notre approche. Cependant, aucune référence bibliographique ne met en évidence la fixation des protéines SR au niveau des sites d'épissage 3'. On en déduit alors de la même façon que le motif de la protéine 9G8 serait présent à la position 5493.

On constate pour la protéine SC35, la présence de 6 domaines de reconnaissance. Ces 6 domaines correspondent tous à la présence de pics dans le graphique des données de sortie de la M-SVM. On peut ainsi considérer que le motif de la protéine SC35 serait présent à six reprises sur la séquence du HIV-1 aux positions 5356, 5401, 5422, 5528, 5550 et 5570. On retrouve les quatre domaines de reconnaissance obtenus précédemment et deux nouveaux domaines de reconnaissance. Le fait de retrouver ces quatre motifs permet de valider nos résultats et de poursuivre l'analyse. Les deux nouveaux motifs retrouvés aux positions 5356 et 5570 correspondent tous deux à des sites biologiques. Ces sites biologiques sont les sites d'épissage A3 et A5 et sont situés respectivement aux positions 5360 et 5557. Le fait de valider biologiquement nos résultats, nous donne confiance dans les autres motifs trouvés. La mise en évidence de deux nouveaux motifs, nous permet de conclure que le lissage réalisé est meilleur que le précédent. Ce lissage permet de mettre en avant de nouvelles informations car l'on obtient des informations supplémentaires

9.4.2 Émergence d'hypothèses nouvelles

D'après l'ensemble des résultats de M-SVM, on observe la prédiction de motifs de la protéine SC35 au niveau de l'ESE et de l'ESS correspondant aux connaissances bibliographiques. Cependant, la mise en évidence de prédictions de motifs au niveau des sites d'épissage 3' n'a pas été prouvé expérimentalement et rien ne laisse penser à ce jour que les protéines SR pourraient se fixer au niveau des sites accepteurs d'épissage ou à leurs

proximité. Il peut s'agir soit, de faux positifs mis en évidence par notre approche, soit de nouvelles données biologiques encore mal connues actuellement. L'absence de références bibliographiques laisse supposer que l'on se trouve en présence de faux positifs. Les faux positifs pourraient correspondre à une composition en bases des séquences nucléiques de la protéine SC35 très répandue dans le génome du HIV-1. Cependant, la présence des motifs de la protéine SC35 à proximité de tous les sites d'épissage 3', laisse entrevoir la possibilité de leurs existences.

Si on se tourne à nouveau vers la bibliographie, on constate que les protéines SR sont impliquées au niveau de la reconnaissance précoce du site d'épissage. Elles stabilisent à cette étape, la fixation de U1 au site donneur 5' et la fixation de U2AF à l'enchaînement poly-pyrimidique qui est adjacent au site accepteur 3'. Il apparaît donc possible que les protéines SR présentent des motifs de reconnaissance adjacents aux sites d'épissage 3'. Ces sites de fixation pour les protéines SR pourraient permettre de stabiliser la présence de ces protéines à proximité des enchaînements poly-pyrimidiques. Ceci peut favoriser dans un second temps l'interaction entre le facteur U2AF et l'enchaînement poly-pyrimidique qui est adjacent au site accepteur 3'. Le fait de trouver les motifs des protéines SR légèrement décalés du site d'épissage 3' renforce l'hypothèse émise aujourd'hui. La présence de sites de fixations des protéines SR à proximité des sites d'épissage 3' faciliterait la reconnaissance précoce du site d'épissage. Cette hypothèse issue de l'approche *in silico* nécessite naturellement des expériences actuellement en cours, pour corroborer nos conclusions.

9.5 Conclusions

L'approche intégrée que KOALAB illustre est appropriée à la recherche de motifs biologiques. Son utilisation permet au biologiste un résultat optimal. En effet, les méthodes algorithmiques discrètes et statistiques se compensent. Dans notre cas d'étude qui concerne des motifs hétérogènes, l'approche par M-SVM est la plus pertinente car permettant de retrouver les positions déjà isolées expérimentalement. Néanmoins, il serait intéressant d'utiliser notre méthodologie sur un problème moins orienté vers une méthode plutôt qu'une autre. Il s'agirait alors de rechercher des motifs partiellement hétérogènes avec des régions particulièrement bien conservées. On serait alors en mesure de voir si les méthodes se complètent le cas échéant. En effet, si les méthodes algorithmiques et statistiques se complètent, étant chacune à une extrémité d'un gradient méthodologique, il persiste une incertitude concernant leur efficacité pour un domaine d'application se situant entre les deux extrêmes qu'elles délimitent.

Pour combler ce manque méthodologique, il est intéressant à plus ou moins court terme d'intégrer à KOALAB une machine HMM qui possède son domaine d'application dans la zone encore incertaine entre `grappe` et la SVM. KOALAB serait alors une plateforme logiciel pertinente pour la recherche de motifs biologiques. Avant de mettre ce projet en oeuvre, il est dans un premier temps important de pouvoir quantifier l'intérêt des biologistes pour un tel logiciel. Si KOALAB trouve les faveurs des biologistes, il sera alors possible d'envisager de piloter KOALAB par des *workflow* qui permettront de relier KOALAB en temps qu'objet web, à des bases de données biologiques structurées afin que l'entraînement de KOALAB soit de plus en plus pertinent. On pourrait alors décrire

des scénarii méthodologiques a priori pour chaque problème biologique que KOALAB exécuterait automatiquement.

Néanmoins, avant d'envisager la pérennité du logiciel, il est important d'effectuer à court terme une bioanalyse complète des résultats de recherche de motifs de régulation sur le génome de HIV-1. Nous nous sommes dans ce chapitre restreints à une illustration des résultats que l'on pouvait obtenir sans nous étendre sur l'ensemble du génome. Cependant, le laboratoire MAEM possède de nombreux résultats qui pourraient confirmer les résultats *in silico* actuels. L'outil ne possède actuellement qu'une vocation d'approche exploratoire qui ne permet que de délimiter les zones expérimentalement intéressantes. Une généralisation de nos résultats sur d'autres domaines du génome qui sont extrêmement bien étudiés comme dans [Marchand *et al.*, 2002], permettrait de transformer notre approche exploratoire en une méthode de prédiction de sites de régulation.

Dans l'optique de généralisation, il est sans aucun doute pertinent de généraliser la recherche des sites de fixation à d'autre protéines SR que SC35 et 9G8. L'approche consiste alors à utiliser autant de résultats SELEX que de protéines SR considérées comme ayant une influence sur la régulation de l'épissage. Il est ainsi envisageable de considérer les protéines ASF/SF2 ou SRp40. Un banque complète d'apprentissage permettra alors d'utiliser tout le potentiel de la SVM multi-classe. KOALAB serait alors un outil qui permettrait de mettre en évidence *in silico* les motifs chevauchant pour deux protéines et ainsi de mettre en évidence les zones fonctionnelles d'interactions. On isolerait de ce fait, par le biais des sites de fixation des protéines, les ESE et les ESS qui contrôlent la régulation de l'épissage. Il sera alors intéressant de comparer notre méthodologie avec celle de ESEfinder [Cartegni *et al.*, 2003].

Quatrième partie

Modélisation formelle : Analyse de la fonctionnalité des motifs de régulation d'épissage

La recherche automatique des motifs de fixation des protéines SR est nécessaire mais non suffisante pour comprendre la régulation de l'épissage. Il est en effet important de vérifier l'aspect fonctionnel des motifs que l'on détecte au moyen de modèles statistiques. La démarche naturelle est alors de vérifier chaque motif expérimentalement afin de valider la démarche employée. Cette vérification est sans aucun doute la plus pertinente d'un point de vue biologique. Néanmoins elle reste souvent difficile à mettre en oeuvre et peu économique en temps et en argent. Cette approche fait l'objet de nombreux travaux que nous n'aborderons pas au cours de cette thèse. Par ailleurs, les approches de classification mettent en évidence des hypothèses biologiques qu'il est également nécessaire de valider afin d'accroître nos connaissances biologiques d'un phénomène. Les validations expérimentales des connaissances issues d'analyses de données biologiques ou des modèles statistiques sont donc nécessaire à la compréhension du système biologique.

Face à ce type de validation, une alternative intéressante est de tester fonctionnellement les motifs au moyen d'une modélisation formelle. Elle permettrait dans notre cas de justifier des effets régulateurs de certains motifs et paramètres biologiques. Comme nous l'avons mentionné précédemment dans le Chapitre 3, cette approche formelle ne se substitue en aucun point à l'expérience mais possède comme seule ambition d'aider les expérimentateurs à la validation du système tel qu'il est perçu. C'est dans ce but que nous allons tenter dans cette partie de modéliser la régulation de l'épissage alternatif dans laquelle les protéines SR ont une importance prépondérante. Le modèle formel va également nous permettre d'aborder un des aspects prometteurs de la biologie des systèmes : le transfert d'échelle en biologie. L'épissage alternatif est en effet un modèle biologique particulièrement bien adapté pour aborder ce concept d'un point de vue théorique. Le processus d'épissage agit en effet sur plusieurs échelles biologiques qui sont souvent associées à des techniques expérimentales d'investigation différentes. Ainsi, l'épissage alternatif peut être abordé du point de vue moléculaire mais aussi cellulaire. On peut en effet s'intéresser à la régulation de l'épissage sur un seul et même site. Ce mécanisme peut alors être appréhendé par des travaux de biologie moléculaire. Nous considérerons alors cette régulation comme locale. Mais l'épissage alternatif peut être aussi régulé sur plusieurs sites simultanément. Nous considérerons cette dernière régulation comme globale. Ces deux modes de régulation sont inter-dépendants suivant le concept même de transfert d'échelles. Cette dépendance est encore mal connue car difficile d'accès expérimentalement. Néanmoins une modélisation formelle teste certaines hypothèses d'interactions entre ces deux échelles ce qui permet une extrapolation des effets de la régulation moléculaire sur la régulation cellulaire de l'épissage.

Notre proposition pour appréhender cette nouvelle problématique ambitieuse de la biologie des systèmes repose sur un raisonnement par contraintes. Nous considérons pour cela que chaque échelle possède des contraintes différentes. La gestion des différentes échelles consiste alors à modéliser les interactions entre les différentes contraintes pour ensuite raisonner sur le modèle du système biologique dans son intégralité. Suivant ce postulat, nous sommes en mesure de reproduire par des contraintes les comportements issus directement des hypothèses biologiques et ce malgré les incertitudes propres aux transferts d'échelles associées à l'épissage alternatif. Notre démarche de modélisation repose sur deux parties distinctes. Dans un premier temps, nous modéliserons les interactions moléculaires associées au comportement local d'un site d'épissage. Nous illustrerons nos propos par un

modèle formel de la régulation de l'épissage alternatif par les protéines SR dans le Chapitre 10. Ce modèle est basé sur les hypothèses issues de nombreuses expériences. Une fois le modèle local et les hypothèses associées validées, il est possible d'intégrer le comportement local dans une deuxième étude qui modélise le comportement global incorporant plusieurs sites de régulation d'épissage comme décrit dans le Chapitre 11. L'objectif final est alors de proposer un outil formel aux biologistes pour tester les connaissances concernant la régulation de l'épissage alternatif par les protéines SR sur plusieurs échelles et ce pour appréhender le processus biologique dans sa totalité.

Chapitre 10

Modéliser la régulation d'un site d'épissage par les protéines SR

Les travaux expérimentaux réalisés au laboratoire MAEM [15] ont permis de mettre en évidence l'importance des protéines SR dans la régulation de l'épissage alternatif (voir Chapitre 1). [Graveley, 2000] insiste également sur l'impact des protéines SR sur la dynamique de la régulation post-transcriptionnelle par le contrôle du spliceosome. Dans le but de formaliser ces hypothèses, nous avons élaboré un modèle formel. Nous considèrerons dans notre démarche de modélisation que les protéines SR sont affectées à deux fonctions distinctes. Elles peuvent activer ou réprimer l'épissage alternatif à un site donné. Actuellement, l'impact de ces protéines est majoritairement étudié sur les sites accepteurs d'épissage (SA). Nous nous focaliserons donc uniquement sur cette régulation, tout en étant conscient dans la suite des travaux que nous n'avons pas considéré toutes les régulations possibles. Néanmoins, [Graveley, 2000] met en avant l'importance qualitative des seuls sites accepteurs dans la régulation de l'épissage par les protéines SR. Par ailleurs, d'un point de vue expérimental, la complexité de la régulation du processus nécessite de restreindre le domaine d'étude à un site accepteur d'épissage seul. Disposant de ce type de données, nous restreindrons le référentiel du modèle à la même échelle.

Dans ce contexte, une première section présentera le modèle de la régulation du site accepteur A3 de HIV-1 par les protéines SR. Après la formulation des hypothèses biologiques adéquates, il nous sera possible de formaliser mathématiquement le système biologique sous la forme d'un système d'équations différentielles ordinaires. Le modèle ainsi conceptualisé, nous nous emploierons dans une deuxième section à valider qualitativement notre démarche de modélisation. Le comportement qualitatif du modèle dépend fortement des valeurs des paramètres des réactions cinétiques. Le manque d'informations cinétiques à notre disposition confirme alors l'intérêt de la programmation par contraintes hybrides (`hybrid cc`, voir Chapitre 4) pour modéliser ce système.

[15] Laboratoire de Maturation des Arn et Enzymologie Moléculaire

10.1 Formalisation du modèle de régulation du site A3

La formalisation du modèle de régulation résulte de divers travaux et collaborations. Dans ce contexte d'échanges, Delphine Ropers et Christiane Branlant du laboratoire MAEM à Nancy, ont isolé au fur et à mesure de l'élaboration du modèle les hypothèses biologiques à prendre en compte. Par ailleurs, une collaboration avec Hidde de Jong du projet HELIX de l'INRIA Rhône-Alpes, nous a permis d'isoler le formalisme le plus approprié à notre problématique. Le résultat de ces échanges est un modèle formel simple élaboré à partir de connaissances expérimentales.

10.1.1 Connaissances expérimentales concernant le site A3

Dans le cas du génome de HIV-1, la régulation de l'épissage est un processus complexe qui nécessite l'intervention de 4 sites donneurs (SD) et de 8 sites accepteurs (SA), qui permettent de la synthèse d'une quarantaine d'ARN matures [Purcell & Martin, 1993], Cette complexité combinatoire est contrôlée par la régulation de la sélection des sites accepteurs d'épissage ([Caputi *et al.*, 1999*a*, Purcell & Martin, 1993]). Des facteurs protéiques contrôlent cette régulation par le biais de sites de fixation. Nous nous focaliserons ici sur l'étude de la régulation du site accepteur A3 seul. La Figure 10.1 rappelle les connaissances acquises expérimentalement au laboratoire MAEM par Lilia Ayadi, Delphine Ropers et Christiane Branlant sur la régulation du site A3. Ces travaux expérimentaux mettent en évidence plusieurs paramètres qu'il est important de prendre en compte afin de modéliser au mieux le comportement du site A3.

- La structure secondaire : Le site A3 est localisé sur une structure dite tige-boucle qui limite son accessibilité aux différents facteurs d'épissage, ce qui réduit son efficacité.
- La présence de séquences inhibitrices d'épissage : 2 éléments ESS2p et ESS2 répriment le site A3 en fixant respectivement les protéines hnRNP H et hnRNP A1.
- La présence de séquences activatrices d'épissage : L'élément ESEt active l'utilisation du site A3 en fixant les protéines ASF/SF2 et SC35.
- Une compétition pour la fixation des protéines SC35 et hnRNP A1 sur l'élément ESS2 : Tout comme les protéines hnRNP A1, les protéines SC35 sont capables de se fixer sur l'élément ESS2. Une hypothèse est donc qu'une fixation de la protéine SR sur ESS2 empêche la fixation de la protéine hnRNP A1 et vice versa.

L'étude expérimentale a également révélé qu'à la différence de l'élément ESS2, la régulation exercée par l'élément ESS2p ne dépend pas des protéines SR. La protéine SRp40 active également l'utilisation du site A3, mais l'état actuel des connaissances ne permet pas d'inférer sur son mode d'action. De plus, sur 10 protéines SR identifiées chez l'homme, seulement 4 d'entre elles ont été testées expérimentalement sur le site A3. Dans l'optique de formuler un modèle prédictif qui aura pour but de tester les hypothèses issues de l'expérimentation, nous nous focaliserons sur un modèle rendant compte des variations d'efficacité d'épissage au site A3, en fonction des quantités de protéines SR ASF/SF2 et SC35 et de la protéine hnRNP A1. Il est donc pour cela nécessaire de faire abstraction des paramètres de régulation indépendants de ces deux types de protéines comme la

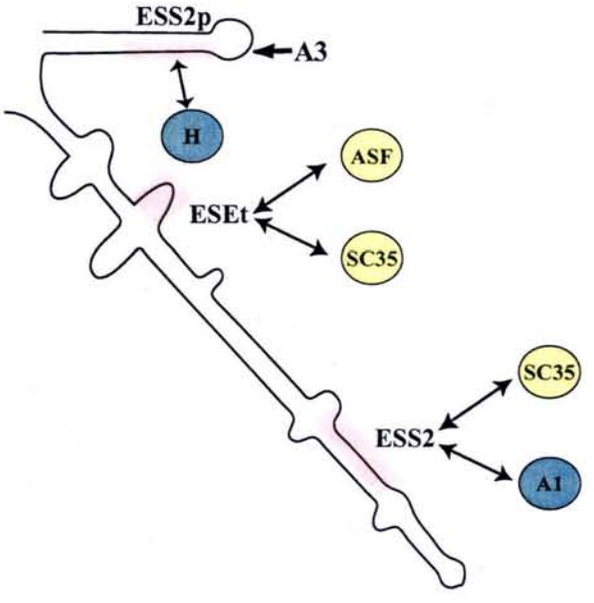

FIG. 10.1 – **Régulation de l'épissage au site A3.** Modèle de régulation proposé à l'issue de l'étude expérimentale réalisée au MAEM par Delphine Ropers, Lilia Ayadi et Christiane Branlant. Les régions roses sur l'ARN correspondent aux sites de fixation des protéines isolées expérimentalement. Les protéines en jaunes correspondent aux facteurs d'activation de l'épissage au site A3 (SC35 et ASF pour la protéine ASF/SF2). Les protéines schématisées en vert correspondent à des facteurs d'inhibition d'épissage au site A3.

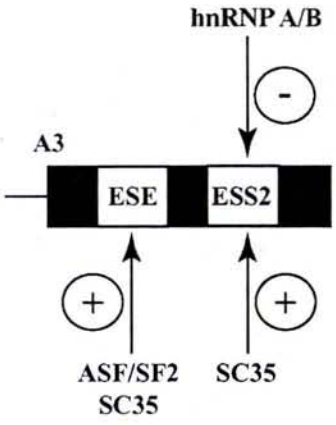

FIG. 10.2 – **Représentation des acteurs de la régulation du site A3.** Les protéines se lient à deux sites de fixation : ESEt qui est un site activateur d'épissage et ESS2 qui est un site inhibiteur d'épissage. ESEt fixe les protéines SR ASF/SF2 et SC35 qui ont un rôle activateur de ce site symbolisé par (+). ESS2 fixe les protéines hnRNP A/B et SC35. hnRNP A/B réprime l'épissage alternatif (-)et SC35 l'active.

concentration de hnRNP H se fixant sur l'élément ESS2p ou la concentration en facteurs spliceosomaux. La quantité de ces derniers facteurs est en effet non limitante pour la cinétique des réactions d'épissage *in vitro* d'après [Audibert *et al.*, 2002]. La restriction du référentiel du modèle nous impose également de négliger dans la modélisation l'effet activateur de la protéine SRp40, dont le mode d'action est jusqu'alors mal connu. Le système simplifié qui en résulte est représenté dans la Figure 10.2.

10.1.2 Hypothèses biologiques

C'est dans ce contexte biologique simplifié que nous représenterons le site A3 où l'épissage peut être réprimé par les protéines hnRNP A/B après fixation sur l'élément ESS2. De récentes études expérimentales effectuées au laboratoire MAEM ([Ropers, 2003]) montrent qu'un autre élément ESEt peut activer l'épissage du site A3 après fixation des protéines ASF/SF2 et SC35. Une hypothèse forte issue de ces expériences concerne alors la fixation ambivalente de la protéine SC35 avec hnRNP A/B sur le site ESS2. De manière plus générale, nous considérerons alors que le ratio des protéines hnRNP A/B sur SR est un facteur prépondérant de l'efficacité d'épissage sur le site A3. Il est donc possible de modéliser la régulation de l'épissage alternatif par les protéines SR dans le contexte restreint du site A3. Il est pour cela nécessaire de considérer les hypothèses biologiques suivantes résumées dans la Figure 10.3 :

- Nous étudions seulement un site d'épissage. Cette hypothèse forte sous-entend que les autres interactions externes au site seront négligées ou assimilées comme des variables forçantes. Ainsi, nous sommes en mesure de considérer le modèle à la même échelle que les données expérimentales en notre disposition. Ces données correspondent à des ratios d'ARN prémessager matures (ARN mature) sur les ARN prémessager natifs ou immatures.

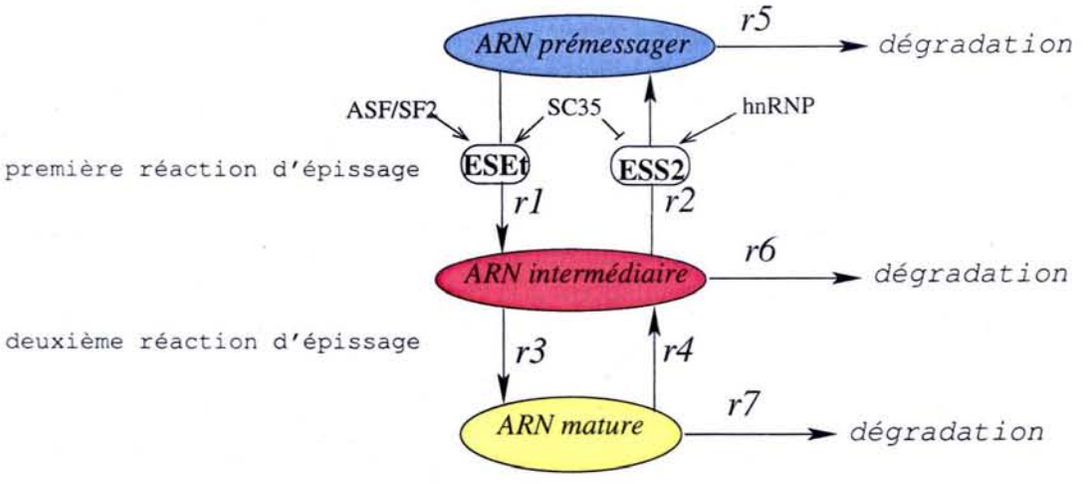

FIG. 10.3 – **Schéma conceptuel de la régulation du site A3.** .

- Nous considérons au cours de ce modèle que l'épissage se déroule en deux réactions (voir Figure 1.4) qui relient les ARN immatures (`rna`), les ARN intermédiaires (`irna`) et ARN matures (`mrna`). Les ARN intermédiaires correspondent aux ARN immatures qui pourront être activés par les protéines. Les ARN matures correspondent ici aux ARN matures et aux ARN en lariat.
- Nous considérons que les concentrations en protéines de régulation sont saturantes. Pour rendre compte des conditions expérimentales, elles seront alors constantes malgré leurs fixations à l'ARN durant la régulation de l'épissage.
- La régulation du site A3 est contrôlée par les sites de fixation des protéines SR ESEt et ESS2. Ces deux sites sont considérés comme cinétiquement indépendants. En effet, d'une part l'élément ESEt est seul responsable de l'activation du site A3 par ASF/SF2, d'autre part, la délétion de l'élément ESS2 n'empêche pas l'activation du site A3 par la protéine fixée à l'élément ESEt et vice versa.
- Les protéines SR ASF/SF2 et SC35 activent la première réaction d'épissage en liant à l'élément ESEt. Une hypothèse de travail de notre modèle est de considérer un phénomène de compensation fonctionnelle entre ces deux protéines. Ainsi, nous considérons que seul l'activation de l'élément ESEt compte quelle que soit la protéine SR qui s'y lie. On considérera alors la somme des protéines SR se liant à l'élément ESEt comme paramètre important.
- Le mode d'inhibition du site A3 par la protéine hnRNP A1 fixée à l'élément ESS2 est mal connu, mais les données expérimentales tendent à confirmer qu'elles empêchent la reconnaissance du site d'épissage par les facteurs spliceosomaux, ce qui bloque le spliceosome, et qui conduit donc à une inhibition de la première et de la deuxième étape d'épissage. Les protéines hnRNP A/B peuvent donc inhiber la première réaction d'épissage par une fixation à l'élément ESS2. Parallèlement, nous considèrerons que SC35 peut aussi se fixer à l'élément ESS2. Mais cette fixation aura comme fonction d'inhiber l'inhibition imposée par hnRNP A/B. Par cette fixation active indirectement l'épissage contrairement à l'élément ESEt où SC35 active directement l'épissage. C'est une hypothèse de conceptualisation forte issue de travaux

expérimentaux que nous allons tester par la modélisation.

- Le manque d'informations actuelles concernant la cinétique de l'épissage nous incite à formuler la cinétique du modèle de manière similaire à la cinétique du modèle de Monod ([Monod, 1950]). Ce dernier modèle considère un modèle à compartiment dont la cinétique est pilotée par un taux de croissance associé à un compartiment. Ce taux est dépendant de la concentration en éléments externes suivant une fonction estimée à partir d'une cinétique de Michaelis-Menten. Nous assimilerons les mêmes hypothèses de formalisation adaptées à la régulation particulière du site A3 par les protéines SR saturantes. Cette hypothèse reste grossière d'un point de vue biochimique mais permet de modéliser le système biologique avec une certaine flexibilité pour compenser notre manque de connaissances.

Ces hypothèses ont pour but de finaliser un travail expérimental. Les travaux ont en effet mis en évidence diverses hypothèses de régulation de l'épissage qui sont difficiles à vérifier rapidement. La modélisation permet ici de tester celles-ci rapidement afin de justifier ou non de recherches expérimentales plus approfondies.

10.1.3 Modèle mathématique

La modélisation du processus biologique est généralement obtenue par une analogie avec un modèle de système existant possédant des caractéristiques communes dont les lois sont mieux connues. Par exemple, la comparaison du coeur avec une pompe hydraulique a prouvé que la circulation sanguine relevait des lois de l'hydraulique. Dans notre cas, aucun modèle mathématique de l'épissage n'a encore été établi à cette échelle. Il a donc fallu rechercher un système qui s'y apparente le mieux. Les hypothèses biologiques peuvent être représentées par un système d'ODE inspiré du modèle de Monod ([Monod, 1950]). Dans ce modèle, les taux de croissance dépendent des concentrations externes qui sont les conditions limitantes du système. La dynamique du système est contrôlée par des cinétiques de type Michaeliennes. Dans notre cas, les facteurs limitants seront les concentrations en protéines régulatrices que l'on considérera dans un premier temps comme constantes. Cette approche est généralement utilisée avec succès dans les modèles écologiques et elle est particulièrement bien adaptée à la description de systèmes qui sont que partiellement connus.

Nous pouvons décrire la cinétique du système associée aux hypothèses biologiques avec 7 réactions déjà représentées dans la Figure 10.3. Les symboles que nous utiliseront par la suite seront répertoriés dans le Tableau 10.1.

La réaction r_1 représente la transformation de l'ARN prémessager (`rna`) en ARN intermédiaire (`irna`). Cette réaction nécessite une complémentarité entre les protéines ASF/SF2 et SC35 sur l'élément ESEt. Suivant cette hypothèse nous considérons alors que la somme des deux protéines activatrices SR est un paramètre important de l'activation de l'élément ESEt. Nous représentons ce taux de réaction par une fonction de type Michaelis-Menten dépendante de la quantité d'ARN immature, et contrôlée par la somme des protéines ASF/SF2 et SC35. La forme générique de la fonction de Michaelis-Menten (voir [Murray, 2002] pour détails) est :

TAB. 10.1 – **Symboles et unités des variables et paramètres du modèle local de régulation d'épissage du site A3**

Symboles	Variables et Paramètres	Unités
rna	ARN Immature	μM
$irna$	ARN Intermediaire	μM
$mrna$	ARN Mature	μM
ASF	Proteine ASF/SF2	μM
SC	Proteine SC35	μM
R	Proteine hnRNP A/B	μM
φ_{ESEt}	Taux de fixation maximale pour l'élément ESE	s^{-1}
φ_R	Taux de fixation maximale de hnRNP A/B	s^{-1}
k_{ESEt}	Coefficient de demi-saturation pour l'élément ESE	μM
k_{SC}	Coefficient de demi-saturation de SC35	μM
k_R	Coefficient de demi-saturation de hnRNP A/B	μM
κ	Constante de réaction	s^{-1}
κ'	Constante de réaction	s^{-1}
λ	Coefficient de dégradation	s^{-1}

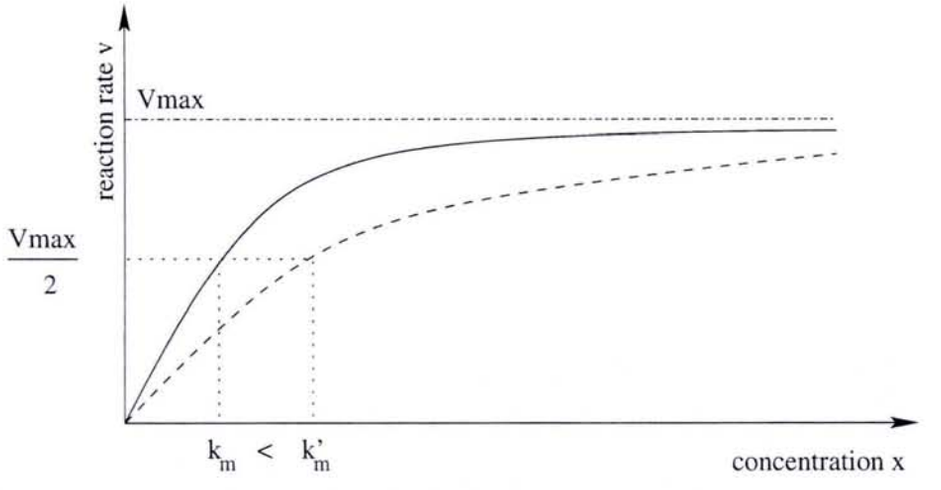

FIG. 10.4 – **Représentation de la cinétique de Michaelis-Menten.**

$$v = \frac{v_{\max} \cdot x}{k_m + x}$$

La courbe représentant la relation entre v et x est donnée dans la Figure 10.4. Dans ce cas générique, $v_{\max}$ est le taux de croissance maximale et la constante de Michaelis k_m est la valeur pour laquelle v correspond à la moitié de sa valeur maximale. Ainsi pour $x = ASF + SC$ et les notations du Tableau 10.1, on obtient :

$$r_1 = \frac{\varphi_{ESE} \cdot (ASF + SC)}{k_{ESE} + (ASF + SC)}\, rna$$

La réaction r_2 représente la transformation de l'ARN intermédiaire en ARN prémessager. Elle permet de mettre en relation les fonctions antagonistes des protéines hnRNP A/B et SC35 après fixation sur l'élément ESS2. Dans ce cas, on représente une fixation de la protéine hnRNP A/B à l'élément ESS2, inhibée par la quantité de protéine SC35 ([Voit, 2000]). La protéine SC35 seul n'active pas l'épissage. Elle active qu'indirectement en empêchant l'inhibition. Nous utiliserons alors le même formalisme que précédemment. Toutefois, nous sommes dans ce cas dans un cadre d'inhibition compétitive (voir [Heinrich & Schuster, 1996] pour détails) entre deux éléments x et y. La forme générique de cette fonction est de la forme :

$$v = \frac{v_{\max} \cdot x}{k_m(1 + \frac{y}{k_y}) + x}$$

Le facteur $(1 + \frac{y}{k_y})$ augmente la valeur de k_m, voir la Fig.10.4 pour illustration. Avec $x = R$ et $y = SC$ et la notation du tableau 10.1, on obtient :

$$r_2 = \frac{\varphi_R \cdot R}{k_R(1 + \frac{SC}{k_{SC}}) + R}\, irna$$

La réaction r_3 représente la transformation de l'ARN intermédiaire en ARN mature (`mrna`). Nous considérons pour cette réaction une cinétique de premier ordre avec une constante cinétique de réaction κ. De manière similaire, r_4 est une réaction qui représente la cinétique qui lie les ARN matures aux ARN intermédiaires.

$$r_3 = \kappa \cdot irna, \qquad r_4 = \kappa' \cdot mrna.$$

r_5, r_6 et r_7 représentent les réactions de dégradation respectivement pour les ARN immatures, intermédiaires et matures. On considère alors à ce niveau que les acides nucléiques se dégradent avec un taux similaire λ :

$$r_5 = \lambda \cdot rna, \quad r_6 = \lambda \cdot irna, \quad r_7 = \lambda \cdot mrna.$$

Nous pouvons alors formaliser le processus de régulation d'épissage au site A3 par le système d'équations différentielles (ODE) (voir Figure 10.3 pour illustration).

$$\begin{aligned}
\frac{d(rna)}{dt} &= r_2 - r_1 - r_5, \\
\frac{d(irna)}{dt} &= r_1 + r_4 - r_2 - r_3 - r_6, \\
\frac{d(mrna)}{dt} &= r_3 - r_4 - r_7,
\end{aligned}$$

ce qui correspond formellement à :

$$\begin{aligned}
\frac{d(rna)}{dt} &= \frac{\varphi_R \cdot R}{k_R(1+\frac{SC}{k_{SC}})+R} irna - \frac{\varphi_{ESE} \cdot (ASF+SC)}{k_{ESE}+(ASF+SC)} rna - \lambda \cdot rna, \\
\frac{d(irna)}{dt} &= \frac{\varphi_{ESE} \cdot (ASF+SC)}{k_{ESE}+(ASF+SC)} rna - \frac{\varphi_R \cdot R}{k_R(1+\frac{SC}{k_{SC}})+R} irna \\
&\quad -\kappa \cdot irna + \kappa' \cdot mrna - \lambda \cdot irna, \\
\frac{d(mrna)}{dt} &= \kappa \cdot irna - \kappa' \cdot mrna - \lambda \cdot mrna.
\end{aligned}$$

Nous supposons ici le même taux de dégradation pour les trois acides nucléiques. De la même façon, nous supposons l'absence de dégradation de l'ensemble des protéines SR. Cette hypothèse forte se justifie par la cinétique faible des protéines compte tenu de la quantité saturante présente dans le milieu de réaction.

10.2 Etude du comportement qualitatif

Le modèle mathématique de la régulation du site A3 peut être directement formalisé dans le langage de programmation par contraintes `Hybrid cc`, déjà présenté dans la Section 4.4.1. Il est alors possible d'implémenter le modèle dans un programme `hcc` pour l'exécuter. Néanmoins, malgré une cohérence avec les résultats expérimentaux, cette simulation du modèle ne constitue pas une preuve suffisante de sa validation. En effet, une simulation numérique reste seulement un échantillon du comportement du système parmi un vaste ensemble de possibilités. Pour cette raison, nous considérons comme essentiel de valider le modèle avant toutes simulations.

Les données biologiques à notre disposition ne sont pas en nombre suffisant pour pouvoir générer un modèle prédictif quantitativement. Un tel modèle nécessite en effet de nombreuses données pour identifier les paramètres du modèle puis le valider quantitativement. Dès lors, nous nous focaliserons sur une validation qualitative de notre approche qui se situe en amont de la validation quantitative. Cette démarche s'intéresse principalement aux signes des variations du système plutôt qu'à l'ordre de grandeur de ces variations. Il existe déjà de nombreuses démarches théoriques qui effectuent une modélisation qualitative ([Thomas *et al.*, 1995, Jong *et al.*, 2001, Bernard & Gouzé, 1999]), comme nous l'avons déjà mentionné dans la Section 4.3.

Appliqués à divers systèmes biologiques, ces techniques permettent cependant seulement de ne pas invalider les modèles correspondants. En effet, si d'un point de vue théorique *un modèle biologique ne peut être validé*, il est par ailleurs possible de *prouver que le modèle ne peut pas être invalidé* dans l'état. Dans le souci de simplifier la terminologie, nous considérerons la non-invalidation comme une validation, tout en gardant à l'esprit le caractère éphémère de ce terme. Pour valider notre approche, nous utiliserons quelques techniques issues de l'étude qualitative de modèles afin d'analyser le comportement du modèle de régulation du site A3. Cette étude est le fruit d'une collaboration avec Olivier Bernard du projet COMORE de l'INRIA Sophia-Antipolis.

10.2.1 Variables d'observations : indicateurs du comportement qualitatif

Une des premières approches qualitatives possibles est de comparer le modèle formel avec les résultats expérimentaux. L'activité biologique d'un site d'épissage après expérimentations se mesure avec le rapport de protéines produites après épissage sur la quantité d'ARN insufflée dans l'expérience. Ce rapport permet de quantifier qualitativement le rendement d'épissage sur un site donné après stimulation avec le ratio :

$$\textit{efficacité d'épissage} = \frac{\textit{Protéine}}{\textit{ARNimmature}} \approx \frac{\textit{ARN mature}}{\textit{ARN immature}}$$

Cette efficacité représente le résultat du processus biologique. Il est donc nécessaire que le modèle formel représente correctement les variations de l'efficacité. Le système mathématique converge vers un unique état d'équilibre stable qui est la dégradation totale de toutes les variables. Les variables indiquant les concentrations en ARN ne possèdent donc pas de valeurs d'équilibre positives. Néanmoins, nous considérons qu'au cours du processus, les réactions d'épissage atteignent rapidement un équilibre. Cette hypothèse se justifie biologiquement si l'on considère que les fixations des protéines régulatrices sont rapides par rapport aux processus de régulation dans sa globalité. Cette hypothèse permet de décrire un comportement du système à l'équilibre de réactions qui se caractérise par :

$$r_1 = r_2, \qquad r_3 = r_4,$$

ce qui est équivalent à :

$$\frac{\varphi_{ESE}(ASF + SC)}{k_{ESE} + (ASF + SC)}\, rna \;=\; \frac{\varphi_R \cdot R}{k_R(1 + \frac{SC}{k_{SC}}) + R}\, irna, \qquad \kappa \cdot irna \;=\; \kappa' \cdot mrna.$$

Si nous définissons l'*efficacité d'épissage* (*efficiency*) telle qu'elle est définie par les expériences par :

$$\textit{efficiency}(t) = \frac{mrna(t)}{rna(t)}$$

Nous obtenons la formule suivante qui décrit l'efficacité d'épissage à son équilibre :

$$\textit{efficiency}_{eq} = \frac{\kappa \cdot \varphi_{ESE}(ASF + SC)(k_R \cdot k_{SC} + k_R \cdot SC + R \cdot k_{SC})}{\kappa'(k_{ESE} + ASF + SC) \cdot \varphi_R \cdot R \cdot k_{SC}}$$

TAB. 10.2 – **Efficacité d'épissage des constructions autour du site A3 après stimulations par les protéines SR** (données du laboratoire MAEM). La stimulation s'effectue avec un ajout d'une quantité variable de protéines SR dans le sérum cellulaire (contenant initialement des protéines SR en concentration naturelle).

$0\ \mu M$	SC35 $(+0,2\mu M)$	SC35 $(+0,4\mu M)$	ASF/SF2 $(+0,2\mu M)$	ASF/SF2 $(+0,4\mu M)$
0,2	5,54	24	0,41	0,62
0,44	7,74	17	1,16	1,12

Malgré l'absence d'équilibre pour les variables d'états du modèle autre que la dégradation totale, il existe un équilibre pour la variable d'observation qui est l'efficacité d'épissage. La variable d'observation n'est pas directement contrôlée par les contraintes du système comme les variables d'état. C'est une variable qui rend compte de l'état d'un système sans prendre part à la régulation de celui-ci. Les observateurs peuvent être directement liés aux variables d'état. On fait alors ici l'hypothèse que les valeurs mesurées de l'efficacité d'épissage correspondent à des valeurs obtenues à l'équilibre du système décrit précédemment. Il est donc intéressant d'analyser cette variable qui rend compte de l'état du système et qui correspond également à une donnée expérimentale. C'est donc une manière indirecte d'analyser les propriétés d'un modèle formel qui n'ont pas été directement insufflées par la régulation des variables d'état.

Suivant la formule de l'efficacité donnée par le modèle, le comportement qualitatif est le suivant (voir Figure 10.5 pour illustration) :

- l'efficacité d'épissage est une fonction croissante de la quantité de protéines activatrices d'épissage comme SC et ASF qui correspondent respectivement aux concentrations de SC35 et ASF/SF2.
- l'efficacité d'épissage est une fonction décroissante de la variable R. L'efficacité décroît donc pour une augmentation de la concentration de protéine hnRNP A/B.

Les résultats expérimentaux montrent quant à eux le comportement représenté dans le tableau 10.2 qui se résume par :

- le rapport expérimental $(mrna/rna)_{eq}$ à l'équilibre de réaction augmente pour une augmentation de la concentration des protéines activatrices.
- le rapport expérimental $(mrna/rna)_{eq}$ à l'équilibre de réaction décroît lors d'une augmentation de la concentration de protéines inhibitrices d'épissage.

Le comportement de la variable observateur est encourageant. Le comportement qualitatif est cohérent avec les résultats expérimentaux à disposition. L'extraction de cette variable du modèle possède plusieurs avantages. Elle est premièrement utile pour identifier quantitativement les paramètres avec les donnés expérimentales. Cette étape peut en effet se résumer à une optimisation de fonction. Par ailleurs, l'efficacité d'épissage étant une variable d'observation, elle peut rendre compte de l'état du système de régulation au niveau du site A3. Incorporer cette fonction dans un modèle de régulation d'épissage à plus grande échelle est assimilé à introduire une abstraction du comportement du site A3 à l'équilibre. En effectuant cette opération, nous formulons l'hypothèse sous-jacente que la phase de transition qui précède la phase d'équilibre est négligeable. La variable

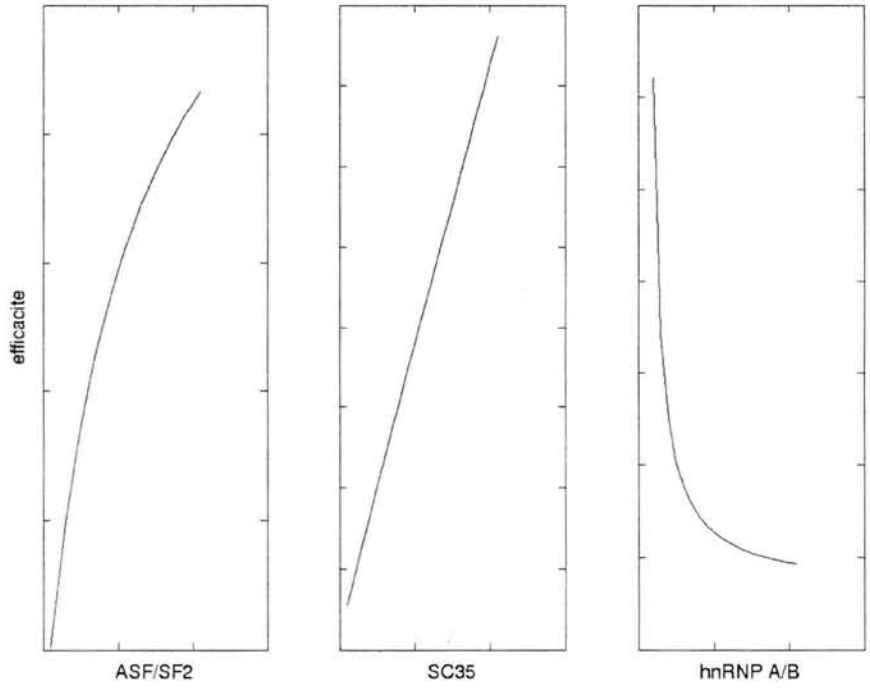

FIG. 10.5 – **Comportement qualitatif de l'efficacité d'épissage au site A3.** Les trois graphiques représentent le comportement de la variable d'observation *efficacité* pour une augmentation des trois protéines régulatrices de l'épissage au site A3. On observe une augmentation de l'observateur pour une augmentation de ASF/SF2 et SC35 et une diminution de la valeur de l'observateur pour une diminution de la protéine hnRNP A/B.

d'observation est donc une candidate idéale à une *time-scale abstraction* [Kuipers, 1988].

Néanmoins, pour utiliser cette hypothèse, il est important de ne pas valider uniquement à l'état d'équilibre avec l'analyse d'une variable d'observation. Il est nécessaire de conforter les résultats préliminaires de validation par d'autres démarches qui puissent prendre en compte notamment des étapes de transition vers l'équilibre de réactions.

10.2.2 Caractéristiques qualitatives du système mathématique

La validation qualitative du modèle par les observateurs est encourageante mais n'est pas suffisante pour valider ou invalider le modèle. Comme nous venons de le mentionner, nous n'avons pas tenu compte de l'état transitoire du modèle. Même si d'après nos hypothèses, cette étape n'est pas prépondérante dans le processus d'épissage, elle peut être la source d'informations supplémentaires concernant le système biologique. Nous utiliserons pour cela une nouvelle méthodologie de validation qualitative. Celle-ci doit être en mesure d'extraire les propriétés du système dans les différents états dans lesquels il peut se trouver. On procède alors à une analyse du modèle mathématique dont le but est d'extraire des propriétés qui doivent être vérifiées expérimentalement pour valider le système et d'autres qui deviendront de nouvelles hypothèses biologiques à tester.

Dans cette partie, nous proposons donc d'analyser qualitativement les principales propriétés mathématiques du système d'équations différentielles décrites précédemment afin de valider le comportement du modèle quelles que soient les valeurs numériques des paramètres qui interviennent au cours de la simulation. Nous nous inspirerons des travaux de [Bernard & Gouzé, 1998, Bernard & Gouzé, 2002]. L'avantage de cette approche est de pouvoir raisonner qualitativement sur le système dans son ensemble. Cette approche extrait des propriétés du système même si celui-ci se situe dans une phase transitoire, contrairement à l'utilisation des observateurs que nous avons mentionnés dans la précédente section.

Afin d'analyser le comportement qualitatif du modèle, nous supposons que les protéines régulatrices ne varient pas en concentration. Cela se justifie par le fait qu'elles soient en quantité suffisamment importante pour être en conditions saturantes. Dans le cas d'une concentration en protéine constante, les signes de la matrice Jacobienne du système mathématique sont les suivants :

$$J = \begin{pmatrix} - & + & 0 \\ + & - & + \\ 0 & + & - \end{pmatrix}$$

Le modèle est dit linéaire et coopératif. : Les signes des dérivées qui sont représentés dans cette matrice Jacobienne sont nuls ou fixés. Par définition, si les signes extra-diagonaux de la matrice sont positifs ou nuls, le système mathématique correspondant est dit coopératif à interactions monotones [Bernard & Gouzé, 1998]. Ces conditions sont satisfaites sur notre système. On est alors en mesure d'appliquer des méthodes de raisonnement qualitatif qui permettent de prédire le comportement du modèle. Il est pour cela nécessaire de représenter le comportement par un graphe de transitions entre les différents états qualitatifs que va adopter le système.

Protocole générique d'analyse qualitative

L'analyse qualitative se base sur les signes des dérivées secondes d'un système d'ODE. Dans le cadre d'un système coopératif, il est possible de décrire les différentes variations des signes des dérivées secondes du système ainsi que les différents états qualitatifs que va adopter le système avant de converger vers son état final.

Ainsi pour chaque ODE du système, on peut lui associer un signe (+) ou (-) qui correspond au signe de la dérivée seconde. On est alors en mesure de décomposer un système en différents états qualitatifs suivant les signes des dérivées secondes. Pour un système à deux variables, il existe 4 états qualitatifs possibles (voir Figure 10.6 pour illustration). On est capable d'associer un comportement qualitatif à chaque état.

- une dérivée seconde positive d'une variable x caractérise une croissance qualitative. On le représente par l'état qualitatif (+).
- une dérivée seconde négative d'une variable x caractérise une décroissance qualitative. On le représente par l'état qualitatif (-).

La transition entre deux états est elle aussi associée à un comportement qualitatif particulier :

$$\begin{array}{lccc} \textit{Cas (1)} & (+) & \rightarrow M \rightarrow & (-) \\ \textit{Cas (2)} & (-) & \rightarrow m \rightarrow & (+) \end{array}$$

Dans le cas (1), le fait de passer d'une croissance d'une variable (production) à une décroissance (consommation), signifie que la variable connaît qualitativement un pic de croissance que l'on caractérise par **M**. Inversement, dans le cas (2), la variable connaît un minimum de croissance que l'on caractérise par **m**. Ces éléments qualitatifs sont facilement mesurables expérimentalement. La validation qualitative revient alors à observer un comportement donné comme un enchaînement de pic (M) et de creux (m) d'une variable au cours d'un processus biologique et de le mettre en relation avec les expériences.

Graphe de transition d'une variable du modèle

Selon la méthodologie énoncée précédemment, on analyse le comportement qualitatif d'une variable. Il est possible d'effectuer cette approche sur toutes les variables d'un système d'ODE. L'analyse qualitative correspond alors à relier entre eux les différents comportements qualitatifs en fonction des propriétés du système mathématique. Il est alors possible de faire un graphe entre les états qualitatifs que nous avons isolés en indiquant le sens des transitions possibles. Nous proposons de l'illustrer sur un système de deux ODE représentant le comportement des variables x_1 et x_2. Les transitions peuvent alors être représentées comme dans la Figure 10.6.

Dans la représentation du graphe, M_x représente le passage par un pic de maximum pour la variable x, et m_x représente le passage par un minimum pour la variable x.

Vers un graphe de transitions qualitatives du modèle de régulation d'épissage au site A3

On effectue la démarche précédente sur un système à trois ODE. Il est ainsi possible de représenter les variations qualitatives du système dans sa totalité. Pour un système

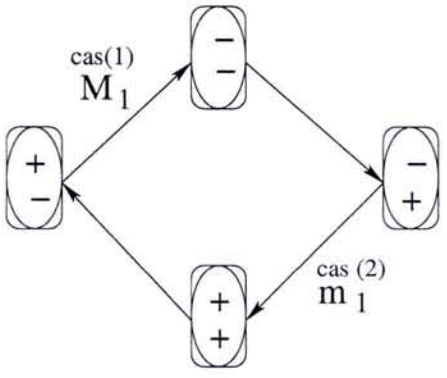

FIG. 10.6 – **Graphe de transition partiel pour un système à deux variables.** M_1 représente le passage par un pic de maximum pour la variable x_1. m_1 représente le passage par un minimum pour la variable x_1.

mathématique composé de trois ODE avec trois variables, on peut décrire le système biologique avec 8 états qualitatifs différents.

Nous avons appliqué la méthode d'analyse qualitative sur le modèle de régulation du site A3. Avec le précédent protocole, en décomposant variable par variable, il est possible d'analyser toutes les transitions possibles entre les états qualitatifs qui décrivent les variations des différents ARN. Par soucis de clarté dans le graphe, nous simplifions le système

$$\begin{pmatrix} rna \\ irna \\ mrna \end{pmatrix} \quad \text{par} \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

La Figure 10.7 représente ainsi les transitions entre les états qualitatifs avec des notations des extrêmes différentes. M_1 correspond alors à un maximum de production pour *rna*. m_1 correspond à un minimum de production pour *rna*. La notation est similaire pour M_2 et m_2 qui sont associés à *irna* et pour M_3 et m_3 à *mrna*. Par cette approche, on décrit ainsi tous les comportements qualitatifs possibles de notre modèle.

Le système mathématique possède d'autres propriétés que celles représentées dans le graphe de transition, avec notamment la conservation de la matière et le taux de dégradation des ARN $\lambda \geq 0$ qui est le même quel que soit l'état des ARN.

$$v = rna + irna + mrna \qquad \text{avec} \qquad \dot{v} = -\lambda v$$

Dans ce contexte, $\dot{v} < 0$ car pour un système biologique : $v > 0$. Il existe donc une partie du graphe de transition qui est non atteignable (voir Figure 10.7). Le graphe confirme alors un unique état final du modèle biologique. En effet, compte tenu du domaine inaccessible, toutes les transitions convergent vers l'état d'équilibre :

$$\begin{pmatrix} - \\ - \\ - \end{pmatrix}$$

qui correspond à une dégradation totale de tous les ARN. Afin de valider qualitativement le modèle, on peut s'intéresser à un scénario précis. Les conditions biologiques dans lesquelles

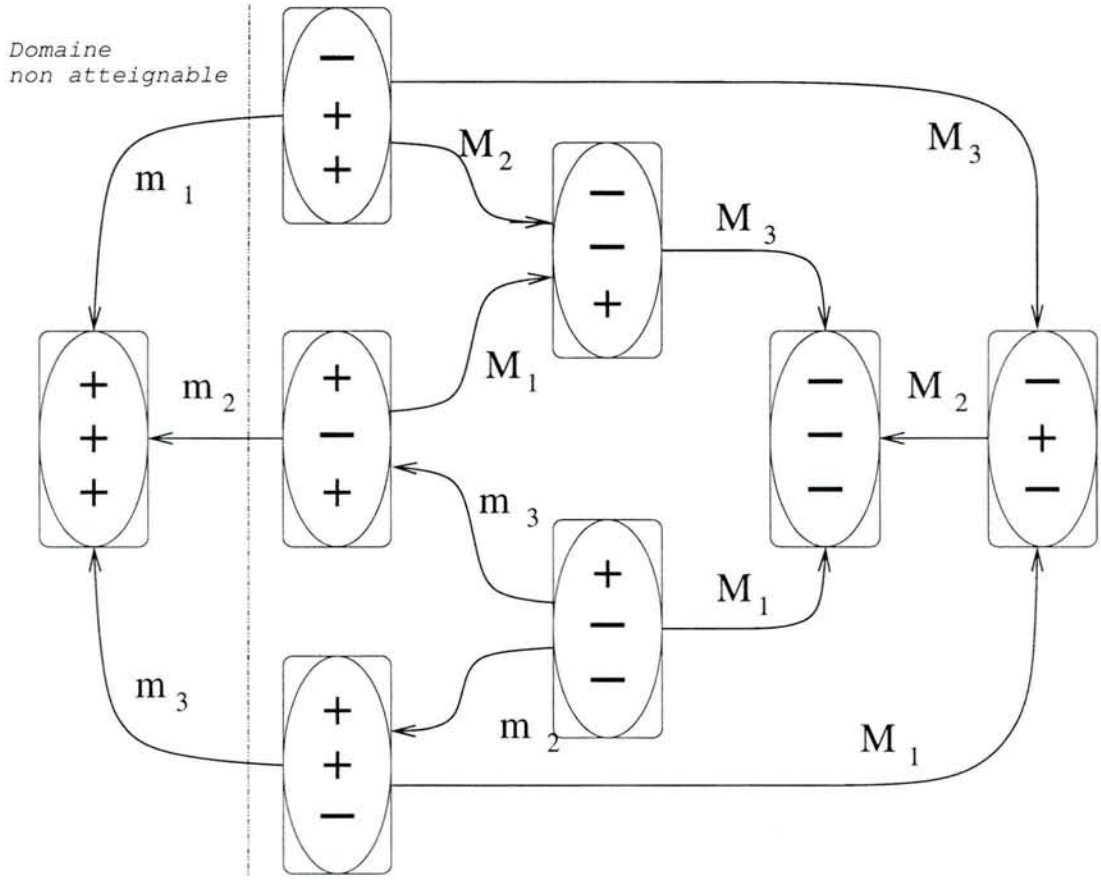

FIG. 10.7 – **Graphe de transition du comportement qualitatif du modèle.**

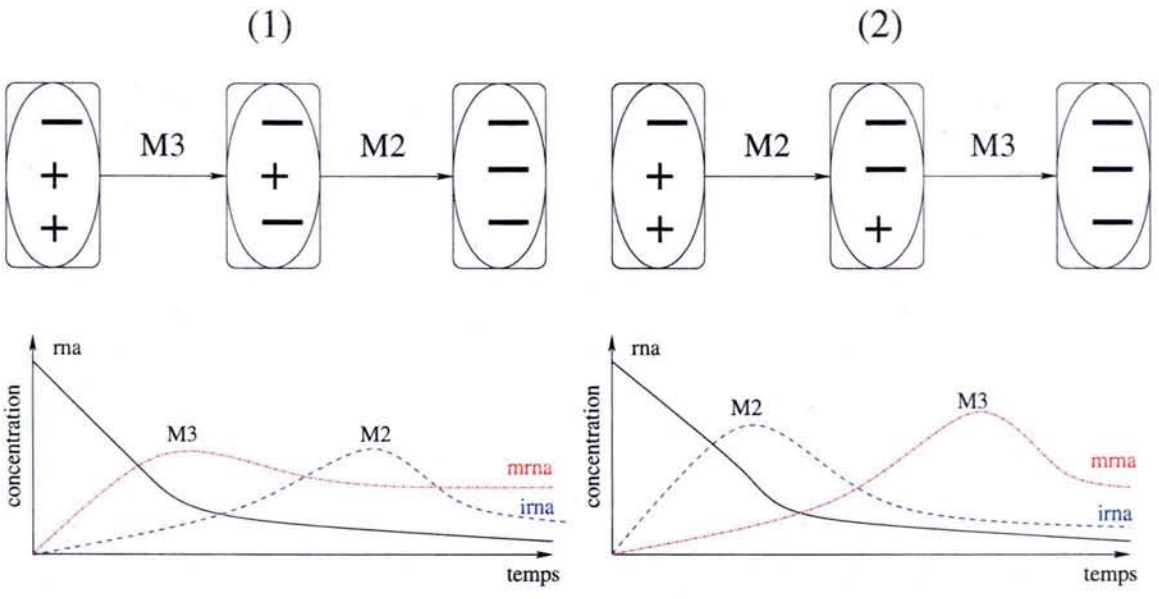

FIG. 10.8 – **Comportement qualitatif du modèle de régulation au site A3.** Pour une initialisation correspondant à une concentration d'ARN immatures (*rna*) seul, il existe deux comportements qualitatifs possibles. (1) correspond à un pic de production (M3) de ARN mature (*mrna*) suivi d'un pic (M2) d'ARN intermédiaires (*irna*). (2) correspond à un pic de production (M2) d'ARN intermédiaires (*irna*) suivi d'un pic (M3) de ARN matures (*mrna*).

on se place au début d'expériences correspondent à la présence d'ARN immature seul. Ce qui correspond au vecteur :

$$x_0 = \begin{pmatrix} rna_0 \\ 0 \\ 0 \end{pmatrix} \quad \text{soit l'état qualitatif initial} \quad \dot{x_0} = \begin{pmatrix} - \\ + \\ + \end{pmatrix}$$

Il existe dans cette position peu de transitions qualitatives parmi celles proposées par le système mathématique qui correspondent à la réalité expérimentale. Ces transitions ainsi que les comportements qualitatifs associés sont représentées dans la Figure 10.8 par un pic de production d'une variable.

Le modèle formel décrit donc deux comportements cinétiques possibles correspondant aux hypothèses et aux propriétés biologiques que nous avons insufflées initialement. Afin de valider notre approche, il suffit de faire des expériences cinétiques afin d'observer si au cours de l'épissage, quel que soit le type de régulation par les protéines SR, on observe un pic de production d'ARN mature suivi d'ARN immature (cas (1)) ou un pic de production d'ARN immature suivi d'un pic d'ARN mature (cas (2)). Actuellement, les résultats expérimentaux tendent à valider le modèle par la transition qualitative (2). Mais de nouvelles expérimentations sont nécessaires pour valider notre approche. Ainsi quelles que soient les valeurs numériques de production d'ARN, seul le comportement qualitatif des expériences sera suffisant pour valider le modèle formel.

10.3 Conclusions

La régulation de l'épissage alternatif est un processus complexe qui nécessite des expérimentations biologiques très chères en temps. C'est une des raisons qui justifie le manque de données expérimentales actuelles. Les expériences mettent principalement en valeur le comportement qualitatif du système biologique. Cette propriété expérimentale ajoutée à l'incertitude des paramètres quantitatifs nous amène naturellement à formuler un modèle qualitatif de la régulation de l'épissage alternatif. L'utilisation de la programmation concurrente par contraintes permet notamment d'avoir cette démarche. Nous avons développé dans ce chapitre une validation de notre modèle qui s'apparente aux approches standard des biomathématiques. *Hybrid cc* permet de les utiliser sur notre problème biologique. Nous envisageons alors de généraliser ces intéressantes propriétés qualitatives grâce à `hcc` sur un système de régulation plus complexe. Ce modèle ainsi que son implémentation en *Hybrid cc* à fait l'objet de diverses publications comme [Eveillard *et al.*, 2002, Eveillard *et al.*, 2003] et [Eveillard *et al.*, 2004]. Dans le contexte de modélisation qualitative, la formulation mathématique nous permet de raisonner sur un système sans données numériques fiables. Cette capacité à raisonner permet d'extraire de nouvelles hypothèses biologiques qui devront être vérifiées par les expériences futures, comme la succession d'évènements qualitatifs, ou le comportement de variables d'observation face à une stimulation par les protéines SR. L'exemple de ce modèle illustre alors les avantages de modéliser un système biologique malgré les incertitudes expérimentales afin de tester diverses hypothèses biologiques. Cette démarche peut permettre ainsi au biologiste de restreindre le protocole expérimental pour se concentrer particulièrement sur le test des hypothèses les plus plausibles. Le modèle qualitatif devient alors un outil de recherche exploratoire *in silico* en biologie et ce dans un contexte dynamique difficilement accessible par les expériences.

Notre modèle partiellement validé qualitativement nécessite néanmoins d'autres expériences pour confirmer nos premiers résultats encourageants. En effet, les premiers résultats montrent que le modèle n'est pas faux. Il reste à généraliser cette conclusion pour des conditions expérimentales différentes. De nouvelles expériences pourraient également nous permettre d'envisager une validation quantitative. Cette optique nous permettrait, après identification statistique des paramètres du modèle, de proposer un modèle prédictif de la régulation de l'épissage par les protéines SR.

Cette dernière remarque replace l'approche de modélisation formelle dans un protocole biologique. La survie d'un modèle plutôt qu'un autre dépend des efforts des biologistes expérimentaux pour tester celui-ci. La formulation des hypothèses biologiques, la validation ou plus précisément les critères nécessaires à la validation d'un modèle sont directement liés aux expériences. Le type de modélisation que nous avons présenté, apparaît alors comme un protocole pré-expérimental dans une démarche globale de compréhension d'un système biologique. Dans cette optique diverses perspectives sont possibles. Concernant la régulation de l'épissage alternatif, il pourrait être pertinent de formuler divers modèles génériques qui prennent en compte les différentes inhibitions et activations d'épissage. Il serait alors possible de proposer aux biologistes de piloter les interactions de leur choix afin d'en mesurer formellement les propriétés dynamiques et qualitatives. Ce test formel serait alors une étape supplémentaire et simplificatrice dans un protocole déjà très complexe.

Chapitre 11

Modélisation intégrative des régulations de l'épissage alternatif

La modélisation que nous venons de présenter possède un comportement qualitatif cohérent avec les données expérimentales à notre disposition. L'analyse qualitative qui permet de valider notre approche repose principalement sur le comportement mathématique du système continu qui représente le système biologique. L'intérêt de la modélisation repose ici principalement sur le fait de pouvoir *in silico* comparer le résultat de l'agencement des hypothèses biologiques avec une réalité expérimentale.

Néanmoins la compréhension de la mécanique associée à un site seul ne permet pas d'aborder toute la complexité de l'épissage alternatif dans sa globalité. En effet, un des enjeux biologiques de l'étude de l'épissage alternatif est de démontrer le lien entre la régulation du processus moléculaire avec la régulation d'un organisme vivant. Il est pour cela nécessaire de caractériser les conséquences de l'épissage sur les échelles biologiques supérieures. L'hypothèse d'une régulation d'épissage comme élément prépondérant dans la régulation de HIV-1 est alors une justification suffisante pour étudier ce processus biologique. C'est dans cette optique que des travaux expérimentaux ont récemment abordé la complexité de la régulation de l'épissage sous un angle multi-site. Mais les hypothèses biologiques restent difficilement formulables. Dans ce cas, la modélisation peut aller au-delà du simple procédé de vérification en proposant un test préliminaire à toutes expériences des hypothèses empiriques qualitatives. C'est cet aspect de la modélisation formelle du vivant que nous allons illustrer dans ce chapitre avec une extension du modèle de régulation existant vers les échelles biologiques supérieures. Notre approche consiste donc à étendre la modélisation de la régulation d'épissage sur plusieurs sites avant d'injecter cette extension dans le cycle de vie de HIV-1.

Les expériences biologiques ne permettent pas encore d'appréhender complètement la régulation de l'épissage sur différentes échelles. Nous sommes alors face à certaines incertitudes que doit gérer le modèle sur plusieurs échelles. La capacité à étendre notre modèle sur des échelles biologiques différentes est directement liée aux formalismes de modélisation (voir Figure 4.1). Afin de modéliser sur plusieurs échelles, il est donc important dans un premier temps d'isoler un formalisme adéquat à une modélisation mathématique mais également capable de gérer certaines incertitudes expérimentales. Nous aborderons cette démarche dans la section 11.1. Le formalisme permettra également de raisonner sur le

système multi-échelle afin de le valider qualitativement comme nous le montrerons dans la section 11.2. Ces approches théoriques consolidées, il nous sera possible de justifier formellement de l'enjeu de la régulation de l'épissage alternatif sur le cycle de vie de HIV-1 dans la section 11.3.

11.1 Utilisation des contraintes hybrides pour modéliser sur plusieurs échelles

Il existe différents formalismes pour modéliser le vivant comme nous l'avons mentionné dans le Chapitre 4. Certains possèdent la capacité d'intégrer des incertitudes expérimentales. C'est le cas des réseaux qualitatifs [Jong, 2004] qui permettent de raisonner en fonction de valeurs seuils. L'incertitude concerne alors les valeurs numériques des paramètres. L'approche par programmation concurrente avec contraintes hybrides permet également de gérer une incertitude comme nous l'avons vu dans la section 4.4. Avec ce paradigme de programmation, il est possible de considérer des incertitudes numériques au niveau des valeurs des seuils des variables comme dans le cas des réseaux qualitatifs, mais il est également possible de considérer des comportements par défaut. Cette dernière propriété est un avantage non négligeable afin de modéliser le système sur plusieurs échelles. En effet, les connaissances actuelles des systèmes biologiques sur plusieurs échelles permettent seulement d'appréhender la tendance du comportement global qui peut être alors considérée comme le comportement par défaut. Par ailleurs, `Hybrid cc` permet de formaliser les systèmes continus avec des équations différentielles ordinaires qu'il interprète comme des contraintes continues. `Hybrid cc` semble donc être un formalisme adéquat à la modélisation multi-échelle.

L'article [Eveillard *et al.*, 2004] résume les propriétés de `Hybrid cc` dans le cadre de la modélisation de systèmes biologiques telles que nous l'avons déjà mentionné dans la section 4.4.1. Ces propriétés sont ensuite mises à profit dans le cadre de la modélisation de la régulation de l'épissage alternatif par les protéines SR au site A3 de HIV-1. La validation qualitative est brièvement mentionnée. Elle nous permet d'envisager l'insertion du comportement du modèle de la régulation sur le site A3, que nous considérerons par la suite comme le modèle local, dans un modèle qui prend en compte plusieurs sites d'épissage, que nous considèrerons comme le modèle global. Afin d'intégrer le modèle local dans le modèle global, il est important de considérer des contraintes différentes en fonction des échelles auxquelles on se place. On utilise une abstraction du modèle local avec l'efficacité d'épissage vu dans la section 10.2.1. Cet observateur est une abstraction du comportement local à l'équilibre. Utiliser cette variable correspond à faire l'hypothèse que la phase de transition est rapide. On effectue alors une abstraction du temps d'exécution d'un processus d'une petite échelle dans un processus de plus grande échelle (*time-scale abstraction*) comme cela est proposé dans [Kuipers, 1988]. Ce procédé correspond à considérer que le processus de régulation local est très rapide par rapport aux processus de régulation d'épissage sur plusieurs sites. Une illustration de cette modélisation sur plusieurs échelles est présentée dans le papier sur un cas d'étude de 3 sites accepteurs d'épissage parmi lesquels on retrouve le site A3.

ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Theoretical Computer Science

ELSEVIER

Theoretical Computer Science ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

www.elsevier.com/locate/tcs

A multi-scale constraint programming model of alternative splicing regulation☆

Damien Eveillard[a,b,*], Delphine Ropers[b], Hidde de Jong[c], Christiane Branlant[b], Alexander Bockmayr[a]

[a] *LORIA, Université Henri Poincaré, BP 239, 54506 Vandoeuvre-lès-Nancy, France*

[b] *Laboratoire de Maturation des ARN et Enzymologie Moléculaire, UMR 7567 CNRS-UHP, BP 239, 54506 Vandoeuvre-lès-Nancy, France*

[c] *INRIA Rhône-Alpes, Helix Project, 655 avenue de l'Europe, Montbonnot, 38334 Saint Ismier, France*

Abstract

Alternative splicing is a key process in post-transcriptional regulation, by which different mature RNA can be obtained from the same premessenger RNA. The resulting combinatorial complexity contributes to biological diversity, especially in the case of the human immunodeficiency virus HIV-1. Using a constraint programming approach, we develop a model of the alternative splicing regulation in HIV-1. Our model integrates different scales (single site vs. multiple sites), and thus allows us to exploit several types of experimental data available to us.

Keywords: Computational biology; Constraint programming; Modeling; Hybrid system; Alternative splicing

1. Introduction

Molecular biology is concerned with the study of three types of biological macromolecules: DNA, RNA, and proteins. Each of these molecules can initially be viewed as a string on a finite alphabet: DNA resp. RNA are nucleic acids, made up of nucleotides A,C,G,T resp. A,C,G,U. Proteins are sequences of amino acids. There exist twenty amino acids, which may be represented by an alphabet of 20 letters. Molecular biology studies the information flow from DNA to RNA, and from RNA to

☆ Part of this work was done within the ARC INRIA "Process Calculi and Biology of Molecular Networks", `http://contraintes.inria.fr/cpbio`

* Corresponding author.
E-mail address: eveillar@loria.fr (D. Eveillard).

0304-3975/$ - see front matter
doi:10.1016/j.tcs.2004.03.062

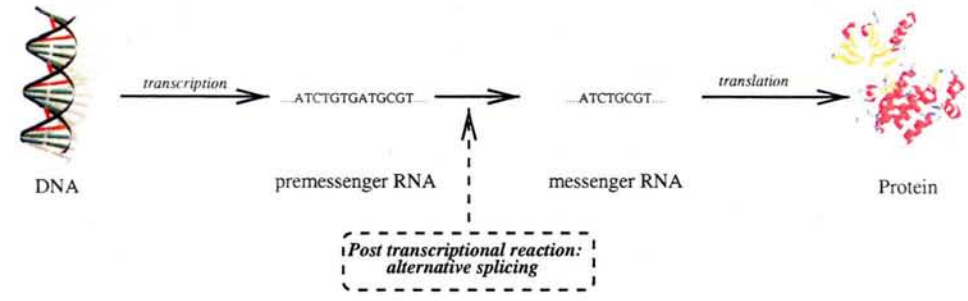

Fig. 1. Information flow in molecular biology.

proteins, see Fig. 1. In a first step, called *transcription*, a substring of DNA ("gene") is transformed into messenger RNA (mRNA). In the second step, called *translation*, the mRNA is translated into a protein, where each triplet of nucleotides encodes one amino acid ("genetic code").

In eukaryotes (i.e., organisms whose cells contain membrane-bound nuclei) and viruses, transcription is followed by another process, which is *alternative splicing* [15]. In a first step, the DNA molecule yields a *premessenger* RNA molecule, by constructing a single-stranded copy of the double-stranded DNA, and by replacing T's with U's. The premessenger RNA may be decomposed into a sequence of substrings called exons and introns. During *splicing*, introns are removed. The remaining exons are concatenated and yield the final *messenger* or *mature* RNA (mRNA). *Alternative* splicing means that through the elimination of selected introns and exons, different mature RNA may be obtained from the same premessenger RNA. In other words, through alternative splicing one and the same gene may code for a variety of proteins.

In our work, we are interested in the alternative splicing regulation of the human immunodeficiency virus (HIV-1). The different proteins that are obtained through this phenomenon play a crucial role in the virus life cycle. Our goal in developing a computational model of the alternative splicing regulation is to get a better understanding of the virus life cycle, in particular the transition from the early to the late phase.

Recent biological studies [8] show that the alternative splicing regulation in HIV-1 depends on a certain class of proteins, so-called *SR proteins* (SR stands for Serine ARginine rich). These proteins can be divided into two functional classes: they may activate or they may inhibit splicing. The knowledge currently available from experiments is limited. Each experiment focuses on one particular splicing site. In a first approach, we therefore model SR regulation in this restricted context. Using differential equations, we develop a continuous model for the regulation of the A3 splicing site in HIV-1. The qualitative behavior of the model depends on the values of the reaction kinetic parameters. Experimental results available to us validate this first approach in the equilibrium phase. In a second step, we integrate the continuous single-site model into a more global multi-site model that expresses the discrete switch from one splicing site to another. This model goes beyond currently available experimental data, and thus may indicate directions for further biological research. Our ultimate goal is to obtain a model that can be validated qualitatively both on the scale of a single splicing site and

on the scale of the whole HIV-1, and which represents the global effect of alternative splicing in the HIV-1 life cycle.

We build our models in a constraint programming framework [2,3]. Constraint programming seems well-suited for modeling biological systems because it allows one to handle partial or incomplete information. Each constraint gives one piece of information on the system that is studied. The overall knowledge is accumulated in the constraint store. The constraint engine available in constraint programming systems operates on the constraint store. It may add new information to the store or check whether some property is entailed by the information present in the store. While a constraint model may be refined whenever additional biological knowledge becomes available, it allows one to make useful inferences even from partial and incomplete information. Therefore, constraint programming seems to be a natural computational approach to face the current situation in systems biology as it is described by Palsson [18]: "Because biological information is incomplete, it is necessary to take into account the fact that cells are subject to certain constraints that limit their possible behaviors. By imposing these constraints in a model, one can then determine what is possible and what is not, and determine how a cell is likely to behave, but never predict its behavior precisely."

The organization of the paper is as follows: we start in Section 2 with a description of the biological process of alternative splicing regulation. Based on a number of biological hypotheses, we develop in Section 3 a continuous model of the regulation at one splicing site. This model includes competition and compensation of different proteins on two binding sites, ESE and ESS2. The single-site model is validated in a qualitative way by extracting from the model a splice efficiency function, which can be measured in experiments. In Section 4, we briefly present the hybrid concurrent constraint programming language `Hybrid cc` [9,10], and explain how it can be used for modeling dynamic biological systems. In Section 5, we first simulate the single-site continuous model in this language. Then we derive a more global model involving three generic splicing sites, which may be generalized to multiple sites. This means that we model at two different scales, using the splice efficiency function as a time-scale abstraction of the local model of one site in the more global context of different sites. The three-site model uses the constraint solving and default reasoning facilities of `Hybrid cc`. This allows us to make predictions on the global behavior even in the absence of detailed local information on some of the splicing sites.

2. Alternative splicing: a biological problem for formals methods

2.1. The biological problem of alternative splicing regulation

The regulation of the splicing process depends on different sites on the premessenger RNA. The first one is the *donor site SD*, located at the end of one exon, see Fig. 2. Its main characteristic is a `GU` nucleic acid sequence motif. The other site is the *acceptor site SA* located at the beginning of the next exon, which is characterized by an `AG` motif. Together, they define the intron to be excised from the premessenger RNA. They permit the binding of a huge ribonucleoproteic complex: the spliceosome.

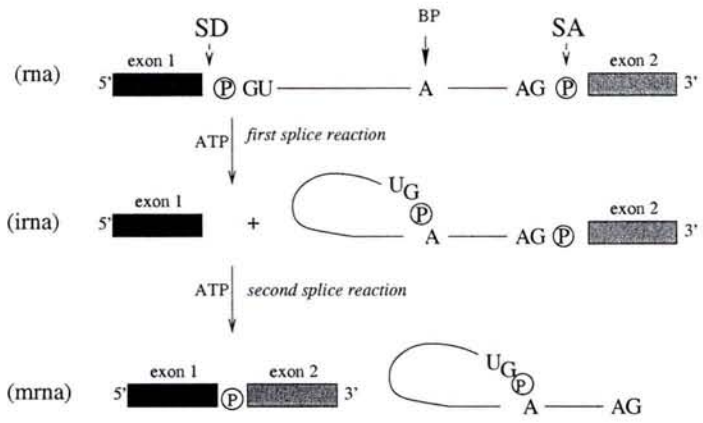

Fig. 2. The splicing process operates in the intronic region of a premessenger RNA that lies between two exons. The exons are delimited by the SD and SA binding sites. A first reaction cuts the RNA at the SD binding site. A second reaction cuts in the SA binding site. Each reaction requires ATP energy.

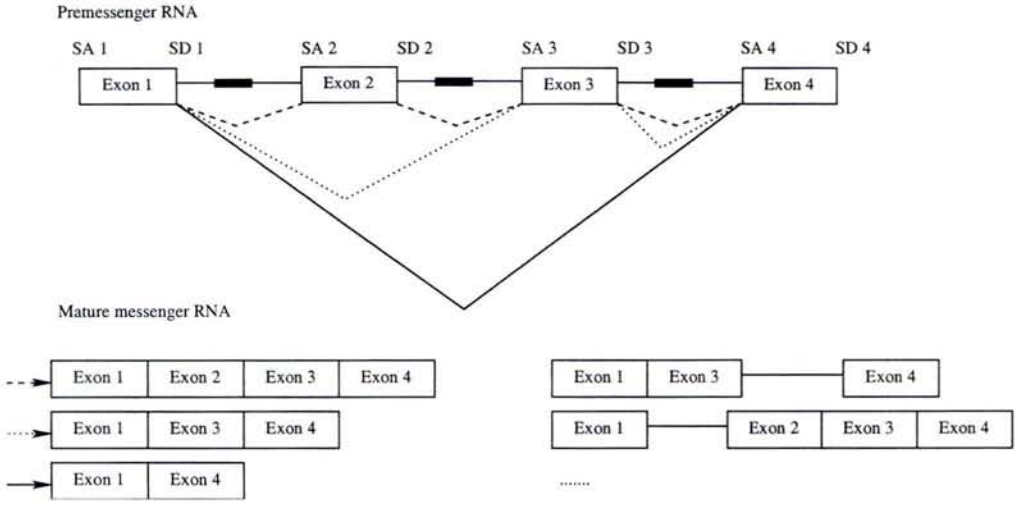

Fig. 3. Obtaining different mature RNA from the same premessenger RNA.

This complex is partially activated by another motif, the *branching point BP*. This is another binding site contained inside the intron. The three sites permit the regulation of the splicing process by activation of the spliceosome complex. The key to the regulation is the choice of one acceptor and one donor site. The splicing activity is determined by additional signals which activate or repress the splicing process.

Understanding the splicing process is a fundamental problem in molecular biology. As illustrated by Fig. 3, various messenger RNA can be obtained from a unique premessenger RNA through the elimination of different introns and exons, and the junction of the remaining exonic sequences. This process depends on the choice of the donor and the acceptor sites.

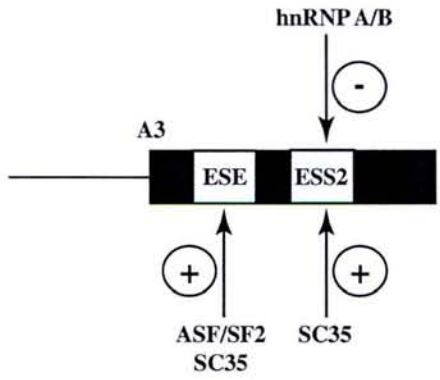

Fig. 4. Regulatory elements of the A3 splicing site. The exon delimited by the A3 acceptor site contains the ESE and ESS2 binding sites, which bind ASF/SF2, SC35 and hnRNP A/B proteins. These regulatory elements activate or repress the splicing reaction on the A3 site.

2.2. Alternative splicing in the context of HIV-1

In the life cycle of the human immunodeficiency virus HIV-1, splicing plays an important role. The viral RNA either remains unchanged to serve as genomic RNA for new virions, or it is spliced to allow for the production of virion proteins [25]. In the HIV-1 case, the alternative splicing regulation involves 4 donor sites (SD) and 8 acceptor sites (SA), which may yield 40 mature messenger RNAs [19]. This diversity is achieved by regulating the selection of the acceptor sites [17,19]. Protein factors such as hnRNP and SR proteins control the regulation via specific binding sites on the premessenger RNA. In general, SR proteins activate the splicing process by initializing the splicing machinery.

In our study, we focus on the acceptor site A3. Inside the A3 splicing site, we distinguish two protein binding sites, ESE and ESS2, see Fig. 4. Splicing can be repressed by hnRNP A/B proteins via the ESS2 binding site [4,7]. Splicing can be activated by the SR proteins SC35 and ASF/SF2 via the ESE binding site [20,21]. However, SC35 can also bind to the ESS2 site. The hypothesis underlying our model is that the ratio of hnRNP A/B and SR proteins determines the splice efficiency at the A3 site.

3. Modeling one splicing site

3.1. Biological hypotheses

We model the regulation by SR proteins in the restricted context of the A3 splicing site under the following hypotheses, see Fig. 4:

- We study only one splicing site. Thus, we consider regulation at the scale corresponding to our experimental results, which are measurements of the splice efficiency given as the ratio of mature RNA over premessenger RNA.

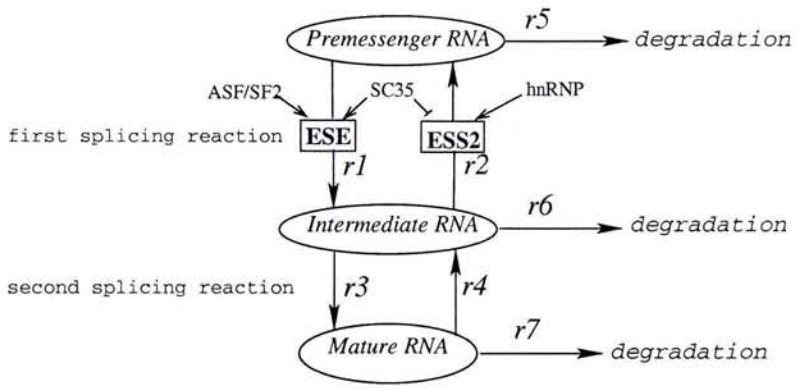

Fig. 5. Schematic representation of the splicing site regulation. The two splicing reactions are composed of 7 kinetic reactions.

- We suppose that the splicing process involves two reactions, relating three functional classes of RNA, see Fig. 2: immature RNA (`rna`), intermediate RNA (`irna`), and mature RNA (`mrna`). Intermediate RNA corresponds to immature RNA activated by proteins. Mature RNA corresponds to mature RNA and introns in lariat.
- The protein concentration in experiments is saturated. Therefore, we assume that it is constant, despite the binding of proteins to the RNA during regulation.
- SR proteins regulate the splicing process by initialization of the splicing machinery.
- Regulation is controlled by the ESE and ESS2 binding sites, which are independent.
- The SR proteins ASF/SF2 and SC35 may activate the first splicing reaction by binding to the site ESE. We assume that these two proteins compensate each other.
- The hnRNP A/B proteins may inhibit the first splicing reaction by binding to the site ESS2. On the other hand, if the SC35 proteins bind to ESS2, this inhibits the hnRNP A/B effect. Therefore we have a competitive inhibition between hnRNP A/B and SC35.

These biological hypotheses are summarized in Fig. 5.

3.2. Mathematical model

Our biological hypotheses can be represented by a system of ordinary differential equations inspired from a model by Monod [14]. In this model, the rate increase depends on the external concentrations, which are limiting factors, and is controlled by a Michaelis–Menten-type kinetics. In our case, we assume that the regulatory protein concentrations are the limiting factors. Such an approach is generally used in ecological modeling, and is well-suited to describe systems that are only partially known.

Table 1
Symbols and units for the biological variables and parameters

Symbol	Variables and parameters	Unit
rna	Immature RNA	μM
irna	Intermediate RNA	μM
mrna	Mature RNA	μM
ASF	Protein ASF/SF2	μM
SC	Protein SC35	μM
R	Protein hnRNP A/B	μM
φ_{ESE}	Maximal affinity for the enhancer	s^{-1}
φ_R	Maximal affinity of hnRNP A/B	s^{-1}
k_{ESE}	Half saturation coefficient for the enhancer	μM
k_{SC}	Half saturation coefficient for SC35	μM
k_R	Half saturation coefficient for hnRNP A/B	μM
κ	Reaction rate	s^{-1}
κ'	Reaction rate	s^{-1}
λ	Degradation coefficient	s^{-1}

The single-site model that we obtain will later be integrated into a larger multi-site model, see Section 5. We will describe the splicing process by seven kinetic reactions. The symbols used are given in Table 1.

The reaction r_1 represents the transformation of premessenger RNA to intermediate RNA. It requires cooperation between ASF/SF2 and SC35 proteins for the regulation of ESE. Since we assume compensation, only the sum of the two activator proteins is important. We represent the reaction rate by a Michaelis–Menten function depending on the quantity of immature RNA, and controlled by the sum of the proteins ASF/SF2 and SC35. The generic form of the Michaelis–Menten function, see e.g. [16], is:

$$v = \frac{v_{\max} x}{k_m + x}.$$

The curve expressing the relationship between v and x is given in Fig. 6. Here, $v_{\max}$ is the maximum rate, and the Michaelis constant k_m is the value at which v is half maximal. With $x = ASF + SC$ and the notation from Table 1, we get:

$$r_1 = \frac{\varphi_{ESE}(ASF + SC)}{k_{ESE} + (ASF + SC)} rna.$$

The reaction r_2 represents the transformation of intermediate RNA to premessenger RNA. It captures the antagonistic function of hnRNP A/B and SC35 proteins on the site ESS2. We use a similar function as before. However, we now have a *competitive inhibition*, see e.g. [13], between two species x and y. The generic form becomes

$$v = \frac{v_{\max} x}{k_m(1 + \frac{y}{k_y}) + x}.$$

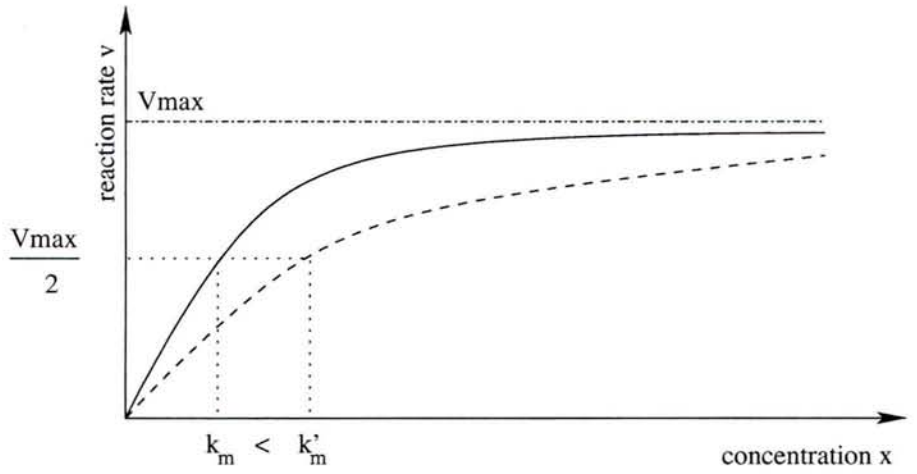

Fig. 6. Michaelis–Menten function.

The factor $(1 + y/k_y)$ increases the value of k_m, see Fig. 6 for illustration. With $x = R$ and $y = SC$ and the notation from Table 1, we get

$$r_2 = \frac{\varphi_R R}{k_R(1 + \frac{SC}{k_{SC}}) + R} irna.$$

The reaction r_3 represents the transformation of intermediate RNA to mature RNA (mrna). We assume for this reaction a simple first-order kinetics with a constant parameter κ. Similarly, r_4 represents the reaction which transforms mature RNA to intermediate RNA:

$$r_3 = \kappa irna, \quad r_4 = \kappa' mrna.$$

r_5, r_6 and r_7 respectively, represent the degradation reaction of immature RNA, intermediate RNA and mature RNA. Different RNAs decrease proportionally to the same degradation factor λ

$$r_5 = \lambda rna, \quad r_6 = \lambda irna, \quad r_7 = \lambda mrna.$$

We formalize the splicing process at site A3 by the system of differential equations, see again Fig. 5,

$$\frac{d(rna)}{dt} = r_2 - r_1 - r_5,$$
$$\frac{d(irna)}{dt} = r_1 + r_4 - r_2 - r_3 - r_6,$$
$$\frac{d(mrna)}{dt} = r_3 - r_4 - r_7,$$

which corresponds to

$$\frac{\mathrm{d}(rna)}{\mathrm{d}t} = \frac{\varphi_R R}{k_R(1+\frac{SC}{k_{SC}})+R}\, irna - \frac{\varphi_{ESE}(ASF+SC)}{k_{ESE}+(ASF+SC)}\, rna - \lambda rna,$$

$$\frac{\mathrm{d}(irna)}{\mathrm{d}t} = \frac{\varphi_{ESE}(ASF+SC)}{k_{ESE}+(ASF+SC)}\, rna - \frac{\varphi_R R}{k_R(1+\frac{SC}{k_{SC}})+R}\, irna - \kappa irna + \kappa' mrna - \lambda irna,$$

$$\frac{\mathrm{d}(mrna)}{\mathrm{d}t} = \kappa irna - \kappa' mrna - \lambda mrna.$$

3.3. *Validation of the regulatory system*

The mathematical model of regulation at the acceptor site A3 can be directly simulated in the constraint programming language `Hybrid cc`, as will be shown in Section 4. However, it should first be validated with respect to existing biological knowledge [1].

In our model, the RNA concentrations do not reach an equilibrium, i.e., a state in which no further net change is occurring, but continue to decrease until total degradation of RNA. However, we may assume that the splicing reactions quickly reach an equilibrium. In the equilibrium phase, we have $r_1 = r_2$, $r_3 = r_4$, which is equivalent to

$$\frac{\varphi_{ESE}(ASF+SC)}{k_{ESE}+(ASF+SC)}\, rna = \frac{\varphi_R R}{k_R(1+\frac{SC}{k_{SC}})+R}\, irna, \quad \kappa irna = \kappa' mrna.$$

If we define the *splice efficiency* by

$$efficiency(t) = \frac{mrna(t)}{rna(t)},$$

we obtain the following formula for the splice efficiency in the equilibrium phase:

$$efficiency_{\mathrm{eq}} = \frac{\kappa\varphi_{ESE}(ASF+SC)(k_R k_{SC} + k_R SC + R k_{SC})}{\kappa'(k_{ESE}+ASF+SC)\varphi_R R k_{SC}}.$$

According to our formula, the splice efficiency is

- an increasing function of the activators SC and ASF.
- a decreasing function of the inhibitor R.

Experimental results show that

- $(mrna/rna)_{\mathrm{eq}}$ increases with an increase of activator proteins.
- $(mrna/rna)_{\mathrm{eq}}$ decreases with an increase of inhibitor proteins.

Thus, the results of our model correlate with available experimental data. Therefore, we may consider the model to be qualitatively validated under the hypotheses described in Section 3.1. We next consider simulation in the concurrent constraint language `Hybrid cc`.

4. Hybrid concurrent constraint programming

To model alternative splicing regulation, we will use hybrid concurrent constraint programming, Hybrid cc [9,10]. The general idea of *constraint programming* for system modeling is that the user specifies constraints on the behavior of the system that is being studied. Each constraint expresses some partial information on the system state. The constraint solver may check constraints for consistency or infer new constraints from the given ones. In *concurrent constraint programming* (cc), different computational processes may run concurrently. Interaction is possible via the *constraint store*. The store contains all the constraints currently known about the system. A process may *tell* the store a new constraint, or *ask* the store whether some constraint is entailed by the information currently available, in which case further action is taken [22]. One major difficulty in the original cc framework is that cc programs can detect only the presence of information, not its absence. To overcome this problem, Saraswat et al. [23] proposed to add to the cc paradigm a sequence of phases of execution. At each phase, a cc program is executed. At the end, absence of information is detected, and used in the next phase. This results in a synchronous reactive programming language, Timed cc. But, the question remains how to detect negative information instantaneously. Default cc extends cc by a negative ask combinator if c else A, which imposes the constraints of A unless the rest of the system imposes the constraint c. Logically, this can be seen as a default. Introducing phases as in Timed cc leads to Timed Default cc [24]. Only one additional construct is needed: hence A, which starts a copy of A in each phase after the current one.

Hybrid cc [9,10], is an extension of Default cc over continuous time. First continuous constraint systems are allowed, i.e., constraints may involve differential equations that express initial value problems. Second, the hence operator is interpreted over continuous time. It imposes the constraints of A at every real time instant after the current one. The evolution of a system in Hybrid cc is piecewise continuous, with a sequence of alternating point and interval phases. All discrete changes take place in a point phase, where a simple Default cc program is executed. In a continuous phase, computation proceeds only through the evolution of time. The interval phase, whose duration is determined in the previous point phase, is exited as soon as the status of a conditional changes [10]. Table 2 summarizes the basic combinators of Hybrid cc.

It has been argued in [2,3] that Hybrid cc is well-suited for modeling dynamic biological systems. In addition to the general discussion in [3], we illustrate here by a number of small examples, how the basic combinators of Hybrid cc can be applied naturally to the study of biological systems.

4.1. Interval constraints and continuous dynamics

The Hybrid cc language that we are using is based on interval constraints [5]. This means that variables are defined over an interval of real numbers, and computations are done in interval arithmetic. This is very useful in biology, where typically parameters and values are not exactly known.

Table 2
Combinators of `Hybrid cc`

Agents	Propositions
c	c holds now
`if` c `then` A	if c holds now, then A holds now
`if` c `else` A	if c does not hold now, then A holds now
`new` X `in` A	the variable X is local to A (hiding)
(A,B)	both A and B hold now
`hence` A	A holds at every instant after now
`always` A	same as (A, `hence` A)
`unless(`c`)` A `else` B	same as (`if` c `then` B, `if` c `else` A)

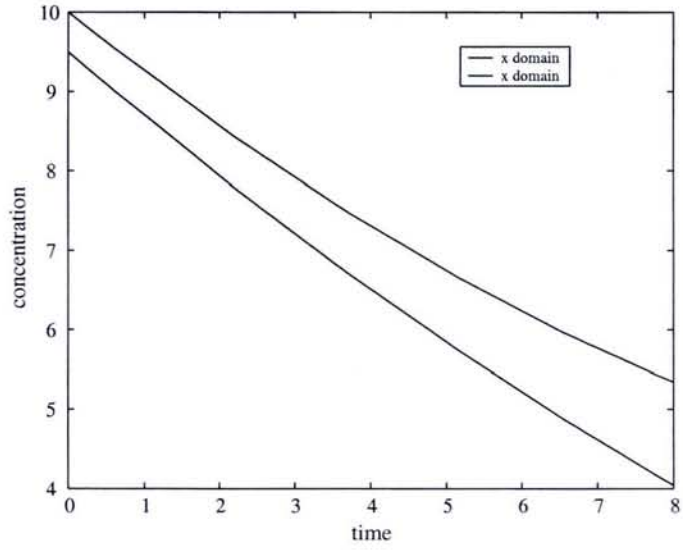

Fig. 7. Enclosure for the dynamics of a molecular species with linear kinetics.

We illustrate this by a very simple example in `Hybrid cc` involving a single constraint on an interval variable x, see Fig. 7. Since we are reasoning about dynamical systems, we use the `always` A combinator, expressing that A holds at every time instant.

```
interval x;
x = [9.5,10];
always { x' = -(2*x)/(15+x);
       }
sample(x);
```

4.2. Parallel composition

Hybrid cc allows for parallel composition of constraints. (A,B) imposes the constraints of both A and B. Operationally, the program (A,B) behaves like the simultaneous execution of both A and B. A and B may share common variables, and thus communicate via the constraint store.

We illustrate parallel composition by a small Hybrid cc program specifying a Michaelis–Menten kinetics. Consider two molecular species X and Y with concentrations x and y, and suppose X is transformed into Y. The initial concentration of X lies in the interval $[14, 14.5]$. The production rate of Y depends on the concentration of X according to the formula $y' = (v_{\max} * x)/(k_m + x)$, for some constants $v_{\max}$ and k_m. The concentration of X is reduced at the same rate. We add constraints $x, y \geqslant 0$ to say that concentrations are non-negative, and constraints $s = x + y$, $s' = 0$ to express conservation of matter. The constraint solver computes enclosures for x and y, see Fig. 8. In particular, we can observe that at the end of the experiment, the concentration of y will be greater than the concentration of x. Interval constraints are particularly useful in sensibility studies, where we can easily test the importance of one variable compared to the others.

```
#define km 1.5
#define vmax  2
interval x,y,s;
x = [14,14.5];              /* Initialization */
y = 0;
always {
   x' = -(vmax*x)/(km+x); /* Michaelis-Menten kinetics */
   y' =  (vmax*x)/(km+x);
   x >= 0;                /* Non-negative concentrations */
   y >= 0;
   s = x + y;             /* Conservation of matter */
   s' = 0;
}
sample(x, y);
```

4.3. Conditionals and discrete change

In general, the dynamics of a system will depend on conditions. In Hybrid cc, we may use the combinator if c then A expressing that if c holds now, then A holds now. This allows one to make discrete changes to switch from one dynamics to another. The next program models the situation that the transformation of X to Y gets activated if a certain protein P reaches a threshold, see Fig. 9 (top) for illustration.

```
interval x, y, p;
x=[14,14.5];
y=0;
```

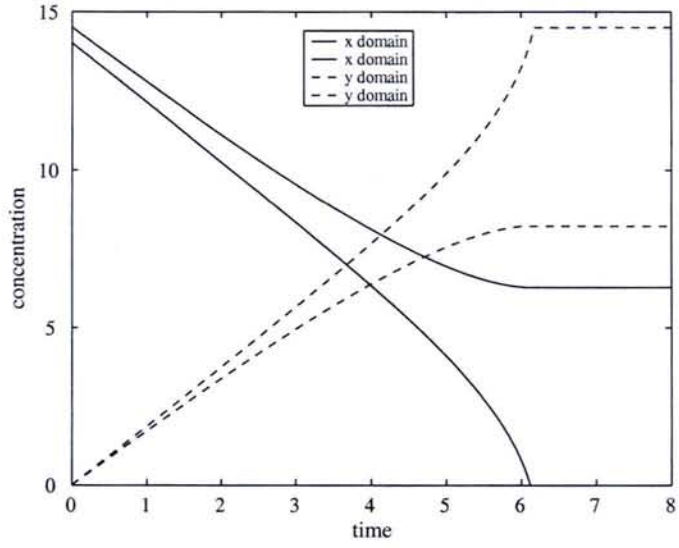

Fig. 8. Enclosures for the dynamics of two molecular species with Michaelis–Menten kinetics. Due to the constraints $x, y \geqslant 0, (x + y)' = 0$, the domain bounds get constant when the lower bound of x reaches 0.

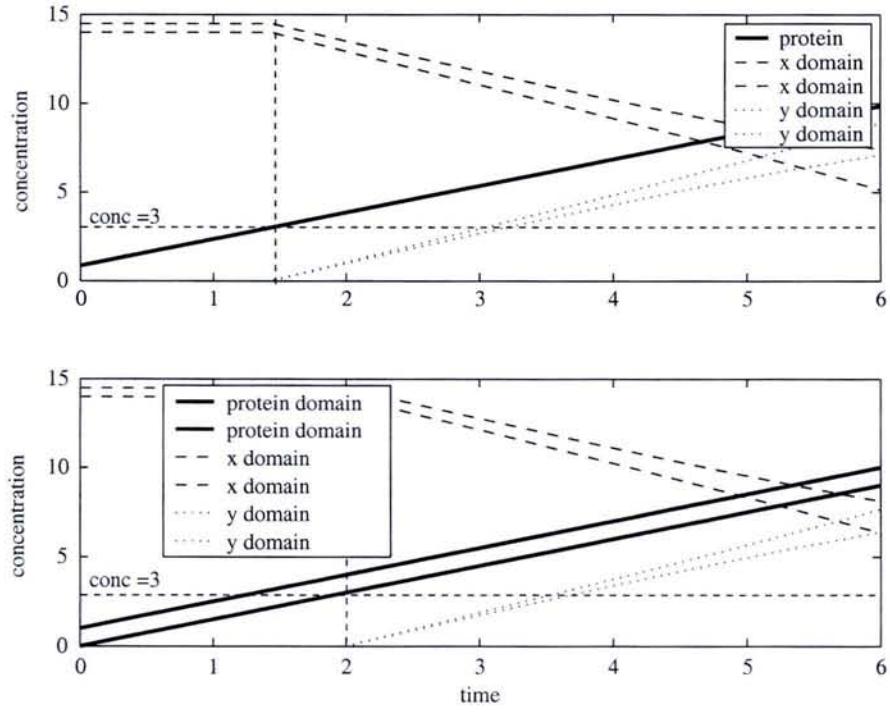

Fig. 9. Switching behavior for conditional (top) and default combinator (bottom).

```
p=0.75;
always {
   p' = 1.5;
   if (p >= 3)
      { x' = -(vmax*x)/(km+x);
        y' =  (vmax*x)/(km+x);
   }
   if (p < 3)
      { x' = 0;
        y' = 0;
   }
}
sample(p, x ,y);
```

Here, we have assumed that there is no uncertainty on the initial value of p. Without this hypothesis, the constraint solver cannot decide between the two alternatives $p \geqslant 3$ and $p < 3$. In order to handle conditions in the presence of uncertainty, we use default reasoning that we describe next.

4.4. *Default behavior*

The default combinator `if` c `else` A (or `unless`(c) A) expresses that A holds now, if c will not hold now. Operationally, this means that the current store on quiescence does *not* entail c. Note that `unless`(c) A is not equivalent to `if` $\neg c$ `then` A. If A is executed, this may have two reasons:

- The current store entails $\neg c$ (in this case `unless`(c) A behaves like `if` $\neg c$ `then` A), or
- the current store neither entails c nor $\neg c$, i.e., it is not known whether or not c holds. In this case, A is executed *by default*.

A is not executed, if the current store entails c.

We use the same example as before. The only difference is that the variable p representing the protein concentration is initialized with the interval $[0, 1]$. As we can see in Fig. 9 (bottom), the reaction gets activated when the *lower* bound for p reaches the threshold.

```
interval x, y, p;
x=[14,14.5];
y=0;
p=[0,1];
always {
   p' = 1.5;
   if ( p >= 3 )
      { x' = -(vmax*x)/(km+x);
        y' =  (vmax*x)/(km+x);
   }
```

```
    unless (p >= 3)
       { x' = 0;
         y' = 0;
    }
}
sample(p, x ,y);
```

The default combinator is a convenient way of handling incomplete knowledge in biology. In particular, we will use it in our multi-site model of alternative splicing regulation in Section 5.2.

5. Modeling the alternative splicing regulation with Hybrid cc

5.1. Single-site model: local modeling

The single-site model from Section 3.2 with experimental values can be expressed directly in `Hybrid cc`.

```
# define Pese 0.01              # define kr 0.01
# define Psc 0.2                # define k 0.19
# define Pr 0.4                 # define kk 0.01
# define kese 0.35              # define SC 2
# define ksc 2                  # define ASF 1.75
# define R 0.35

interval t, rna, irna, mrna;
t=0; rna = 0.06; irna = 0; mrna = 0;
always{
   rna'  =  (Pr*R*irna)/(kr*(1+(SC/ksc))+R)
           -(Pese*(ASF+SC)*rna)/(kese+ASF+SC)
           -delta*rna;
   irna' =  (Pese*(ASF+SC)*rna)/(kese+ASF+SC)
           -(Pr*R*irna)/(kr*(1+(SC/ksc))+R)
            -k*irna+kk*mrna-delta*irna;
   mrna' = k*irna-kk*mrna-delta*mrna;
}
sample(rna, irna, mrna);
```

During the simulation, we obtain for the splice efficiency *mrna/rna* the equilibrium predicted in Section 3.3, see Fig. 10. Under our hypotheses, which include protein competition and compensation, the model correctly simulates the alternative splicing activity at site A3. This supports the hypotheses made in the model such as the role of the ESE and ESS2 binding sites.

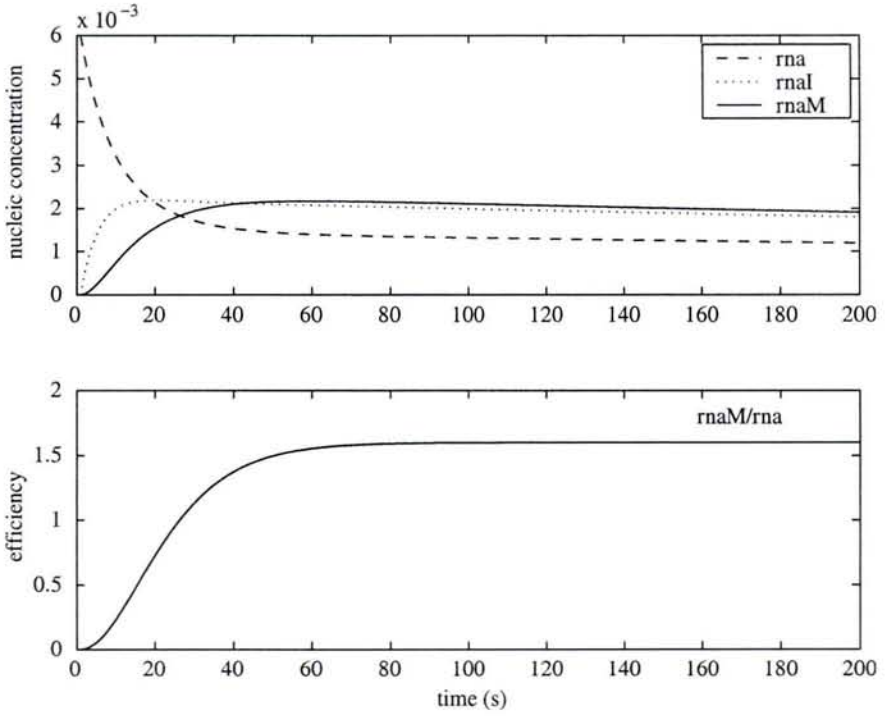

Fig. 10. Variation of RNA pool and splice efficiency in the splicing reaction.

5.2. *Three-site model: global modeling*

A realistic model of alternative splicing has to reflect the combinatorial complexity discussed in Section 2.2. Assuming that regulation is modular [12], the single-site model may be seen as one module inside a larger framework. The qualitative validation given in Section 3.3 justifies the introduction of the single-site model into a larger-scale model involving several splicing sites. To illustrate this, we consider the generic example of three acceptor sites (A3, A4 and A7) associated with one donor site (SD), see Fig. 11.

Using time-scale abstraction, the behavior at one splicing site is captured by a single function, the splice efficiency, which depends on the protein concentrations. This function is used in a larger-scale global model that describes the choice between three acceptor sites A3, A4 and A7. In the HIV-1 case, the A4 site is the default splicing site. Only if the efficiency of A3 (`effA3`) or A7 (`effA7`) gets larger than the efficiency of A4 (`effA4`), regulation switches to the other state. The sites A3, A4, and A7 exhibit three generic behaviors, see also Fig. 12:

- A3 is a regulated site with known behavior.
- A7 is a regulated site with unknown behavior.
- A4 is an unregulated site, i.e., the behavior does not depend on protein concentrations.

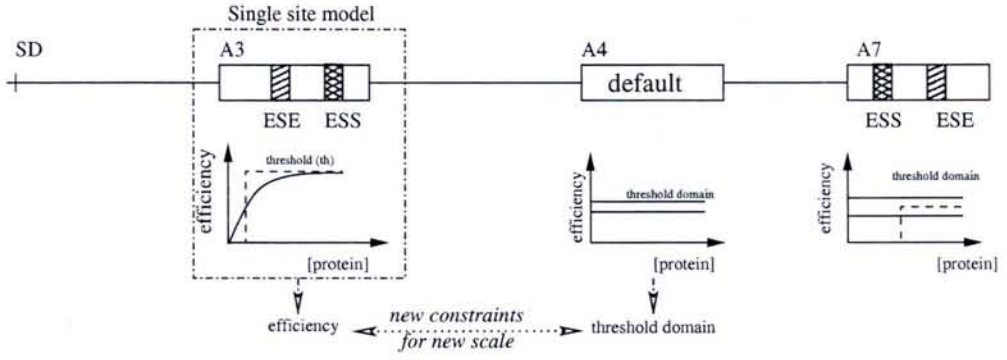

Fig. 11. Single-site model inside a more general multi-site regulation model.

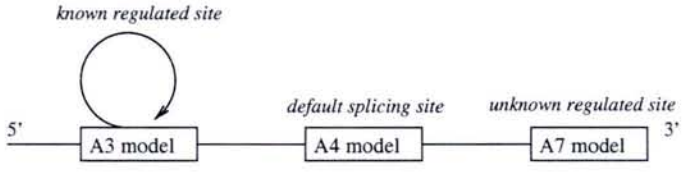

Fig. 12. Biological information on three acceptor sites A3, A4, A7.

Current biological experiments give information on the local behavior at one site. However, modeling the local behavior is not enough. In order to understand the global splicing process, we must integrate several types of knowledge. On the one hand, we have information on the local behavior at individual acceptor sites. On the other hand, we have some information on the global behavior, like the default role of A4 or the competition between different acceptor sites. Constraint programming allows us to integrate this information, and to produce a global model.

Recent work [6] shows the linearity of the splicing kinetics. Thus, on the larger scale, we may consider splicing as a linear process described by three systems of ordinary differential equations. For each acceptor site Ai, $i \in \{3, 4, 7\}$, we introduce one system with four differential equations:

- r_{i1} represents the consumption of immature RNA if Ai is dominating.
- r_{i3} represents the production of mature RNA at A3.
- r_{i4} represents the production of mature RNA at A4.
- r_{i7} represents the production of mature RNA at A7.

k_{ij} is the kinetic constant for reaction r_{ij}.

A4 is the default splicing site. It is dominating unless the splice efficiency of A3 or A7 gets larger than the splice efficiency of A4. If this happens, A7 becomes the default splicing site unless the efficiency of A3 gets larger than the efficiency of A7, see Fig. 13. The local behavior at A3 has been described by the single-site model given in Section 5.1. This model predicts the splice efficiency of A3 depending on the protein concentrations.

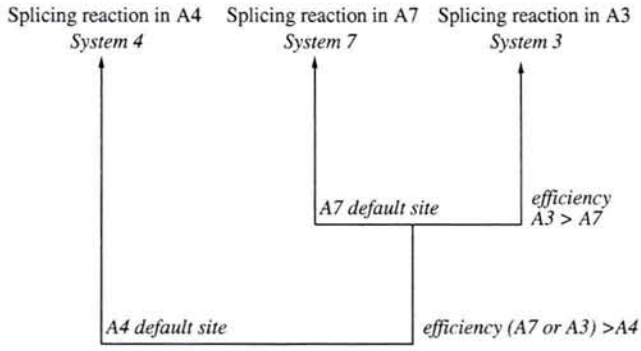

Fig. 13. Choice of the acceptor site A3, A4 or A7 depending on the splice efficiency.

In the Hybrid cc program given below, the concentration of SC35 is increased linearly. Depending on the corresponding variation of the splice efficiency at A3, the three-site model exhibits different behaviors, characterized by the choice of one of the three differential equation systems. The default behavior discussed before can be expressed naturally in Hybrid cc using the combinator unless(c) A.

```
#define c1 2
#define c2 0.5
#define c3 0.8
#define c4 1
#define c5 0.9
#define c6 0.1
#define R 3.5

interval t, prot, effA3, effA4, effA7, rna, mrnaA3,
    mrnaA4, mrnaA7;
t = 0;
rna = 10;
mrnaA3 = 0; /*known regulated acceptor site */
mrnaA4 = 0; /*unregulated acceptor site*/
mrnaA7 = 0; /*unknown regulated acceptor site*/

always { t' = 10;
prot = 0.1*t;                  /* protein variation */

/* A3 efficiency depends on the protein concentration
   A3 efficiency represents the local behavior of A3
   (observer of A3) */
effA3 = c1*(prot+c2)*(c3*prot+c4)/(c5*(prot+c6));

6 <= effA4; effA4 <= 8;  /* effA4 : efficiency domain of A4*/
7 <= effA7; effA7 <= 9;  /* effA7 : efficiency domain of A7*/
}
```

```
/* The behavior depends on the efficiency of
   the 3 acceptor sites*/

always {
/* if A3 or A7 dominant */
if (effA3 >= effA4 || effA7 >= effA4) {
      if (effA7 <= effA3)  {        /* splicing on A3 */
          rna' = -0.51 * rna - 0.01*rna;
       mrnaA3' =  0.4  * rna - 0.1*mrnaA3;  /* A3 kinetics */
       mrnaA4' =  0.01 * rna - 0.1*mrnaA4;  /* A4 kinetics */
       mrnaA7' =  0.1  * rna - 0.1*mrnaA7;  /* A7 kinetics */
      };
      unless ((effA7 <= effA3)) { /*default splicing on A7*/
          rna' = -0.51 * rna - 0.01*rna;
       mrnaA3' =   0.1 * rna - 0.1*mrnaA3; /* A3 kinetics */
       mrnaA4' =  0.01 * rna - 0.1*mrnaA4; /* A4 kinetics */
       mrnaA7' =   0.4 * rna - 0.1*mrnaA7; /* A7 kinetics */
      };
   };
/* default splicing on A4 */
unless (effA3 >= effA4 || effA7 >= effA4) {
     rna' = -0.32 * rna - 0.01*rna;
  mrnaA3' = -0.01 * rna - 0.1*mrnaA3; /* A3 kinetics */
  mrnaA4' =  0.3  * rna - 0.1*mrnaA4; /* A4 kinetics */
  mrnaA7' = -0.01 * rna - 0.1*mrnaA7; /* A7 kinetics */
   };
};
sample(prot, effA3, rna, mrnaA3, mrnaA4, mrnaA7);
```

According to the semantics of the default combinator, the A4 site will be chosen if the solver cannot deduce that (effA3≥effA4) or (effA7≥effA4). This may have *two* reasons:

- (effA3≥effA4) or (effA7≥effA4) is false, i.e., (effA3<effA4) and (effA7 <effA4), or
- it is not known whether (effA3≥effA4) or (effA7≥effA4) holds (default behavior).

Thus A4 is the default site if the splice efficiency of A3 and A7 is not sufficiently high. If A3 or A7 dominate A4, then A7 is the default splicing site, unless A3 dominates A7.

Simulation in Hybrid cc yields the behavior shown in Fig. 14. First mrnA4 is produced, i.e., the default site A4 is active. When effA3 passes the upper threshold for effA4, site A7 gets activated, and mrnA7 is produced. Finally, when effA3 further increases and passes the upper threshold for effA7, site A3 gets activated and we observe production of mrnA3, while the concentrations of mrnA4 and mrnA7 become stationary.

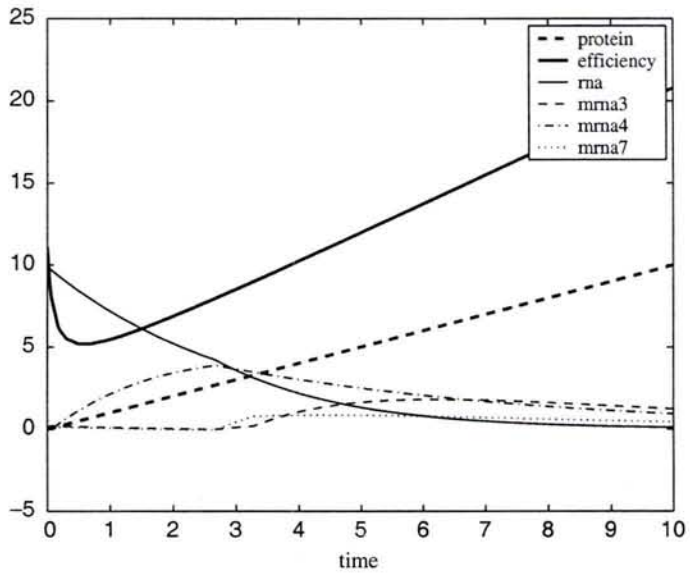

Fig. 14. Variation of mRNA production depending on a variation of SR proteins.

Basically, the model gives to the biologist three qualitative states: first splicing at the A4 site, second splicing at the A7 site, and finally splicing at the A3 site. The constraint programming system can compute enclosures for the three biological states, despite the variation in the concentrations of SR proteins. The enclosure is an important qualitative information to extend the single-site to a multi-site model. `Hybrid cc` permits a qualitative validation of the model, although the currently available information on the alternative splicing regulation in HIV-1 is incomplete.

6. Conclusion and further research

Our approach combines mathematical and computational methods. Mathematical analysis allows us to validate the single-site model in a qualitative way, based on the experimental data obtained in our group. The validation shows the consistency of our biological hypotheses. In a second step, we can extract the splice efficiency as a time-scale abstraction of the local behavior at one site inside a more global model involving different sites. For the experimental biologist, the single-site model may serve as a computational tool to evaluate his knowledge on a fine-grained biological process.

On the computational side, the constraint solving and default reasoning capabilities of `Hybrid cc` allow us to exploit as much as possible the incomplete knowledge of our system. Default behavior may compensate the lack of experimental data. Using constraint programming, we can delimit with our model the possible splicing behavior. This provides a powerful tool for qualitative validation.

Combining mathematical analysis and computational methods is the key to extending the single-site model to a multi-site model as described in this paper. It leads to the

qualitative validation represented by the extraction of the splice efficiency function. The splice efficiency characterizes the modularity of the regulation. Thus, the one-site behavior is represented in the three-site model, based on the single-site splice efficiency. The extraction of a suitable criterion on the smaller scale is crucial to understanding an experimental process from a systems biology perspective. Furthermore, constraints can be used to handle the problem of missing data in time-scale abstraction of a single-site model in a more global multi-site model. Different scales usually correspond to biological experiments yielding different types of results. Despite the variety of possible experiments, these must be integrated into a global model in order to better understand the biological process.

Modeling alternative splicing requires a close interaction between biological and computational approaches. In the context of alternative splicing regulation, we are currently working on new experimental data for the quantitative validation of our models. On the computational side, we have integrated our model into a general model of the HIV-1 life cycle [11]. Preliminary results show that the modification of a splice constant may induce different behaviors in the HIV-1 life cycle model. Using the extended model, we may validate several biological hypotheses on the global effect of alternative splicing in the full HIV-1 life cycle.

Acknowledgements

The authors would like to thank Arnaud Courtois for his comments on a draft of this paper.

References

[1] O. Bernard, J.-L. Gouzé, Nonlinear qualitative signal processing for biological systems: application to the algal growth in bioreactors, Math. Biosci. 157 (1999) 357–372.
[2] A. Bockmayr, A. Courtois, Modeling biological systems in hybrid concurrent constraint programming (abstract), in: The Second Internat. Conf. Systems Biology, ICSB'01, Pasadena, CA, 2001, p. 106.
[3] A. Bockmayr, A. Courtois, Using hybrid concurrent constraint programming to model dynamic biological systems, in: 18th Internat. Conf. on Logic Programming, ICLP'02, Copenhagen, Lecture Notes in Computer Science, Vol. 2401, Springer, Berlin, 2002, pp. 85–99.
[4] M. Caputi, M. Mayeda, A. Krainer, A. Zahler, hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing, EMBO 18 (14) (1999) 4060–4067.
[5] B. Carlson, V. Gupta, Hybrid cc and interval constraints, in: Hybrid Systems: Computation and Control, HSCC'98, Lecture Notes in Computer Science, Vol. 1386, Springer, Berlin, 1998, pp. 80–95.
[6] A. Audibert, D. Weil, F. Dautry, In vivo kinetics of mRNA splicing and transport in mammalian cells, Molecular Cell Biol. 22 (2002) 6706–6718.
[7] F. Del Gatto-Konczak, M. Olive, M. Gesnel, R. Breathnach, hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer, Molecular Cell. Biol. 19 (1) (1999) 251–260.
[8] B. Graveley, Sorting out the complexity of SR protein functions, RNA 6 (2000) 1197–1211.
[9] V. Gupta, R. Jagadeesan, V. Saraswat, Computing with continuous change, Sci. Comput. Programming 30 (1–2) (1998) 3–49.
[10] V. Gupta, R. Jagadeesan, V. Saraswat, D.G. Bobrow, Programming in hybrid constraint languages, in: Hybrid Systems II, Lecture Notes in Computer Science, Vol. 999, Springer, Berlin, 1995, pp. 226–251.

[11] B. Hammond, Quantitative study of the control of HIV-1 gene expression, J. Theoret. Biol. 163 (1993) 199–221.
[12] L. Hartwell, J. Hopfield, S. Leibler, A. Murray, From molecular to modular cell biology, Nature 402 (1999) C47–C52.
[13] R. Heinrich, S. Schuster, The Regulation of Cellular Systems, Thomson Publishing, New York, 1996.
[14] J. Monod, La technique des cultures continues. Théorie et applications, Ann. Inst. Pasteur 79 (1950) 390–410.
[15] M. Moore, C. Query, P. Sharp, Splicing of precursors to mRNA by the spliceosome, in: R. Gesteland, J. Atkins (Eds.), The RNA World, Cold Spring Harbor Laboratory Press, New York, 1993, pp. 303–357.
[16] J.D. Murray, Mathematical Biology I, An Introduction, 3rd Edition, Springer, Berlin, 2002.
[17] M. O'Reilly, M. McNally, K. Beemon, Two strong 5' splice sites and competing, suboptimal 3' splice sites involved in alternative splicing of human immunodeficiency virus type 1 RNA, Virology 213 (2) (1995) 373–385.
[18] B. Palsson, The challenges of in silico biology, Natur. Biotechnol. 18 (2000) 1147–1150.
[19] D. Purcell, M. Martin, Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity, J. Virol. 67 (11) (1993) 6365–6378.
[20] D. Ropers, Etude expérimentale du rôle des protéines SR dans la régulation de l'épissage de l'ARN du virus HIV-1, responsable de l'immunodéficience humaine, et modélisation mathématique de ces régulations, Ph.D. Thesis, University Henri Poincaré, Nancy, France.
[21] D. Ropers, L. Ayadi, S. Jacquenet, A. Méreau, D. Thomas, A. Mougin, P. Bilodeau, M. Stolzfus, R. Gattoni, J. Stévenin, C. Branlant, Differential effects of the SR proteins 9G8, SC35, ASF/SF2 and SRp40 on the utilization of the A1 to A5 splicing sites of HIV-1 RNA, J. Biol. Chem., in press.
[22] V.A. Saraswat, Concurrent Constraint Programming, MIT Press, Cambridge, MA, 1993.
[23] V.A. Saraswat, R. Jagadeesan, V. Gupta, Foundations of timed concurrent constraint programming, in: The Ninth Symp. Logic in Computer Science, LICS'94, Paris, IEEE, New York, 1994, pp. 71–80.
[24] V.A. Saraswat, R. Jagadeesan, V. Gupta, Timed default concurrent constraint programming, J. Symbol. Comput. 22 (5/6) (1996) 475–520.
[25] H. Tang, K. Kuhen, F. Wong-Staal, Lentivirus replication and regulation, Annual Rev. Genet. 33 (1999) 133–170.

11.2 Construction et analyse de modèle multi-échelles

Après les travaux que nous venons de présenter, la modélisation du processus d'épissage sur plusieurs échelles apparaît comme possible formellement avec l'utilisation de la programmation par contraintes. Dans cette optique, nous construirons dans un premier temps un modèle qui représente le choix alternatif entre différents sites d'épissage que nous considérerons comme un modèle multi-sites.

11.2.1 Hypothèses et conceptualisation de la régulation multi-site

La conceptualisation nécessite de faire différentes hypothèses avant de conceptualiser le modèle. Nous proposons en effet de modéliser la régulation *in vitro* de l'épissage alternatif par les protéines SR et hnRNP A1 dans un contexte de régulation multi-site gouvernant les sites accepteurs A3, A4, A5 et A7. Nous représenterons la dynamique dans une représentation de système à compartiment clos (voir 10.1.2 pour illustrations et références). Les compartiments représenteront les ensembles des acides nucléiques avant et après le processus biologique. Cette abstraction est relativement simpliste mais elle permet néanmoins de se placer dans un référentiel cohérent avec les connaissances expérimentales actuelles du phénomène. Les hypothèses que nous formulons correspondent à des contraintes décrites dans les résultats expérimentaux de [Purcell & Martin, 1993]. Les variables d'états sont les ARN pré-messagers *rna* qui deviendront des ARN matures qui après traduction deviendront diverses protéines. Ces ARN sont associés aux protéines :

- Tat de la phase précoce qui est composée de deux exons Tat1 et Tat2 (*tat1-2*)
- Rev de la phase précoce (*rev*)
- Nef de la phase précoce (*nef*)
- Tat de la phase tardive qui est composé d'un exon Tat1 (*tat1*)

Le processus de régulation d'épissage alternatif dépend de la concentration en protéines régulatrices comme les protéines SC35, ASF/SF2 et hnRNP A1. Nous négligeons lors de ce modèle la régulation inhérente aux sites donneurs. Les travaux de [O'Reilly *et al.*, 1995] corroborent cette hypothèse forte en affirmant que les sites donneurs sont moins régulés que les sites accepteurs. Une modélisation concernant les sites accepteurs seuls sont de plus en corrélation avec les connaissances actuelles du processus de régulation de l'épissage. Les protéines de régulation activent ou inhibent les sites accepteurs A3, A4, A5 et A7. Nous proposons de restreindre l'éventail des sites accepteurs à ces quatre sites qui sont d'une importance qualitative non négligeable dans la régulation de l'épissage.

Les détails de ce modèle sont exposés dans [Bockmayr *et al.*, *a*] présenté plus loin. Le modèle construit grâce à des contraintes hybrides est un automate hybride (voir 4.4.1) qu'il est possible d'analyser qualitativement par une forme de *model checking*. Il est nécessaire pour cela d'isoler une instanciation particulière du modèle concernant la variation d'une protéine régulatrice. Après analyse, il est possible d'interroger le système avec différentes requêtes de type CTL à réponse positives ou négatives (voir [Chabrier & Fages, 2003] pour application à la biologie).

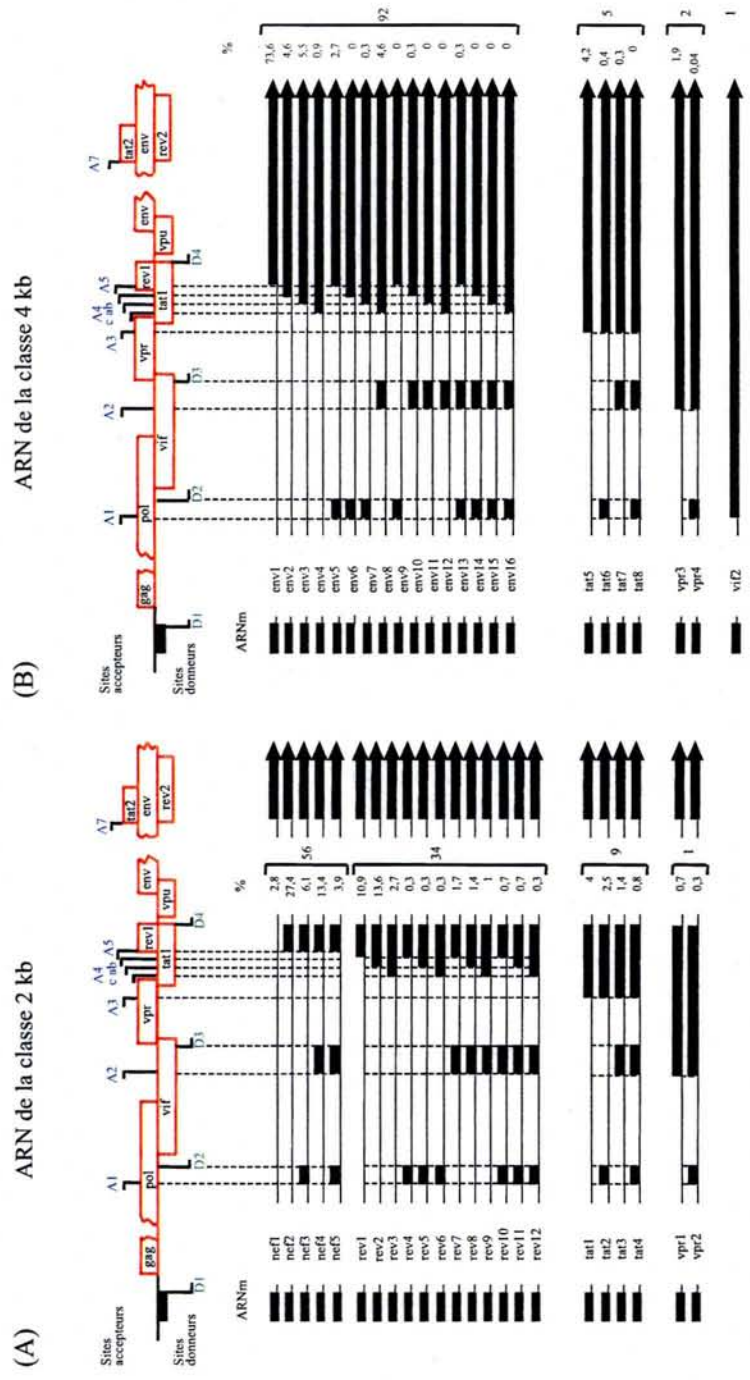

FIG. 11.1 – **Composition en exons et abondance relative des ARNm du virus HIV-1** d'après [Purcell & Martin, 1993]. Les différents variants des ARNm sont exprimés ainsi que les sites d'épissage correspondants.

Building and Analysing an Integrative Model of HIV-1 RNA Alternative Splicing

A. Bockmayr[1], A. Courtois[1], D. Eveillard[1,2], and M. Vezain[1]

[1] LORIA, UMR 7503 (CNRS, INRIA, Universités de Nancy), BP 239
54506 Vandœuvre-lès-Nancy, France
[2] Laboratoire de Maturation des ARN et Enzymologie Moléculaire,
UMR 7567 CNRS-UHP, BP 239, 54506 Vandoeuvre-lès-Nancy, France
e-mail: {bockmayr|acourtoi|eveillar|vezain}@loria.fr

Abstract. We present a multi-site model describing the alternative use of the RNA splicing sites A3, A4, A5 and A7 in the human immunodeficiency virus HIV-1. Our goal is to integrate experimental data obtained on individual splicing sites into a global model of HIV-1 RNA alternative splicing. We give a qualitative validation of our model, and analyse the possible impact of variations of regulatory protein concentrations on virus multiplication.

1 Introduction

The life cycle of the human immunodeficiency virus HIV-1 involves the production of different kinds of proteins. Among them the Tat and Nef proteins, which are lethal to the cell by inducing the apoptotic process. Their regulation is based on alternative splicing. Today, the control of the virus life cycle is only partially known. In order to improve existing knowledge on this finely regulated process, additional molecular biological experiments are needed. At present, the alternative splicing regulation is understood through experiments focusing on one particular splicing site. The global splicing behaviour is not studied experimentally, despite its importance for the HIV-1 life cycle. To overcome this difficulty, we propose an integrative modeling approach of alternative splicing regulation. Using an hybrid automaton with default reasoning, we construct a model that integrates the regulation at individual splicing sites into a multi-site model. Based on this modeling approach, we can study the impact of the regulation at one activator site, characterised by single-site experiments, on the production of the lethal HIV-1 proteins. In order to validate our approach, the model that covers different biological scales is analysed in a qualitative way.

The organisation of the paper is as follows. We start in Sect. 2 with a biological description of alternative splicing regulation. Based on a number of biological hypotheses in Sect. 3, we introduce in Sect. 4 an hybrid automaton model to describe the behaviour of the splicing process in a multi-site context. In Sect. 5, we analyse this model in a qualitative way, studying the impact of increasing the concentration of hnRNP A1 proteins. Based on biological queries and a model

checking approach, the model is validated in Sect. 6 in a qualitative way. Sect. 7 contains some conclusions and perspectives for future work.

Compared to our earlier work, the main contributions of this paper are as follows. While in [1, 2] we introduced our general modeling methodology, based on hybrid concurrent constraint programming, we develop here for the first time a realistic model of alternative splicing regulation in HIV-1, which integrates the sites A3, A4, A5, and A7. Furthermore, we give a qualitative analysis of this model to verify and discover different biological properties. Although further refinements are possible and desirable, this model can already be used to study *in silico* interesting biological questions, in particular the impact of variations of SR and hnRNP A1 regulatory protein concentrations on the virus life cycle.

2 Alternative Splicing: Regulation in Various Contexts

For eukaryotes and retroviruses such as HIV-1, the splicing reaction is a maturation process of the premessenger RNA (pre-mRNA), by elimination of non-coding sequences, introns, and junction of coding sequences, exons. The mature RNA (mRNA) can then be translated into a protein. The splicing reaction takes place in a large ribonucleoproteic complex called the spliceosome, which recognises signals on the pre-mRNA. Among them are the donor site (SD) and the acceptor site (SA), which border the non-coding intron [3]. Some pre-mRNAs are alternatively spliced, using competing splicing sites, which allows for the production of several different mRNAs from the same pre-mRNA. The selection of an alternative splicing site generally depends on a regulation ensured by the binding of an activator and/or an inhibitory protein. Members of the Serine-Arginine rich (SR) protein family (such as ASF/SF2 and SC35) are activator proteins, which help to recognise the regulated splicing site by the spliceosome. In contrast, the hnRNP A1 protein is well-known to inhibit splicing sites; in some cases, it acts by masking the site to the spliceosome [4].

For HIV-1 RNA splicing, the situation is more complex: the unique viral pre-mRNA contains 4 donor sites (D1 to D4) and 8 acceptor sites (A1 to A7), see Fig. 1. They can be used in different combinations to produce about 40 mRNAs, during 2 steps of the viral infection [5]. This controls the HIV-1 protein expression necessary to the production of a new virion. In the early phase, the viral pre-mRNA undergoes multiple splicing; the mature mRNAs produced have in common a splicing event between sites D4 and A7. Among them, *tat* mRNAs are also spliced at site A3 and contain two exons, coding an activator of the viral transcription, Tat. Other early mRNAs are the *nef* mRNAs spliced at site A5, and the *rev* mRNAs, spliced at any of the sites A4c, A4a, or A4b [5].

When enough Rev protein is synthesised, a shift from the early to the late phase of the infection occurs [6]. Indeed, the protein binds to the Rev Responsive Element (RRE) localised in the D4-A7 intron before this intron is spliced, and only incompletely spliced mRNAs characteristic of the late phase are produced. They all contain the D4-A7 intron and some of them are spliced at site A3,

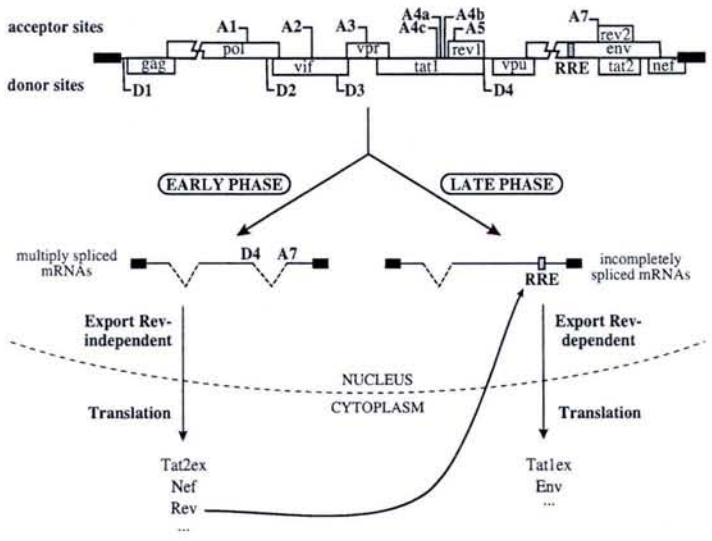

Fig. 1. Alternative splicing of HIV-1 RNA. Boxes delimitate the sequence of encoded proteins (ORF). In the mRNAs, spliced introns are designated by dotted lines, and exons by thick lines.

giving a Tat protein encoded by only one exon. Other sites (A4c, A4a, A4b and A5) are used to produce the *env* mRNAs [5].

The competing sites A3 to A5 are used to produce major viral proteins. Experimental studies show that these sites are regulated. Site A3 can be inhibited by the binding of the protein hnRNP A1 to the ESS2 (Exon Splicing Silencer) element, or activated by fixation of SR proteins to the ESEt (Exon Splicing Enhancer) element. SC35 proteins can also bind to the ESS2 element. They are expected to activate splicing at the site A3 by competing with the hnRNP A1 proteins binding to the ESS2 element.

3 Modeling Multi-Site Alternative Splicing Regulation

We model the regulation by SR and hnRNP A1 proteins in the multi-site context (A3, A4, A5, A7) under several hypotheses. We consider the biological process of alternative splicing in an *in vitro* context, which corresponds to the experimental knowledge. Our model describes the dynamics of a closed system of nucleic acid concentrations. The state variables are the pre-mRNAs (*rna*) which become mRNAs (*tat1, tat2, rev, env, nef*) of different proteins. The splicing process depends on the SC35, ASF/SF2 and hnRNP A1 regulatory proteins. They activate four splicing acceptors sites A3, A4, A5, and A7. We neglect the regulation at the donor sites because these seem to be less regulated than the acceptor sites [7, 5].

The four acceptor sites are in competition for the regulatory protein allocation. We consider this regulation level only on a single-site scale. Thus we assume the local behaviour as a continuous competition between four acceptor sites. The choice of one of them depends on the score of the *splice efficiency* defined as the ratio of the mature RNA and pre-mRNA at each splicing site [8]. For example, if the efficiency at A7 increases above a certain threshold τ, we assume that the A7 acceptor site is activated. We assume next that the splicing regulation of HIV-1 produces early mRNAs in the *early phase* of the life cycle. Under these cellular conditions, the combination of several acceptor sites produces different kinds of mature RNAs. Only the acceptor site with the highest efficiency is chosen by the spliceosomal complex. More precisely,

- splice activation of the A7 and the A3 site produces the mRNA *tat2* for the Tat protein with 2 exons.
- splice activation of the A7 and the A4 site produces the mRNA *rev* for the Rev protein.
- splice activation of the A7 and the A5 site produces the mRNA *nef* for the Nef protein.

We suppose that the concentration of the Rev protein is proportional to the Rev nucleic acid quantity (*rev*). The Rev protein represses the A7 splicing site. We assume in our local model that the quantity of the Rev nucleic acid reduces the efficiency of the A7 splicing site. This hypothesis corresponds to a time-scale abstraction of the translation to proteins, in order to focus on the dynamics of the splicing regulation.

If the A7 efficiency decreases below a certain threshold, the A7 splicing site is deactivated. Now we consider the biological system to be in the *late phase*. Without splicing activation at the A7 site:

- splice activation at the A3 site produces the mRNA *tat1* for the Tat protein with 1 exon.
- splice activation at the A4 or A5 site produces the mRNA *env* for the Env protein.

The choice between different acceptor sites depends on the concentration of regulatory proteins that control the splicing site activity. Modeling the regulation of the A3 site [2] allowed us to extract the splice efficiency eff_3, defined as the ratio of the mature RNA and pre-mRNA at equilibrium, as an observer variable of the local behaviour at the acceptor site A3. Depending on the concentration of hnRNP A1, eff_3 is given by the formula

$$eff_3(hnRNPA1) = \frac{\eta \cdot P_{ESE_t} \cdot (ASF+SC35) \cdot (k_{ESS2a} \cdot (k_{ESS2b}+SC35)+hnRNPA1 \cdot k_{ESS2b})}{\eta' \cdot (k_{ESE}+ASF+SC35) \cdot P_{ESS2} \cdot hnRNPA1 \cdot k_{ESS2b}} \quad (1)$$

A similar approach for the acceptor site A7 yields an efficiency function eff_7, integrating the effects of SR and Rev proteins:

$$eff_7(hnRNPA1, rev) = \frac{\frac{P_{ESE3} \cdot ASF}{k_{ESE3b} \cdot (1+\frac{hnRNPA1}{k_{ESE3}})+ASF} + \frac{P_{ESS3ab} \cdot SC35}{k_{ESS3ab}+SC35}}{\frac{P_{ISS} \cdot hnRNPA1}{k_{ISS}+hnRNPA1} + \frac{P_{ESS3b} \cdot hnRNPA1}{k_{ESS3b}+hnRNPA1} + \frac{P_{RRE} \cdot rev}{k_{RRE}+rev}} \quad (2)$$

In both formulas, $P_x > 0$ (resp. $k_x > 0$) are Michaelis-Menten constants that correspond to the maximal affinity (resp. half-saturation coefficient) at the binding site x. The concentrations of the regulatory proteins ASF/SF2, SC35, and hnRNP A1 are denoted by ASF, $SC35$, $hnRNPA1$, respectively. Finally, η, η' are kinetic constants.

The A4 and A5 acceptor sites are not well-studied yet, due to the complexity of the corresponding nucleic acid domain [9]. Their local behaviour is captured by the splice efficiencies eff_4 and eff_5, which are defined as interval constants. To model the relative interaction between those sites, we use default reasoning. Following [10], we describe A5 as a preferential activated site in the presence of regulatory proteins, and A4 as the default splicing site.

4 Integrative Model

To model the global alternative splicing regulation according to the hypotheses introduced in Sect. 3, we use an hybrid automaton with default reasoning, called $\mathcal{A}$, which is given in Fig. 2. There are five states corresponding to the production

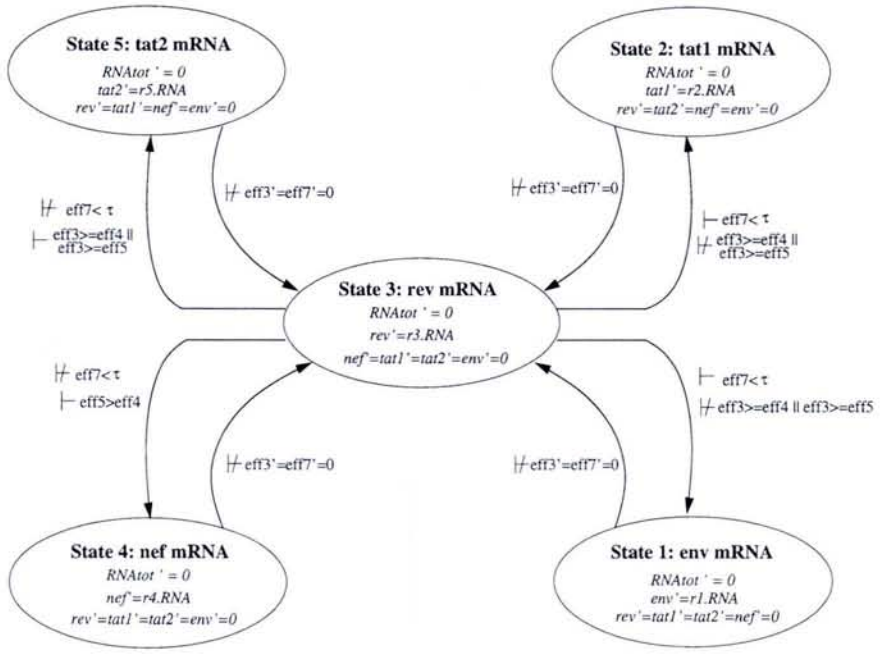

Fig. 2. Hybrid automaton $\mathcal{A}$ for the alternative splicing regulation in a multi-site context

(after splicing) of the mRNAs *rev, env, nef, tat1, tat2* for the proteins Rev, Env, Nef, Tat with 1 exon, and Tat with 2 exons, respectively. In each state, the dynamics is described by a set of differential equations, and a law for matter conservation.

The transitions from one state to another depend on constraints. We will consider two types of transitions from a state i to a state j:

- a transition $i \xrightarrow{\vdash c} j$ occurs if in state i the condition c holds.
- a transition $i \xrightarrow{\nvdash c} j$ occurs if in state i the condition c does not hold. This may have two reasons: either $\neg c$ holds or c is unknown (default reasoning).

  Equivalently, we will say that $\neg c$ holds *by default*, and then write $i \xrightarrow{\vdash^d \neg c} j$

For reasons of readability, we give in Fig. 2 only the default transitions. A complete description of the automaton can be found in Tab. 1.

To be in state $i \in \{1, \ldots, 5\}$, the constraint $\mathcal{D}_i(hnRNPA1, rev)$ from Tab. 1 has to be satisfied. Here, $eff_3(hnRNPA1)$ (resp. $eff_7(hnRNPA1, rev)$) are abbreviated by eff_3 (resp. eff_7). The vector $\boldsymbol{u} = (env, tat1, rev, nef, tat2)^T \in \mathbb{R}^5_{\geq 0}$ evolves in state i according to the system of differential equations

$$du_i/dt = r_i \cdot rna, \quad du_j/dt = 0, \text{ for } j \neq i \qquad (\mathcal{S}_i)$$

with rate constants $r_1, \ldots, r_5 > 0$. We also assume that the total quantity of RNA, $rna_{tot} = rna + env + tat1 + tat2 + nef + rev$ is invariant, i.e.

$$drna_{tot}/dt = 0 \qquad (3)$$

Using default reasoning, the continuous dynamics in state i is given by the system of differential equations $\mathcal{S}_i$, together with the equations (1) and (2) from Sect. 3, and (3).

Table 1. Formal description of the states of the hybrid automaton $\mathcal{A}$

State i	$\mathcal{D}_i(hnRNPA1, rev)$	Continuous system
1	$(eff_7 < \tau) \wedge (eff_3 < eff_5) \wedge (eff_3 < eff_4)$	$\vdash^d$ (1), (2), (3), $(\mathcal{S}_1)$
2	$(eff_7 < \tau) \wedge [(eff_3 \geq eff_5) \vee (eff_3 \geq eff_4)]$	$\vdash^d$ (1), (2), (3), $(\mathcal{S}_2)$
3	$(eff_7 \geq \tau) \wedge (eff_3 < eff_5 < eff_4)$	$\vdash^d$ (1), (2), (3), $(\mathcal{S}_3)$
4	$(eff_7 \geq \tau) \wedge (eff_3 < eff_4 \leq eff_5)$	$\vdash^d$ (1), (2), (3), $(\mathcal{S}_4)$
5	$(eff_7 \geq \tau) \wedge [(eff_3 \geq eff_5) \vee (eff_3 \geq eff_4)]$	$\vdash$ (1), (2), (3), $(\mathcal{S}_5)$

Throughout the paper, we assume that

- $\forall t \in \mathbb{R}_{\geq 0} : rna_{tot} > tat_1 + tat_2 + rev + env + nef$
- all parameters are strictly positive $(\mathcal{H}_{\mathcal{P}})$

The hybrid automaton $\mathcal{A}$ has been implemented in the hybrid concurrent constraint programming language `Hybrid cc` [11], see the part `GENERIC MODEL` in the program given in Fig. 3. In `Hybrid cc`, the pair of constraints $c \vdash^d A, \neg c \vdash B$ is noted as `unless` $\neg c$ A `else` B. Furthermore, f' denotes the time derivative df/dt of f. As argued in [1], constraint programming with *default logic* is well-suited for modeling biological systems because it allows us to handle partial or incomplete information. Each constraint gives one piece of information on the system that is studied. Using logical combinators, the hybrid automaton uses this information for the transition from one state to the other.

5 Model Analysis

Integrated modeling of local and global behaviours allows us to study alternative splicing regulation for different parameter variations. We may introduce in the hybrid automaton variations of hnRNP A1, ASF/SF2 or SC35 proteins based on the corresponding production rates. In order to validate our integrative model, we apply an infinite-time analysis to the hybrid automaton. This kind of analysis aims at answering biological queries about the system, without knowing a priori about numerical parameterisation and initial values. For the purpose of this paper, the generic model in Sect. 4 has been specialised by introducing a function $hnRNPA1(t)$ describing the variation of hnRNP A1:

$$d\,hnRNPA1(t)/dt = k_h \cdot hnRNPA1(t), \quad \text{with} \quad k_h > 0 \tag{4}$$

Other definitions of $hnRNPA1(t)$ would be possible. For our analysis, we only need that $hnRNPA1(t)$ is strictly growing, and that $hnRNPA1(0) > 0$. Indeed, the sign of the partial derivative $\partial eff_3/\partial hnRNPA1$ depends only on the sign of the derivative $hnRNPA1'$. So $hnRNPA1' > 0$ implies $\partial eff_3/\partial hnRNPA1 < 0$. The sign of $\partial^2 eff_7/\partial hnRNPA1\,\partial rev$ depends on the signs of $hnRNPA1'$ and rev'. So $hnRNPA1' > 0$ and $rev' \geq 0$ implies $\partial^2 eff_7/\partial hnRNPA1\,\partial rev < 0$. As eff_3 and eff_7 are monotonic, Tab. 2 below will not change, and the same analysis can be performed.

Note that since $hnRNPA1$ is now totally defined on $\mathbb{R}_{\geq 0}$, the default operator $\vdash^d$ is not needed anymore in the specialised program, i.e. we could replace the statement `unless` $\neg c$ A `else` B, with `if` c `then` A `else` B.

Our analysis of the hybrid automaton proceeds in several steps. To determine the possible behaviour in each individual state, we first analyse the signs of the (partial) derivatives and the limits of eff_3 and eff_7. From that we construct a transition table, which is then used to obtain transition graphs. Finally, we eliminate indeterminism by adding constraints that have been inferred during the analysis. Applications to biological queries about the system will be given in Sect. 6.

Variations of eff_3 and eff_7. Let $x \nearrow y$ (resp. $\searrow, \rightarrow$) denote that a given function is strictly increasing (resp. decreasing, constant), with greatest lower

```
hnRNPA1=...; tat2=...; tat1=...; rev=...; env=...; nef=...;
      /*** VARIABLE INITIALISATION at t=0 with interval constants ***/

always{                                          /*** GENERIC MODEL ***/
  unless (eff7 >= tau) {
    tat2'=0;  rev'=0;  nef'=0;
    unless (eff3 >= eff5 || eff3 >= eff4) {                  //State 1//
      tat1'=0;  env'=r1*(ARNtot-tat1-tat2-rev-env-nef);
    }
    else {                                                   //State 2//
      tat1'=r2*(ARNtot-tat1-tat2-rev-env-nef);
      env'=0;
    };
  }
  else {
    env'=0;  tat1'=0;
    unless(eff3 >= eff4 || eff3 >= eff5) {
      tat2'=0;
      unless (eff5 >= eff4) {                                //State 3//
        rev'=r3*(ARNtot-tat1-tat2-rev-env-nef);  nef'=0;
      }
      else {                                                 //State 4//
        nef'=r4*(ARNtot-tat1-tat2-rev-env-nef); rev'=0;
      };
    }
    else {                                                   //State 5//
      tat2'=r5*(ARNtot-tat1-tat2-rev-env-nef);  rev'=0;  nef'=0;
    };
  };
  eff3=(eta*PESEt*(ASF+SC35)*...;
  eff4=[v1,v3];                                    // 0 <= v1 <= v3
  eff5=[w1,w3];                                    // 0 <= w1 <= w3
  eff7=(PESE3*ASF/(KESE3b*...;
}

always {                                  /*** SPECIALISED MODEL ***/
  eff4=[v2,v2];  eff5=[w2,w2];          //v2 in [v1,v3], w2 in [w1,w3]
  hnRNPA1'=kh*hnRNPA1;
}
```

Fig. 3. Hybrid cc program implementing the hybrid automaton

bound x and least upper bound y. Let h_0 be the concentration of *hnRNPA1* at time $t = 0$, and r_0 (resp. r_{t_k}) the concentration of *rev* at time $t = 0$ (resp. $t = t_k$). Based on an analysis of the signs of the (partial) derivatives of eff_3 and eff_7 and using the lower and upper bounds

$$
\begin{aligned}
\alpha &= \mathit{eff}_3(h_0) \\
\beta &= lim_{hnRNPA1 \to +\infty} \mathit{eff}_3 = \tfrac{(ASF+SC35)\cdot P_{ESE_t}\cdot \eta}{\eta'\cdot(k_{ESE}+ASF+SC35)\cdot P_{ESS2}} \\
\gamma &= \mathit{eff}_7(h_0, r_0) \\
\delta_{t_k} &= lim_{hnRNPA1 \to +\infty} \mathit{eff}_7(r_{t_k}) \\
&= \tfrac{P_{ESS3_{ab}}\cdot SC35\cdot(k_{RRE}+r_{t_k})}{(k_{ESS3_{ab}}+SC35)(P_{ISS}\cdot k_{RRE}+P_{ISS}\cdot r_{t_k}+P_{ESS3_b}\cdot k_{RRE}+P_{ESS3_b}\cdot r_{t_k}+P_{RRE}\cdot r_{t_k})} \\
\varepsilon &= lim_{(hnRNPA1,rev)\to(+\infty,+\infty)} \mathit{eff}_7 = \tfrac{P_{ESS3_{ab}}\cdot SC35}{(P_{RRE}+P_{ISS}+P_{ESS3_b})\cdot(k_{ESS3_{ab}}+SC35)}
\end{aligned}
$$

the qualitative behaviour in the different states is given in Tab. 2.

Table 2. Qualitative behaviour in the different states

	t	*hnRNPA1*	*rev*	eff_3	eff_7
States1, 2, 4, 5	$0 \nearrow +\infty$	$h_0 \nearrow +\infty$	$r_{t_k} \to r_{t_k}$	$\alpha \searrow \beta$	$\gamma \searrow \delta_{t_k}$
State3			$r_0 \nearrow +\infty$		$\gamma \searrow \varepsilon$

For all $(hnRNPA1, rev) \in \mathbb{R}_{>0} \times \mathbb{R}_{\geq 0}$, we have $\varepsilon < \mathit{eff}_7(hnRNPA1, rev) < \gamma$ and $\varepsilon < \delta_{t_k} < \gamma$.

Using the table above, we can determine under which conditions the automaton will stay definitely in a given state i, denoted $i \circlearrowleft_\infty$:

State 1: Since eff_7 and eff_3 decrease, the system will stay in state 1 once it has reached it.
State 2: The system will stay in state 2, if $\beta \geq \max\{\mathit{eff}_4, \mathit{eff}_5\}$.
State 3: The system will stay in state 3, if $\varepsilon \geq \tau$.
State 4: The system will stay in state 4, if $\delta_{t_{k'}} \geq \tau$, for some $t_{k'} > 0$.
State 5: The system will stay in state 5, if $\delta_{t_{k''}} \geq \tau$, for some $t_{k''} > 0$, and $\beta \geq \max\{\mathit{eff}_4, \mathit{eff}_5\}$.

Transitions from state i to state j. Next we analyse which transitions are possible in the hybrid automaton. Tab. 3 summarises these results. Col. 1 indicates the transition $i \to j$ in question. Col. 2 gives necessary conditions for this transition, for a given time t. Cols. 3 to 5 recall the behaviour of *rev*, eff_3, eff_7 in state i. As a conclusion, Col. 6 indicates whether or not transition $i \to j$ is possible. Failing conditions in Col. 2 are underlined.

Table 3. Table of transitions

$T_{i \to j}$	Conditions on eff_3 and eff_7 for $i \to j$ transition	rev^i	$eff_3{}^i$	$eff_7{}^i$	possible
1 $\circlearrowleft_\infty$	$(eff_7 < \tau) \wedge (eff_3 < eff_5) \wedge (eff_3 < eff_4)$	$\to$	$\searrow$	$\searrow$	yes
$1 \to 2$	$\underline{eff_3 \nearrow} \wedge\ (eff_3 = eff_5{}^- \vee eff_3 = eff_4{}^-)$	$\to$	$\searrow$	$\searrow$	no
$1 \to 3$	$\underline{eff_7 \nearrow} \wedge\ (eff_7 = \tau^-)$	$\to$	$\searrow$	$\searrow$	no
$1 \to 4$	$\underline{eff_7 \nearrow} \wedge\ (eff_7 = \tau^-)$	$\to$	$\searrow$	$\searrow$	no
$1 \to 5$	$\underline{eff_7 \nearrow} \wedge\ \underline{eff_3 \nearrow} \wedge\ (eff_7 = \tau^-)$ $\wedge\ (eff_3 = eff_5{}^- \vee eff_3 = eff_4{}^-)$	$\to$	$\searrow$	$\searrow$	no
$2 \to 1$	$eff_3 \searrow \wedge\ (eff_3 = eff_5 \vee eff_3 = eff_4)$	$\to$	$\searrow$	$\searrow$	yes
2 $\circlearrowleft_\infty$	$(eff_7 < \tau) \wedge (eff_3 \geq eff_5 \vee eff_3 \geq eff_4)$ $\wedge\ (\beta \geq \max\{eff_4, eff_5\})$	$\to$	$\searrow$	$\searrow$	yes
$2 \to 3$	$\underline{eff_7 \nearrow} \wedge\ eff_3 \searrow \wedge\ (eff_7 = \tau^-)$ $\wedge\ (eff_3 = eff_5 \vee eff_3 = eff_4)$	$\to$	$\searrow$	$\searrow$	no
$2 \to 4$	$\underline{eff_7 \nearrow} \wedge\ \underline{eff_3 \nearrow} \wedge\ (eff_7 = \tau^-)$ $\wedge\ (eff_3 = eff_5 \vee eff_3 = eff_4)$	$\to$	$\searrow$	$\searrow$	no
$2 \to 5$	$\underline{eff_7 \nearrow} \wedge\ (eff_7 = \tau^-) \wedge (eff_3 \geq eff_5 \vee eff_3 \geq eff_4)$	$\to$	$\searrow$	$\searrow$	no
$3 \to 1$	$\underline{eff_7 \searrow} \wedge (eff_7 = \tau) \wedge (eff_3 < eff_4)$	$\nearrow$	$\searrow$	$\searrow$	yes
$3 \to 2$	$\underline{eff_7 \searrow} \wedge \underline{eff_3 \nearrow} \wedge\ (eff_7 = \tau) \wedge\ (eff_3 = eff_4{}^-)$	$\nearrow$	$\searrow$	$\searrow$	no
3 $\circlearrowleft_\infty$	$(eff_7 \geq \tau) \wedge\ (eff_3 < eff_4 < eff_5) \wedge (\varepsilon \geq \tau)$	$\nearrow$	$\searrow$	$\searrow$	yes
$3 \to 4$	$\underline{eff_4 \nearrow} \vee\ \underline{eff_5 \searrow}$	$\nearrow$	$\searrow$	$\searrow$	no
$3 \to 5$	$\underline{eff_3 \nearrow} \wedge\ (eff_7 \geq \tau) \wedge (eff_3 = eff_4{}^-)$	$\nearrow$	$\searrow$	$\searrow$	no
$4 \to 1$	$\underline{eff_7 \searrow} \wedge (eff_7 = \tau) \wedge\ (eff_3 < eff_5)$	$\to$	$\searrow$	$\searrow$	yes
$4 \to 2$	$\underline{eff_7 \searrow} \wedge \underline{eff_3 \nearrow} \wedge\ (eff_7 = \tau) \wedge\ (eff_3 = eff_5{}^-)$	$\to$	$\searrow$	$\searrow$	no
$4 \to 3$	$(eff_7 \geq \tau) \wedge (\underline{eff_4 \searrow} \vee\ \underline{eff_5 \nearrow})$	$\to$	$\searrow$	$\searrow$	no
4 $\circlearrowleft_\infty$	$(eff_7 \geq \tau) \wedge (eff_3 < eff_5 \leq eff_4) \wedge (\delta_{k'} \geq \tau)$	$\to$	$\searrow$	$\searrow$	yes
$4 \to 5$	$\underline{eff_3 \nearrow} \wedge\ (eff_7 \geq \tau) \wedge\ (eff_3 = eff_5{}^-)$	$\to$	$\searrow$	$\searrow$	no
$5 \to 1$	$eff_7 \searrow \wedge\ eff_3 \searrow \wedge (eff_7 = \tau)$ $\wedge\ (eff_3 = eff_5 \vee eff_3 = eff_4)$	$\to$	$\searrow$	$\searrow$	yes
$5 \to 2$	$eff_7 \searrow \wedge (eff_7 = \tau) \wedge (eff_3 \geq eff_5 \vee eff_3 \geq eff_4)$	$\to$	$\searrow$	$\searrow$	yes
$5 \to 3$	$eff_3 \searrow \wedge (eff_7 \geq \tau) \wedge\ (eff_3 = eff_4) \wedge\ (eff_4 < eff_5)$	$\to$	$\searrow$	$\searrow$	yes
$5 \to 4$	$eff_3 \searrow \wedge (eff_7 \geq \tau) \wedge\ (eff_3 = eff_5) \wedge\ (eff_5 \leq eff_4)$	$\to$	$\searrow$	$\searrow$	yes
5 $\circlearrowleft_\infty$	$(eff_7 \geq \tau) \wedge (eff_3 \geq eff_5 \vee eff_3 \geq eff_4)$ $\wedge\ (\delta_{t_{k''}} \geq \tau) \wedge (\beta \geq \max\{eff_4, eff_5\})$	$\to$	$\searrow$	$\searrow$	yes

The expression τ^- means that the limit τ is reached from below. eff_3 and eff_7 are abbreviations for $eff_3(hnRNPA1)$ and $eff_7(hnRNPA1, rev)$.

Finally we observe that for $i \in \{1, .., 4\}$, transitions $i \to i+1$ are not allowed. So $\delta_{t_{k''}}$ and $\delta_{t_{k'}}$ are unique if they exist, and $\delta_{t_{k''}} \geq \delta_{t_{k'}}$.

Transition graphs. To classify the qualitative behaviour of the hybrid automaton, we introduce the following five constraints, which have been obtained during the analysis given before:

$c_1 \equiv (\mathit{eff}_4 < \mathit{eff}_5)$, $c_2 \equiv (\beta \geq \max\{\mathit{eff}_4, \mathit{eff}_5\})$, $c_3 \equiv (\varepsilon \geq \tau)$,
$c_4 \equiv (\delta_{t_{k'}} \geq \tau)$, $c_5 \equiv (\delta_{t_{k''}} \geq \tau)$

Each of these constraints may be satisfied or not, depending on the initial values and the parameterisation. This yields $2^5 = 32$ theoretically possible behaviours. Since $\delta_{t_{k''}} \geq \delta_{t_{k'}}$, we have $c_4 \Rightarrow c_5$. Furthermore, if c_1 holds, then c_4 is not relevant. Therefore 20 different behaviours remain, see Tab. 4. Here, the relation symbols $<, \geq$ indicate whether a constraint or its negation holds. An irrelevant value is marked by "$-$". After simplification, we obtain 11 possible transition graphs, which are given in Fig. 4.

Table 4. The 20 behaviours depending on constraints. *Def states* means that the system will stay in the corresponding state, once it has reached it.

c_1	c_2	c_3	c_4	c_5	Def states	c_1	c_2	c_3	c_4	c_5	Def states	c_1	c_2	c_3	c_4	c_5	Def states
$<$	$<$	$\geq$	$-$	$\geq$	1, 3	$<$	$<$	$\geq$	$-$	$<$	1, 3	$<$	$<$	$<$	$-$	$\geq$	1
$<$	$<$	$<$	$-$	$<$	1	$<$	$\geq$	$\geq$	$-$	$\geq$	1, 2, 3, 5	$<$	$\geq$	$\geq$	$-$	$<$	1, 2, 3
$<$	$\geq$	$<$	$-$	$\geq$	1, 2, 5	$<$	$\geq$	$<$	$-$	$<$	1, 2	$\geq$	$<$	$\geq$	$\geq$	$\geq$	1, 4
$\geq$	$<$	$\geq$	$<$	$\geq$	1	$\geq$	$<$	$\geq$	$<$	$<$	1	$\geq$	$<$	$<$	$\geq$	$\geq$	1, 4
$\geq$	$<$	$<$	$<$	$\geq$	1	$\geq$	$<$	$<$	$<$	$<$	1	$\geq$	$\geq$	$\geq$	$\geq$	$\geq$	1, 2, 4, 5
$\geq$	$\geq$	$\geq$	$<$	$\geq$	1, 2, 5	$\geq$	$\geq$	$\geq$	$<$	$<$	1, 2	$\geq$	$\geq$	$<$	$\geq$	$\geq$	1, 2, 4, 5
$\geq$	$\geq$	$<$	$<$	$\geq$	1, 2, 5	$\geq$	$\geq$	$<$	$<$	$<$	1, 2						

Adding new constraints. Note that 7 graphs are indeterministic. If the automaton is initialised in state 5, the successor state could be 1, 2, 3 or 4. In order to simplify the queries in Sect. 6, we eliminate this indeterminism by introducing additional constraints. Define
$t_m = \min\{t \in \mathbb{R}_{\geq 0} \mid \mathit{eff}_3(\mathit{hnRNPA1}(t)) = \mathit{eff}_4 \text{ or } \mathit{eff}_3(\mathit{hnRNPA1}(t)) = \mathit{eff}_5\}$, if the minimum exists, and $t_m = -1$, otherwise. Given the constraints

$$c_6 \equiv (t_m \geq 0 \wedge \mathit{eff}_7(\mathit{hnRNPA1}(t_m), r_0) \geq \tau) \text{ and}$$
$$c_7 \equiv (t_m \geq 0 \wedge \mathit{eff}_7(\mathit{hnRNPA1}(t_m), r_0) = \tau),$$

the successor state of 5 is $\begin{cases} 1, & \text{if } c_7 \\ 2, & \text{if } \neg c_6 \\ 3 \text{ or } 4, & \text{if } c_6 \wedge \neg c_7 \end{cases}$

Together with $c_1, \ldots, c_5$, the constraints c_6, c_7 allow us to define $25 = 3 \times 7 + 4$ deterministic transition graphs, which cover the possible behaviours in the deterministic case.

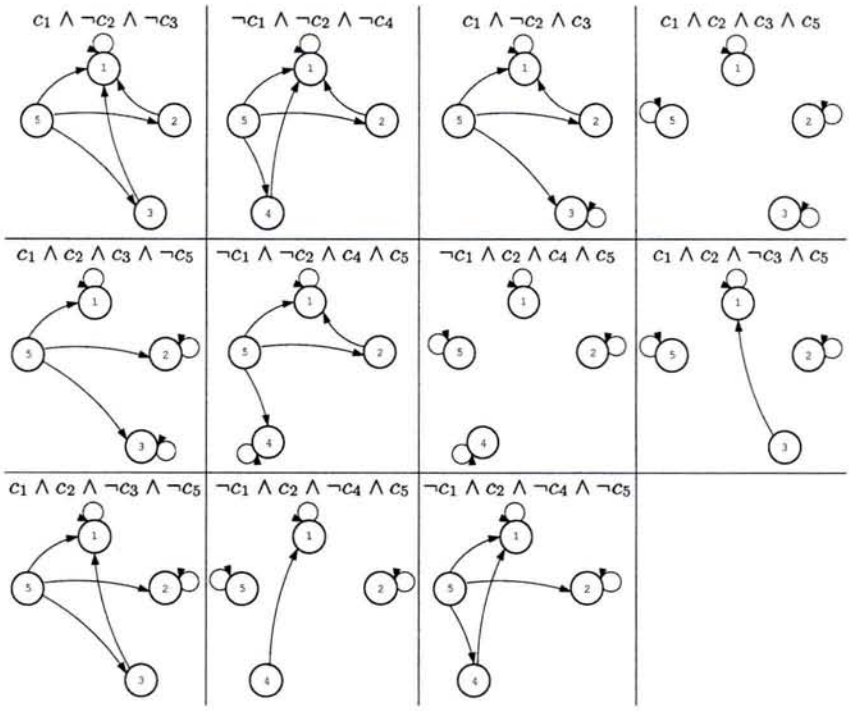

Fig. 4. Transition graphs

6 Biological Queries and Property Synthesis

Now we are able to answer biological queries, which could be expressed in CTL [12]. The result will be either a failure $\varnothing$, which means that the property is always false, or parameterisation and initialisation conditions will be *synthesised* that correspond to the desired behaviour. These conditions involve either the constraints $c_1, \ldots, c_7$ or the initialisation hypotheses $\mathcal{H}_i, i = 1, \ldots, 5$, expressing that $\mathcal{D}_i(h_0, r_0)$ is satisfied.

Suppose for example that we want to determine how one can reach state 2. From Tab. 4, we can see that are two possibilities: either the automaton starts already in state 2 ($\mathcal{H}_2$), or it starts in state 5 ($\mathcal{H}_5$), and then transits to state 2. In order to get a transition from 5 to 2, constraint $\neg c_6$ must be satisfied. This leads to an expression of the form

$$\mathcal{H}_2 \vee [\, \mathcal{H}_5 \wedge \neg c_6 \wedge \mathcal{C}\,].$$

To determine $\mathcal{C}$, we number the 11 transition graphs in Tab. 4 from left to right and from top to bottom, keeping only the graphs 1, 2, 3, 5, 6, 9, 11 which enable the transition $5 \rightarrow 2$. Thus $\mathcal{C}$ is a disjunction of constraints describing the remaining graphs :

$$\mathcal{C} = \begin{pmatrix} c_1 \wedge \neg c_2 \wedge \neg c_3 \\ \vee\ \neg c_1 \wedge \neg c_2 \wedge \neg c_4 \\ \vee\ c_1 \wedge \neg c_2 \wedge c_3 \\ \vee\ c_1 \wedge c_2 \wedge c_3 \wedge \neg c_5 \\ \vee\ \neg c_1 \wedge \neg c_2 \wedge c_4 \wedge c_5 \\ \vee\ c_1 \wedge c_2 \wedge \neg c_3 \wedge \neg c_5 \\ \vee\ \neg c_1 \wedge c_2 \wedge \neg c_4 \wedge \neg c_5 \end{pmatrix}$$

After simplification of $\mathcal{C}$, we obtain

$$\mathcal{H}_2 \vee \left[\mathcal{H}_5 \wedge \neg c_6 \wedge \begin{pmatrix} c_1 \wedge \neg c_5 \\ \vee\ \neg c_2 \wedge c_5 \\ \vee\ \neg c_4 \wedge \neg c_5 \end{pmatrix} \right]$$

Further examples of biological queries and answers are given in Tab. 5.

Table 5. Synthesising conditions to satisfy the goals

Query/Goal	Conditions
1) Reach state 3	$[\mathcal{H}_5 \wedge c_1 \wedge (\neg c_2 \vee \neg c_5) \wedge c_6 \wedge \neg c_7] \vee \mathcal{H}_3$
1') Reach and exit from state 3	$\neg c_3 \wedge [(\mathcal{H}_5 \wedge c_1 \wedge (\neg c_2 \vee \neg c_5) \wedge c_6 \wedge \neg c_7) \vee \mathcal{H}_3]$
2) Transit from {1,2} to {3,4,5}	$\varnothing$
3) Reach state 2	$\mathcal{H}_2 \vee [\mathcal{H}_5 \wedge \neg c_6 \wedge$ $(c_1 \wedge \neg c_5 \vee \neg c_2 \wedge c_5 \vee \neg c_4 \wedge \neg c_5)]$

Query 1 yields the well-known qualitative biological conditions to be in the state which produces the *rev* mRNA [13]. In order to be validated, our model should give a positive answer to this query and produce some sufficient conditions.

Query 2 corresponds to a question where we expect a negative answer. Similarly to Query 1, the result is a priori well-known. It concerns the crucial switch between the early and late phase in the virus life cycle. In a validated model, the virus cannot switch back from the late to the early phase. Thus the negative answer agrees with biological knowledge.

Correct answers to such queries give a partial validation of the integrative model. Furthermore, the hybrid automaton allows us to synthesise sufficient conditions for a given property.

Since our model is in accordance with existing biological knowledge, we may also ask a different type of query, where we do not know the answer yet. Query 3 can be seen as a biological exploration based on a formal model. The answer shows that, given the production of hnRNP A1 proteins, state 2 can be reached only from state 5. Thus, *tat1* mRNA production is only possible after *tat2* mRNA production. Our model suggests that by increasing the concentration of the repressor proteins hnRNP A1, the system may switch from the early to the late phase. This observation corresponds to a new biological hypothesis, which should be verified experimentally.

7 Conclusions and Perspectives

Our approach combines single-site and multi-site modeling approaches of the alternative splicing regulation. The integration is achieved by an hybrid automaton with default reasoning, in accordance with available biological knowledge. The splice efficiency is used as a time-scale abstraction of the local behaviour at one site inside a more global multi-site model. For the experimental biologist, this integrative model serves as a computational tool to study a fine-grained biological process on different scales. In particular, one can analyze the effect of the local regulation at one site on the global regulation involving different sites. On the one hand, the above queries may validate the biological hypotheses underlying the model. On the other hand, they may suggest future experiments by focusing on one particular process of interest, which remains difficult for classical approaches.

Our integrative model is partially validated by biological queries on its qualitative behaviour. An hybrid automaton appears to be an efficient model to represent the switching conditions between the early and the late phase, despite the lack of numerical values.

The present qualitative analysis concerns only one instantiation of the generic hybrid automaton, which consists in increasing the concentration of hnRNP A1 proteins. Thus our model is validated only in this situation. In order to generalise our approach, we could study the effect of other regulatory proteins in order to extract novel biological hypotheses, which may initiate new experimental work.

In a virological study of the HIV-1 life cycle, Hammond [14] describes the variation of the proteins translated from mRNAs. His study does not take into account the effect of alternative splicing. In future work, we plan to introduce the alternative splicing regulation in Hammond's model of the virus life cycle. The variation of the viral proteins can be represented by additional constraints in the hybrid automaton. Using this approach, our goal is to develop a more dynamic model of alternative splicing regulation, and to quantify the effects of this complex biological process in the HIV-1 life cycle.

References

1. Bockmayr, A., Courtois, A.: Using hybrid concurrent constraint programming to model dynamic biological systems. In: 18th International Conference on Logic Programming, ICLP'02, Copenhagen. Springer LNCS 2401 (2002) 85–99
2. Eveillard, D., Ropers, D., Jong, H.d., Branlant, C., Bockmayr, A.: Multiscale modeling of alternative splicing regulation. In: Computational Methods in Systems Biology, CMSB'03, Rovereto, Italy. Springer LNCS 2602 (2003) 75–87. Long version to appear in *Theoretical Computer Science.*
3. Moore, M., Query, C., Sharp, P.: Splicing of precursors to messenger RNAs by the spliceosome. In: The RNA World. Cold Spring Harbor Laboratory Press (1993)
4. Smith, C.W., Valcarcel, J.: Alternative pre-mRNA splicing: the logic of combinatorial control. Trends in Biochemical Sciences **25** (2000) 381–388

5. Purcell, D., Martin, M.: Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication and infectivity. J. Virol. **67** (1993) 6365–78
6. Hope, T.: The ins and outs of HIV rev. Arch. Biochem. Biophys. **365** (1999) 186–91
7. O'Reilly, M., McNally, M., Beemon, K.: Two strong 5' splice sites and competing, suboptimal 3' splice sites involved in alternative splicing of human immunodeficiency virus type 1 RNA. Virology **213** (1995) 373–85
8. Si, Z., Amendt, B.A., Stoltzfus, C.M.: Splicing efficiency of human immunodeficiency virus type 1 tat RNA is determined by both a suboptimal 3' splice site and a 10 nucleotide exon splicing silencer element located within tat exon 2. Nucleic Acids Res **25** (1997) 861–7
9. Swanson, A.K., Stoltzfus, C.M.: Overlapping *cis* sites used for splicing of HIV-1 env/nef and rev mRNAs. J. Biol. Chem. **273** (1998) 34551–7
10. Schaal, H., Freund, M., Kammler, S., Asang, C., Caputi, M.: A bidirectional SR protein-dependent exonic splicing enhancer regulates *rev*, *env*, *vpu* and *nef* gene expression. In: Eukaryotic mRNA Processing. (2003)
11. Gupta, V., Jagadeesan, R., Saraswat, V.: Computing with continuous change. Science of Computer Programming **30** (1998) 3–49
12. Chabrier, N., Fages, F.: Symbolic model checking of biochemical networks. In: Computational Methods in Systems Biology, CMSB'03, Rovereto, Italy. Springer LNCS 2602 (2003) 149–162
13. Pongoski, J., Asai, K., Cochrane, A.: Positive and negative modulation of human immunodeficiency virus type 1 Rev function by *cis* and *trans* regulators of viral RNA splicing. J. Virol. **76** (2002) 5108–20
14. Hammond, B.J.: Quantitative study of the control of HIV-1 gene expression. J. Theor. Biol. **163** (1993) 199-221

11.2.2 Analyse qualitative formelle de modèles intégrés

L'analyse qualitative du modèle multi-site démontre que la transition entre la phase précoce et la phase tardive de l'épissage ne présente pas d'incohérences avec les connaissances expérimentales. Il est impossible pour le système de passer en phase précoce à partir de la phase tardive. Cette requête doit être négative. C'est une approche qui permet de valider le modèle. Certaines hypothèses concernant la dynamique du système sont issues de l'analyse de ce modèle. Le modèle permet d'interroger le système d'un point de vue dynamique. Le comportement dynamique est directement issu des hypothèses statiques. Valider le comportement dynamique par rapport à des expériences est donc le gage d'une bonne compréhension statique du système. Dans l'optique de mieux comprendre la régulation de l'épissage alternatif et ces conséquences sur le comportement global, le modèle multi-site qui peut correspondre à une construction expérimentale, est donc un outil efficace pour tester *a priori* des hypothèses de régulation.

Le modèle multi-site étant validé, on peut supposer que le comportement local sur un comportement multi-site est partiellement validé. Le comportement du passage entre la phase précoce et tardive peut donc être assimilé par l'automate hybride. Il peut être intéressant de se placer dans un cadre biologique tel que le cycle de vie de HIV-1 afin de tester l'effet de notre modélisation de la régulation de l'épissage sur le virus HIV-1.

11.3 Influence de la régulation de l'épissage alternatif dans le cycle de vie de HIV-1

Le modèle multi-site est incomplet. De nombreuses hypothèses sont trop stringentes par rapport à la réalité biologique du cycle de vie de HIV-1. Néanmoins, le modèle possède la vertu de pouvoir analyser et valider le comportement multi-site avant de l'intégrer dans un modèle multi-échelle dont le référentiel biologique est de plus grande dimension. Il convient dans ce but de formaliser un automate hybride qui tienne compte des facteurs biologiques que nous avons mentionné dans le Chapitre 2, comme l'amplification de la transcription par la protéine Tat, ou l'accélération du passage de la phase précoce à la phase tardive par fixation de la protéine Rev à l'élément RRE.

Afin de modéliser plus finement le cycle cellulaire du virus, nous nous sommes inspirés du modèle de [Hammond, 1993]. Ce modèle est formalisé comme un système continu à compartiments. Le système dans ce formalisme est relativement précis et possède de nombreuses valeurs numériques concernant les paramètres cinétiques. Néanmoins, il ne formalise pas la régulation de l'épissage alternatif et il ne permet pas de représenter significativement le passage de la phase précoce vers la phase tardive qui est une étape essentielle dans le cycle de vie du virus. En collaboration avec Myriam Vezain, nous avons intégré le modèle multi-site avec le modèle de [Hammond, 1993]. Il en résulte un automate hybride dont la structure est représentée dans la Figure 11.2. On y retrouve les différentes étapes de production des protéines virales.

L'absence de valeurs numériques concernant les paramètres cinétiques, nous oblige à nous restreindre uniquement au comportement qualitatif du modèle multi-échelle. Les valeurs numériques mentionnées sur les graphiques 11.3 et 11.4 sont donc uniquement

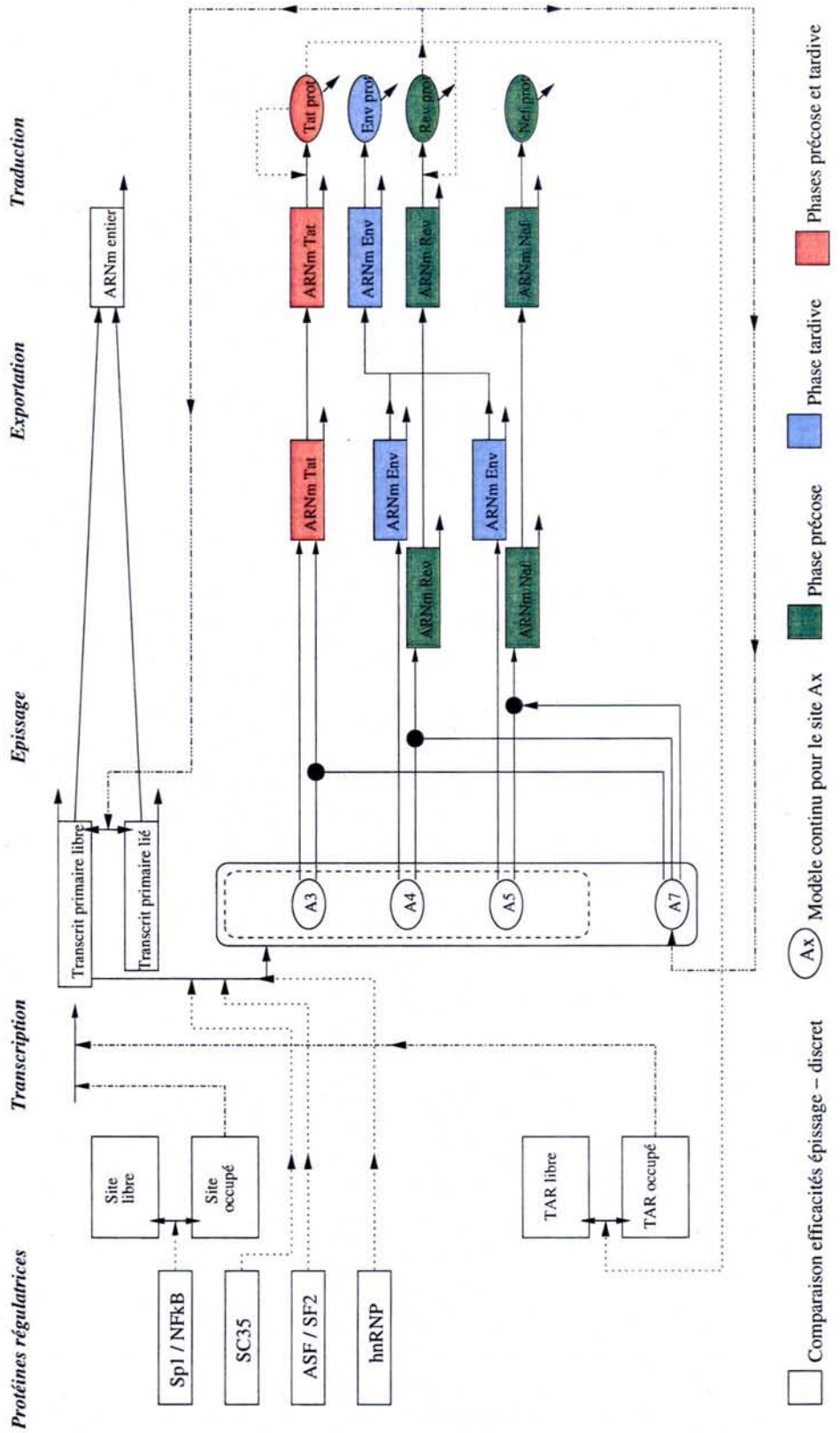

FIG. 11.2 – **Schéma conceptuel du modèle multi-échelles du cycle de vie de HIV-1.**

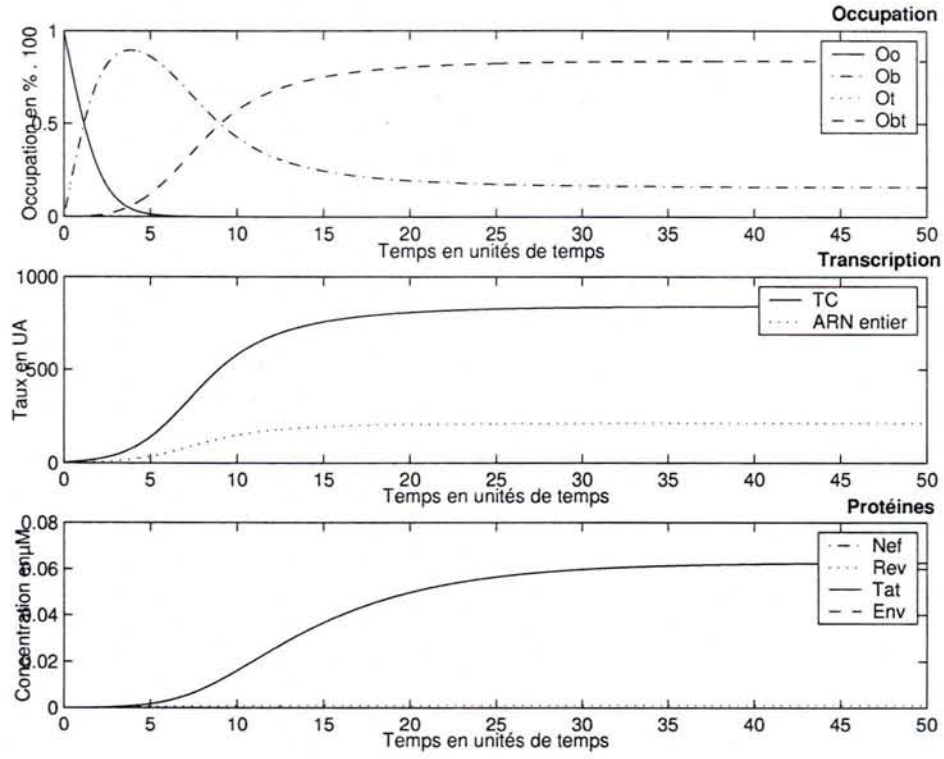

FIG. 11.3 – **Comportement dynamique du modèle multi-échelle**. On observe la variation de 3 paramètres. L'**occupation** correspond à la variation (en échelle $\times 100$ %) de l'occupation des principaux sites de régulation. O_o correspond l'occupation d'aucun site, O_b correspond à l'occupation du site de fixation des protéines SP1 et NFκB, O_t correspond à l'occupation du site TAR qui fixe la protéine Tat. La **transcription** au cours du cycle de vie de HIV-1 s'observe grâce à la variable TC qui correspond au taux de transcription du virus et de la quantité d'ARN non épissé qui deviendra le génome viral. Le modèle multi-échelles permet d'apprécier la variation du taux de production de **protéines** virales.

qualitativement indicatives et elles ne reflètent en aucun cas une quelconque prédiction numérique du comportement multi-échelles. Malgré ce biais, il est cependant possible d'observer qualitativement certaines variations et une modification de comportements pour des variations de quantité de protéines SR.

Avec les valeurs numériques issues de [Hammond, 1993], il est possible avec notre modèle multi-échelle de retrouver le comportement qualitatif du modèle de cycle de vie de HIV-1, comme l'illustre la Figure 11.3. Le comportement est qualitativement similaire au modèle de [Hammond, 1993] concernant la variation de l'occupation et de la transcription. Les valeurs numériques ne correspondent pas au modèle du cycle de vie, ce qui corrobore le fait que pour de nouvelles contraintes ajoutées au système, il soit important de re-paramétrer le modèle. Le modèle multi-échelle permet d'analyser les variations des productions des différentes protéines virales comme la protéine Tat dans la Figure 11.3.

Une fois le comportement du cycle de vie cohérent avec les données existantes, il est

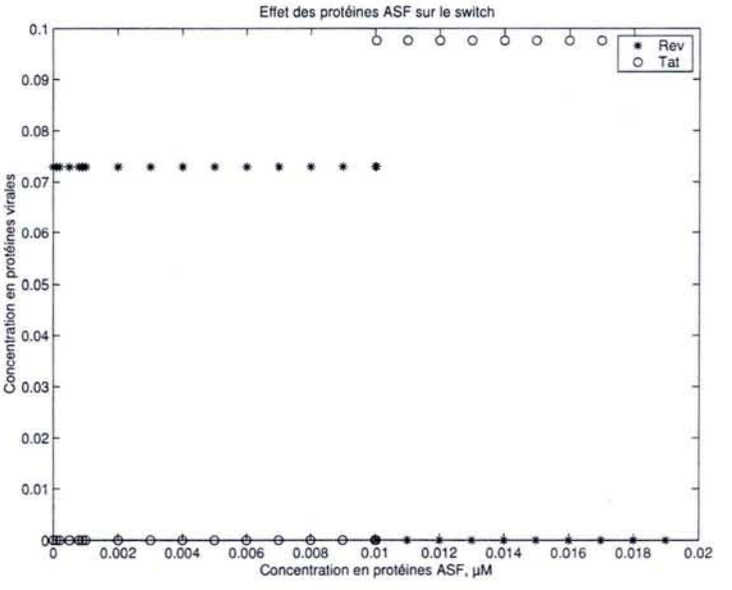

FIG. 11.4 – **Effet de l'augmentation de la quantité de protéine ASF/SF2 sur le cycle viral.** Une augmentation de la quantité d'ASF/SF2 modifie le type de protéine produite. Le virus produit initialement la protéine Rev (∗). Pour une quantité suffisamment grande de protéine régulatrice ASF/SF2 (seuil ici de 0,01), le virus produit la protéine Tat (∘). A partir de ce seuil, le taux de production de Rev s'annule pour laisser une production constante de Tat.

intéressant de tester les effets de la régulation locale sur le comportement du cycle de vie du virus. La variation des protéines de régulation affecte le cycle de manière inégale. Nous proposons de tester la variation de la protéine ASF/SF2 sur le cycle du virus. Les effets d'une augmentation de la quantité de protéine SR sont représentés dans la Figure 11.4.

On assiste alors à une modification du comportement qualitatif (*switch*) pour une augmentation de la protéine SR. Ces protéines, nous l'avons, vu agissent au niveau de la régulation moléculaire et peuvent donc produire des conséquences au niveau viral. La modification peut être de grande importance si l'on considère que la protéine Rev est une protéine virale qui conditionne le passage du virus de la phase précoce vers la phase tardive. L'augmentation de la quantité de protéine ASF/SF2, dans les conditions biologiques qui sont celles du modèle, permettrait de ralentir le passage en phase tardive.

Il est cependant important de donner la juste valeur de cette hypothèse. Elle résulte de la modélisation des hypothèses statiques existantes. Le modèle n'étant pas encore validé et il ne peut être considérée que comme une piste expérimentale à exploiter. Néanmoins, il est possible de formuler que de manière théorique la régulation de l'épissage au niveau des sites de régulation possède indéniablement des conséquences sur la régulation du cycle de vie du virus. La modélisation formelle multi-échelles atteste de l'importance de cet axe de recherche dans la compréhension de la dynamique du virus.

11.4 Conclusions

Cette étude de modélisation formelle n'est pas aboutie. Le modèle multi-échelle n'est pas validé. En effet, on est actuellement juste en mesure d'observer le comportement par simulation sans pouvoir encore raisonner. Cela s'explique par la difficulté de ce modèle qui est un automate hybride trop complexe pour les outils de validation à notre disposition. Des théoriciens de la modélisation travaillent actuellement sur les outils d'analyse qualitative sous forme de *model checking* qui pourraient valider la formalisation du modèle multi-échelles.

Le modèle multi-échelles illustre donc seulement un cas particulier de la modélisation sur plusieurs échelles qui a été initié et validé avec le modèle de régulation d'épissage multi-sites. L'intérêt de ce type de modèle est de représenter les résultats que donnent les travaux expérimentaux actuels. En effet, un travail formel sur les expériences et les hypothèses qui en découlent permettent souvent d'aller plus loin que la simple description de résultats expérimentaux. L'analyse formelle permet de mettre en lumière nombre de conséquences biologiques qui peuvent être triviales mais également importantes. Dans ce contexte, le modèle peut être assimilé par les biologistes comme un outil formel qui permet de tester les conséquences des hypothèses qu'ils sont amenés à formuler. L'objectif sera alors de restreindre les expériences sur celles qui permettent de valider la perception du système biologique. D'un point de vue biologique, malgré l'absence de nombreux paramètres, le modèle multi-échelles démontre que la régulation de l'épissage alternatif au niveau des sites d'épissage possède des conséquences sur le cycle de vie du virus HIV-1.

Cinquième partie

Discussion

Chapitre 12

Vers une méthodologie biologique *in silico*

Le contexte pluridisciplinaire de ces travaux soulève des questions pratiques qui ont été développées à la fin de chaque chapitre. Il soulève également des perpectives méthodologiques. Notre démarche possède la caractéristique de couvrir différents domaines informatiques afin de répondre à une demande biologique. Cet aspect transversal se retrouve également dans la nature des connaissances biologiques que nous avons été amenés à traiter.

12.1 Approche méthodologique transversale

Notre méthodologie pour étudier la régulation de l'épissage alternatif va au-delà des domaines prédéfinis par l'enseignement académique. Dans notre approche se croisent des concepts statistiques et formels dans le seul but de répondre à une problématique biologique. Notre démarche *in silico* est donc l'occasion de faire un bilan des différentes interactions disciplinaires qui permettent de proposer des solutions.

12.1.1 Approche informatique intégrée

L'approche que nous avons utilisée repose sur les modélisations statistique et formelle qui sont des domaines informatiques relativement disjoints. Néanmoins, avec l'interaction de ces domaines, il est possible d'extraire un protocole qui est associé à la gestion des connaissances. A partir des résultats expérimentaux, il est important d'extraire l'information qui servira de connaissance biologique. Cette connaissance est souvent intuitive en fonction de la nature des expériences. En effet, les résultats sont souvent des conséquences des protocoles expérimentaux. Par exemple, lors d'une migration sur gel retard, les résultats sont directement interprétés : la migration du complexe dépend de ces propriétés. Mais lors d'expériences telles que les expériences SELEX qui sont semi-automatiques, il est important d'analyser pour extraire la connaissance. Une démarche de classification est dans ce cadre la plus appropriée afin de comprendre les données pour générer différentes hypothèses concernant la structure de l'information qui est sous-jacente. Les outils de

classification sont alors plus ou moins adaptés pour déterminer cette structure.

Les connaissances extraites, les modèles statistiques permettent de les généraliser comme nous l'avons fait avec les méthodes à noyaux comme les SVM. Néanmoins, cette étape sera pertinente si et seulement si les connaissances peuvent être discriminées. Cela correspond à avoir une démarche de classification rigoureuse qui démontre que les données expérimentales contiennent des connaissances discriminantes.

Le fait de généraliser est une étape de validation en soi. Ceci est illustré dans notre cas d'étude où les résultats généralisés vont permettre de mettre en évidence des régions sur les génomes qui possèdent des sites de régulation d'épissage. La cohérence de ces résultats estimés avec les données biologiques est le gage d'une approche rigoureuse et de la valeur des connaissances qui ont été extraites. La discrimination revêt dans ce cas un rôle de prédiction qui permet de valider les hypothèses. C'est un autre type de validation que celle qui est recherchée avec la conceptualisation de modèles formels. Dans un sens, ces deux approches sont relativement complémentaires pour valider les connaissances d'un système vivant puisque les modèles formels permettent de valider les hypothèses d'un point de vue qualitatif et les modèles statistiques d'un point de vue quantitatif. On effectue ainsi une synthèse des connaissances actuelles pour en vérifier le comportement dans un cadre dynamique. Pour ce faire, on intègre l'approche dans la méthodologie *in silico* suivante :

- Extraction des connaissances biologiques de données expérimentales avec un **modèle de classification**.
- Généralisation des connaissances avec un **modèle de discrimination**.
- Validation des connaissances dans un contexte dynamique avec un **modèle formel**.

Cette méthodologie générale que nous proposons pour étudier un problème biologique est possible uniquement par l'utilisation de formalismes qui permettent de passer d'un domaine statistique au domaine formel de l'informatique. Chaque formalisme est issu de travaux théoriques informatiques. Ainsi, derrière chaque outil, il existe des théorèmes ou lemmes qui définissent des domaines d'applications concernant le traitement et la nature des informations que l'outil peut étudier. Pour illustration, on peut citer les approches connexionnistes type SVM, les machines probabilistes et les machines basées sur l'algorithmique des mots qui sont toutes les trois employées dans la recherche de motifs biologiques. Elles ne raisonnent cependant pas sur le même type d'information. En effet les SVM raisonnent sur des vecteurs qui sont moins spécifiques que les autres méthodes qui raisonnent sur les séquences. La nature de la connaissance et l'objectif de la méthode sont des critères de premier ordre dans le choix de l'outil que l'on va utiliser. Malheureusement, cette réflexion est souvent absente des démarches bioinformatiques qui appliquent les outils de traitement de séquences biologiques sans maîtriser les fondements théoriques de chaque méthode.

Il apparaît donc après ces travaux de thèse que pour chaque problème biologique, il subsiste une formalisation du problème qui permet d'obtenir des résultats plus ou moins appropriés. Il convient alors de déterminer dans un premier temps le problème biologique pour ensuite appliquer la méthode ou les formalismes qui correspondent. La démarche inverse ne donnera pas de résultats satisfaisants. Ainsi, si l'on est capable de faire une cartographie des problèmes biologiques, il serait intéressant de superposer un diagramme des méthodes et objectifs correspondants. Cette mise en corrélation des méthodes et des objets d'étude est sans doute un des enjeux des méthodologies bioinformatiques de demain

pour proposer la méthode optimale à chaque problème. C'est par cette structuration du domaine que l'application des méthodes bioinformatiques sera reconnue comme un outil expérimental au même titre que les autres approches issues de la chimie combinatoire. Par ailleurs, cette définition des problèmes biologiques en corrélation des méthodes, permettra de fournir des bases de travail solides pour les théoriciens du domaine permettant la conceptualisation de problèmes fondamentaux en bioinformatique.

12.1.2 Une nouvelle méthodologie biologique

Si les méthodes bioinformatiques sont reconnues par les biologistes, il est intéressant de pouvoir les situer dans la méthodologie biologique. Les résultats que l'on obtient par l'utilisation de méthodes informatiques doivent être validés par les expériences. Quelles que soient les validations que nos approches puissent faire, les hypothèses que l'on arrive à extraire d'un système vivant doivent être expérimentées à nouveau afin de produire de nouveaux résultats. La méthodologie est donc une approche qui se situe en amont des expériences, ce qui permet de la considérer comme un protocole pré-expérimental. Sa fonction dans l'approche globale du biologiste est de restreindre les expériences possibles. L'ajout de ce protocole permet alors de gagner en rentabilité puisque de nombreuses hypothèses peuvent être testées *in silico* pour en certifier la consistance, avant d'être testées *in vitro*. De la même manière les modèles statistiques peuvent permettre d'isoler les régions de génomes qui sont pertinents à tester comme nous l'avons vu avec KOALAB. La méthodologie bioinformatique sera donc d'autant plus efficace que les interactions entre les différentes méthodes seront fortes. Les futures expériences biologiques devront alors intégrer l'approche *in silico* afin de planifier le temps entre les expériences et les formalisations informatiques et ceci dans le but d'optimiser la compréhension d'un processus biologique. Cette conclusion peut être illustrée par les expériences SELEX qui fournissent des données d'apprentissage de première qualité aux machines d'apprentissage statistique. Une interaction efficace entre les expériences et les méthodes informatiques permet d'obtenir une approche efficace, et ce malgré la difficulté du problème biologique.

12.1.3 Critiques de l'approche bioinformatique

Néanmoins, la méthodologie *in silico* comporte de nombreuses difficultés. Certaines sont inhérentes au caractère pluridisciplinaire de la bioinformatique. En effet, la réussite de la modélisation repose sur le degré d'interactions entre la biologie et l'informatique. Or les objectifs des deux domaines ne sont pas naturellement complémentaires. La biologie est demandeuse de résultats rapides et pertinents pour avancer dans la connaissance des systèmes vivants. L'informatique est quant à elle demandeuse de problèmes séduisants qui permettront de développer de nouvelles théories ou d'appliquer des outils de haute technologie existants. Par une caricature excessive, les biologistes ont le sentiment de se faire déposséder de leurs problèmes alors que les informaticiens ont le sentiment de restreindre leurs compétences à la mise à disposition de technologies. La méthodologie bioinformatique se situant à l'interface, doit préserver l'équilibre subtil entre une recherche efficace biologiquement mais également théoriquement satisfaisante. Pour pallier ce problème, la formalisation de la biologie est encore une fois une alternative efficace. Elle permet une

base de discussion commune entre les deux communautés. L'interaction lors de la modélisation du problème biologique permet à chaque domaine d'influer sur le problème commun.

La formalisation commune entre les biologistes et les informaticiens empêche également de dévier le problème bioinformatique vers une problématique purement théorique. Il est en effet relativement aisé de travailler sur des concepts sans avoir recours à des données expérimentales. Or les résultats de biologie expérimentale restreignent fortement les domaines de réflexion théorique. Pour illustration, il est possible de raisonner de manière formelle aux techniques de discrimination sans que les résultats soient significatifs Par ailleurs, la problématique biologique peut contraindre de manière trop spécifique une méthode informatique. Dans ce cas, les résultats ne seront pas généralisables et la théorie pas suffisamment efficace pour que la méthode bioinformatique développée puisse être appliquée à des problèmes voisins. Pour illustration, le développement d'une méthode qui permet de retrouver le site de fixation de la protéine ASF/SF2 doit pouvoir être facilement généralisable aux autres protéines SR.

12.2 Perspectives

Malgré ces incertitudes, il est possible d'envisager diverses perspectives de travail. Elles représentent des futurs axes de recherche qui concernant les différentes thématiques que nous avons abordées au cours de ces travaux.

12.2.1 Outils d'aide à la décision pour les expériences SELEX

L'approche de classification permet d'extraire des connaissances biologiques des données SELEX mais également de proposer une représentation des motifs qui soit plus cohérente avec la réalité biologique. La difficulté des expériences SELEX réside donc dans la détermination du nombre de cycles expérimentaux optimal. L'approche de classification que nous avons proposée peut être perçue comme un outil d'aide à la décision pour l'expérimentateur. Il convient alors dans cette optique, de mieux caractériser les seuils statistiques qui permettent de séparer des ensembles de séquences nucléiques homogènes. Par exemple, à partir d'un certain seuil statistique d'homogénéité des séquences dans un ensemble, on pourra considérer que le nombre de cycles d'expériences SELEX est optimum. Afin d'optimiser ce critère, il est important de travailler sur la dissémination de l'information dans les ensembles homogènes au cours des différents cycles de sélection.

Actuellement, notre approche permet de classer les séquences nucléiques issues des expériences SELEX dans des ensembles statistiquement homogènes. Nous avons vu que notre approche permet de retrouver une structure classificatoire sous-jacente qui est motivée par des critères de structure secondaire. Il est possible d'associer d'autres critères dans la matrice de comparaison des séquences nucléiques. Cette opération permettrait de faire une véritable analyse de données SELEX. On peut dans cette optique ajouter des critères biologiques qualitatifs comme des classes de structures secondaires associées aux séquences nucléiques, ou des propriétés physico-chimiques. Il ressortira de cette analyse des connaissances biologiques qui permettront d'expliquer de manière plus fine la struc-

ture de classification. Cet aspect méthodologique est relativement facilement implantable. Le logiciel qui en découle sera alors un complément graphique de premier choix à l'outil d'aide à la décision des expériences SELEX.

12.2.2 Exploitation de classification par la discrimination automatique

L'exploitation des données SELEX par la classification permet d'isoler des ensembles homogènes qui sont discriminants. On est alors en mesure de généraliser la connaissance sous-jacente aux données par une méthode de discrimination comme celle qui est notamment présentée dans KOALAB. La machine, une fois l'apprentissage statistique effectué, est donc spécifique au problème biologique que représentent les données biologiques. En utilisant les données SELEX pour toutes les protéines SR, notre approche discriminante est semblable aux outils comme ESEfinder de [Cartegni *et al.*, 2003]. La démarche diffère cependant par la nature de la méthode employée. ESEfinder est un outil de *pattern matching*, alors que nous utilisons avec les SVM un outil de *pattern recognition*. Il ne nous est donc pas nécessaire de spécifier les motifs que l'on recherche avant d'effectuer la recherche dans un génome. Néanmoins, l'étape de classification préliminaire est dans ce cas primordiale.

L'utilisation de la méthode de classification en tant que module de pré-traitement des données SELEX dans KOALAB est donc une des perspectives intéressantes pour valider les connaissances sous-jacentes à la classification, en retrouvant les motifs connus dans les génomes, comme les sites ESE ou ESS identifiés expérimentalement. Cette validation peut être étendue par l'agglomération d'une machine HMM dans KOALAB. Il sera alors possible de rechercher des motifs qui seront cette fois-ci mieux définis car ils auront été pré-analysés par l'approche de classification-discrimination. L'emploi de cette méthode permettra de valider une fois de plus nos connaissances. Dans un cadre plus général, KOALAB possèderait alors un échantillonnage des méthodes les plus significatives dans la recherche de motifs biologiques dans les génomes. L'intégration de ces méthodes sur une même plate-forme logicielle laisse envisager la possibilité d'une analyse théorique et pratique des différents résultats pour un type de motifs biologiques. On sera alors en mesure de quantifier les avantages des méthodes les unes par rapport aux autres en fonction de l'hétérogénéité nucléique des motifs ou de leurs conservations. Les biologistes disposeraient alors d'un outil graphique ergonomique qui leur permettrait de piloter des technologies de haut niveau appropriées à la problématique biologique.

12.2.3 Détection statistique et fonctionnelle de motifs

Les motifs biologiques qui sont détectés par les modèles statistiques de discrimination sont ensuite validés par les expériences. Les expérimentateurs associent des propriétés biologiques à ces sites en fonction de leur position par rapport à des éléments ayant un rôle de balise dans un génome. Dans notre cas, la présence d'un site de fixation de protéine SR dans un exon sera considérée comme un élément ESE.

Il est envisageable à court terme de proposer une démarche automatique pour associer des propriétés des motifs que nous aurons identifiés par analyse discriminante. Il sera

ensuite important de prendre en compte divers paramètres comme la structure des ARN qui permet de cacher ou d'exposer des sites de fixation des protéines SR, et le cas échéant les rapprocher dans l'espace des sites d'épissage. Il faudra également tenir compte de la proximité de sites de fixation des protéines régulatrices. Cette proximité peut présenter des conséquences contradictoires. Si les sites fixent des protéines similaires, ils pourront avoir un rôle dans la polymérisation. Par ailleurs, si les sites fixent des protéines antagonistes, l'encombrement stérique donnera lieu à une compétition inhibitrice. Tous ces paramètres sont des contraintes qu'il est possible de modéliser grâce à des outils d'optimisation de contraintes ou des HMM. On serait alors en mesure de générer une carte fonctionnelle des sites de régulation d'épissage. Cette carte permettrait alors de guider *in silico* les expérimentateurs dans les génomes eucaryotes encore inexploités.

12.2.4 Modélisation générique par des systèmes de contraintes

Les sites de régulation de l'épissage possèdent des propriétés fonctionnelles qui sont actuellement déterminées exclusivement par des expériences. Nous venons de mentionner que la fonction était étroitement corrélée à la position de ces motifs par rapport aux autres motifs de régulation, mais aussi par rapport aux sites d'épissage. Les différentes combinaisons des positions relatives des sites de fixation des protéines déterminent les interactions fonctionnelles entre les sites. On peut considérer la présence de systèmes génériques qui correspondent à ces combinaisons. Ainsi, si l'on considère les protéines deux à deux, on peut identifier les phénomènes de compensation, d'inhibition, d'inhibition compétitive, sur-activation. Ces différentes possibilités correspondent aux fonctions qu'il faudra combiner pour avoir une représentation de la régulation locale de l'épissage. Cette combinaison est la source de la complexité combinatoire de la régulation de l'épissage alternatif.

Il est possible de modéliser formellement ces interactions avec un formalisme qualitatif de manière à représenter sans valeurs numériques le comportement des régulations. On isolerait ainsi les différentes régulations élémentaires qui sont possibles que l'on considère comme des modules élémentaires. Les modules peuvent être agencés dans un modèle pour simuler le comportement de la régulation de l'épissage dans une région. On peut caractériser ces modules dans un logiciel qui permettra de les agencer via une interface graphique. L'expérimentateur pourra alors construire un modèle qualitatif de régulation d'épissage uniquement en manipulant ces modules élémentaires de régulation qui seront alors des *pattern* fonctionnels mais également des systèmes de contraintes. Cet outil sera alors une démarche de modélisation pré-expérimentale qui permet d'observer rapidement les conséquences des hypothèses formulées. Il servira aussi de base commune de réflexions entre biologistes et informaticiens, initiant ainsi le développement d'un modèle plus fin.

12.2.5 Analyse automatique des propriétés d'un système biologiques

L'utilisation de la programmation par contraintes hybrides permet de formuler les systèmes biologiques comme des automates hybrides. Cette formalisation permet de représenter des systèmes formalisés par des contraintes continus comme les équations dif-

férentielles, ou des contraintes discretes. Le langage de programmation possède donc une certaine flexibilité qui permet de simuler des systèmes, et ce malgré les incertitudes encore existantes. Mais un des avantages des contraintes et de permettre un raisonnement sur le système ainsi modélisé. Ce raisonnement permet la validation du système qui est l'étape la plus critique dans le processus de modélisation. Il est actuellement possible de simuler facilement les systèmes programmés en *Hybrid cc*. Mais comme nous l'avons mentionné dans les chapitres précédents, la simulation n'est pas suffisante pour valider un modèle. Il faut raisonner. L'avantage des modèles formalisés par des contraintes et de composer naturellement un automate hybride. Dès lors il existe de nombreuses approches théoriques qui permettent de raisonner sur les automates telles que les procédés de vérification. Les modèles biologiques représentent alors une forme d'inspiration pour de nouvelles méthodes théoriques d'analyses. Un nouveau domaine informatique s'ouvre donc à la biologie pour valider les modèles.

Mais cette opération de vérification est souvent délicate pour les non-spécialistes. Il convient donc de proposer des protocoles de vérification automatique dans des logiciels de programmation afin que les modèles puissent conjointement être formalisés, testés puis validés par la même personne indépendamment de ces compétences informatiques. Cette opération de vérification testera des propriétés connues des expérimentateurs mais proposera également des hypothèses de comportement qui sont inhérentes au modèle. Le protocole de modélisation pourra donc dans cette optique proposer automatiquement des hypothèses biologiques à tester. Ces hypothèses satisfaites renforceraient la validation *a posteriori*.

Bibliographie

[Aizerman *et al.*, 1964] Aizerman, M., Braverman, E. & Rozonoer, L. (1964) Theoretical foundations of the potential function and pattern recognition learning. *Automaton and Remote Control,* **25**, 821–837.

[Akutsu *et al.*, 1999] Akutsu, T., Miyano, S. & Kuhara, S. (1999) Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium of Biocomputing*, (Altman, R., Lauderdale, K., Dunker, A., Hunter, L. & Klein, T., eds), vol. 4, pp. 17–28 World Scientific Publishing, Singapore.

[Audibert *et al.*, 2002] Audibert, A., Weill, D. & Dautry, F. (2002) In vivo kinetics of mrna splicing and transport in mammalian cells. *Molecular and Cellular Biology,* **22** (19), 6706–6718.

[Baeza-Yates & Navarro, 2004] Baeza-Yates, R. & Navarro, G. (2004) *Text searching : theory and practice.* Physica-Verlag, Heidelberg to appear.

[Bailey & Elkan, 1994] Bailey, T. L. & Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymer. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* pp. 28–36.

[Bailey & Gribskov, 1998] Bailey, T. L. & Gribskov, M. (1998) Methods and statistics for combining motif match scores. *Journal of Computational Biology,* **5**, 211–221.

[Baldi & Brunak, 2001] Baldi, P. & Brunak, S. (2001) *Bioinformatics The machine learning approach.* second edition,, MIT press.

[Bartlett, 1998] Bartlett, P. (1998) The sample complexity of pattern classification with neural networks : the size of the weigths is more important than the size of the network. *IEEE transactions on Information Theory,* **44** (2), 525–536.

[Batt *et al.*, 2003] Batt, G., Jong, H. d., Geiselmann, J. & Page, M. (2003) Qualitative analysis of genetic regulatory networks : a model-checking approach,. In *Working Notes of Seventeenth International Workshop on Qualitative Reasoning, QR-03*, (Bredeweg, B. & Salles, P., eds), pp. 31–38, Brasilia, Brazil.

[Ben-Hur *et al.*, 2001] Ben-Hur, A., Horn, D., Siegelmann, H. T. & Vapnik, V. (2001) Support vector clustering. *Journal of machine learning research,* **2**, 125–137.

[Berget, 1995] Berget, S. M. (1995) Exon recognition in vertebrate splicing. *J Biol Chem,* **270** (6), 2411–4.

[Berget & Sharp, 1977] Berget, S. M. & Sharp, P. A. (1977) A spliced sequence at the 5'-terminus of adenovirus late mRNA. *Brookhaven Symp Biol,* **29**, 332–44.

[Bernard & Gouzé, 1995*a*] Bernard, O. & Gouzé, J.-L. (1995*a*) Transient behavior of biological loop models with application to the droop model. *Mathematical Bioscience,* **127**, 19–43.

[Bernard & Gouzé, 1995*b*] Bernard, O. & Gouzé, J.-L. (1995*b*) Transient behaviour of biological models as a tool of qualitative validation - application to the droop model and to a n-p-z model. *Journal of Biological Systems,* **4** (3), 303–314.

[Bernard & Gouzé, 1998] Bernard, O. & Gouzé, J.-L. (1998). Global qualitative behavior of a class of non linear biological systems ; application to the qualitative validation of phytoplankton growth models.

[Bernard & Gouzé, 1999] Bernard, O. & Gouzé, J.-L. (1999) Non-linear qualitative signal processing for biological systems : application to the algal growth in bioreactors. *Mathematical Biosciences,* **157**, 357–372.

[Bernard & Gouzé, 2002] Bernard, O. & Gouzé, J.-L. (2002) Global qualitative description of a class of nonlinear dynamical systems. *Artificial Intelligence,* **136**, 29–59.

[Black, 2000] Black, D. L. (2000) Protein diversity from alternative splicing : a challenge for bioinformatics and post-genome biology. *Cell,* **103** (3), 367–70.

[Bockmayr & Courtois, 2001] Bockmayr, A. & Courtois, A. (2001) Modeling biological systems in hybrid concurrent constraint programming. In *2nd International Conference on Systems Biology, ICSB'01* vol. (abstract)„ Caltech.

[Bockmayr & Courtois, 2002] Bockmayr, A. & Courtois, A. (2002) Using hybrid concurrent constraint programming to model dynamic biological systems. In *18th International Conference on Logic Programming, ICLP'02* vol. 2401, pp. 85–99 LNCS Springer, Copenhagen.

[Bockmayr *et al.*, *a*] Bockmayr, A., Courtois, A., Eveillard, D. & Vezain, M. Building and analysis an integrative model of HIV-1 RNA alternative splicing. submit to CMSB'04.

[Boser *et al.*, 1992] Boser, B., Guyon, I. & Vapnik, V. (1992) A training algorithm for optimal margin classifiers. In *COLT'92* pp. 144–152.

[Boyer & Moore, 1977] Boyer, R. & Moore, S. (1977) A fast string searching algorithm. *Communication of the ACM,* **20**, 762–772.

[Burke *et al.*, 1996] Burke, D., Scates, L., Andrews, K. & Gold, L. (1996) Bent pseudoknots and novel RNA inhibitors of type 1 human immunodeficiency virus (HIV-1) reverse transcriptase. *J. Mol. Biol.,* **264**, 650–666.

[Caputi *et al.*, 1999*a*] Caputi, M., Mayeda, A., Krainer, A. & Zahler, A. (1999*a*) hnRNPA proteins are required for inhibition of HIV-1 pre-mRNA splicing. *EMBO,* **18** (14), 4060–4067.

[Caputi *et al.*, 1999*b*] Caputi, M., Mayeda, A., Krainer, A. R. & Zahler, A. M. (1999*b*) hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing. *EMBO J,* **18** (14), 4060–7.

[Carlson & Gupta, 1998] Carlson, B. & Gupta, V. (1998) Hybrid cc and interval constraints. In *Hybrid Systems : Computation and Control, HSCC'98* pp. 80 – 95 Springer, LNCS 1386.

[Cartegni *et al.*, 2003] Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q. & Krainer, A. R. (2003) ESEfinder : a web resource to identify exonic splicing enhencers. *Nucleic Acids Research,* **31** (13), 3568–3571.

[Cavaloc *et al.*, 1999*a*] Cavaloc, Y., Bourgeois, C. F., Kister, L. & Stevenin, J. (1999*a*) The splicing factors 9g8 and srp20 transactivate splicing through different and specific enhancers. *RNA,* **5**, 468–483.

[Cavaloc *et al.*, 1999*b*] Cavaloc, Y., Bourgeois, C. F., Kister, L. & Stévenin, J. (1999*b*) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA,* **5** (3), 468–83.

[Cavaloc *et al.*, 1994] Cavaloc, Y., Popielarz, M., Fuchs, J. P., Gattoni, R. & Stévenin, J. (1994) Characterization and cloning of the human splicing factor 9G8 : a novel 35 kDa factor of the serine/arginine protein family. *EMBO J,* **13** (11), 2639–49.

[Chabrier & Fages, 2003] Chabrier, N. & Fages, F. (2003) Symbolic model checking of biochemical networks. In *Computational Methods in Systems Biology*, (Priami, C., ed.), vol. 2602, pp. 149–162 LNCS, Rovereto.

[Chaudhary *et al.*, 1991] Chaudhary, N., McMahon, C. & Blobel, G. (1991) Primary structure of a human arginine-rich nuclear protein that colocalizes with spliceosome components. *Proc Natl Acad Sci U S A,* **88** (18), 8189–93.

[Clément & Codani, 1997] Clément, C. & Codani, J.-J. (1997) Lassap, a large scale sequence comparaison package. *CABIOS,* **2**, 137–147.

[Cornish-Bowden, 1995] Cornish-Bowden, A. (1995) *Fundamentals of Enzyme Kinetics.* Portland Press, London.

[Cortes & Vapnik, 1995] Cortes, C. & Vapnik, V. (1995) Support-vector networks. *Machine Learning,* **20**, 273–297.

[Crochemore *et al.*, 1994] Crochemore, M., Czumaj, A., Gasieniec, L., Jarominek, S., Lecroq, T., Plandowski, W. & Rytter, W. (1994) Speeding up two string matching algorithms. *Algorithmica,* **12** (4/5), 247–267.

[Danos & Laneve, 2003] Danos, V. & Laneve, C. (2003) Graphs for core molecular biology. In *Computational Methods in Systems Biology*, (Priami, C., ed.), vol. 2602, pp. 34–46 LNCS, Rovereto.

[Deutsch & Long, 1999] Deutsch, M. & Long, M. (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res,* **27** (15), 3219–28.

[Dreyfuss *et al.*, 2002] Dreyfuss, G., Kim, V. N. & Kataoka, N. (2002) Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol,* **3** (3), 195–205.

[Durbin *et al.*, 1998] Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998) *Probabilistic models of proteins and nucleic acids.* Cambridge university press.

[Duret & Abdeddaim, 1999] Duret, L. & Abdeddaim, S. (1999) Multiple alignments for structural, functional, or phylogenetic analyses of homologous sequences. In *Bioinformatics : Sequence, structure, and databanks.*, (Higgins, D. & Taylor, W., eds), The pratical approach series. Oxford university press pp. 51–76.

[Edwards *et al.*, 2001] Edwards, R., Siegelmann, H., Aziza, K. & Glass, L. (2001) Symbolic dynamics and computation in model gene networks. *Chaos,* **11** (1), 160–169.

[Elisseef *et al.*, 1999] Elisseef, A., Guermeur, Y. & Paugam-Moisy, H. (1999). Margin error and generalization capabilities of multi-class discriminant models. Technical report NC-TR-99-051-R NeuroCOLT2.

[Eperon *et al.*, 2000] Eperon, I. C., Makarova, O. V., Mayeda, A., Munroe, S. H., Cáceres, J. F., Hayward, D. G. & Krainer, A. R. (2000) Selection of alternative 5' splice sites : role of u1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol Cell Biol,* **20** (22), 8303–18.

[Eveillard & Guermeur, 2002*a*] Eveillard, D. & Guermeur, Y. (2002*a*) Traitement statistique des résultats SELEX. In *JOBIM*, (Nicolas, J. & Thermes, C., eds), pp. 277–283, St Malo.

[Eveillard & Guermeur, 2002*b*] Eveillard, D. & Guermeur, Y. (2002*b*) Statistical processing of SELEX results. In *ISMB*, Edmonton.

[Eveillard *et al.*, *a*] Eveillard, D., Larhlimi, A., Ropers, D., Billaut, S. & Peyrefitte, S. KOALAB : a new method for regulatory research, illustration on alternative splicing regulation in HIV-1. soumis à JOBIM'04.

[Eveillard *et al.*, 2004] Eveillard, D., Ropers, D., de Jong, H., Branlant, C. & Bockmayr, A. A multi-scale constraint programming model of alternative splicing regulation. in press in Journal of Theoretical Computer Science.

[Eveillard *et al.*, 2002] Eveillard, D., Ropers, D., Jong, H. d. & Bockmayr, A. (2002) Modeling the effects of SR proteins on alternative splicing. In *Symposium on Macromolecular Network.*

[Eveillard *et al.*, 2003] Eveillard, D., Ropers, D., Jong, H. d., Branlant, C. & Bockmayr, A. (2003) Multiscale modeling of alternative splicing regulation. In *Computational Methods in Systems Biology, CMSB'03*, (Priami, C., ed.), vol. 2602, pp. 75–87 Springer LNCS, Rovereto, Italy,.

[Feng & Doolitle, 1987] Feng, D. & Doolitle, R. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution,* **25**, 351–360.

[Fu & Maniatis, 1992] Fu, X. D. & Maniatis, T. (1992) Isolation of a complementary DNA that encodes the mammalian splicing factor SC35. *Science,* **256** (5056), 535–8.

[Ge & Manley, 1990] Ge, H. & Manley, J. L. (1990) A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA *in vitro. Cell,* **62** (1), 25–34.

[Ge *et al.*, 1991] Ge, H., Zuo, P. & Manley, J. L. (1991) Primary structure of the human splicing factor ASF reveals similarities with *Drosophila* regulators. *Cell,* **66** (2), 373–82.

[Golbeter, 1995] Golbeter, A. (1995) A model for circadian oscillations in the *drosophila* period protein (per). *Proc. R. Soc. Lond.,* **261**, 319–324.

[Gold *et al.*, 1997] Gold, L., Brown, D., He, Y.-Y., Shtaland, T., Singer, B. & Wu, Y. (1997) From oligonucleotide shapes to genomic SELEX : novel biological regulatory loops. *Proc. Natl Acad. Sci. USA,* **94**, 59–64.

[Gouzé, 1998] Gouzé, J.-L. (1998) Positive and negative circuits in dynamical systems. *J. Biol. Syst.,* **6** (1), 11–15.

[Graveley, 2000] Graveley, B. R. (2000) Sorting out the complexity of SR protein functions. *RNA,* **6** (9), 1197–211.

[Graveley *et al.*, 1998] Graveley, B. R., Hertel, K. J. & Maniatis, T. (1998) A systematic analysis of the factors that determine the strength of pre-mrna splicing enhancers. *EMBO,* **17** (22), 6747–6756.

[Guermeur, 2002] Guermeur, Y. (2002) Combining discriminant models with new multi-class svms. *Pattern Analysis and Applications,* **5** (2), 168–179.

[Guermeur *et al.*, 2002] Guermeur, Y., Elisseeff, A. & Zelus, D. (2002). Bounding the capacity measure of multi-class discriminant models. technical report NC-TR-2002-123 NeuroCOLT2.

[Gupta *et al.*, 1998] Gupta, V., Jagadeesan, R. & Saraswat, V. (1998) Computing with continuous change. *Science of Computer programming,* **30** (1–2), 3–49.

[Gupta *et al.*, 1995] Gupta, V., Jagadeesan, R., Saraswat, V. & Bobrow, D. G. (1995) Programming in hybrid constraint languages. In *Hybrid Systems II* pp. 226–251 Springer, LNCS 999.

[Hammond, 1993] Hammond, B. J. (1993) Quantitative study of the control of HIV-1 gene expression. *J. Theor. Biol.,* **163** (199-221).

[Hartigan & Wong, 1979] Hartigan, J. A. & Wong, M. A. (1979) A k-means clustering algorithm. *Applied Statistics,* **28**, 100–108.

[Haussler, 1999] Haussler, D. (1999). Convolution kernels on discrete structures. Technical Report UCS-CRL-99-10 UC Santa Cruz citeseer.nj.nec.com/haussler99convolution.html.

[Hawkins, 1988] Hawkins, J. D. (1988) A survey on intron and exon lengths. *Nucleic Acids Res,* **16** (21), 9893–908.

[Heinrich & Schuster, 1996] Heinrich, R. & Schuster, S. (1996) *The Regulation of Cellular Systems.* Chapman and Hall, New York.

[Higgins *et al.*, 1994] Higgins, D., Thompson, J., Gibson, T., Thompson, J. & Higgins, D.G.and Gibson, T. (1994) CLUSTAL W : improving the sensitivity of progressive-multiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* **22**, 4673–4680.

[Hope, 1999] Hope, T. (1999) The ins and outs of HIV Rev. *Arch Biochem Biophys,* **365** (2), 186–91.

[Hornig *et al.*, 1986] Hornig, H., Aebi, M. & Weissmann, C. (1986) Effect of mutations at the lariat branch acceptor site on beta-globin pre-mRNA splicing *in vitro. Nature,* **324** (6097), 589–591.

[Jaakkola *et al.*, 1999] Jaakkola, T., Diekhans, M. & Haussler, D. (1999) Using the fisher kernel method to detect remote protein homologies. In *ISMB* pp. 149–158, Heidelberg.

[Jacquenet, 2001] Jacquenet, S. (2001). *Etude de la regulation de l'épissage du transcrit primaire du virus de l'immunodéficience humaine type I.* PhD thesis, Henri Poincaré. 15 armoire du kiki (pour l'instant chez le stappi).

[Jacquenet *et al.*, 2001] Jacquenet, S., Ropers, D., Bilodeau, P. S., Damier, L., Mougin, A., Stoltzfus, C. M. & Branlant, C. (2001) Conserved stem-loop structures in the hiv-1 rna region containing the a3 3' splice site and its cis-regulatory element : possible involvement in rna splicing. *Nucleic Acids Res,* **29** (2), 464–78.

[Jong, 2004] Jong, H. (2004). Qualitative simulation and related approaches for the analysis of dynamical systems. Technical Report 5128 INRIA.

[Jong, 2002] Jong, H. d. (2002) Modeling and simulation of genetic regulatory systems : a litterature review. *Journal of Computational Biology,* **9** (1), 67–103.

[Jong *et al.*, 2001] Jong, H. d., Geiselmann, J., Hernandez, C. & Page, M. (2001) Genetic network analyzer : a tool for the qualitative simulation of genetic regulatory networks. *Bioinformatics,* **19** (3), 336–344.

[Jong & Page, 2000] Jong, H. d. & Page, M. (2000) Qualitative simulation of large and complex genetic regulatory systems. In *European Conference on Artificial Intelligence,* (Horn, W., ed.), vol. 14, IOS press, Amsterdam.

[Karp *et al.*, 1999] Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. & Krummenacker, M. (1999) Ecocyc : encyclopedia of *escherichia coli* genes and metabolism. *Nucleic Acids Research,* **27** (1), 55–58.

[Krainer *et al.*, 1990] Krainer, A. R., Conway, G. C. & Kozak, D. (1990) The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites. *Cell,* **62** (1), 35–42.

[Krainer *et al.*, 1991] Krainer, A. R., Mayeda, A., Kozak, D. & Binns, G. (1991) Functional expression of cloned human splicing factor SF2 : homology to RNA-binding proteins, U1 70K, and *Drosophila* splicing regulators. *Cell,* **66** (2), 383–94.

[Kucherov & Rusinowitch, 1997] Kucherov, G. & Rusinowitch, M. (1997) Matching a set of strings with variable length don't cares. *Theoretical Computer Science,* **178**, 129–154.

[Kuipers, 1988] Kuipers, B. (1988) Qualitative simulation using time-scale abstraction. *International Journal of Artificial Intelligence in Engineering,* **3** (4), 185–191.

[Kuipers, 1994] Kuipers, B. (1994) *Qualitative reasoning : Modeling and simulation with incomplete knowledge.* MIT Press, Cambridge, MA.

[Lawrence & Reilly, 1990] Lawrence, C. & Reilly, A. (1990) An expectation maximization (em) algorithm for the idnetification and characterization of common sites in unaligned biopolymer sequences. *Proteins : Structure, Function, and Genetics.*

[Legay, 1973] Legay, J.-M. (1973) La méthode des modèles, état actuel de la méthode expérimentale. In *Informatique et biosphère.* Paris pp. 1–76.

[Legendre & Legendre, 2000] Legendre, P. & Legendre, L. (2000) *Numerical ecology.* second english edition edition,, Elsevier.

[Lejeune *et al.*, 2001] Lejeune, F., Cavaloc, Y. & Stevenin, J. (2001) Alternative splicing of intron 3 of the serine /arginine-rich protein 9g8 gene. *Journal of Biological Chemistry,* **276** (March 16), 7850–7858.

[Leloup & Golbeter, 1998] Leloup, J.-C. & Golbeter, A. (1998) A model for the circadian rythms in *drosophila* incorporating the formation of a complex between the per and tim proteins. *J. Biol. Rhythms,* **13** (1), 70–87.

[Liu *et al.*, 1998] Liu, H.-X., Zhang, M. & Krainer, A. R. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual sr proteins. *Genes & Development,* **12**, 1998–2012.

[Mackay & Crossley, 1998] Mackay, J. P. & Crossley, M. (1998) Zinc fingers are sticking together. *Trends Biochem Sci,* **23** (1), 1–4.

[Maniatis & Reed, 2002] Maniatis, T. & Reed, R. (2002) An extensive network of coupling among gene expression machines. *Nature,* **416** (6880), 499–506.

[Manley & Tacke, 1996] Manley, J. L. & Tacke, R. (1996) SR proteins and splicing control. *Genes Dev,* **10** (13), 1569–79.

[Marchand *et al.*, 2002] Marchand, V., Mereau, A., Jacquenet, S., Thomas, D., Mougin, A., Gattoni, R., Stevenin, J. & Branlant, C. (2002) A janus splicing regulatory element modulates HIV-1 tat and rev mRNA production by coordination of hnRNP A1 cooperative binding. *J Mol Biol,* **323** (4), 629–52.

[McAdams & Shapiro, 1995] McAdams, H. & Shapiro, L. (1995) Circuit simulation of genetic network. *Science,* **269**, 650–656.

[McPherson *et al.*, 2001] McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., Fulton, R., Kucaba, T. A., Wagner-McPherson, C., Barbazuk, W. B., Gregory, S. G., Humphray, S. J., French, L., Evans, R. S., Bethel, G., Whittaker, A., Holden, J. L., McCann, O. T., Dunham, A., Soderlund, C., Scott, C. E., Bentley, D. R., Schuler, G., Chen, H. C., Jang, W., Green, E. D., Idol, J. R., Maduro, V. V., Montgomery, K. T., Lee, E., Miller, A., Emerling, S., Kucherlapati, Gibbs, R., Scherer, S., Gorrell, J. H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P. J., Catanese, J. J., Nowak, N., Osoegawa, K., Qin, S., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V. G., Kirsch, I. R., Reid, T., Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J. F., Hawkins, T., Myers, R. M., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N. E., Cox, D. R., Haussler, D., Kent, W. J., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X. N., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H. S., Sakaki, Y., Shimizu, N., Asakawa, S. *et al.* (2001) A physical map of the human genome. *Nature,* **409** (6822), 934–41.

[Milner, 1993] Milner, R. (1993) The polyadic pi-calculus : a tutorial. In *Logic and Algebra of Specification, Proceedings of International NATO Summer School* pp. 428–440.

[Modrek & Lee, 2002] Modrek, B. & Lee, C. (2002) A genomic view of alternative splicing. *nature genetics,* **30**, 13–19.

[Monod, 1950] Monod, J. (1950) La technique des cultures continues. Théorie et applications. *Ann. Inst. Pasteur,* **79**, 390–410.

[Monod & Jacob, 1961] Monod, J. & Jacob, F. (1961) General conclusions : teleomic mechanisms in cellular metabolism, growth, and differentiation. In *Cold Spring Harbor Symp. Quant. Biol.* vol. 26, pp. 389–401 Cold Spring Harbor NY.

[Moore *et al.*, 1993] Moore, M., Query, C. & Sharp, P. (1993) Splicing of precursors to messenger rnas by the spliceosome. In *The RNA World.* Cold Spring Harbor Laboratory Press.

[Murray, 2002] Murray, J. (2002) *Mathematical biology I : an introduction.* third edition,, Springer-Verlag.

[Murray, 2003] Murray, J. (2003) *Mathematical biology II : spatial models and biomedical applications.* third edition,, Springer-Verlag.

[Needleman & Wunsch, 1970] Needleman, S. & Wunsch, C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two prteins. *Journal of Molecular Biology,* **48**, 443–453.

[O'Reilly *et al.*, 1995] O'Reilly, M., McNally, M. & Beemon, K. (1995) Two strong 5' splice sites and competing, suboptimal 3' splice sites involved in alternative splicing of human immunodeficiency virus type 1 rna. *Virology,* **213** (2), 373–85.

[Palsson, 2000] Palsson, B. (2000) The challenges of in silico biology. *Nature Biotechnology,* **18**, 1147–1150.

[Palsson, 2002] Palsson, B. (2002) In silico biology through "omics". *nature biotechnology,* **20**, 649–650.

[Pavlakis & Felber, 1990] Pavlakis, G. N. & Felber, B. K. (1990) Regulation of expression of human immunodeficiency virus. *New Biol,* **2** (1), 20–31.

[Pavé, 1994] Pavé, A. (1994) *Modélisation en biologie et en écologie.* Aléas.

[Purcell & Martin, 1993] Purcell, D. & Martin, M. (1993) Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication and infectivity. *Journal of Virology,* **67** (11), 6365–78.

[Query *et al.*, 1995] Query, C. C., Strobel, S. A. & Sharp, P. A. (1995) The branch site adenosine is recognized differently for the two steps of pre-mRNA splicing. *Nucleic Acids Symp Ser,* **33**, 224–5.

[Query *et al.*, 1996] Query, C. C., Strobel, S. A. & Sharp, P. A. (1996) Three recognition events at the branch-site adenine. *EMBO J,* **15** (6), 1392–402.

[Regev *et al.*, 2001] Regev, A., Silverman, W. & Shapiro, E. (2001) Representation and simulation of biochemical processes using the pi-calculus process algebra. In *Pacific Symposium of Biocomputing* vol. 6, pp. 459–470.

[Roberts & Smith, 2002] Roberts, G. C. & Smith, C. W. (2002) Alternative splicing : combinatorial output from the genome. *Curr Opin Chem Biol,* **6** (3), 375–83.

[Ropers, 2003] Ropers, D. (2003). *Etude expérimentale du rôle des protéines SR dans la régulation de l'épissage de l'ARN du virus HIV-1, responsable de l'immunodéficience humaine, et modélisation mathématique de ces régulations.* PhD thesis, Université Nancy 1.

[Roscigno & Garcia-Blanco, 1995] Roscigno, R. F. & Garcia-Blanco, M. A. (1995) SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome. *RNA,* **1** (7), 692–706.

[Roulet *et al.*, 2002] Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J., Mermod, N. & Bucher, P. (2002) Higth-throughput selex-sage method for quantitative modeling of transcription-factor binding sites. *nature biotechnology,* **20**, 831–835.

[Rowen *et al.*, 2002] Rowen, L., Young, J., Birditt, B., Kaur, A., Madan, A., Philipps, D. L., Qin, S., Minx, P., Wilson, R. K., Hood, L. & Graveley, B. R. (2002) Analysis of the human neurexin genes : alternative splicing and the generation of protein diversity. *Genomics,* **79** (4), 587–97.

[Saigo *et al.*, 2004] Saigo, H., Vert, J.-P., Akutsu, T. & Ueda, N. Protein homology detection using string alignment kernels. appear in Bioinformatics.

[Saraswat, 1993] Saraswat, V. A. (1993) *Concurrent constraint programming.* MIT Press.

[Saraswat *et al.*, 1994] Saraswat, V. A., Jagadeesan, R. & Gupta, V. (1994) Foundations of timed concurrent constraint programming. In *9th Symp. Logic in Computer Science, LICS'94, Paris* pp. 71 – 80 IEEE.

[Saraswat *et al.*, 1996] Saraswat, V. A., Jagadeesan, R. & Gupta, V. (1996) Timed default concurrent constraint programming. *Journal of Symbolic Computation,* **22** (5/6), 475–520.

[Schaal & Maniatis, 1999*a*] Schaal, T. D. & Maniatis, T. (1999*a*) Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol Cell Biol,* **19** (1), 261–73.

[Schaal & Maniatis, 1999*b*] Schaal, T. D. & Maniatis, T. (1999*b*) Selection and characterization of pre-mRNA splicing enhancers : identification of novel SR protein-specific enhancer sequences. *Mol Cell Biol,* **19** (3), 1705–19.

[Screaton *et al.*, 1995] Screaton, G. R., Cáceres, J. F., Mayeda, A., Bell, M. V., Plebanski, M., Jackson, D. G., Bell, J. I. & Krainer, A. R. (1995) Identification and characterization of three members of the human SR family of pre-mRNA splicing factors. *EMBO J,* **14** (17), 4336–49.

[Smith & Valcarcel, 2000] Smith, C. W. & Valcarcel, J. (2000) Alternative pre-mrna splicing : the logic of combinatorial control. *Trends In Biochemical Sciences,* **25** (8), 381–388.

[Smith & Valcárcel, 2000] Smith, C. W. & Valcárcel, J. (2000) Alternative pre-mRNA splicing : the logic of combinatorial control. *Trends Biochem Sci,* **25** (8), 381–8.

[Smith & Waterman, 1981] Smith, T. & Waterman, M. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology,* **47** (1), 195–197.

[Smolen *et al.*, 2000] Smolen, P., Baxter, D. & Burne, J. (2000) Modeling transcriptional control in gene networks : method, recent results, and future directions. *Bull. Math. Biol.,* **62**, 247–292.

[Snoussi, 1989] Snoussi, E. (1989) Qualitative dynamics of piecewise-linear differential equations : a discrete mapping approach. *Dynam. Stabil. Syst.,* **4** (3–4), 189–207.

[Soret *et al.*, 1998] Soret, J., Gattoni, R., Guyon, C., Sureau, A., Popielarz, M., Le Rouzic, E., Dumon, S., Apiou, F., Dutrillaux, B., Voss, H., Ansorge, W., Stévenin, J. & Perbal, B. (1998) Characterization of SRp46, a novel human SRsplicing factor encoded by a PR264/SC35 retropseudogene. *Mol Cell Biol,* **18** (8), 4924–34.

[Sterner *et al.*, 1996] Sterner, D. A., Carlo, T. & Berget, S. M. (1996) Architectural limits on split genes. *Proc Natl Acad Sci U S A,* **93** (26), 15081–5.

[Sun *et al.*, 1996] Sun, F., Galas, D. & Waterman, M. S. (1996) A mathematical analysis of in vitro molecular selection-amplification. *Journal of Molecular Biology,* **258**, 650–660.

[Sun *et al.*, 2003] Sun, Y. F., Fan, X. D. & Li, Y. D. (2003) Identifying splicing sites in eukaryotic RNA : support vector machine approach. *Comput. Biol. Med.,* **33** (1), 17–29.

[Sànchez & Thieffry, 2001] Sànchez, L. & Thieffry, D. (2001) A logical analysis of *drosophila* gap genes. *J. Theor. Biol.,* **211**, 115–141.

[Tacke *et al.*, 1997] Tacke, R., Chen, Y. & Manley, J. L. (1997) Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation : creation of an SRp40-specific splicing enhancer. *Proc Natl Acad Sci U S A,* **94** (4), 1148–53.

[Tacke & Manley, 1995*a*] Tacke, R. & Manley, J. L. (1995*a*) The human splicing factors asf/sf2 and sc35 possess distinct, functionally significant rna binding specificities. *EMBO,* **14**, 3540–3551.

[Tacke & Manley, 1995*b*] Tacke, R. & Manley, J. L. (1995*b*) The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J,* **14** (14), 3540–51.

[Tang *et al.*, 1999] Tang, H., Kuhen, K. L. & Wong-Staal, F. (1999) Lentivirus replication and regulation. *Annu Rev Genet,* **33**, 133–70.

[Thieffry & Jong, 2002] Thieffry, D. & Jong, H. d. (2002) Modélisation, analyse et simulation des réseaux génétiques. *Médecine/Sciences,* **18**, 492–502.

[Thieffry & Thomas, 1998] Thieffry, D. & Thomas, R. (1998) Qualitative analysis of gene networks. In *Pacific Symposium on Biocomputing* vol. 3, pp. 77–88.

[Thomas *et al.*, 1995] Thomas, R., Thieffry, D. & Kaufman, M. (1995) Dynamical behaviour of biological regulatory networks - i. biological role of feedback loops and pratical use of the concept of the loop-characteristic state. *Buletin of Mathematical Biology,* **57** (2), 247–276.

[Tuerk & Gold, 1990] Tuerk, C. & Gold, L. (1990) Systematic evolution of ligands by exponential enrichment : rna ligands to bacteriophage t4 dna polymerase. *Science,* **249**, 505–510.

[Tyson, 1999] Tyson, J. (1999) Models of cell cycle control in eukaryotes. *J. Biotechnol.,* **71** (1–3), 239–244.

[Tyson *et al.*, 1996] Tyson, J., Novak, B., Odell, G., Chen, K. & Thron, C. (1996) Chemical kinetic theory : understanding cell-cycle regulation. *Trends Biochem. Sci.,* **21** (3), 89–96.

[Vant-Hull *et al.*, 1998] Vant-Hull, B., Payano-Baez, A., Davis, R. H. & Gold, L. (1998) The mathematics of selex against complex targets. *Journal of Molecular Biology,* **278**, 579–597.

[Vapnik, 1982] Vapnik, V. (1982) *Estimation od Dependences Based on Empirical Data.* Springer-Verlag.

[Vapnik, 1998] Vapnik, V. (1998) *Statistical learning theory.* John Wiley & Sons, Inc, N.Y.

[Venter *et al.*, 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C. *et al.* (2001) The sequence of the human genome. *Science,* **291** (5507), 1304–51.

[Voit, 2000] Voit, E. (2000) *Computational Analysis of Biochemical Systems.* Cambridge University Press.

[Wain-Hobson *et al.*, 1985] Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S. & Alizon, M. (1985) Nucleotide sequence of the aids virus, lav. *Cell,* **40** (1), 9–17.

[Weston & Watkins, 1998] Weston, J. & Watkins, C. (1998). Multi-class support vector machines. technical report CSD-TR-98-04 Royal Holloway, University of London, Department of Computer Science.

[Will & Lührmann, 2001] Will, C. L. & Lührmann, R. (2001) Spliceosomal UsnRNP biogenesis, structure and function. *Curr Opin Cell Biol,* **13** (3), 290–301.

[Wong *et al.*, 1997] Wong, P., Gladney, S. & Keasling, J. (1997) Mathematical model of the *lac* operon : inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnol. Prog,* **13**, 132–143.

[Xiao & Manley, 1997] Xiao, S. H. & Manley, J. L. (1997) Phosphorylation of the ASF/SF2 RS domain affects both protein-protein and protein-RNA interactions and is necessary for splicing. *Genes Dev,* **11** (3), 334–44.

[Xiao & Manley, 1998] Xiao, S. H. & Manley, J. L. (1998) Phosphorylation-dephosphorylation differentially affects activities of splicing factor ASF/SF2. *EMBO J,* **17** (21), 6359–67.

[Zahler *et al.*, 1992] Zahler, A. M., Lane, W. S., Stolk, J. A. & Roth, M. B. (1992) SR proteins : a conserved family of pre-mRNA splicing factors. *Genes Dev,* **6** (5), 837–47.

[Zahler *et al.*, 1993*a*] Zahler, A. M., Neugebauer, K. M., Lane, W. S. & Roth, M. B. (1993*a*) Distinct functions ofSR proteins in alternative pre-mRNA splicing. *Science,* **260** (5105), 219–22.

[Zahler *et al.*, 1993*b*] Zahler, A. M., Neugebauer, K. M., Stolk, J. A. & Roth, M. B. (1993*b*) Human SR proteins and isolation of a cDNA encoding SRp75. *Mol Cell Biol,* **13** (7), 4023–8.

[Zhang & Wu, 1996] Zhang, W. J. & Wu, J. Y. (1996) Functional properties of p54, a novel SR protein active in constitutive and alternative splicing. *Mol Cell Biol,* **16** (10), 5400–8.

[Zhang *et al.*, 2003] Zhang, X. H.-F., Heller, K. A., Hefter, I. & Leslie, C. S. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Research,* **13**, 2637–2650.

[Zuker, 2003] Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res,* **31** (13), 3406–3415.

Résumé

L'épissage des transcrits de l'ARN polymérase II chez les eucaryotes permet la production des ARN messagers codant les protéines. Les possibilités d'épissage alternatifs d'un grand nombre de transcrits permet de générer plusieurs ARN messagers distincts à partir d'un même gène. Cette régulation post-transcriptionnelle, associée à une forte capacité combinatoire, est par ailleurs très utilisées par les virus tels que HIV-1. La compréhension des mécanismes qui contrôlent l'épisage est une problématique centrale dans les recherches biologiques actuelles. De récentes études expérimentales ont mis en évidence le rôle des protéines SR. Ces protéines régulent l'épissage en se fixant à des sites privilégiés des ARN immatures. Néanmoins, l'identification expérimentale de ces sites de fixation reste difficile. Afin de faciliter les démarches expérimentales et mieux comprendre les mécanismes de cette régulation, notre approche, développée dans ces travaux de thèse, vise à mettre en oeuvre un protocole *in silico* à partir de modèles statistiques et formels. Cette approche repose sur deux techniques de modélisation. Une modélisation statistique, nous permet d'extraire les hypothèses biologiques que nous représentons ensuite dans une modélisation formelle pour en vérifier la cohérence.

L'approche SELEX consiste à identifier expérimentalement dans une collection de séquences aléatoires celles ayant une affinité pour la protéine étudiée. Nous disposons de données obtenues pour les protéines SR. Nous avons montré que l'application des modèles statistiques de classification est particulièrement adaptée à l'exploitation des données SELEX permettant la caractérisation dans notre cas d motifs reconnus par les protéines SR. Nous avons alors développé le logiciel, KOALAB, qui permet de localiser la présence des sites potentiels dans les ARN par discrimination. Les technologies employées reposent sur des modèles statistiques comme les machines d'apprentissage statistique telles que les SVM, et sur la recherche algorithmique discrète de mots. Enfin, dans le but de caractériser fonctionnellement ces motifs, nous avons employé une modélisation formelle. Pour ce faire, nous nous sommes tout d'abord focalisé sur la régulation d'un site d'épissage seul. Ce modèle est formalisé dans un langage de programmation par contraintes concurrentes hybrides (`Hybrid cc`) et validé qualitativement par rapport aux données expérimentales. Il est alors possible d'envisager une modélisation qui intègre les régulations d'un site d'épissage dans un contexte biologique plus large comme la régulation simultanée de plusieurs sites. D'un point de vue théorique, ce nouveau modèle toujours formalisé en `Hybrid cc` est une base théorique intéressante pour envisager l'étude d'un système concernant plusieurs échelles biologiques. D'un point de vue biologique, la modélisation formelle permet de tester *in silico* les influences de la régulation locale de l'épissage comme le rôle des protéines SR sur un comportement plus globale tel que le cycle de vie du virus HIV-1.

Université HENRI POINCARÉ NANCY 1

FACULTE DES SCIENCES & TECHNIQUES

RAPPORT DE SOUTENANCE

Concernant la thèse de **Doctorat de l'Université Henri Poincaré, Nancy 1**

en Biologie Moléculaire

Présentée par : EVEILLARD Damien

Date de la soutenance : 14 Mai 2004

Mr Damien EVEILLARD a présenté au jury ses travaux sur la modélisation de la régulation de l'épissage alternatif.

L'exposé a été jugé clair et pédagogique, ce qui a été d'autant plus apprécié qu'il s'agissait d'un exercice de synthèse particulièrement ardu : il fallait couvrir de nombreux domaines, à la fois pour situer les problèmes biologiques et informatiques et pour décrire les résultats des multiples techniques abordées.
Le candidat a fait la preuve qu'il possédait une bonne maîtrise de ces problèmes encore largement ouverts, discutant de questions difficiles à la fois en biologie (épissage du virus HIV1) et en informatique (méthode d'apprentissage SVM, modélisation par systèmes de contraintes).

Une longue séance de questions a suivi, dont un certain nombre était de vraies questions scientifiques ouvertes, qui reflétaient l'intérêt du jury pour les multiples prolongements possibles de ce travail. Au cours de ses réponses, qui ont été jugées tout à fait satisfaisantes, Damien EVEILLARD a témoigné d'une réelle connaissance pluridisciplinaire et d'un constant souci d'honnêteté dans son travail.

Le jury encourage le candidat à faire fructifier cet investissement via un séjour de recherche post-doctoral, en cherchant à prolonger l'étude sur la modélisation par systèmes de contraintes, qui a été jugée prometteuse.

résident du Jury - Nom, Prénom et signature : NICOLAS Jacques

lembres du Jury - Nom, Prénom et signature :

THERMES, Claude

BRANLANT Christiane

BOCKMAYR Alexander A. Bockmayr

VIARI Alain

B. De par décision du Conseil d'administration de l'Université en date du 2 décembre 2002, les jurys de soutenance de thèse de l'UHP Nancy 1 n'attribuent plus aucune mention à compter du 1er janvier 2003.

78A **A retourner au Service de la Scolarité de la Faculté des Sciences & Techniques dans les meilleurs délais.**

Monsieur **EVEILLARD Damien**

DOCTORAT de l'UNIVERSITE HENRI POINCARE, NANCY 1

en BIOLOGIE MOLECULAIRE

VU, APPROUVÉ ET PERMIS D'IMPRIMER N° 934

Nancy, le 1er juin 2004

Le Président de l'Université

UNIVERSITÉ HENRI POINCARÉ, NANCY 1 * PRESIDENT

J.P. FINANCE

Université Henri Poincaré, Nancy I
24-30 rue Lionnois - B.P. 3069 - 54013 NANCY Cédex
Tél. : 03 83 68 20 00 - Fax : 03 83 68 21 00

Résumé

L'épissage des transcrits de l'ARN polymérase II chez les eucaryotes permet la production des ARN messagers codant les protéines. Les possibilités d'épissage alternatifs d'un grand nombre de transcrits permet de générer plusieurs ARN messagers distincts à partir d'un même gène. Cette régulation post-transcriptionnelle, associée à une forte capacité combinatoire, est par ailleurs très utilisées par les virus tels que HIV-1. La compréhension des mécanismes qui contrôlent l'épisage est une problématique centrale dans les recherches biologiques actuelles. De récentes études expérimentales ont mis en évidence le rôle des protéines SR. Ces protéines régulent l'épissage en se fixant à des sites privilégiés des ARN immatures. Néanmoins, l'identification expérimentale de ces sites de fixation reste difficile. Afin de faciliter les démarches expérimentales et mieux comprendre les mécanismes de cette régulation, notre approche, développée dans ces travaux de thèse, vise à mettre en oeuvre un protocole *in silico* à partir de modèles statistiques et formels. Cette approche repose sur deux techniques de modélisation. Une modélisation statistique, nous permet d'extraire les hypothèses biologiques que nous représentons ensuite dans une modélisation formelle pour en vérifier la cohérence.

L'approche SELEX consiste à identifier expérimentalement dans une collection de séquences aléatoires celles ayant une affinité pour la protéine étudiée. Nous disposons de données obtenues pour les protéines SR. Nous avons montré que l'application des modèles statistiques de classification est particulièrement adaptée à l'exploitation des données SELEX permettant la caractérisation dans notre cas d motifs reconnus par les protéines SR. Nous avons alors développé le logiciel, KOALAB, qui permet de localiser la présence des sites potentiels dans les ARN par discrimination. Les technologies employées reposent sur des modèles statistiques comme les machines d'apprentissage statistique telles que les SVM, et sur la recherche algorithmique discrète de mots. Enfin, dans le but de caractériser fonctionnellement ces motifs, nous avons employé une modélisation formelle. Pour ce faire, nous nous sommes tout d'abord focalisé sur la régulation d'un site d'épissage seul. Ce modèle est formalisé dans un langage de programmation par contraintes concurrentes hybrides (`Hybrid cc`) et validé qualitativement par rapport aux données expérimentales. Il est alors possible d'envisager une modélisation qui intègre les régulations d'un site d'épissage dans un contexte biologique plus large comme la régulation simultanée de plusieurs sites. D'un point de vue théorique, ce nouveau modèle toujours formalisé en `Hybrid cc` est une base théorique intéressante pour envisager l'étude d'un système concernant plusieurs échelles biologiques. D'un point de vue biologique, la modélisation formelle permet de tester *in silico* les influences de la régulation locale de l'épissage comme le rôle des protéines SR sur un comportement plus globale tel que le cycle de vie du virus HIV-1.