Projet présenté par l'UMR 7567 CNRS-UHP (MAEM) et l'UMR 7503 CNRS-INPL-INRIA-UHP (LORIA)

Développement et utilisation d'approches informatiques et théoriques pour l'analyse des liens existant entre défauts d'épissage et maladies génétiques

Plan du rapport

Le présent rapport fait le bilan des activités menées pendant la période de janvier à décembre 2005 sur le projet intitulé: "Développement et utilisation d'approches informatiques et théoriques pour l'analyse des liens existant entre défauts d'épissage et maladies génétiques" dans le cadre du programme décrypthon. Le document est divisé en trois parties: (1) la description générale du projet (responsables, participants, objectifs généraux), (2) le rapport scientifique (les travaux effectués, en cours et restant à réaliser seront précisés), (3) le rapport financier.

1. Description générale du projet

1.1 Introduction

L'état actuel des connaissances les plus récentes sur l'expression du génome humain montre que beaucoup de gènes humains conduisent en moyenne à la production de 7 à 8 formes différentes d'ARNm (Carninci et al., 2005). Cette grande diversité d'expression des gènes est liée en grande partie à l'épissage alternatif des transcrits (Clark and Thanaraj, 2002). L'épissage est en effet une étape clef de la maturation des transcrits, produits à partir de l'ADN, par l'ARN polymérase II. Cette étape de l'expression des gènes consiste en l'élimination de séquences internes des ARN pré-messagers, les introns, avec ligation des séquences exoniques. Elle permet la production des ARN messagers matures qui définissent la séquence en acides aminés des protéines synthétisées par les ribosomes. La possibilité d'utiliser des sites d'épissage de manière alternative sur un même ARN prémessager permet la production de différentes protéines à partir d'un gène donné, ceci en fonction de l'état de différenciation cellulaire ou du stade du développement par exemple. La machinerie d'épissage (complexe ribonucléoprotéique) fonctionne sur la base de la reconnaissance de signaux présents dans les ARN pré-messagers. Les régulations de l'épissage et, en particulier, de l'épissage alternatif font appel à de nombreux facteurs protéiques qui peuvent influencer l'identité du produit final de l'épissage.

Au sein des gènes, les séquences correspondant à des sites d'épissage, ou à des éléments régulant l'épissage alternatif peuvent être altérés par des mutations. De nombreuses maladies génétiques sont liées à des défauts d'épissage dû à des mutations inactivant des séquences régulant l'efficacité d'utilisation des sites d'épissage (Corcos and Solier, 2005). A titre d'exemple, le nombre de mutations ponctuelles identifiées en 1999 touchant des sites d'épissage dans le gène de la dystrophine, était estimé à 34% du nombre total des mutations répertoriées (70 des 203 mutations) (Tuffery-Giraud et al., 1999). En plus de défauts d'épissage liés à la présence de mutations ponctuelles, différentes maladies génétiques sont dues à des défauts d'épissage liés à la présence d'amplification de séquences trinucléotidiques (entre 50 et 11000 répétitions) (Cummings and Zoghbi, 2000). C'est en particulier le cas des dystrophies myotoniques de type 1 et 2, qui sont respectivement liées à la séquestration de protéines jouant un rôle clé dans des régulations de l'épissage alternatif par les séquences répétées CUG et CCUG (Ranum and Day, 2004). Les protéines produites par plusieurs gènes sont alors modifiés, ce qui conduit à la maladie.

1.2 Objectifs généraux

L'objectif du projet est d'analyser les relations existant entre défauts d'épissage et maladies génétiques, au niveau des déterminants de séquences (structure primaire) et de structures 2D et 3D des éléments des ARNm impliqués dans certains dysfonctionnements de l'épissage. Notre objectif est double. Le premier est de développer des méthodes

permettant d'identifier les sites d'épissage (par apprentissage statistique) ou de façon plus spécifique de distinguer les introns des exons. Le second est de développer des méthodes de modélisation moléculaire permettant de comprendre les bases moléculaires de la formation de complexes ARN/protéine spécifiques conduisant à des défauts d'épissage. Une meilleure prédiction par informatique des séquences introniques et exoniques dans les séquences génomiques humaines devrait permettre de prédire si des mutations identifiées dans les gènes sont susceptibles de conduire à un défaut d'épissage. La modélisation des complexes formés entre les séquences répétées CUG et CCUG et protéines de régulation de l'épissage devrait permettre de comprendre les bases moléculaires des dystrophies myotoniques. A long terme, ces développements et leurs applications à des maladies génétiques pourraient avoir des applications dans le domaine du diagnostic et du dépistage de maladies génétiques liées à des défauts d'épissage et dans la conception de drogues permettant d'atténuer les symptômes des dystrophies myotoniques.

1.3 Responsables et participants

Responsables:

- Christiane Branlant, Directeur de Recherche CNRS ; UMR 7567 : équipe de Maturation des ARN
- Fabrice Leclerc, Chargé de Recherche CNRS; UMR 7567
- Yann Guermeur, Chargé de Recherche CNRS; LORIA UMR 7503: équipe MODBIO

Laboratoires partenaires :

- LORIA, équipe MODBIO
- UHP UMR CNRS 7567 « Maturation des ARN et Enzymologie Moléculaire »

Participants:

- Christiane Branlant, Directeur de Recherche CNRS ; UMR 7567 (15%)
- Fabrice Leclerc, Chargé de Recherche CNRS; UMR 7567 (30%)
- Yann Guermeur, Chargé de Recherche CNRS; UMR 7503 (15%)
- Petar Mitrasinovic, chercheur modélisateur (7 mois* / 24 mois à partir du 10/01/2005, 100%, AFM)
- Nathalie Marmier-Gourrier, ingénieur biologiste (24 mois à partir du 01/04/2005, 100%, AFM)
- Delphine Autard, ingénieur bioinformaticien, AFM (6 mois à partir du 1/10/2005, 100%, AFM)
- Emmanuel Monfrini, chercheur postdoctoral statisticien, AFM (18 mois à partir du 1/10/2005, 100%, AFM)
- Raphaël Bolze, ingénieur de Recherche CNRS-STIC, projet décrypthon (24 mois à partir du 01/02/2005, 50%)
- Juan Alexander Padron García, 3ème cycle, Université de La Havane, Cuba (3 mois, 50%, échange dans le cadre de la convention entre l'UHP et l'Université de La Havane)

* durée effective du travail réalisé depuis janvier 2005 pour le contrat initial de 24 mois.

Nous remercions les personnes d'IBM-France impliqués dans le programme décrypthon avec lesquels nous avons collaboré (Fabien Lomet, Pierre-Jean Ponenti, Emmanuel d'Oncieu, Pascal Sempé, Jean-Christophe Mestres, Anne-Marie Armbruster, Yves Blanchet ainsi que le nouveau contact IBM: Karine Brua).

2. Rapport scientifique

2.1 Développement et application de méthodes SVM pour la recherche de sites d'épissage fonctionnels constitutifs et régulés

L'objectif de cette partie du projet est de développer des outils performants d'analyse et de prédiction des sites d'épissage et de leurs régulations. La première étape est la construction d'une banque de données de sites d'épissage et des séquences qui les environnent, ceci pour des sites ayant des caractéristiques bien définies. La seconde étape sera la mise en œuvre de la méthode d'apprentissage. Cette partie du projet a été débutée de façon conjointe au MAEM et au LORIA, ceci grâce à la participation des personnels recrutés, Mlle Delphine Autard, co-encadrée par F. Leclerc et Y. Guermeur et M. Emmanuel Monfrini encadré par Y. Guermeur et N Marmier Gourrier, ceci en s'appuyant sur les compétences de l'équipe de C Branlant en matière de régulation de l'épissage alternatif. Les travaux réalisés et ceux en cours portent sur : (1) l'évaluation et le choix des données d'épissage qui seront exploitées dans la suite du projet, (2) la création d'une banque de données locale obtenue par extraction et annotations des données sur des sites d'épissage et leurs séquences environnantes, (3) le développement d'un programme de préparation des données pour l'apprentissage, (4) la mise en œuvre de la méthode d'apprentissage (tests et portage) destinée à l'identification introns/exons. Ce projet a été débuté plus tardivement et la durée effective de travail sur cette partie du projet par les personnels recrutés a été de 3 mois.

2.1.1 Evaluation et choix des banques de données d'épissage

Il existe différentes banques de données d'épissage publiées à ce jour, qui sont *a priori* pertinentes pour la réalisation du projet, ceci en termes de quantité, de qualité et d'accessibilité des données. Pour garantir la performance et la pertinence des outils développés basés sur les techniques d'apprentissage, l'homogénéité des données expérimentales est importante et une seule et même source de données sera donc utilisée pour la suite du projet.

Après examen du contenu d'une 10 de banques de données d'épissage (EID, ExInt, HS3D, AEDB, AltSplice, AltExtron, AsMamDB, ProSplicer, SpliceDB, HASDB, voir Annexe 1), la banque AltSplice (http://www.ebi.ac.uk/asd/altsplice/index.html) a été retenue, car elle présente les meilleures garanties selon les critères de qualité, quantité et accessibilité des données. Cette banque est maintenue par l'EBI ("European Bioinformatics Institute") qui est une organisation académique à but non-lucratif rattaché à l'EMBL ("European Molecular Biology Laboratory"), ce qui garantit l'accessibilité des données et des mises à jour. De plus, cette banque de données est désormais intégrée dans un projet plus global de l'EBI-EMBL appelé ATD pour "Alternate Transcript Diversity Database" (http://www.ebi.ac.uk/atd) sur la diversité de transcrits isoformes, à l'échelle des génomes, en particulier, du génome humain. Ceci garantit une pérennité d'accès et d'utilisation des données. La banque AltSplice (Clark and Thanaraj, 2002) répertorie 16 293 gènes avec leur(s) profil(s) d'épissage : les exons et introns, ainsi que leur position et le nombre de transcrits sont identifiés pour chaque gène. Bien que les informations sur les exons soumis à l'épissage alternatif ne soient pas explicites, le répertoire de l'ensemble des introns et exons possibles sera obtenu par le développement d'outils spécifiques (voir sections 2.1.2 et 2.1.3).

2.1.2 Création d'une banque des données locale d'épissage

Un outil a été créé pour extraire les données de la banque AltSplice et générer une banque de données locale. Les données sont réparties en 3 types de fichiers : "exons", "exons alternatifs" et "introns". Les exons pour lesquels il existe peu de données (par exemple dans les cas d'exons dont les transcrits correspondant n'ont été détectés qu'une ou 2 fois expérimentalement) sont classés à part comme "exons non confirmés". Cette classe d'exons ne sera pas utilisée pour les tests d'apprentissage de façon indépendante.

Une banque de données sur les séquences régulatrices de l'épissage (éléments activateurs exoniques ou introniques : ESE-"Exon Splicing Enhancer" et ISE-"Intron Splicing Enhancer", éléments inhibiteurs exoniques ou introniques : ESI-"Exon Splicing Silencer" et ISI-"Intron Splicing Silencer") a également été construite. Toutefois, comme elle présente un nombre restreint de séquences (115 au total : 52 ESE, 19ISS, 29ISE, et 7ISS), elle ne sera pas utilisée pour le moment dans les tests d'apprentissage pour éviter un biais.

2.1.3 Création d'un outil d'annotation des données pour l'apprentissage

Un programme écrit en C++ a été développé afin d'exploiter les données de la banque locale, pour l'apprentissage par la méthode de machines à vecteurs support à catégories multiples (M-SVM). Ce programme permet de créer des fichiers d'annotation exploitables par la M-SVM : il annote de façon automatique les introns, les exons constitutifs, les exons alternatifs et de façon optionnelle les sites d'épissage et les zones de régulation de l'épissage.

Afin d'équilibrer l'apprentissage et de permettre à la M-SVM d'apprendre à reconnaître les zones non introniques (les gènes contenant une grande majorité d'introns), nous avons choisi de couper des morceaux de séquences autour des exons (-250pb et +250pb de part et d'autre de l'exon). Dans le cas où au moins deux exons sont trop proches pour définir des séquences distinctes, des séquences contenant plusieurs exons sont constituées.

Enfin, pour créer les fichiers d'apprentissage de la M-SVM, une fenêtre glissante de taille fixée à 251 nucléotides est utilisée. C'est le nucléotide situé au milieu de cette fenêtre qui sera annoté. Pour chaque séquence, nous annotons donc les 125 nucléotides introniques précédents l'exon, la totalité de l'exon et les 125 nucléotides suivants l'exon. L'égalité en taille entre les exons et les introns, nécessaire pour une représentation équivalente exons/introns, est alors à peu près atteinte.

2.1.4 Mise en œuvre de la méthode d'apprentissage

Nous disposons de deux bases très complètes d'apprentissage : l'une formée par les "exons confirmés" et l'autre par une base ajoutant à la première l'information contenue dans les "exons non confirmés". Malgré la fiabilité plus réduite des données de la 2ème base, le gain potentiel en apprentissage reste à déterminer et justifie l'utilisation des 2 bases pour effectuer des apprentissages de façon indépendante. Ces 2 études parallèles (sur les 2 bases) nécessitent de nouveaux développements informatiques et mathématiques permettant une amélioration des temps de calcul, ainsi qu'un accès à des moyens informatiques conséquents. Ce n'est qu'alors qu'il sera possible de définir le choix du noyau et d'optimiser les autres hyperparamètres de la M-SVM, et ce, indépendamment, pour chacune des deux bases.

Nous nous proposons d'opérer un premier découpage des gènes, en nous basant sur les prédictions élargies (probabilité de validation abaissée à 0,01) d'exons par le logiciel GenScan. En effet, ce premier découpage permet d'une part de diminuer la taille de la

séquence considérée et d'autre part de concentrer les travaux de la M-SVM autour d'une zone dans laquelle on suppose la présence d'un ou plusieurs exons.

Deux méthodes de post-traitement peuvent alors être envisagées :

- exploiter les résultats de la M-SVM afin de confirmer ou d'infirmer les "candidats exons" proposés par GenScan (logiciel basé sur les HMM qui est donc plus sensible aux sites 5' et 3' que la M-SVM).
- opérer un apprentissage (et une prédiction) en deux temps grâce à la M-SVM (stacked generalization (Wolpert, 1992)). Il s'agit de couper les bases d'apprentissage en deux, d'effectuer l'apprentissage de la M-SVM sur la première partie, de calculer les scores obtenus sur la seconde partie de la base, et de refaire un second apprentissage en se basant sur ces scores.

Comme nous l'avons signalé plus haut, la nécessité de dédier un modèle au problème posé, l'importance d'une bonne qualité d'apprentissage et la possibilité de l'effectuer sur deux bases très complètes demandent un fort investissement, sur le plan mathématique et informatique évidemment, afin d'optimiser à la fois les temps de calcul et la précision des résultats obtenus, mais aussi sur le plan matériel puisque les ressources de stockage et les capacités de calcul nécessaires seront très importantes.

Les développements mathématiques nécessaires sont les suivants :

- élaboration d'un noyau dédié à ce modèle (gaussien étendu, HMM, PairHMM (Shawe-Taylor J., 2004), ...).
- important travail au niveau de la programmation mathématique permettant une optimisation des temps de calcul (collaboration avec des spécialistes comme Jean-Luc Lamotte du LIP6 et Anatoli Iouditski de l'université J.Fourier).

Les développements nécessaires sur le plan informatique sont les suivants :

- Reprogrammation de la M-SVM basée sur les travaux en programmation mathématique.
- Eventuellement le choix et l'intégration d'un solveur linéaire plus rapide et plus performant.

Les moyens matériels nécessaires sont les suivants :

- capacité de stockage des bases d'apprentissage de l'ordre de 2 To (dépendant du choix d'une méthode en une ou deux passes de la M-SVM).
- Paralléliser au maximum les calculs

Notons, enfin, que la partie liée aux choix de la base, du noyau ainsi qu'au réglage des hyperparamètres nécessitera un accès direct aux sources du programme, point qu'il faudra évidemment prendre en compte au moment du portage des algorithmes sur la grille.

2.2 Modélisation de la structure 3D des ARN à séquences répétées CUG et CCUG et leur interaction avec des protéines de régulation de l'épissage

L'objectif de cette 2ème partie du projet est la modélisation de complexes ARN/protéines impliqués dans les dystrophies myotoniques (DM) et plus particulièrement la modélisation des interactions entre les séquences répétées de type CH(H)G, $(CUG)_n$ et $(CCUG)_n$, avec les protéines de régulation connues pour se fixer sur ces séquences (Kino et al., 2004). Le modèle actuellement proposé pour expliquer les Dystrophie Myotoniques 1 et 2 repose sur la séquestration, sur les séquences répétées CUG et CCUG, de protéines, qui sont des régulateurs de l'épissage alternatif , les protéines de la famille CELF, et de la famille "muscleblind" (MBNL). En accord avec cette hypothèse, des données récentes obtenues avec un modèle murin suggèrent que les manifestations cliniques multiples des DM sont

liées à une altération de la régulation de l'épissage de plusieurs gènes pour lesquels les protéines de la famille CELF, en particulier CUG-BP1 ou ETR3, et la protéine MBN-1 ont un rôle antagoniste (Kanadia et al., 2003, Ho et al., 2005a, Ladd et al., 2005). La surexpression de CUG-BP1 (Ho et al., 2005a) ou la diminution de l'expression de MBNL-1 (Kanadia et al., 2003, Ladd et al., 2005) induisent, en effet, des manifestations cliniques caractéristiques des DM. Le rôle toxique des répétitions de type CH(H)G serait donc dû à un déséquilibre entre activation et répression de l'épissage alternatif, provoqué par une modification des concentrations en protéines CELF et MBNL libres, ceci du fait qu'elles sont liées aux séquences répétitions.

Les objectifs initiaux prévoyaient la modélisation des complexes formées entre les séquences répétées de type CH(H)G et les protéines CUG-BP1 et PKR. En raison de la démonstration récente du rôle prépondérant des protéines MBNL (Kanadia et al., 2003, Jiang et al., 2004, Kino et al., 2004, Ho et al., 2005b, Ladd et al., 2005), nous avons intégré, dans les objectifs du projet, la modélisation des interactions avec ces protéines. Des données structurales obtenues très récemment sur un duplex ARN comportant 6 répétitions CUG (Mooers et al., 2005) nous ont également conduit à réviser la stratégie de modélisation des ARN à séquences répétées $(CUG)_{n,}$ ceci afin d'intégrer ces nouvelles données à nos modèles.

Cette partie du projet a été menée au MAEM par Juan Alexander Padron (Université de La Havane) pour la partie de modélisation 3D des protéines, par Fabrice Leclerc pour la partie modélisation des ARN, par Raphaël Bolze et Fabrice Leclerc pour le portage de l'application sur la grille et par Nathalie Marmier-Gourrier pour la partie expérimentale. Petar Mitrasinovic à réalisé des analyses structurales d'ARN contenant des répétitions mais sa participation au projet a été interrompue à la fin du mois de juillet 2005 du fait qu'il n'acceptait plus de tenir compte des objectifs fixés dans le cadre du programme Décrypthon. La durée effective de travail pour les personnels recrutés a donc été de 7 mois pour Petar Mitrasinovic et 8 mois pour Nathalie Marmier-Gourrier.

2.2.1 Modélisation des ARN à séquence répétée CUG

Les séquences répétées $(CUG)_n$ se replient pour former de longues structures tige-boucle qui piégent des facteurs régulant l'épissage (Jasinska et al., 2003, Sobczak et al., 2003). La protéine CUG-BP1 semble se lier aux régions en simple-brin (extrémités des séquences répétées ou boucle terminale) (Michalowski et al., 1999), alors que la protéine MBNL1 se lierait aux régions double-brin (tige) (Miller et al., 2000, Kino et al., 2004), néanmoins aucune étude expérimentale ne prouve cette hypothèse. C'est ce que N Marmier-Gourrier essaye de démontrer expérimentalement. Selon des données obtenues avec des sondes enzymatiques, les répétitions CUG peuvent se replier de 2 façons différentes (Sobczak et al., 2003). Lorsque le nombre de répétitions est pair (2n) et que la tige-boucle la plus longue possible est formée, la boucle terminale est une boucle à 4 nucléotides ("tetraloop"). Lorsque le nombre de répétitions est impair (2n+1) et que la tige-boucle la plus longue possible est formée, la boucle terminale est une boucle à 3 nucléotides ("triloop"). Des données RMN (Leppert et al., 2004) et de microscopie électronique (Michalowski et al., 1999) suggèrent que ces séquences forment un structure double-brin au niveau de la tige où 2 paires Watson-Crick G-C et C-G sont séparées de façon régulière par un mésappariemment U:U. Des régions en simple-brin peuvent apparaître aux extrémités 5' et 3' des répétitions, lorsque des formes alternatives de la tige sont formées et elles pourraient constituer des sites de fixation pour CUG-BP1 à la base de la tige-boucle. Les 2 difficultés majeures rencontrées pour générer un modèle 3D d'ARN à séquences répétées CUG

résidaient dans l'absence de données expérimentales sur la nature du repliement de la boucle terminale CUG et la nature des mésappariemments U:U.

L'ensemble des données disponibles ont été intégrées comme contraintes géométriques pour modéliser la structure des ARN $(CUG)_n$ ceci en utilisant une approche par satisfaction de contraintes, implémentée dans le programme Mc-Sym (Major et al., 1991, Gautheret and Cedergren, 1993, Gautheret et al., 1993). Dans un premier temps, un petit modèle d'ARN à 3 répétitions CUG de séquence r(GG$(CUG)_3$CC) comportant une courte région double-brin et une boucle terminale de type "triloop" a été construit (Fig. 1). En l'absence de contraintes expérimentales sur le repliement de la boucle terminale (aucune structure 3D de boucle terminale CUG n'est connue), une analyse systématique et exhaustive des structures 3D connues des boucles terminales "triloop" a été effectuée. L'objectif était d'annoter et de classifier ces boucles en fonction des repliements observés et d'en déduire des règles générales applicables à la prédiction du repliement et de la structure 3D de la boucle terminale CUG. Cette analyse a été réalisée sur un emsemble de 90 structures 3D expérimentales (obtenues par diffraction des rayons-X ou par RMN) répertoriées dans la banque SCOR (http://scor.lbl.gov/) et dont les coordonnées atomiques ont été obtenues de la banque PDB (http://www.rcsb.org/pdb/). Les résultats de cette analyse sont présentés de façon schématique en Annexe 2 : le repliement prédit pour la boucle terminale CUG est de type YYN[3] (Fig. 1, voir Annexe 2 pour la nomenclature).

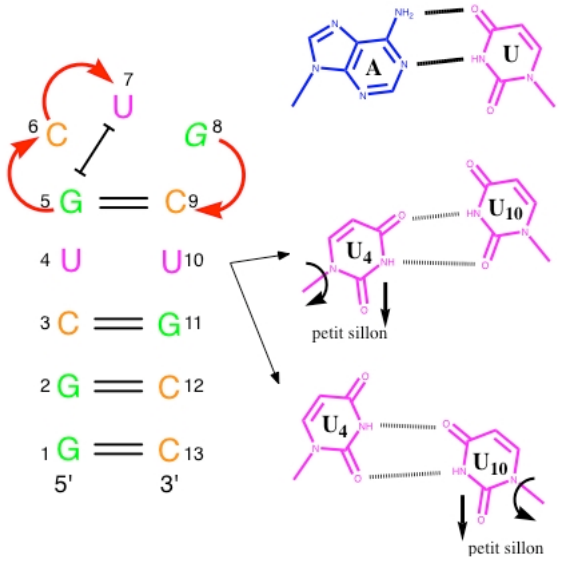


Figure 1: Annotation structurale de la séquence r(GG$(CUG)_3$CC) modélisée. La structure 2D est représentée en utilisant la nomenclature de Leontis & Westhof d'annotation des structures d'ARN. Les appariements Watson-Crick G :C sont indiqués par un double trait, l'empilement entre bases (en dehors des régions double brin) par un trait barré aux extrémités et les changements d'orientation de brin par un flèche rouge. L'appariement U_4 :U_{10} peut être inexistant ou bien correspondre à l'un des 2 appariements représentés de façon schématique sous l'appariement A :U Watson-Crick utilisé comme référence. Dans le premier appariement U_4 :U_{10}, U_4 est déplacé vers le petit sillon alors que dans le second, c'est U_{10} qui est déplacé vers le petit sillon par rapport à une paire Watson-Crick.

En l'absence de données expérimentales sur la nature des mésappariemments U:U, 3 cas de figure ont été envisagés (Fig. 1): (1) l'absence d'appariement fort, (2) la présence d'un appariement U:U où le U en 5' (U_4, Fig. 1) est déplacé vers le petit sillon, (3) la présence d'un appariement U :U où le U en 3' (U_{10} , Fig. 1) est déplacé vers le petit sillon. Les modèles 3D d'ARN obtenus à partir de ces contraintes présentent des structures avec de

possibles déformations locales du squelette phosphodiester (Fig. 2). Lorsqu'aucun appariement fort U_4:U_{10} existe (les 2 bases sont empilées de façon standard dans la double hélice), aucune déformation notable du squelette phosphodiester n'est présente (Fig. 2A). Lorsque l'appariement U_4:U_{10} est formé : une légère déformation du squelette phosphodiester est présente, à la position correspondant au nucléotide déplacé vers le petit sillon (U_4, Fig. 2B ou U_{10}, Fig. 2C).

Afin de déterminer la pertinence des 3 modèles obtenus (stabilité en solution, influence de contre-ions, etc) et plus particulièrement de la nature du mésappariemment U:U qui est considéré comme un facteur clé dans la reconnaissance ARN/protéine, des simulations par dynamique moléculaire ont été réalisées. Les résultats obtenus concordent avec les données structurales récentes sur un duplex ARN $(CUG)_6$: la géométrie du mésappariemment U:U observée dans les simulations est proche de celle dans la structure expérimentale du duplex (Mooers et al., 2005). Toutefois, les simulations suggèrent un assez grande flexibilité du mésappariemment U:U qui peut passer d'un appariemment où le U en 3' est déplacé vers le petit sillon à celui où le U en 5' est déplacé vers le petit sillon (dans la structure expérimentale du duplex, seul l'appariemment où le U en 3' est déplacé vers le petit sillon est observée).

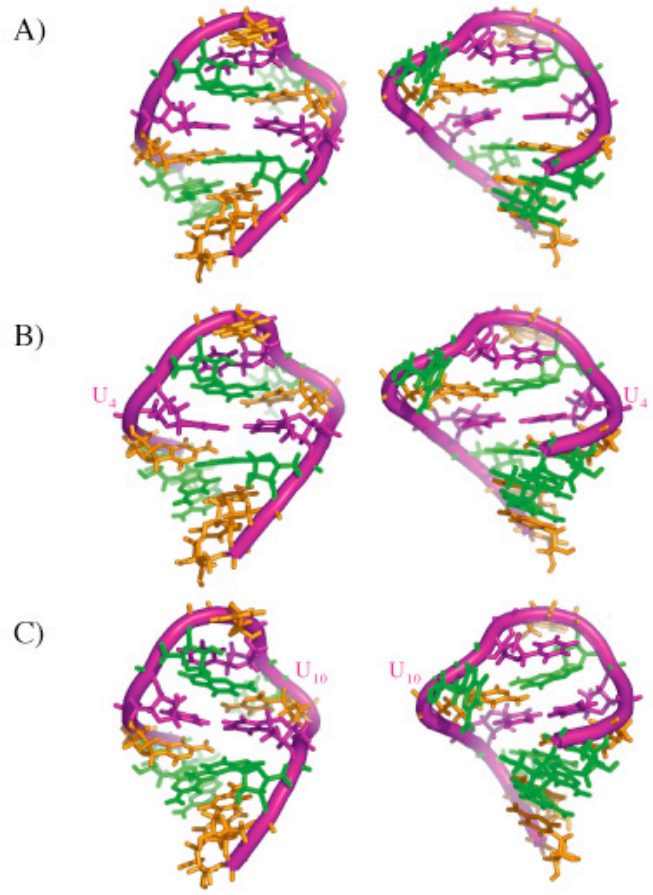


Figure 2: **Modèles 3D de la tige-boucle r(GG(CUG)$_3$CC)**. A) Modèle sans appariement fort U_4 :U_{10} (gauche: vue du petit sillon, droite: vue du grand sillon) . B) Modèle avec avec un appariement U_4 :U_{10} avec U_4 déplacé vers le petit sillon. C) Modèle avec avec un appariement U_4 :U_{10} avec U_{10} déplacé vers le petit sillon. Le squelette phosphodiester est représenté par un tube rose, les nucléotide sont colorés selon le code utilisé dans la figure 1.

Afin d'identifier les conformations représentatives prises en solution par des ARN comportant 16 (boucle terminale "tetraloop") ou 17 répétitions (boucle terminale "triloop")

CUG étendus ont été construits à partir des modèles établis pour un petit nombre de répétition. Puis lorsque la structure 3D d'un duplex de 6 répétitions CUG a été publié, le modèle de l'ARN $(CUG)_{17}$ a été modifié en utilisant les coordonnées atomiques du duplex $2x(CUG)_6$ (code PDB: 1ZEV) auquel ont été ajoutées 4 répétitions CUG en double-brin (en tenant compte de la symétrie du duplex) et une répétition correspondant à la boucle terminale CUG qui avait déjà été modélisée précédemment. La structure 3D d'un modèle ARN $r(CUG)_{17}$ est présentée ci-dessous (Fig. 3). La même approche sera utilisée pour construire l'ARN comportant 16 répétitions CUG : $r(CUG)_{16}$ (avec une boucle terminale "tetraloop"), et générer des conformations représentatives en solution par dynamique moléculaire. Ces modèles d'ARN seront ensuite utilisés pour les simulations de "docking" ARN/protéine.

Pour cette partie de modélisation des ARN, les ARN à séquences répétées CCUG restent encore à modéliser. Les répétitions CUG se structurent également sous la forme de longues tige-boucle, mais à la différence des répétitions CUG, elles possèdent 2 mésappariemments consécutifs C:U qui alternent avec 2 paires Watson-Crick G-C et C-G. Pour leur modélisation, la difficulté par rapport aux séquences CUG est cependant moindre dans la mesure où une structure 3D expérimentale comportant ces 2 mêmes mésappariemments consécutifs C:U dans le contexte d'un duplex ARN existe déjà (code PDB: 165D). La poursuite des travaux de modélisation des ARN sera effectuée par le nouveau postdoc modélisateur qui sera recruté en 2006 (en remplacement de Petar Mitrasinovic).

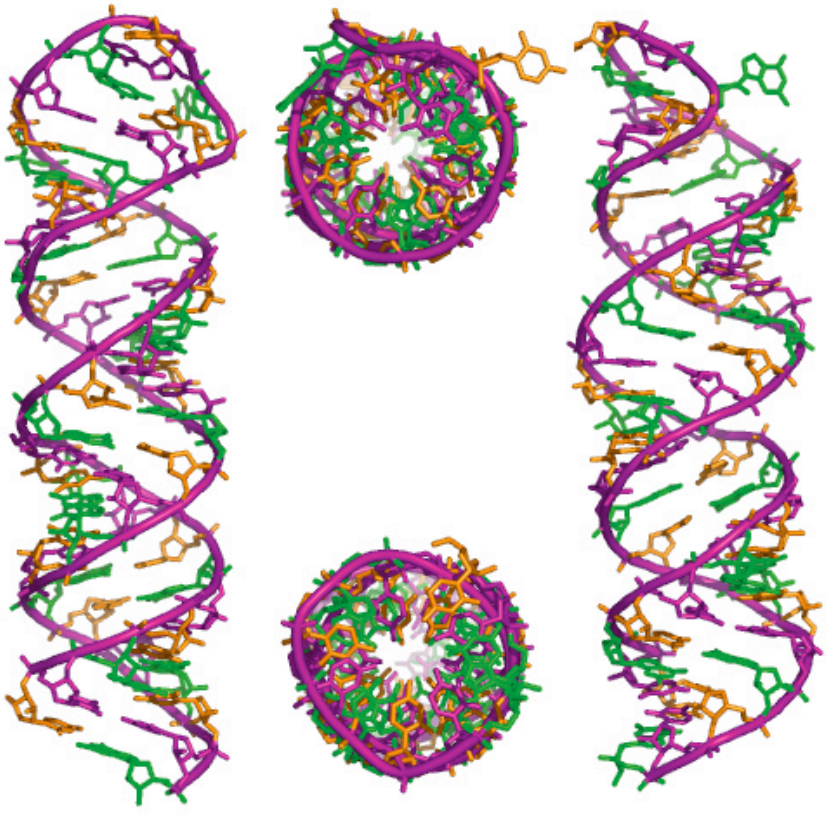


Figure 3 : Modèle 3D d'un ARN $r(CUG)_{17}$. Les 2 vues avec de la structure tige-boucle représentées longitudinalement montrent les sillons et la plus grande accessibilité du petit sillon. Les 2 vues axiales montrent une structure tige-boucle très rectiligne un peu moins régulière qu'une double hélice standard d'ARN. La région double-brin de l'ARN a été construite à partir des coordonnées atomiques de la structure 3D d'un duplex$2x(CUG)_6$ (code PDB : 1ZEV).

2.2.2 Modélisation 3D des protéines

Un modèle de structure 3D des domaines RRM1 et RRM2 de liaison à l'ARN de CUG-BP1 a été construit par J-A. Padron, ceci par une méthode de modélisation par homologie (Swiss-PDB Server) (Schwede et al., 2003). Pour la construction du modèle, J Padron a utilisé 3 structures expérimentales correspondant à ses 3 plus proches homologues structuraux: les protéines Hud et Huc, qui lient les extrémités 3'UTR d'ARNm à courte durée de vie (2 variants : PDB ID : 1GE2, chaîne A et 1FXL, chaîne A ; PDB ID : 1FNX, chaîne H), (Fig. 4). Une comparison du modèle 3D (Fig. 4A) avec plusieurs de ses homologues est montrée ci-dessous (Fig. 4B). D'autres protéines à domaines RRM, telles que la nucléoline ont un repliement identique : dans le cas de la nucléoline, le mode d'interaction ARN/protéine a été caratérisé (PDB ID : 1FJE, Fig. 4C). Le modèle de CUG-BP1 sera raffiné suite à la détermination récente d'une structure 3D en solution d'un des domaines RRM par RMN (code PDB : 2CPZ).

En ce qui concerne les autres protéines pour lesquelles nous avons prévu de modéliser le complexe formé avec les séquences répétées de type C(H)HG, figurait la protéine PKR, dont la structure 3D expérimentale est déjà connue (code PDB: 1QU6). Cependant, des données récentes suggèrent que l'affinité de PKR pour les séquences répétées CUG est très faible. PKR sera donc utilisée comme contrôle et système test dans les simulations de "docking".

Quant à la protéine MBNL-1, dont la modélisation ne faisait pas parti des objectifs initiaux, la construction d'un modèle 3D a été envisagée. Une tentative de modélisation a déjà été effectuée en collaboration avec le professeur Y. Zhou (University at Buffalo, The State University of New York), mais l'absence de similarités significative avec des protéines de structures 3D connues n'a pas permis de générer de modèles 3D satisfaisant. La cristallisation de la protéine MBNL-1 est donc envisagée au laboratoire par X Manival CR1 CNRS et C Charron IR CNRS, ceci afin d'en déterminer la structure 3D expérimentale par diffraction des rayons-X. Pour cela, un clonage de la protéine pour sa production en grande quantité a été réalisépar N Marmier-Gourrier (voir section 2.2.4).

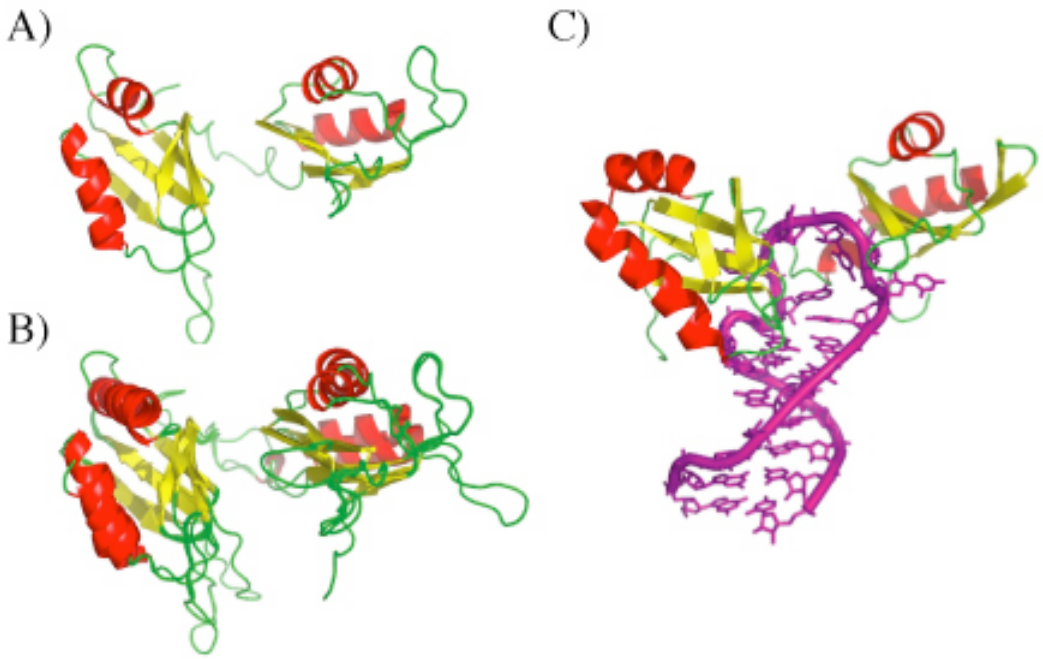


Figure 4: Structures 3D de protéines de laison à l'ARN à domaines RRM. A) Modèle 3D de CUG-BP1. B) Superposition du modèle 3D de CUG-BP1 avec ses homologues Huc et Hud. C) Complexe formé entre la nucléoline (protéine à domaines RRM) et sa séquence de reconnaissance du pre-ARNr (PDB ID : 1FJE).

2.2.3 Modélisation des complexes ARN/protéines

Une version mono-processeur de l'application destinée aux simulations de "docking" a été portée par Raphaël Bolze avec le concours de Fabrice Leclerc. Des tests de simulations sur la grille ont été effectués sur un système test. Le portage de la version multi-processeur de l'application est presque achevé. Des tests de simulations avec cette version de l'application qui sera utilisée en production seront effectués. La phase de production avec la soumission des simulations sur les complexes associant les ARN à séquences répétées de type C(H)HG et leurs protéines de liaison pourra alors débuter. La poursuite de ces travaux sur le "docking" ARN/protéine sera assurée par le nouveau chercheur postdoct modélisateur qui sera recruté en 2006. Le complexe entre l'ARN $(CUG)_{17}$ et la protéine PKR sera le premier système test utilisé pour la version multi-processeur de l'application. Le complexe entre l'ARN $(CUG)_{17}$ et la protéine CUG-BP1 sera ensuite modélisé. Les autres complexes ARN/protéine seront modélisés une fois que des modèles modèles satisafaisants ou des structures 3D auront été obtenus pour les protéines.

2.2.4 Mise en œuvre des techniques expérimentales pour la valisation des modèles

La validation expérimentale des modèles 3D ARN/protéines nécessite la production des molécules d'ARN et de protéines dont la liaison sera étudiée *in vitro* par des techniques de biologie moléculaire (empreintes chimiques et enzymatiques, gel retard, etc) qui sont utilisées couramment au MAEM. La première étape est donc le clonage des protéines CUG-BP1 et MBNL-1 et des séquences ARN cibles.

N Marmier-Gourrier a réalisé des constructions permettant de produire ces deux protéines chez *E. coli*, ainsi qu'une série de plasmides permettant de produire par transcription *in vitro* des ARN cibles contenant des nombres variables de répétitions CUG.

Les plasmides utilisés pour cloner les gènes de CUG-BP1 et MBNL1 (isoforme EXP40) (pGEX-6P1 et pGEX-4T1 Amersham Pharmacia Biotech) permettent de produire des protéines possédant une extension GST en N-terminal. Cette extension peut être éliminée par l'action de protéases spécifiques. De nombreux tests de productions ont déjà été réalisés (tests de la performance de différentes souches d'*E. coli* et de différentes conditions de fermentation). Les productions en grandes quantités s'avèrent difficiles du fait de problèmes de solubilité des protéines recombinantes. N Marmier-Gourrier continue à optimiser les conditions de production des 2 protéines. La possibilité de produire des domaines de ces protéines (domaines RRM de CUG-BP et fragment de MBNL-1) est actuellement testée. En effet, les travaux de Kino *et al.* (2004) ont montré que la protéine MBNL-1 tronquée de 101 ou 122 acides aminés, est capable d'interagir avec l'ARN dans un test triple-hybride. Nous espérons que cette diminution de taille favorisera la production de la protéine, en particulier pour les besoins de la cristallisation.

En ce qui concerne les partenaires ARN de ces protéines, N Marmier-Gourrier a construit des plasmides permettant de produire des ARN comportant 16 et 17 répétitions (Sobczak et al., 2003) mais aussi des répétitions plus longues $(CUG)_{51}$ et $(CUG)_{54}$ qui sont plus proches de la situation observée *in vivo*. Pour leur production, la technique décrite par Takahashi *et al.* (1999) a été utilisée. N Marmier Gourrier a commencé à étudier la structure secondaire adoptée par ces répétitions en solution. Les premières données ont montré que les analyses ne peuvent pas être réalisées par extension d'amorce par la transcriptase inverse, car du fait de la trop grande stabilité de la structure formée par les répétitions, les extensions d'amorces par la réverse transcriptase ne se sont pas processives. Les expériences vont donc être réalisées en marquant radioactivement les ARN en 5' terminal, et en faisant agir des RNases spécifiques de régions en double brin ou en simple brin, ceci

d'abord en absence de protéine, puis en présence de chaque protéine. Pour comparaison, N Marmier-Gourrier a aussi réalisé une construction permettant de produire par transcription *in vitro* un fragment du pré-ARNm de la Troponine T contenant des sites de fixation de CUG-BP et de MBNL-1. En effet, Ho et *al*. (Ho et al., 2004) avaient montré que l'épissage de l'exon 5 de la troponine T humaine est régulé de façon antagoniste par ces deux protéines, et que des sites de fixation de MBNL-1 se situent en amont de l'exon 5 et que des sites de fixation de CUG-BP1 se situent en aval de ce même exon. Un examen plus approfondi des séquences en aval de l'exon 5 ont permis de remarquer que la séquence contenant les sites de fixation de CUG-BP1 peuvent aussi constituer des sites reconnus par MBNL-1, ceci sur la base du consensus (YGCUG/UY), proposé par Ho *et al*. (2004). N Marmier-Gourrier a donc produit des constructions permettant la production d'ARN contenant les séquences reconnues par les deux protéines, ainsi que des variants dans lesquels un ou plusieurs sites sont mutés. Des expériences de retard sur gel ont été réalisées avec les protéines recombinantes produites, mais les conditions expérimentales doivent encore être améliorées du fait des problèmes de solubilité de ces protéines.

Ces travaux seront poursuivis par Nathalie Marmier-Gourrier afin de pouvoir tester et raffiner les modèles 3D des complexes ARN/protéine obtenus par docking. En raison des contraintes liées à la mise en œuvre de ces techniques, sa 1ère année sera entièrement focalisée sur la validation de la 2nde partie du projet (qui a été amorcée en premier). Lors de la 2nde année de son contrat, en fonction des résultats obtenus par le groupe de Y Guermeur sur la première partie du projet, elle pourra aussi participer à la validation des résultats obtenus par ce groupe.

2.3 Autres activités

Fabrice Leclerc a participé à l'animation organisée par la Coordination Départementale de Meurthe-et-Moselle du Téléthon à l'occasion de la fête de la science la journée du 15 septembre 2005, sur invitation de M. Pascal Laurent, responsable local. Une présentation simple du projet a été réalisée. Il a également participé, en tant que scientifique, à l'émission de radio "Grand défi : 40h pour le Téléthon" sur l'antenne de RCM (radio associative à Thiaville-sur-Meurthe) le 3 décembre 2005.

2.4 Références

Carninci, P et al. "The transcriptional landscape of the mammalian genome.". *Science* 309 (2005): 1559-1563.

Clark, F, and TA Thanaraj. "Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human.". *Hum Mol Genet* 11 (2002): 451-464.

Corcos, L, and S Solier. "[Alternative mRNA splicing, pathology and molecular therapeutics]". *Med Sci (Paris)* 21 (2005): 253-260.

Cummings, CJ, and HY Zoghbi. "Fourteen and counting: unraveling trinucleotide repeat diseases.". *Hum Mol Genet* 9 (2000): 909-916.

Wolpert, D.H. "Stacked Generalization". *Neural Networks* 5 (1992): 241-249.

Gautheret, D, and R Cedergren. "Modeling the three-dimensional structure of RNA.". *FASEB J* 7 (1993): 97-105.

Gautheret, D et al. "Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets.". *J Mol Biol* 229 (1993): 1049-1064.

Ho, TH et al. "Transgenic mice expressing CUG-BP1 reproduce splicing mis-regulation observed in myotonic dystrophy.". *Hum Mol Genet* 14 (2005a): 1539-1547.

Ho, TH et al. "Muscleblind proteins regulate alternative splicing.". *EMBO J* 23 (2004): 3103-3112.

Ho, TH et al. "Colocalization of muscleblind with RNA foci is separable from mis-regulation of alternative splicing in myotonic dystrophy.". *J Cell Sci* 118 (2005b): 2923-2933.

Jasinska, A et al. "Structures of trinucleotide repeats in human transcripts and their functional implications.". *Nucleic Acids Res* 31 (2003): 5463-5468.

Jiang, H et al. "Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons.". *Hum Mol Genet* 13 (2004): 3079-3088.

Kanadia, RN et al. "A muscleblind knockout model for myotonic dystrophy.". *Science* 302 (2003): 1978-1980.

Kino, Y et al. "Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats.". *Hum Mol Genet* 13 (2004): 495-507.

Krawczak, M et al. "The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences.". *Hum Genet* 90 (1992): 41-54.

Ladd, AN et al. "Dynamic balance between activation and repression regulates pre-mRNA alternative splicing during heart development.". *Dev Dyn* 233 (2005): 783-793.

Leppert, J et al. "Identification of NH...N hydrogen bonds by magic angle spinning solid state NMR in a double-stranded RNA associated with myotonic dystrophy.". *Nucleic Acids Res* 32 (2004): 1177-1183.

Major, F et al. "The combination of symbolic and numerical computation for three-dimensional modeling of RNA.". *Science* 253 (1991): 1255-1260.

Michalowski, S et al. "Visualization of double-stranded RNAs from the myotonic dystrophy protein kinase gene and interactions with CUG-binding protein.". *Nucleic Acids Res* 27 (1999): 3534-3542.

Miller, JW et al. "Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy.". *EMBO J* 19 (2000): 4439-4448.

Mooers, BH et al. "The structural basis of myotonic dystrophy from the crystal structure of CUG repeats.". *Proc Natl Acad Sci U S A* 102 (2005): 16626-16631.

Ranum, LP, and JW Day. "Pathogenic RNA repeats: an expanding role in genetic disease.". *Trends Genet* 20 (2004): 506-512.

Schwede, T et al. "SWISS-MODEL: An automated protein homology-modeling server.". *Nucleic Acids Res* 31 (2003): 3381-3385.

Shawe-Taylor J., Cristianini N. "Kernel methods for Pattern Analysis". (2004):

Sobczak, K et al. "RNA structure of trinucleotide repeats associated with human neurological diseases.". *Nucleic Acids Res* 31 (2003): 5469-5482.

Tuffery-Giraud, S et al. "Point mutations in the dystrophin gene: evidence for frequent use of cryptic splice sites as a result of splicing defects.". *Hum Mutat* 14 (1999): 359-368.

3. Rapport financier

bénéficiaire	UHP	CNRS
somme versée (21/02/2005) contrat PM	52 440€	
somme dépensée	30 590€	
somme non dépensée avant fin 2005 (contrat PM)	21 850€	
somme versée (21/02/2005) contrat NM-G et fontionnement		41 626,12€
somme dépensée contrat NM-G		21 084€
somme dépensée fonctionnement		414,41€
somme non dépensée avant fin 2005 (fonctionnement)		9585,59€
somme versée (6/09/2005) contrats DA et EM	92100,12€	
somme dépensée	23044,04€	

Les initiales utilisées dans la tableau font référence aux personnels recrutés dans le cadre du projet (PM : Petar Mitrasinovic, NM-G : Nathalie Marmet-Gourrier, DA : Delphine Autard, EM : Emmanuel Monfrini).

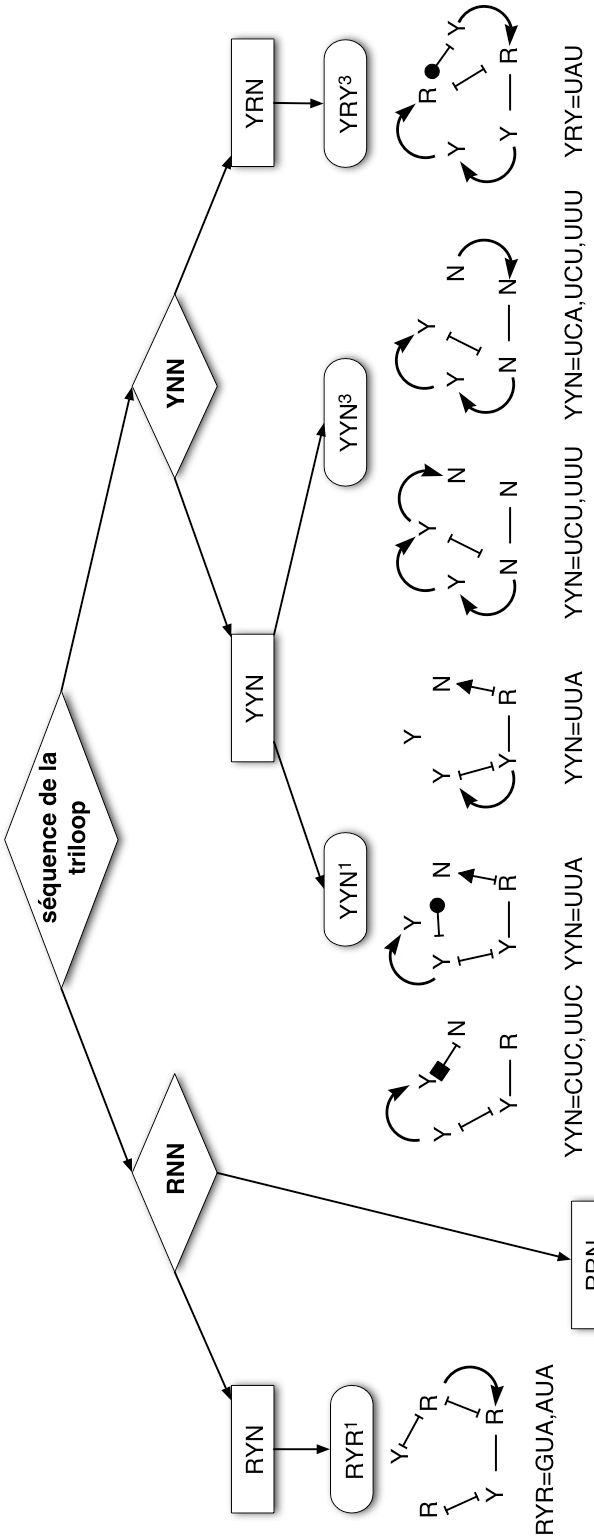
4. Annexes

4.1 Annexe 1

La liste des banques de données d'épissage qui ont été évaluées est fournie ci-dessous avec indication de l'hyperlien permettant d'y accéder :

EID (http://www.meduohio.edu/bioinfo/eid/index.html),
ExInt (http://sege.ntu.edu.sg/wester/exint/),
HS3D (http://www.sci.unisannio.it/docenti/rampone/),
AEDB (http://www.ebi.ac.uk/asd/index.html),
AltSplice (http://www.ebi.ac.uk/asd/index.html),
AltExtron (http://www.ebi.ac.uk/asd/index.html),
AsMamDB (http://166.11.30.65/AsMamDB/),
ProSplicer (http://prosplicer.mbc.nctu.edu.tw/),
SpliceDB (http://www.softberry.com/berry.phtml?topic=splicedb&group=data&subgroup=spldb),
HASDB (http://www.bioinformatics.ucla.edu/%7Esplice/HASDB/).

4.2 Annexe 2



Annexe 2 : Classification structurale des triloops dans des structures tige-boucle. La triloop CUG correspond à une séquence de type YNN/YYN/YYN[3]. La représentation schématique des re0pliements des triloops est donnée selon la nomenclature de Leontis et Westhof (2001).