HAL open science

Comparer des structures de gènes pour la prédiction de transcrits alternatifs codants chez l'humain, la souris et le chien

Nicolas Guillaudeux

▶ To cite this version:

Nicolas Guillaudeux. Comparer des structures de gènes pour la prédiction de transcrits alternatifs codants chez l'humain, la souris et le chien. Bio-Informatique, Biologie Systémique [q-bio.QM]. Université Rennes 1, 2021. Français. NNT : . tel-03593091v1

HAL Id: tel-03593091

https://theses.hal.science/tel-03593091v1

Submitted on 24 Jan 2022 (v1), last revised 1 Mar 2022 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Nicolas GUILLAUDEUX

Comparer des structures de gènes pour la prédiction de transcrits alternatifs codants chez l'humain, la souris et le chien

Thèse présentée et soutenue à Rennes, le 16 décembre 2021
Unité de recherche : Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), UMR 6074

Rapportrices avant soutenance :

Sèverine BERARD	Maîtresse de conférence, Université de Montpellier
Elodie LAINE	Maîtresse de conférence, Sorbonne Université

Composition du Jury :

Président :	Christian DIOT	Directeur de recherche, INRAE, Institut Agro Rennes
Examinateurs :	Sèverine BERARD	Maîtresse de conférence, Université de Montpellier
	Elodie LAINE	Maîtresse de conférence, Sorbonne Université
	Nicolas LARTILLOT	Directeur de recherche, CNRS, Université de Lyon 1
	Jean-Stéphane VARRE	Professeur, Université de Lille
Dir. de thèse :	Olivier DAMERON	Professeur, Université de Rennes 1
Co-dir. de thèse :	Catherine BELLEANNEE	Maîtresse de conférence, Université de Rennes 1
	Samuel BLANQUART	Chargé de recherche, Inria Rennes

Remerciements

Je ne suis pas spécialement un grand écrivain, ces remerciements ne seront donc pas très étendus et je tiens à m'excuser pour toutes les personnes que j'aurai oublié.

Tout d'abord, je tiens à remercier chacun des membres du jury d'avoir accepté de participer à l'examen de cette thèse. Notamment, je tiens à remercier Sèverine Bérard et Elodie Laine pour leur rapport détaillé et leurs nombreux conseils pour faire évoluer ce travail avec une mise en évidence plus précise de mes contributions et développements.

Je remercie également mes encadrants de thèse Catherine Belleannée et Samuel Blanquart de m'avoir accompagné dans cette aventure à la fois périlleuse et dangereuse. Toutes ces conversations scientifiques, techniques et le blabla quotidien m'ont été d'une grande aide dans le développement de cette rigueur nécessaire au travail d'un chercheur. Catherine, je n'oublierai pas la promesse du début de thèse : "Mais pourquoi on le fait ?", qui a bien sur un rôle dans l'avancement de la méthode et de résultats qu'on a pu obtenir. Bien entendu, Samuel, ton esprit de rebelle nous a propulsé sur de divers sujets de discussions et de débats. Je remercie également et bien entendu mon directeur de thèse, Olivier Dameron, et la joie administrative de cette direction.

Que serait une thèse sans toutes les personnes avec qui nous travaillons et nous discutons ? Et bien pas grand chose, alors bien entendu, merci à toutes les personnes qui ont eu le courage de partager leurs expériences, leurs soutiens tant moral que physique (oui il est très bon ton chocolat Clara). De nombreuses personnes sont à remercier pour cela : Kévin, Camille, Clara (encore), Victor, Méziane, Méline, Lolita, Lucas, Sarah, et à toutes les personnes que j'oublie (mais qui se reconnaîtrons j'en suis sûr).

En somme, merci aux équipes de Symbiose (Dyliss, Genscale et GenOuest) pour ce superbe environnement de travail que vous nous offrez. Merci à Marie, sans toi nous serions perdus dans les méandres de l'administratif. Merci également au service informatique du labo pour résoudre nos petits soucis techniques (promis je ne referai plus de boucles, ni n'installerai Conda sous Fedora). A ce niveau, il est important de remercier celui qui a toujours été à mes côtés du début à la fin, sur qui mes nerfs et mon humeur ont fait fureur... Mon petit Raptor. Parce que si il y a bien une *clever girl* dans le monde (parce que oui j'étais obligé de mettre au moins une référence à cette œuvre de culture que je

ne cesserai pas de revendiquer) c'est bien toi. Toi et tes caprices du refus de travailler, de te connecter, de te rallumer, de fonctionner... mais après tout je n'ai pas non plus été délicat avec ton système d'exploitation.

Ensuite, il semble évident que je remercie toutes les personnes de ma famille qui m'ont accompagné et écouté tout au long de cette thèse. Parce que oui, je suis fier de remercier mes parents, mes quatre chipies de sœurs, mes grand-parents et mon arrière grand-mère. Mais surtout j'aimerais dédier ce travail de thèse à la personne qui m'a le plus encouragé, et qui a toujours été là pour moi, qui m'a toujours dit que j'irai jusqu'au bout et qui croyait en moi. Ma grand-mère Evelyne, ça a été une année difficile mais je sais qu'aujourd'hui où que tu sois, tu es fier de cette route que je parcours.

Pour conclure sur ces remerciements, je remercie Aurélie qui m'accompagne au quotidien et qui a supporté mes crises de nerfs et mes humeurs changeantes selon les conditions de la journée et de l'avancement de chaque étape de la thèse.

Sommaire

Introduction **11**

1 Contexte biologique et état de l'art **15**

1.1 Mécanismes biologiques de la production de transcrits 18

1.1.1 Génome des organismes eucaryotes 18

1.1.2 Transcription du gène en transcrit, puis traduction en protéine . . . 19

1.1.2.1 Transcription du gène en ARN 20

1.1.2.2 Épissage 21

1.1.2.3 Traduction 22

1.1.3 Diversité des transcrits d'un gène eucaryote 23

1.1.3.1 Quantification de la diversité des transcrits 23

1.1.3.2 Mécanismes de la diversité des transcrits 24

1.2 Données bioinformatiques concernant les gènes et les transcrits 27

1.2.1 Données publiques, annotation et génome de référence 27

1.2.2 Formats des données 29

1.2.3 Représentation des données 30

1.2.4 Séquençage de l'ADN et de l'ARN 31

1.2.4.1 Lectures courtes 31

1.2.4.2 Lectures longues 33

1.2.4.3 Autres technologies de séquençage ciblées 34

1.3 Travaux sur l'identification et la comparaison de transcriptomes 35

1.3.1 Observer les transcrits 35

1.3.2 Méthodes de prédiction de transcrits 35

1.3.3 Identifier les évènements d'épissage, CRAC 37

1.3.4 Aligner des transcrits par programmation dynamique 38

1.3.5 Aligner des transcrits chez plusieurs espèces, SplicedFamAlign . . . 41

1.3.6 Construire des graphes d'épissage multi-espèces, ThorAxe 42

1.3.6.1 Graphe d'épissage multi-espèces 42

1.3.6.2 Nœuds et alignements multiples 43

1.3.6.3 Orthologie des sous-exons 43
1.3.7 Génomique comparative entre deux espèces et prédiction de sites et transcrits orthologues, CG-alcode 44
1.3.7.1 Considérer la séquence complète du gène 44
1.3.7.2 Aligner des modèles de gènes 44
1.3.7.3 Identifier l'orthologie des transcrits et prédire des transcrits 45
1.4 Cas d'étude : l'humain, la souris et le chien 48
1.4.1 Choix des organismes d'étude 48
1.4.2 Intérêt de passer à une comparaison multi-espèces 49
1.4.3 Identifier un ensemble de gènes conservés 50

2 Représentation, comparaison et prédiction de structures de gènes et de transcrits 53
2.1 Représentation de la structure des gènes et des transcrits 56
2.1.1 Unités lexicales des modèles : sites fonctionnels et blocs codants . . 56
2.1.2 Modèle de structure de gène 58
2.1.3 Modèle de structure de transcrit 58
2.2 Comparaison des structures de gènes et de transcrits entre deux espèces : définitions et prédictions 60
2.2.1 Comparaison des structures de deux gènes orthologues et prédiction de sites fonctionnels 60
2.2.2 Comparaison des structures de transcrits et prédiction de transcrits 62
2.3 Vers le multi-espèce : comparaison de gènes orthologues sur trois espèces . 66
2.3.1 Comparer plus de deux espèces : graphes de sites fonctionnels entre gènes orthologues 66
2.3.2 Identification de transcrits structurellement conservés (groupes de CDS orthologues) : graphes de transcrits 66

3 Analyse de transcrits de l'humain, de la souris et du chien : prédictions des transcrits par génomique comparative 69
3.1 Données utilisées pour l'humain, la souris et le chien 72
3.2 Prédiction de transcrits à partir de 2 167 triplets de gènes orthologues et de 18 109 transcrits connus 73
3.2.1 Exemple du gène CREM 73
3.2.2 Complétion des transcriptomes : 6 861 transcrits prédits 75

3.2.3 Espèces modèles et non modèles 77
3.3 Analyse de la fiabilité des prédictions 77
3.3.1 Recherche dans des bases de données 78
3.3.2 Recherche dans des données de lectures de séquençage : jonctions d'exons spécifiques 81
3.3.2.1 Identification des jonctions d'exons spécifiques 82
3.3.2.2 Jeux de données de lectures 83
3.3.2.3 Préparation des données en vue de l'alignement avec des lectures courtes 83
3.3.3 Résultats de la recherche des transcrits prédits dans des bases de données auxiliaires 88
3.3.4 Résultats de la recherche des transcrits prédits dans des jeux de données de RNA-seq 88
3.3.4.1 Résultats obtenus 88
3.3.4.2 Couverture des jonctions d'exons 89
3.3.5 Résultats des deux méthodes sur les données complètes 90

4 Identification d'un ensemble de gènes de structure d'épissage conservés entre l'humain, la souris et le chien 93
4.1 Comparaison trois-espèces 96
4.1.1 Relations d'orthologie trois espèces 96
4.1.2 Graphes de sites fonctionnels : interprétation phylogénétique de la conservation des sites chez trois espèces 96
4.1.3 Graphes de transcrits : identification des groupes de CDS orthologues chez trois espèces 98
4.2 Construction des graphes à partir des données 100
4.2.1 Hypothèses des comparaisons de paires de gènes 100
4.2.2 Hypothèses des graphes de sites fonctionnels 102
4.2.3 Hypothèses des graphes de transcrits 103
4.3 Exemple du gène CREM et illustration des divergences 106
4.4 Application des graphes à la détection de structures strictement conservées 110
4.4.1 253 gènes de structure d'épissage conservée 111
4.4.2 Transcriptomes des gènes structurellement conservés 112
4.4.2.1 CDS en copie unique : Analyse des 135 triplets de gènes . 114

4.4.2.2 CDS en copies multiples : Analyse des 118 triplets de gènes 115
4.4.3 Base de données de 253 triplets de gènes structurellement conservés 118

5 Applications et perspectives 121
5.1 Génomique comparative 122
5.1.1 Comparaison des transcrits dans leur intégralité : analyse des régions UTR 122
5.1.2 Comparaison des éléments régulateurs 125
5.2 Transcriptomique comparative 126
5.3 Évaluation comparative des méthodes d'alignement épissé 128
5.4 Poursuite de la méthode de comparaison multi-espèces 129

Conclusion 131

Bibliographie 135

Liste des acronymes 143

Liste des figures 145

Liste des tableaux 149

Liste des algorithmes 151

Annexes 153
Annexe 1 : Comparaisons de la paire de gènes CREM entre l'humain et le chien 154
Annexe 2 : Comparaisons de la paire de gènes CREM entre la souris et le chien 156
Annexe 3 : Identifiants attribués aux transcrits prédits 158
Annexe 4 : Utilisation de BedTools intersect pour la recherche de transcrits prédits dans les données auxiliaires 159
4.1 : Description du contenu d'un fichier au format BED12 adapté aux transcrits 159
4.2 : Exécution de BedTools intersect pour la recherche de transcrits prédits dans des bases de données auxiliaires 160
4.3 : Exécution de BedTools intersect pour la recherche de jonctions d'exons spécifiques aux transcrits prédits 160

Annexe 5 : Détails des échantillons de tissus utilisés pour la confortation des jonctions d'exons spécifiques 161
5.1 : Échantillons de tissus utilisés chez l'humain 161
5.2 : Échantillons de tissus utilisés chez la souris 162
5.3 : Échantillons de tissus utilisés chez le chien 163
Annexe 6 : Détails des prédictions et des alignements sur un triplet de gènes . . 164
Annexe 7 : Détails de la composition des tables de la base de données 167

Introduction

En biologie moléculaire, un dogme fondamental issu de la biologie des bactéries et des archées impose que le produit d'un gène soit une et une seule protéine. Cependant, chez les eucaryotes, les mécanismes d'expression du gène sont plus complexes et le dogme "un gène exprime une protéine" n'est plus applicable. En effet, l'expression des gènes est régulée par des phénomènes qui vont permettre à ces gènes d'exprimer plusieurs protéines n'ayant pas les mêmes fonctions (Kelemen et al. 2013 ; Blencowe 2017 ; Sulakhe et al. 2019). Ces phénomènes correspondent à la transcription alternative et à l'épissage alternatif, et ils concernent la plupart des gènes eucaryotes (E. T. Wang et al. 2008 ; Chaudhary et al. 2019). Depuis la découverte des exons (Gilbert 1978), les fragments qui portent l'information fonctionnelle des gènes, il a été démontré que leur sélection alternative induit une combinatoire de transcrits possibles (Smithers et al. 2019). A eux deux, la transcription et l'épissage alternatifs permettent aux 20 000 gènes connus chez l'humain de produire plus de 237 000 transcrits (Ensembl 2021, Howe et al. 2021). Par la suite, un certain nombre de ces transcrits vont être traduits en protéines quantifiées à hauteur du million de protéines différentes nommées *isoformes*. Les transcrits peuvent permettre des interactions spécifiques avec des protéines et des ligands (Ellis et al. 2012 ; Sulakhe et al. 2019), une localisation subcellulaire spécifique (Kelemen et al. 2013), une expression différentielle dans des tissus spécifiques (Kelemen et al. 2013 ; E. T. Wang et al. 2008 ; Ellis et al. 2012), et ils peuvent être spécifiques aux stades de développement, à l'âge (Kalsotra et Cooper 2011 ; Mazin et al. 2013) et au sexe (Oliva et al. 2020). Dans cette thèse, nous nous intéressons à la question ouverte : "*Quels sont les transcrits codants exprimables par un gène ?*".

Les techniques de séquençage permettent de détecter les transcrits exprimés à partir d'échantillons biologiques. Afin de traiter ces données, des méthodes bioinformatiques ont été mises en place dans le but de pouvoir estimer les transcrits qu'un gène est capable de produire. Pour cela, les données sont issues de différents projets. Entre 1988 et 2003, le premier génome humain a été séquencé. Ce projet a ensuite été suivi par le *1000 genome project* entre 2008 et 2010 puis par le *100000 genome project* entre 2012 et

2018. De nombreux autres projets émergent ainsi dans le but de compléter et d'affiner les connaissances sur les transcrits, mais obtenir des transcrits complets est difficile. En effet, certains transcrits sont rares, peu exprimés, ou dépendants de conditions physiologiques spécifiques. Ainsi, les connaissances sur les transcrits existants issus des analyses de séquençage ne sont pas exhaustives et de nouvelles méthodes alternatives sont nécessaires. Dans cette thèse, nous nous sommes intéressés à compléter cette connaissance grâce aux méthodes de génomique comparative, avec un focus sur les régions codantes des transcrits codants, ces régions qui ont la capacité de pouvoir être traduites en protéines.

D'un point de vue évolutif, on peut comparer les gènes entre espèces au travers de différentes ressources afin de déterminer les gènes orthologues, puis les isoformes orthologues (Zambelli et al. 2010). On décrit un caractère orthologue comme un caractère hérité d'un ancêtre commun et présent en copie unique chez les descendants. Dans cette thèse, pour déterminer les transcrits exprimables par un gène, nous nous sommes appliqués à décrire une méthodologie de comparaison de la séquence des gènes sur plus de deux espèces. Nous avons appliqué la comparaison sur un ensemble de trois espèces : l'humain, la souris et le chien. Cette comparaison à l'échelle multi-espèces permet de s'interroger sur la conservation de la structure d'épissage des gènes et sur la conservation de l'ensemble des transcrits exprimables par un gène. Pour cela, nous nous sommes basés sur une méthode de génomique comparative pré-existante (Blanquart et al. 2016) de comparaisons de structures de gènes entre deux espèces qui avaient été appliquées à l'humain et la souris. Cette méthode compare une paire de gènes orthologues, c'est-à-dire un gène chez une espèce avec le gène orthologue chez l'autre espèce. La méthode détermine si un transcrit exprimé par le gène d'une espèce peut-être exprimé par le gène de l'autre espèce. Si oui, un transcrit orthologue est prédit chez l'autre espèce. La méthode propose une représentation abstraite des gènes et des transcrits (Ouangraoua et al. 2012).

Dans cette thèse, nous avons étendu la méthode pour permettre une comparaison multi-espèces. Elle repose sur une représentation des structures orthologues sous forme de graphes. Nous l'avons appliqué à trois espèces, l'humain, la souris et le chien, à la recherche de gènes dont la structure en exons, les sites fonctionnels d'épissage et les transcrits sont conservés. Déterminer un répertoire de transcrits productibles et conservés par un gène permet de comprendre à une plus grande échelle la régulation des gènes eucaryotes. Avec cette étude, on peut ainsi apporter une première réponse à la question : *Si des gènes partagent une structure similaire entre plusieurs espèces, est-ce qu'ils partagent également*

les mêmes transcrits ?

Dans le Chapitre 1, nous décrivons le contexte de la thèse et l'état de l'art sur les mécanismes responsables de la production des transcrits à partir des gènes (la transcription alternative et l'épissage alternatif), sur les méthodes qui permettent de déterminer la séquence ARN des transcrits à partir d'échantillons biologiques (les méthodes de séquençage à haut débit) et les méthodes qui permettent de prédire des transcrits complets et/ou des structures orthologues en se basant sur les connaissances et sur leur structure épissée (les méthodes d'alignements épissés). Nous présentons également les représentations usuelles de la structure d'un gène et des transcrits ainsi que leur composition.

Dans le Chapitre 2, nous proposons un formalisme décrivant les structures d'épissage des gènes et des transcrits. Nous décrivons comment ces structures sont comparées entre deux gènes et nous proposons des définitions de l'orthologie des transcrit basées sur la conservation de ces structures. Cette définition de l'orthologie entre transcrits dont la structure est conservée permet de prédire de nouveaux transcrits. Nous décrivons également les méthodes développées durant la thèse pour la comparaison multi-espèces en s'appuyant sur des structures de graphes.

Dans le Chapitre 3, nous appliquons la méthode de prédiction de transcrits orthologues à des données disponibles pour trois espèces, l'humain, la souris et le chien. Le chapitre décrit, de plus, la recherche de traces de l'existence de ces transcrits prédits parmi des données auxiliaires (autres bases de données de transcrits annotés et données de lectures issues de séquençage).

Dans le Chapitre 4, nous détaillons la construction des graphes de sites orthologues et de transcrits orthologues obtenus à partir de nos données sur les trois espèces, et nous expliquons la topologie des graphes obtenus. Nous montrons comment nous identifions un ensemble de gènes dont la structure et le potentiel codant, les protéines productibles, sont conservés chez les trois espèces.

Enfin, dans le Chapitre 5, nous exposons les perspectives qui peuvent découler des méthodes et les résultats obtenus, notamment concernant les comparaisons des régions non traduites des transcrits codants.

Chapitre 1

Contexte biologique et état de l'art

Ce chapitre présente le contexte biologique et un état de l'art concernant les différents points examinés dans cette thèse. Nous introduisons d'abord les notions biologiques concernées par la thèse : les génomes, les gènes et les transcrits, et les mécanismes nécessaires à la production des transcrits à partir des gènes. Ensuite nous présentons les données bioinformatiques disponibles concernant les transcriptomes : les banques de données publiques et les formats d'annotations. Enfin, nous décrivons plusieurs travaux connexes à nos recherches concernant la prédiction de transcrits ou la représentation des transcrits d'un gène.

Sommaire

1.1 Mécanismes biologiques de la production de transcrits 18
1.1.1 Génome des organismes eucaryotes 18
1.1.2 Transcription du gène en transcrit, puis traduction en protéine 19
1.1.3 Diversité des transcrits d'un gène eucaryote 23
1.2 Données bioinformatiques concernant les gènes et les transcrits 27
1.2.1 Données publiques, annotation et génome de référence 27
1.2.2 Formats des données 29
1.2.3 Représentation des données 30
1.2.4 Séquençage de l'ADN et de l'ARN 31
1.3 Travaux sur l'identification et la comparaison de transcriptomes 35
1.3.1 Observer les transcrits 35
1.3.2 Méthodes de prédiction de transcrits 35
1.3.3 Identifier les évènements d'épissage, CRAC 37
1.3.4 Aligner des transcrits par programmation dynamique 38
1.3.5 Aligner des transcrits chez plusieurs espèces, SplicedFamAlign 41
1.3.6 Construire des graphes d'épissage multi-espèces, ThorAxe . . . 42
1.3.7 Génomique comparative entre deux espèces et prédiction de sites et transcrits orthologues, CG-alcode 44
1.4 Cas d'étude : l'humain, la souris et le chien 48
1.4.1 Choix des organismes d'étude 48
1.4.2 Intérêt de passer à une comparaison multi-espèces 49
1.4.3 Identifier un ensemble de gènes conservés 50

Définitions des termes du chapitre

Un *génome* est un ensemble de séquences d'ADN comprenant l'ensemble de l'information génétique.
Un *gène* est une région du génome pouvant être transcrite en ARN.
La *transcription* est un processus de copie d'un gène (ADN) en ARN.
Le *transcriptome* est l'ensemble des transcrits du génome d'un organisme. Dans la thèse, nous parlons du transcriptome d'un gène comme l'ensemble des transcrits exprimables par un gène.
La *traduction* est le processus permettant à certains transcrits d'être traduits en protéine, la région décrivant la protéine est dite codante (CDS, *coding sequence*).
Un *ARNm* ou *transcrit* est un produit d'expression issu d'un gène via la transcription et l'épissage formé par une combinatoire d'exons et pouvant être traduit en protéine.
Un *exon* est un segment d'un gène transcrit en ARN et retenu dans un transcrit mature.
Un *intron* est un segment d'ARN entre deux exons contigus et qui est supprimé lors du processus d'*épissage* produisant un transcrit mature, ou ARN.
Les *sites fonctionnels* correspondent aux codons *start* (qui selon nos conventions seront notés "[") et *stop* ("]") délimitant un CDS et aux sites donneurs ("<") et accepteurs d'épissage (">") formant le premier et le dernier dinucléotides d'un intron.
La *structure d'un gène* désigne la succession des exons et des sites fonctionnels qui le composent, permettant d'identifier la composition en introns et en exons des transcrits du gène en fonction de l'ensemble des CDS exprimés.
Un *CDS* (*coding sequence*) est la région d'un ARNm qui peut être traduite en protéine, depuis un codon *start* jusqu'à un codon *stop*, séparés par un nombre entier de codons consécutifs et sans codon *stop* intermédiaire en phase.

1.1 Mécanismes biologiques de la production de transcrits

1.1.1 Génome des organismes eucaryotes

Le génome est l'ensemble du matériel génétique d'un organisme stocké sous forme de chromosomes. Ces chromosomes sont contenus, chez les eucaryotes, dans le noyau des cellules qui constituent les tissus de l'organisme. Chaque chromosome est constitué de deux molécules d'acide désoxyribonucléique (ADN) sous la forme de deux brins complémentaires, représentant une partie de l'information génétique, et de protéines, notamment des histones responsables de la condensation de l'ADN. Chaque brin est composé d'unités d'information que l'on nomme nucléotides : l'adénine (A), la thymine (T), la guanine (G) et la cytosine (C). Une complémentarité entre les nucléotides existe et permet une liaison stable entre les deux brins complémentaires d'ADN : A se lie à T et G se lie avec C. D'une manière plus formelle, on peut décrire un brin d'ADN comme étant une séquence sur un alphabet Σ de quatre lettres avec $\Sigma = \{A, C, G, T\}$. Chaque brin se lit dans le sens allant d'une extrémité nommée "5'" vers l'extrémité nommée "3'". Les deux brins d'ADN sont orientés parallèlement mais en sens opposé, on parle ainsi de brins d'ADN complémentaires et antiparallèles.

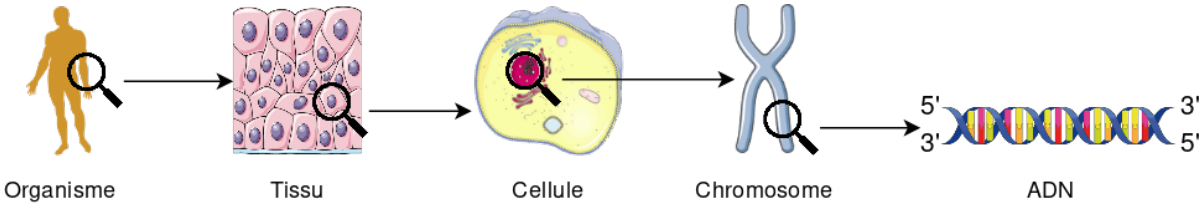


FIGURE 1.1 – L'information génétique chez un organisme eucaryote. Une molécule d'ADN double brin forme un chromosome localisé dans le noyau des cellules eucaryotes. Un chromosome est formé de deux molécules d'ADN complémentaires (paires complémentaires $A - T$, $G - C$) et antiparallèle.

L'information génétique portée par l'ADN a une organisation structurée. En effet, l'ADN contient notamment des segments d'intérêt particulier nommés *gènes*. On distingue deux catégories de gènes, des *gènes codants* et des *gènes non codants*. Les gènes codants produisent des *transcrits* dits *codants* et pouvant être traduits en *protéines*. Les gènes non codants ne produisent que des *transcrits* dits *non codants* qui ne sont pas traduits, et qui sont responsables de nombreuses régulations au sein de l'organisme. Chez l'humain, on

estime qu'il y aurait plus de 20 000 gènes codants pour des protéines, ce qui représente 1 à 2% des nucléotides du génome complet. La grande majorité du génome est composée d'ADN non codant (dont les gènes non codants).

1.1.2 Transcription du gène en transcrit, puis traduction en protéine

Chaque gène eucaryote est capable d'exprimer des acides ribonucléiques (ARN) ou *transcrits*. Pour cela, deux mécanismes sont nécessaires : la transcription et l'épissage. Le premier produit une copie de l'ADN sous forme d'un ARN complémentaire, le *pré-ARN*, alors que le second va être responsable de la maturation de l'*ARN*, en y sélectionnant des segments particuliers, les *exons*, par rapport à d'autres segments éliminés, les *introns*. Dans le cas des gènes codants, on obtient un *transcrit mature* ou *ARN messager* (ARNm) à partir d'un pré-ARN messager (pré-ARNm). Celui-ci contient une *séquence codante* (CDS, *coding sequence*) pouvant être traduite en protéine. Le processus de traduction de l'ARNm en protéine est enfin réalisé par le ribosome qui traduit la séquence ARN du CDS (codon par codon du codon *start* au codon *stop*) pour former une protéine (Figure 1.2). Soulignons enfin que la synthèse des transcrits est un processus hautement régulé, depuis son initiation jusqu'à sa destruction.

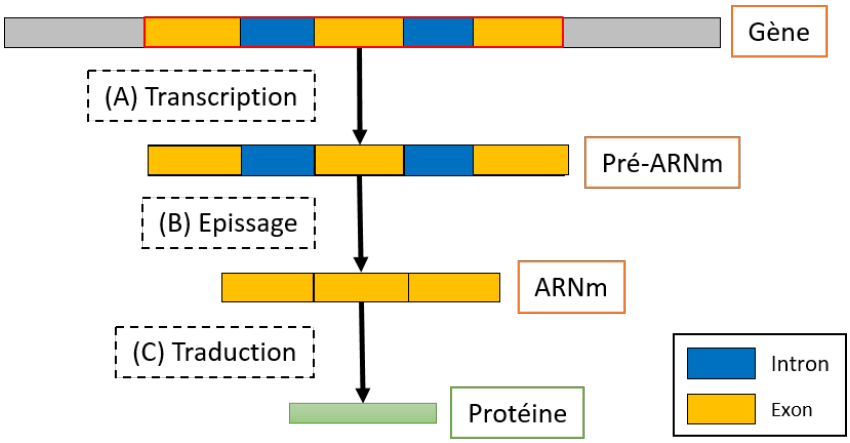


FIGURE 1.2 – Les mécanismes généraux d'expression des gènes codants. A partir d'un gène codant, la transcription (A) synthétise une séquence complémentaire nommée pré-ARNm qui contient certains segments qui vont être sélectionnés, les exons, au détriment de certains segments qui vont être éliminés par l'épissage (B), les introns. Ce processus produit un ARNm qui peut ensuite être traduit pour former une protéine (C).

1.1.2.1 Transcription du gène en ARN

Divers éléments de la séquence d'ADN d'un gène sont exposés pour permettre d'initier la transcription de l'ADN en *pre-ARNm* (Figure 1.2(A)). Le *promoteur*, un segment d'ADN en amont de la région transcrite, est l'un de ces éléments. Le promoteur est responsable de la régulation de l'expression du gène. Les promoteurs peuvent fixer des protéines que l'on nomme *facteurs de transcription*. Ils vont avoir une activité activatrice de la transcription, on parle d'éléments *enhancer*, ou au contraire avoir une activité inhibitrice de la transcription, on parle d'éléments *silencer*. Lorsque le promoteur est actif, une ARN polymérase (ARN pol II dans le cas des ARNm) se fixe sur le site d'initiation de la transcription et commence la transcription (Figure 1.3). L'ARN pol II va construire la séquence ARN complémentaire du gène. A la différence de l'ADN, l'ARN ne comporte pas de thymine (T) qui est remplacée par l'uracile (U). Ainsi, l'ARN pol II synthétisera A à la place de T, C à la place de G, G à la place de C et U à la place de A. Dans le cadre de la thèse, par simplicité, nous utiliserons T dans la description de toutes les séquences nucléotidiques. A l'issue de la transcription, une molécule d'ARN dite pré-ARNm a été synthétisée, c'est une copie d'une région du gène.

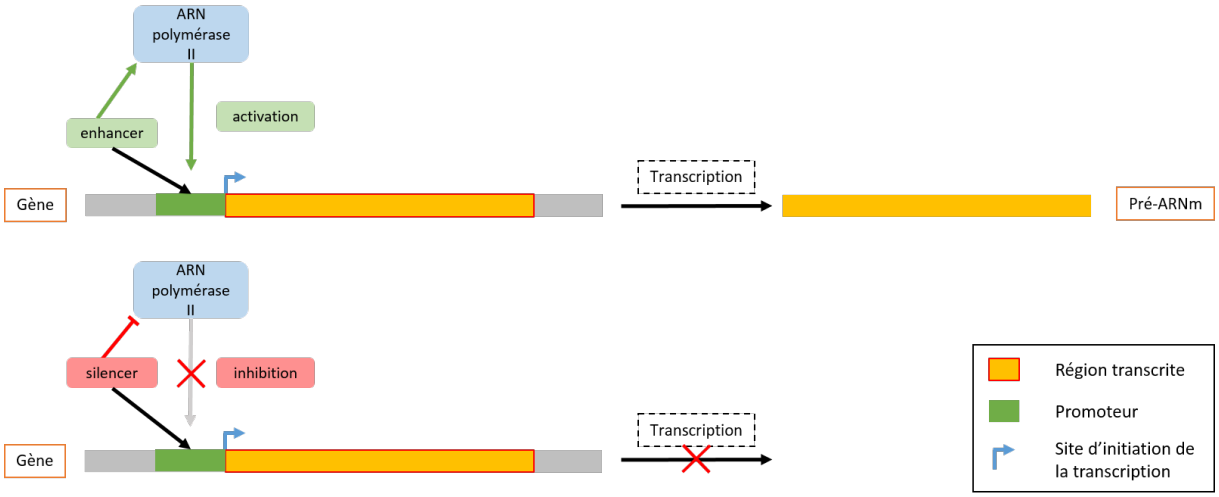


FIGURE 1.3 – Régulation de la transcription. Si des *enhancer* se fixent au niveau du promoteur, zone du gène représentée en vert, l'ARN polymérase II se positionne pour synthétiser un pré-ARNm. Si des *silencer* sont recrutés, la transcription n'est pas réalisée. La région transcrite, en jaune, du gène va du site d'initiation de la transcription jusqu'à un dernier site, dit de terminaison de transcription. La transcription produit une copie ARN complémentaire de cette région, le pré-ARNm.

1.1.2.2 Épissage

Le pré-ARNm comporte deux types de segments, les *introns* et les *exons*. Les introns correspondent à des séquences qui sont excisées et qui ne seront pas présentes dans le transcrit dit mature. Les exons, quand à eux, sont des séquences qui sont conservées et concaténées entre elles pour former le transcrit mature : l'ARNm. Ce mécanisme est nommé l'*épissage* et utilise entre autre les *sites d'épissage*, qui délimitent les introns (Figure 1.2(B)). Ces sites d'épissage comportent en particulier des dinucléotides extrêmement conservés et constituant les extrémités des introns. Dans le cas majoritaire de l'épissage U2-dépendant (Sibley et al. 2016), en 5' de l'intron, on retrouve le dinucléotide "GT", débutant le *site donneur d'épissage*, et du côté 3', on retrouve le dinucléotide "AG", terminant le *site accepteur d'épissage*. Les dinucléotides "GT"/"AG" sont présents dans 99% des introns humains. Ces dinucléotides sont reconnus par des ribonucléoprotéines nucléaires (*small nuclear ribonucleoprotein*, snRNP) qui vont avoir un rôle dans la définition d'un intron et d'un exon en agissant de manière à former un complexe nommé *spliceosome* responsable du mécanisme d'épissage (Kelemen et al. 2013).

Une fois épissé, l'ARN est une structure instable qui doit être stabilisé, c'est ce qu'on nomme la maturation. Pour être maturé, une coiffe est ajoutée à l'extrémité 5' et une queue poly(A) (plusieurs dizaines voire centaines d'adénosines) est synthétisée en 3' (Figure 1.4). Ces nucléotides ajoutés à l'ARNm ne font donc pas partie de la séquence du gène.

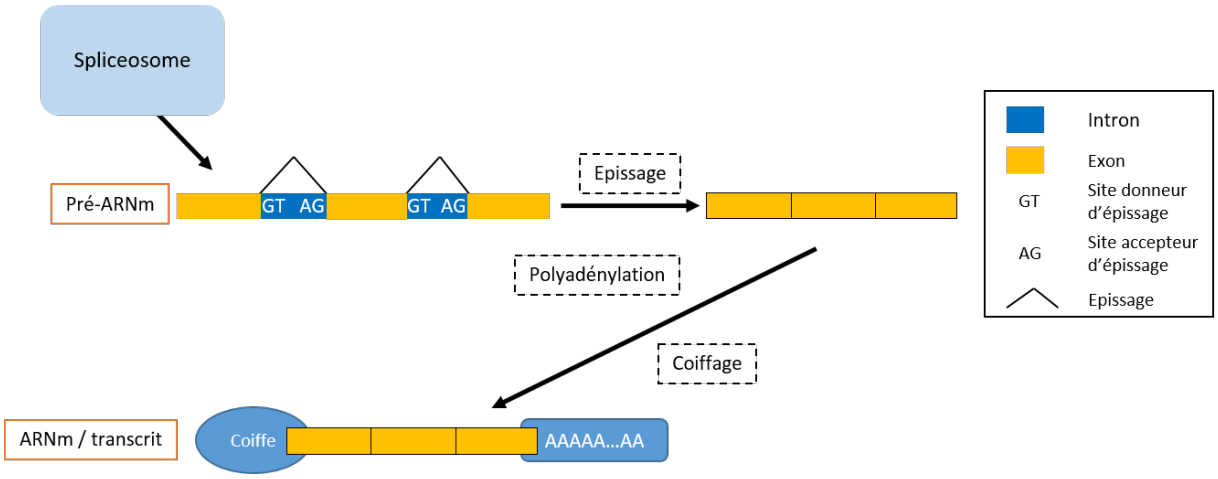


Figure 1.4 – Maturation d'un pré-ARNm. Le spliceosome excise les introns et concatène les exons en localisant les sites d'épissage. La maturation implique enfin l'ajout d'une coiffe à l'extrémité 5' et d'une queue poly(A) à l'extrémité 3' pour former le transcrit mature.

1.1.2.3 Traduction

Les gènes codants expriment des transcrits codants, les ARNm, et des transcrits non codants. Si le transcrit maturé est codant, il peut être traduit en protéine. C'est le mécanisme de la traduction (Figure 1.5). Par définition, les transcrits non codants ne sont pas traduits en protéines. Ces transcrits ont une fonction régulatrice sur le maintien du génome, la régulation des transcrits codants, la différenciation cellulaire, le choix de la lignée cellulaire, de l'organogenèse et de l'homéostasie tissulaire (Mattick et Makunin 2006 ; Marchese et al. 2017 ; Schmitz et al. 2016). Afin d'être traduits par le ribosome, les transcrits matures codants possèdent une séquence codante (CDS, *coding sequence*) qui est lue par triplets consécutifs de nucléotides, les codons. Chaque codon, à l'exception des 3 codons *stop* possibles, correspond à un acide aminé, l'unité de base d'une protéine. Le premier codon qui est lu par le ribosome est le codon d'initiation ou codon *start*, "ATG", aussi traduit en méthionine. Le dernier est le codon de terminaison ou codon *stop*, "TAA", "TAG" ou "TGA". Ces 3 codons ne codent pour aucun acide aminé. On parle de code génétique pour désigner la correspondance entre les 64 codons possibles et les 20 acides aminés. Les nucléotides précédents et suivants de ces deux codons, en amont du codon *start* et en aval du codon *stop*, ne sont pas traduits et composent les régions 5' et 3' non traduites de l'ARNm (les UTR, *untranslated region*).

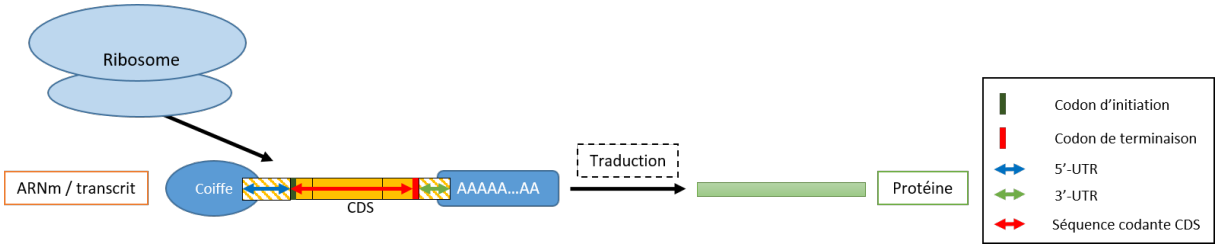


Figure 1.5 – Traduction d'un ARNm. Le ribosome traduit le CDS de l'ARNm du codon d'initiation jusqu'au codon de terminaison. Chaque codon, à l'exception du codon *stop*, est traduit en un acide aminé constituant la protéine.

La transcription et l'épissage sont des étapes qui sont réalisées dans le noyau des cellules eucaryotes. Une fois maturé, l'ARNm est exporté en dehors du noyau, dans le cytoplasme et l'appareil de Golgi, où il pourra être traduit par le ribosome puis dégradé. Excepté en ce qui concerne la traduction, ces étapes sont aussi valables pour les ARN non codants. Si un ARN est mal synthétisé (par exemple en incluant un codon *stop* prématuré), le processus de dégradation NMD (*Non sense Mediated Decay*) se met en place afin d'éviter

la production de protéines toxiques. Lorsqu'un ARNm est arrivé en fin de vie, il est dégradé par plusieurs voies possibles comme par exemple un décoiffage, une perte de sa queue poly(A) ou encore par l'action de micro ARN régulateurs (ARNmi). Ainsi, la transcription, l'épissage et la traduction constituent les étapes clés de l'expression des gènes (Figure 1.6).

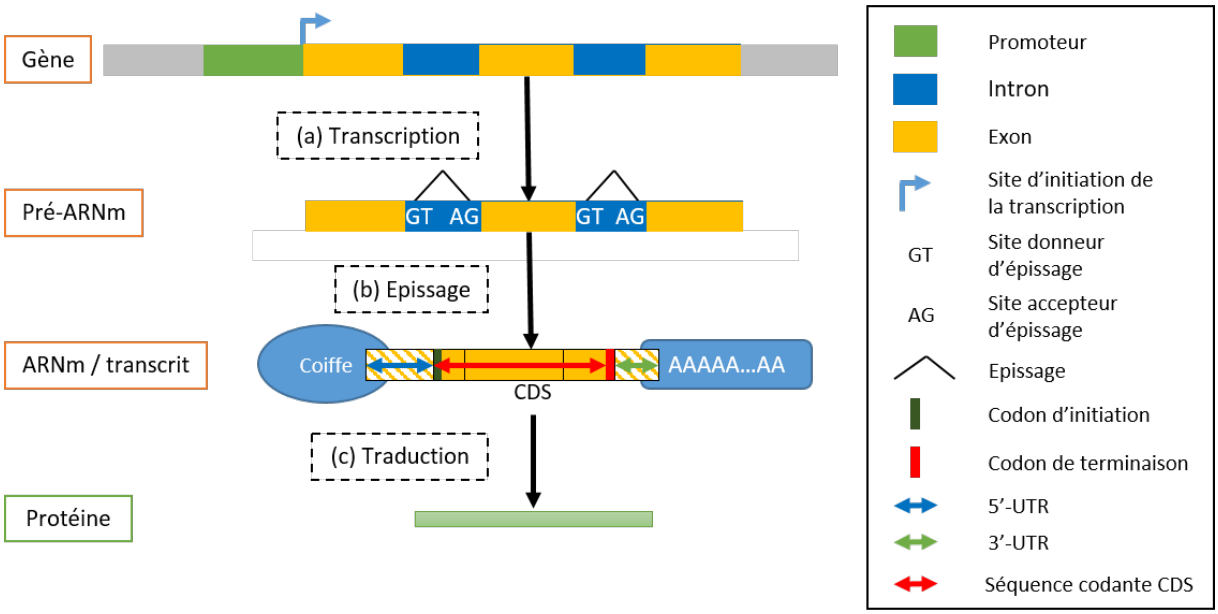


FIGURE 1.6 – Schéma de l'expression d'un gène. (a) Un gène est transcrit, à partir du site d'initiation de la transcription suite à l'activation du promoteur, en pré-ARNm par le mécanisme de la transcription. (b) L'épissage forme le transcrit mature, en concaténant les exons entre eux et en épissant les introns, grâce aux sites d'épissage. L'ARN est enfin stabilisé par l'ajout d'une coiffe en 5' et d'une queue poly(A) en 3'. (c) L'ARNm, mature, peut ensuite être traduit pour former une protéine.

Notons enfin que l'expression d'un gène consiste en la production de grandes quantités de ces copies que sont les ARNm. Par la suite, on parlera par simplicité d'un transcrit pour dénommer tout un ensemble de transcrits identiques traités simultanément dans une cellule.

1.1.3 Diversité des transcrits d'un gène eucaryote

1.1.3.1 Quantification de la diversité des transcrits

Chez l'humain, on estime qu'il y aurait plus de 20 000 gènes codants pour des protéines. On dénombre aussi près de 24 000 gènes non codants. Au total, l'humain exprime-

rait plus de 230 000 transcrits différents (Tableau 1.1) recensés dans la base de données Ensembl (Howe et al. 2021). D'après les données Gencode (Frankish et al. 2021), environ 90 000 de ces transcrits sont codants (*GENCODE - Human Release Statistics* 2021). D'après l'état de l'art, les observations expérimentales réalisées dénombrent actuellement en moyenne entre 5 et 7 transcrits par gène (Steijger et al. 2013 ; Tung et al. 2020). De ce fait, un gène ne donne majoritairement pas un seul mais plusieurs transcrits.

	Homo sapiens
Nombre de gènes codants	20 442
Nombre de gènes non codants	23 982
Nombre de transcrits connus	237 081

Table 1.1 – Nombre de gènes codants et de transcrits connus chez l'humain dans la base de données Ensembl en mars 2021.

Par exemple, le gène CREM humain possède 48 transcrits actuellement décrits dans la base de données publique Ensembl dont 34 codent pour des protéines et 14 sont non codants (Table 1.2).

1.1.3.2 Mécanismes de la diversité des transcrits

Plusieurs transcrits permettent donc la production de différentes protéines pour un même gène. Elles sont appelées protéines isoformes (H.-D. Li et al. 2021 ; Nilsen et Graveley 2010). Deux phénomènes sont à l'origine de cette diversité de transcrits : la *transcription alternative* et l'*épissage alternatif*.

Ces mécanismes sont extrêmement régulés pour permettre le bon fonctionnement de l'organisme. Par exemple, si des mutations sont présentes dans les séquences, un dysfonctionnement intervient pouvant conduire à des cancers ou à des maladies rares (Taneri et al. 2012).

Transcription alternative. Plusieurs promoteurs dits alternatifs peuvent coexister au sein de la séquence du gène et permettre la synthèse de plusieurs pré-ARNm. Dans la Figure 1.7(A), deux promoteurs et deux sites d'initiation de la transcription sont présents, pouvant initier la transcription de deux exons différents. Ces deux sites permettent de produire deux pré-ARNm différents. Notons que des promoteurs différents peuvent conduire à la formation d'ARNm possédant le même CDS, mais dont les régions UTR sont différentes. Au cours de la thèse, nous discuterons de ce dernier cas précis (voir section 4.4

Nom du transcrit	Identifiant *Ensembl* du transcrit	Taille en nucléotides	Taille protéique	Biotype	Identifiant CCDS
CREM-210	ENST00000374721.7	2020	269aa	Protein coding	-
CREM-217	ENST00000439705.5	2306	248aa	Protein coding	CCDS7182
CREM-204	ENST00000345491.7	2220	300aa	Protein coding	CCDS7180
CREM-208	ENST00000361599.8	2170	270aa	Protein coding	CCDS7185
CREM-212	ENST00000374728.7	1933	221aa	Protein coding	CCDS7183
CREM-207	ENST00000356917.9	1895	108aa	Protein coding	CCDS53518
CREM-205	ENST00000348787.6	1821	221aa	Protein coding	CCDS7183
CREM-203	ENST00000344351.5	1625	95aa	Protein coding	CCDS53522
CREM-206	ENST00000354759.7	1589	248aa	Protein coding	CCDS7182
CREM-202	ENST00000342105.7	1541	245aa	Protein coding	CCDS7186
CREM-218	ENST00000460270.5	1416	95aa	Protein coding	CCDS53522
CREM-211	ENST00000374726.7	1207	137aa	Protein coding	CCDS7184
CREM-228	ENST00000473940.5	1182	120aa	Protein coding	CCDS7187
CREM-237	ENST00000488328.5	1049	109aa	Protein coding	CCDS53519
CREM-229	ENST00000474362.5	1030	96aa	Protein coding	CCDS53523
CREM-220	ENST00000463314.5	964	138aa	Protein coding	CCDS58075
CREM-201	ENST00000337656.8	900	299aa	Protein coding	CCDS7181
CREM-214	ENST00000395887.7	849	282aa	Protein coding	CCDS58074
CREM-239	ENST00000489321.5	731	137aa	Protein coding	CCDS7184
CREM-213	ENST00000374734.7	711	236aa	Protein coding	CCDS31181
CREM-238	ENST00000488741.5	651	102aa	Protein coding	CCDS53521
CREM-230	ENST00000474931.5	626	112aa	Protein coding	CCDS7188
CREM-224	ENST00000468236.5	582	125aa	Protein coding	CCDS58076
CREM-242	ENST00000490511.1	469	113aa	Protein coding	CCDS53520
CREM-236	ENST00000487763.5	429	121aa	Protein coding	CCDS53517
CREM-216	ENST00000429130.7	1476	345aa	Protein coding	-
CREM-231	ENST00000479070.5	939	312aa	Protein coding	-
CREM-226	ENST00000469949.6	721	59aa	Protein coding	-
CREM-234	ENST00000484283.5	660	219aa	Protein coding	-
CREM-235	ENST00000487132.5	655	186aa	Protein coding	-
CREM-215	ENST00000427847.6	635	156aa	Protein coding	-
CREM-221	ENST00000463960.5	582	193aa	Protein coding	-
CREM-243	ENST00000494479.5	563	115aa	Protein coding	-
CREM-244	ENST00000495301.1	555	78aa	Protein coding	-
CREM-241	ENST00000490460.5	1074	62aa	Nonsense mediated decay	-
CREM-245	ENST00000495960.5	895	46aa	Nonsense mediated decay	-
CREM-222	ENST00000464475.1	453	53aa	Nonsense mediated decay	-
CREM-240	ENST00000489388.5	1022	Pas de protéine	Processed transcript	-
CREM-227	ENST00000472813.1	955	Pas de protéine	Processed transcript	-
CREM-225	ENST00000469517.1	910	Pas de protéine	Processed transcript	-
CREM-248	ENST00000497686.1	839	Pas de protéine	Processed transcript	-
CREM-219	ENST00000461968.5	747	Pas de protéine	Processed transcript	-
CREM-209	ENST00000374711.5	722	Pas de protéine	Processed transcript	-
CREM-233	ENST00000482646.5	715	Pas de protéine	Processed transcript	-
CREM-247	ENST00000496626.5	596	Pas de protéine	Processed transcript	-
CREM-246	ENST00000496019.5	563	Pas de protéine	Processed transcript	-
CREM-223	ENST00000466251.5	485	Pas de protéine	Processed transcript	-
CREM-232	ENST00000482633.5	415	Pas de protéine	Retained intron	-

TABLE 1.2 – Ensemble des transcrits du gène CREM humain décrits en 2021 par la base de données Ensembl. Les biotypes sont définis : "*Protein coding*" correspond à un transcrit codant pour une protéine, "*Nonsens mediated decay*" correspond à un transcrit ayant un codon *stop* prématuré, "*Processed transcript*" et "*Retained intron*" correspondent à des transcrits qui ne sont pas codants.

page 110).

Epissage alternatif. Le mécanisme d'épissage alternatif a été décrit pour la première fois en 1978 (GILBERT 1978). Ce phénomène concerne les gènes multi-exoniques (qui

contiennent plusieurs exons). Le spliceosome peut épisser, c'est-à-dire sélectionner des introns différents à partir d'un même pré-ARNm. Cette sélection crée des combinaisons d'exons différentes et conduit à la formation de transcrits matures différents (Figure 1.7(B)). De manière générale, on estime que dans environ 90% des cas l'expression des gènes des mammifères (Bao et al. 2019) et dans 95% des cas l'expression des gènes humains (Pan et al. 2008) implique l'épissage alternatif.

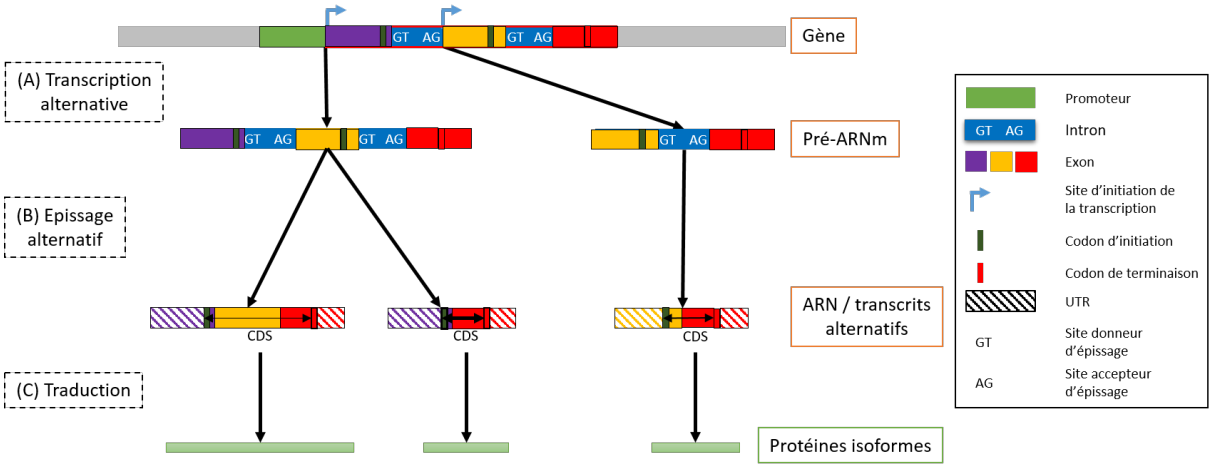


Figure 1.7 – Mécanismes de l'expression alternative des gènes. Un gène exprime des transcrits différents à partir du mécanisme de transcription alternative (A) et d'épissage alternatif (B). Ces transcrits peuvent être traduit en protéines (C) si ils correspondent à des transcrits codants. Le gène possède deux sites d'initiation de la transcription alternatifs (flèches bleues), et deux types de pré-ARNm peuvent être transcrits. L'épissage alternatif du premier peut retenir ou non l'exon jaune. Les trois ARNm productibles incluent trois CDS différents encodant trois protéines isoformes. Par simplicité, la figure représente les régions du gène comme ayant une seule implication future, promotrice, exonique, intronique, codante, UTR. Cependant et du fait de la combinatoire générée, les régions du gène peuvent avoir plusieurs implications : la même région du gène illustré va être transcrite puis épissée (premier exon, en violet) ou bien ne pas être transcrite (promoteur alternatif, seconde flèche bleue) : cette région est selon le contexte un intron ou un promoteur. De plus, un même exon codant pourra dans certain contexte ne pas être traduit, ou être partiellement traduit (exon jaune).

Evènements de l'épissage alternatif. Il existe plusieurs types d'évènements d'épissage alternatifs responsables de la diversité des transcrits : le saut d'exon, la rétention d'intron, le site d'épissage 3' alternatif, le site d'épissage 5' alternatif ou encore les exons mutuellement exclusifs (Figure 1.8). Le *saut d'exon* dénomme la situation où un exon est

excisé avec ses introns flanquants. Il constitue le type d'évènement le plus répandu chez les vertébrés parmi les évènements d'épissage alternatifs (Y. Wang et al. 2015 ; Sammeth et al. 2008). La *rétention d'introns* intervient lorsqu'un intron n'est pas épissé et va être conservé dans le transcrit mature. C'est un mécanisme qui intervient principalement dans les régions UTR chez l'humain (Alexandre Favoretto Galante et al. 2004) et est l'évènement d'épissage alternatif majoritaire chez les plantes (Chaudhary et al. 2019). Les *sites d'épissage 3' et 5' alternatifs* vont sélectionner différemment l'extrémité d'un intron résultant en des variations de la taille des exons qui vont potentiellement changer la séquence codante. Les *exons mutuellement exclusifs* désignent le fait que plusieurs choix d'exons sont possibles et que seulement l'un de ces choix est utilisé, à l'exclusion des autres. Ces types d'évènements sont illustrés avec la Figure 1.8. L'*épissage constitutif* correspond enfin au cas où tous les exons du pré-ARN sont retenus par l'épissage. De même un exon dit constitutif est présent dans l'intégralité des transcrits alternatifs d'un gène.

Notons enfin qu'avec l'expression alternative, un même segment de gène peut avoir plusieurs rôles : transcrit ou non transcrit, exonique ou intronique, codant ou non codant (Figure 1.7). De plus, un même transcrit codant peut parfois être traduit de différentes manières, conduisant à des protéines différentes. On parle alors de *traduction alternative*. Au total, l'expression alternative d'un gène résulte de ces trois mécanismes que sont la transcription alternative, l'épissage alternatif et la traduction alternative.

1.2 Données bioinformatiques concernant les gènes et les transcrits

Le traitement automatique des données génétiques repose sur l'organisation de bases de connaissances, de définitions et de représentations des objets biologiques et de formats standards décrits dans cette partie.

1.2.1 Données publiques, annotation et génome de référence

L'ensemble des connaissances disponibles sur les gènes et les transcrits sont stockées dans des bases de données publiques telles que Ensembl (Howe et al. 2021) pour les vertébrés, la base de données de l'UCSC (*University of California, Santa Cruz*) (Navarro Gonzalez et al. 2021) pour près de 100 organismes ou encore CCDS (*Consensus coding*

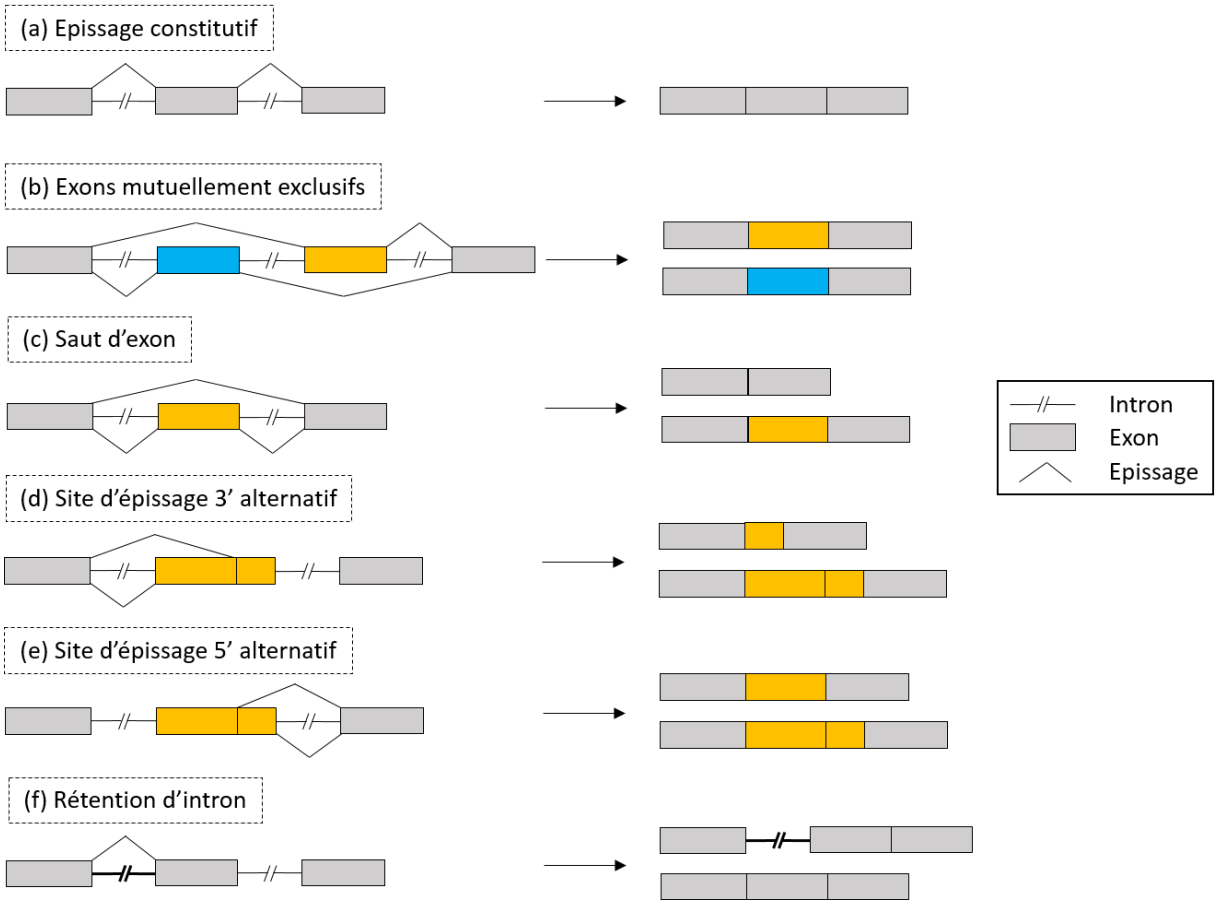


FIGURE 1.8 – Types d'évènements de l'épissage alternatif. Chaque évènement d'épissage alternatif (à gauche) est réalisé sur les pré-ARNm et participe à la synthèse d'ARN messagers matures (à droite). A gauche on schématise une population de pré-ARNm dont une partie subit un épissage particulier, et l'autre partie un autre épissage.

sequence) (PUJAR et al. 2018) spécialisée pour l'humain et la souris et nettoyée manuellement. Nous utiliserons ces bases dans la thèse.

Ces données sont basées sur un *génome de référence*. Chaque organisme séquencé dispose d'un génome de référence qui lui est propre. Un génome de référence correspond au génome d'un individu de l'espèce considéré comme référence pour l'espèce. Un génome de référence correspond ainsi à une version d'un génome issue d'un assemblage qui est susceptible d'évoluer avec de nouvelles versions d'assemblage. En *Janvier 2018*, le génome de référence humain est hg38 (GRCh38), celui de la souris est mm10 (GRCm38) et celui du chien est canfam3.1. En *Juillet 2021*, le génome de référence de la souris a été modifié et est passé à la version mm39 (GRCm39).

Un génome de référence n'est pas forcément représentatif de tous les individus. En

plus de comporter potentiellement des erreurs, un génome de référence n'est pas forcément complet et nécessite de nouveaux assemblages. Si on prend l'exemple de l'humain, son génome est divisé en 23 paires de chromosomes qui ont été séquencés en 2003 (Craig Venter et al. 2001 ; Collins et al. 2003) mais pas entièrement. C'est en 2020 que le chromosome X humain a été entièrement séquencé en considérant les régions télomériques (extrémité d'un chromosome contenant de nombreuses répétitions, impliquées dans le processus de vieillissement cellulaire) et les régions centromériques (jointure des chromatides) non assemblées précédemment (Miga et al. 2020).

Une fois assemblé, le génome doit être analysé afin de décrire les séquences qui composent sa structure telles que les gènes, les régions intergéniques, les régions régulatrices, les fonctions associées. Cela fait partie de ce qu'on nomme l'*annotation*. Ainsi, chaque élément caractérisé sera localisé à un endroit précis, le *locus*, par des coordonnées sur le génome de référence.

1.2.2 Formats des données

Les données décrivant les gènes et les transcrits sont généralement distribuées sous des formats spécifiques. On retrouve notamment les formats FASTA et GTF. Le format *FASTA* est un format de fichier texte qui contient en plus de la séquence décrite une information sur la séquences. Il est composé de deux lignes, la première commence par le symbole ">" et est généralement suivi de l'identifiant de la séquence (nom du gène, nom du transcrit, nom de la protéine) et de commentaires (non obligatoires). La seconde ligne correspond à la séquence de nucléotides pour un gène ou un transcrit, ou encore d'acides aminés pour une protéine. Par exemple (Figure 1.9), la première ligne du fichier FASTA du gène CREM humain contient l'identifiant Ensembl, l'espèce, le chromosome où est situé le gène, la version du génome de référence, les positions du gène sur ce chromosome et le sens de lecture du gène.

```
> ENSG00000095794 homo_sapiens 10 GRCh38 35126791 35212958 1
CTAGGCCCCGCCCCCTGACCCGCACCTTCCCGCCGCCCCTCCCCGGTTCCATTTCATTGT
TGGATTGTGGCGCTTCACTCCTGCTGGCGGCCGGCAGGGGGCGGAGTTCGAGCCTGGATT
TTTTTCCTCGGGGCCTCCCCCGGGAGGCCGTCCCGGCGTGGGGGAGGGGAGGACGGGGCG
```

Figure 1.9 – Extrait d'un fichier FASTA pour le gène CREM humain.

Le Format *GTF* (*General Transfer Format*) ou *GFF* (*General Feature Format*) est un format de fichier qui contient l'ensemble des informations concernant des objets (des

gènes, des transcrits) et leurs compositions (les exons, les codons *start*, les codons *stop*, les régions non traduites par exemple) en présentant leur position génomique au sein du génome de référence. Chaque ligne représente la description de l'objet. Chaque description est organisée en 9 colonnes séparées par des tabulations. Par exemple, la Figure 1.10 présente un extrait d'information du gène CREM humain au format GTF.

1. le chromosome considéré ;
2. la source de l'information (nom de la base de données, de la méthode d'origine, etc.) ;
3. l'objet concerné (gène, transcrit, exon, CDS, codon *start*, *stop*, etc.) ;
4. la position de début de l'objet sur le génome de référence ;
5. la position de fin de l'objet sur le génome de référence ;
6. le score de l'objet (facultatif) ;
7. le brin d'ADN pour le sens de lecture de l'information : "+" pour le sens $5' \rightarrow 3'$ (sens), "-" pour le sens $3' \rightarrow 5'$ (antisens) ;
8. le cadre de lecture considéré (facultatif) : "0", "1" ou "2" où "0" indique que la première base de la séquence est la première base du codon, 1 et 2 pour indiquer que la deuxième ou la troisième base de la séquence respectivement sont la première base du codon ;
9. les attributs de l'objet (informations complémentaires).

```
10    Ensembl    gene          35126791    35212958    .   +   .   gene_id "ENSG00000095794"; transcript_nb "21";
10    Ensembl    transcript    35126923    35179847    .   +   .   gene_id "ENSG00000095794"; transcript_id "ENST00000374726"; exon_nb "3";
10    Ensembl    exon          35148372    35148491    .   +   .   gene_id "ENSG00000095794"; transcript_id "ENST00000374726"; exon_number "1";
10    Ensembl    exon          35178889    35178986    .   +   .   gene_id "ENSG00000095794"; transcript_id "ENST00000374726"; exon_number "2";
10    Ensembl    exon          35179134    35179329    .   +   .   gene_id "ENSG00000095794"; transcript_id "ENST00000374726"; exon_number "3";
```

FIGURE 1.10 – Extrait simplifié d'un fichier GTF pour le gène CREM humain.

1.2.3 Représentation des données

Il existe différentes façons de représenter visuellement les données à partir des données génomiques. Les plus communes sont celles présentées dans les visualiseurs de génome (*Genome browser*) comme celui d'Ensembl ou celui de l'UCSC. Le premier représente chaque exon de transcrit par rapport au gène en utilisant des blocs pleins pour les exons ou les parties d'exons codants et des blocs vides pour les exons ou les parties d'exons non

codants (UTR) (Figure 1.11a). Un exemple de cette représentation pour une partie des transcrits du gène CREM est présenté dans la Figure 1.11b. Le visualiseur de génome de l'UCSC adopte une représentation similaire. Les zones épissées y sont représentées par un trait avec une orientation indiquant le sens de lecture du gène ($5' \rightarrow 3'$) par rapport au brin considéré. Les exons ou les parties d'exons non codants (UTR) sont présentés comme des blocs pleins à l'échelle et donc de taille réduite comparée à celle des introns (1.11c). Dans la Figure 1.11d, les transcrits du gènes CREM présentés sont orientés par rapport à leur brin d'ADN lu.

Il existe également une représentation compacte de l'ensemble des transcrits issus d'un gène, sous forme de graphes. Ces graphes sont appelés *graphes d'épissage* (*splicing graph*, Heber et al. 2002). Chaque nœud correspond à un exon et chaque arête correspond aux liaisons possibles entre les exons, les jonctions entre les exons : les *jonctions d'exons*. Par exemple, dans la Figure 1.12, le graphe d'épissage permet de visualiser les compositions possibles en exons des transcrits. Le graphe montre qu'on peut relier l'exon gris à l'exon jaune pour former un transcrit et relier l'exon gris à l'exon bleu puis l'exon bleu à l'exon jaune pour former un second transcrit. Les graphes d'épissage permettent ainsi de visualiser les combinaisons possibles d'exons sans pour autant afficher les transcrits possibles. Les graphes d'épissage sont utilisés dans de nombreuses méthodes. Par exemple, la méthode ASGAL (*Alternative Splicing Graph ALigner, Denti et al. 2018*) qui vise à aligner des lectures RNA-seq sur un graphe d'épissage afin de détecter et de prédire de nouveaux évènements d'épissage. Elle ne permet pas de faire des prédictions directes de transcrits.

1.2.4 Séquençage de l'ADN et de l'ARN

1.2.4.1 Lectures courtes

Pour tenter d'observer les transcrits exprimés par un gène, de nombreuses méthodes ont été mises en place. Récemment, des technologies de séquençage à haut débit, dites de seconde génération, se sont développées afin de décoder les séquences d'ADN.

Elles sont basées sur des fragments de séquences nommés lectures (*reads*) et présentent moins de 0,1% de taux d'erreur. Des millions de lectures sont ainsi séquencées en parallèle pour chaque échantillon, ce nombre dépendant de la profondeur (nombre de fois qu'un nucléotide est séquencé) et de la couverture (proportion de la séquence que l'on a réussi à séquencer et à assembler) de séquençage. Il existe différentes technologies (Roche 454, ABI SOLiD, Illumina Solexa). La technologie Illumina qui est la plus répandue (Bronner

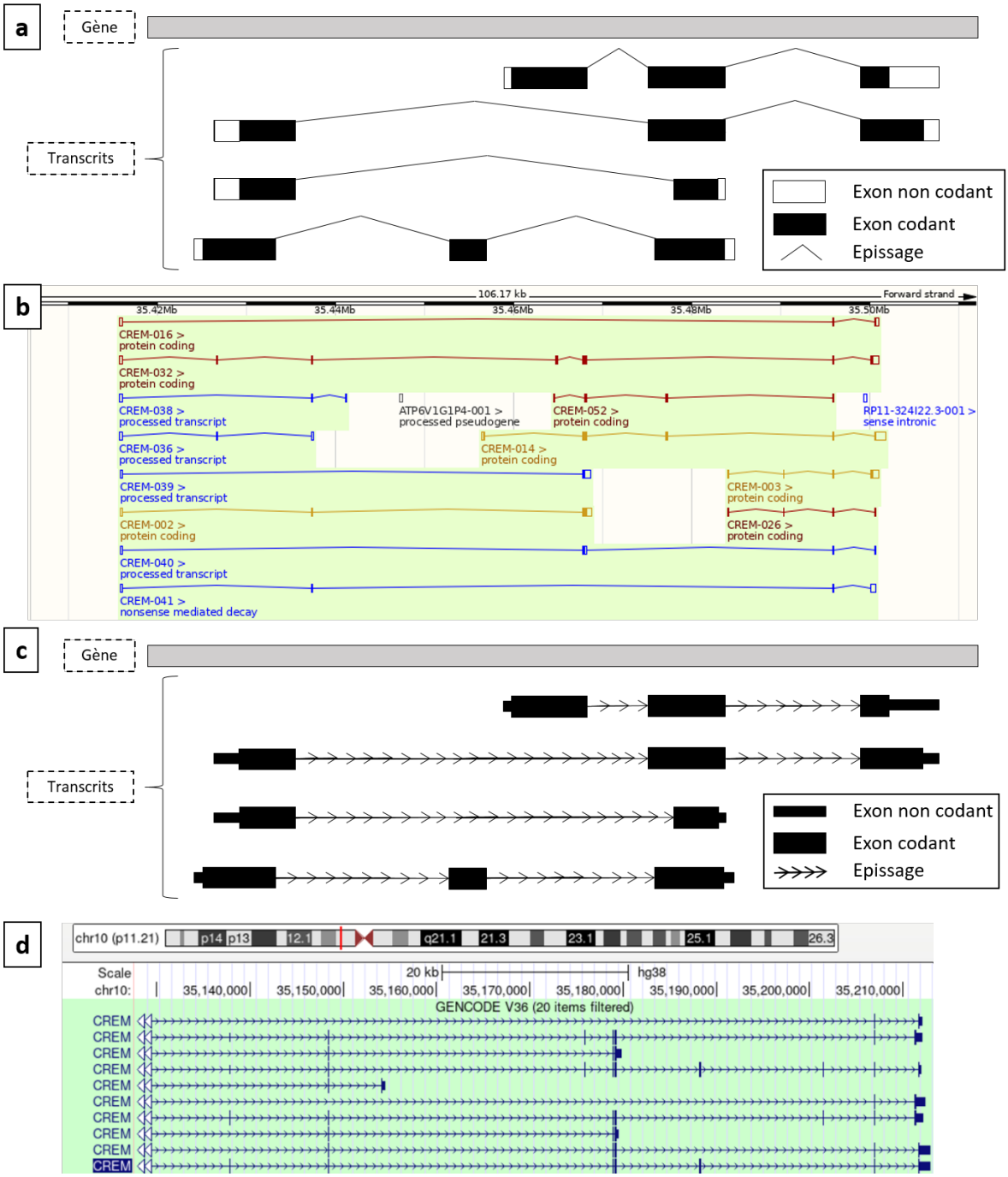


FIGURE 1.11 – Représentation des transcrits dans les visualiseurs de génomes. a) Représentation schématique dans le visualiseur d'Ensembl. b) Exemple d'une parties des transcrits du gène CREM dans le visualiseur d'Ensembl. c) Représentation schématique dans le visualiseur de l'UCSC. d) Exemple d'une partie des transcrits du gène CREM dans le visualieur de l'UCSC.

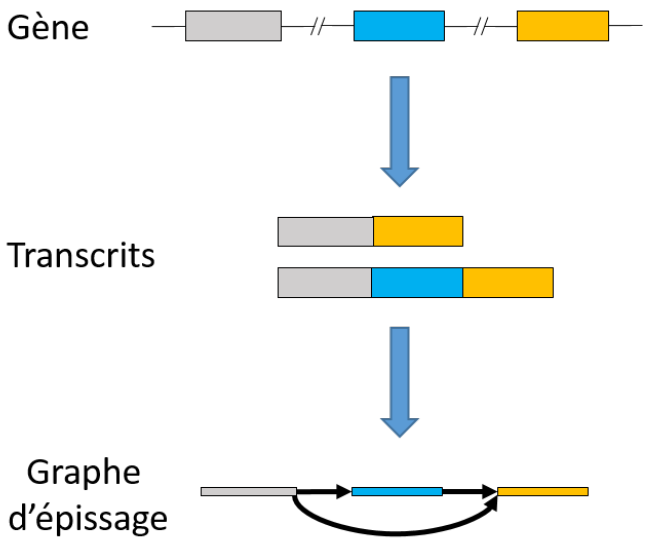


FIGURE 1.12 – Construction d'un graphe d'épissage (adaptée de HEBER et al. 2002). Un gène exprime des transcrits selon des combinaisons d'exons produites par l'épissage alternatif. Chaque exon définit un nœud du graphe et chaque arête représente les jonctions d'exons connues. Un transcrit complet est un chemin dans le graphe.

et al. 2014).

En plus de séquencer l'ADN, ces méthodes permettent également de séquencer l'ARN d'un échantillon et caractériser ainsi un transcriptome en identifiant et en quantifiant les transcrits exprimés dans cet échantillon. On parle de méthodes de *RNA-seq*. Elles permettent également d'étudier les régulations post-transcriptionnelles des ARN (maturation, épissage, évaluation du taux de traduction, épigénétique).

Les limites du RNA-seq avec des lectures courtes résultent de la profondeur de séquençage puisqu'on ne détectera l'ARN exprimé qu'à partir d'un certain seuil malgré les procédures d'amplification. On peut aussi évoquer la limite résultant de la taille des lectures, entre 30 et 150 nucléotides, qui ne couvre pas l'intégralité de l'ARN et ne donne pas d'information directe sur les transcrits complets longs de centaines, voir de milliers de nucléotides.

1.2.4.2 Lectures longues

En parallèle, des méthodes de troisième génération se développent en se basant sur de longues lectures (jusqu'à plusieurs centaines de milliers de nucléotides) dans le but de séquencer notamment de grands fragments d'ADN et des transcrits dans leur intégralité.

Deux technologies sont mises en avant : Pacific Biosciences (PacBio) et Oxford Nanopores Technologies (ONT). Seulement, cette nouvelle technologie comporte un taux d'erreur plus conséquent en terme d'insertions et de délétions avec environ 10% à l'heure actuelle (Sessegolo et al. 2019) bien que des méthodes de correction tentent de corriger ces erreurs (Marchet et al. 2020).

1.2.4.3 Autres technologies de séquençage ciblées

Il existe d'autres méthodes de séquençage basées sur des techniques de PCR (Réaction de polymérisation en chaîne) dont le but est d'amplifier les molécules d'ADN ou d'ARN en utilisant des amorces, c'est-à-dire un fragment complémentaire de la séquence ciblée. Par exemple la RACE-PCR (amplification rapide des extrémités d'ADNc) est une méthode basée sur la PCR dont le but est d'obtenir la séquence complète des ARN, c'est-à-dire obtenir les séquences codantes et non codantes (UTR) des transcrits codants notamment mais aussi les transcrits non codants comme les ARNlnc. La méthode produit une copie complète de l'ARN en ADNc (ADN complémentaire) qui est ensuite amplifiée et séquencée.

De plus, comme évoqué précédemment, ces technologies séquencent le transcriptome à un instant donné, et dans une condition donnée et les quantifient avec un seuil de profondeur de détection. Or, tous les transcrits ne s'expriment pas au même moment, dans les mêmes conditions et dans les mêmes proportions. En 2013, on estimait que 21% du transcriptome humain complet pouvait être connu de ces façons (Steijger et al. 2013).

Ces technologies de séquençages ciblés sont très utiles pour valider des hypothèses concernant l'expression d'un ARN et obtenir des transcrits complets avec peu d'erreur. Cependant, elles requièrent des a priori (les amorces) sur les transcrits recherchés, et ne sont pas automatisables à très hauts débits.

1.3 Travaux sur l'identification et la comparaison de transcriptomes

Dans cette thèse, nous nous intéressons à la problématique de l'identification du transcriptome d'un organisme, c'est-à-dire à caractériser quels sont tous les transcrits qu'un organisme est capable de produire. Par la suite, nous parlerons de "*transcriptome d'un gène*" dans cette thèse, pour parler de l'ensemble des transcrits qu'un gène est capable d'exprimer. Pour cela, nous nous basons sur de la comparaison de structures de gènes à une échelle multi-espèces afin de prédire des transcrits et d'identifier les isoformes orthologues d'un gène. Nous cherchons également à constituer un jeu de données de gènes dont la structure est conservée chez plusieurs espèces et pour lesquels nous identifions les isoformes orthologues et vérifions la conservation des transcriptomes. Dans cette section, nous présentons les méthodes existantes permettant d'aborder ces questions.

1.3.1 Observer les transcrits

Étudier les transcrits permet de comprendre les processus moléculaires responsables du développement normal d'un organisme. Cela peut aussi permettre de comprendre les problèmes qui peuvent entraver son bon fonctionnement et pouvant conduire à des maladies rares par exemple. Pour cela, il est nécessaire de pouvoir identifier fonctionnellement les mutations et d'explorer la diversité des transcriptomes dans différents tissus ou chez différents individus.

Enfin, tous les transcrits qu'un gène peut produire ne sont pas exprimés au même moment ou dans le même tissu de l'organisme, et on peut chercher à déterminer les conditions physiologiques de leur expression.

1.3.2 Méthodes de prédiction de transcrits

Si les méthodes de séquençage sont aujourd'hui les méthodes les plus utilisées pour étudier l'expression des gènes, elles ne sont donc cependant pas exhaustives. Pour compléter ces approches, il est utile de comprendre quels sont les transcrits qu'un gène peut exprimer et de pouvoir évaluer si l'environnement permet son expression. C'est à ce premier point, l'identification des transcrits réalisables par un gène, que nous nous sommes intéressés. Comme dit précédemment, une limitation des méthodes de séquençage est le fait que l'on séquence un objet biologique (tissu, cellule) à un moment donné et dans une

condition donnée. Seulement, les transcrits ne s'expriment pas tous en même temps, ni dans les mêmes proportions, ni dans les mêmes conditions. C'est pourquoi des méthodes complémentaires sont nécessaires afin de pouvoir découvrir l'ensemble du transcriptome d'un organisme.

Une première approche pourrait être de partir de la séquence brute d'un gène et de retrouver sa structure intron/exon en utilisant les signaux de sites d'épissage (les donneurs et accepteurs d'épissage). Comme présenté dans la section 1.1.2.2 (page 21), ces signaux ne sont pas toujours identiques aux dinucléotides "GT" et "AG", qui constituent le signal le plus fort de motifs très dégénérés. Partir de la séquence brute et rechercher de façon *ex nihilo* des signaux dinucléotidiques conduirait à une explosion combinatoire des possibilités et rend cette approche peu réaliste. Sur CREM, le gène contient 5 945 occurrences de "AG" et 4 865 de "GT" et des millions d'introns candidats pourraient ainsi être proposés alors que les 48 transcrits connus mettent en évidence seulement quelques dizaines d'introns.

Une autre approche pourrait être de se baser sur la connaissance actuelle en terme d'exons pour prédire les combinaisons possibles pour former des transcrits candidats. Si on prend l'exemple du gène CREM humain qui comporte 20 exons connus, il en résulte $2^{20} - 1$ combinaisons d'exons possibles soit 1 048 575 transcrits possibles. Cette explosion combinatoire ne reflète pas la réalité du transcriptome, car la majorité des exons sont constitutifs (présents dans tous les transcrits alternatifs) et ne participent donc pas à la combinatoire effectivement générée. En effet, seuls 48 transcrits sont observés (Table 1.2). Ainsi, toutes les propositions de combinaisons d'exons ne conduisent pas à un transcrit mature. On peut, de plus, prendre en considération les différentes contraintes structurelles, notamment concernant le CDS dans le cas des transcrits codants (codons *start* suivit d'un nombre entier de codons sans codon *stop* intermédiaire) pour éliminer des transcrits non réalisables, mais cela ne suffit pas à réduire complètement l'explosion combinatoire.

Se baser sur les connaissances actuelles semble alors être une démarche envisageable pour tenter de compléter les données et c'est sur ce principe que s'appuient de nombreuses méthodes. Ces méthodes se basent principalement sur la recherche d'évènements liés à l'épissage alternatif ou encore sur la conservation des exons entre espèces, ce qui concerne généralement peu d'espèces (Muller et al. 2021). De plus, en l'état actuel de nos connaissances, aucune de ces méthodes ne s'applique à reconstituer des transcrits entiers. Dans les sections suivantes, quatre méthodes sont illustrées et leurs principes : CRAC (Philippe et al. 2013), SplicedFamAlign (SFA, Jammali et al. 2019), ThorAxe (Zea et al. 2021) et CG-alcode (Blanquart et al. 2016). Ces méthodes abordent la pro-

blématique de l'étude des transcriptomes selon les questions suivantes : 1) Quels sont les évènements d'épissage observés ? 2) Comment définir des transcrits similaires entre plusieurs espèces ? 3) Comment aligner les transcrits observés sur le génome d'une espèce ? La dernière méthode citée, CG-alcode, est le socle sur laquelle repose les travaux de cette thèse.

1.3.3 Identifier les évènements d'épissage, CRAC

Un certain nombre de méthodes prennent en entrée des lectures issues de séquençages et avec le but de reconstruire une annotation de la structure des transcrits à partir d'un génome de référence. Parmi ces méthodes, il existe des méthodes d'alignement standard en RNA-seq, comme CRAC (Philippe et al. 2013), BWA (H. Li et Durbin 2009 ; H. Li et Durbin 2010), SOAP2 (R. Li et al. 2008), Bowtie (Langmead et al. 2009), GASSST (Rizk et Lavenier 2010) mais aussi des méthodes de prédiction de jonctions d'épissage, GSNAP (Wu et Nacu 2010), TopHat (Trapnell et al. 2009), MapSplice (K. Wang et al. 2010), TopHat-fusion (Kim et Salzberg 2011).

Ces méthodes intègrent des données de lectures RNA-seq pour déterminer les évènements d'épissage. Elles permettent également d'étudier d'autres caractéristiques comme des mutations (SNP), des insertions ou des délétions (indel) ou des substitutions (polymorphisme), comme le fait par exemple CRAC. Ces méthodes requièrent un génome de référence (voir section 1.2.1 page 27) et une collection de lectures, indépendemment de l'annotation du génome. Chaque lecture est subdivisée en sous-séquences de longueur k, appelées *k-mers*. CRAC considère une longueur de $k = 22$ nucléotides soit une probabilité de 10^{-4} d'avoir un alignement à une mauvaise position sur le génome humain de référence (Philippe et al. 2013). Ces *k-mers* sont ensuite alignés sur le génome de référence pour déterminer leur emplacement. Une suite de *k-mers* alignés de manière consécutive sur le génome indique l'emplacement d'un exon (Figure 1.13).

L'alignement des *k-mers* sur le génome de référence permet de plus de déterminer des zone de "cassure", où se situe une mutation par exemple. En effet, une suite de sites sans *k-mers* alignés correspond à une cassure, par exemple correspondant à k sites pour un SNP, lequel se situe après le dernier *k-mer* aligné (Figure 1.14).

CRAC permet d'identifier les évènements d'épissage présents dans les transcrits séquencés et leurs jonctions d'exons. Une cassure correspond à un intron, soit aux sites du génome de référence sans *k-mers* alignés et aux $k - 1$ dernières positions de l'exon précédant l'intron (Figure 1.13). Les *k-mers* chevauchant la jonctions d'exons ne peuvent

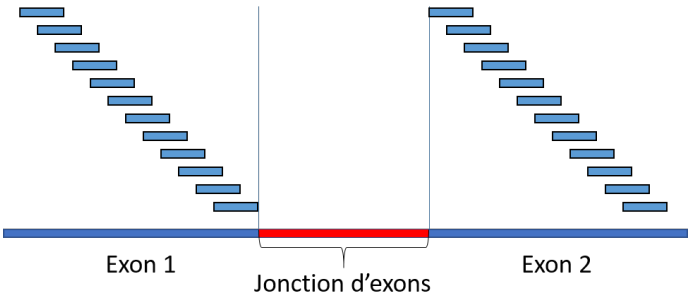


FIGURE 1.13 – Alignement des *k-mers* sur le génome. Les *k-mers* s'alignent sur les zones transcrites en ARNm. Ils ne s'alignent pas dans les zones introniques. La cassure obtenue délimitent la jonction entre les deux exons.

pas être alignés sur le génome.

Il existe d'autres méthodes basées sur des *k-mers* comme CRAC mais qui n'utilisent pas de génome de référence. Ce type de méthode basée sur des *k-mers* est généralisée sous la forme de graphe de De Bruijn où un nœud est un *k-mer* et une arête est un chevauchement de *k-mers* sur $k-1$ nucléotides.

1.3.4 Aligner des transcrits par programmation dynamique

Ces méthodes utilisent en entrée des données d'annotation de transcrits et un génome de référence. Leur but est de reconstituer la structure du gène à partir de ces données d'annotation par programmation dynamique. L'alignement réalisé correspond à un *alignement épissé* (*spliced alignment*), c'est-à-dire, qu'on aligne la séquence épissée du CDS d'un transcrit partiel ou complet sur une séquence génomique. Il s'agit donc de retrouver les positions des exons du transcrit sur le génome.

Le *problème de l'alignement épissé* (*spliced alignment problem*, SAP) a été décrit et caractérisé en trois classes (JAMMALI et al. 2019, Figure 1.15), SAP1, SAP2 et SAP3 :

— Le SAP 1 considère l'identification d'une zone ciblée sur le génome de la séquence du CDS d'un transcrit en recherchant par similarité de séquence l'alignement épissé. Seuls les alignements de séquences exoniques sont optimisés. Le meilleur alignement est celui qui aura le score le plus élevé de similarité. Exemple de méthodes SAP1 : BLAT (KENT 2002), SOAPsplice (HUANG et al. 2011), HSA (BU et al. 2013).

— Le SAP 2 inclut également la détection des signaux d'épissage correspondant aux dinucleotides "GT" et "AG" délimitant dans la grande majorité des cas les extrémi-

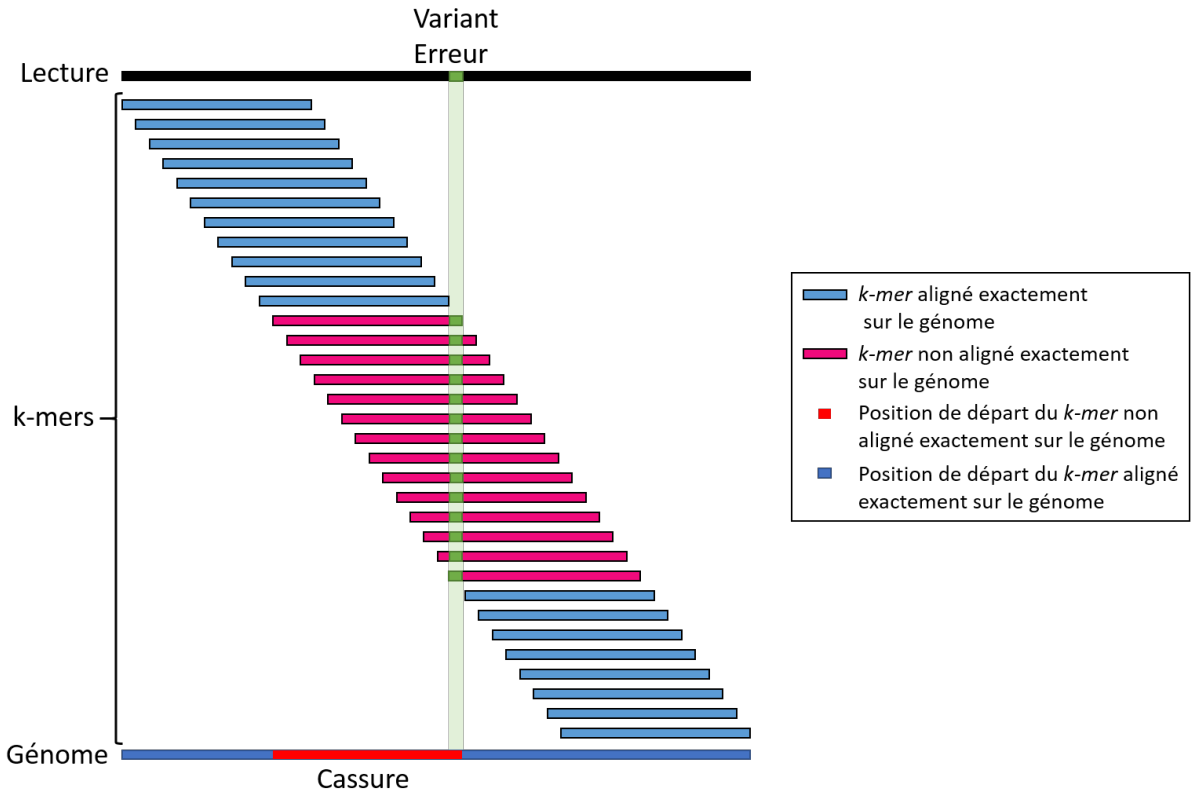


FIGURE 1.14 – Représentation de la méthode CRAC sur l'alignement des *k-mers* en présence d'une variation de type SNP (*single nucleotide polymorphism* (adaptée de PHILIPPE et al. 2013). Sur la figure, les *k-mers* ont une longueur $k = 22$ nucléotides. 14 *k-mers* portent la variation et ne s'alignent pas parfaitement avec le génome de référence, induisant la zone de coupure des alignements des *k-mers* sur 14 positions consécutives du génome. Les *k-mers* adjacents à la coupure sont alignés à l'identique à des positions consécutives.

tés des introns pour déduire les limites des exons. L'alignement de meilleur score obtiendra la meilleure similarité de séquence entre les exons et le génome et délimitera des introns potentiels par ces signaux d'épissage. Exemple de méthodes SAP2 : Splign (KAPUSTIN et al. 2008), MGAlign (LEE et al. 2003), GMAP (WU et WATANABE 2005).

— Le SAP 3 intègre de plus la structure d'épissage connue des séquences du gène et du transcrit en entrée pour maximiser le score. Il prend en compte les sites d'épissage des transcrits connus et les jonctions d'exons connues dans les annotations en plus de la similarité de séquence et des signaux d'épissage pour obtenir le meilleur alignement. SplicedFamAlign (SFA, JAMMALI et al. 2019) est la première implé-

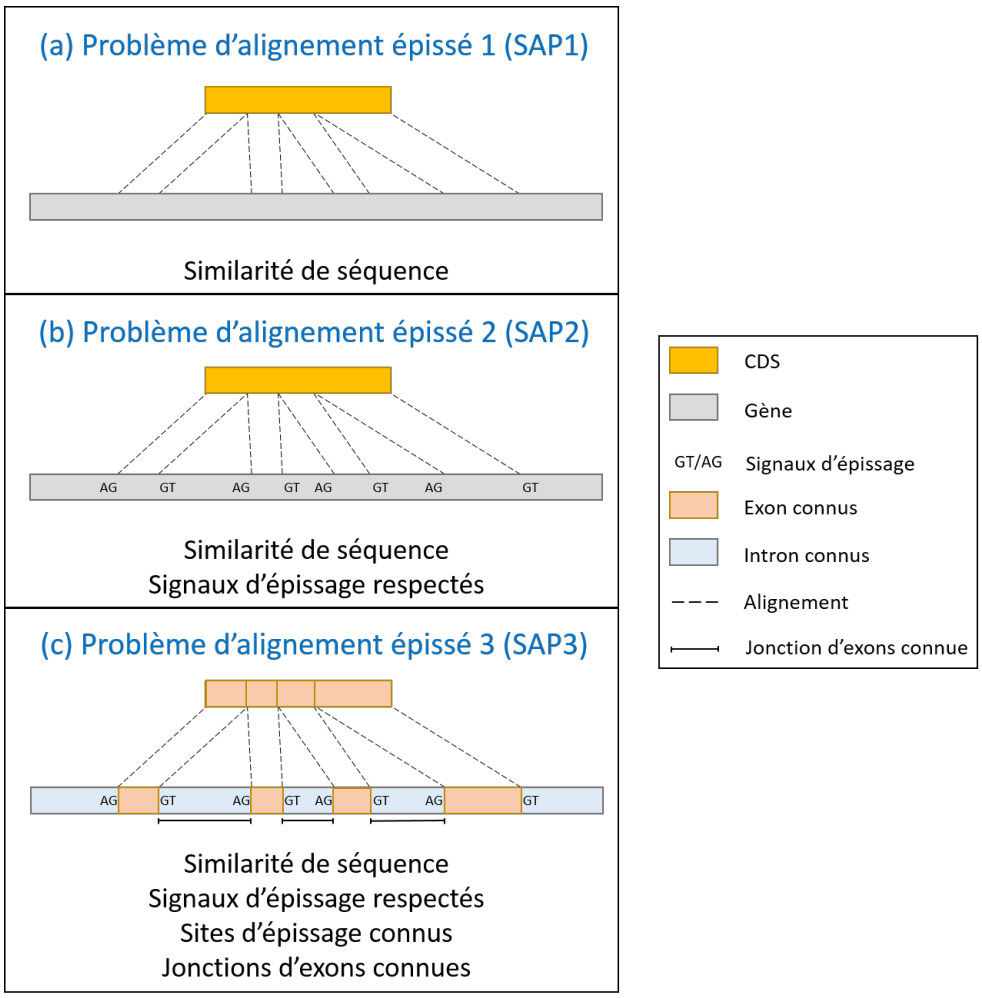


FIGURE 1.15 – Détails des trois problèmes d'alignement épissé (SAP) d'après les descriptions de JAMMALI et al. 2019. SAP3 garantit un bon alignement car il atteste de la similarité de séquence (SAP1), du respect des signaux d'épissage (SAP2) et du fait que certains des signaux sont connus et que les jonctions d'exons sont connues (SAP3).

mentation d'un algorithme SAP3 et les auteurs montrent que l'approche identifie les alignements épissés mieux que les méthodes SAP1 et SAP2.

Des méthodes telles que SFA permettent de déterminer de nouveaux transcrits avec une précision plus importante en considérant ce qui est déjà connu.

1.3.5 Aligner des transcrits chez plusieurs espèces, SplicedFamAlign

On peut ici distinguer deux cas d'utilisation des méthodes d'alignements épissés. Dans le premier cas, on aligne la séquence d'ARN/ADNc d'une espèce sur le génome de référence de cette même espèce. Dans le second cas, on aligne la séquence d'une espèce sur le génome d'une autre espèce. Si un transcrit d'une espèce est détecté sur le génome de référence d'une autre espèce, alors ce transcrit possède une séquence dont la structure d'épissage est homologue à celle détectée dans le génome de l'autre espèce. Cette homologie de séquences et de structure peut amener à désigner un couple de transcrits orthologues chez les deux espèces.

SplicedFamAlign (Jammali et al. 2019), présenté précédemment, a également un module permettant l'analyse de l'orthologie des structures d'épissage, c'est-à-dire d'étudier un même gène chez différentes espèces (voir section 1.4.3 page 50). En plus d'identifier la structure des gènes en réalisant des alignements épissés, entre le génome et les transcrits d'une même espèce, SFA identifie des CDS orthologues partageant des structures épissées homologues. Deux CDS provenant de deux espèces différentes sont jugés orthologues si ils ont les mêmes exons avec préservation de la colinéarité des exons (Figure 1.16). La méthodes permet ainsi de construire des ensembles de CDS orthologues partagés entre plusieurs espèces.

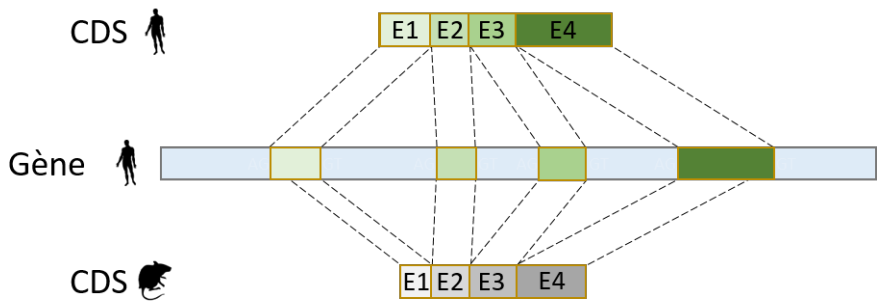


Figure 1.16 – SplicedFamAlign : identification de CDS orthologues (adaptée de Jammali et al. 2019). Deux CDS sont alignés sur un même génome au même *locus*. Si chaque exon des CDS est aligné avec le même *locus* sur le gène, alors les deux CDS sont estimés orthologues : ils partagent la même structure épissée. Ainsi des CDS orthologues présentent des exons orthologues, en copie unique et colinéaires.

1.3.6 Construire des graphes d'épissage multi-espèces, ThorAxe

1.3.6.1 Graphe d'épissage multi-espèces

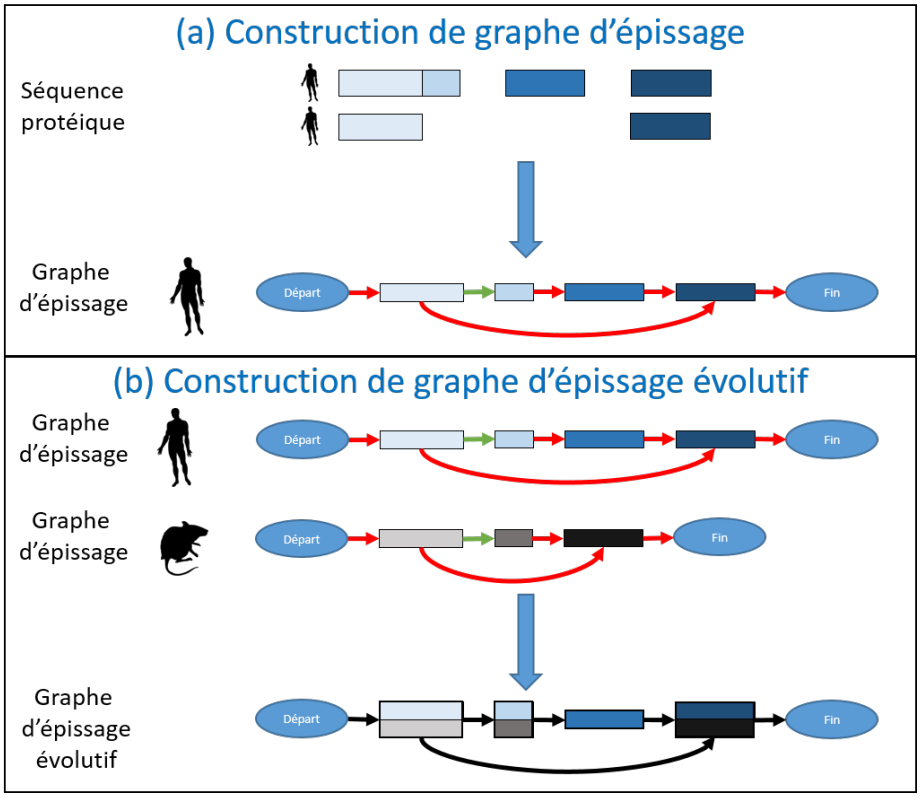


FIGURE 1.17 – Construction d'un graphe d'épissage évolutif (adaptée de ZEA et al. 2021). (a) Un graphe d'épissage illustre pour chaque espèce les sous-exons à partir des séquences protéiques. Deux sous-exons consécutifs appartenant au même *locus* du gène sont liés par une arête induite (flèche verte) et deux sous-exons séparés par un intron sont liés par une arête structurelle (flèche rouge). (b) Pour l'ensemble des gènes orthologues, un alignement global multiple de séquences est réalisé pour chaque sous-exon. Les sous-exons alignés forment un regroupement d'exons épissés (s-exons). L'ensemble des liens entre les s-exons forme le graphe d'épissage évolutif. Sur l'exemple, on pourra lire que l'exon bleu foncé est spécifique à l'humain et absent chez la souris.

ThorAxe (ZEA et al. 2021) est une méthode de description de la variabilité du transcriptome complet entre plusieurs espèces réalisée à partir de graphes d'épissage multi-espèces : les *graphes d'épissage évolutifs*. La méthode identifie les unités exoniques (*sous-exon*, correspondant au bloc minimal d'un exon pouvant former un transcrit) de chaque gène. Puis, la méthode établie les correspondances (orthologies) entre les sous-exons des gènes orthologues pour comparer les gènes et identifier les segments exoniques partagés

et représenter ainsi leur histoire évolutive.

1.3.6.2 Nœuds et alignements multiples

Dans un premier temps, la méthode s'illustre par la représentation implicite d'un graphe d'épissage par gène et par espèce, où chaque nœud représente un intervalle génomique qui correspond à un exon ou un fragment d'exon (appelé *sous-exon*) et les arêtes sont divisées en deux catégories, les arêtes *induites* et les arêtes *structurelles*. Les arêtes induites correspondent à la liaison directe de sous-exons consécutifs et les arêtes structurelles représentent la liaison d'exons séparés par des introns. Pour construire ces graphes, l'algorithme se base sur les séquences protéiques en entrée pour reconstituer les transcrits. Par exemple, dans la Figure 1.17(a), le premier exon peut-être découpé en deux sous-exons indiquant que le deuxième sous-exon peut-être utilisé alternativement dans deux transcrits (extrémités alternatives en 3' de l'exon). Dans ce cas, l'arête entre ces deux sous-exons est une arête induite. Pour les autres exons, ils sont toujours utilisés entièrement. Ce sont des sous-exons complets séparés par des introns. Dans ce cas, les arêtes qui les séparent sont des arêtes structurelles.

1.3.6.3 Orthologie des sous-exons

Dans un second temps, un alignement global multiple de séquences est réalisé pour chaque ensemble de sous-exons de chaque espèce et lorsque des sous-exons sont alignés entre les espèces, ils sont qualifiés d'*exons épissés* (*s-exons*, *spliced exons*). Ainsi, un graphe est construit où chaque nœud représente un s-exon et chaque arête représente le lien entre les s-exons. Ce graphe représente le *graphe d'épissage évolutif*. Un exemple adapté de l'article de ThorAxe est montré en Figure 1.17(b). Avec de nombreuses espèces, plusieurs graphes sont possibles et la méthode attribue un score à chaque graphe en tenant compte du score de l'alignement multiple, des arêtes induites et des arêtes structurelles. ThorAxe est une méthode qui permet de décrire les variations des transcriptomes inter-espèces à partir de ces graphes et ainsi de décrire une notion d'orthologie appliquée au niveau des exons.

1.3.7 Génomique comparative entre deux espèces et prédiction de sites et transcrits orthologues, CG-alcode

Enfin, une autre catégorie de méthode d'identification des transcrits est basée sur les principes de la génomique comparative comme la méthode CG-alcode (Comparative Genomics for ALternative CODing in Eukaryote genes, Blanquart et al. 2016). Selon ces principes, les prédictions de transcrits sont établies en exploitant la conservation de séquences entre gènes orthologues. De plus, c'est une méthode de reconstruction du transcriptome d'un gène assistée par la connaissance. En effet, la méthode transpose les informations connues pour le gène d'une espèce source sur le gène orthologue d'une espèce cible et permet d'identifier quels orthologues des transcrits du gène source le gène cible est capable de produire. Pour être identifiés comme orthologues, deux transcrits doivent notamment partager l'ensemble de leurs sites d'épissage. Cela conduit à la fois à identifier l'orthologie entre les transcrits connus du gène cible et les transcrits connus du gène source, et à prédire de nouveaux transcrits réalisables par le gène cible.

1.3.7.1 Considérer la séquence complète du gène

La méthode utilise en entrée la séquence complète de deux gènes orthologues et l'ensemble de leurs transcrits codants connus. La première étape consiste à projeter les transcrits d'un gène sur la séquence de ce gène pour identifier les fragments exoniques (appelés *blocs codants*) et ainsi reconstituer la structure intron/exon du gène (Figure 1.19(a)). Les blocs codants correspondent aux segments exoniques qui sont appelées sous-exons dans la méthode ThorAxe. La structure du gène est formalisée par une liste de symboles représentant les sites fonctionnels ("[" pour les codons *start*, "]" pour les codons *stop*, "<" pour les sites donneurs et ">" pour les sites accepteurs d'épissage) et les blocs codants (caractères alphabétiques : A, B, etc.). Chaque site fonctionnel et chaque bloc codant est de plus associé à sa coordonnée sur le génome de référence. Par exemple : $M = [A <> B <> C]$ est le modèle d'un gène contenant trois exons codants. Cette étape est réalisée pour les deux gènes orthologues indépendamment.

1.3.7.2 Aligner des modèles de gènes

La deuxième étape consiste à aligner localement par *BLAST* (Altschul et al. 1990) les blocs codants et les sites fonctionnels du gène d'une espèce sur la séquence complète du gène orthologue de l'autre espèce afin de retrouver les blocs codants conservés partagés par

les deux gènes orthologues. Cette étape constitue un point clé de la méthode, puisqu'elle permet de prédire l'orthologie des blocs codants et des sites fonctionnels et de prédire des transcrits orthologues (Figure 1.19(b)). Les alignements révèlent de plus des séquences inconnues conservées, et donc considérées comme orthologues aux séquences connues. L'alignement est fait en ciblant la séquence nucléotidique du gène de l'autre espèce en utilisant un contexte de vingt nucléotides additionnels de part et d'autre du bloc codant de manière à aligner les parties conservées des sites d'épissage. Le score de *E-value* pour les *BLAST* est calibré à 10^{-5} pour les blocs de plus de 20 nucléotides et à 10^{-1} dans les autres cas. Ensuite, les blocs des deux modèles de gènes sont nommés en fonction des résultats de l'alignement : deux blocs codants alignés porteront la même dénomination indiquant leur orthologie. Des blocs non alignés auront leur propre dénomination (1.19(c)).

Sites fonctionnels alignés. Pour chaque site connu d'un gène source, on dispose, en sortie de CG-alcode, des informations suivantes sur le gène orthologue (le gène cible) : 1) si le site n'est pas aligné sur le gène cible, alors ce site n'a pas d'orthologue dans le gène orthologue, 2) si le site est aligné alors il y a plusieurs cas possibles (Figure 1.18) :

— si le site est aligné contre des *gaps*, alors il n'y a pas de site orthologue ;
— si le site est aligné avec un motif (site et contexte nucléotidique) non fonctionnel, alors il n'y a pas de site orthologue ;
— si le site est aligné avec un motif (site et contexte nucléotidique) fonctionnel, alors on a détecté la présence dans le gène cible d'un site orthologue. Deux sous cas se présentent alors : soit le site identifié était déjà connu dans le gène cible, les deux sites alignés sont alors qualifiés de sites orthologues, et ils porteront le même nom, soit il était non connu jusqu'alors, et on le qualifie d'orthologue prédit. Il apparaîtra dans le modèle du gène cible.

Blocs codants alignés. Un bloc d'un gène source est aligné sur un gène orthologue si on aligne à la fois les sites qui flanquent le bloc et la séquence du bloc contre la séquence du gène orthologue. Deux blocs alignés sont alors qualifiés de blocs orthologues.

1.3.7.3 Identifier l'orthologie des transcrits et prédire des transcrits

Enfin, à partir des modèles d'un gène source et d'un gène orthologue (gène cible), la méthode va estimer si l'orthologue d'un transcrit connu du gène source peut être produit par le gène orthologue (1.19(d)). Pour cela, CG-alcode va vérifier la présence, dans le

Non	site non aligné	G_i >AG G_j Ø	Pas de site orthologue
Oui	site aligné contre des *gap*	G_i >AG G_j - -	Pas de site orthologue
	site aligné contre un motif non fonctionnel	G_i >AG G_j GG	Pas de site orthologue
	site aligné contre un motif fonctionnel connu	G_i >AG G_j >AG	Site orthologue connu
	site aligné contre un motif fonctionnel connu	G_i >AG G_j >AG?	Site orthologue connu

FIGURE 1.18 – Alignement d'un site fonctionnel contre la séquence d'un gène orthologue

modèle du gène cible, des blocs codants et sites fonctionnels orthologues nécessaires à sa production par le gène cible. Si tous les blocs codants et tous les sites fonctionnels orthologues sont présents dans le gène orthologue alors le transcrit est considéré comme réalisable par le gène orthologue. Sinon, il manque dans le gène orthologue un ou des éléments nécessaires pour produire l'orthologue du transcrit, et il est donc considéré comme non réalisable. Si le transcrit réalisable par le gène cible était déjà présent parmi les transcrits connus du gène cible alors les transcrits sont qualifiés de transcrits orthologues. Dans le cas contraire, un nouveau transcrit est prédit chez le gène cible. Cette méthode a été appliquée aux génomes de l'humain et de la souris, à partir de données de références (les transcrits connus) issues de la base de données CCDS (PUJAR et al. 2018). Cette thèse repose sur la méthode CG-alcode et ses capacités à prédire des relations d'orthologie entre transcrits (notion définie par ZAMBELLI et al. 2010), à prédire de nouveaux transcrits chez une espèce, orthologues à des transcrits connus chez une autre espèce et à enfin comparer des structures de gènes orthologues chez deux espèces par génomique comparative.

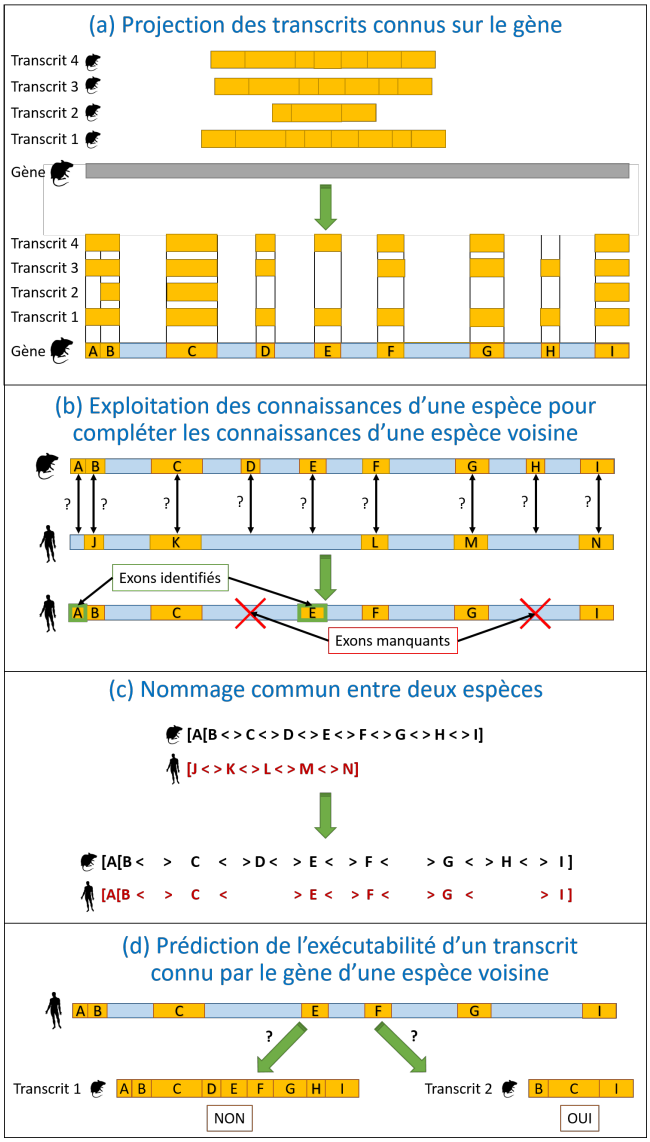


FIGURE 1.19 – Principes de la méthode CG-alcode. (a) Les transcrits connus d'un gène sont projetés sur celui-ci pour reconstituer sa structure en sites d'épissage et en blocs codants. (b) les blocs codants d'un gène sont alignés localement sur l'autre gène pour mettre en évidence des absences (croix rouge) ou des présences (cadre vert) de séquences homologues. (c) Les blocs codants sont dénommés de façon identiques entre les deux espèces si ils sont alignés, indiquant leur caractère orthologue. Pour qu'un orthologue d'un transcrit connu soit exprimable dans un gène, chaque bloc du transcrit doit avoir un orthologue dans ce gène. Par exemple, le transcrit humain $[BCI]$ est composé de trois blocs, chacun ayant un orthologue chez la souris. On conclut que le transcrit $[BCI]$ de la souris et le transcrit $[BCI]$ humain sont orthologues. (d) Prédiction de l'exécutabilité des transcrits issus d'un gène par un gène orthologue.

1.4 Cas d'étude : l'humain, la souris et le chien

1.4.1 Choix des organismes d'étude

D'une manière générale en biologie, les espèces sont étudiées avec des objectifs bien précis. La souris (*Mus musculus*) est une espèce modèle. Elle est notamment étudiée dans un but médical afin d'améliorer la santé humaine. De ce fait, concernant ces deux organismes, de nombreuses informations génétiques sont disponibles dans des bases de données. Ces organismes sont continuellement étudiés pour affiner les connaissances. La souris n'est pas la seule à être considérée comme modèle. En ce sens, on retrouve également d'autres organismes : la bactérie *Escherichia coli*, les levures *Saccharomyces cerevisiae* et *Schizosaccharomyces pombe*, le nématode *Caenorhabditis elegans* et la drosophile *Drosophila melanogaster*. Toutes ces espèces sont devenues des modèles par le fait qu'elles sont peu coûteuses et faciles à reproduire rapidement (Matthews et Vosshall 2020).

Plus récemment, le chien (*Canis lupus familiaris*), est devenu une espèce très étudiée. Notamment cette espèce vit et partage le même environnement que l'homme et est proche phylogénétiquement. Cette espèce est ainsi utilisée comme nouveau modèle pour l'étude des maladies humaines en particulier car les races canines, en tant qu'isolat génétique, présentent des pathologies propres résultant d'un fort taux d'homozygotie. Certaines races canines peuvent servir de modèle pour certaines pathologies humaines (Kirkness et al. 2003 ; Hoeppner et al. 2014 ; Le Béguec et al. 2018). Pour ces raisons, le chien est parfois considéré comme une espèce modèle émergente. Au cours de cette thèse, nous avons utilisé des données génétiques du chien au travers une collaboration avec l'IGDR (Institut Génétique & Développement de Rennes).

Dans cette thèse, nous examinons des données issues de l'humain, de la souris et du chien. D'un point de vue évolutif, les lignées de l'humain et de la souris se sont séparées il y a 75 à 80 millions d'années alors que la divergence avec le chien a eu lieu il y a 90 à 110 millions d'années (Koren et al. 2007 ; Scheidt et al. 2017 ; Foley et al. 2016, Figure 1.20). Les données qui concernent ces espèces dans la base de données Ensembl sont présentées dans la Table 1.3. On remarque en particulier que 2 à 4 fois moins de transcrits sont connus chez la souris et le chien respectivement, comparé à l'humain, pour un nombre de gènes codants comparables. Par conséquent, il est vraisemblable qu'un grand nombre de transcrits peuvent être découverts chez la souris et le chien, tels qu'ils soient des orthologues de transcrits connus exclusivement chez l'humain.

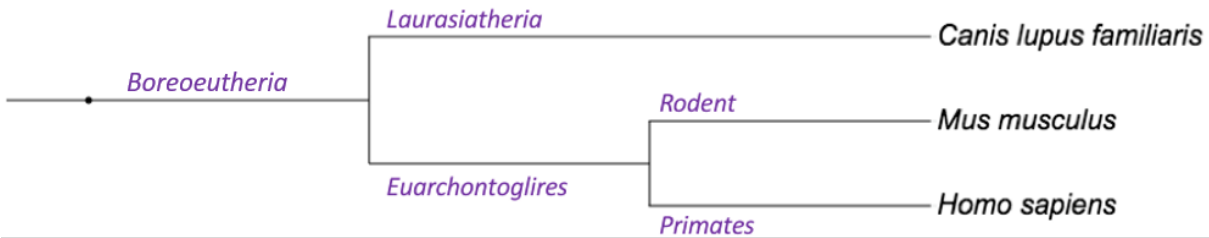


FIGURE 1.20 – Relations phylogénétiques entre l'humain, la souris et le chien. L'humain (*Homo sapiens*) et la souris (*Mus musculus*) sont plus proches par rapport au chien (*Canis lupus familiaris*). *Euarchontoglires* est l'ancêtre commun de l'humain et de la souris. *Boreoeutheria* est l'ancêtre commun aux trois espèces.

Espèces	*Homo sapiens*	*Mus musculus*	*Canis lupus familiaris*
Nombre de gènes codants	20 442	22 468	20 257
Nombre de gènes non codants	23 982	16 060	10 081
Nombre de transcrits connus	237 081	142 434	60 994

TABLE 1.3 – Nombre de gènes et de transcrits connus chez l'humain, la souris et le chien sur la base de données Ensembl en Mars 2021.

1.4.2 Intérêt de passer à une comparaison multi-espèces

Lorsqu'on compare deux espèces, comme avec la méthode CG-alcode, on peut identifier un caractère présent chez une espèce $E1$ mais pas chez une autre espèce $E2$ et ainsi prédire sa potentielle existence chez $E2$. Seulement, si ce caractère n'est connu d'aucune des espèces, celui-ci ne pourra jamais être prédit. En augmentant le nombre d'espèces étudiées, on peut ainsi accumuler plus de connaissances sur les gènes et être le plus complet possible sur leur structure. Par exemple, dans la Figure 1.21, l'humain a trois exons connus (A, X et C) et la souris et le chien ont chacun deux exons (D et E chez la souris, F et G chez le chien). Si on compare l'humain avec l'une des deux autres espèces avec CG-alcode, on peut prédire l'exon X chez le chien mais pas chez la souris, et observer que l'exon C humain correspond à l'exon E chez la souris et à l'exon G chez le chien Par contre, si on compare uniquement la souris et le chien, on met en évidence la correspondance entre leurs deux exons (D coïncide avec F et E avec G) mais on ne peut pas observer la prédiction d'un exon correspondant à l'exon X humain. Ici, la succession des comparaisons de paires de gènes humain-souris, humain-chien et souris-chien, avec CG-alcode aboutit donc à la juxtaposition des trois couples de gènes : "AXC"-"AC", "AXC"-"AXC", "DE"-"DE". La mutualisation des relations d'orthologie (qui sera faite avec des graphes de sites fonctionnels dans cette thèse), quant à elle, permet d'obtenir

une comparaison plus globale des structures des trois gènes humain-souris-chien : "AC"-"AC"-"AXC". Ainsi, passer à une échelle multi-espèce permet d'obtenir des connaissances plus complètes sur la structure d'un gène entre espèces.

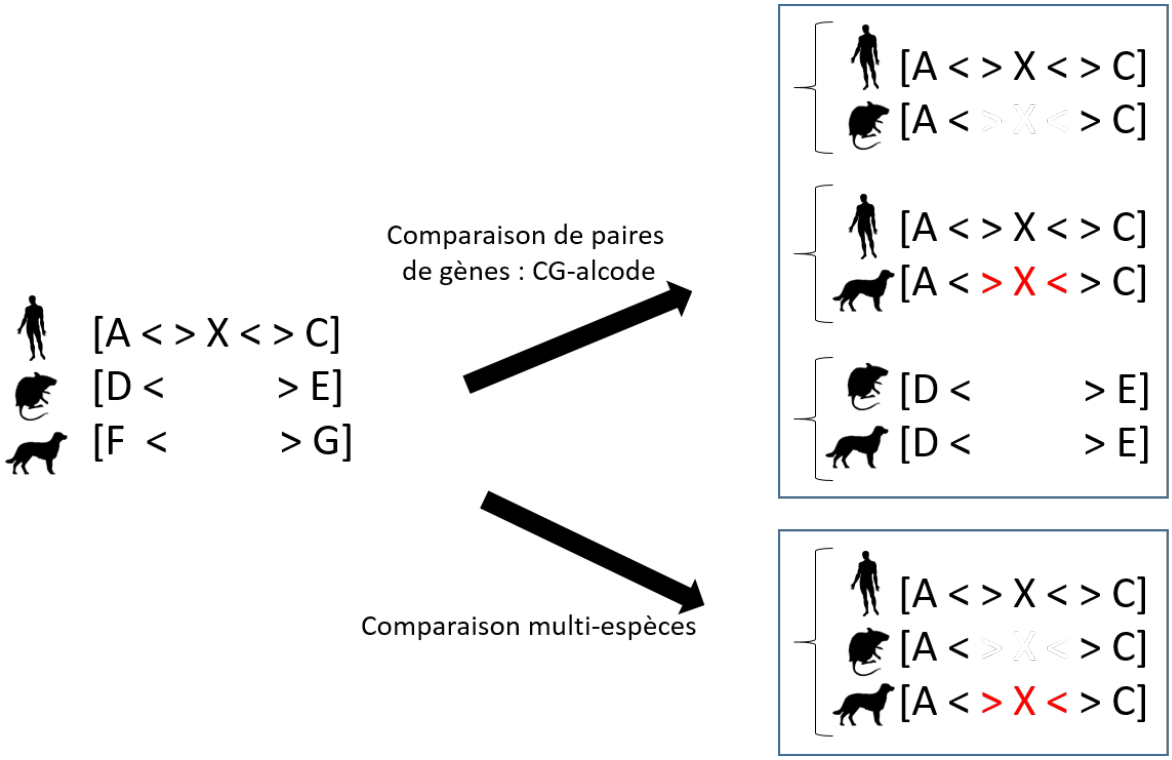


FIGURE 1.21 – Passage à une échelle multi-espèces pour comparer des structures de gènes. L'humain possède 3 exons connus (A, X et C), la souris et le chien ont chacun 2 exons connus (D et E chez la souris, F et G chez le chien). La comparaison via CG-alcode avec l'humain contre chacune des deux autres espèces permet de prédire l'exon X chez le chien. Les comparaisons avec CG-alcode sans l'humain ne permettent pas de mettre en évidence l'exon X. La comparaison de paires de gènes ne permet pas de mettre directement en correspondance les exons entre les trois espèces. La vision trois espèces, mise en œuvre dans cette thèse, révèle quant à elle la correspondance des exons A-D-F ainsi que C-E-G, et fait apparaître l'exon X chez le chien.

1.4.3 Identifier un ensemble de gènes conservés

En phylogénétique, on définit un *caractère orthologue* comme un caractère présent chez un ancêtre commun et hérité en copie unique chez les descendants. Dans cette thèse, on s'intéresse à l'orthologie de plusieurs types de caractères tels que les gènes, les transcrits ou encore les sites fonctionnels (voir Chapitre 2).

En pratique, il existe un certain nombre de gènes partagés entre les espèces. D'après Ensembl Compara (Breschi et al. 2017), 80% des gènes codants humains ont un orthologue chez la souris, et réciproquement, 72% des gènes codants de la souris ont un orthologue chez l'humain. Si il existe de nombreuses ressources décrivant des gènes orthologues, il n'existe pas de méthode décrivant formellement des isoformes orthologues, des exons orthologues et des sites d'épissage orthologues. Dans la thèse, on souhaite identifier parmi les gènes partagés entre l'humain, la souris et le chien, des gènes produisant les mêmes protéines isoformes. D'après l'état de nos connaissances, il n'existe pas d'ensemble publié présentant des gènes aux structures d'épissage (l'ensemble des introns et exons possibles) conservées entre plusieurs espèces. Au cours de la thèse, nous avons constitué un ensemble de gènes qui ont leur structure d'épissage conservée entre les trois espèces, et nous analysons la conservation de leurs transcriptomes.

Conclusion du chapitre

Dans ce chapitre nous avons présenté le contexte biologique et un état de l'art de la thèse, à savoir l'étude des transcrits productibles par un gène eucaryote. Si les méthodes de séquençage permettent l'observation des transcrits exprimés à un moment donné et dans une condition donnée, de nouvelles méthodes alternatives sont nécessaires pour compléter les connaissances sur le transcriptome d'un organisme. La thèse s'appuie sur une méthode de prédiction de transcrits orthologues par comparaison de paires de gènes orthologues, CG-alcode, pour étendre l'étude à une comparaison multi-espèces concernant trois espèces : l'humain, la souris et le chien. Nous décrivons formellement la notion de sites fonctionnels et de transcrits (Chapitre 2). Nous appliquons la méthode de comparaison de structure à un ensemble de gènes orthologues afin de prédire de nouveaux transcrits chez les trois espèces d'étude et nous examinons la validité de ces prédictions (Chapitre 3). Nous identifions et examinons un ensemble particulier de gènes dont la structure d'épissage et les transcrits alternatifs sont conservés (Chapitre 4). Enfin, nous décrivons les perspectives possibles des développements effectués et des résultats obtenus au cours de cette thèse (Chapitre 5).

Chapitre 2

Représentation, comparaison et prédiction de structures de gènes et de transcrits

Dans ce chapitre, nous présentons les formalismes et les méthodes qui ont été développés pour la thèse. Ils permettent d'appréhender trois problèmes principaux. Le premier concerne la représentation de la structure d'un gène et d'un transcrit à partir de leurs annotations. Le deuxième présente la comparaison des gènes et des transcrits entre *deux espèces* dans le but de prédire des sites fonctionnels, des blocs codants et des transcrits, et de définir les structures conservées des gènes et des transcrits. Enfin, le troisième définit la comparaison *multi-espèces* de gènes orthologues, à partir de graphes. En particulier, le graphe des transcrits permet de regrouper les transcrits des différentes espèces partageant une même structure codante (le même CDS). On identifie ainsi des groupes de CDS orthologues.

Sommaire

2.1 Représentation de la structure des gènes et des transcrits . . 56

2.1.1 Unités lexicales des modèles : sites fonctionnels et blocs codants 56

2.1.2 Modèle de structure de gène 58

2.1.3 Modèle de structure de transcrit 58

2.2 Comparaison des structures de gènes et de transcrits entre deux espèces : définitions et prédictions 60

2.2.1 Comparaison des structures de deux gènes orthologues et prédiction de sites fonctionnels 60

2.2.2 Comparaison des structures de transcrits et prédiction de transcrits 62

2.3 Vers le multi-espèce : comparaison de gènes orthologues sur trois espèces 66

2.3.1 Comparer plus de deux espèces : graphes de sites fonctionnels entre gènes orthologues 66

2.3.2 Identification de transcrits structurellement conservés (groupes de CDS orthologues) : graphes de transcrits 66

Définitions des termes du chapitre

Un *CDS* est la séquence exonique d'un ARNm qui peut être traduite en protéine, depuis un codon *start* jusqu'à un codon *stop*, séparés par un nombre entier de codons consécutifs et sans codon *stop* intermédiaire en phase.
Les *blocs codants* sont les segments exoniques qui correspondent aux intervalles entre les sites fonctionnels, connus ou prédits. Il peut s'agir d'un exon codant complet. Ils sont représentés par des lettres ("A", "B", etc.).
Les *sites fonctionnels* correspondent aux codons *start* ("[") et *stop* ("]") et aux sites donneurs ("<") et accepteurs d'épissage (">") formant le premier et le dernier dinucléotides d'un intron.
La *structure d'un gène* est la succession des blocs codants et des sites fonctionnels qui le composent, permettant d'identifier la composition en introns et en exons des CDS, connus et prédits, d'un gène.
Le *modèle de structure d'un gène* est la représentation abstraite de la structure d'un gène, sous la forme d'une collection ordonnée d'unités lexicales représentant les sites fonctionnels et les blocs codants.
Le *modèle de structure d'un transcrit* est la représentation abstraite des blocs codants du gène qui le composent, délimités par deux sites fonctionnels : le codon *start* et le codon *stop*. Ainsi, ce modèle est focalisé sur le CDS (il ne prend pas en compte les régions non traduites, les UTR), et la séquence du transcrit est la concaténation des séquences des blocs codants le composant.
Un *caractère orthologue* est un caractère commun à deux espèces en copie unique, issu d'un évènement de spéciation et hérité de l'ancêtre commun le plus récent de ces deux espèces. On applique ici l'orthologie à trois types de caractères : au niveau du gène, du transcrit et du site fonctionnel.
Un *groupe de CDS orthologues* est un ensemble de transcrits issus d'espèces différentes et ayant la même structure de CDS. C'est une composante connexe des graphes de transcrits.

2.1 Représentation de la structure des gènes et des transcrits

Comme présenté dans la section 1.1.2 (page 19), les pré-ARNm sont structurés en introns et en exons. Le site donneur d'épissage (généralement "GT") et le site accepteur d'épissage (généralement "AG") délimitent chaque intron pour qu'il puisse être supprimé pendant l'épissage, pour ne laisser que des exons constituant le transcrit mature ou ARNm. Un transcrit mature codant contient un CDS (*coding sequence*) qui peut être traduit en protéine. Ce CDS, composé d'une succession de codons, démarre au codon *start* (généralement "ATG") et se termine avec un codon *stop* ("TAG", "TGA" ou "TAA"), et ne contient pas de codon *stop* intermédiaire en phase. Les définitions de structures de gènes et de transcrits proposées dans cette thèse reposent sur des formalismes introduits par Ouangraoua et al. 2012 et employés notamment dans le travail de Blanquart et al. 2016 pour décrire des modèles de gènes et de transcrits.

2.1.1 Unités lexicales des modèles : sites fonctionnels et blocs codants

On définit les sites d'épissage donneur et accepteur ainsi que les codons *start* et *stop* comme des *sites fonctionnels*. Les *exons* sont définis à partir des transcrits. Les exons d'un transcrit sont les segments issus de différents *loci* génomiques que l'on nomme des *loci exoniques*. Un *locus* exonique est un *locus* génomique contenant un ou plusieurs exons alternatifs. Il est entouré de zones introniques. Un exon peut donc être éventuellement fragmenté en plusieurs parties, ce qu'on nomme des *blocs codants*. Par exemple, dans la Figure 2.1, le *locus* exonique peut être utilisé différemment pour former quatre exons alternatifs possibles. On distingue trois intervalles pour ce *locus* correspondant aux blocs codants 'B', 'C' et 'D'. Les sites fonctionnels et les blocs codants définissent les *unités lexicales*. Ces unités lexicales sont les éléments de base pour définir des modèles de structure de gènes et de transcrits (voir Table 2.1).

Définition 1 (Sites fonctionnels)**.** Les *sites fonctionnels* correspondent aux sites donneur (représenté par '<') et accepteur (représenté par '>') d'épissage, qui délimitent les extrémités d'un intron, et aux codons *start* (représenté par '[') et *stop* (représenté par ']'), qui délimitent les extrémités d'un CDS. Ils peuvent être connus ou prédits. Les sites fonctionnels prédits sont des orthologues de sites fonctionnels connus, obtenus par génomique

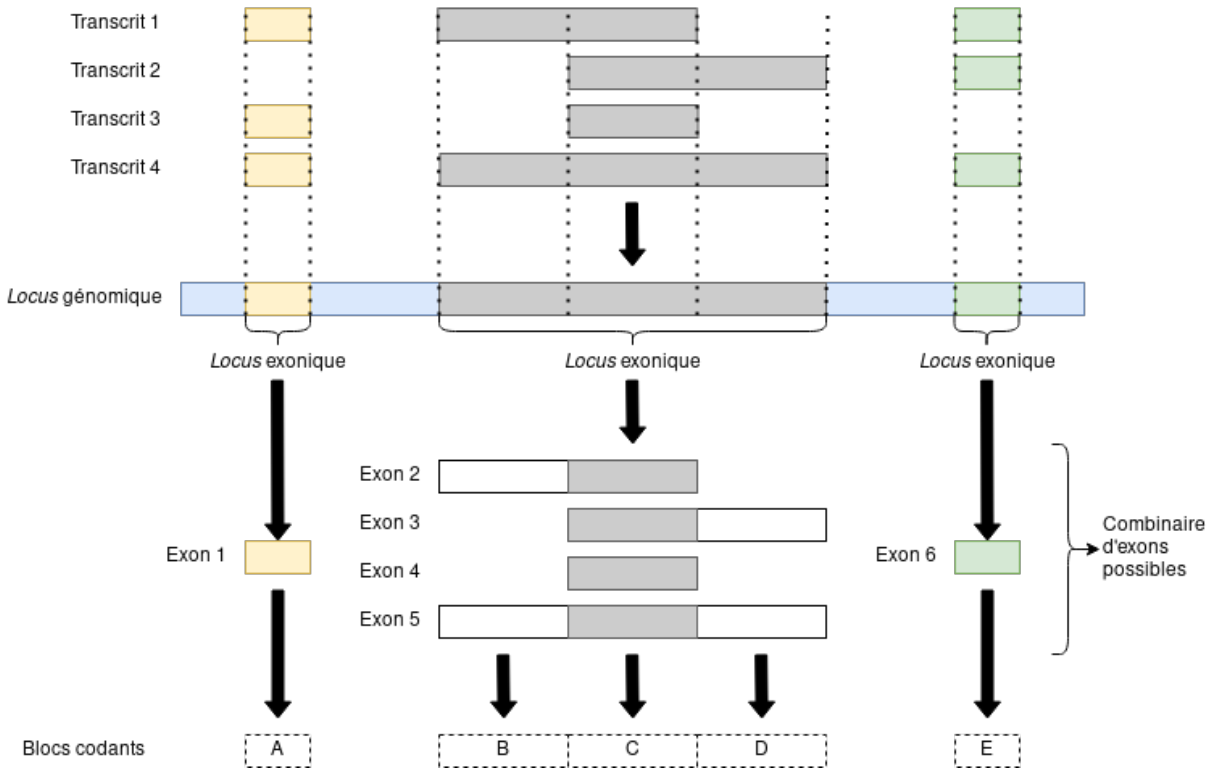


FIGURE 2.1 – Définition des blocs codants d'un exon. Le *locus* génomique présenté contient des *loci* exoniques pouvant donner un (jaune et vert) ou plusieurs (gris) exons. Le *locus* exonique gris peut donner quatre exons différents. Trois des quatre exons peuvent être découpés en plusieurs fragments génomiques : les blocs codants. Le *locus* exonique gris peut donc donner un exon (C) avec un bloc codant ('$\{C\}$'), deux exons (BC et CD) avec deux blocs codants ($\{$'B',‘C'$\}$ et $\{$'C',‘D'$\}$) et un exon (BCD) avec trois blocs codants ($\{$'B', 'C', 'D'$\}$). Ces différentes combinaisons sont dues à des sites d'épissage 3', en aval du bloc 'C', et 5', en amont du bloc 'C', alternatifs (voir section 1.1.3.2 page 24). L'assemblage des différents exons permet de constituer différents transcrits alternatifs.

comparative.

Définition 2 (Blocs codants)**.** Les *blocs codants* (représentés par des caractères alphabétiques 'A', 'B', etc.) correspondent aux intervalles génomiques entre des sites fonctionnels (connus et prédits). Ils désignent des exons complets ou des fragments d'exons présents dans au moins un CDS d'un gène. Un ou plusieurs blocs codants forment un exon (voir Figure 2.1 et section 2.2 page 60).

Unité lexicale	**Valeur**	**Représentation structurelle**
Codon d'initiation (*start*)	ATG	[
Codon de terminaison (*stop*)	TAA, TAG, TGA	]
Site donneur d'épissage	GT	$<$
Site accepteur d'épissage	AG	$>$
Bloc codant	Intervalle génomique	A, B, C, etc.

TABLE 2.1 – Représentation des principales unités lexicales nécessaires à la construction de modèles de structure de gènes et de transcrits.

2.1.2 Modèle de structure de gène

On représente la structure intron/exon d'un transcrit à partir du codon *start* de l'exon codant de l'extrémité 5' du CDS jusqu'au codon *stop* de l'exon codant de l'extrémité 3' du CDS. En utilisant les unités lexicales correspondant aux sites impliqués dans la formation de l'ensemble des CDS d'un gène, on peut créer un modèle de structure de gène.

Définition 3 (Modèle de structure de gène)**.** Le *modèle de structure de gène* M_i^G du gène i est composé d'une liste de N_i unités lexicales, où chaque unité lexicale du gène i, $K_{i,m}$, correspond soit à l'un des sites fonctionnels soit à un bloc codant. La liste est ordonnée suivant les coordonnées génomiques, indépendamment du sens de lecture.

Par exemple, dans la Figure 2.2, le modèle de structure du gène i est représenté par $M_i^G = [A <> B <> C] > D]$ et celui du gène j par $M_j^G = [E <> F <> G]$.

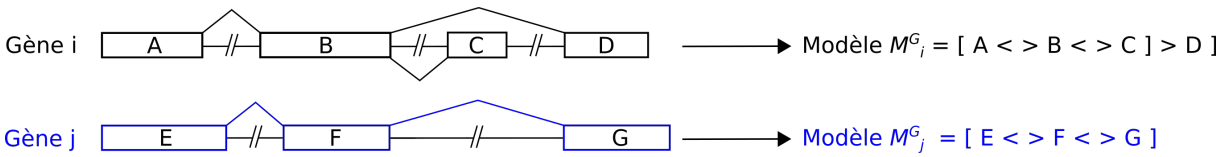


FIGURE 2.2 – Exemple de modèles de structure de gènes. Chaque modèle représente les sites fonctionnels (codons *start* : '[', codon *stop* : ']', sites donneur d'épissage : '<', site accepteur d'épissage : '>' et les blocs codants ('A', 'B', etc.) qui composent un modèle de structure du gène sous forme d'une liste ordonnée d'unités lexicales (à droite).

2.1.3 Modèle de structure de transcrit

Dans cette thèse, l'étude des transcrits se base sur les régions codantes (CDS) des transcrits. Par simplicité et jusqu'à ce que les cas de régions non traduites (UTR) soient considérées, on désignera par "transcrit" le CDS du transcrit.

Définition 4 (Modèle de structure de transcrit)**.** Étant donné un gène i et un de ses transcrits $T_{i,u}$, le *modèle de structure de transcrit* $M^T_{i,u}$ de $T_{i,u}$ représente la structure exonique de la partie CDS du transcrit. $M^T_{i,u}$ correspond à une liste composée d'un sous-ensemble d'unités lexicales $K_{i,m}$ provenant du modèle de gène M^G_i. La première et la dernière unités lexicales de $M^T_{i,u}$ correspondent aux codons *start* et *stop*. Chaque exon du transcrit $T_{i,u}$ est constitué d'un ou plusieurs blocs codants adjacents, et les unités lexicales $K_{i,m}$ correspondantes font partie du modèle $M^T_{i,u}$. Enfin, l'épissage des introns flanquant ces exons nécessitant la présence de leurs sites d'épissage, alors les unités lexicales $K_{i,m}$ correspondantes sont considérées dans $M^T_{i,u}$. La liste est ordonnée suivant les coordonnées génomiques, indépendamment du sens de lecture.

Par exemple, étant donné les modèles de gènes définis dans la Figure 2.2, la Figure 2.3 présente les transcrits réalisables par les gènes i et j. Leurs modèles de structure sont représentés par $M^T_{i,1} = [A <> B <> D]$, $M^T_{i,2} = [A <> B <> C]$ et $M^T_{j,1} = [E <> F <> G]$.

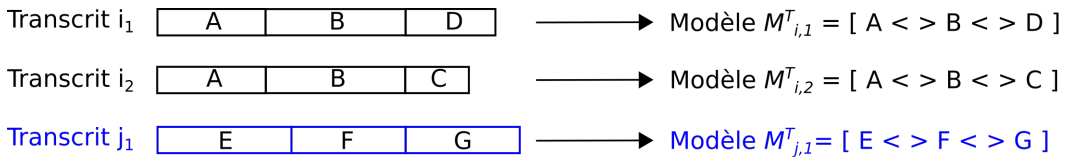


FIGURE 2.3 – Exemple de modèles de structure de transcrits issus des gènes de la Figure 2.2. Chaque modèle représente la concaténation des blocs codants ("A", "B", etc.) et fait apparaître le codon *start* ("[") et le codon *stop* ("]") délimitant un CDS.

Techniquement, les données de départ sont les transcrits connus d'un gène, indiqués sous format GTF. Ainsi, chaque unité lexicale présente dans un transcrit connu est associée à ses coordonnées sur le génome de référence et le modèle de structure de gène est obtenu à partir des modèles de structure de ses transcrits connus. Par exemple, le modèle du gène M^G_i a pu être obtenu à partir des deux modèles de structure des transcrits $M^T_{i,1}$ et $M^T_{i,2}$ alors que pour le modèle du gène M^G_j, seul le transcrit $M^T_{j,1}$ permet sa construction (Figure 2.4 et Figure 1.19(a)).

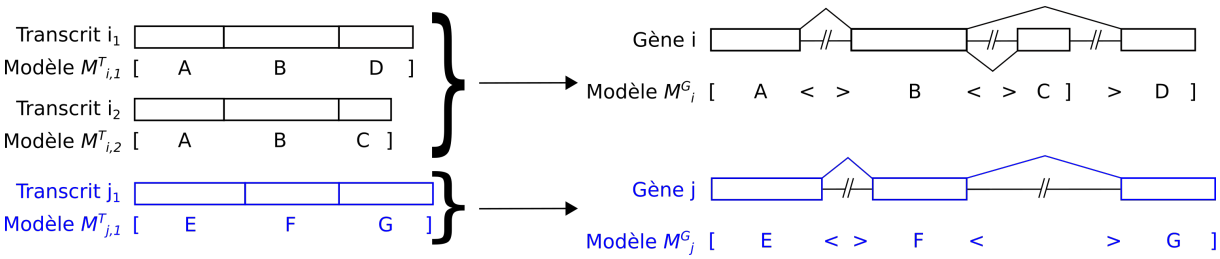


FIGURE 2.4 – Construction des modèles de structure de gènes. Chaque exon des transcrits est projeté sur le gène, en fonction des coordonnées annotées sur un assemblage, pour déterminer les blocs codants et lister les sites fonctionnels permettant de construire le modèle de gène.

2.2 Comparaison des structures de gènes et de transcrits entre deux espèces : définitions et prédictions

2.2.1 Comparaison des structures de deux gènes orthologues et prédiction de sites fonctionnels

Étant donné deux gènes orthologues i et j chez deux espèces, ainsi que leurs transcrits connus, chaque modèle de structure de gène, M_i^G et M_j^G, est d'abord construit à partir de la structure de ses transcrits connus (Figure 2.4). Le modèle de structure du gène i rassemble ainsi les éléments constituant chacun des transcrits connus de i : les blocs codants et les sites fonctionnels apparaissant dans au moins un transcrit. La comparaison d'une paire de gènes, M_i^G et M_j^G, consiste à examiner si chaque unité lexicale $K_{i,m}$ de M_i^G (et *vice-versa* pour les unités lexicales $K_{j,n}$ de M_j^G) est conservée ou non dans le gène j orthologue, ce qui est fait en alignant les blocs codants et les sites fonctionnels du gène i contre la séquence génomique du gène j et réciproquement comme décrit dans BLANQUART et al. 2016 (voir section 1.3.7.2 page 44). Lorsqu'on prédit la relation d'orthologie entre les deux unités lexicales $K_{i,m}$ et $K_{j,n}$ alignées, si $K_{j,n}$ n'était pas une unité lexicale connue alors elle est ajoutée pour compléter M_j^G et est appelée *unité lexicale orthologue prédite*. On retrouve ce cas par exemple dans la Figure 2.5 avec les unités lexicales "$>$", "C" et "]" qui correspondent à des unités lexicales orthologues prédites dans le gène j à partir des transcrits connus du gène i. Dans les cas où les unités lexicales $K_{i,m}$ et $K_{j,n}$ représentent des blocs codants orthologues, la même lettre est utilisée pour indiquer qu'un même bloc

est partagé entre deux gènes, ce qui constitue un caractère orthologue.

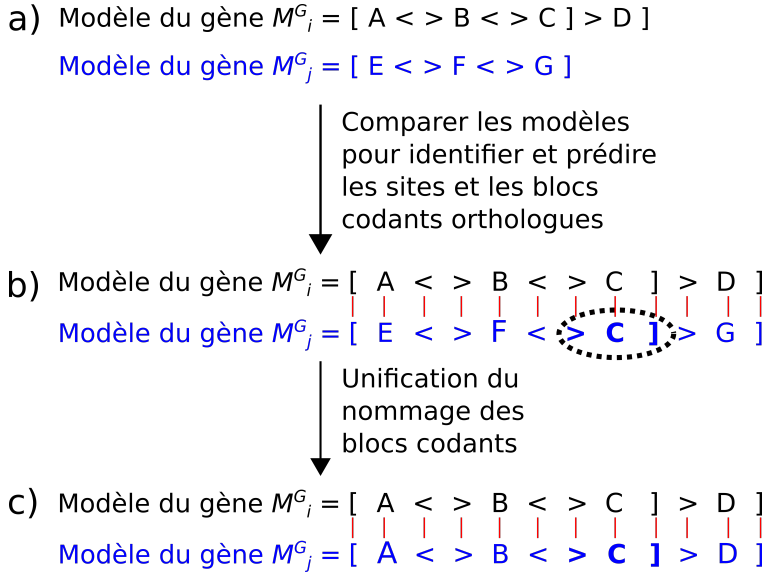


FIGURE 2.5 – Comparaison de deux modèles de gènes. Chaque séquence composant un gène (blocs codants et sites fonctionnels) est alignée sur un autre gène orthologue. Les unités lexicales qui s'alignent sur les séquences conservées sont prédites orthologues et les blocs orthologues sont unifiés par une même lettre (trait rouge). Les unités lexicales qui s'alignent sur une autre séquence qui n'était pas connues sont prédites orthologues (traits en pointillés). Les séquences qui ne s'alignent pas n'ont pas de relation d'orthologie prédite. Ici, les éléments '>', 'C' et ']' du gène i ont été prédits chez le gène j.

Il est intéressant de mentionner qu'un site fonctionnel ou un bloc codant prédit dans le gène j n'appartient à aucun des transcrits connus de ce gène. Un site fonctionnel/-bloc codant prédit dans le gène j à partir du gène i correspond à une séquence associée orthologue à un site fonctionnel/bloc codant appartenant à au moins un des transcrits connus du gène i. Cette mise en évidence de nouveaux blocs codants par une approche de génomique comparative ouvre la voie à la prédiction de transcrits. La comparaison d'une paire de gènes conduit aux définitions suivantes de la conservation des structures de gènes.

Définition 5 (Unités lexicales orthologues)**.** Soit $\mathcal{A}(a, b)$ la relation indiquant que deux unités lexicales a et b sont considérées alignées par CG-alcode. Deux unités lexicales alignées, $K_{i,m}$ du gène i et $K_{j,n}$ du gène j, définissent une paire d'*unités lexicales orthologues*, désignée par $\mathcal{A}(K_{i,m}, K_{j,n})$ (et réciproquement $\mathcal{A}(K_{j,n}, K_{i,m})$, la relation d'alignement des deux unités lexicales étant symétrique). Dans le cas de blocs codants, $K_{i,m}$ et $K_{j,n}$ sont désignés par une même lettre.

Définition 6 (Unité lexicale prédite)**.** Une unité lexicale, $K_{i,m}$ du gène i, qui s'aligne sur la séquence du gène j avec des sites qui ne correspond à aucune unité lexicale connue dans ce gène définit une nouvelle unité lexicale du gène j que l'on appelle *unité lexicale prédite*. Cette unité lexicale prédite est orthologue à l'unité lexicale $K_{i,m}$ du gène i alignée.

Dans le cas des sites d'épissage, deux dinucléotides identiques doivent être alignés pour établir une prédiction, et on considère ainsi d'autres cas que "AG" et "GT". Dans le cas des codons *stop*, les motifs alignés attendus sont "TAA", "TAG" et "TGA". Pour les codons *start*, deux "ATG" doivent être alignés. Lors de l'alignement de sites connus, il n'y a pas de contrôle des séquences des motifs. Ainsi, "TAA" peut être aligné avec "TGA" par exemple.

Définition 7 (Gènes structurellement conservés)**.** Deux gènes i et j dont les modèles de structure de gènes M_i^G et M_j^G ne contiennent que des paires d'unités lexicales orthologues $\mathcal{A}(K_{i,m}, K_{j,n})$ définissent une paire de *gènes structurellement conservés*. Ainsi, l'ordre des unités lexicales orthologues entre les deux gènes est conservé. En conséquence, après renommage des blocs codants orthologues, M_i^G et M_j^G sont syntaxiquement égaux.

Par exemple, dans la Figure 2.5, après comparaison des gènes i et j, le bloc codant 'C' et ses sites fonctionnels flanquants sont prédits dans le gène j et identifiés comme orthologues au bloc 'C' connu dans le gène i. On a ainsi $M_i^G = M_j^G = [A <> B <> C] > D]$. Ces deux gènes sont donc identifiés comme étant structurellement conservés.

2.2.2 Comparaison des structures de transcrits et prédiction de transcrits

La comparaison des modèles de structure peut être appliquée à l'ensemble des transcrits de deux gènes orthologues. Cela permet à la fois d'identifier les relations d'orthologie entre les transcrits connus dont la structure est conservée, et de prédire des transcrits orthologues à un transcrit connu et de même structure. Une comparaison entre les CDS de deux gènes orthologues permet d'évaluer si chaque transcrit connu $T_{i,u}$ du gène i (et réciproquement pour les transcrits $T_{j,v}$ du gène j) possède un CDS structurellement conservé dans le gène orthologue j. En d'autres termes, on recherche si un transcrit $T_{j,v}$ existe dans j avec la même structure de CDS que $T_{i,u}$: $M_{i,u}^T = M_{j,v}^T$.

Si c'est le cas, nous déduisons une relation d'orthologie de CDS entre les transcrits connus $T_{i,u}$ et $T_{j,v}$ (Figure 2.6). Ils feront partie du même *groupe de CDS orthologues*.

Sinon, on examine si un nouveau transcrit $T_{j,v}$ peut être prédit dans le gène j. Pour cela, il faut que chaque unité lexicale impliquée dans le modèle de structure du transcrit $M^T_{i,u}$ ait une unité lexicale orthologue dans le modèle M^G_j de gène j (Figure 2.7).

Définition 8 (CDS structurellement conservés)**.** Deux transcrits $T_{i,u}$ provenant du gène i et $T_{j,v}$ provenant du gène j dont les modèles de structure $M^T_{i,u}$ et $M^T_{j,v}$ ne contiennent que des paires d'unités lexicales orthologues, $\mathcal{A}(K_{i,m}, K_{j,n})$, définissent une paire de *CDS structurellement conservés*. $M^T_{i,u}$ et $M^T_{j,v}$ sont syntaxiquement égaux et l'ordre des unités lexicales orthologues entre les deux CDS est conservé.

Par exemple, dans la Figure 2.6, les transcrits connus $T_{i,1}$ du gène i et $T_{j,1}$ du gène j sont composés uniquement de paires d'unités lexicales orthologues. $T_{i,1}$ et $T_{j,1}$ sont donc des CDS structurellement conservés : $T_{i,1} = T_{j,1} = [A <> B <> D]$. De même, dans la Figure 2.7, les deux transcrits connus étiquetés "1" dans les gènes i et j sont considérés comme des CDS orthologues structurellement conservés puisqu'ils partagent le même modèle $M^T_{i,1} = M^T_{j,1} = [A <> B <> D]$.

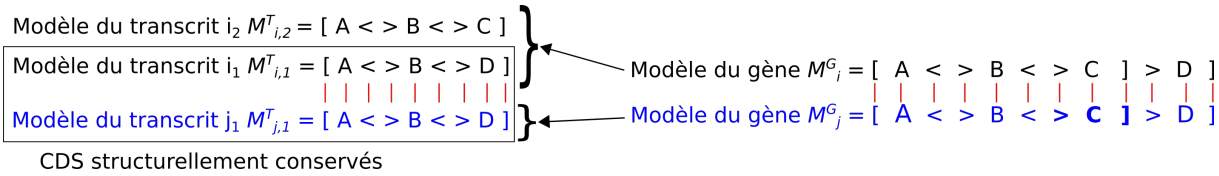


FIGURE 2.6 – Paire de CDS connus structurellement conservés. Le transcrit i_1 et le transcrit j_1 ont les mêmes unités lexicales (trait rouge) indiquant leur relation d'orthologie.

Définition 9 (Séquence d'un transcrit)**.** Étant donnés deux gènes orthologues i et j et un transcrit $T_{i,u}$ provenant du gène i, si chaque unité lexicale de $M^T_{i,u}$ a une unité lexicale orthologue dans le modèle de gène M^G_j, alors une séquence homologue à $T_{i,u}$, appelée $\mathcal{S}(M^T_{i,u}, M^G_j)$, est jugée exprimable dans j. La séquence $\mathcal{S}(M^T_{i,u}, M^G_j)$ consiste en la concaténation des séquences des blocs codants impliqués dans le gène j et désignés comme orthologues par les modèles de gènes.

Par exemple, dans la Figure 2.7, le transcrit étiqueté "2" dans le gène i est formé des exons A, B et C. Tous ces exons ont des orthologues dans le gène j, où les exons A et B sont connus, et l'exon C est prédit. Ainsi, la séquence $\mathcal{S}(M^T_{i,2}, M^G_j)$ est composée d'une concaténation des séquences du gène j désignées par A, B et C.

Définition 10 (Modèle de transcrit exécutable)**.** Si la séquence résultante $\mathcal{S}(M^T_{i,u}, M^G_j)$ forme un CDS, alors le modèle de transcrit $M^T_{i,u}$ est appelé *exécutable étant donné le modèle de gène* M^G_j *et la séquence du gène* j, que l'on désigne par $\mathcal{E}(M^T_{i,u}, M^G_j)$. Pour rappel, cela signifie que la séquence obtenue contient un nombre entier de codons et qu'il n'y a pas de codon *stop* intermédiaire en phase entre le codon *start* et le codon *stop*. Cela indique qu'un transcrit $T_{j,v}$, ayant la séquence CDS $\mathcal{S}(M^T_{i,u}, M^G_j)$ et la même structure de CDS que $M^T_{i,u}$, est potentiellement exprimable dans le gène j. C'est-à-dire que $T_{j,v}$ et $T_{i,u}$ sont des CDS orthologues structurellement conservés.

Définition 11 (Transcrit orthologue prédit)**.** Si un transcrit exécutable, désigné par $T_{j,v}$, n'appartient pas déjà à l'ensemble des transcrits connus du gène j considéré, on l'appelle alors un *transcrit prédit*.

Par exemple, dans la Figure 2.7, le transcrit marqué "2" dans le gène j est un transcrit prédit, impliquant l'exon prédit C. De plus, les séquences concaténées A, B et C du gène j forment un CDS. Ainsi, $M^T_{i,2}$ est *exécutable* par le modèle M^G_j du gène j : $\mathcal{E}(M^T_{i,2}, M^G_j)$. On peut aussi noter que $M^T_{i,2} = M^T_{j,2} = [A <> B <> C]$ avec $\mathcal{E}(M^T_{i,2}, M^G_j)$.

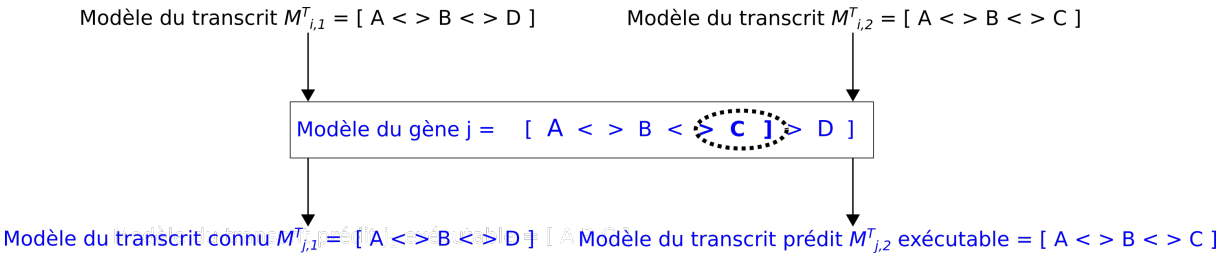


FIGURE 2.7 – Transcrit orthologue prédit et exécutabilité des modèles des transcrits du gène i par le modèle du gène j. Les transcrits $T_{i,1}$ et $T_{i,2}$ sont exécutables par le modèle du gène j. Le transcrit $T_{i,1}$ correspond au modèle du transcrit $T_{j,1}$ connu. Le transcrit $T_{i,2}$ n'a pas de modèle de transcrit correspondant connu chez le gène j et $T_{j,2}$ est qualifié de prédit.

À la fin de la comparaison de paires de gènes entre deux gènes orthologues i et j, nous disposons donc d'un alignement d'une paire de modèles de gènes M^G_i et M^G_j, de relations d'orthologie entre les unités lexicales, notées $\mathcal{A}(K_{i,m}, K_{j,n})$, d'un ensemble de transcrits prédits, et de relations d'orthologie entre les transcrits (vérifiant, $M^T_{i,u} = M^T_{j,v}$, $\mathcal{E}(M^T_{i,u}, M^G_j)$ et $\mathcal{E}(M^T_{j,v}, M^G_i)$), avec les transcrits $T_{i,u}$ et $T_{j,v}$ pouvant être connus ou prédits. La Figure 2.8 récapitule l'ensemble des étapes présentées précédemment.

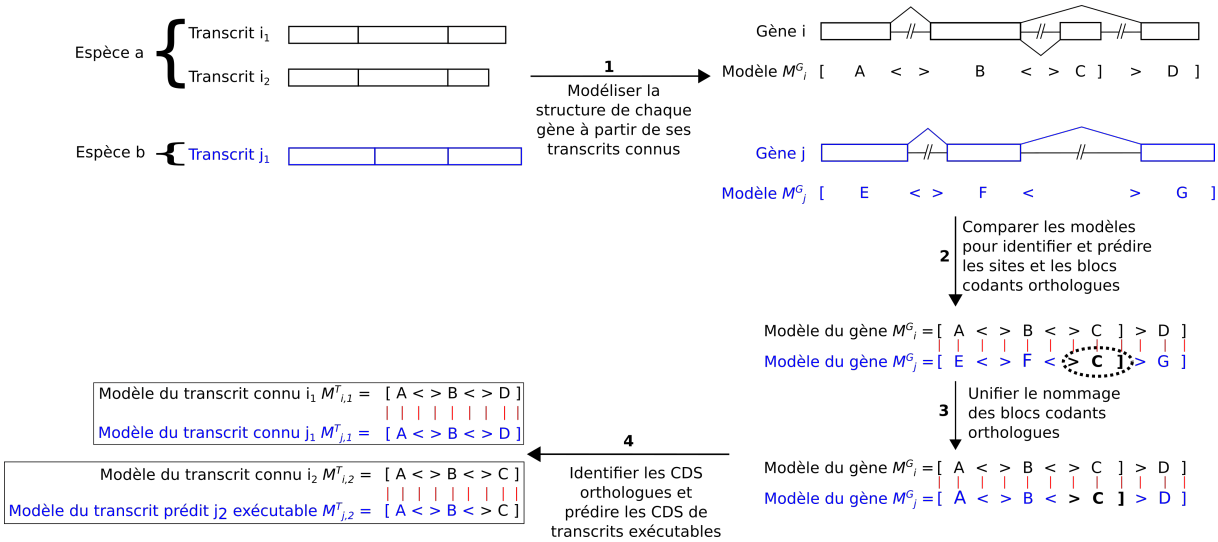


FIGURE 2.8 – Identification de la structure des gènes et des transcrits par comparaison de gènes et prédiction des CDS des transcrits. Cette méthode prend en entrée des transcrits connus et donne en sortie i) des modèles de structure de gènes et de transcrits et ii) des CDS prédits ainsi que iii) des paires de CDS orthologues. (1) Les transcrits connus d'un gène permettent de construire sa structure en blocs codants et sites d'épissage. (2) Les alignements de chaque site fonctionnel et de chaque bloc codant d'un gène contre la séquence d'un autre gène permet i) d'identifier l'orthologie entre les sites fonctionnels connus et les blocs codants pour deux espèces (les éléments orthologues sont représentés dans une même colonne indiquée par une ligne rouge) et ii) de révéler des sites fonctionnels et des blocs codants auparavant inconnus (boîte en pointillés). (3) Dans chaque gène, chaque paire de blocs codants orthologues reçoit un nom commun. Le renommage du nom des blocs codants du gène j est montré. (4) Les éléments révélés permettent de prédire un CDS orthologue d'un CDS connu. Ici, le transcrit j_2, orthologue du CDS du transcrit i_2 connu dans le gène i, est prédit exécutable par le gène j. En effet, comme le bloc codant 'C' et ses sites flanquants '$> C]$' sont prédits présents dans le gène j, le transcrit i_2 apparaît comme réalisable dans le gène j. On prédit enfin des paires de transcrits connus orthologues, par exemple les transcrit i_1 et j_1 partagent la même structure $[A <> B <> D]$.

2.3 Vers le multi-espèce : comparaison de gènes orthologues sur trois espèces

2.3.1 Comparer plus de deux espèces : graphes de sites fonctionnels entre gènes orthologues

Comme décrit ci-dessus, une comparaison d'une paire de gènes entre deux gènes orthologues i et j produit des paires d'unités lexicales orthologues alignées, $\mathcal{A}(K_{i,m}, K_{j,n})$, et des unités lexicales sans orthologue identifié, désignées par $\mathcal{A}(K_{i,m}, -)$, où "$-$" représente un *gap*. Pour chaque unité lexicale, cela définit si elle est partagée par les deux gènes ou si elle est spécifique à un gène, respectivement. Ainsi, étant donné trois gènes orthologues i, j et k dans trois espèces, une unité lexicale partagée par les trois espèces peut être identifiée par trois paires d'alignement : $\mathcal{A}(K_{i,m}, K_{j,n})$, $\mathcal{A}(K_{i,m}, J_{k,o})$, et $\mathcal{A}(K_{j,n}, K_{k,o})$, indiquant que $\{K_{i,m}, K_{j,n}, K_{k,o}\}$ est un triplet d'unités lexicales orthologues. Afin de représenter une comparaison d'unités lexicales au niveau du triplet de gènes, nous avons défini un *graphe de sites fonctionnels*, $\mathcal{G}^{SF}$.

Chaque nœud de $\mathcal{G}^{SF}$ représente une unité lexicale correspondant à un site fonctionnel de l'un des trois gènes. Tous les sites fonctionnels impliqués dans les transcrits connus et prédits sont pris en compte pour construire le graphe. Chaque arête de $\mathcal{G}^{SF}$ relie un site fonctionnel $K_{i,m}$ d'un gène i à un site fonctionnel $K_{j,n}$ d'un autre gène j si $\mathcal{A}(K_{i,m}, K_{j,n})$ (Figure 2.9). De plus, la relation d'orthologie est orientée et deux relations réciproques, "i *aligné avec* j" et "j *aligné avec* i", définissent l'orthologie entre deux sites des gènes i et j. Les blocs ne sont pas considérés dans $\mathcal{G}^{SF}$. On construit un graphe des sites fonctionnels par triplet de gènes orthologues chez trois espèces. Chaque composante connexe de $\mathcal{G}^{SF}$ représente un ensemble de sites orthologues. La comparaison des gènes se base sur l'alignement des sites fonctionnels et des blocs codants. Les graphes ne mettent en évidence que les orthologies de sites fonctionnels. Les orthologies des blocs codants en découlent.

2.3.2 Identification de transcrits structurellement conservés (groupes de CDS orthologues) : graphes de transcrits

Un graphe a été conçu pour comparer les transcrits des gènes entre trois espèces, le *graphe de transcrits*, $\mathcal{G}^{T}$. Chaque nœud de $\mathcal{G}^{T}$ correspond à un transcrit (connu ou prédit) de l'une des trois espèces. Chaque arête de $\mathcal{G}^{T}$ relie un transcrit $T_{i,u}$ d'un gène i à un

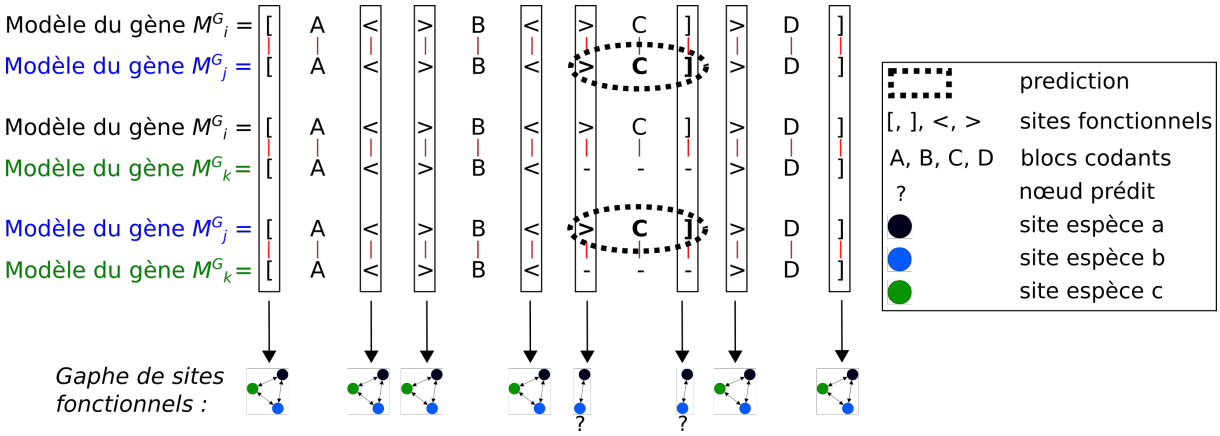


FIGURE 2.9 – Graphe de sites fonctionnels : identification des unités lexicales de gènes partagées entre plusieurs espèces. Les paires d'alignements des modèles de structure de gènes indiquent les relations d'orthologie des sites fonctionnels et des blocs codants et permet de prédire de nouveaux éléments réalisables (par exemple, le bloc C et les sites fonctionnels qui l'entourent "$> C]$" dans le gène j, voir Figure 2.8). Les relations d'orthologie sont rassemblées dans un graphe de sites fonctionnels multi-espèces (bas de la figure), où les nœuds sont des sites fonctionnels, les arêtes sont des relations d'orthologie et "?" indique les sites prédits.

transcrit $T_{j,v}$ d'un gène j si leurs CDS sont orthologues : $\mathcal{E}(M^T_{i,u}, M^G_j)$ et $\mathcal{E}(M^T_{j,v}, M^G_i)$, c'est-à-dire $M^T_{i,u} = M^T_{j,v}$ (Figure 2.10). Nous avons construit un graphe des transcrits par triplet de gènes orthologues chez trois espèces. Chaque composante connexe de $\mathcal{G}^T$ représente un *groupe de CDS orthologues* structurellement conservés.

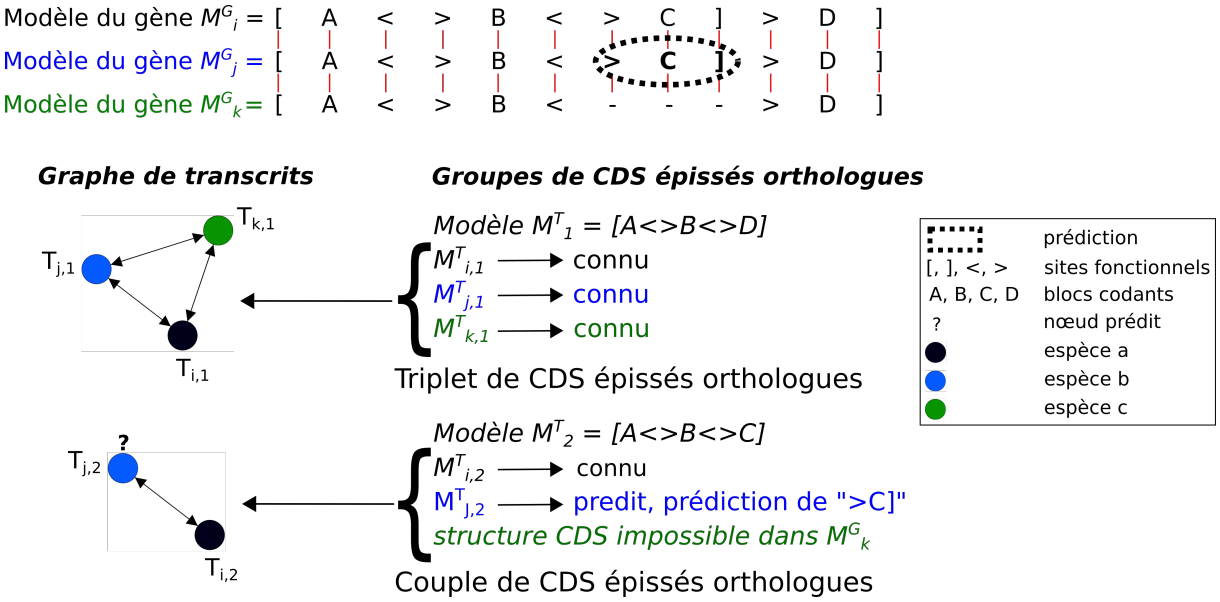


FIGURE 2.10 – Graphe de transcrits : identification de groupes de CDS orthologues structurellement conservés chez trois espèces. Les graphes de transcrits ont été construits en utilisant les relations d'orthologie entre transcrits obtenues pour chaque paire de comparaisons. L'exemple illustre deux groupes de CDS orthologues. Un premier groupe comprend trois CDS orthologues connus ($T_{i,1}$, $T_{j,1}$, $T_{k,1}$). Un deuxième groupe de deux CDS est constitué d'un CDS connu dans le gène i ($T_{i,2}$) et d'un CDS prédit dans le gène j ($T_{j,2}$, où "?" indique la prédiction du CDS). Le gène k ne peut pas former le CDS orthologue en raison de l'absence de l'ensemble de sites fonctionnels "$> C]$" nécessaires dans ce gène.

Conclusion du chapitre

Dans ce chapitre, nous avons décrit des formalismes pour représenter les structures des gènes et des transcrits et nous avons présenté une méthode de comparaison multi-espèces par graphes. Cette méthode repose sur des comparaisons de paires de gène pour définir des ensembles de sites fonctionnels ou de transcrits orthologues. Ces relations d'orthologie permettent de construire des graphes représentant un ensemble (ici de taille trois) de gènes orthologues. Le graphe de sites fonctionnels se base sur les sites fonctionnels afin de décrire la structure des sites d'épissages et les codons *start* et *stop* d'un gène partagé par plusieurs espèces. Le graphe de transcrits s'applique aux CDS (la structure exonique codante d'un transcrit) et décrit les groupes de CDS orthologues, dont la structure est conservée, d'un ensemble de gènes orthologues.

Chapitre 3

Analyse de transcrits de l'humain, de la souris et du chien : prédictions des transcrits par génomique comparative

Dans ce chapitre, nous abordons l'une des questions posées par notre étude : un gène, quels transcrits ? Dans un premier temps, nous nous intéressons à la prédiction de transcrits orthologues en utilisant une méthode de génomique comparative. Cette méthode repose sur la modélisation de la structure des gènes définie par les transcrits connus et les alignements des sites fonctionnels et des blocs codant entre deux gènes orthologues. Pour chaque transcrit connu du premier gène, l'analyse de la structure du gène orthologue permet de déterminer si un transcrit orthologue peut être produit ou non par ce gène. Ces analyses permettent à la fois de prédire des paires de transcrits orthologues connus partageant la même structure, et de prédire de nouveaux transcrits. Avec ces prédictions, on est capable de compléter le répertoire de transcrits des gènes. Dans un second temps, nous avons cherché à valider ces transcrits prédits, et nous leur avons attribué un degré de confiance à partir de deux approches. La première vise à rechercher les transcrits prédits directement dans d'autres bases de données, différentes de celles ayant fourni nos données initiales. La seconde explore des jeux de données de lectures alignées pour y rechercher une trace caractéristique des transcrits prédits, leurs jonctions d'exons spécifiques. Ces analyses ont été appliquées à trois espèces : l'humain, la souris et le chien.

Sommaire

3.1 Données utilisées pour l'humain, la souris et le chien 72

3.2 Prédiction de transcrits à partir de 2 167 triplets de gènes orthologues et de 18 109 transcrits connus 73

3.2.1 Exemple du gène CREM 73

3.2.2 Complétion des transcriptomes : 6 861 transcrits prédits 75

3.2.3 Espèces modèles et non modèles 77

3.3 Analyse de la fiabilité des prédictions 77

3.3.1 Recherche dans des bases de données 78

3.3.2 Recherche dans des données de lectures de séquençage : jonctions d'exons spécifiques 81

3.3.3 Résultats de la recherche des transcrits prédits dans des bases de données auxiliaires 88

3.3.4 Résultats de la recherche des transcrits prédits dans des jeux de données de RNA-seq 88

3.3.5 Résultats des deux méthodes sur les données complètes 90

Définitions des termes du chapitre

La *structure d'un gène* est la succession des blocs codants et des sites fonctionnels qui le composent, permettant d'identifier la composition en introns et en exons du gène en fonction de l'ensemble des CDS exprimés.
Le *modèle de structure d'un gène* est la représentation abstraite de la structure d'un gène, sous la forme d'une collection ordonnée d'unités lexicales de sites fonctionnels et de blocs codants.
Les *sites fonctionnels* correspondent aux codons *start* ("[") et *stop* ("]") et aux sites donneurs ("<") et accepteurs d'épissage (">") formant le premier et le dernier dinucléotides d'un intron.
Les *blocs codants* sont les segments génomiques qui composent les exons utilisés dans les parties codantes (CDS) des transcrits codants. Ils sont représentés par des lettres ("A", "B", etc.).
L'*alignement de modèles de structure de gènes* est l'alignement des blocs codants et des sites fonctionnels d'un gène sur la séquence d'un gène orthologue.
Le *modèle de structure d'un transcrit* est la représentation abstraite des blocs codants du gène qui le composent, délimités par deux sites fonctionnels : le codon *start* et le codon *stop*. Ainsi, ce modèle est focalisé sur le CDS (il ne prend pas en compte les régions non traduites, les UTR), et la séquence du transcrit est la concaténation des séquences des blocs codants le composant.
Un *caractère orthologue* est un caractère commun à deux espèces en copie unique, issu d'un évènement de spéciation et hérité de l'ancêtre commun le plus récent à ces deux espèces.

3.1 Données utilisées pour l'humain, la souris et le chien

Dans le cadre de cette thèse, nous avons utilisé des données pour l'humain, la souris et le chien provenant de deux bases de données : Ensembl (Howe et al. 2021) dans sa version 90 et CCDS (Pujar et al. 2018) en janvier 2018. A cette date, les versions d'assemblage de ces trois espèces correspondaient à *GRCh38.p10* pour l'humain, *GRCm38.p5* pour la souris et *CanFam3.1* pour le chien.

La sélection des données a été effectuée de manière à obtenir les gènes orthologues (relation bijective dénommées "*1-to-1*", "un à un", dans la terminologie de Ensembl Compara) entre ces trois espèces respectant un certain nombre de critères. Tout d'abord, les gènes de l'humain et de la souris ont été récupérés dans la base de données CCDS, une base de données nettoyée manuellement et spécifique à ces deux espèces. Pour ces espèces, seuls ont été sélectionnés les gènes possédant au moins deux transcrits alternatifs. Nous sélectionnons ainsi des gènes pour lesquels l'expression (transcription et épissage) alternative est avérée et dont les annotations de transcrits sont fiables (transcrits complets). Ces couples de gènes humain/souris ont à nouveau été filtrés pour ne garder que des gènes ayant un orthologue également présents chez le chien. Du fait que le chien est une espèce modèle émergente moins documentée que l'humain et la souris, nous avons considéré un gène si il pouvait exprimer, cette fois-ci, au moins un transcrit connu dans la base de données Ensembl. Les transcrits sélectionnés correspondent aux transcrits codants (ou ARNm) associés à ces gènes. Dans la suite, nous nommerons ces transcrits : *transcrits connus*. Notons également que seuls les CDS des transcrits seront comparés, et que par simplicité nous ferons le plus souvent référence au terme de "transcrit" pour parler des "prédictions", des "orthologues", *etc.*, alors qu'il s'agira au sens strict de CDS.

Au total, en janvier 2018, ce sont 2 167 gènes orthologues partagés un-à-un entre l'humain, la souris et le chien qui ont été extraits. Ces gènes exprimant chez les trois espèces 18 109 transcrits connus dont 8 374 chez l'homme, 6 511 chez la souris et 3 224 chez le chien.

Cette thèse analyse ces données pour répondre aux questions définies en introduction.

3.2 Prédiction de transcrits à partir de 2 167 triplets de gènes orthologues et de 18 109 transcrits connus

Dans le Chapitre 2 et dans la section 1.3.7 (page 44), nous avons présenté la méthode CG-alcode permettant de comparer des structures de gènes et de transcrits. Cette méthode repose notamment sur de la prédiction de sites fonctionnels et de blocs codants à partir d'une approche de génomique comparative pour compléter les modèles de gènes avec des sites orthologues prédits. Ces modèles d'alignement de gènes permettent de tester si un transcrit connu dans un gène chez une espèce a peut-être un *transcrit exécutable* par le modèle de son gène orthologue chez une espèce b. Si c'est le cas et qu'aucun des transcrits connus chez b n'a ce modèle de transcrit, alors ce transcrit est un *transcrit prédit*.

3.2.1 Exemple du gène CREM

Pour illustrer la méthode de prédiction, nous prenons l'exemple du gène CREM (modulateur de l'élément de réponse à l'adénosine monophosphate cyclique - *Cyclic adenosine monophosphate Responsive Element Modulator*). La comparaison de la paire de gènes CREM orthologues chez l'humain et la souris est présentée à la Figure 3.1. Le premier point qui est à considérer avant tout concerne le modèle de structure des deux gènes (haut de la Figure 3.1). Si on regarde ces deux modèles, on observe que des sites fonctionnels ne sont pas alignés comme par exemple le bloc codant "O" dans le gène humain n'apparaît pas dans le gène de la souris. Dans le modèle d'alignement $\mathcal{A}$ des gènes, on note $\mathcal{A}(O, -)$ où "O" est le bloc spécifique à l'humain et "$-$" définit un *gap*). Concrètement, cela signifie que si un transcrit humain nécessite ce bloc codant O alors, étant donné qu'il n'est pas retrouvé dans le modèle du gène de la souris, le transcrit n'est pas exécutable par le modèle du gène de la souris.

Dans la comparaison de l'humain vers la souris, sur les 21 transcrits humains connus, 3 nécessitent ce bloc codant "O". Ainsi, ces 3 transcrits humains ne sont pas exécutables par le modèle de structure du gène CREM de la souris et sont considérés comme spécifiques au gène CREM humain. Concernant les autres transcrits du gène CREM humain, on retrouve 7 transcrits qui sont exécutables par le modèle de structure du gène de la souris et qui correspondent à des CDS déjà connus chez la souris. On infère donc 7 *relations d'orthologie* entre transcrits connus. Par construction, ces transcrits partagent le même

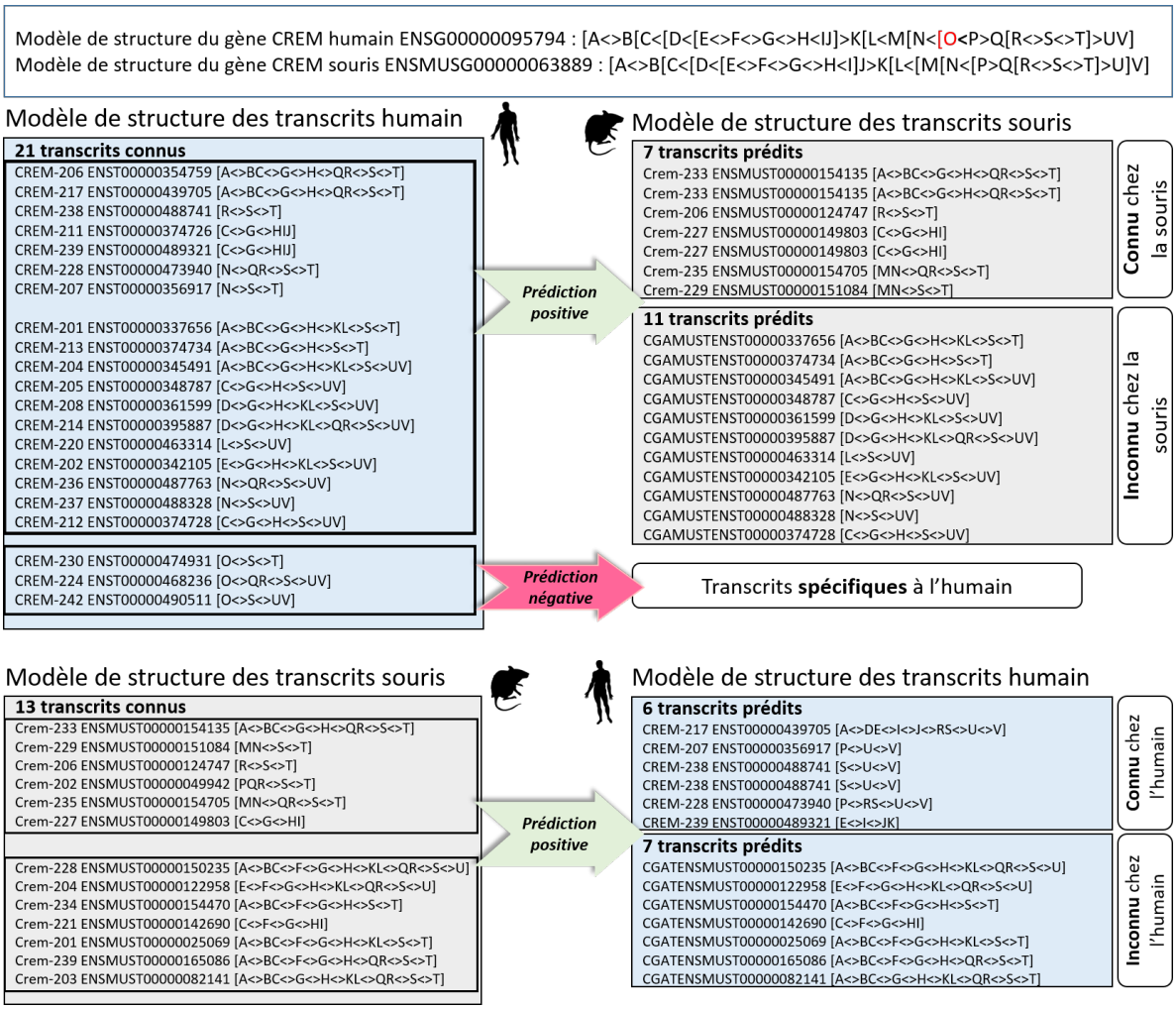


FIGURE 3.1 – Prédiction de l'exécutabilité des transcrits connus du gène CREM. Les modèles obtenus pour les gènes CREM de l'humain et de la souris sont présentés en haut de la figure. La comparaison des transcrits de l'humain avec le gène de la souris concerne 21 transcrits connus. 18 sont exécutables par le modèle du gène CREM de la souris (indiqués par la flèche "*Prédiction positive*"), dont 7 sont déjà connus parmi les transcrits de la souris et 11 ne le sont pas et sont donc prédits. 3 transcrits humains ne sont pas exécutables par le gène CREM de la souris (indiqués par la flèche "*Prédiction négative*") et sont spécifiques au gène CREM humain. 13 transcrits connus de la souris sont exécutables chez l'humain dont 6 correspondent à un orthologue connu et 7 ont un orthologue prédit chez l'humain.

modèle de structure de transcrit sont donc liés par une *relation d'orthologie* entre transcrits connus. Enfin, on retrouve 11 transcrits qui sont exécutables chez la souris mais dont les modèles de structure ne correspondent à aucun transcrit connu de la souris, ils sont alors

prédits chez le gène CREM de la souris. Ainsi, on obtient, de plus, 11 relations d'orthologie entre un transcrit connu chez l'humain et un transcrit prédit chez la souris.

Dans la comparaison de la souris vers l'humain, les 13 transcrits de la souris sont tous exécutables par le modèle du gène humain. 6 ont déjà leur modèle identifié parmi les transcrits connus humain ce qui permet d'inférer 6 relations d'orthologie entre transcrits connus. 7 transcrits de la souris ont un modèle non identifié chez l'humain bien qu'étant exécutable chez l'humain. Ils sont donc prédits et forment une relation d'orthologie entre transcrits connus chez la souris et transcrits prédits chez l'humain.

On réalise ces comparaisons de paires de gènes orthologues pour les trois espèces (humain ↔ souris, humain ↔ chien, souris ↔ chien) ce qui représente un total de 6 comparaisons de gènes pour prédire des transcrits (voir Figure 3.1). L'Annexe 1 montre la comparaison du gène CREM entre l'humain et le chien, et l'Annexe 2 montre la comparaison entre la souris et le chien. Avec toutes les paires de gènes comparées, on obtient des modèles d'alignements de paires de gènes incluant les connaissances actuelles issues des transcrits connus, et de nouveaux sites et transcrits orthologues prédits par génomique comparative.

3.2.2 Complétion des transcriptomes : 6 861 transcrits prédits

Comme présenté dans la section 2.2.2 (page 62) et illustré dans le paragraphe précédent, nous avons appliqué la méthode de prédiction de transcrits par comparaison de paires de gènes sur l'ensemble des 2 167 triplets de gènes orthologues entre l'humain, la souris et le chien. Les résultats obtenus à partir de ces comparaisons sont illustrés dans la Figure 3.2. Au total, nous avons pu prédire 6 861 transcrits exécutables.

La Figure 3.2 montre notamment l'origine des prédictions réalisées chez l'humain. On observe 1 223 transcrits qui ont été prédits au moins à partir des transcrits connus chez la souris (certains d'entre eux ont également pu être prédits en considérant des transcrits connus chez le chien, voir ci-après) et 317 à partir des transcrits connus exclusivement chez le chien. Ainsi, 1 540 transcrits prédits sont ajoutés au transcriptome de l'humain qui était composé de 8 374 transcrits connus, ce qui représente un ajout de 18,4% par rapport à son transcriptome connu. On parvient donc à compléter le transcriptome humain pourtant très étudié.

Chaque prédiction réalisée peut provenir de deux sources de données. Par exemple, pour l'humain, on peut prédire un même transcrit à la fois à partir de la souris (comparaison humain-souris) et à partir du chien (comparaison humain-chien). On s'intéresse dans

ce chapitre aux transcrits prédits, que l'on cherche ici à nommer. Ainsi, pour nommer un transcrit prédit on ne mentionnera qu'une seule des espèces sources possédant un transcrit orthologue connu. Par convention, pour les trois espèces, l'humain est prioritaire sur la souris et la souris prioritaire sur le chien. Donc, pour des transcrits prédits chez l'humain, on nomme la souris comme source avant le chien, et les identifiants produits pour les transcrits prédits seront construit avec un préfixe CGA (pour CG-alcode), suivit de l'identifiant de l'espèce où la prédiction est réalisée, suivit enfin de l'identifiant Ensembl du transcrit orthologue à l'origine de la prédiction (voir Annexe 3). Certains des 1 223 transcrits prédits chez l'humain à partir de la souris auraient pu également être prédits à partir du chien. En revanche, les 317 transcrits prédits prédit chez l'humain à partir du chien sont des prédictions exclusives au chien.

Au total, ce sont 6 861 transcrits prédits (1 540 chez l'humain, 2 112 chez la souris et 3 209 chez le chien) qui, ajoutés aux 18 109 transcrits connus des trois espèces, représentent une augmentation du nombre d'annotations de 37,9% pour les trois espèces confondues.

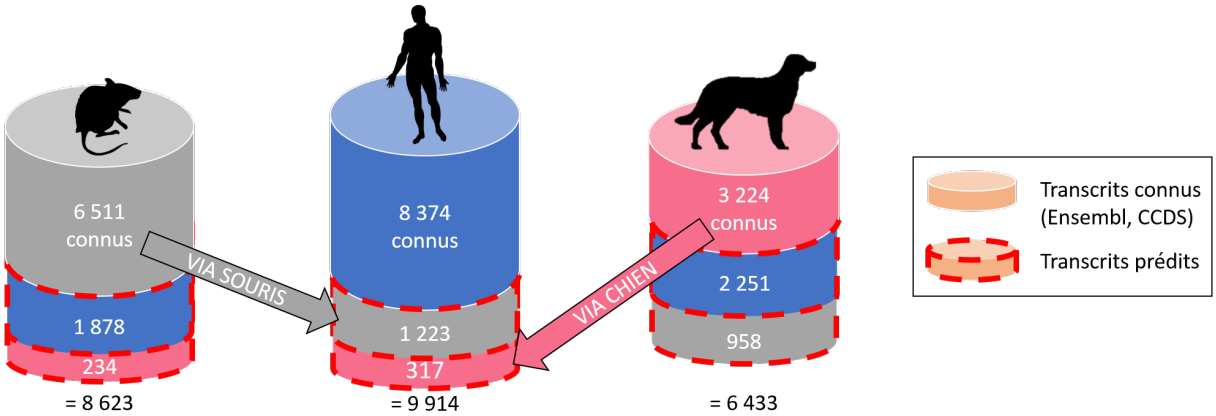


FIGURE 3.2 – Prédiction de transcrits à partir de transcrits connus. La somme de chaque colonne donne le nombre de transcrits connus et prédits chez chacune des espèces. Le milieu et le bas de chaque colonne correspondent aux prédictions effectuées à partir des deux autres espèces. Par exemple, 8 374 transcrits sont connus chez l'humain, 317 ont été prédits chez l'humain en utilisant exclusivement les connaissances sur les 3 224 connus chez le chien et 1 223 ont été prédits au moins à partir des 6 511 transcrits connus chez la souris.

3.2.3 Espèces modèles et non modèles

Les 6 861 prédictions ne sont pas distribuées de façon homogène (Figure 3.2). En effet, les transcriptomes de base (CCDS et Ensembl 90) de l'humain et de la souris sont composés de 8 374 et 6 511 transcrits connus, ce qui est bien supérieur aux 3 224 transcrits connus pour le chien, une espèce modèle émergente et moins documentée. On s'attendrait donc à obtenir beaucoup de prédictions chez le chien. En effet, on a vu précédemment (Figure 3.2) que 1 540 transcrits sont prédits chez l'humain (18,4% de son transcriptome de base). Or pour le chien, ce sont 3 209 transcrits qui sont prédits, soit une augmentation de 99,5% de son transcriptome de base. La souris, quant à elle, a 2 112 transcrits prédits (32,4% de son transcriptome de base). L'humain, l'espèce la plus documentée, est bien l'espèce ayant eu le moins de transcrits prédits à l'inverse du chien, l'espèce la moins documentée.

3.3 Analyse de la fiabilité des prédictions

Bien que les transcrits soient prédits en posant un certain nombre de garanties (exécutabilité, conservation du CDS, présence d'un codon *start* et d'un codon *stop*, nombre entier de codons et absence de codon *stop* intermédiaire pour ne pas décaler le cadre de lecture de la séquence), il est important d'estimer la pertinence de nos prédictions vis à vis des connaissances en biologie. Pour cela, on va rechercher si certains des transcrits prédits peuvent être retrouvés dans d'autres ensembles de données, ce qui conforterait l'ensemble des résultats de la méthode. Tout d'abord, nous recherchons si ces transcrits prédits sont présents dans des bases de données publiques. Sinon, nous tentons d'identifier dans des données issues de séquençage (les lectures) un élément discriminant de nos transcrits prédits : les *jonctions d'exons spécifiques* aux prédictions, que nous définissons dans la section 3.3.2 (page 81). Avec ces analyses, nous avons attribué à chaque transcrit prédit une annotation parmi les quatre suivantes : *confirmé*, *possible*, *réalisable*, *non réalisable* (Figure 3.3). Chacune de ces quatre annotations étant mutuellement exclusive.

On définit ces catégories de la façon suivante :

Définition 12. Un *transcrit confirmé* est un transcrit prédit retrouvé parmi les transcrits connus d'une base de données.

Définition 13. Un *transcrit possible* est un transcrit prédit dont aucune jonction d'exons

n'est caractérisée comme spécifique, c'est-à-dire que toutes les jonctions d'exons du transcrit prédit sont déjà observées dans d'autres transcrits connus.

Définition 14. Un *transcrit réalisable* est un transcrit prédit possédant au moins une jonction d'exons spécifique, et tel que toutes ses jonctions spécifiques sont soutenues par des données de lectures issues de séquençage.

Définition 15. Un *transcrit non réalisable* est un transcrit prédit possédant au moins une jonction d'exons spécifique non soutenue par des données de lectures issues de séquençage.

3.3.1 Recherche dans des bases de données

Nous avons utilisé les données provenant de la version 90 d'Ensembl pour réaliser nos prédictions, nous pouvons donc nous appuyer soit sur des données plus récentes de Ensembl soit sur des bases de données indépendantes. Nous avons comparé nos transcrits prédits avec trois nouvelles versions d'Ensembl (96, 98, 102) pour les trois espèces et la version 103 pour l'humain et le chien (la plus récente utilisée au moment de la rédaction). La version 103 n'a pas été utilisée pour la souris compte tenu du fait qu'elle repose sur une nouvelle version d'assemblage. Sur cette nouvelle version (GRCm39), les coordonnées génomiques sont donc différentes de celles de la version précédente (GRCm38) sur laquelle repose nos résultats. On ne peut donc pas comparer directement nos résultats basés sur l'ancienne version avec la nouvelle version d'assemblage de la souris. En effet, on utilise des fichiers au format GTF comme entrée pour la recherche de transcrits prédits dans d'autres sources de données. Cette recherche est donc basée sur des coordonnées génomiques exactes et non de l'identité de séquences génomiques.

Nous avons également utilisé les données présentes dans la base de données de l'UCSC pour les trois espèces. En plus de cela, nous avons utilisé les données de XBSeq pour l'humain et la souris (Chen et al. 2015), un outil mis en place pour mesurer les données d'expression, et de FEELnc pour le chien (Wucher et al. 2017), un outil pour annoter les ARN longs non-codants (lncRNA) et décrivant également des transcrits ARNm prédits chez le chien.

Afin de pouvoir comparer nos prédictions aux données de ces bases, nous avons employé le logiciel BedTools (Quinlan et Hall 2010) dans sa version 2.18 avec son option d'*intersection de fichiers*.

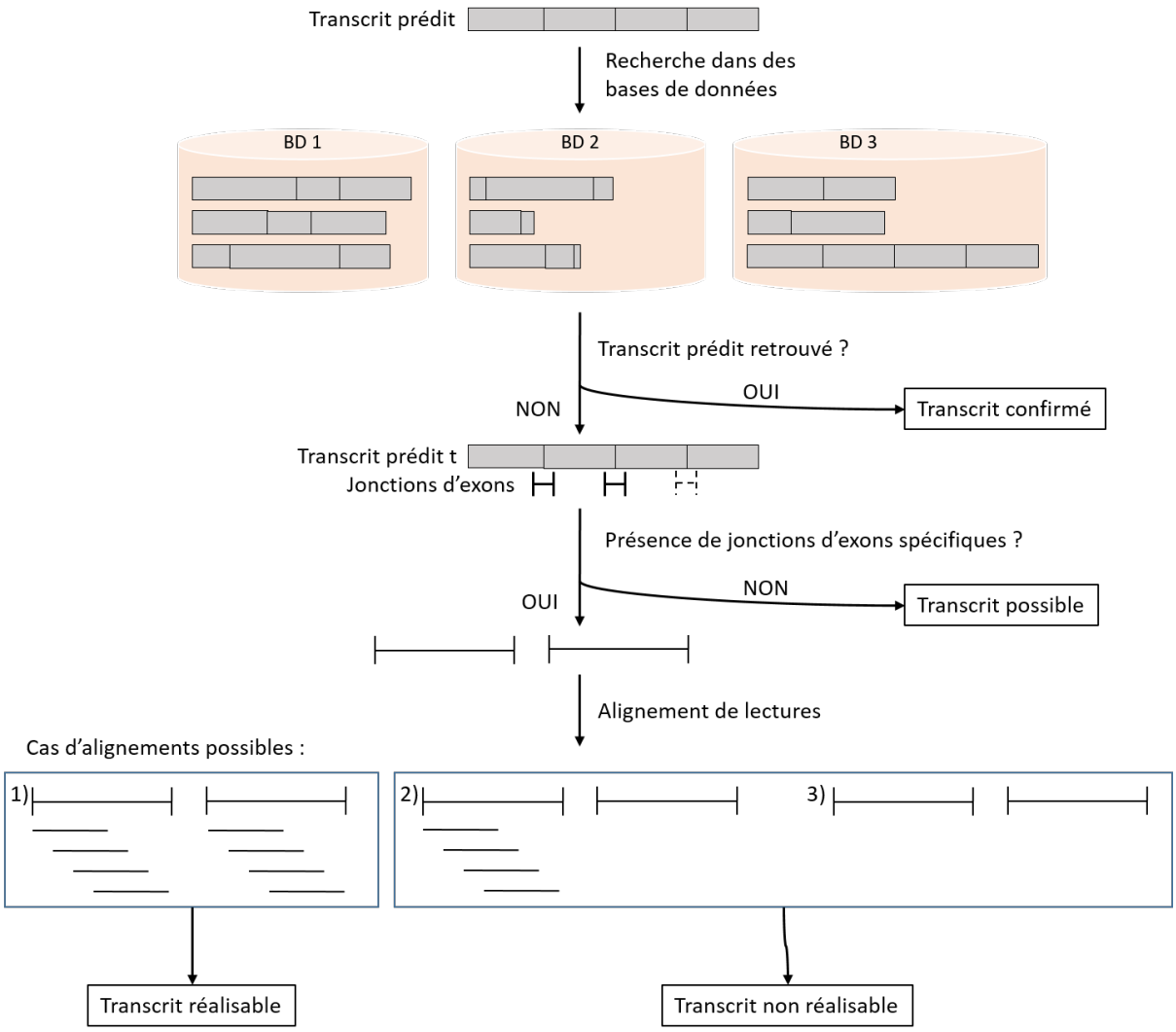


FIGURE 3.3 – Protocole de vérification des transcrits prédits. Étape 1 : recherche du transcrit prédit t dans les bases de données. Si le transcrit t y est retrouvé, il est considéré comme *confirmé*. Étape 2 : recherche de la présence de jonctions d'exons spécifiques à t. Si toutes les jonctions d'exons de t sont déjà observées parmi les transcrits connus, le transcrit est considéré *possible*. Si il existe des jonctions d'exons spécifiques et que des lectures peuvent s'aligner sur toutes ces jonctions d'exons spécifiques, le transcrit est considéré *réalisable*, sinon il est *non réalisable*.

Chaque fichier de base de données, des fichiers GTF, ont été convertis au format BED12, un format de fichier tabulé qui décrit, dans notre cas, pour chaque ligne, un transcrit et sa composition en exons (voir Annexe 4.1).

De plus, pour chaque transcrit, on ne conserve que les informations concernant son CDS pour avoir une correspondance parfaite entre un CDS connu dans une base de don-

nées et un CDS qu'on a prédit. L'Annexe 4.1 donne un exemple de représentation d'un transcrit au format BED12.

L'outil *BedTools intersect* vérifie les correspondances entre nos prédictions et les transcrits des bases de données (voir Annexe 4.2). Si l'une de ces bases contient un transcrit qui, décrit au format BED12, correspond exactement à un transcrit prédit, alors le transcrit prédit est annoté comme *confirmé*. L'ensemble des étapes du pipeline employé pour cette étape de vérification est présenté dans la Figure 3.4 et dans l'Algorithme 1 .

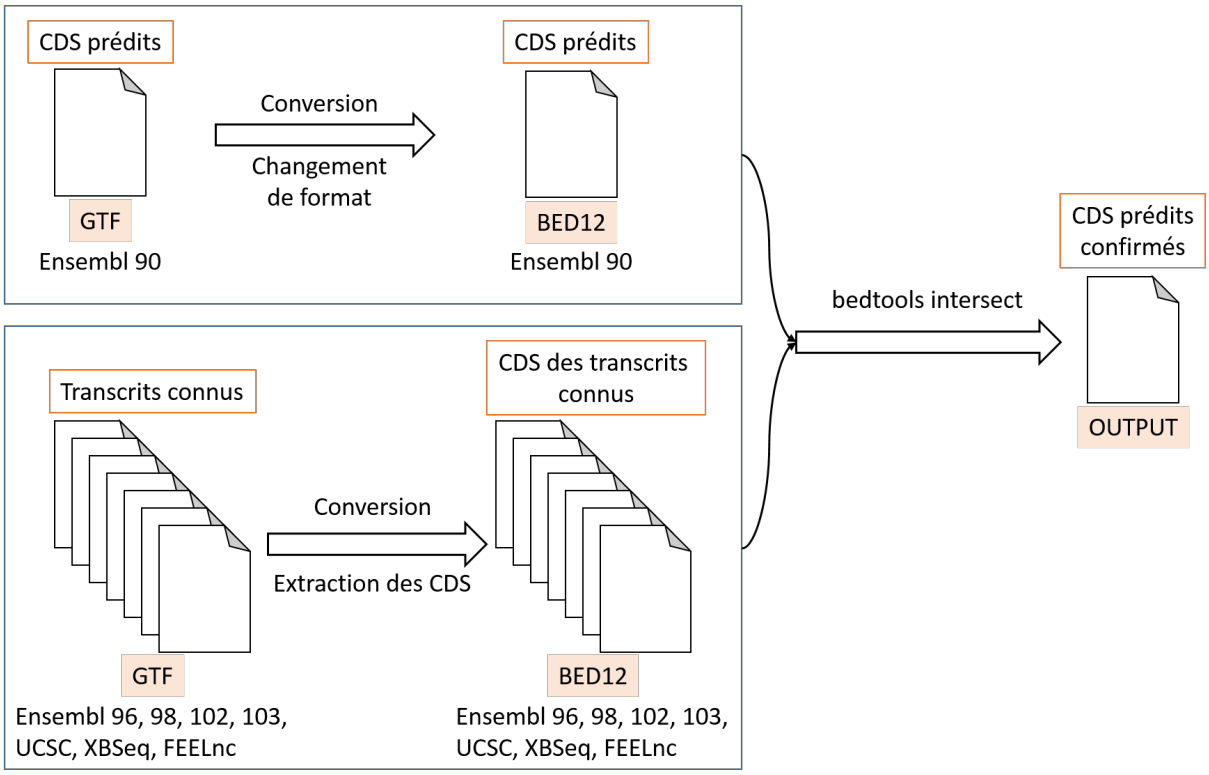


FIGURE 3.4 – Confirmation des transcrits prédits à partir des transcrits connus dans les bases de données.

Algorithme 1 Confirmation des transcrits prédits à partir de transcrits connus auxiliaires

Entrée: CDS prédits, transcrits connus auxiliaires
Sortie: CDS étiquetés "confirmé"
  pour tout transcrit connu auxiliaire **faire**
    extraire CDS
    convertir au format BED12
  fin pour
  pour tout CDS prédit **faire**
    exporter au format BED12
  fin pour
  pour tout CDS prédit **faire**
    BedTools intersect
    si CDS prédit entièrement aligné à un CDS connus **alors**
      CDS prédit étiqueté "confirmé"
    fin si
  fin pour

3.3.2 Recherche dans des données de lectures de séquençage : jonctions d'exons spécifiques

La seconde méthode de vérification des prédictions, pour attribuer un degré de confiance à nos transcrits prédits, se base sur l'identification d'éléments appartenant spécifiquement à un transcrit prédit : les *jonctions d'exons spécifiques*. Une jonction d'exons est le lien entre deux exons consécutifs. Nous définissons une *jonction d'exons spécifique* à un transcrit par les définitions suivantes en considérant l'ordre des exons sur le brin sens (Figure 3.6). Pour le brin antisens, l'ordre s'inverse pour les définitions.

Définition 16 (Précédence d'exons)**.** Étant donnés deux exons e_i et e_j d'un transcrit T définis par leurs positions de début a et de fin b sur un génome de référence, $e_i = [a_i, b_i]$ et $e_j = [a_j, b_j]$. On dit que e_i est situé avant e_j sur le transcrit, c'est-à-dire que e_i précède e_j, si $b_i < a_j$.

Définition 17 (Exons consécutifs)**.** On dit que deux exons e_i et e_j sont consécutifs dans un ensemble de transcrits τ si e_i et e_j sont contigus sur un des transcrits de τ, c'est-à-dire si il existe un transcrit $T \in \tau$, constitué des exons $e'_1, ..., e'_n$, et un index k tel que $e'_k = e_i$ et $e'_{\mathrm{k+1}} = e_j$. On note le couple d'exons consécutifs, sans exon intermédiaire, par $e_i \prec e_j$ dans T.

Définition 18 (Jonction d'exons)**.** Une jonction d'exons est le couple de positions $[b_i, a_j]$ entre deux exons consécutifs $e_i \prec e_j$ (Figure 3.5).

FIGURE 3.5 – Illustration d'une jonction d'exons. La jonction d'exons est représentée par la double flèche rouge séparant les deux exons e_i et e_j. Les nucléotides aux positions génomiques b_i et a_j sont consécutifs dans le transcrit possédant la jonction $e_i \prec e_j$.

Définition 19 (Jonction d'exons spécifiques)**.** Une jonction d'exons JE est dite spécifique à un transcrit T par rapport à un corpus τ, si $JE \in T$ et $JE \notin \tau$. Une jonction d'exons spécifique pointe un couple d'exons apparaissant consécutifs dans T mais pas dans les transcrits du corpus τ (Figure 3.6).

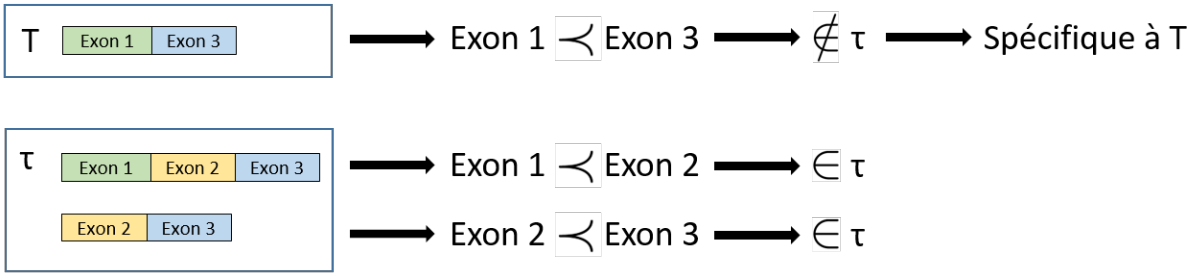


FIGURE 3.6 – Mise en évidence d'une jonction d'exons spécifique. Les jonctions "exon 1 $\prec$ exon 2" et "exon 2 $\prec$ exon 3" appartiennent aux transcrits connus d'un corpus τ. La jonction "exon 1 $\prec$ exon 3" appartient à un transcrit T mais pas aux jonctions de τ, elle est une jonction spécifique de T comparé au corpus τ.

3.3.2.1 Identification des jonctions d'exons spécifiques

Pour obtenir des jonctions d'exons spécifiques, on extrait l'ensemble des jonctions d'exons codants issues des transcrits connus dans Ensembl version 90 (version utilisée pour les prédictions de transcrits, voir section 3.1 page 72) ainsi que l'ensemble des jonctions d'exons des transcrits prédits (voir section 3.2 page 72). Une fois les jonctions d'exons extraites pour les deux ensembles de données sous deux fichiers BED12, on identifie les jonctions d'exons du fichier des prédictions qui ne sont pas dans les jonctions d'exons du fichier Ensembl 90. Pour cela, deux listes Python sont créées, l'une regroupant les

jonctions d'exons connues et la seconde les jonctions d'exons prédites. On identifie pour chaque jonction d'exons prédite si elle est contenue dans la liste des jonctions d'exons connues. Si elle n'est pas dans cette liste, alors on définie cette jonction d'exons comme ayant été prédite et comme spécifique à nos données de prédictions. Par exemple, la Figure 3.7 montre chez le chien les deux transcrits connus issus du fichier Ensembl 90 et le transcrit prédit issu du fichier de prédictions. Le transcrit prédit possède une jonction d'exons spécifique qui n'appartient à aucun des deux transcrits connus d'Ensembl 90. Cette jonction relie le bloc codant G au bloc codant I. La jonction d'exons qui relie le bloc codant G au bloc codant J dans un des transcrits connus n'est pas la même. Le bloc codant J se situe 6 nucléotides en aval du début du bloc codant I. Si un transcrit prédit ne possède aucune jonction d'exons spécifique, nous ne rechercherons pas d'autres preuves biologique, et le transcrit est considéré comme *transcrit possible* pour le gène.

3.3.2.2 Jeux de données de lectures

Pour vérifier les jonctions d'exons spécifiques, nous avons utilisé des jeux de données de lectures couvrant une grande quantité de tissus chez l'humain (Annexe 5.1, D. Wang et al. 2019), la souris (Annexe 5.2, Söllner et al. 2017) et le chien (Annexe 5.3, Le Béguec et al. 2018 ; Wucher et al. 2017 ; Hoeppner et al. 2014). Hormis pour la souris, ces jeux de données étaient disponibles sous le format BAM, un format de fichier de lectures alignées sur leur génome de référence. Les fichiers de lectures de la souris étaient disponibles au format FASTQ, le format de lectures brutes non alignées sur le génome de la souris. Un alignement des lectures a été réalisé avec le logiciel d'alignement STAR, conformément au protocole décrit par les auteurs, pour obtenir les fichiers au format BAM. Chacun des fichiers BAM a été aligné sur le fichier de jonctions d'exons spécifiques avec BedTools (Annexe 4.3).

3.3.2.3 Préparation des données en vue de l'alignement avec des lectures courtes

Selon la taille des exons, on augmente à 65 nucléotides en amont pour le premier exon ou en aval pour le second exon. Si l'exon a une taille inférieur à 65 nucléotides, c'est sa taille complète qui est utilisée. Ce choix résulte du fait que l'on souhaite aligner des lectures courtes qui font généralement une centaine de nucléotides en taille. En attribuant 65 nucléotides de chaque côté de la jonction d'exons, on s'assure de pouvoir aligner des lectures complètes et que le tiers de la lecture au moins chevauche un exon. Les jonctions

FIGURE 3.7 – Exemple d'une jonction d'exons spécifique pour le transcrit prédit *CGACAFTENST*00000257468 visualisée sous le logiciel IGV. (a) Deux transcrits sont connus dans Ensembl 90 (fichier "*dog_ensembl_tot.bed*") et un transcrit *CGACAFTENST*00000257468 a été prédit (fichier *dog_cgalcode.bed*). Le transcrit prédit *CGACAFTENST*00000257468 contient une jonction spécifique (fichier *specific_junction.bed*, en rouge) n'appartenant à aucun des transcrits connus dans Ensembl 90. (b) Cette spécificité est due à la liaison entre le bloc codant G et le bloc codant I. (c) Le bloc codant J est situé à 6 nucléotides en aval du premier nucléotide du bloc codant I. I, J et l'accepteur alternatif appartiennent au même exon.

d'exons sont stockées dans un fichier au format BED12 où chaque ligne correspond à une jonction d'exons. Le format est le même que pour les transcrits (voir Annexe 4.1) mais adapté aux jonctions d'exons. Ainsi, en colonne 10, on aura le chiffre 2 pour indiquer qu'on s'intéresse aux deux exons de la jonction d'exons. En colonne 11, généralement on aura "65, 65" pour indiquer l'extension en amont et en aval et la colonne 12 donnera les positions relatives en considérant cette extension (0, taille du fragment génomique entre les deux extensions sur le génome) (Figure 3.8).

Certaines lectures s'alignent par fragments sur les jonctions d'exons ou sur un seul des deux exons. Un nettoyage est réalisé pour ne conserver que les lectures qui s'alignent au

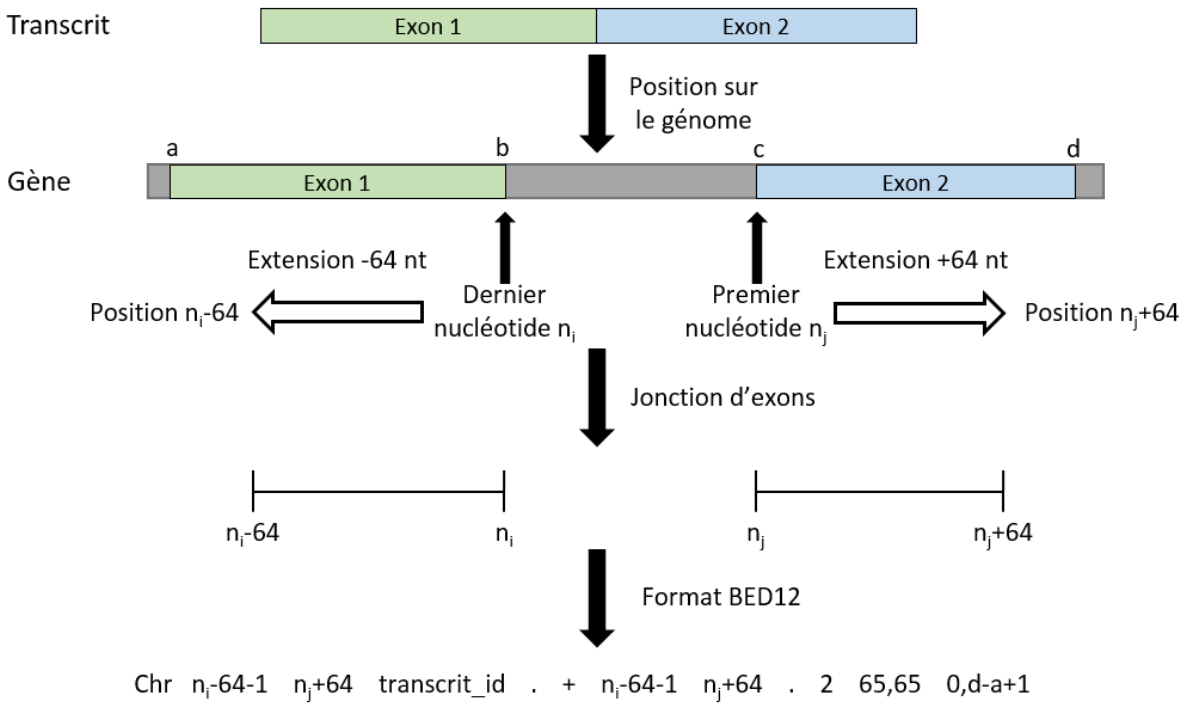


FIGURE 3.8 – Obtention des jonctions d'exons de transcrits au format BED12. La jonction d'exon correspond au dernier nucléotide du premier exon (n_i) et au premier nucléotide du second exon (n_j). Chaque extrémité est augmenté de 64 nucléotides ($n_i - 64$, $n_j + 64$) pour obtenir des jonctions d'exons de 130 nucléotides. Chaque jonction d'exons est stockée au format BED12 en considérant ces extensions.

moins sur les deux exons impliqués dans la jonction d'exons (chevauchement), en vérifiant la correspondance parfaite des coordonnées des lectures avec les coordonnées des exons. On évite ainsi d'avoir des alignements sur un seul des deux exons impliqués de la jonction d'exons. Par exemple, la Figure 3.9 illustre un alignement des lectures issu du fichier de l'échantillon de la glande surrénale chez le chien sur la jonction d'exons spécifique du transcrit prédits
$CGACAFTENST00000257468$.

Une fois cet alignement effectué, on vérifie si toutes les jonctions d'exons spécifiques d'un transcrit prédit ont été couvertes par des lectures. Ainsi, si un transcrit prédit n'a aucune ou seulement certaines jonctions d'exons couvertes par des lectures, alors le transcrit prédit est étiqueté comme *non réalisable* selon les ensembles de données de lectures. A l'inverse, si toutes les jonctions d'exons spécifiques sont couvertes par des lectures, alors le transcrit prédit est étiqueté comme *réalisable*. Ceci met en évidence dans les données de

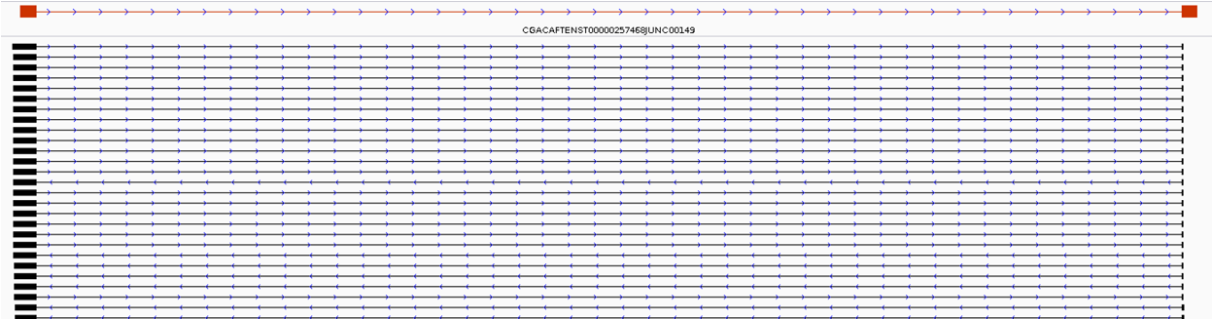


FIGURE 3.9 – Alignement parfait des lectures issues de l'échantillon de glande surrénale chez le chien sur la jonction d'exons spécifique du transcrit prédit $CGACAFTENST00000257468$.

séquençage une "signature" du transcrit prédit nécessaire mais non suffisante pour qualifier le transcrit prédit comme confirmé. L'ensemble du pipeline détaillé est présenté à la Figure 3.10 et dans l'Algorithme 2.

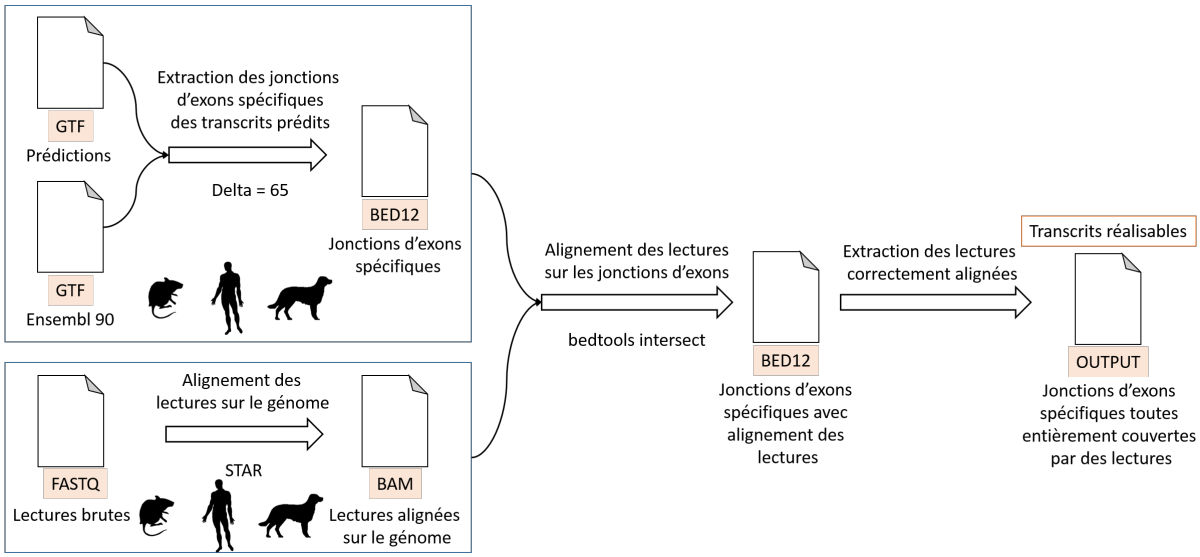


FIGURE 3.10 – Vérification des transcrits prédits réalisables en alignant des lectures courtes sur les jonctions d'exons spécifiques pour l'humain, la souris et chien.

Algorithme 2 Recherche de jonctions d'exons spécifiques dans des données expérimentales

Entrée: CDS prédits non étiquetés "confirmé", données de lectures issues de séquençage
Sortie: CDS étiquetés "possible", "réalisable" ou "non réalisable"

```
pour tout CDS connu faire
    extraire les jonctions d'exons dans une liste
fin pour
pour tout CDS prédit faire
    extraire les jonctions d'exons dans une liste
fin pour
pour toute jonction d'exons de CDS prédit faire
    si jonction d'exons ∉ liste jonctions d'exons de CDS connus alors
        ajouter jonction d'exons dans la liste des jonctions d'exons spécifiques
    fin si
    si CDS prédit sans jonction d'exons spécifique alors
        CDS prédit étiqueté "possible"
    fin si
fin pour
exporter liste jonctions d'exons spécifiques au format BED12
pour toute jonction d'exons spécifique faire
    BedTools intersect avec données de lectures (BAM)
    si jonctions d'exons spécifiques toutes couvertes par des lectures alors
        CDS prédit étiqueté "réalisable"
    sinon {jonctions d'exons spécifiques non toutes couvertes par des lectures}
        CDS prédit étiqueté "non réalisable"
    fin si
fin pour
```

3.3.3 Résultats de la recherche des transcrits prédits dans des bases de données auxiliaires

A partir de la méthode de recherche des transcrits prédits dans des bases de données, nous avons pu obtenir les résultats présentés dans la Table 3.3. Premièrement, 32,5% des transcrits prédits chez le chien sont confirmés avec les bases de données notamment pour Ensembl version 98 et FEELnc. Pour FEELnc, il s'agit de nouvelles données qui ont été obtenues par l'Institut Génétique & Développement de Rennes (IGDR). Pour ce qui concerne l'humain et la souris, 17,5% et 16,4% respectivement des transcrits prédits sont confirmés. Au total, nous avons montré que 1 659 transcrits prédits (269 chez l'humain, 346 chez la souris et 1 044 chez le chien) sur les 6 861, soit 24,2%, sont validés par recherche dans des bases de données auxiliaires. Ayant été obtenus de manière indépendante à notre méthode et à nos données initiales, ils sont annotés *confirmés*.

Espèces	*Humain*	*Souris*	*Chien*	Total
Nombre de transcrits prédits	1 540	2 112	3 209	6 861
Trouvés dans Ensembl 96	162	294	1	
Trouvés dans Ensembl 98	+3	+9	+720	
Trouvés dans Ensembl 102	+5	+4	+0	
Trouvés dans Ensembl 103	+8	-	+0	
Trouvés dans XBSeq	+85	+35	-	
Trouvés dans FEELnc	-	-	+321	
Trouvés dans UCSC	+6	+4	+2	
Nombre total de transcrits prédits retrouvés	269	346	1 044	1 659
dans des bases de données	17,5%	16,4%	32,5%	24,2%

TABLE 3.1 – Vérification des transcrits prédits pour les 2 167 triplets de gènes orthologues à partir de bases de données

3.3.4 Résultats de la recherche des transcrits prédits dans des jeux de données de RNA-seq

3.3.4.1 Résultats obtenus

Les 5 202 transcrits prédits qui n'ont pas été confirmés avec les bases de données ont été confrontés aux données de lectures issues de séquençage RNA-seq. Il s'agit là de rechercher des "signatures" des transcrits prédits dans les lectures. Plus précisément, si un transcrit possède des jonctions d'exons $e_i \prec e_j$ qui lui sont spécifiques, le fait que

les lectures témoignent de l'existence de ces enchaînement d'exons $e_i \prec e_j$ supporte le transcrit prédit. Avec les alignements de lectures disponibles, nous obtenons les résultats présentés dans la Table 3.2. On retrouve en moyenne 1,3 jonctions d'exons spécifiques par transcrit lorsque les transcrits prédits en possèdent (3761/2897, Table 3.2). Si toutes les jonctions d'exons spécifiques d'un transcrit prédit sont couvertes par des lectures, le CDS du transcrit est dit *réalisable*. Sur les 5 202 transcrits prédits analysés, 2 897 (55,7%) contiennent au moins une jonction d'exons spécifique dont 1 732 (59,8%), sont *réalisables*. Un exemple de résultat d'alignement est présenté en Figure 3.11 où on observe le nombre de lectures qui s'alignent sur des jonctions d'exons spécifiques selon les tissus considérés. Les transcrits prédits dans lesquels toutes les jonctions d'exons spécifiques n'ont pas été couvertes par des lectures sont qualifiés de *non réalisables* sachant les jeux de données de lectures utilisés. Enfin, les transcrits prédits qui ne contiennent aucune jonction d'exons spécifique sont qualifiés de *possibles*. L'ensemble de leurs jonctions d'exons étant déjà connu, l'enchaînement des exons du transcrit prédit est possible.

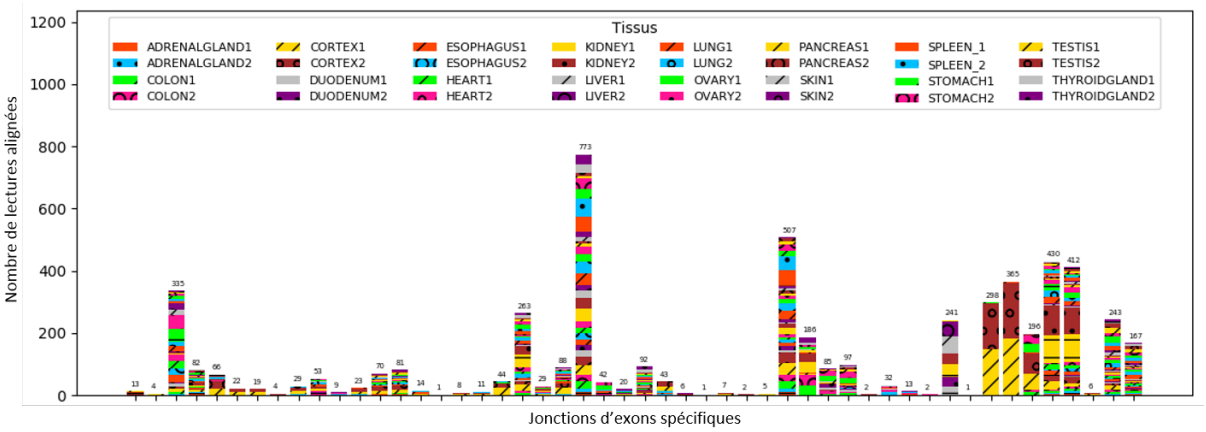


FIGURE 3.11 – Alignement de lectures sur des jonctions d'exons spécifiques. Chaque colonne en abscisse représente une jonction d'exons spécifique et les ordonnées donnent le nombre de lectures alignées. Cet exemple montre un extrait de 50 jonctions d'exons spécifiques de transcrits prédits chez l'humain.

3.3.4.2 Couverture des jonctions d'exons

Nous cherchons maintenant à déterminer quelles couvertures des jonctions d'exons sont attendues, c'est-à-dire le nombre de lectures de séquençage alignées sur une jonction d'exons spécifique. Pour cela, nous avons réalisé un alignement des lectures sur les jonctions d'exons non spécifiques (jonctions d'exons appartenant à des transcrits connus

Espèces	*Humain*	*Souris*	*Chien*	Total
Transcrits prédits non trouvés dans les bases de données	1 271	1 766	2 165	5 202
Transcrits prédits sans jonctions d'exons spécifiques	547	751	1 007	2 305
Nombre de jonctions d'exons spécifiques	939	1 397	1 425	3 761
Transcrits prédits avec des jonctions d'exons spécifiques	724	1 015	1 158	2 897
Sans lectures alignées (aucune jonction spécifique retrouvée)	292	468	289	1 049
Avec et sans lectures alignées (seulement certaines jonctions spécifiques retrouvées)	33	49	34	116
Avec lectures alignées (toutes les jonctions spécifiques retrouvées)	**399**	**498**	**835**	**1 732**
Transcrits trouvés avec des lectures alignées	55,1%	49,1%	72,1%	59,8%

TABLE 3.2 – Vérification des transcrits prédits pour les 2 167 triplets de gènes orthologues avec des jonctions d'exons spécifiques trouvées dans des données de lectures issues de séquençage.

de la version 90 d'Ensembl). Nous avons testées les jonctions d'exons des transcrits pour un sous-ensemble de 253 triplets de gènes parmi les 2 167 dont les caractéristiques et l'obtention seront l'objet du prochain chapitre.

Pour cet ensemble de données, nous avons extrait 3 407, 3 350 et 3 125 jonctions d'exons pour l'humain, la souris et le chien réciproquement. Nous avons aligné des lectures sur 3 387, 3 337 et 3 066 jonctions d'exons chez ces trois espèces. Ces jeux de données de lectures couvrent donc 99,4%, 99,6% et 98,1% des jonctions d'exons appartenant aux transcrits connus. Ainsi, les jeux de données sont assez complets, sans l'être totalement. En terme de couverture, on retrouve, comme pour les jonctions d'exons spécifiques, des jonctions d'exons peu couvertes (moins de 10 lectures) quelque soit l'espèce (Figure 3.12). On peut donc considérer que des jonctions d'exons spécifiques prédites ayant une couverture de moins de 10 lectures sont supportées par les données de séquençage.

3.3.5 Résultats des deux méthodes sur les données complètes

Grâce à ces deux méthodes, nous avons pu confirmer 1 659 (24,2%) des 6 861 transcrits prédits grâce aux données de bases de données auxiliaires. 2 305 (33,6%) des transcrits prédits sont considérés comme possibles du fait que leurs enchaînements d'exons soient déjà connus. 1 732 (25,2%) des transcrits prédits ont toutes leurs jonctions d'exons spécifiques couvertes par des lectures issues de données de RNA-seq et sont réalisables sachant ces données de séquençage. Ainsi, 3 391 (49,4%) transcrits prédits ont soit été retrouvés dans les bases de données soit considérés comme réalisables par des jonctions d'exons retrouvées dans des données de lectures (Figure 3.13).

Des résultats similaires sont obtenus pour le sous-ensemble de 253 gènes étudiés au chapitre suivant. Parmi les 1 029 transcrits prédits, 34% sont retrouvés dans les bases

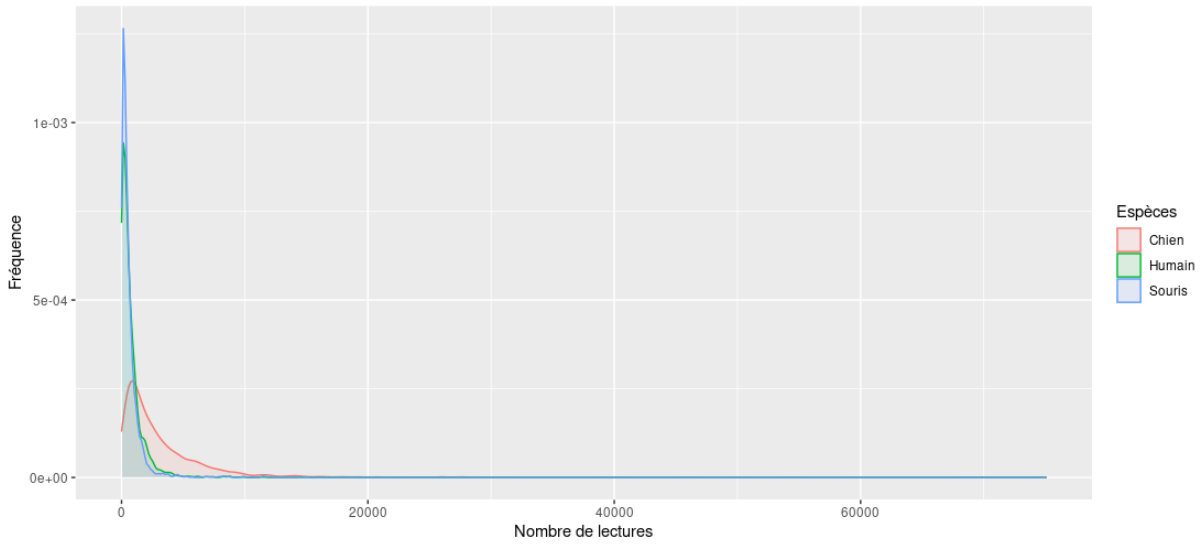


FIGURE 3.12 – Fréquence de la couverture en lectures des jonctions d'exons connues.

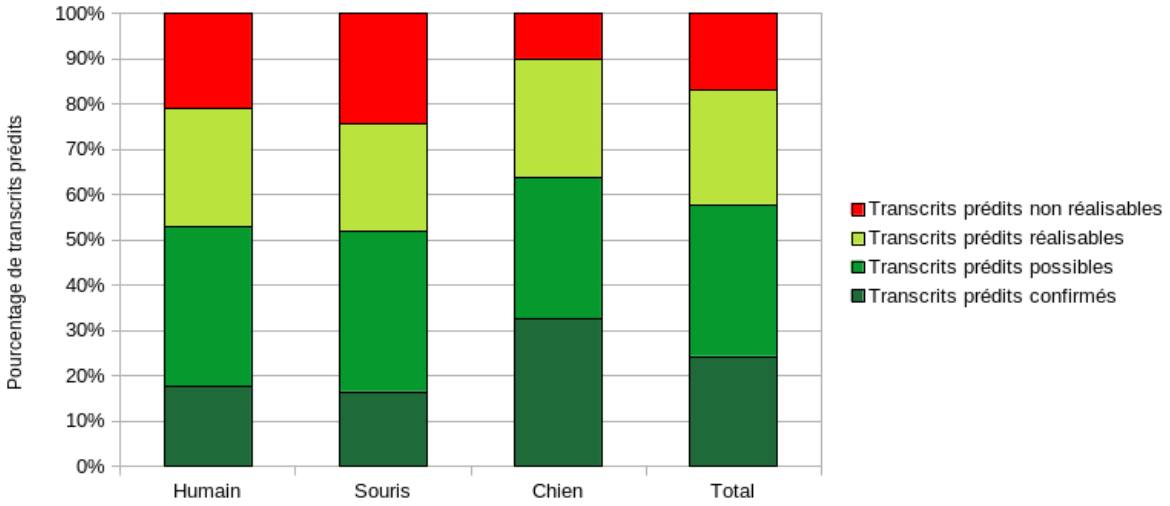


FIGURE 3.13 – Résultats de la confirmation expérimentale de l'existence des transcrits prédits. Un transcrit prédit confirmé est un transcrit prédit présent dans une base de données autre que celle considérée pour établir les prédictions ; un transcrit réalisable est un transcrit prédit dont toutes ses jonctions d'exons spécifiques sont supportées par des lectures issues de séquençage ; un transcrit possible est un transcrit prédit qui ne possède aucune jonction d'exons spécifique et un transcrit non réalisable est un transcrit prédit dont au moins une jonction d'exons spécifique n'est pas supportée par des lectures.

de données auxiliaires (Table 3.3). Parmi les 679 prédictions restantes, 394 transcrits possèdent des jonctions d'exons spécifiques et pour 225 d'entre eux (64%) des lectures sont

alignées avec toutes les jonctions d'exons spécifiques (Table 3.4). Ces résultats indiquent que nos prédictions sont soutenues par les données expérimentales considérées.

Espèces	*Humain*	*Souris*	*Chien*
Nombre de transcrits prédits	180	249	600
Trouvé dans les données Ensembl 96	27	43	0
Trouvé dans les données Ensembl 98	+2	+1	+183
Trouvé dans les données Ensembl 102	+0	+1	+0
Trouvé dans les données Ensembl 103	+3	-	+0
Trouvé dans les données XBSeq	+8	+5	-
Trouvé dans les données FEELnc	-	-	+74
Trouvé dans les données UCSC	+2	+1	+0
Nombre total de transcrits trouvés	42 23.3%	51 20.5%	257 42.8%

Table 3.3 – Vérification des transcrits prédits pour les 253 triplets de gènes orthologues à partir de bases de données

Espèces	*Humain*	*Souris*	*Chien*
Transcrits prédits non trouvés dans les bases de données	138	198	343
Transcrits prédits sans jonctions d'exons spécifiques	59	78	148
Transcrits prédits avec des jonctions d'exons spécifiques	79	120	195
Sans lectures alignées	32	59	48
Avec lectures alignées	47	61	147
Transcrits trouvés avec des lectures alignées	59.5%	50.8%	75.4%

Table 3.4 – Vérification des transcrits prédits pour les 253 triplets de gènes orthologues avec des jonctions d'exons spécifiques trouvées dans des données de lectures

Conclusion du chapitre

Dans ce chapitre, nous avons appliqué une méthode de comparaison de gènes sur un ensemble de 2 167 triplets de gènes orthologues de l'humain, la souris et le chien pour prédire des transcrits orthologues. Nous avons pu prédire au total 6 861 CDS avec une base de 18 109 transcrits connus chez ces trois espèces. A partir de notre méthode de recherche de traces de nos transcrits prédits dans des bases de données et dans des jeux de données de lectures issues de séquençage, nous avons pu donner du crédit à 3 391 d'entre eux, soit 49,4%.

Chapitre 4

Identification d'un ensemble de gènes de structure d'épissage conservés entre l'humain, la souris et le chien

Le chapitre précédent présente la prédiction de transcrits chez l'humain, la souris et le chien et l'analyse de fiabilité des prédictions. Dans ce chapitre, nous traitons de la comparaison multiple de gène sur les trois espèces à la recherche de gènes pouvant exprimer exactement le même ensemble de protéines chez les trois espèces. Pour cela, nous avons recherché des gènes ayant une structure d'épissage strictement conservée : mêmes sites fonctionnels, mêmes exons. Ces gènes orthologues conservés peuvent ainsi produire le même ensemble de CDS : chaque CDS observé chez une espèce est potentiellement réalisable chez les deux autres espèces, ce qui constitue un groupe de CDS orthologues (contenant des transcrits connus et prédits). Cet ensemble de gènes strictement conservés peut ensuite permettre d'observer des variations plus fines de séquences, notamment des variations de la structure des régions non traduites des transcrits : les UTR (*untranslated region*). Ces variations résultent du phénomène de transcription alternative.

Sommaire

4.1 Comparaison trois-espèces 96

4.1.1 Relations d'orthologie trois espèces 96

4.1.2 Graphes de sites fonctionnels : interprétation phylogénétique de la conservation des sites chez trois espèces 96

4.1.3 Graphes de transcrits : identification des groupes de CDS orthologues chez trois espèces 98

4.2 Construction des graphes à partir des données 100

4.2.1 Hypothèses des comparaisons de paires de gènes 100

4.2.2 Hypothèses des graphes de sites fonctionnels 102

4.2.3 Hypothèses des graphes de transcrits 103

4.3 Exemple du gène CREM et illustration des divergences . . . 106

4.4 Application des graphes à la détection de structures strictement conservées 110

4.4.1 253 gènes de structure d'épissage conservée 111

4.4.2 Transcriptomes des gènes structurellement conservés 112

4.4.3 Base de données de 253 triplets de gènes structurellement conservés 118

Définitions des termes du chapitre

La *structure d'un gène* est la succession des blocs codants et des sites fonctionnels qui le composent, permettant d'identifier la composition en introns et en exons du gène en fonction de l'ensemble des CDS exprimés.
Les *sites fonctionnels* correspondent aux codons *start* ("[") et *stop* ("]") et aux sites donneurs ("<") et accepteurs d'épissage (">") formant le premier et le dernier dinucléotides d'un intron.
Un *CDS* est la séquence exonique d'un ARNm qui peut être traduite en protéine, depuis un codon *start* jusqu'à un codon *stop*, séparés par un nombre entier de codons consécutifs et sans codon *stop* intermédiaire en phase.
Un *caractère orthologue* est un caractère commun à deux espèces en copie unique, issu d'un évènement de spéciation et hérité de l'ancêtre commun à ces deux espèces. On applique ici ce terme à trois types de caractères : au niveau du gène, du transcrit et du site fonctionnel. Deux sites fonctionnels sont orthologues si ils sont alignés. Deux transcrits sont orthologues si tous les sites fonctionnels et les blocs codants des deux CDS sont conservés.
Un *gène de structure d'épissage conservé* est composé d'un ensemble de sites fonctionnels partagés entre les trois espèces, dits orthologues. On parle de conservation stricte : aucun site fonctionnel n'a divergé.
Une *classe d'équivalence* du graphe des transcrits est constituée d'un ensemble de transcrits qui, chez une espèce, partagent un même CDS.
Un *groupe de CDS orthologues* est un ensemble de transcrits issus d'espèces différentes et ayant la même structure de CDS. C'est une composante connexe des graphes de transcrits.
Un *groupe de CDS orthologues complet* est un groupe contenant un CDS représentant pour chacune des trois espèces. Dans le cas trois espèces, il s'agit donc d'un triplet.

4.1 Comparaison trois-espèces

4.1.1 Relations d'orthologie trois espèces

Dans le Chapitre 2, nous avons décrit des comparaisons de paires de gènes basées sur la modélisation de la structure d'épissage des gènes à partir de leurs transcrits connus. Pour rappel, cette modélisation permet de révéler, grâce aux alignements de séquences, les *relations d'orthologie* qui relient les blocs exoniques et les sites fonctionnels (codons *start* et *stop*, sites donneurs et accepteurs d'épissage) entre les modèles de structure de deux gènes orthologues. Ainsi, pour chaque site fonctionnel présent chez une espèce, on est capable de dire si il possède un orthologue chez l'autre espèce ou si il est spécifique à cette espèce.

Les comparaisons de paires de gènes permettent ensuite d'associer des paires de CDS orthologues entre deux espèces. Si deux transcrits entre deux espèces partagent une même structure d'épissage décrite par un même modèle de transcrit, ils sont considérés comme orthologues : ils appartiennent au même groupe de CDS orthologues. On peut noter qu'il pourrait exister des CDS orthologues ayant divergé : quelques sites fonctionnels auraient été modifiés durant l'évolution. Dans le cadre de cette thèse, nous nous somme basé sur une définition de conservation stricte des transcrits orthologues.

Nous avons étendu la comparaison à une échelle multi-espèces (voir section 2.3 page 66) en utilisant les informations obtenues des comparaisons de paires de gènes pour construire des graphes. Nous appliquons ici la méthode à une comparaison de trois espèces comme première étape à une application multi-espèces (Figure 4.1).

Pour cela, deux types de graphes sont construits : les graphes de sites fonctionnels et les graphes de transcrits. Le pipeline de construction des graphes est résumé à la Figure 4.2. Il rappelle comment la comparaison de paires de gènes, nécessaire pour prédire les relations d'orthologies entre sites fonctionnels et entre transcrits, est utilisée pour construire les graphes.

4.1.2 Graphes de sites fonctionnels : interprétation phylogénétique de la conservation des sites chez trois espèces

Étant donné des gènes orthologues et leurs transcrits connus sur trois espèces, des relations d'orthologie sont identifiés pour les sites fonctionnels grâce aux comparaisons de paires de gènes. Des graphes de sites fonctionnels sont alors construits (voir section

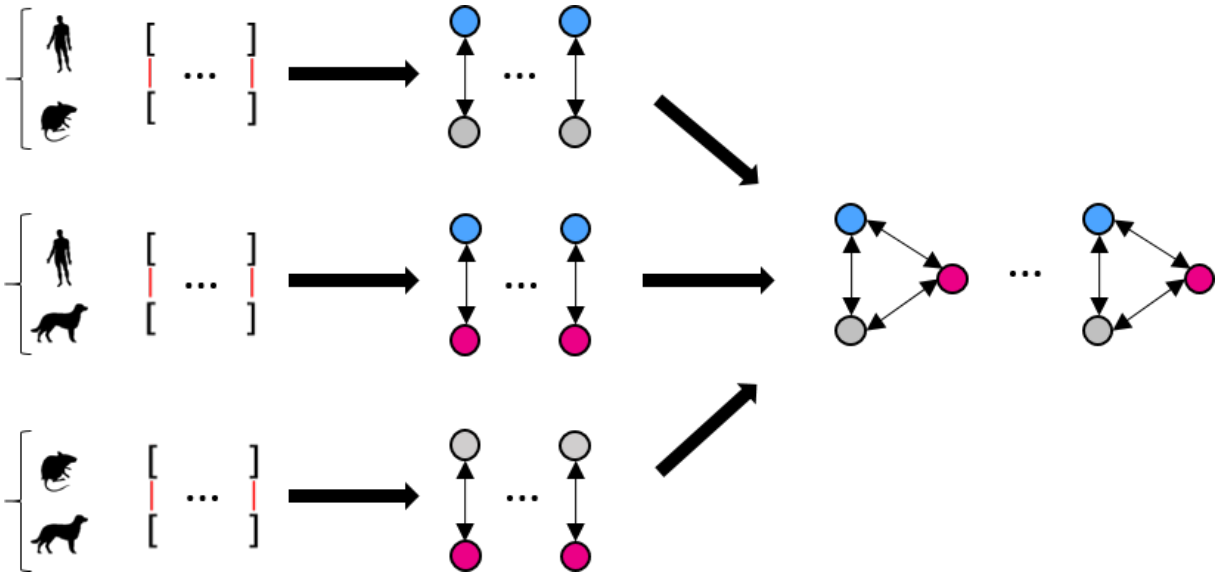


FIGURE 4.1 – Passage des alignements de paires de gènes à un graphe multi-espèces. L'alignement de deux sites fonctionnels entre deux espèces est représenté par un trait rouge. Chaque rond coloré est un nœud du graphe de sites et chaque arête est une relation d'orthologie identifiée.

2.3.1 page 66 et Annexe 6) pour établir la comparaison des trois espèces. On est ainsi capable d'identifier si un site fonctionnel est partagé entre plusieurs espèces (triplet ou couple de sites fonctionnels dans le graphe) ou bien si il est spécifique à une seule espèce (singleton dans le graphe). Ainsi, ces graphes peuvent donner lieu à une interprétation phylogénétique. Si un site fonctionnel est retrouvé chez les trois espèces, cela suggère qu'il était possiblement présent dès leur ancêtre commun (Figure 4.3a). Un site non observé chez toutes les espèces sera interprété différemment selon les relations phylogénétiques entre les espèces. De cette façon, on considère qu'il a été perdu chez l'humain ou la souris lorsqu'il est absent chez ces espèces. Alors qu'on peut supposer qu'il est apparu chez l'ancêtre de l'humain et de la souris lorsqu'il est absent chez le chien (Figure 4.3b). Enfin, si un site fonctionnel n'est présent que chez une espèce, on peut supposer qu'il est apparu chez cette espèce ou bien qu'il aurait également pu être perdu chez l'ancêtre commun de l'humain et de la souris lorsqu'il n'est présent que chez le chien (Figure 4.3c).

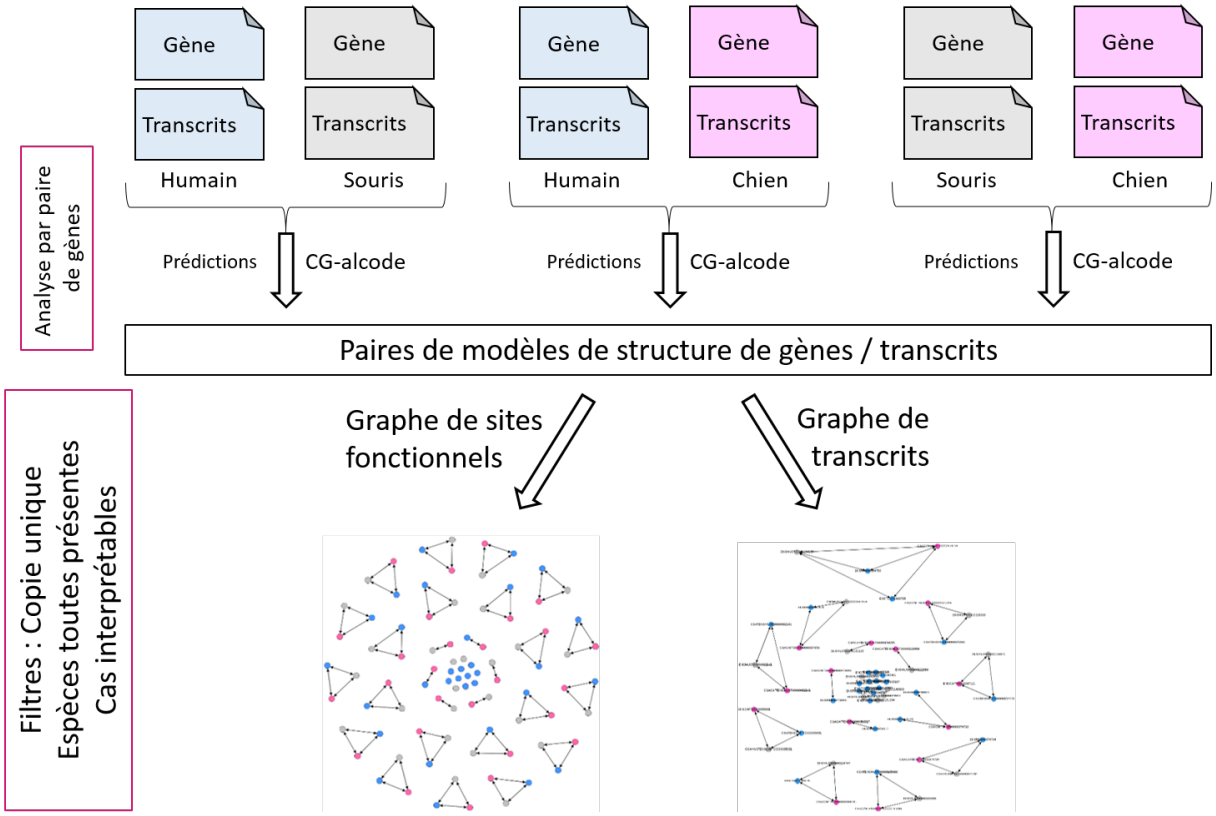


FIGURE 4.2 – Pipeline de construction des graphes de sites fonctionnels et de transcrits pour un triplet de gènes orthologues

4.1.3 Graphes de transcrits : identification des groupes de CDS orthologues chez trois espèces

Sur les trois espèces, un graphe de transcrits est construit. Il relie les transcrits orthologues (connus et prédits) ayant la même structure d'épissage (voir section 2.3.2 page 66). On peut, comme pour les sites fonctionnels, interpréter ce graphe en termes phylogénétiques. Ainsi, si un transcrit possède une même structure d'épissage du CDS conservée chez les trois espèces, alors ce transcrit a pu être exprimé depuis le gène de l'ancêtre commun aux trois espèces (Figure 4.3a). Au contraire, si ce transcrit n'est pas retrouvé chez les trois espèces, alors soit il n'est plus exprimable par le gène (chez l'humain et la souris) ou n'a jamais existé (chez le chien) (Figure 4.3b) soit il est exprimé spécifiquement chez une espèce et n'existait pas chez les ancêtres communs *Boreoeutheria* ou *Euarchontoglires* (Figure 4.3c).

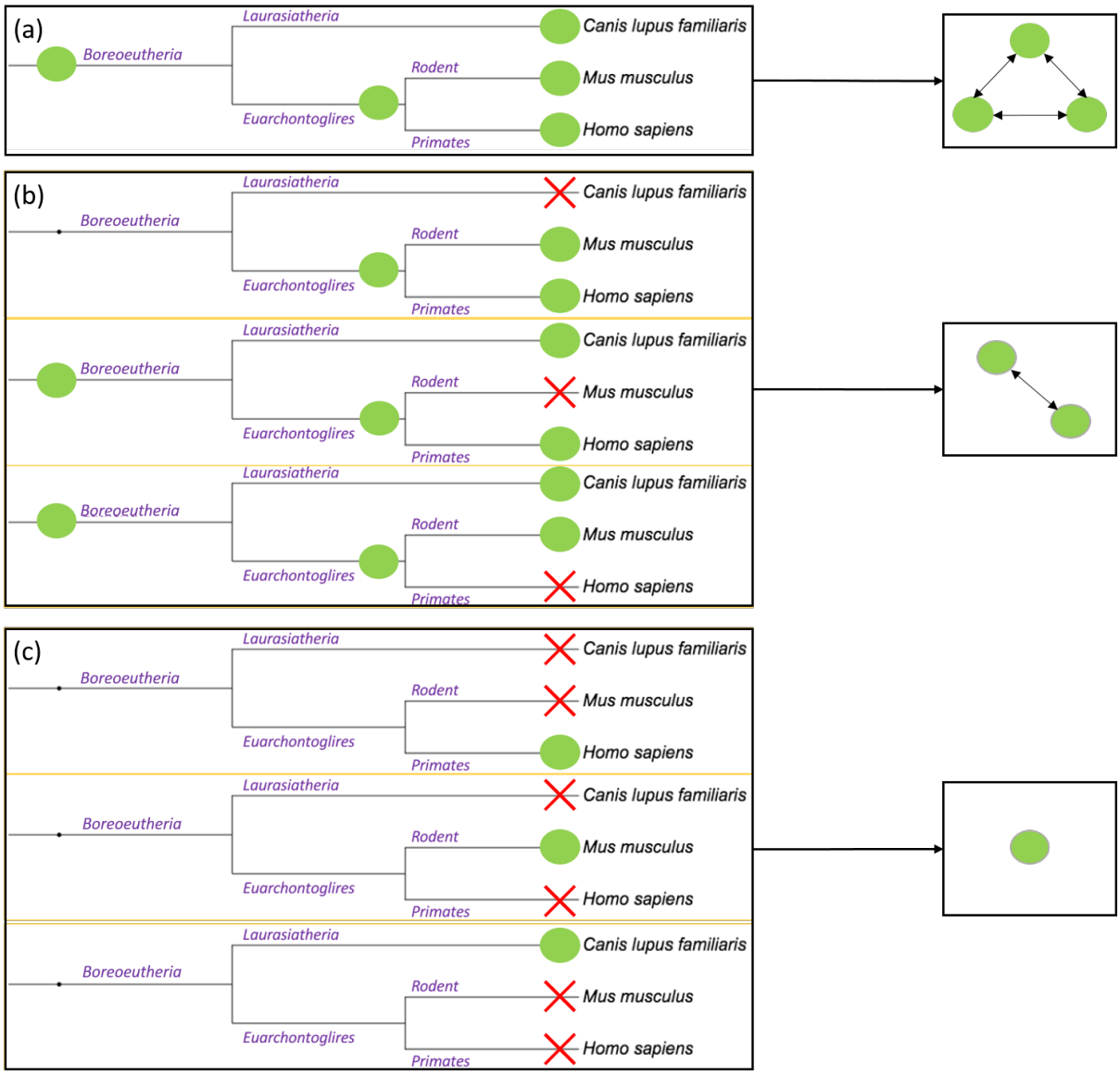


FIGURE 4.3 – Interprétations phylogénétiques des graphes de sites fonctionnels et de transcrits sur l'humain, la souris et le chien. Un rond vert correspond à la présence d'un caractère chez une espèce ou un ancêtre commun : *Boreoeutheria* et *Euarchontoglires*. Une croix rouge correspond à la perte d'un caractère chez une espèce. (a) Cas d'un caractère présent chez les trois espèces depuis leur ancêtre commun *Boreoeutheria*, représenté par un triplet dans le graphe. (b) Cas d'un caractère absent chez une espèce, représenté par un couple dans le graphe. (c) Cas d'un caractère présent chez une seule espèce, représenté par un singleton dans le graphe.

4.2 Construction des graphes à partir des données

Notre approche est sélective, un certain nombre de filtres sont appliqués pour sélectionner les gènes à analyser. Étant donné un ensemble de gènes sur k espèces, la première série de contraintes appliquées vise à sélectionner des gènes sur lesquels on dispose d'orthologie "un-à-un" dans les espèces considérées et possédant suffisamment de transcrits connus pour que l'expression alternative soit avérée (voir section 3.1 page 72 et section 4.2.1). La deuxième série de contraintes, qui fait l'objet de la section 4.4 (page 110), a pour but d'identifier un ensemble de gènes ayant la même structure d'épissage, conservée chez chacune des trois espèces : humain, souris et chien. Chacun de ces gènes ayant les mêmes exons et les mêmes sites fonctionnels partagés par les trois espèces, il est ainsi supposé produire le même ensemble de CDS. Autrement dit, chaque CDS réalisé sur une des espèces est réalisable par les deux autres espèces et appartient à un groupe de CDS orthologue complet, de taille de trois classes d'équivalence.

La Figure 4.1 montre un graphe de sites fonctionnels attendu lors de cette sélection. Au départ, les trois alignements de paires de gènes permettent d'obtenir des relations d'orthologie entre sites fonctionnels par paires d'espèces. La combinaison de toutes les relations d'orthologie permet d'obtenir un graphe, dont les nœuds sont des sites fonctionnels et les arêtes des relations d'orthologie. Sur l'exemple, toutes les composantes connexes sont sous forme de triplets, indiquant que tous les sites fonctionnels sont partagés chez les trois espèces. Ce type de graphe sélectionné est aussi applicable au graphe des transcrits mais avec des conditions précisées dans les paragraphes suivants.

4.2.1 Hypothèses des comparaisons de paires de gènes

Plusieurs hypothèses sont posées pour réaliser les comparaisons de paires de gènes : la colinéarité des gènes, la bijection des relations d'orthologie entre sites fonctionnels et les transcrits.

Colinéarité des sites orthologues des gènes. Pour pouvoir construire des graphes, la méthode de comparaison de paires de gènes suppose la colinéarité des exons des gènes orthologues pour produire des modèles de structure de gènes et de transcrits. L'hypothèse est que l'ordre des exons d'un gène est conservé chez ses orthologues. Si l'ordre n'est pas conservé entre les deux gènes orthologues, alors on dit qu'il n'y pas de conservation de la colinéarité du gène chez les deux espèces et les gènes orthologues ne sont pas comparés.

Par exemple, dans la Figure 4.4, l'ordre des exons gris et bleu n'est pas le même entre le gène humain et le gène de la souris, en haut de la figure. Au contraire, l'ordre est identique sur ceux du bas. De ce fait, les deux gènes orthologues en haut de la figure ne sont pas comparés et ne permettront pas de produire des modèles d'alignement de gènes. Ces gènes sont donc éliminés en première phase de la comparaison de paires de gènes.

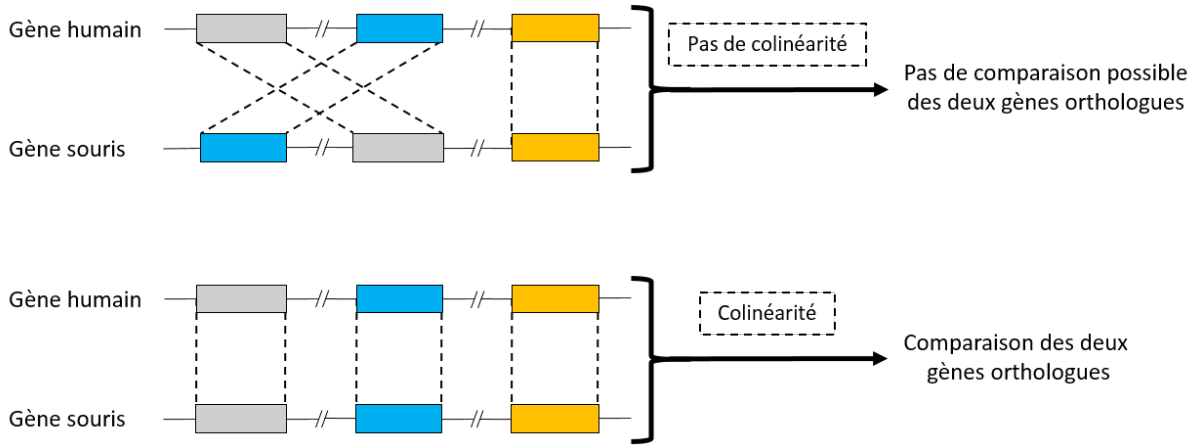


FIGURE 4.4 – Hypothèse pour la comparaison de paires de gènes : conservation de la colinéarité des exons des gènes

Bijection des relations d'orthologie entre sites. Seuls les gènes disposant d'exons en copie unique (exons non répétés, dupliqués, au sein d'un même gène) dans un même gène sont considérés. Ainsi, chaque comparaison de paires de gènes aligne chaque site fonctionnel d'un gène avec un site fonctionnel d'un gène orthologue, et réciproquement. Par exemple, dans la Figure 4.5, le gène humain et son orthologue chez la souris possèdent 6 sites fonctionnels. Chaque site fonctionnel est aligné de manière bijective (un site humain aligné avec un site souris et ce même site souris aligné avec le même site humain).

Filtrage des transcrits connus. Certains transcrits connus ne sont pas considérés si leur CDS ne respecte pas certains standards : c'est le cas si un des sites fonctionnels ne respecte pas les contraintes suivantes : ATG en codon *start* et TAA, TAG ou TGA en codon *stop*. Par exemple, le transcrit $ENSCAFT000000005928$ du gène CREM chez le chien a son CDS débutant par GAC. De ce fait, ce transcrit ne sera analysé.

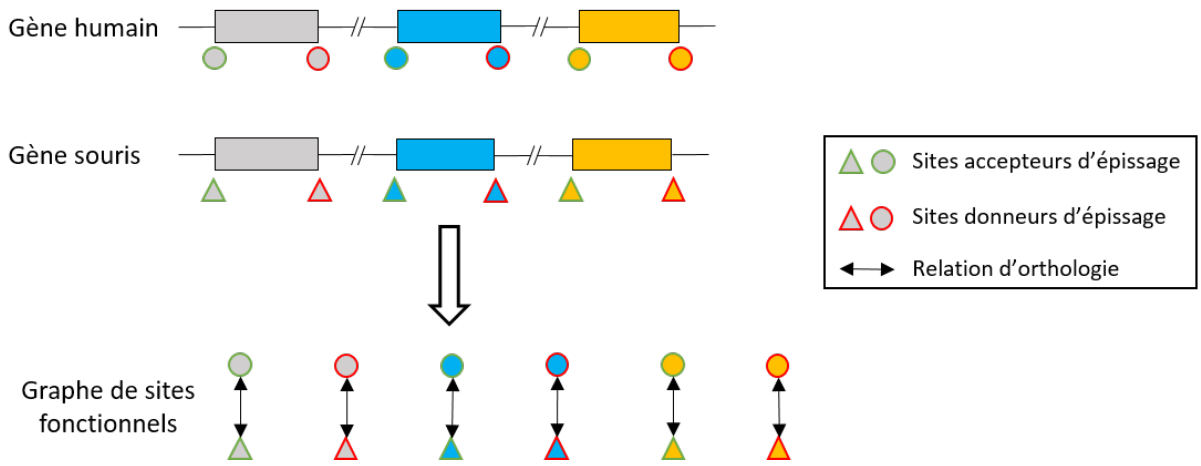


FIGURE 4.5 – Bijection des relations d'orthologie pour la construction de graphes de sites fonctionnels

4.2.2 Hypothèses des graphes de sites fonctionnels

Pour rappel, un graphe de sites fonctionnels représente l'orthologie des sites fonctionnels d'un triplet de gènes orthologues.

Sites fonctionnels utilisés. Si un gène a un seul transcrit ou peu de transcrits et qu'ils sont non valides (mauvais codon *start*/*stop*, nombre multiple de codons non respecté, codon *stop* prématuré), alors aucun modèle de structure de transcrits n'est construit pour l'espèce. Seuls les sites fonctionnels appartenant à au moins un modèle de CDS connu ou prédit sont utilisés pour construire les graphes de sites fonctionnels. Ainsi, des sites fonctionnels prédits dans un modèle de gène, mais n'appartenant à aucun CDS connu ou prédit valide, ne seront pas représentés dans le graphe. Par conséquence, si aucun transcrit n'est valide pour une espèce, l'espèce sera absente du graphe obtenu. Au final, nous sélectionnons les graphes ayant toutes les espèces représentées. Dans notre cas, les trois espèces d'étude humain, souris et chien doivent être présentes dans chaque graphe.

Anomalies de relations d'orthologie. Lors des multiples alignements de paires de gènes, il arrive que certains sites fonctionnels soient alignés entre certaines paires d'espèces mais pas avec d'autres paires d'espèces. Par exemple, dans la Figure 4.6, l'accepteur d'épissage est aligné entre l'humain et la souris et entre la souris et le chien mais n'est pas aligné entre l'humain et le chien du fait du comportement indépendant de l'alignement BLAST entre les comparaisons de paires de gènes. Par exemple, l'alignement de meilleur

score ne place pas les nucléotides "AG" du chien en face des nucléotides "AG" de l'humain. On peut avoir "AG" aligné avec "$A-G$" (où "$-$" représente un *gap*). Les graphes de sites fonctionnels qui présentent ce genre de cas ne seront pas sélectionnés pour rechercher des gènes conservés entre les trois espèces.

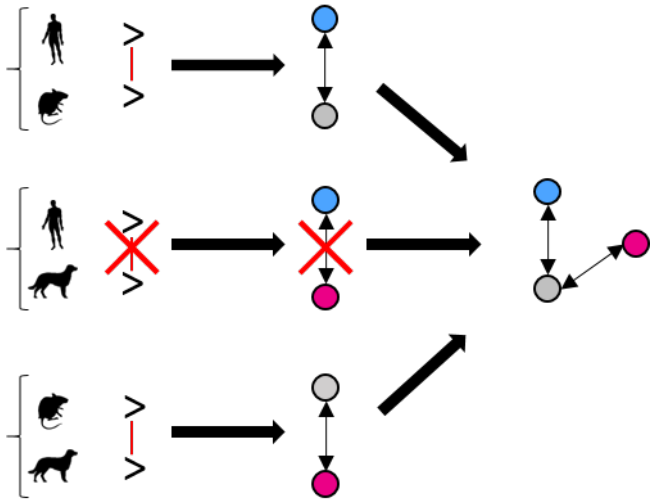


FIGURE 4.6 – Cas de sites fonctionnels non alignés entre certaines espèces.

Interprétation phylogénétique sur les graphes de sites fonctionnels. Au final, trois types de composantes connexes sont considérés dans un graphe de sites fonctionnels $\mathcal{G}^{FS}$: un triplet (Figure 4.7a) représente un site fonctionnel partagé par les trois espèces, un couple (Figure 4.7b) représente un site fonctionnel partagé par deux espèces, un singleton (Figure 4.7c) correspond à un site fonctionnel présent chez une seule espèce. Seuls les triplets de gènes orthologues dont les graphes sont constitués exclusivement de composantes connexes en forme de singletons, de couples et/ou de triplets sont pris en compte pour l'analyse dans le cas des graphes de sites fonctionnels. Les autres cas, expliqués précédemment, ne sont pas interprétés dans la suite des analyses (Figure 4.7d).

L'Algorithme 3 présente l'idée générale de construction du graphe de sites fonctionnels.

4.2.3 Hypothèses des graphes de transcrits

Pour rappel, un graphe de transcrits concerne un triplet de gènes orthologues. Il relie les transcrits du gène ayant la même structure de CDS.

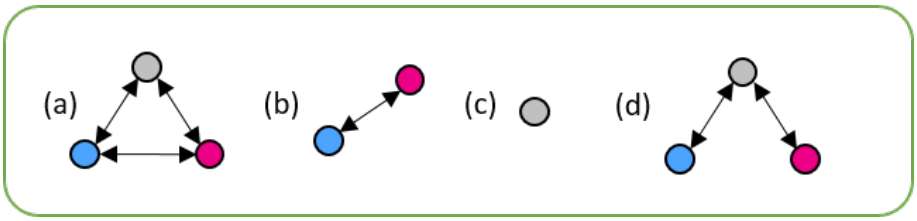


FIGURE 4.7 – Composants présents dans les graphes de sites fonctionnels. Les composants considérés sont : (a) triplets, (b) doublons, (c) singletons. Les composants ambigus comme (d) ne sont pas considérés.

Algorithme 3 Construction d'un graphe de sites fonctionnels pour un gène partagé par k espèces

Entrée: alignements des $k \times (k-1)$ paires de gènes via CG-alcode
Sortie: graphe de sites fonctionnels G^{SF}
  pour toutes les $k \times (k-1)$ paires de gènes $[M_i^G, M_j^G]$ **faire**
    pour tout site $\in M_i^G$ **faire**
      si site $\in$ CDS valide **alors**
        ajouter le site à l'ensemble des nœuds de G^{SF}
        si site $\in M_i^G$ aligné à un site $\in M_j^G$ **alors**
          si site de $M_j^G \in$ CDS valide **alors**
            ajouter le site de M_j^G à l'ensemble des nœuds de G^{SF}
            ajouter l'arête entre site de M_i^G et site de M_j^G
          fin si
        fin si
      fin si
    fin pour
  fin pour

Relation d'orthologie 1-à-1, 1-à-plusieurs. Dans ce cadre, on s'intéresse à deux types d'orthologie. Dans l'orthologie "$1-$à-1" (*1-to-1*), chaque transcrit en correspondance est liée par une relation d'orthologie bijective. Dans l'orthologie "$1-$à$-plusieurs$" ($1-to-many$), il y a plus d'un transcrit orthologue associé. Dans ce second cas, on n'est pas en mesure de déterminer quelle paire de transcrits sont orthologues sur la base de la région codante seule.

Interprétation phylogénétique sur les graphes de transcrits. On cherche à déterminer des CDS orthologues. On considère donc, comme pour les sites fonctionnels, les graphes ayant des cas d'interprétation phylogénétiques simples. Ainsi, chaque composante connexe doit avoir la forme de singleton, de couple ou de triplets de CDS (Figure 4.7).

Cependant, comme on vient de l'évoquer, il existe aussi des transcrits différents encodant un même CDS (relations d'orthologie "1-à-plusieurs"). Les graphes qui les contiennent sont alors plus complexes et seront considérés dans un second temps (Figure 4.8d).

Il est à noter que chacune des composantes connexes d'un graphe de transcrits correspond à ce qu'on définit comme un *groupe de CDS orthologues* (voir section 2.2.2 page 62). Les cas de triplets sont ceux qui correspondent à la conservation trois espèces du CDS du transcrit.

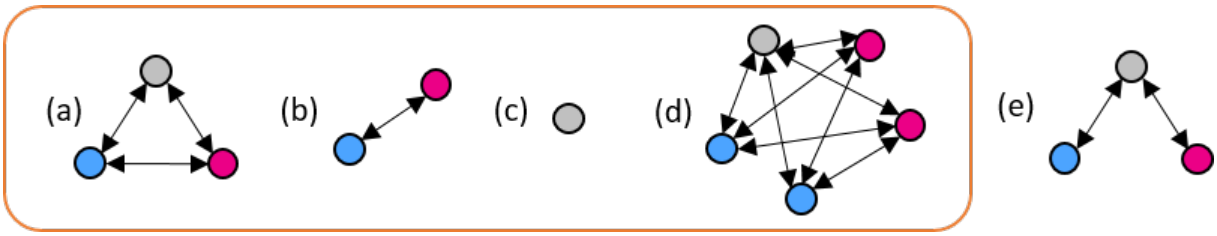


FIGURE 4.8 – Composants présents dans les graphes de transcrits. Les composants considérés sont : (a) triplets, (b) doublons, (c) singletons dans les graphes de transcrits, (d) composantes connexes avec des transcrits encodant un même CDS (transcription alternative). On appellera les CDS, présents plusieurs fois dans ces triplets (d), des CDS en copies multiples. Ceux-ci seront étudiés en Chapitre 5. Les composants ambigus comme (e) ne sont pas considérées.

Un graphe ne contenant que des triplets de sites fonctionnels (Figure 4.7a) ou de transcrits (Figure 4.8a) indique que chaque site fonctionnel ou transcrit possède un orthologue dans chacune des autres espèces. Dans le cas des graphes de sites fonctionnels $\mathcal{G}^{SF}$, un tel gène possède un modèle de structure de gène partagé par les trois espèces définissant un triplet de *gènes structurellement orthologues*. Un graphe de transcrits $\mathcal{G}^{T}$ contenant uniquement des triplets de transcrits implique que chaque CDS possède un CDS orthologue dans chacune des autres espèces. Cela indique que chacun des trois gènes orthologues peut exprimer le même ensemble de CDS. Toutes ses composantes connexes sont des groupes de CDS orthologues complets. Chaque CDS est en copie simple dans ce cas. Dans le cas des graphes de transcrits contenant des CDS en copies multiples, les groupes de CDS orthologues peuvent contenir plusieurs fois le même CDS chez la même espèce. On définit alors une *classe d'équivalence* pour chaque CDS de chaque espèce (voir Figure 4.17 page 117). Ainsi, si chaque espèce a une classe d'équivalence non vide, cela forme aussi un groupe de CDS orthologue complet et revient à la même conclusion que précédemment. Cette partie sera discutée au Chapitre 5.

L'Algorithme 4 présente l'idée générale de construction du graphe de transcrits.

Algorithme 4 Construction d'un graphe de transcrit pour un gène partagé par k espèces

Entrée: alignements des $k \times (k-1)$ paires de gènes via CG-alcode et les relations d'orthologie entre CDS (connus et prédits)
Sortie: graphe de transcrits G^T
  pour toutes les $k \times (k-1)$ paires de gènes $[M_i^G, M_j^G]$ **faire**
    pour tout transcrit T_m du gène i **faire**
      si CDS $\in T_m$ valide **alors**
        ajouter CDS $\in M_{i,m}^T$ à l'ensemble des nœuds de G^T
        si il existe un T_n du gène j tel que $M_{i,m}^T$ syntaxiquement égal à $M_{j,n}^T$ **alors**
          si CDS $\in T_n$ valide **alors**
            ajouter CDS $\in M_{j,n}^T$ à l'ensemble des nœuds de G^T
            ajouter l'arête entre T_m et T_n à G^T
          fin si
        fin si
      fin si
    fin pour
  fin pour

4.3 Exemple du gène CREM et illustration des divergences

Si on reprend l'exemple du gène CREM, suite aux différents filtres, la Figure 4.9 montre son graphe de sites fonctionnels obtenus entre l'humain, la souris et le chien à partir des alignements de paires de gènes. Pour rappel, seuls les sites fonctionnels appartenant à des CDS valides sont présents dans le graphe de sites fonctionnels. Dans ce graphe, 9 sites fonctionnels sont présents chez une seule espèce (6 chez l'humain et 3 chez la souris), 3 sont partagés entre deux espèces (2 couples humain-chien et 1 couple souris-chien, cercle rouge dans la Figure 4.9) et 24 sont conservés entre les trois espèces (triplets dans la Figure 4.9). Grâce à ce graphe, on peut conclure que les structures ont divergé. Il ne constitue pas un cas de triplet de gènes structurellement conservés. En l'occurence, durant l'évolution des trois espèces, on constate que les espèces ont divergé notamment chez l'humain et la souris (plusieurs nouveaux sites fonctionnels). Le chien a tous ses sites fonctionnels déjà observés au moins chez une autre espèce et pourrait avoir conservé la structure ancestrale.

A partir de ce graphe, on peut aligner les modèles de structure de gènes des trois

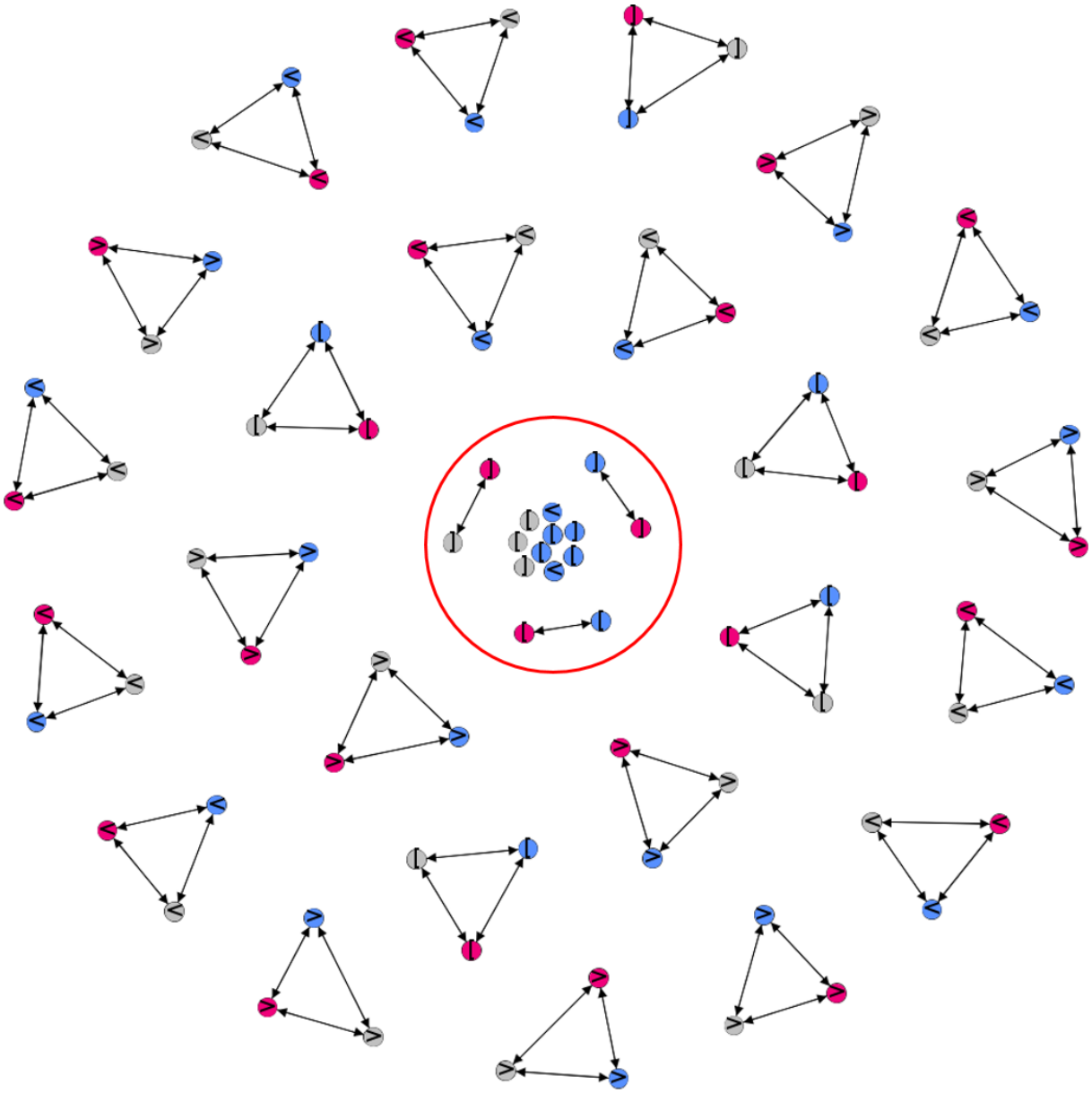


FIGURE 4.9 – Représentation du graphe de sites fonctionnels du gène CREM chez l'humain, la souris et le chien. Le graphe comporte 24 sites fonctionnels partagés entre les trois espèces (composantes connexes à trois sommets ou triplets), 3 partagés entre deux espèces et 9 qui sont retrouvés spécifiquement chez une seule espèce (composantes connexes contenues dans le cercle rouge). Les nœuds des sites fonctionnels sont colorés en bleu, gris et rouge respectivement pour l'humain, la souris et le chien. On observe que des sites fonctionnels ont divergé (apparition/perte) chez les trois espèces.

espèces. Pour le gène CREM, les modèles sont présentés à la Figure 4.10. Les modèles montrent ainsi les différences entre les trois espèces. Les principales différences pour ce gène sont des apparitions de nouveaux blocs exoniques (chez l'humain trois blocs : D, O et V, chez la souris deux blocs : M, P) ou des disparitions de blocs (chez la souris un bloc : J).

```
Humain   [A<>B[C<[D<[E<>F<>G<>H<I J]>K[L<[M[N<[O<[P>Q[R<>S<>T]>U V]
Souris   [A<>B[C<[D<[E<>F<>G<>H<I]]>K[L<[M[N<[O<[P>Q[R<>S<>T]>U]
Chien    [A<>B[C<[D<[E<>F<>G<>H<I J]>K[L<[M[N<[O<[P>Q[R<>S<>T]>U]
```

FIGURE 4.10 – Construction des modèles de structure de gènes pour trois espèces à partir du graphe de sites fonctionnels. En rouge figurent les sites fonctionnels existant sous forme de singletons et en vert les sites fonctionnels existant sous forme de couples. En bleu sont présentés les blocs qui ne sont pas présents chez les trois espèces.

La Figure 4.11 montre le graphe de transcrits, avec CDS valides, du gène CREM où deux groupes de CDS orthologues présentent des cas de transcrits ayant un même CDS chez l'humain (cadres verts dans la Figure 4.11). Ces cas sont expliqués plus loin dans le chapitre. Comme pour son graphe de sites fonctionnels, le graphe de transcrits du gène CREM n'a pas que des triplets de transcrits, ce qui fait que le répertoire de transcrits n'est pas conservé entre les trois espèces. En effet, étant donné que les modèles de gènes des trois espèces ne sont pas structurellement identiques, les gènes n'ont pas la possibilité de produire les mêmes transcrits. Par exemple, le bloc O, uniquement présent dans le modèle du gène CREM humain, ne sera possédé par aucun transcrit de la souris ou du chien. Pour l'humain trois transcrits connus l'utilisent ($ENST00000474931$: $[O <> S <> T]$, $ENST00000468236$: $[O <> QR <> S <> UV]$ et $ENST00000490511$: $[O <> S <> UV]$). Ces CDS ne peuvent exister chez la souris ou chez le chien et sont spécifiques à l'humain. Autre exemple, le bloc J, uniquement présent dans les modèles de l'humain et du chien, ne pourra pas être présent dans des transcrits de la souris. Deux transcrits humains connus ($ENST00000374726$: $[C <> G <> HIJ]$, $ENST00000489321$: $[C <> G <> HIJ]$) et un transcrit du chien prédit ($CGACAFTENST00000374726$: $[C <> G <> HIJ]$) l'utilisent. On peut supposer, ici, l'apparition d'un codon *stop* prématuré chez la souris "$HI]$" alors que chez l'humain et le chien les exons ancestraux sont conservés "$HIJ]$". En conclusion, des sites qui ont divergé (cadre rouge dans la Figure 4.11), qui se sont perdus ou qui sont apparus entraîneront des CDS de transcrits divergés, perdus ou créés.

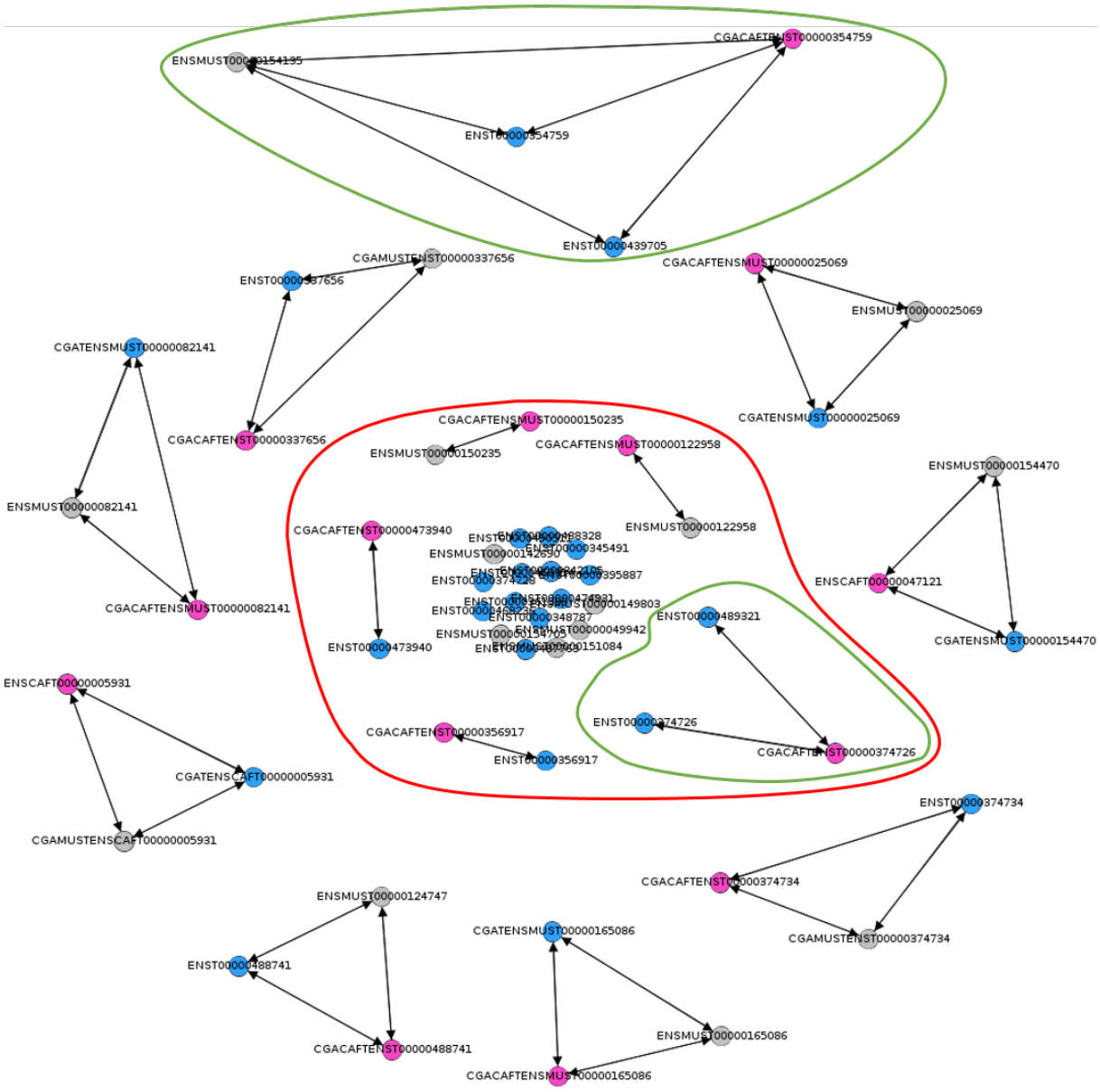


FIGURE 4.11 – Représentation du graphe de transcrits du gène CREM chez l'humain, la souris et le chien. Le graphe comporte 9 composantes connexes de transcrits impliquant les trois espèces (composantes connexes en triplets), 5 composantes connexes de transcrits impliquant deux espèces et 17 composantes connexes de transcrits impliquant une seule espèce (composantes connexes contenus dans le cadre rouge). Certains transcrits encodent le même CDS (cadres verts). Les nœuds sont colorés en bleu, gris et rose selon s'il s'agit de transcrits de l'humain, de la souris et du chien. On déduit que les transcriptomes des gènes CREM ont divergé chez les trois espèces.

4.4 Application des graphes à la détection de structures strictement conservées

Le but de cette partie est d'identifier un jeu de gènes dont la structure d'épissage est conservée entre l'humain, la souris et le chien.

A partir des 2 167 triplets de gènes orthologues de départ et des comparaisons de paires de gènes réalisées et présentées dans le Chapitre 3, nous avons construit 2 141 graphes basés sur les relations d'orthologie entre sites fonctionnels. En effet, 26 triplets de gènes orthologues ne passent pas les filtres posés au niveau de la méthode de comparaison de paires de gènes (problème de colinéarité ou de bijection, voir section 4.2.1 page 100). Sur ces 2 141, 1 661 graphes ont été sélectionnés pour être analysés (voir section 4.2.2 page 102) car respectant toutes les hypothèses relatives aux graphes. Au final, nous avons retenu des graphes possédant soit des sites fonctionnels spécifiques à un gène d'une seule espèce, soit des sites fonctionnels partagés dans les gènes de deux ou trois espèces comme expliqué précédemment (Figure 4.7). Les résultats à chaque étape du pipeline sont présentés à la Figure 4.12.

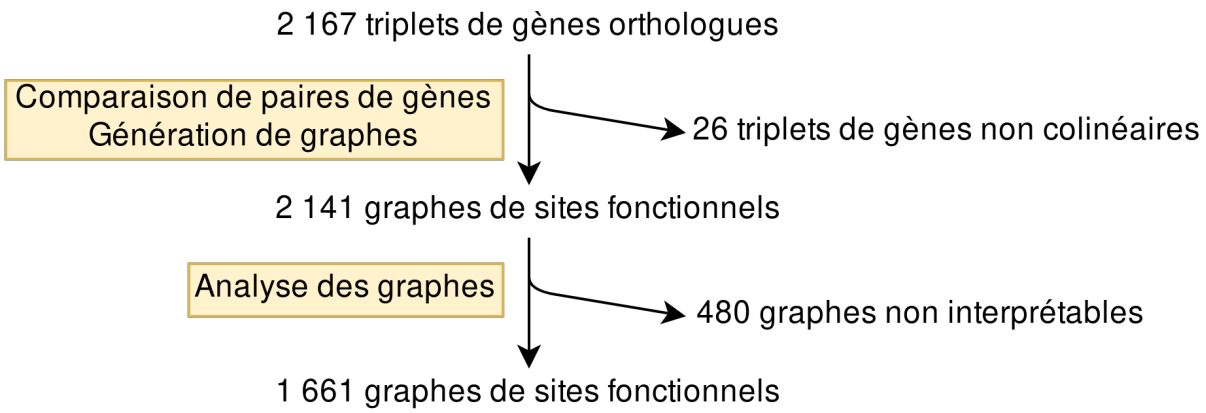


FIGURE 4.12 – Pipeline d'analyse des graphes de sites fonctionnels. Au départ 2 167 triplets de gènes orthologues sont définis pour l'humain, la souris et le chien. 2 141 graphes de sites fonctionnels sont obtenus à partir des comparaisons de paires de gènes les 26 autres ont échoué (gènes non colinéaires entre les trois espèces et orthologie des sites fonctionnels non bijective). 1 661 sont analysables (sites fonctionnels en copie unique, les trois espèces présentes et tous les cas sont interprétables).

4.4.1 253 gènes de structure d'épissage conservée

Parmi les 1 661 triplets de gènes, on en identifie 253 (15,2%) ayant des graphes de sites fonctionnels où chaque site fonctionnel est partagé chez les trois espèces (Figure 4.13). Ces 253 triplets de gènes orthologues définissent, par leur représentation sous forme de graphe de sites fonctionnels ne comportant que des triplets, des gènes structurellement conservés. Les autres gènes présentent au moins un site fonctionnel spécifique à une espèce, ou partagé dans seulement deux espèces sur trois. Par exemple, le gène CREM n'est pas retenu parmi les 253 gènes (voir section 4.3 page 106). Le diagramme de Venn (Figure 4.13) indique la topologie des graphes de sites fonctionnels obtenus pour l'ensemble des données. 220 graphes sont situés à l'intersection "Deux espèces" - "Une espèce", c'est-à-dire que pour chacun de ces 220 gènes, aucun site fonctionnel n'est partagé entre les 3 espèces : des sites fonctionnels peuvent être partagés entre deux espèces mais jamais avec une troisième espèce. 74 cas n'ont aucun site fonctionnel aligné. On peut supposer des artefacts de la comparaison de paires de gènes, un problème avec les données ou l'occurrence de *frameshifts* (décalage du cadre de lecture et codons *stop* divergés) durant l'évolution et nous empêchant de déterminer une paire de transcrits orthologues. Les explorations approfondies pour analyser et traiter ces cas n'ont pas été réalisées dans la thèse. 520 graphes ont des composantes en triplets, couples et singletons comme l'exemple du gène CREM (Figure 4.9) présenté dans la section 4.3 (page 106). Cette figure révèle beaucoup de divergences. Elles ne sont pas interprétées dans cette thèse, le focus ayant été mis sur des gènes conservés.

Analyse des 253 gènes. Les 253 triplets de gènes orthologues partagent une même structure d'épissage, c'est-à-dire qu'on est capable de retrouver tous les sites fonctionnels (codon *start*, codon *stop*, donneur et accepteur d'épissage) alignés chez les trois espèces. De ce fait, tous les exons codants sont partagés entre les trois espèces. Sur cette base, ces gènes orthologues ont donc théoriquement le même potentiel de transcription, c'est-à-dire qu'ils ont en commun tous les sites nécessaires pour réaliser les mêmes CDS orthologues et donc le même ensemble de protéines isoformes. Si on regarde les termes de la *Gene Ontology* afin de classer les gènes, on observe que ces 253 gènes n'appartiennent pas à une catégorie de gènes en particulier, aucun processus n'est surreprésenté (voir Figure 4.14).

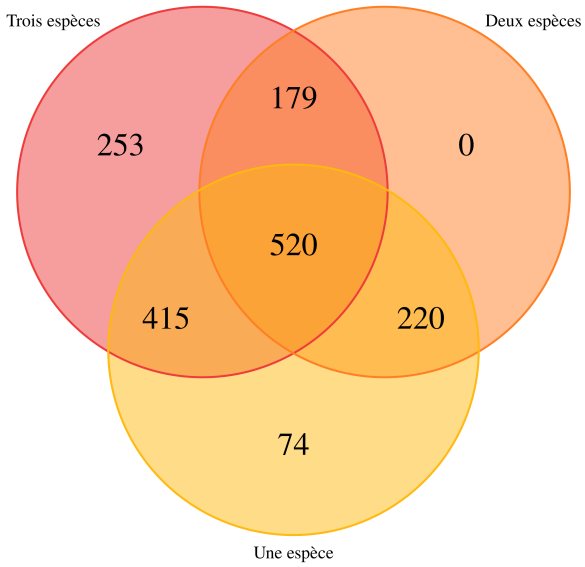


FIGURE 4.13 – Diagramme de Venn montrant la distribution des graphes de sites fonctionnels sur les 1 661 triplets de gènes orthologues. Chaque cercle représente les gènes ayant des sites fonctionnels communs partagés entre trois ou deux espèces ou spécifiques à une seule espèce. Tous les sites fonctionnels sont partagés par les trois espèces pour 253 gènes. Ces gènes sont structurellement conservés.

4.4.2 Transcriptomes des gènes structurellement conservés

Nous avons construit des graphes de transcrits en considérant les 1 661 triplets de gènes orthologues ayant servi dans le paragraphe précédent pour l'analyse des structures d'épissage des gènes. En utilisant les mêmes propriétés citées précédemment (uniquement des singletons, couples ou triplets comme composantes connexes, pas de CDS en copies multiples), nous avons obtenu 986 graphes de transcrits. La Figure 4.15 présente la distribution des topologies de ces 986 graphes de transcrits. Sur l'ensemble de ces 986 graphes, 135 triplets de gènes orthologues possèdent des composantes connexes sous forme de triplets. C'est-à-dire que chacun des CDS de ces gènes est retrouvé chez toutes les espèces en copie unique (ce qui suggère que l'ancêtre commun à ces espèces possédait ces structures de CDS). 206 n'ont aucun transcrit partagé à plus de deux espèces, 103 n'ont eu aucune relation d'orthologie déterminée par CG-alcode. En plus des 135 gènes complets cités, 542 $(185 + 261 + 96)$ ont des transcrits partagés entre les trois espèces.

Les 135 graphes de transcrits ont tous les CDS partagés entre les trois espèces en copie unique, c'est-à-dire que la structure de tout CDS d'une espèce est aussi représentée chez les autres espèces. Pour rappel, avec les graphes de sites fonctionnels (voir section précédente

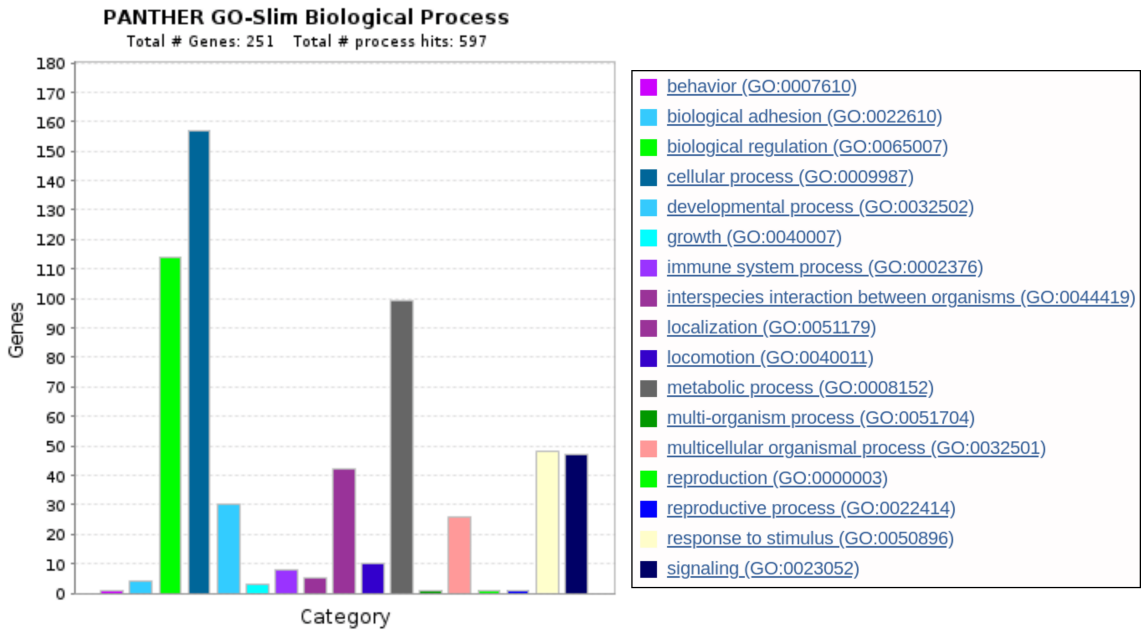


FIGURE 4.14 – Répartition des 253 gènes dans les annotations de processus biologiques de la *Gene Ontology*. 251 gènes ont eu une correspondance dans les annotations. La majorité des gènes se classent dans les catégories "processus cellulaires" ("*cellular process*"), "régulations biologiques" ("*biological regulation*") et "processus métaboliques" ("*metabolic process*").

4.4.1), on a extrait 253 graphes de sites fonctionnels conservés, et on a émis l'hypothèse que ces 253 graphes avaient le même potentiel de transcription. Les 135 gènes identifiés via les graphes de transcrits sont en effet inclus dans ces 253 gènes confirmant l'hypothèse que partager les mêmes CDS implique de partager les mêmes sites fonctionnels.

Les 118 autres gènes restant $(253 - 135)$ comportent quant à eux des cas de CDS en copies multiples chez une espèce. Le gène CREM, qui ne fait pas partie de l'ensemble 253, illustre ces cas de copies multiples (Figure 4.8 et Figure 4.11), où deux composantes connexes présentent deux CDS humains identiques (nœuds bleus dans les cadres verts). Dans chacun des 118 graphes de transcrits, il existe au moins une composante connexe correspondant à de tels cas, c'est-à-dire qu'au moins une composante connexe est un triplet avec des CDS en copies multiples. En considérant ces CDS en copies multiples dans des *classes d'équivalence*, chacune des composantes connexes de ces 118 triplets de gènes est présent sous forme de triplets de classes d'équivalence (voir Figure 4.17). On définit une classe d'équivalence du graphe comme étant l'ensemble du ou des transcrits d'une même espèce ayant un même CDS. On dit que chaque classe d'équivalence ayant

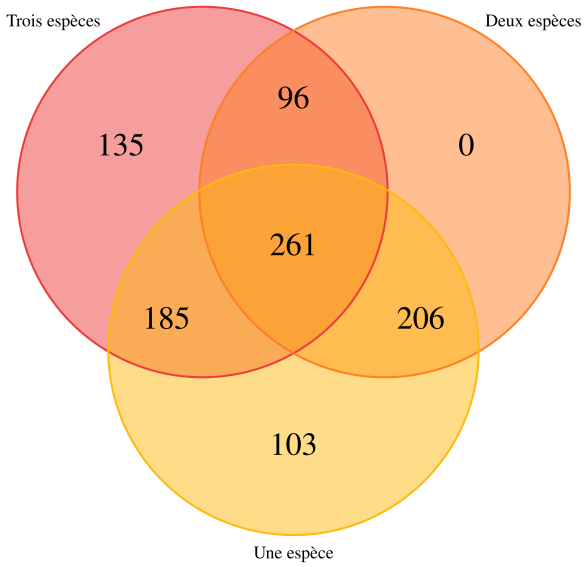


FIGURE 4.15 – Diagramme de Venn montrant la distribution des topologies des graphes de transcrits sur les 986 triplets de gènes orthologues sélectionnés. Chaque cercle représente les gènes ayant des sites fonctionnels communs partagés entre trois ou deux espèces ou spécifiques à une seule espèce. Tous les CDS sont partagés en copie unique chez les trois espèces pour les 135 gènes.

plus d'un transcrit contient n transcrits et $n-1$ transcrits redondants. A ce niveau, chacun des 253 triplets de gènes, dont les structures d'épissage sont conservées, a son répertoire de CDS conservé entre l'humain, la souris et le chien.

4.4.2.1 CDS en copie unique : Analyse des 135 triplets de gènes

135 gènes sont tels que chez chaque espèce, le gène est conservé structurellement et chaque CDS est présent en copie unique. Pour rappel, les graphes de transcrits sont basés à la fois sur les transcrits connus et prédits du gène. On a voulu savoir quel était le nombre de gènes parmi ces 135 gènes pour lesquels tous les transcrits étaient connus chez l'une ou l'autre des espèces. La distribution du nombre de transcrits connus et prédits pour chacun des 135 triplets de gènes est présentée à la Figure 4.16. On observe que seuls 6 gènes du chien sur les 135 n'ont eu aucune prédiction contre 70 (51,9%) chez l'humain et 61 (45,2%) chez la souris. Cela montre que les transcrits connus sont principalement présents chez l'humain et la souris alors que les transcrits prédits sont plus largement présents chez le chien, une espèce émergente dont les données continuent d'arriver dans les bases de données. En effet, 129 gènes (95,6%) sur les 135 gènes du chien ont obtenu

au moins une prédiction. Au total, cet ensemble contient 845 transcrits connus et 541 transcrits prédits, soit $(845 + 541)/3 = 462$ groupes de CDS orthologues (Table 4.1). Il est à noter que 25 transcrits du chien donnés en entrée (présent sur le graphique) n'ont pas été utilisés dans nos graphes car leur CDS n'était pas valide.

4.4.2.2 CDS en copies multiples : Analyse des 118 triplets de gènes

En complément des 135 triplets de gènes en copie unique, 118 gènes sont présents avec au moins un CDS en copies multiples chez une espèces pour un total de 253 triplets de gènes structurellement orthologues. Ces 118 gènes possèdent donc tous des transcrits présentant le même CDS en copies multiples.

Concrètement, comme présenté à la Figure 4.17, nous avons défini, au sein des groupes de CDS orthogues, des *classes d'équivalence* qui correspondent aux transcrits partageant le même CDS chez une espèce. Dans la Figure 4.17a, le triplet de droite correspond à un groupe de CDS orthologues avec un unique transcrit dans chaque classe d'équivalence de chaque espèce. Le triplet de gauche forme un groupe de CDS orthologues avec deux transcrits dans les classes d'équivalence de l'humain et de la souris, et un seul transcrit dans la classe d'équivalence du chien. La Figure 4.17b représente les transcrits complets de ce groupes de CDS orthologues (gènes $ENSG$00000173065, $ENSMUSG$00000037750 et $ENSCAFG$00000031142) pour des classes d'équivalence comprenant plus d'un transcrit. Avec cette représentation, on observe que les CDS des transcrits sont identiques et que leurs régions UTR diffèrent. Avec ces informations, on ne peut pas déterminer quel transcrit humain est orthologue à quel transcrit de la souris sur la base du CDS du transcrit comme seule information. En effet, les CDS sont identiques. Par contre, on peut supposer que le transcrit humain $ENST$00000452648 est orthologue au transcrit de la souris $ENSMUST$0000015571 du fait que les sites d'épissage des introns présents en 5'-UTR peuvent s'aligner. De même, on peut supposer que le transcrit humain $ENST$00000581407 est orthologue au transcrit de la souris $ENSMUST$00000073705, leur 5'-UTR pouvant quasiment s'aligner entièrement. Les extrémités 3'-UTR quant à elles ne permettent pas de déterminer cela.

Sur l'ensemble des 118 triplets de gènes orthologues, on dénombre 1 051 transcrits connus qui permettent de prédire 488 transcrits. En considérant les classes d'équivalence, ce sont 288 CDS redondants qui sont contenus dans ces 118 triplets de gènes. Ainsi, 417 groupes de CDS orthologues $((1\ 051 + 488 - 288)/3)$ sont exprimés par ces 118 triplets de gènes orthologues (Table 4.1).

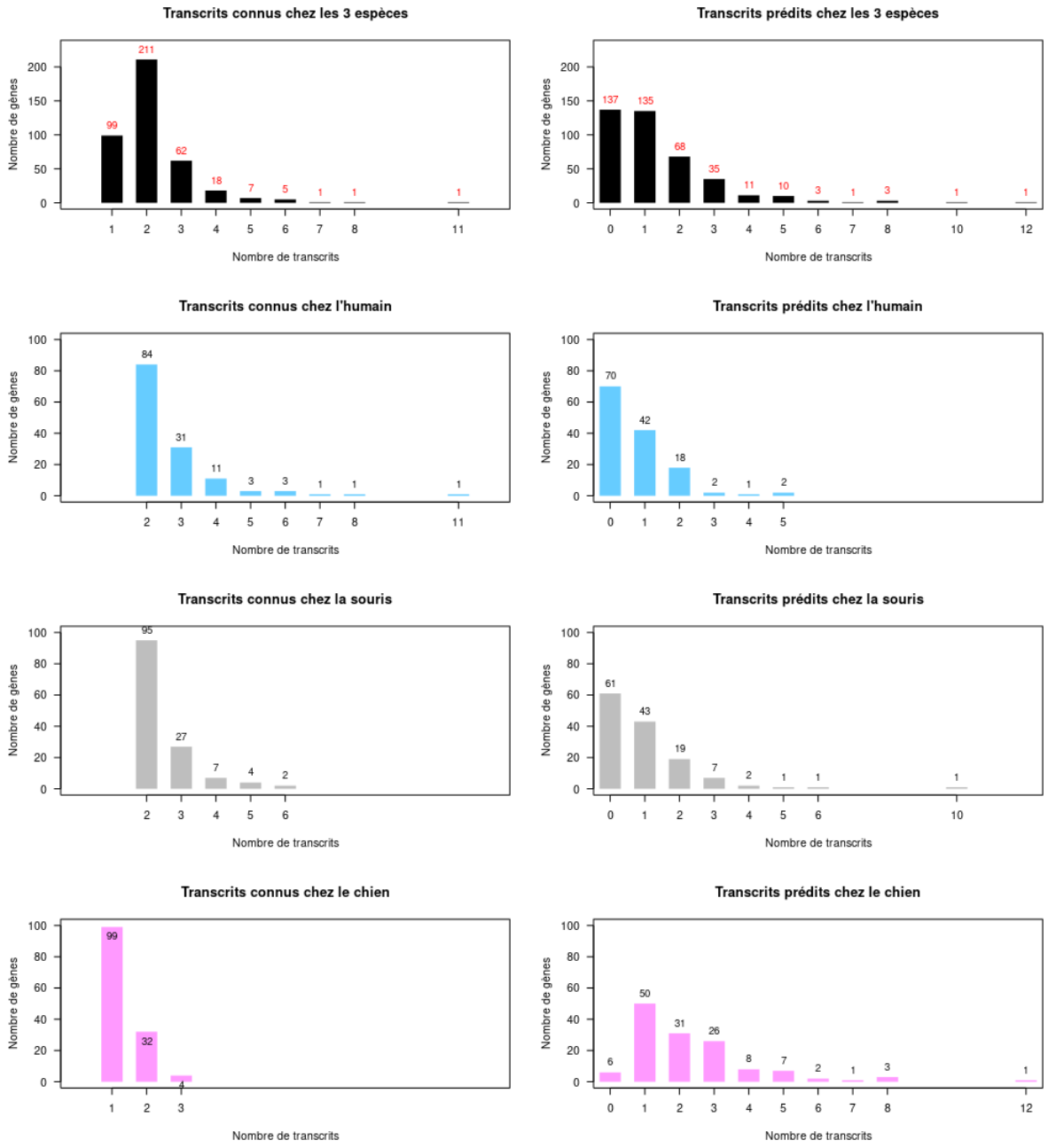


FIGURE 4.16 – Distribution du nombre de transcrits connus (CCDS et Ensembl 90) et prédits dans le cas des 135 triplets de gènes orthologues. Les quatre distributions de gauche correspondent aux transcrits connus, les quatre de droite correspondent aux transcrits prédits. Les deux distributions noires représentent l'ensemble des transcrits chez les trois espèces, les 6 autres correspondent aux distributions chez l'humain (bleu), la souris (gris) et le chien (rose).

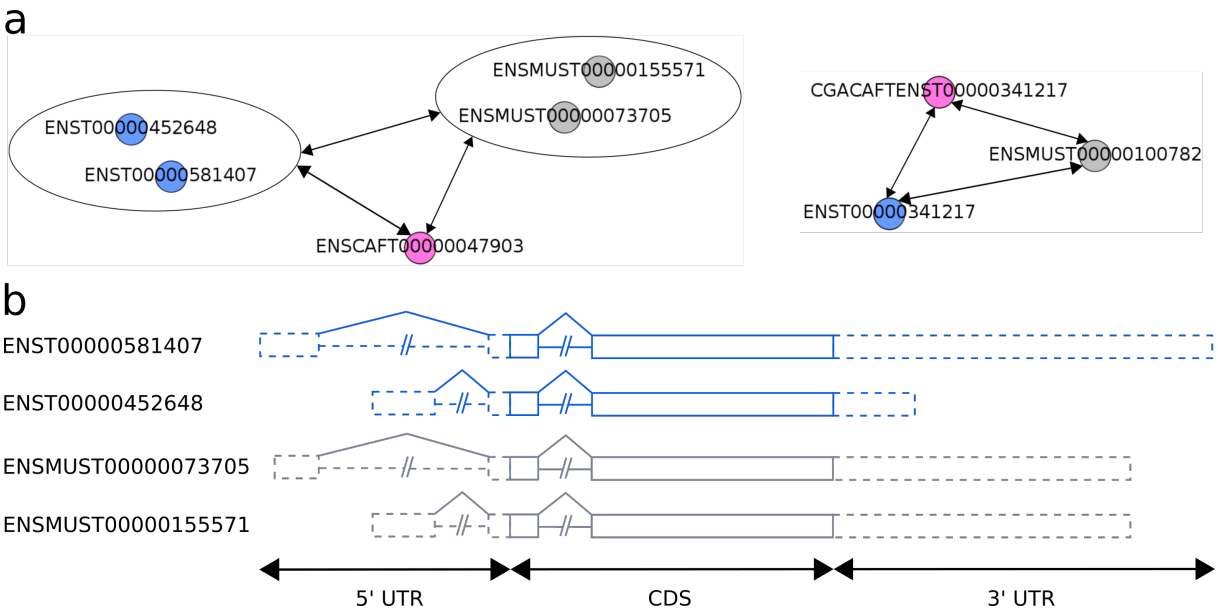


FIGURE 4.17 – Représentation des classes d'équivalence au sein des groupes de CDS orthologues pour le triplet de gène *ENSG*00000173065, *ENSMUSG*00000037750 et *ENSCAFG*00000031142. (a) Caractérisation des classes d'équivalence : le groupe de CDS orthologues de gauche a deux transcrits au sein d'une même classe d'équivalence chez l'humain, de même chez la souris. Pour ce triplet chez le chien, comme pour toutes les espèces du second groupe de CDS orthologues, la classe d'équivalence est composée d'un CDS en copie unique par espèce. (b) Représentation des transcrits impliquant des CDS en copies multiples chez l'humain et chez la souris.

Ensemble de données	Transcrits	Humain	Souris	Chien	Total
253	Transcrits connus	854	762	280	1,896
	CDS prédits	180	249	600	1,029
135 sans CDS redondants	Transcrits connus	364	331	150	845
	CDS prédits	98	131	312	541
118 avec CDS redondants	Transcrits connus	490	431	130	1,051
	CDS prédits	82	118	288	488
	CDS redondants	155	132	1	288
	CDS connus avec UTR manquants	25	10	66	101
114	Classes d'équivalence avec redondance	109	103	1	213

TABLE 4.1 – Caractéristiques des CDS des 253 triplets de gènes structurellement conservés.

De plus, concernant ces 118 triplets de gènes, 101 transcrits connus (Table 4.1) sont dépourvus de régions 5' UTR, 3' UTR, ou les deux, dans les bases de données. Ainsi, si l'on souhaite analyser plus en détail ces gènes au niveau des UTR, nous devons considérer cette absence d'information. Nous avons trouvé 114 triplets de gènes orthologues sur les

118 qui présentent des CDS redondants dans au moins une espèce et dont les régions UTR sont annotées, ce qui représente 213 classes d'équivalence concernées : 109 chez l'humain, 103 chez la souris et 1 seule chez le chien. Pour ces classes d'équivalence, il est possible de chercher à déterminer quels sont les UTR conservés entre deux espèces, tel que illustré à la Figure 4.17. Dans chaque classe d'équivalence de l'humain et de la souris, deux promoteurs alternatifs différents permettent d'initier la transcription de deux transcrits différents encodant le même CDS. Outre la comparaison des UTR conservés orthologues, il devient possible d'examiner la conservation des promoteurs orthologues (voir Chapitre 5)

4.4.3 Base de données de 253 triplets de gènes structurellement conservés

Grâce aux graphes de sites fonctionnels et aux graphes de transcrits, nous avons identifié 253 gènes orthologues dont la structure d'épissage est conservée chez l'humain, la souris et le chien. Ces gènes expriment 1 896 transcrits connus et nous avons prédit 1 029 CDS (Table 4.1). En moyenne, ce sont 2,5 transcrits connus qui sont exprimés par gène allant de 1 (chez le chien) à 13 (chez l'humain). Nous avons prédits une moyenne de 1,3 transcrits par gène allant de 1 (dans chaque espèce) à 12 transcrits prédits (chez le chien). Parmi les 1 029 transcrits prédits, 350 ont été trouvés dans d'autres bases de données que la version 90 d'Ensembl et, pour les autres, 255 présentent des jonctions d'exons spécifiques qui ont été alignées avec des lectures de séquençage (voir Chapitre 3).

Chacun des 253 gènes orthologues code les mêmes structures de CDS épissés dans les trois orthologues. Le protéome des gènes est donc partagé entre les espèces et nous avons identifié 879 groupes de CDS orthologues. Nous avons identifié 114 gènes parmi les 253 où une redondance de CDS s'est produite (45,1% des gènes), ce qui implique 213 classes d'équivalence contenant un CDS en copies multiples (Table 4.1). 8% des CDS sont donc encodés par plusieurs transcrits ($213/3 * 879$). Pour ces gènes, il devient nécessaire d'examiner la conservation des UTR.

Par ailleurs, nous avons construit une base de données SQL, "*transcript_ortho*", pour stocker cet ensemble de 253 triplets de gènes structurellement orthologues. Son schéma relationnel est présenté dans la Figure 4.18. Cette base de données décrit diverses informations telles que :

— la composition en intron/exon des gènes et des transcrits basée sur leur coordonnées

génomiques (contenus dans les tables "exon" et "intron" où "posStart" et "posEnd" définissent la première et la dernière coordonnée génomique),
— les relations d'orthologie entre les gènes des espèces et les relations d'orthologie entre les CDS de transcrits pour les 2 925 transcrits (1 896 connus et 1 029 prédits) (contenus dans les tables "orthoGene", "groupOrthoG" et "pairTOrtho" où "groupGId" définit le groupe de gènes orthologues et "sourceGId" définit l'identifiant du gène),
— les annotations de vérifications expérimentales accordées aux transcrits prédits (contenus dans les tables "typeValidationT" et "sourceValidationT"),
— les sous-catégories de groupes de gènes (ensemble 135, ensemble 118 avec UTR, ensemble 118 sans UTR, contenus dans la table "typeGeneSet" où "nameSet" définit les types d'ensemble),
— les CDS retrouvés dans les groupes de CDS orthologues (contenus dans la table "refCDS" où "idTranscrit" est l'identifiant du transcrit et "idCDS" est l'identifiant du groupe de CDS orthologues),
— les différents CDS apparaissant dans une même classe d'équivalence d'un groupe de CDS orthologue (deux éléments apparaissant avec le même identifiant "idCDS" dans la table "refCDS" et appartenant à la même espèce).

Cette base de données SQL a été implémentée par des scripts Python. Elle a été réalisée en collaboration avec Pierre-Alexis Odye lors d'un stage de M1 informatique que j'ai encadré. Le descriptif des différentes tables et attributs de la base de données est donné en Annexe 7.

Les utilisations potentielles de ces données seront décrites dans le chapitre suivant.

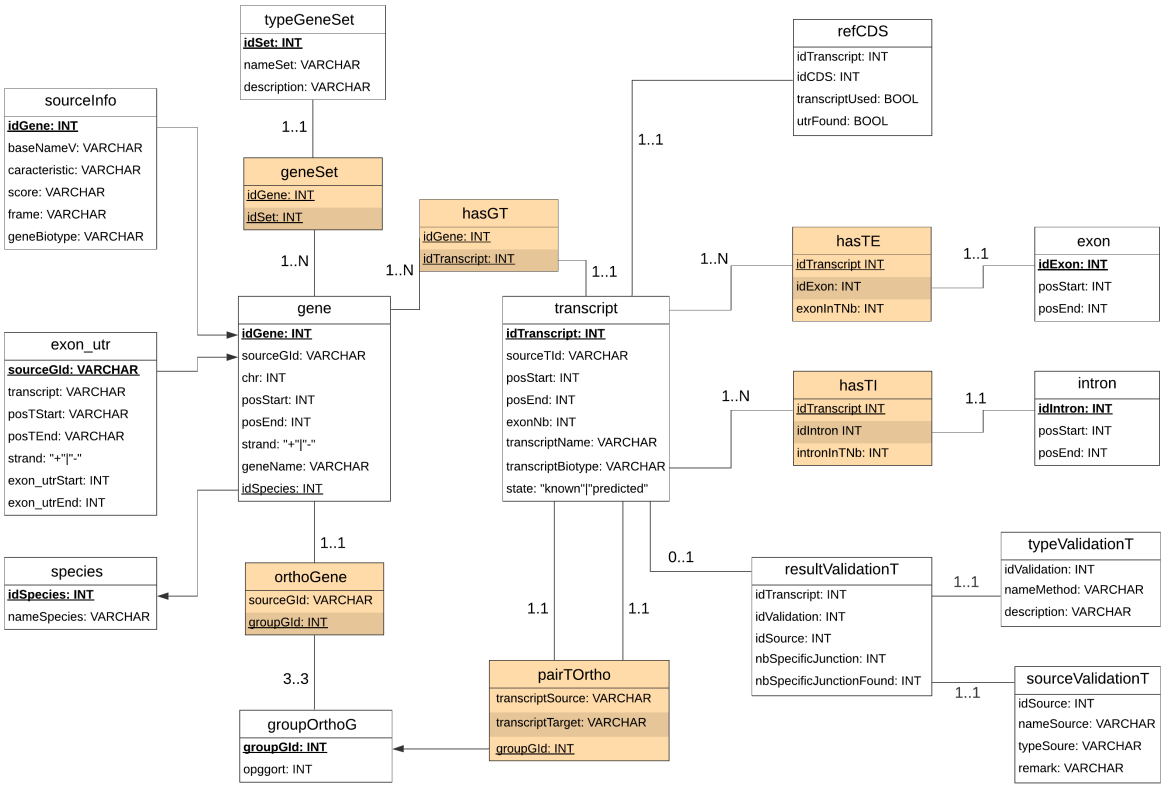


FIGURE 4.18 – Schéma relationnel de la base de données "*transcript_ortho*".

Conclusion du chapitre

Dans ce chapitre, nous avons identifié un ensemble de 253 triplets de gènes orthologues qui présentent des structures de gènes complètement conservés entre l'humain, la souris et le chien. Tous les sites fonctionnels sont partagés. Ces 253 triplets de gènes mettent aussi en avant une conservation des CDS potentiellement exprimés. Dans la description des transcrits complets, 135 triplets de gènes présentent un ensemble de transcrits identiques entre les 3 espèces dans leur partie codante et 118 illustrent des groupes d'orthologie avec de multiples transcrits encodant un même CDS pour une même espèce, résultant de la transcription alternative. Nous avons également identifier 879 groupes de CDS orthologues pour l'ensemble des 253 triplets de gènes.

Chapitre 5

Applications et perspectives

Dans ce chapitre, nous présentons les perspectives découlant de la thèse et les applications possibles du jeu de données de 253 gènes dont la structure et les CDS alternatifs sont conservés.

Sommaire

5.1 Génomique comparative 122

5.1.1 Comparaison des transcrits dans leur intégralité : analyse des régions UTR 122

5.1.2 Comparaison des éléments régulateurs 125

5.2 Transcriptomique comparative 126

5.3 Évaluation comparative des méthodes d'alignement épissé . . 128

5.4 Poursuite de la méthode de comparaison multi-espèces . . . 129

5.1 Génomique comparative

5.1.1 Comparaison des transcrits dans leur intégralité : analyse des régions UTR

Jusqu'à présent, on a pu déterminer la conservation des gènes en fonction de leur potentiel codant. Pour 253 gènes sélectionnées, le potentiel codant est strictement conservé. Il devient alors possible d'interpréter d'autres caractères divergents. Une perspective peut notamment être de se pencher sur une analyse plus fine de la diversité des transcrits concernant les UTR alternatifs. Les UTR d'un transcrit codant, 5'UTR précédant le CDS et 3' UTR suivant le CDS, sont d'importantes régions régulatrices du transcrit. Les UTR régulent la sélection par le ribosome d'un codon *start* donné, l'adressage du transcrit dans un compartiment cellulaire donné ou encore sa stabilité (Mayr 2019). La variabilité des UTR des transcrits résulte du phénomène de transcription alternative d'une part (voir section 1.1.3.2 page 24) impliquant plusieurs sites d'initiation et de terminaison de la transcription présent sur un même gène. D'autre part, il existe aussi la possibilité de mobiliser des codons *start* alternatifs d'un même transcrit, phénomène de traduction alternative, ce qui induit des 5'UTR et des CDS différents (Kazak et al. 2013 ; Kwan et Thompson 2019). Les UTR alternatifs des transcrits n'ont pas été pris en compte dans les comparaisons et les prédictions de transcrits dans les travaux présentés dans ce manuscrit. Cela dit, nos données permettent de comparer les UTR des transcrits encodant des CDS orthologues. Chacun des 253 gènes structurellement conservés codent les mêmes structures de CDS chez les trois espèces. Nous avons identifié 879 groupes de CDS orthologues. Nous avons identifié 114 gènes parmi les 253 (45,1% des gènes) où une redondance de CDS est observée, ce qui implique 213 classes d'équivalence présentant des CDS redondants sur les 879 groupes de CDS orthologues (voir section 4.4.2.2 page 115 et Table 4.1). Pour rappel, une classe d'équivalence est un ensemble de transcrits qui, dans un même gène chez une même espèce, partagent un même CDS (voir Chapitre 4). Pour chacun de ces 213 cas de classes d'équivalence, il existe ainsi plusieurs transcrits sur un même gène qui encodent une même protéine, et dont les UTR diffèrent. Nous avons examiné ces UTR pour les transcrits connus (les transcrits prédits étant réduits à la partie CDS).

La partie 5'-UTR des transcrits est une zone qui contient une certaine diversité d'introns. Quant au côté 3'-UTR, la plupart du temps, peu d'introns sont observés. Par la suite, on parle d'*intron UTR* pour qualifier un intron situé dans une région UTR. Pour des classes d'équivalence allant jusqu'à 6-7 transcrits possibles chez une espèce, nous avons pu

observer que le nombre d'introns peut être assez important. Par exemple, pour le gène de la souris $ENSMUSG00000070780$ (Figure 5.1), nous observons 6 transcrits au sein d'une même classe d'équivalence. Leur côté 5'-UTR est différent pour chacun des 6 transcrits indiquant qu'au moins 6 régions promotrices sont présentes pouvant encoder le même CDS. Pour rappel, une région promotrice est située avant chaque site de début de transcription, et elle a pour fonction d'initier la transcription. Le côté 5'-UTR est complexe puisque la combinatoire d'exons et d'introns est importante : on retrouve entre 1 et 5 introns UTR pour chaque transcrit, et 8 introns différents peuvent être dénombrés sur l'ensemble des 6 transcrits, dans la région UTR.

```
ENSMUST00000113726                    {<  ><          ><  > [<><><>]          }
ENSMUST00000201544  {<>     <><         ><          ><  > [<><><>]          }
ENSMUST00000094757                              {  <  ><  > [<><><>]        }
ENSMUST00000167950        {  <><        ><          ><  > [<><><>]      }
ENSMUST00000200775                {<      ><          ><  > [<><><>]    }
ENSMUST00000113724                            {     <     > [<><><>] }
```

FIGURE 5.1 – Modèles de structure des transcrits contenus dans une classe d'équivalence du gène de la souris $ENSMUSG00000070780$. 6 transcrits sur les 7 connus du gène possèdent un même CDS en copies multiples constituant ainsi une même classe d'équivalence. Sur la figure, seuls les 6 transcrits de la même classe d'équivalence sont représentés. Le CDS est délimité par les codons *start* et *stop* et est encadré sur la figure. Les symboles correspondent aux codons *start* ([) et *stop* (]), aux sites donneur (<) et accepteur (>) d'épissage, et aux sites d'initiation ({) et de terminaison (}) de la transcription. Les sites fonctionnels présents dans plusieurs transcrits et alignés dans une même colonne ont la même coordonnée génomique. Pour chaque transcrit, les sites fonctionnels affichés sont uniquement ceux impliqués dans la structure du transcrit (*i.e.* les sites non impliqués pour ces transcrits sont laissés en blanc).

Dans le même groupe de CDS orthologues (Figure 5.2), le gène humain $ENSG00000163694$ possède 2 transcrits avec le même CDS en copies multiples dans la classe d'équivalence. L'un possède 2 introns UTR et l'autre 3 introns UTR. Pour l'ensemble des deux transcrits humains, ce sont 4 introns UTR différents qui sont identifiés. Quant au gène du chien complétant le groupe de CDS orthologues, un seul transcrit est identifié, avec 1 intron UTR.

Une classe d'équivalence contient des transcrits possédant le même CDS mais des UTR différents. Elle met donc en évidence des transcrits qui se confondent si on les regarde à l'échelle de la partie codante (CDS) mais qui se différencient si on les observe à l'échelle de la séquence entière (CDS et UTR). Une perspective envisageable est de comparer les transcrits dans leur intégralité. Dans le détail, cette thèse met en correspondance

des transcrits sur plusieurs espèces au niveau des CDS aboutissant à l'identification de l'homologie structurelle des CDS de transcrits communs et à la prédiction de CDS orthologues. Cette approche pourrait être étendue à l'échelle du transcrit entier (CDS et UTR) pour identifier de nouveaux exons UTR dans les gènes des autres espèces par alignement des régions UTR et pour prédire de nouveaux transcrits complets orthologues. En l'état actuel de la méthode, on a le moyen d'aligner les exons et les sites des régions UTR des transcrits d'un même gène chez une même espèce, mais pas de faire cet alignement entre les espèces. En effet, CG-alcode utilisé pour la base de ce travail de thèse pour comparer les transcrits entre les espèces ne se base que sur l'alignement de leur CDS et non de leurs UTR.

```
Mouse   ENSMUST00000113726                          { < > <         > < > [<><><>]          }
        ENSMUST00000201544  { < >    < > <            > <         > < > [<><><>]          }
        ENSMUST00000094757                                    { < > < > [<><><>]        }
        ENSMUST00000167950          { < > <            > <         > < > [<><><>]     }
        ENSMUST00000200775                    { <      > <         > < > [<><><>]   }
        ENSMUST00000113724                                 {    <      > [<><><>] }
        ____________________________________________________________________________________
Human   ENST00000381793                                       { < > < > [<><><>]   }
        ENST00000295971                             { < > <      > < > [<><><>] }
        ____________________________________________________________________________________
Dog     ENSCAFT00000025193                                        { < > [<><><>] }
```

FIGURE 5.2 – Modèles de structure des transcrits contenus dans les trois classes d'équivalence humain, souris et chien d'un même groupe de CDS orthogues. Le triplet de gènes orthologue $ENSG00000163694$-$ENSMUSG00000070780$-$ENSCAFG00000015898$ possède un groupe de CDS orthologues avec un même CDS en copies multiples dans la classe d'équivalence de la souris et dans celle de l'humain. 6 transcrits souris et 2 transcrits humains possèdent le même CDS. Le chien a un seul transcrit représentant ce CDS. Les unités lexicales des parties UTR sont alignées sur une même espèce, mais pas entre les espèces. Pour chaque transcrit, les sites fonctionnels affichés sont uniquement ceux impliqués dans la structure du transcrit (*i.e.* les sites non impliqués pour ces transcrits sont laissés en blanc).

Dans l'exemple illustré, il existe une combinatoire complexe résultant de promoteurs alternatifs et d'épissages alternatifs conduisant à la formation de différentes régions 5'UTR de transcrits, lesquels encodent une même protéine. Une perspective importante de cette thèse consisterait à examiner la conservation de ces UTR distincts entre espèces. Chacune des trois espèces partage-t-elle le même ensemble de régions 5'UTR, auquel cas une combinaison d'exons UTR reste à découvrir chez la souris et 6 chez le chien ? Les différents UTR connus sont-ils au contraire distincts entre espèces et certains n'ont pas d'orthologue chez une ou deux espèces ? Ces différents scénarios posent donc la question de l'évolution et ces

différents UTR. Enfin et par ailleurs, est-il possible d'établir les spécificités fonctionnelles liées à ces transcrits ? Et comment est régulée l'expression de différents transcrits encodant une même protéine ? Leurs promoteurs alternatifs d'un même gène opèrent-ils dans des contextes physiologiques différents ? Les promoteurs alternatifs orthologues sont-ils conservés entre espèces ?

5.1.2 Comparaison des éléments régulateurs

L'épissage alternatif est régulé grâce à différents processus aujourd'hui identifiés et impliquant la formation de complexes moléculaires combinant le pré-ARNm avec des protéines ou des micro-ARN complémentaires. Les sites de fixation de ces éléments régulateurs sur le pré-ARNm peuvent être identifiés dans la séquence ADN du gène. On distingue les motifs dits *enhancer*, dont la fonction est de favoriser un évènement, et les motifs *silencer*, dont la fonction est d'empêcher l'évènement (voir section 1.1.2.1 page 20, Kelemen et al. 2013). Ces motifs *enhancer* et *silencer* ne sont pas déterminés entièrement et sont dépendants de conditions d'expression. Ces éléments agissent directement sur la reconnaissance d'un intron et la sélection d'un exon par le spliceosome. La Figure 5.3 présente des exemples possibles de motifs *enhancer* et de motifs *silencer*. D'autres cas peuvent exister comme par exemple la présence de combinaison de motifs *enhancer* et/ou *silencer*. Lorsqu'un motif *enhancer* et un motif *silencer* co-existent, c'est une compétition directe entre les deux motifs où la régulation dépendra de la concentration des éléments et de leur affinité pour ces motifs (Cartegni et al. 2002).

Les données sont aujourd'hui encore peu nombreuses et de nombreuses études sont nécessaires pour compléter nos connaissances sur ces motifs (Halfon 2020). Dans cette thèse, nous avons pu obtenir un ensemble de gènes structurellement conservés. Après avoir pris en compte l'orthologie décrite des sites fonctionnels, nous pourrions ensuite étudier la divergence ou la conservation de ces séquences impliquées dans la régulation de l'expression des transcrits alternatifs. Des études montrent que certains événements d'épissage rencontrent une divergence dans leurs taux d'inclusion (Xiong et al. 2018) ou une divergence dans leurs taux d'expression spécifiques aux tissus (Koren et al. 2007), ce qui suggère une divergence des séquences régulatrices. Nos données identifient les transcrits alternatifs orthologues et les sites d'épissage orthologues. Par conséquent, les niveaux d'inclusion d'exons alternatifs orthologues peuvent être comparés entre espèces et corrélés avec de possibles divergences des motifs régulateurs observées aux environs des sites d'épissage concernés.

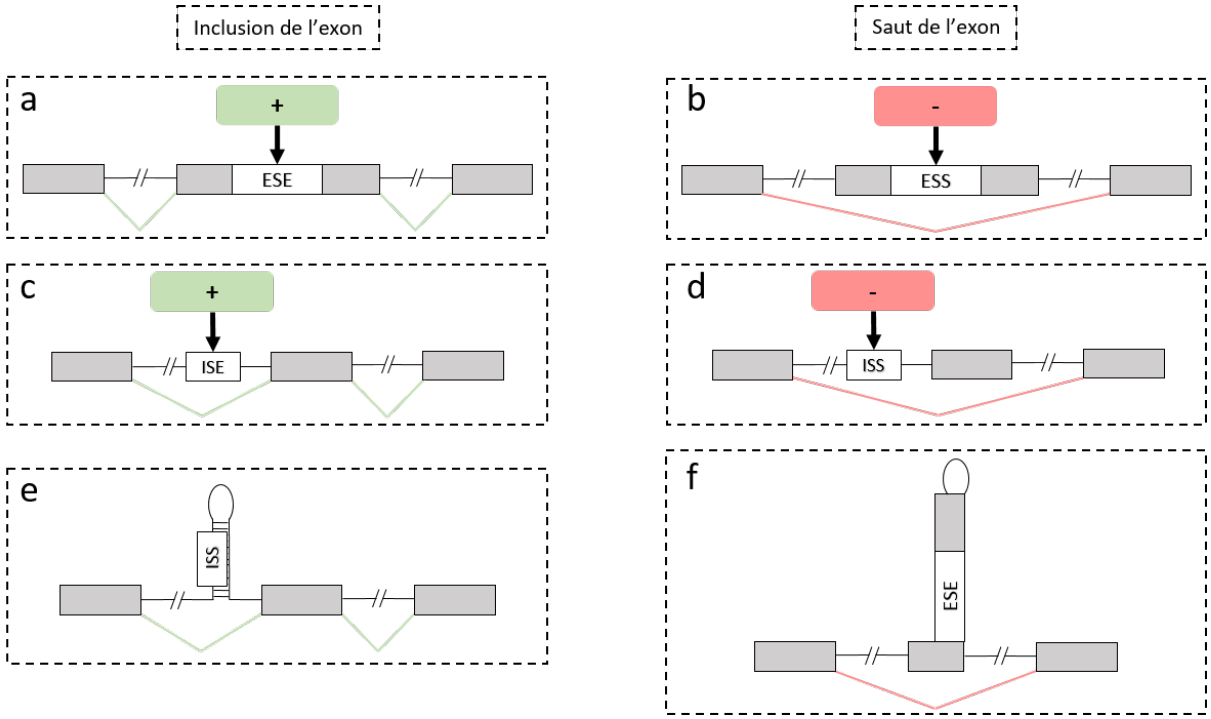


FIGURE 5.3 – Définition d'un intron et d'un exon (adaptée de KELEMEN et al. 2013 et de CARTEGNI et al. 2002). Le motif *enhancer* situé sur l'exon (ESE : *exonic splicing enhancer*, **a**) ou sur l'intron (ISE : *intronic splicing enhancer*, **c**) favorise le recrutement des éléments qui vont reconnaître l'exon (où se situe l'ESE, en aval de l'ISE) pour que celui-ci soit sélectionné. Le motif *silencer* situé sur l'exon (ESS : *exonic splicing silencer*, **b**) ou sur l'intron (ISS : *intronic splicing silencer*, **d**) inhibe le recrutement des éléments qui vont reconnaître l'exon (où se situe l'ESS, en aval de l'ISS) pour que celui-ci ne soit pas reconnu. Des structures secondaires d'ARN sous forme de tiges-boucle peuvent se former et replier l'ARN sous forme double brin, empêchant la reconnaissance de ces éléments qui nécessite que l'ARN soit sous forme simple brin. Par exemple, un ISS n'est pas reconnu et n'inhibe pas la sélection de l'exon (**e**), ou encore un ESE n'est pas accessible est ne permet pas de reconnaître l'exon (**f**).

5.2 Transcriptomique comparative

La transcriptomique comparative est la méthode qui permet de comparer les niveaux d'expression des gènes à travers différentes conditions physiologiques, différents tissus ou même différents individus d'une même espèce. Cette méthode peut être poussée à une toute autre échelle : la comparaison des niveaux d'expression des transcrits alternatifs d'un gène entre les espèces. Par exemple, BRESCHI et al. 2017 compare les transcriptomes exprimés entre l'humain et la souris. Dans leur analyse, ils précisent que le profil d'ex-

pression d'un gène selon les tissus et les conditions d'expression peut être différent chez une espèce. Ces profils d'expression peuvent être partagés ou divergés entre espèces. Dans tous les cas, ces comparaisons entre espèces sont aujourd'hui réalisées à l'échelle du gène, tous ses transcrits alternatifs étant confondus dans la mesure d'expression. Il conviendrait de réaliser ces mesures à l'échelle du transcrit alternatif, ce que pourraient permettre de réaliser nos données.

En particulier, la notion de jonction d'exons spécifique utilisée dans cette thèse (voir section 3.3.2 page 81) correspond à une signature spécifique à un transcrit.

A partir des données d'alignement de lectures sur les jonctions d'exons spécifiques (voir section 3.3.4 page 88), on a pu observer que certaines jonctions avaient de fortes couvertures (Figure 3.11). Avec la description des CDS orthologues produite dans cette thèse, il devient possible de quantifier les niveaux d'expression différentielles de certains CDS alternatifs chez chacune des espèces. Une telle étude de transcriptomique comparative nécessite pour cela d'utiliser des échantillons de tissus pour chacune des espèces et obtenus dans des conditions physiologiques similaires et à une même profondeur de séquençage. En supposant que ces échantillons soient obtenus, nous pourrions quantifier le niveau d'expression d'une jonction d'exons spécifique à un transcrit donné chez les trois espèces afin d'étudier si cette jonction est exprimée différemment chez les trois espèces c'est-à-dire si des espèces sur-expriment ou sous-expriment cette jonction et donc ce transcrit dans des conditions données. Ceci constituerait une première approche possible pour comparer les niveaux d'expression entre espèces à l'échelle du transcrit alternatif.

Prenons l'ensemble du triplets de gènes $ENSG00000001167$ - $ENSMUSG00000023994$ - $ENSCAFG00000001580$ appartenant à l'ensemble des 253 triplets de gènes structurellement conservés. Ce triplet de gènes possède quatre groupes de CDS orthologues où chaque CDS est en copie unique dans chaque espèce (Figure 5.4). Parmi les jonctions d'exons visibles, nous avons sélectionné deux jonctions : la jonction "$F <> H$" spécifique au groupe de CDS orthologues 2 et la jonction "$A <> BC$" spécifique au groupe de CDS orthologues 3 dans la figure. Nous avons ensuite sélectionné les échantillons de tissus en commun entre les trois espèces (Tables **??**, **??**, **??**). Au total, 5 tissus semblent communs aux trois espèces : côlon, cœur, rein, foie et pancréas. Nous avons aligné les échantillons de lectures de ces 5 tissus sur les deux jonctions spécifiques sélectionnées. Les résultats sont présentés dans la Figure 5.5. Les résultats montrent un alignement de lectures sur les deux jonctions pour les trois espèces. Seulement, ces résultats ne sont pas interprétables directement. Il est nécessaire de considérer plusieurs paramètres notamment la profon-

deurs de séquençage, les réplicats ou encore les conditions physiologiques. Si toutes les conditions sont réunies, on peut imaginer normaliser des couvertures et les comparer, ce qui n'est pas possible ici.

Groupe de CDS orthologue 1	CGATENSMUST00000078800	[A<>C<>D<>E<>F<>GH<>I<>J<>K]
	ENSMUST00000078800	[A<>C<>D<>E<>F<>GH<>I<>J<>K]
	ENSCAFT00000002475	[A<>C<>D<>E<>F<>GH<>I<>J<>K]
Groupe de CDS orthologue 2	CGATENSMUST00000159063	[A<>D<>E<>F<>H<>I<>J<>K]
	ENSMUST00000159063	[A<>D<>E<>F<>H<>I<>J<>K]
	CGACAFTENSMUST00000159063	[A<>D<>E<>F<>H<>I<>J<>K]
Groupe de CDS orthologue 3	ENST00000341376	[A<>BC<>D<>E<>F<>GH<>I<>J<>K]
	ENSMUST00000046719	[A<>BC<>D<>E<>F<>GH<>I<>J<>K]
	CGACAFTENST00000341376	[A<>BC<>D<>E<>F<>GH<>I<>J<>K]
Groupe de CDS orthologue 4	ENST00000353205	[A<>D<>E<>F<>GH<>I<>J<>K]
	ENSMUST00000162460	[A<>D<>E<>F<>GH<>I<>J<>K]
	CGACAFTENST00000353205	[A<>D<>E<>F<>GH<>I<>J<>K]

FIGURE 5.4 – Groupes de CDS orthologues pour le triplet de gènes $ENSG00000001167$-$ENSMUSG00000023994$-$ENSCAFG00000001580$. Quatre groupes de CDS orthologues sont exprimés par les trois espèces (gauche), les identifiants de transcrit sont indiqués (centre) ainsi que les modèles des transcrits (droite). Certains groupes de transcrits orthologues possèdent des jonctions d'exons qui leurs sont spécifiques, la jonction entre les exons A et BC (notée "$A <> BC$" dans le modèle de transcrit) est spécifique au groupe 3, la jonction entre les exons F et H (notée "$F <> H$") est spécifique au groupe 2. Ces deux exemples sont indiqués en rouge sur la figure.

Ainsi, à partir de jeux de données séquencés dans des tissus et des conditions physiologiques similaires, on pourra comparer entre espèces, pour ce jeu de données de 253 gènes, les niveaux d'expression à l'échelle des transcrits alternatifs, identifiés par leurs jonctions d'exons spécifiques.

5.3 Évaluation comparative des méthodes d'alignement épissé

Une autre perspective de cette thèse peut être d'améliorer les méthodes de génomiques comparatives ou d'alignement d'épissé (MEYER et al. 2020). Dans leur travail avec *SplicedFamAlign* (JAMMALI et al. 2019), les auteurs ont pris en considération la structure d'épissage des transcrits ainsi que les sites d'épissage connus dans un gène cible afin de rechercher des orthologues épissés grâce à un alignement de séquences par programmation dynamique (voir Chapitre 1). Avec *ThorAxe* (ZEA et al. 2021), les auteurs représentent le

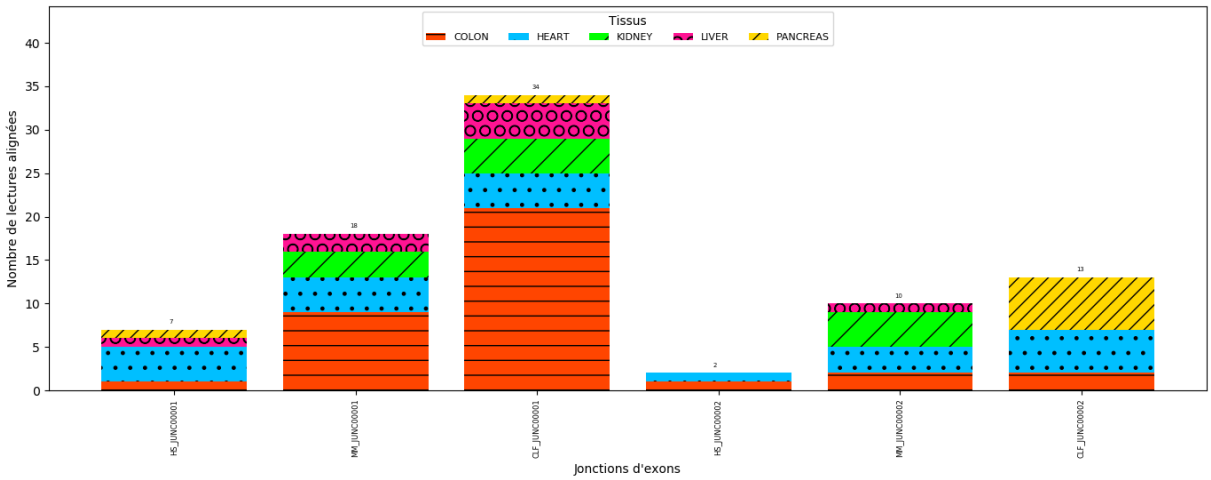


FIGURE 5.5 – Alignement de lectures sur des jonctions d'exons spécifiques. Chaque colonne en abscisse représente une jonction d'exons testée et en ordonnée le nombre de lectures qui y sont alignées. Les trois jonctions de gauche correspondent à la jonction d'exons "$F <> H$" du groupe de CDS orthologues 2 (Figure 5.4) et les trois jonctions de droite correspondent à la jonction d'exons "$A <> BC$" du groupe de CDS orthologues 3 (Figure 5.4).

transcriptome multi-espèce d'un gène par le biais de graphes dont les nœuds sont les segments exoniques partagés entre espèces. Un chemin dans le graphe partagé par plusieurs espèces peut correspondre à des transcrits orthologues partagés.

Notre ensemble de 253 triplets de gènes orthologues conservés obtenu au cours de la thèse, peut permettre d'évaluer, d'améliorer et de comparer des méthodes informatiques telles que *SplicedFamAlign* ou encore *ThorAxe* visant à identifier la conservation des transcrits entre espèces et visant à identifier des ensembles de transcrits orthologues.

5.4 Poursuite de la méthode de comparaison multi-espèces

La méthode de comparaison trois-espèces employée sur l'humain, la souris et le chien est une première étape de comparaison multi-espèces. Deux questions interviennent à la suite de ce travail : La méthode peut-elle être employée avec trois espèces plus éloignées phylogénétiquement ? La méthode peut-elle être employée à une échelle de plus de trois espèces ?

L'étude réalisée repose sur une implémentation pré-existante d'alignement sur un gène

orthologue de blocs codants et de sites connus d'un gène source (Blanquart et al. 2016), méthode appliquée à des espèces proches phylogénétiquement. Si on souhaite étudier de nouvelles espèces plus éloignées, les résultats peuvent changer : les scores des alignements des Blast ne seront pas les mêmes et on s'attend à perdre en conservation sur des espèces très éloignées. En particulier, plus les espèces seront divergées, plus la probabilité sera grande pour qu'un site fonctionnel au moins ait divergé. Dans ce cas on ne pourra pas identifier comme orthologues certains couples de transcrits, car leur structure d'épissage ne sera plus jugée conservée. En d'autres termes, porter notre analyse à des cas d'espèces plus éloignées requerra de savoir interpréter les divergences possibles des exons orthologues : apparition de larges indel, disparition et réapparition compensatrice d'un site d'épissage, apparition de décalage de phase du CDS, mutation des codons *start* et *stop* entraînant une variation de longueur d'exons C et N terminaux orthologues. Notre étude n'a considéré comme orthologues que les transcrits ayant conservé tous leurs sites fonctionnels. Identifier les transcrits orthologues dont certains sites ont divergé pourrait permettre d'étendre la méthode à des espèces plus éloignées.

Le deuxième point à évoquer concerne le fait de pouvoir employer cette méthode à une échelle de plus de trois espèces. On pourrait imaginer, par exemple, comparer à des espèces de primates, ou à d'autres mammifères, ou à des espèces plus éloignées de l'humain. La structure de graphe, introduite pour gérer les comparaisons multi-espèces, permet d'envisager de comparer plus de trois espèces. Les limites reposent principalement sur la combinatoire, et le fait que la méthode va devoir réaliser $(N \times (N-1))$ analyses indépendantes de paires de gènes pour une exploration sur N espèces, et valider à chaque fois les hypothèses de conservation (bijection, colinéarité) avec un risque d'échec dépendant des alignements réalisés.

Une stratégie envisageable serait de ne plus avoir recours à de multiples comparaisons de paires de gènes, mais à des alignements multiples des séquences des gènes pour inférer les modèles de structure.

Conclusion

Au cours de cette thèse, nous avons conçu et mis en œuvre une méthode de comparaison multi-espèces de gènes orthologues dans le but de contribuer à répondre à la question ouverte : *quels sont les transcrits exprimables d'un gène ?* En particulier, nous nous sommes intéressés à rechercher des gènes ayant des structures d'épissage conservées chez plusieurs espèces, et à la recherche de gènes capables d'exprimer un transcriptome également conservé chez ces espèces. La modélisation multi-espèces des gènes et des transcrits repose sur une structure de graphe. L'ensemble des développements et des résultats produits au cours de cette thèse sont présentés dans la Figure 5.6.

Cette comparaison multi-espèces a été appliquée à trois espèces : l'humain, la souris et le chien. 2 167 gènes orthologues partagés par ces trois espèces ont été sélectionnés, l'ensemble exprimant 18 109 transcrits connus en janvier 2018. La comparaison multi-espèces a notamment permis de prédire 6 861 transcrits non connus et potentiellement exprimables par ces gènes. Une analyse a été mise en place pour tester la validité de ces prédictions en fonction d'autres sources de connaissance et de données de séquençage. Elle révèle que parmi les transcrits prédits, plus de 24% sont déjà connus dans d'autres bases de données et plus de 25% possèdent des jonctions spécifiques d'exons sur lesquels s'alignent des lectures de séquençage (voir Chapitre 3).

La comparaison multi-espèces s'appuie sur des graphes de sites fonctionnels (codon *start*, codon *stop*, sites donneur et accepteur d'épissage) spécifiant les relations d'orthologie entres sites, ceci afin d'analyser la structure des régions codantes des gènes. Au total, 253 triplets de gènes ont été identifiés chez l'humain, la souris et le chien, ayant tous leurs sites fonctionnels composants leur structure d'épissage partagés chez les trois espèces. Les structures partagées par ces 253 triplets de gènes orthologues peuvent être interprétées comme conservées depuis l'ancêtre commun à ces trois espèces. A partir de là, nous avons émis l'hypothèse que le transcriptome de ces gènes pouvait également être conservé du fait que tous les sites fonctionnels permettant d'exprimer les exons codants pour des protéines (CDS) sont présents chez chaque espèce. Nous avons testé cette hypothèse en examinant les graphes de transcrits spécifiant les relations d'orthologie entre CDS chez les trois espèces. Les graphes de transcrits sont composés des composantes connexes réunissant les

transcrits ayant un même CDS dont la structure est conservée chez les différentes espèces. On appelle cette composante connexe un groupe de CDS orthologues. Au total, pour les 253 gènes, on dénombre 879 groupes de CDS orthologues et donc 879 protéines isoformes possibles partagées par l'humain, la souris et le chien. Pour ces 253 gènes, aucun isoforme n'est par conséquent spécifique à une ou deux des espèces.

De plus, pour 135 de ces 253 gènes, tous les CDS ont une relation d'orthologie bijective ("*un-à-un*"), et donc un et un seul transcrit porteur de ce CDS dans chaque espèce. Les 118 autres gènes restants, quant à eux, ont aussi une conservation de leurs CDS chez les trois espèces mais possèdent plusieurs transcrits encodant un même CDS chez une même espèce, résultant d'une activité de transcription alternative. Ces transcrits "redondants" dans une même espèce ont une même partie codante et des UTR différents. Examiner la conservation des UTR pour déterminer quels sont les transcrits complets orthologues est une perspective de cette thèse.

Les résultats principaux de la thèse sont :

— la formalisation de l'orthologie structurelle entre gènes et entre transcrits,
— la prédiction de transcrits chez l'humain, la souris et le chien (6 861 transcrits prédits),
— la mise en œuvre d'une méthode de recherche de transcrits prédits dans des données auxiliaires,
— l'identification de relations d'orthologie structurelle entre des transcrits connus,
— la définition d'une méthode de comparaison de structure d'épissage multi-espèces (appliquée à trois espèces),
— l'identification d'un ensemble de 253 gènes dont la structure d'épissage est conservée entre les trois espèces,
— l'identification des différents transcrits exprimables pour chaque gène (inventaire des différentes protéines réalisables) au travers de groupes de CDS orthologues,
— la distribution des données sous la forme d'une base de données.

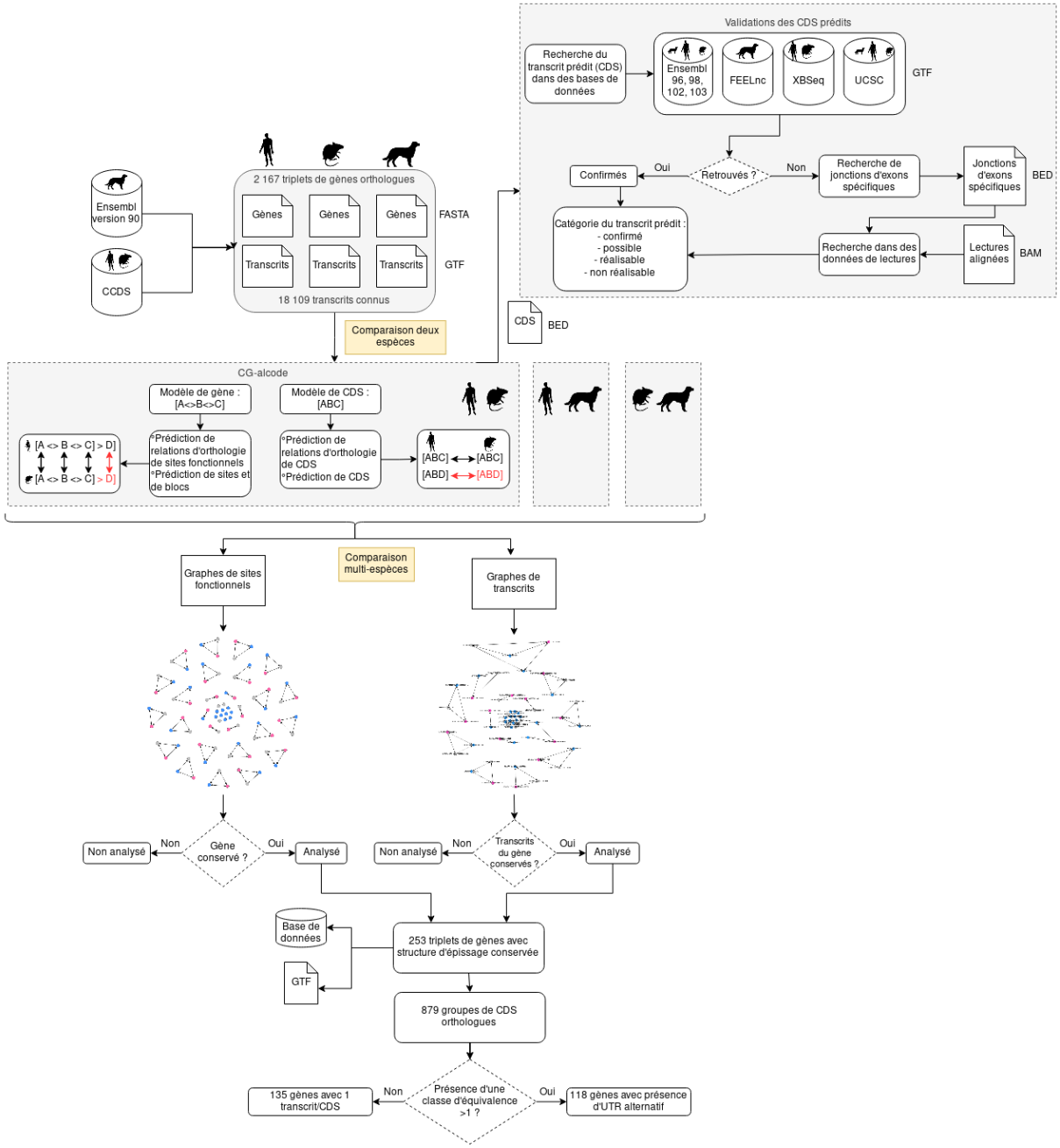


FIGURE 5.6 – Pipeline général des réalisations de la thèse

Bibliographie

Alexandre Favoretto Galante, Pedro, Noboru JO Sakabe, Natanja Kirschbaumslager et Sandro DE José Souza (2004), « Detection and evaluation of intron retention events in the human transcriptome », in : doi : 10.1261/rna.5123504.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers et David J. Lipman (1990), « Basic local alignment search tool », in : *Journal of Molecular Biology* 215.*3*, p. 403-410, issn : 00222836, doi : 10.1016/S0022-2836(05)80360-2.

Bao, Suying, Daniel F. Moakley et Chaolin Zhang (jan. 2019), *The Splicing Code Goes Deep*, doi : 10.1016/j.cell.2019.01.013.

Blanquart, Samuel, Jean-Stéphane Varré, Paul Guertin, Amandine Perrin, Anne Bergeron et Krister M Swenson (nov. 2016), « Assisted transcriptome reconstruction and splicing orthology », in : *BMC Genomics* 17.*10*, p. 786, issn : 1471-2164, doi : 10.1186/s12864-016-3103-6.

Blencowe, Benjamin J. (juin 2017), *The Relationship between Alternative Splicing and Proteomic Complexity*, doi : 10.1016/j.tibs.2017.04.001.

Breschi, Alessandra, Thomas R. Gingeras et Roderic Guigó (juil. 2017), *Comparative transcriptomics in human and mouse*, doi : 10.1038/nrg.2017.19.

Bronner, Iraad F., Michael A. Quail, Daniel J. Turner et Harold Swerdlow (2014), « Improved protocols for Illumina sequencing », in : *Current Protocols in Human Genetics* 80.*SUPPL.80*, issn : 19348258, doi : 10.1002/0471142905.hg1802s80.

Bu, Jingde, Xuebin Chi et Zhong Jin (déc. 2013), « HSA : A Heuristic Splice Alignment Tool », in : *BMC Systems Biology* 7.*2*, p. 1-6, issn : 17520509, doi : 10.1186/1752-0509-7-S2-S10.

Cartegni, Luca, Shern L. Chew et Adrian R. Krainer (2002), *Listening to silence and understanding nonsense : Exonic mutations that affect splicing*, doi : 10.1038/nrg775.

Chaudhary, Saurabh, Waqas Khokhar, Ibtissam Jabre, Anireddy S.N. Reddy, Lee J. Byrne, Cornelia M. Wilson et Naeem H. Syed (mai 2019), *Alternative splicing and protein diversity : Plants versus animals*, doi : 10.3389/fpls.2019.00708.

Chen, Hung I.Harry, Yuanhang Liu, Yi Zou, Zhao Lai, Devanand Sarkar, Yufei Huang et Yidong Chen (juin 2015), « Differential expression analysis of RNA se-

quencing data by incorporating non-exonic mapped reads », in : *BMC Genomics* 16.*7*, S14, ISSN : 14712164, DOI : 10.1186/1471-2164-16-S7-S14.

COLLINS, Francis S., Eric D. GREEN, Alan E. GUTTMACHER et Mark S. GUYER (avr. 2003), *A vision for the future of genomics research*, DOI : 10.1038/nature01626.

CRAIG VENTER, J. et al. (fév. 2001), « The sequence of the human genome », in : *Science* 291.*5507*, p. 1304-1351, ISSN : 00368075, DOI : 10.1126/science.1058040.

DENTI, Luca, Raffaella RIZZI, Stefano BERETTA, Gianluca Della VEDOVA, Marco PREVITALI et Paola BONIZZONI (déc. 2018), « ASGAL : Aligning RNA-Seq data to a splicing graph to detect novel alternative splicing events », in : *BMC Bioinformatics* 19.*1*, p. 1-21, ISSN : 14712105, DOI : 10.1186/s12859-018-2436-3.

ELLIS, Jonathan D. et al. (juin 2012), « Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks », in : *Molecular Cell* 46.*6*, p. 884-892, ISSN : 10972765, DOI : 10.1016/j.molcel.2012.05.037.

FOLEY, Nicole M., Mark S. SPRINGER et Emma C. TEELING (juil. 2016), *Mammal madness : Is the mammal tree of life not yet resolved ?*, DOI : 10.1098/rstb.2015.0140.

FRANKISH, Adam et al. (jan. 2021), « GENCODE 2021 », in : *Nucleic Acids Research* 49.*D1*, p. D916-D923, ISSN : 13624962, DOI : 10.1093/nar/gkaa1087, URL : https://www.gencodegenes.org..

GENCODE - Human Release Statistics (2021), URL : https://www.gencodegenes.org/human/stats.html (visité le 27/07/2021).

GILBERT, Walter (1978), *Why genes in pieces ?*, DOI : 10.1038/271501a0.

HALFON, Marc S. (mar. 2020), *Silencers, Enhancers, and the Multifunctional Regulatory Genome*, DOI : 10.1016/j.tig.2019.12.005.

HEBER, Steffen, Max ALEKSEYEV, Sing-Hoi SZE, Haixu TANG et Pavel A. PEVZNER (juil. 2002), « Splicing graphs and EST assembly problem », in : *Bioinformatics* 18.*suppl_1*, S181-S188, ISSN : 1367-4803, DOI : 10.1093/bioinformatics/18.suppl_1.S181.

HOEPPNER, Marc P et al. (2014), « An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts », in : *PLoS ONE* 9.*3*, e91172, ISSN : 19326203, DOI : 10.1371/journal.pone.0091172.

HOWE, Kevin L. et al. (jan. 2021), « Ensembl 2021 », in : *Nucleic Acids Research* 49.*D1*, p. D884-D891, ISSN : 13624962, DOI : 10.1093/nar/gkaa942.

HUANG, Songbo, Jinbo ZHANG, Ruiqiang LI, Wenqian ZHANG, Zengquan HE, Tak-Wah LAM, Zhiyu PENG et Siu-Ming YIU (2011), « SOAPsplice : Genome-Wide ab initio

Detection of Splice Junctions from RNA-Seq Data », in : *Frontiers in Genetics* 2.*JULY*, p. 46, ISSN : 1664-8021, DOI : 10.3389/fgene.2011.00046.

JAMMALI, Safa, Jean David AGUILAR, Esaie KUITCHE et Aïda OUANGRAOUA (2019), « SplicedFamAlign : CDS-to-gene spliced alignment and identification of transcript orthology groups », in : *BMC Bioinformatics* 20.*Suppl 3*, ISSN : 14712105, DOI : 10.1186/s12859-019-2647-2.

KALSOTRA, Auinash et Thomas A. COOPER (oct. 2011), *Functional consequences of developmentally regulated alternative splicing*, DOI : 10.1038/nrg3052.

KAPUSTIN, Yuri, Alexander SOUVOROV, Tatiana TATUSOVA et David LIPMAN (mai 2008), « Splign : Algorithms for computing spliced alignments with identification of paralogs », in : *Biology Direct* 3.*1*, p. 1-13, ISSN : 17456150, DOI : 10.1186/1745-6150-3-20.

KAZAK, Lawrence et al. (fév. 2013), « Alternative translation initiation augments the human mitochondrial proteome », in : *Nucleic Acids Research* 41.*4*, p. 2354-2369, ISSN : 03051048, DOI : 10.1093/nar/gks1347.

KELEMEN, Olga, Paolo CONVERTINI, Zhaiyi ZHANG, Yuan WEN, Manli SHEN, Marina FALALEEVA et Stefan STAMM (2013), « Function of alternative splicing », in : *Gene* 514.*1*, p. 1-30, ISSN : 03781119, DOI : 10.1016/j.gene.2012.07.083.

KENT, W. J. (mar. 2002), « BLAT—The BLAST-Like Alignment Tool », in : *Genome Research* 12.*4*, p. 656-664, ISSN : 1088-9051, DOI : 10.1101/gr.229202.

KIM, Daehwan et Steven L. SALZBERG (août 2011), « TopHat-Fusion : An algorithm for discovery of novel fusion transcripts », in : *Genome Biology* 12.*8*, R72, ISSN : 14747596, DOI : 10.1186/gb-2011-12-8-r72.

KIRKNESS, Ewen F. et al. (sept. 2003), « The dog genome : Survey sequencing and comparative analysis », in : *Science* 301.*5641*, p. 1898-1903, ISSN : 00368075, DOI : 10.1126/science.1086432.

KOREN, Eli, Galit LEV-MAOR et Gil AST (2007), « The emergence of alternative 3' and 5' splice site exons from constitutive exons », in : *PLoS Computational Biology* 3.*5*, p. 0895-0908, ISSN : 15537358, DOI : 10.1371/journal.pcbi.0030095.

KWAN, Thaddaeus et Sunnie R. THOMPSON (avr. 2019), « Noncanonical translation initiation in eukaryotes », in : *Cold Spring Harbor Perspectives in Biology* 11.*4*, a032672, ISSN : 19430264, DOI : 10.1101/cshperspect.a032672.

LANGMEAD, Ben, Cole TRAPNELL, Mihai POP et Steven L. SALZBERG (mar. 2009), « Ultrafast and memory-efficient alignment of short DNA sequences to the human

genome », in : *Genome Biology* 10.*3*, R25, ISSN : 14747596, DOI : 10.1186/gb-2009-10-3-r25.

LE BÉGUEC, Céline et al. (déc. 2018), « Characterisation and functional predictions of canine long non-coding RNAs », in : *Scientific Reports* 8.*1*, ISSN : 20452322, DOI : 10.1038/s41598-018-31770-2.

LEE, Bernett, Shoba RANGANATHAN et Tin TAN (2003), *MGAlign, a Reduced Search Space Approach to the Alignment of mRNA Sequences to Genomic Sequences*, rapp. tech., p. 474-475, DOI : 10.11234/GI1990.14.474.

LI, H. et R. DURBIN (juil. 2009), « Fast and accurate short read alignment with Burrows-Wheeler transform », in : *Bioinformatics* 25.*14*, p. 1754-1760, ISSN : 1367-4803, DOI : 10.1093/bioinformatics/btp324.

— (mar. 2010), « Fast and accurate long-read alignment with Burrows-Wheeler transform », in : *Bioinformatics* 26.*5*, p. 589-595, ISSN : 13674803, DOI : 10.1093/bioinformatics/btp698.

LI, Hong-Dong, Changhuo YANG, Zhimin ZHANG, Mengyun YANG, Fang-Xiang WU, Gilbert S OMENN et Jianxin WANG (mai 2021), « IsoResolve : predicting splice isoform functions by integrating gene and isoform-level features with domain adaptation », in : *Bioinformatics* 37.*4*, sous la dir. d'Inanc BIROL, p. 522-530, ISSN : 1367-4803, DOI : 10.1093/bioinformatics/btaa829.

LI, R., Y. LI, K. KRISTIANSEN et J. WANG (mar. 2008), « SOAP : short oligonucleotide alignment program », in : *Bioinformatics* 24.*5*, p. 713-714, ISSN : 1367-4803, DOI : 10.1093/bioinformatics/btn025.

MARCHESE, Francesco P., Ivan RAIMONDI et Maite HUARTE (oct. 2017), *The multidimensional mechanisms of long noncoding RNA function*, DOI : 10.1186/s13059-017-1348-2.

MARCHET, Camille, Pierre MORISSE, Lolita LECOMPTE, Arnaud LEFEBVRE, Thierry LECROQ, Pierre PETERLONGO et Antoine LIMASSET (mar. 2020), « ELECTOR : evaluator for long reads correction methods », in : *NAR Genomics and Bioinformatics* 2.*1*, DOI : 10.1093/NARGAB/LQZ015.

MATTHEWS, Benjamin J. et Leslie B. VOSSHALL (fév. 2020), *How to turn an organism into a model organism in 10 'easy' steps*, DOI : 10.1242/jeb.218198.

MATTICK, John S. et Igor V. MAKUNIN (avr. 2006), *Non-coding RNA.* DOI : 10.1093/hmg/ddl046.

MAYR, Christine (oct. 2019), « What are 3' utrs doing ? », in : *Cold Spring Harbor Perspectives in Biology* 11.*10*, a034728, ISSN : 19430264, DOI : 10.1101/cshperspect.a034728.

MAZIN, Pavel et al. (2013), « Widespread splicing changes in human brain development and aging », in : *Molecular Systems Biology* 9, ISSN : 17444292, DOI : 10.1038/msb.2012.67.

MEYER, Corentin et al. (déc. 2020), « Understanding the causes of errors in eukaryotic protein-coding gene prediction : a case study of primate proteomes », in : *BMC Bioinformatics* 21 (1), p. 513, ISSN : 14712105, DOI : 10.1186/s12859-020-03855-1.

MIGA, Karen H. et al. (sept. 2020), « Telomere-to-telomere assembly of a complete human X chromosome », in : *Nature* 585.*7823*, p. 79-84, ISSN : 14764687, DOI : 10.1038/s41586-020-2547-7.

MULLER, Ittai B. et al. (déc. 2021), « Computational comparison of common event-based differential splicing tools : practical considerations for laboratory researchers », in : *BMC Bioinformatics* 22.*1*, p. 347, ISSN : 1471-2105, DOI : 10.1186/s12859-021-04263-9.

NAVARRO GONZALEZ, Jairo et al. (jan. 2021), « The UCSC genome browser database : 2021 update », in : *Nucleic Acids Research* 49.*D1*, p. D1046-D1057, ISSN : 13624962, DOI : 10.1093/nar/gkaa1070.

NILSEN, Timothy W. et Brenton R. GRAVELEY (jan. 2010), *Expansion of the eukaryotic proteome by alternative splicing*, DOI : 10.1038/nature08909.

OLIVA, Meritxell et al. (sept. 2020), « The impact of sex on gene expression across human tissues », in : *Science* 369.*6509*, ISSN : 10959203, DOI : 10.1126/SCIENCE.ABA3066.

OUANGRAOUA, Aïda, Krister M. SWENSON et Anne BERGERON (2012), « On the comparison of sets of alternative transcripts », in : *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, t. 7292 LNBI, Springer, Berlin, Heidelberg, p. 201-212, ISBN : 9783642301902, DOI : 10.1007/978-3-642-30191-9_19.

PAN, Qun, Ofer SHAI, Leo J. LEE, Brendan J. FREY et Benjamin J. BLENCOWE (déc. 2008), « Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing », in : *Nature Genetics* 40.*12*, p. 1413-1415, ISSN : 10614036, DOI : 10.1038/ng.259.

Philippe, Nicolas, Mikaël Salson, Thérèse Commes et Eric Rivals (mar. 2013), « CRAC : An integrated approach to the analysis of RNA-seq reads », in : *Genome Biology* 14.*3*, R30, issn : 1474760X, doi : 10.1186/gb-2013-14-3-r30.

Pujar, Shashikant et al. (jan. 2018), « Consensus coding sequence (CCDS) database : A standardized set of human and mouse protein-coding regions supported by expert curation », in : *Nucleic Acids Research* 46.*D1*, p. D221-D228, issn : 13624962, doi : 10.1093/nar/gkx1031.

Quinlan, Aaron R. et Ira M. Hall (jan. 2010), « BEDTools : A flexible suite of utilities for comparing genomic features », in : *Bioinformatics* 26.*6*, p. 841-842, issn : 13674803, doi : 10.1093/bioinformatics/btq033.

Rizk, Guillaume et Dominique Lavenier (août 2010), « GASSST : Global alignment short sequence search tool », in : *Bioinformatics* 26.*20*, p. 2534-2540, issn : 13674803, doi : 10.1093/bioinformatics/btq485.

Sammeth, M, S Foissac et R Guigó (2008), « A General Definition and Nomenclature for Alternative Splicing Events », in : *PLoS Comput Biol* 4.*8*, p. 1000147, doi : 10.1371/journal.pcbi.1000147, url : http://genome.imim.es/astalavista.

Scheidt, Moritz von, Yuqi Zhao, Zeyneb Kurt, Calvin Pan, Lingyao Zeng, Xia Yang, Heribert Schunkert et Aldons J. Lusis (fév. 2017), *Applications and Limitations of Mouse Models for Understanding Human Atherosclerosis*, doi : 10.1016/j.cmet.2016.11.001.

Schmitz, Sandra U., Phillip Grote et Bernhard G. Herrmann (juil. 2016), *Mechanisms of long noncoding RNA function in development and disease*, doi : 10.1007/s00018-016-2174-5.

Sessegolo, Camille, Corinne Cruaud, Corinne Da Silva, Audric Cologne, Marion Dubarry, Thomas Derrien, Vincent Lacroix et Jean-Marc Aury (déc. 2019), « Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules », in : *Scientific Reports* 9.*1*, p. 14908, issn : 2045-2322, doi : 10.1038/s41598-019-51470-9.

Sibley, Christopher R., Lorea Blazquez et Jernej Ule (juil. 2016), *Lessons from non-canonical splicing*, doi : 10.1038/nrg.2016.46.

Smithers, Ben, Matt Oates et Julian Gough (juin 2019), « 'Why genes in pieces ?' - Revisited », in : *Nucleic Acids Research* 47.*10*, p. 4970-4973, issn : 13624962, doi : 10.1093/nar/gkz284.

Söllner, Julia F., German Leparc, Tobias Hildebrandt, Holger Klein, Leo Thomas, Elia Stupka et Eric Simon (déc. 2017), « An RNA-Seq atlas of gene expression in mouse and rat normal tissues », in : *Scientific Data* 4.*1*, p. 170185, issn : 2052-4463, doi : 10.1038/sdata.2017.185.

Steijger, Tamara et al. (déc. 2013), « Assessment of transcript reconstruction methods for RNA-seq », in : *Nature Methods* 10.*12*, p. 1177-1184, issn : 15487091, doi : 10.1038/nmeth.2714.

Sulakhe, Dinanath et al. (2019), *Exploring the functional impact of alternative splicing on human protein isoforms using available annotation sources*, doi : 10.1093/bib/bby047.

Taneri, Bahar, Esra Asilmaz et Terry Gaasterland (fév. 2012), « Biomedical impact of splicing mutations revealed through exome sequencing », in : *Molecular Medicine* 18.*2*, p. 314-319, issn : 10761551, doi : 10.2119/molmed.2011.00126.

Trapnell, Cole, Lior Pachter et Steven L. Salzberg (2009), « TopHat : Discovering splice junctions with RNA-Seq », in : *Bioinformatics* 25.*9*, p. 1105-1111, issn : 13674803, doi : 10.1093/bioinformatics/btp120.

Tung, Kuo Feng, Chao Yu Pan, Chao Hsin Chen et Wen chang Lin (déc. 2020), « Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset », in : *Scientific Reports* 10.*1*, p. 16245, issn : 20452322, doi : 10.1038/s41598-020-73081-5.

Wang, Dongxue et al. (fév. 2019), « A deep proteome and transcriptome abundance atlas of 29 healthy human tissues », in : *Molecular Systems Biology* 15.*2*, issn : 1744-4292, doi : 10.15252/msb.20188503.

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth et Christopher B. Burge (nov. 2008), « Alternative isoform regulation in human tissue transcriptomes », in : *Nature* 456.*7221*, p. 470-476, issn : 00280836, doi : 10.1038/nature07509.

Wang, Kai et al. (août 2010), « MapSplice : Accurate mapping of RNA-seq reads for splice junction discovery », in : *Nucleic Acids Research* 38.*18*, issn : 03051048, doi : 10.1093/nar/gkq622.

Wang, Yan et al. (mar. 2015), « Mechanism of alternative splicing and its regulation », in : *Biomedical Reports* 3.*2*, p. 152-158, issn : 2049-9434, doi : 10.3892/br.2014.407.

Wu, T. D. et S. Nacu (avr. 2010), « Fast and SNP-tolerant detection of complex variants and splicing in short reads », in : *Bioinformatics* 26.*7*, p. 873-881, issn : 1367-4803, doi : 10.1093/bioinformatics/btq057.

Wu, T. D. et C. K. Watanabe (mai 2005), « GMAP : a genomic mapping and alignment program for mRNA and EST sequences », in : *Bioinformatics* 21.*9*, p. 1859-1875, issn : 1367-4803, doi : 10.1093/bioinformatics/bti310.

Wucher, Valentin et al. (2017), « FEELnc : a tool for long non-coding RNA annotation and its application to the dog transcriptome », in : 45.*8*, e57, issn : 0305-1048, doi : 10.1093/nar/gkw1306.

Xiong, Jieyi et al. (avr. 2018), « Predominant patterns of splicing evolution on human, chimpanzee and macaque evolutionary lineages », in : *Hum. Mol. Genet.* 27 (8), p. 1474-1485, issn : 14602083, doi : 10.1093/hmg/ddy058.

Zambelli, Federico, Giulio Pavesi, Carmela Gissi, David S. Horner et Graziano Pesole (oct. 2010), « Assessment of orthologous splicing isoforms in human and mouse orthologous genes », in : *BMC Genomics* 11.*1*, p. 534, issn : 14712164, doi : 10.1186/1471-2164-11-534.

Zea, Diego Javier, Sofya Laskina, Alexis Baudin, Hugues Richard et Elodie Laine (août 2021), « Assessing conservation of alternative splicing with evolutionary splicing graphs », in : *Genome Research* 31.*8*, p. 1462-1473, issn : 15495469, doi : 10.1101/gr.274696.120.

Liste des acronymes

A : Adénine
ADN : Acide désoxyribonucléique
ARN : Acide ribonucléique
ARNm : ARN messager
BAM : Binary Alignment Map
BLAST : Basic Local Alignment Search Tool
BLAT : BLAST-Like Alignment Tool
C : Cytosine
CDS : Coding sequence
CG-alcode : Comparative genomics for alternative coding in eukaryote genes
G : Guanine
NMD : Nonsense Mediated Decay
pré-ARNm : ARN prémessager
RNA-seq : RNA sequencing
T : Thymine
SAP : Spliced alignment problem
snRNP : Small nuclear ribonucleoprotein
UCSC : University of California, Santa Cruz
UTR : Untranslated region

Table des figures

1.1 L'information génétique chez un organisme eucaryote 18
1.2 Les mécanismes généraux d'expression des gènes codants 19
1.3 Régulation de la transcription 20
1.4 Maturation d'un pré-ARNm 21
1.5 Traduction d'un ARNm 22
1.6 Schéma de l'expression d'un gène 23
1.7 Mécanismes de l'expression alternative des gènes 26
1.8 Types d'évènements de l'épissage alternatif 28
1.9 Extrait d'un fichier FASTA pour le gène CREM humain. 29
1.10 Extrait simplifié d'un fichier GTF pour le gène CREM humain 30
1.11 Représentation des transcrits dans les visualiseurs de génomes 32
1.12 Construction d'un graphe d'épissage 33
1.13 Alignement des *k-mers* sur le génome 38
1.14 Représentation de la méthode CRAC sur l'alignement des *k-mers* en présence d'une variation de type SNP (*single nucleotide polymorphism*) 39
1.15 Détails des trois problèmes d'alignement épissé (SAP) 40
1.16 SplicedFamAlign : identification de CDS orthologues 41
1.17 Construction d'un graphe d'épissage évolutif 42
1.18 Alignement d'un site fonctionnel contre la séquence d'un gène orthologue . 46
1.19 Principes de la méthode CG-alcode 47
1.20 Relations phylogénétiques entre l'humain, la souris et le chien 49
1.21 Passage à une échelle multi-espèces pour comparer des structures de gènes 50

2.1 Définition des blocs codants d'un exon 57
2.2 Exemple de modèles de structure de gènes 58
2.3 Exemple de modèles de structure de transcrits 59
2.4 Construction des modèles de structure de gène 60
2.5 Comparaison de deux modèles de gènes 61
2.6 Paire de CDS connus structurellement conservés 63

2.7 Transcrit orthologue prédit et exécutabilité des modèles des transcrits du gène i par le modèle du gène j 64
2.8 Identification de la structure des gènes et des transcrits par comparaison de gènes et prédiction des CDS des transcrits 65
2.9 Graphe de sites fonctionnels : identification des unités lexicales de gènes partagées entre plusieurs espèces 67
2.10 Graphe de transcrits : identification de groupes de CDS orthologues structurellement conservés chez trois espèces 68

3.1 Prédiction de l'exécutabilité des transcrits connus du gène CREM 74
3.2 Prédiction de transcrits à partir de transcrits connus 76
3.3 Protocole de vérification des transcrits prédits 79
3.4 Confirmation des transcrits prédits à partir des transcrits connus dans les bases de données. 80
3.5 Illustration d'une jonction d'exons 82
3.6 Mise en évidence d'une jonction d'exons spécifique 82
3.7 Exemple d'une jonction d'exons spécifique visualisée sous le logiciel IGV . 84
3.8 Obtention des jonctions d'exons de transcrits au format BED12 85
3.9 Alignement parfait des lectures issues d'un échantillon de glande surrénale chez le chien sur une jonction d'exons spécifique d'un transcrit prédit . . . 86
3.10 Vérification des transcrits prédits réalisables en alignant des lectures courtes sur les jonctions d'exons spécifiques pour l'humain, la souris et chien. . . . 86
3.11 Alignement de lectures sur des jonctions d'exons spécifiques 89
3.12 Fréquence de la couverture en lectures des jonctions d'exons connues. . . . 91
3.13 Résultats de la confirmation expérimentale de l'existence des transcrits prédits 91

4.1 Passage des alignements de paires de gènes à un graphe multi-espèces . . . 97
4.2 Pipeline de construction des graphes de sites fonctionnels et de transcrits pour un triplet de gènes orthologues 98
4.3 Interprétation phylogénétique des graphes de sites fonctionnels et de transcrits sur l'humain, la souris et le chien 99
4.4 Hypothèse pour la comparaison de paires de gènes : conservation de la colinéarité des exons des gènes 101

4.5 Bijection des relations d'orthologie pour la construction de graphes de sites fonctionnels . . . 102
4.6 Cas de sites fonctionnels non alignés entre certaines espèces. . . . 103
4.7 Composants présents dans les graphes de sites fonctionnels . . . 104
4.8 Composants présents dans les graphes de transcrits . . . 105
4.9 Représentation du graphe de sites fonctionnels du gène CREM chez l'humain, la souris et le chien . . . 107
4.10 Construction des modèles de structure du gène CREM trois espèces à partir du graphe de sites fonctionnels . . . 108
4.11 Représentation du graphe de transcrits du gène CREM chez l'humain, la souris et le chien . . . 109
4.12 Pipeline d'analyse des graphes de sites fonctionnels . . . 110
4.13 Diagramme de Venn montrant la distribution des graphes de sites fonctionnels sur les 1 661 triplets de gènes orthologues . . . 112
4.14 Répartition des 253 gènes dans les annotations de processus biologiques de la *Gene Ontology* . . . 113
4.15 Diagramme de Venn montrant la distribution des topologies des graphes de transcrits sur les 986 triplets de gènes orthologues sélectionnés . . . 114
4.16 Distribution du nombre de transcrits connus et prédits dans le cas des 135 triplets de gènes orthologues . . . 116
4.17 Représentation des classes d'équivalence au sein des groupes de CDS orthologues pour un triplet de gène . . . 117
4.18 Schéma relationnel de la base de données "*transcript_ortho*". . . . 120

5.1 Modèles de structure des transcrits contenus dans une classe d'équivalence du gène de la souris ENSMUSG00000070780 . . . 123
5.2 Modèles de structure des transcrits contenus dans les trois classes d'équivalence humain, souris et chien d'un même groupe de CDS orthogues. . . . 124
5.3 Définition d'un intron et d'un exon . . . 126
5.4 Groupes de CDS orthologues pour le triplet de gènes ENSG00000001167 - ENSMUSG00000023994 - ENSCAFG00000001580 . . . 128
5.5 Alignement de lectures sur des jonctions d'exons spécifiques . . . 129
5.6 Pipeline général des réalisations de la thèse . . . 133

Liste des tableaux

1.1 Nombre de gènes codants et de transcrits connus chez l'humain dans la base de données Ensembl en mars 2021. 24
1.2 Ensemble des transcrits du gène CREM humain décrits en 2021 par la base de données Ensembl 25
1.3 Nombre de gènes et de transcrits connus chez l'humain, la souris et le chien sur la base de données Ensembl en Mars 2021. 49

2.1 Représentation des principales unités lexicales nécessaires à la construction de modèles de structure de gènes et de transcrits. 58

3.1 Vérification des transcrits prédits pour les 2 167 triplets de gènes orthologues à partir de bases de données 88
3.2 Vérification des transcrits prédits pour les 2 167 triplets de gènes orthologues avec des jonctions d'exons spécifiques trouvées dans des données de lectures issues de séquençage. 90
3.3 Vérification des transcrits prédits pour les 253 triplets de gènes orthologues à partir de bases de données 92
3.4 Vérification des transcrits prédits pour les 253 triplets de gènes orthologues avec des jonctions d'exons spécifiques trouvées dans des données de lectures 92

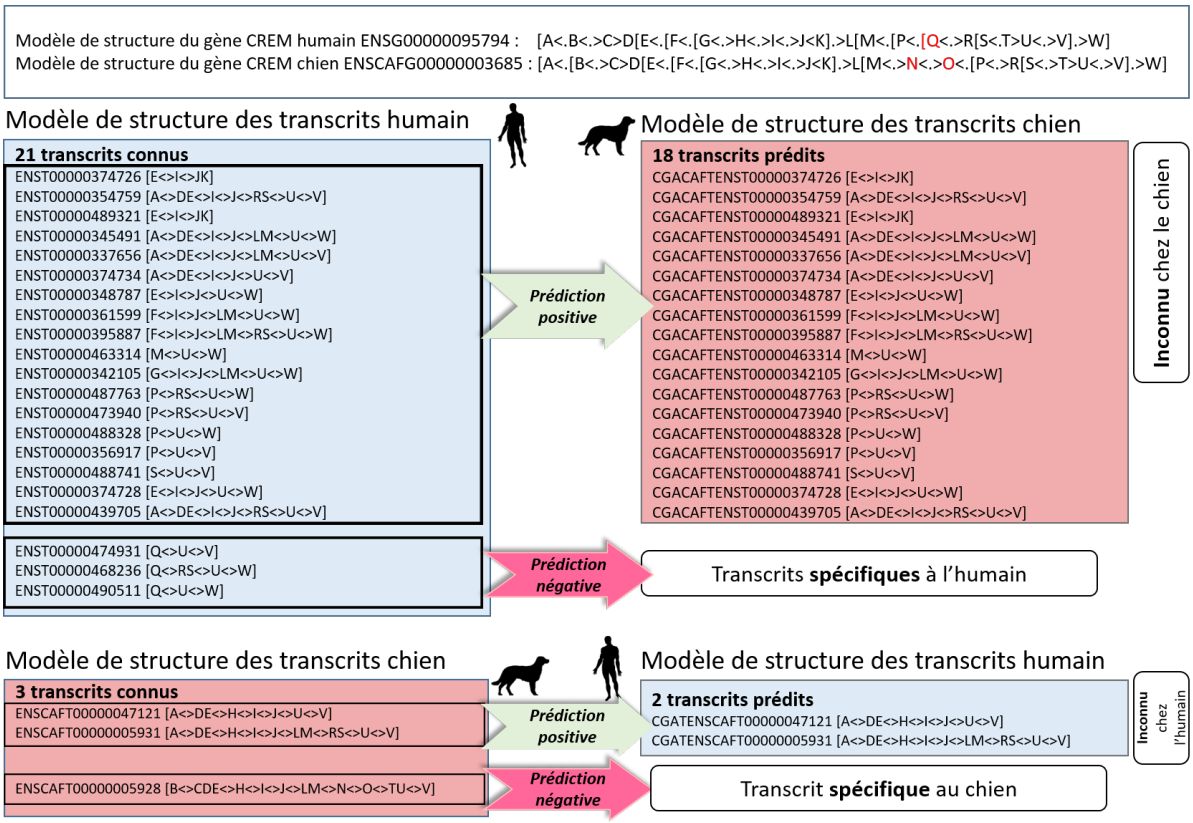
4.1 Caractéristiques des CDS des 253 triplets de gènes structurellement conservés. 117

Liste des algorithmes

1 Confirmation des transcrits prédits à partir de transcrits connus auxiliaires 81
2 Recherche de jonctions d'exons spécifiques dans des données expérimentales 87
3 Construction d'un graphe de sites fonctionnels pour un gène partagé par k espèces 104
4 Construction d'un graphe de transcrit pour un gène partagé par k espèces . 106

Annexes

Annexe 1 : Comparaisons de la paire de gènes CREM entre l'humain et le chien



Prédiction de l'exécutabilité sur le chien des transcrits connus chez l'humain (haut de la figure) et réciproquement (bas de la figure) du gène CREM. Les modèles obtenus pour les gènes CREM de l'humain et du chien sont présentés en haut de la figure.

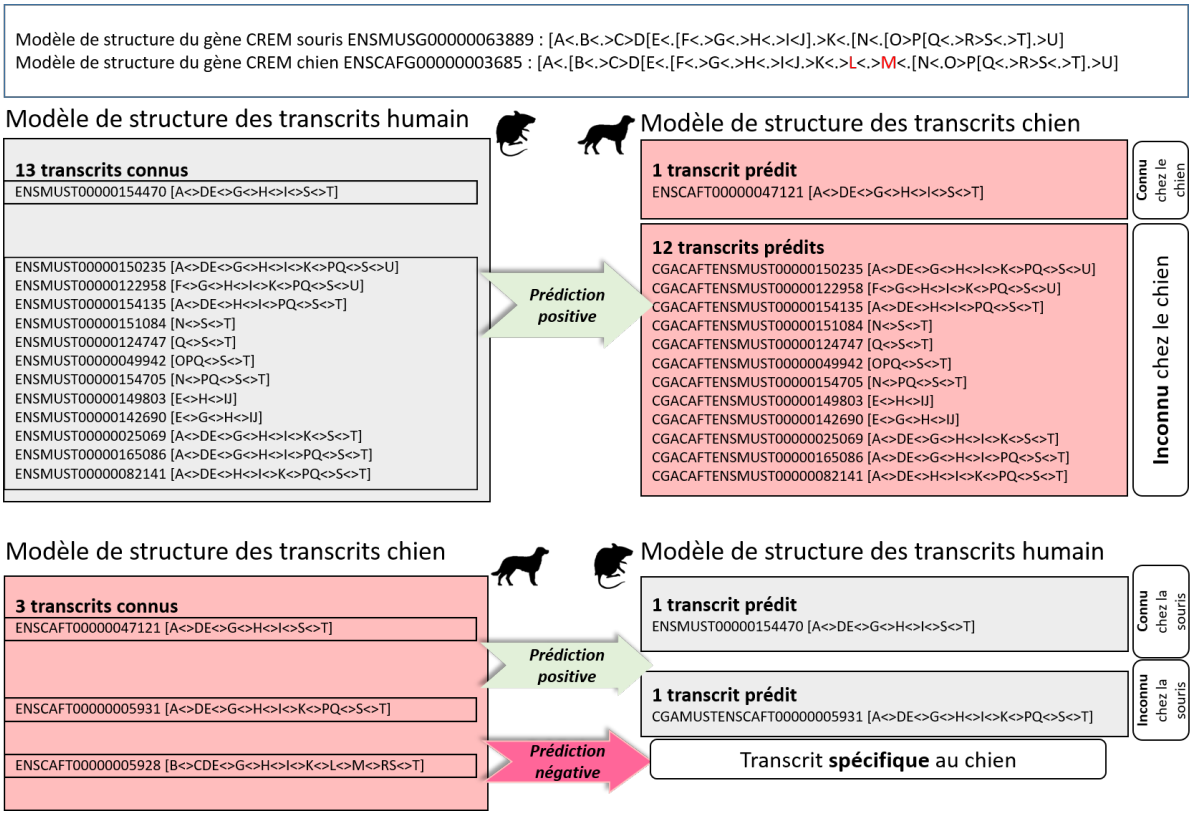
La comparaison des transcrits de l'humain avec le gène du chien concerne 21 transcrits connus. 18 sont exécutables par le modèle du gène CREM du chien. Tous étant inconnus chez le chien, ces 18 transcrits sont donc prédits chez le chien. 3 transcrits de l'humain ne sont pas exécutables par le gène CREM du chien et sont spécifiques au gène CREM humain (ils nécessitent le bloc codant Q qui n'a pas d'orthologue sur le gène du chien).

La comparaison des transcrits du chien avec le gène de l'humain concerne 3 transcrits connus. 2 sont exécutables chez l'humain et sont donc prédits chez l'humain. 1 transcrit n'est pas exécutable par le gène CREM de l'humain et est spécifique au gène CREM du chien (il nécessite les blocs et les sites $N <> 0$, non trouvés chez l'humain). Les sites

fonctionnels et les blocs codants spécifiques à l'une des deux espèces sont indiqués en rouge dans les modèles des gènes.

Bilan : 18 prédictions ont été faites chez le chien et 2 chez l'humain. Sur les 24 CDS différents au total, 20 sont partagés par les deux espèces, 3 sont spécifiques à l'humain et 1 est spécifique au chien.

Annexe 2 : Comparaisons de la paire de gènes CREM entre la souris et le chien



Prédiction de l'exécutabilité sur le chien des transcrits connus chez la souris (haut de la figure) et réciproquement (bas de la figure) du gène CREM. Les modèles obtenus pour les gènes CREM de la souris et du chien sont présentés en haut de la figure.

La comparaison des transcrits de la souris avec le gène du chien concerne 13 transcrits connus. Tous sont exécutables par le modèle du gène CREM du chien, dont 1 déjà connu parmi les transcrits du chien et 12 ne le sont pas et sont donc prédits chez le chien.

La comparaison des transcrits du chien avec le gène de la souris concerne 3 transcrits connus. 2 sont exécutables chez la souris, dont 1 déjà connu parmi les transcrits de la souris et 1 qui ne l'est pas et qui est prédit. 1 transcrit n'est pas exécutable par le gène CREM de la souris et est spécifique au gène CREM du chien (il nécessite les blocs et sites $L <> M$). Les sites fonctionnels et les blocs codants spécifiques à l'une des deux espèces sont indiqués en rouge dans les modèles des gènes.

Bilan : 12 prédictions ont été faites chez le chien et 1 chez la souris. Sur les 16 CDS différents au total, 15 sont partagés par les deux espèces, 1 est spécifique au chien.

Annexe 3 : Identifiants attribués aux transcrits prédits

Un identifiant est attribué à chaque transcrit prédit, en se référant au transcrit source. Il commencera par les lettres "CGA" pour qualifier qu'on l'a prédit par la méthode CGalcode chez un espèce avec l'identifiant Ensembl du transcrit source. Par exemple, si un transcrit souris "ENSMUST000x" est prédit exécutable chez l'humain, il est identifié "CGATENSMUST000x". Si un transcrit chien "ENSCAFT000x" est prédit exécutable chez l'humain sans qu'un transcrit souris n'ait permis de le prédire, il est identifié "CGATENSCAFT000x". On utilise ainsi l'identifiant de base d'Ensembl "ENS" suivit de "MUS" pour la souris, "CAF" pour le chien, et $\emptyset$ pour l'humain et "T" pour indiquer qu'il s'agit d'un transcrit suivi du numéro de référence du transcrit "00000000xxxxxx". Par exemple, le transcrit humain "$CREM - 206$" exprimé par le gène CREM est identifié dans Ensembl par "ENST00000354759" et le transcrit de la souris du gène orthologue "$Crem - 233$" est identifié par "ENSMUST00000154135".

Annexe 4 : Utilisation de BedTools intersect pour la recherche de transcrits prédits dans les données auxiliaires

4.1 : Description du contenu d'un fichier au format BED12 adapté aux transcrits

Description des 12 colonnes du format BED12 appliqué aux transcrits :

1. Nom du chromosome sur lequel est situé le transcrit sur le génome ;
2. Position de départ, à partir de zéro, du transcrit sur le génome ;
3. Position de fin, inclusive, du transcrit sur le génome ;
4. Nom du transcrit décrit ;
5. Score (non utilisé dans notre cas) ;
6. Brin, soit "+" soit "-", du sens de lecture du gène ;
7. Similaire à la colonne 2 ;
8. Similaire à la colonne 3 ;
9. Valeur Rgb (non utilisé dans notre cas) ;
10. Nombre d'exons du transcrit
11. Liste des tailles de chaque exon séparées par des virgules ;
12. Liste des positions de départ de chaque exon séparées par des virgules.

Exemple de représentation du CDS du transcrit $ENST00000291722$ au format BED12 :

```
9 133460066 133468653 ENST00000291722 0 + 133460066 133468653 0 4 121,73,108,91, 0,3416,7854,8496,
```

Le transcrit $ENST00000291722$, exprimé par le gène $ENSG00000160325$, est situé sur le chromosome 9 sur le brin "+" et est situé entre les positions 133 460 067 et 133 468 653 incluses. Il est composé de 4 exons de 121, 73, 108 et 91 nucléotides en taille qui sont situés aux positions relatives 0, 3 416, 7 854 et 8 496.

4.2 : Exécution de BedTools intersect pour la recherche de transcrits prédits dans des bases de données auxiliaires

Nous avons utilisé la ligne de commande suivante pour effectuer la recherche de correspondance entre transcrits prédits et transcrits connus dans des bases de données auxiliaires :

```
bedtools intersect -a prediction.bed -b reference.bed -wa -wb -f 1 -r -split
```

Les options "-wa" et "-wb" servent à donner la correspondance entre les deux transcrits trouvés, "-f 1" précise que le transcrit prédit contenu dans "prediction.bed" doit s'aligner de façon exacte sur le transcrit de référence décrit dans "reference.bed", "-r" précise la réciprocité de l'alignement, afin que le CDS du transcrit connu s'aligne parfaitement sur le transcrit prédit, et "-split" précise que les fichiers sont des fichiers au format BED.

4.3 : Exécution de BedTools intersect pour la recherche de jonctions d'exons spécifiques aux transcrits prédits

Nous avons utilisé la ligne de commande suivante pour effectuer la recherche de correspondance entre les jonctions d'exons spécifiques et les lectures issus de données de séquençage :

```
bedtools intersect -abam file.bam -b junction.bed -wa -wb -split -bed
```

Les options "-wa" et "-wb" servent à donner la correspondance entre la jonction d'exons ciblée et la lecture qui peut s'y aligner, "-split" précise que le fichier de jonctions d'exons est au format BED, "-abam" indique que le premier fichier est au format BAM, et "-bed" demande un fichier de sortie au format BED.

Annexe 5 : Détails des échantillons de tissus utilisés pour la confortation des jonctions d'exons spécifiques

5.1 : Échantillons de tissus utilisés chez l'humain

Échantillons	Tissus
1-2	Glande surrénale (x2)
3-4	Colon (x2)
5-6	Cortex (x2)
7-8	Duodénum (x2)
9-10	Œsophage (x2)
11-12	Cœur (x2)
13-24	Rein (x2)
15-16	Foie (x2)
17-18	Poumon (x2)
19-20	Ovaire (x2)
21-22	Pancréas (x2)
23-24	Peau (x2)
25-26	Rate (x2)
27-28	Estomac (x2)
29-30	Testicule (x2)
31-32	Glande thyroïde (x2)

Liste des 16 échantillons de tissus utilisés pour la recherche de jonctions d'exons spécifiques dans le cas des transcrits prédits chez l'humain. Ces 16 échantillons de tissus sont en duplicat. Ces données proviennent de D. Wang et al. 2019.

5.2 : Échantillons de tissus utilisés chez la souris

Échantillons	Tissus
1-3	Cerveau (x3)
4-6	Colon (x3)
7-9	Duodénum (x3)
10-12	Œsophage (x3)
13-14	Cœur (x2)
15-17	Iléon (x3)
18-20	Jéjunum (x3)
21-23	Rein (x3)
24-26	Foie (x3)
27-29	Muscle quadriceps (x3)
30-32	Pancréas (x3)
33-35	Estomac (x3)
36-38	Thymus (x3)

Liste des 13 échantillons de tissus utilisés pour la recherche de jonctions d'exons spécifiques dans le cas des transcrits prédits chez la souris. Ces 13 échantillons de tissus sont en triplicat à l'exception des échantillons de tissus du cœur qui sont en duplicat. Ces données proviennent de l'article de Söllner et al. 2017.

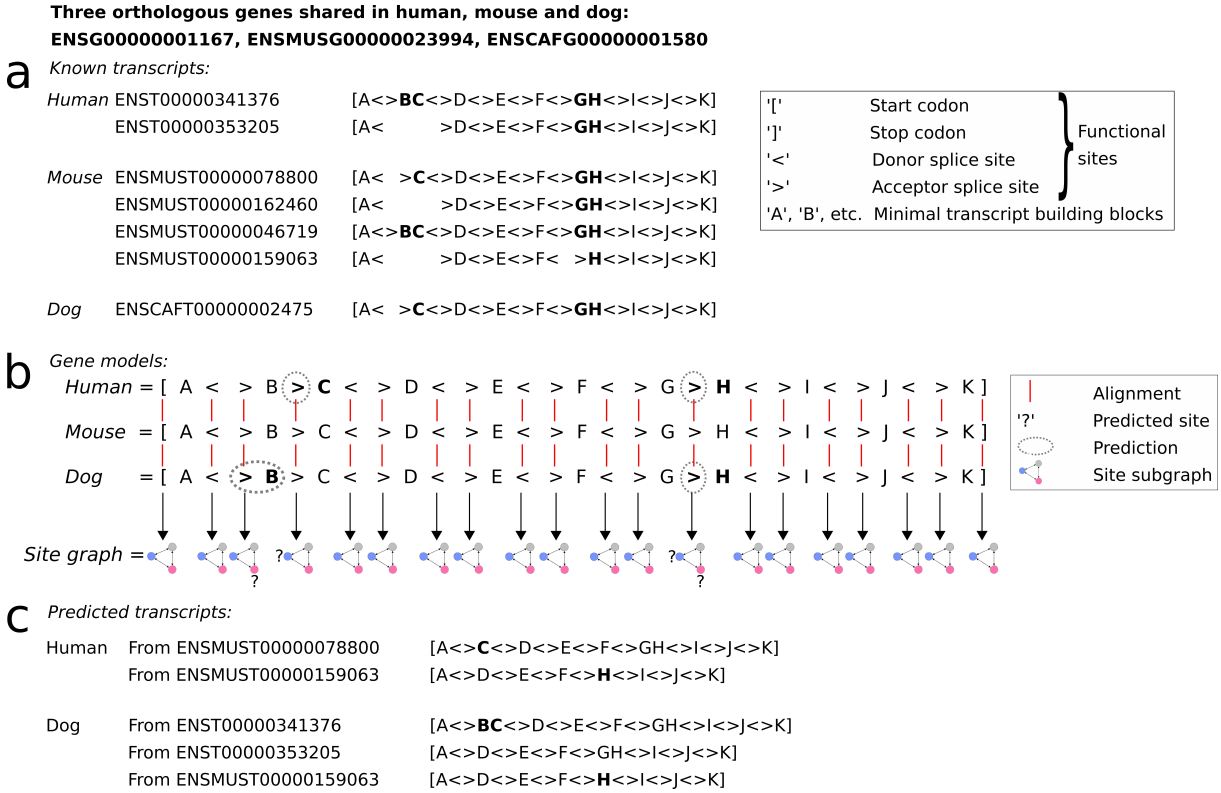
5.3 : Échantillons de tissus utilisés chez le chien

Échantillon	Tissus	Races	Références
1	Glande surrénale	Bouvier bernois	Wucher et al. 2017
2	Sang	Beagle	Hoeppner et al. 2014
3	Cerveau	Beagle	Hoeppner et al. 2014
4	Cervelet	Berger belge	Wucher et al. 2017
5	Cervelet	Grand Bouvier Suisse	Wucher et al. 2017
6	Cortex	Berger belge	Wucher et al. 2017
7	Côlon intestinal	Bouvier bernois	Wucher et al. 2017
8	Follicule pileux	Labrador	Wucher et al. 2017
9	Cœur	Beagle	Hoeppner et al. 2014
10	Jéjunum	Labrador	Wucher et al. 2017
11	Kératinocyte	Beagle	Wucher et al. 2017
12	Rein	Berger belge	Hoeppner et al. 2014
13	Foie	Beagle	Hoeppner et al. 2014
14	Poumon	Beagle	Hoeppner et al. 2014
15	Glande mammaire	Grand Bouvier Suisse	Wucher et al. 2017
16	Muqueuse buccale	Golden retriever	Le Béguec et al. 2018
17	Muqueuse buccale	Labrador	Le Béguec et al. 2018
18	Muqueuse buccale	Caniches	Le Béguec et al. 2018
19	Muscle	Beagle	Hoeppner et al. 2014
20	Nez	Labrador	Wucher et al. 2017
21	Nez	Labrador	Wucher et al. 2017
22	Nez	Labrador	Wucher et al. 2017
23	Bulbe olfactif	Grand Bouvier Suisse	Wucher et al. 2017
24	Ovaire	Beagle	Hoeppner et al. 2014
25	Pancréas	Berger belge	Wucher et al. 2017
26	Rétine	Border Collie	Wucher et al. 2017
27	Peau	Beagle	Wucher et al. 2017
28	Peau	Grand Bouvier Suisse	Wucher et al. 2017
29	Moelle épinière	Grand Bouvier Suisse	Wucher et al. 2017
30	Rate	Berger belge	Wucher et al. 2017
31	Testicule	Beagle	Hoeppner et al. 2014
32	Thymus	Saluki	Wucher et al. 2017

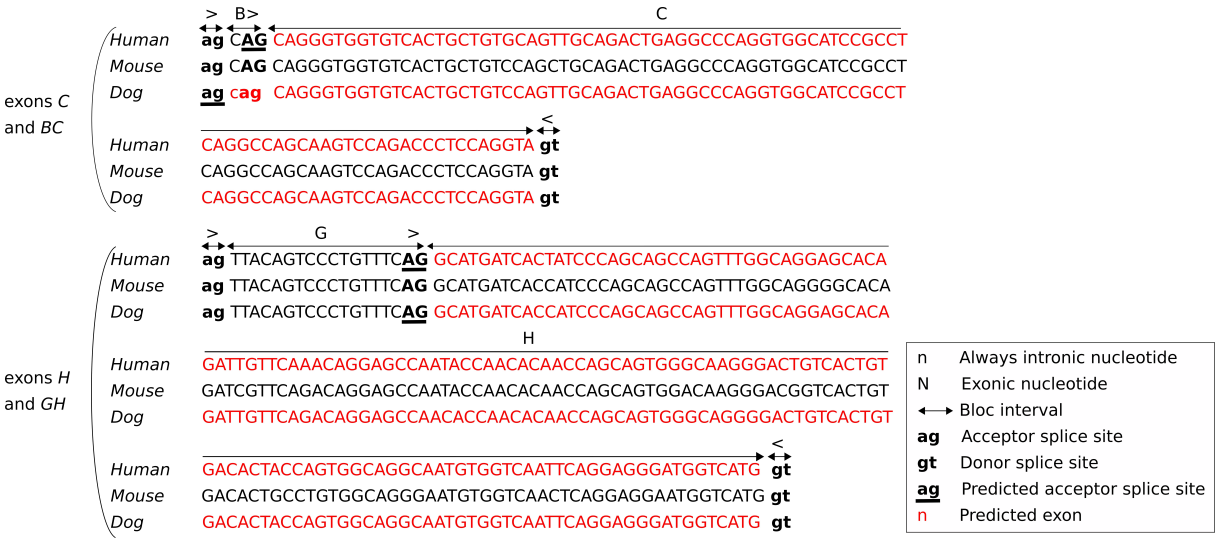
Liste des 32 échantillons de tissus utilisés pour la recherche de jonctions d'exons spécifiques dans le cas des transcrits prédits chez le chien. Ces 32 échantillons de tissus proviennent de trois études publiées (Le Béguec et al. 2018 ; Wucher et al. 2017 ; Hoeppner et al. 2014).

Annexe 6 : Détails des prédictions et des alignements sur un triplet de gènes

L'annexe présente un exemple des prédictions et des alignements réalisés pour le triplet de gènes $ENSG$00000001167 - $ENSMUSG$00000023994 - $ENSCAFG$00000001580. La première figure présente les transcrits connus pour chaque gène (a), les alignements des trois modèles de gènes (b) et les transcrits prédits obtenus (c). La seconde figure illustre l'alignement des blocs codants et des sites fonctionnels "$> B > C <$" et "$> G > H <$" prédits dans la première figure.



Prédiction de sites d'épissage et d'exons, conduisant à la prédiction de transcrits. (a) Modèles des CDS des transcrits connus. Ils font intervenir des *blocs codants*, et des *sites fonctionnels*.Les blocs codants correspondent aux intervalles entre les sites fonctionnels. Un même nom de bloc apparaissant chez plusieurs espèces indique une région conservée et orthologue. Par exemple, deux transcrits connus chez la souris impliquent des exons alternatifs dénommés C et BC, où les deux exons contiennent le bloc 'C', et le bloc 'B' est une extension 5' alternative de l'exon C. Un CDS humain connu implique un exon BC estimé être un orthologue de l'exon BC de la souris, et le CDS canin connu implique un exon C, orthologue de l'exon C de la souris. (b) Alignement de modèles de gènes. Chaque bloc et site d'un gène est aligné avec la séquence d'un gène orthologue, donnant des alignements de paires de gènes. Ces alignements révèlent i) l'orthologie entre les sites (ou blocs codants) déjà connus, ii) l'homologie de séquence des sites (ou blocs) connus avec des *loci* non annotés dans un autre gène, donnant des sites et blocs prédits (bulles en pointillés). Ici, l'alignement des gènes humain-souris révèle la présence d'un homologue humain du site accepteur de C, déclaré comme prédit. Il indique que le gène humain est capable d'exprimer l'exon C seul, sans la partie 'B'. Prédictions supplémentaires : le site accepteur H chez l'humain et le chien, et bloc 'B' plus son site accepteur chez le chien. Le *graphe de sites fonctionnel* résume les relations d'orthologie par paires : un nœud est un site fonctionnel et une arête est une relation d'orthologie. (c) Transcrits prédits. 5 transcrits sont rendues possibles par les prédictions de sites et de blocs.



Détails d'un alignement de séquences : prédiction des sites et blocs accepteurs chez l'humain et le chien. Deux alignements multiples sont présentés ici, fournissant plus de détails sur les prédictions illustrées dans la figure : les alignements des *loci* codant pour les exons C et BC, et les exons H et GH. Les majuscules indiquent les nucléotides impliqués dans les exons connus, et les minuscules indiquent les nucléotides introniques dont on n'a encore jamais observé l'appartenance à un exon. Par exemple, dans les gènes de l'humain et de la souris, le plus long exon BC est connu (en majuscules). Dans le gène du chien cependant, seul l'exon C plus court est connu et les nucléotides en amont ont été observés comme introniques jusqu'à présent (minuscules). Notez que l'exon C n'est pas connu chez l'humain (il n'apparaît dans aucun transcrit humain). Les sites d'épissage sont indiqués en gras, et les sites d'épissage prédits sont soulignés. Par exemple, un motif "AG" chez l'humain a été aligné avec l'accepteur connu "AG" de l'exon C chez la souris, ce qui donne un site d'épissage orthologue prédit chez l'humain ("AG" en gras souligné). Ce motif est maintenant identifié comme un site accepteur de l'exon C chez l'humain. De plus, un motif "ag" chez le chien a été aligné avec l'accepteur "ag" connu de l'exon BC chez la souris, ce qui a permis de prédire un site d'épissage orthologue chez le chien (souligné en gras "ag").Les nucléotides des exons prédits sont indiqués en rouge. Par exemple, un motif "cag" chez le chien a été aligné avec la séquence "CAG" du bloc connu 'B' chez la souris, ce qui donne un bloc prédit ("cag" rouge). En conséquence, des exons C et H plus courts peuvent exister chez l'humain (seuls les exons BC et GH plus longs étaient connus). Chez le chien, deux nouveaux exons peuvent exister, H (seul GH était connu) et BC (seul C était connu).

Annexe 7 : Détails de la composition des tables de la base de données

La base de données, dont le schéma relationnel est présenté à la figure 4.18, est constitué de 19 tables.

La table ***gene*** contient les informations disponibles sur les 253 gènes dont la structure est conservée avec les attributs suivants :

— *idGene* : INT / Identifiant du gène généré pour la base de données ;
— *sourceGId* : VARCHAR / Identifiant Ensembl du gène ;
— *chr* : INT / Chromosome associé ;
— *posStart* : INT / Position génomique de début du gène ;
— *posEnd* : INT / Position génomique de fin du gène ;
— *strand* : ('+'|'-') / Sens de lecture du gène ;
— *geneName* : INT / Nom du gène ;
— *idSpecies* : INT / Identifiant de l'espèce associé au gène.

La table ***hasGT*** présente les relations entre les gènes et les transcrits, pour connaître quels sont les transcrits associés à un gène :

— *idGene* : INT / Identifiant du gène ;
— *idTranscript* : INT / Identifiant du transcrit ;

La table ***transcript*** contient toutes les informations disponibles sur les transcrits :

— *idTranscript* : INT / Identifiant du transcrit généré pour la base de données ;
— *sourceTId* : VARCHAR / Identifiant Ensembl du transcrit ;
— *posStart* : INT / Position génomique de début du transcrit ;
— *posEnd* : INT / Position génomique de fin du transcrit ;
— *nbExon* : INT / Nombre d'exons codants constituant le transcrit ;
— *transcriptName* : VARCHAR / Nom du transcrit ;
— *transcriptBiotype* : VACRHAR / Biotype du transcrit ;
— *state* : ('known'|'predicted') / Statut du transcrit : connu ou prédit.

La table ***hasTE*** contient toutes les relation entre les transcrits et les exons qui les composent ainsi que leur position dans le transcrit :

— *idTranscript* : INT / Identifiant du transcrit ;
— *idExon* : INT / Identifiant de l'exon ;
— *exonInTNb* : INT / Numéro d'ordre de l'exon dans le transcrit.

La table ***exon*** contient toutes les informations génomiques sur les exons :
— *idExon* : INT / Identifiant de l'exon généré pour la base de données ;
— *posStart* : INT / Position génomique de début de l'exon ;
— *posEnd* : INT / Position génomique de fin de l'exon.

La table ***hasTI*** contient toutes les relation entre les transcrits et les introns qui séparent les exons qui composent les transcrits ainsi que leur position dans le transcrit :
— *idTranscript* : INT / Identifiant du transcrit ;
— *idIntron* : INT / Identifiant de l'intron ;
— *intronInTNb* : INT / Numéro d'ordre de l'intron dans le transcrit.

La table ***intron*** contient toutes les informations génomiques sur les introns :
— *idIntron* : INT / Identifant de l'intron généré pour la base de données ;
— *posStart* : INT / Position génomique de début de l'intron ;
— *posEnd* : INT / Position génomique de fin de l'intron.

La table ***orthoGene*** permet de connaître les groupes d'orthologie entre gènes :
— *groupGId* : INT / Identifiant du groupe d'orthologie ;
— *sourceGId* : VARCHAR / Identifiant du gène.

La table ***groupOrthoG*** permet de relier les groupes d'orthologie aux identifiants générés par CG-alcode :
— *groupGId* : INT / Identifiant du groupe d'orthologie ;
— *opggnb* : INT / Numéro de la paire orthologue du groupe d'orthologie (correspond au numéro de la paire de gènes comparée dans l'analyse par paires de CG-alcode).

La table ***pairTOrtho*** contient les relations d'orthologie au niveau du transcrit (source et cible) :
— *transcriptSource* : VARCHAR / Transcrit source ;
— *transcriptTarget* : VARCHAR / Transcrit cible ;
— *groupeGId* : INT / Identifiant correspondant au groupe d'orthologie.

La table ***sourceInfo*** contient les informations secondaires sur les gènes :

- *idGene* : INT / Identifiant du gène ;
- *baseNameV* : VARCHAR / Nom et version de la base de données d'origine du gène ;
- *caracteristic* : VARCHAR / Caractéristique du gène ;
- *score* : VARCHAR / Score du gène ;
- *frame* : VARCHAR / Cadre de lecture du gène ;
- *geneBiotype* : VARCHAR / Biotype du gène.

La table ***species*** contient les espèces étudiées :

- *idSpecies* : INT / Identifiant de l'espèce ;
- *nameSpecies* : VARCHAR / Nom de l'espèce.

La table ***exon_utr*** contient les informations sur les exons UTR :

- *sourceGId* : VARCHAR / Identifiant du gène ;
- *transcript* : VARCHAR / Identifiant du transcrit ;
- *posTStart* : VARCHAR / Position génomique de début du transcrit ;
- *posTEnd* : VARCHAR / Position génomique de fin du transcrit ;
- *strand* : ('+'|'-') / Sens de lecture sur le gène ;
- *exon_utrStart* : INT / Position de début de l'UTR ;
- *exon_utrEnd* : INT / Position de fin de l'UTR.

La table ***resultValidationT*** contient les informations sur les transcrits prédits avec leur statut de validation :

- *idTranscript* : INT / Identifiant du transcrit ;
- *idValidation* : INT / Identifiant de la méthodes de validation ;
- *idSource* : INT / Identifiant de la source de données utilisée ;
- *nbSpecificJunction* : INT / Nombre de jonctions d'exons spécifiques (par défaut 0 si il n'y en a pas) ;
- *nbSpecificJunctionFound* : INT / Nombre de jonctions d'exons spécifiques identifiés avec des lectures.

La table ***sourceValidationT*** contient les information sur les sources de données utilisées pour les validations :

- *idSource* : INT / Identifiant de la source de données utilisée ;

— *nameSource* : VARCHAR / Nom de la source de données utilisée ;
— *typeSource* : VARCHAR / Type de données ('database'|'read') ;
— *remark* : VARCHAR / Informations supplémentaires.

La table ***typeValidationT*** contient les informations sur les méthodes de validations employées :
— *idValidation* : INT / Identifiant de la méthode de validation ;
— *nameMethod* : VARCHAR / Nom de la méthode utilisée : '*annotation*', '*all_junction_known*', '*specific_junction_in_reads*', '*no_evidence*' ;
— *description* : VARCHAR / Descriptions supplémentaires.

La table ***geneSet*** contient les relations entre les gènes et l'identifiant de l'ensemble auquel ils appartiennent :
— *idGene* : INT / Identifiant du gène ;
— *idSet* : INT / Identifiant de l'ensemble ;

La table ***typeGeneSet*** contient la description des ensembles de gènes (ensemble 135 ou 118, avec ou sans transcription alternative avérée) :
— *idSet* : INT / Identifiant de l'ensemble ;
— *nameSet* : VARCHAR / Nom de l'ensemble de gènes ("*Set135*", "*Set118*") ;
— *description* : VARHAR / Description de l'ensemble de gènes ("*transcriptome_conserved*", "*alternative_transcription_true*", "*alternative_transcription_false*").

La table ***refCDS*** contient les informations sur les groupes de CDS orthologues :
— *idTranscript* : INT / Identifiant du transcrit ;
— *idCDS* : INT / Identifiant du groupe de CDS orthologue ;
— *transcriptUsed* : BOOL / Si le transcrit est utilisé dans cette étude ;
— *utrFound* : BOOL / si les UTR sont trouvés dans le transcrit (5' et 3' UTR).

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Comparer des structures de gènes pour la prédiction de transcrits alternatifs codants chez l'humain, la souris et le chien

Mot clés : Bioinformatique, génomique comparative, épissage alternatif, transcription alternative, orthologie, prédiction de transcrits

Résumé : Les organismes vivants sont capables d'exprimer plusieurs transcrits (ou ARN) alternatifs à partir d'un même gène. Ces transcrits sont responsables des mécanismes de régulation de l'organisme, certains sont traduits en protéine. Détecter l'ensemble des transcrits pouvant être exprimés par un gène est aujourd'hui un problème ouvert auquel de nombreuses méthodes informatiques telles que le séquençage des données de l'ARN, les méthodes d'alignements de séquences épissées ou encore les méthodes de génomiques comparative tentent de répondre. Cette thèse propose une méthode de génomique comparative permettant de comparer la séquence de gènes partagés par plusieurs espèces. Il en résulte une méthode de prédiction de transcrits à une échelle multi-espèces, en s'appuyant sur une structure de graphes. Cette méthode a été appliquée à trois espèces (humain, souris, chien). Elle a permis de prédire un nombre important de transcrits et d'identifier un ensemble de gènes conservés entre les trois espèces et partageant les mêmes structures exoniques et les mêmes CDS.

Title: To compare gene structures for prediction of alternative coding transcripts in human, mouse and dog

Keywords: Bioinformatics, comparative genomics, alternative splicing, alternative transcription, orthology, transcript prediction

Abstract: Living organisms are capable of expressing several alternative transcripts (or RNAs) from a single gene. These transcripts are responsible for the regulatory mechanisms of the organism, some of them are translated into protein. Today, detecting the set of transcripts that can be expressed by a gene is an open problem to which many computational methods such as RNA sequencing, spliced sequence alignment methods or comparative genomics methods try to address. This thesis proposes a comparative genomics method to compare the sequence of genes shared by several species. The result is a method for predicting transcripts on a multi-species scale, based on a graph structure. This method was applied to three species (human, mouse, and dog). It allowed to predict a relevant number of transcripts as well as to identify a set of genes that are conserved between the three species and that share both the same exonic structures and the same CDS.