Intégration de l'évolution pour contribuer à l'étude de la relation séquence, structure, fonction des protéines

Laurent Bianchetti

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)

THÈSE

présentée par :

Laurent BIANCHETTI

soutenue le : **2 octobre 2019**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Bioinformatique et Biologie des Systèmes

Intégration de l'évolution pour contribuer à l'étude de la relation séquence, structure, fonction des protéines

THÈSE dirigée par :

Mme DEJAEGERE Annick	Professeur, Université de Strasbourg
Mr POCH Olivier	Docteur, Centre National de la recherche Scientifique (CNRS)

RAPPORTEURS :

Mr CAILLAUD Emmanuel	Professeur, Université de Strasbourg
Mr POTIER Serge	Professeur, Université de Strasbourg

AUTRES MEMBRES DU JURY :

Mr Frédéric DEVAUX	Professeur, Université de la Sorbonne, Paris
Mr Dominique SCHLAEFLI	Dir. Adjoint aux Ressources, Formation Continue
Mr MARULLO Philippe	Chargé de Recherche, BioLaffort, Bordeaux
Mme FRIEDRICH Anne	Maître de Conférences, Université de Strasbourg
Mme CABURET Sandrine	Maître de Conférences, Université Diderot, Paris

Cellule VAE
Université de Strasbourg

Transformer votre expérience en diplôme

Année Universitaire 2018 – 2019

DOSSIER DE DEMANDE DE VALIDATION DES ACQUIS DE L'EXPÉRIENCE

Vu la loi n° 2002-72 du 17 janvier 2002 de modernisation sociale et le code de l'éducation et notamment ses articles R 613-32 et suivants

BIANCHETTI Laurent

Doctorat de Bioinformatique

Déclaration sur l'honneur

Je déclare sur l'honneur que toutes les informations fournies dans ce dossier sont exactes.
J'ai été rendu attentif au fait qu'agrémenter mon travail de citations en omettant d'en citer les sources représente un acte de plagiat. Le plagiat est une atteinte au droit d'auteur et à la propriété intellectuelle. La présente candidature à la validation des acquis de l'expérience en vue de l'obtention diplôme postulé, précité, constitue l'unique demande pour ce diplôme pour la même année civile. Je m'engage également à ne pas présenter plus de trois candidatures à la validation des acquis de l'expérience pour des diplômes, certificats ou titres différents durant la présente année civile.
Fait à Illkirch

Le 10/05/2019

Signature

La loi punit quiconque se rend coupable de fausses déclarations.
Constitue un faux toute altération frauduleuse de la vérité, de nature à causer un préjudice et accomplie par quelque moyen que ce soit, dans un écrit ou tout autre support d'expression de la pensée qui a pour objet ou qui peut avoir pour effet d'établir la preuve d'un droit ou d'un fait ayant des conséquences juridiques. Le faux et l'usage de faux sont punis de trois ans d'emprisonnement et de 45000 euros d'amende (code pénal, art. 441-1) Le fait de se faire délivrer indûment par une administration publique ou par un organisme chargé d'une mission de service public, par quelque moyen frauduleux que ce soit, un document destiné à constater un droit, une identité ou une qualité ou à accorder une autorisation, est puni de deux ans d'emprisonnement et de 30000 euros d'amende » (code pénal art. 441-6)

Sommaire

A. Fiche Analytique

IDENTITE

Monsieur BIANCHETTI Laurent

Né(e) le 24/02/1972

A Saint Martin d'Hères, Isère, FRANCE

Nationalité française

Adresse personnelle 1 rue de la poste

67400 ILLKIRCH FRANCE

Courriel laurent.bianchetti@igbmc.fr

Tél. portable 00 33 (0)6 11 91 35 60

SITUATION ACTUELLE

Dernière fonction exercée en lien avec le diplôme souhaité Ingénieur de Recherche

Nom de la structure Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)

Vous êtes actuellement

☒**en situation d'emploi** fonctionnaire

☐**en recherche d'emploi** Choisissez un élément.

☐**en situation d'inactivité** Choisissez un élément.

Etes-vous reconnu travailleur handicapé ? non

SCOLARITE / FORMATION

Dernière classe suivie master 2, DEA, DESS, doctorat ou équivalent

Niveau du diplôme le plus élevé obtenu titre d'ingénieur diplômé

Intitulé du diplôme le plus élevé obtenu Ingénieur en biotechnologie (ESBS)

B. Tableaux descriptifs du parcours

B1. Expérience salariée, non salariée ou bénévole en rapport avec le diplôme visé (en commençant par la plus récente)

1 Emploi ou fonction bénévole occupée	*2 Nom et lieu de l'entreprise (ou structure) dans laquelle les activités ont été exercées*	*3 Secteur d'activité de l'entreprise ou de la structure*	*4 Statut: 1 salarié 2 travailleur indépendant, artisan ou prof. libérale 3 bénévole*	*5 Temps de travail (indiquer le nombre d'heures effectuées / mois)*	*7 Période d'emploi début - fin (ou « à ce jour »)*	*6 Total des heures effectuées dans cet emploi ou cette fonction**	*8 Principales activités exercées en rapport avec le diplôme ou le titre professionnel visé*	*9 N° du justificatif d'activité*
Ingénieur de Recherche (IR)	Institut National de la Santé et de la Recherche Médicale (INSERM) – Unité 1258 (IGBMC)	Recherche scientifique en biologie structurale, médecine translationelle et neurogénétique, cellules souches et biologie du développement, génomique fonctionelle et cancer	Salarié	Temps complet 151 h/mois	Du 01/02/2004 à ce jour	24908 heures	. Développement de projet de recherche . Bibliographie . Question biologique . Conception et planification de projet . Production et interprétation de résultats . Ecriture de manuscrit . Critique d'article . Encadrement d'étudiants . Enseignement . Veille technologique	1

TOTAL des heures effectuées (total de la colonne 6) : 24908 _____ heures

* En France, la durée légale du travail est fixée à 35 heures par semaine, 151 heures par mois ou 1 607 heures par an. Le dispositif VAE est accessible à toute personne justifiant d'au moins 1an d'expérience professionnelle équivalent temps plein acquise dans des activités salariées, non salariées ou bénévoles en rapport direct avec le diplôme. Il appartient au candidat d'apporter la preuve de la durée de son expérience

B.2. Diplômes, titres, certificats et formations suivies (en commençant par le plus récent

1 ***Diplôme, titre, certificat ou formation (Intitulé exact)***	***2*** ***Pour les diplômes, titres ou certificats :*** ***1 Obtenu*** ***2 Non obtenu***	***3*** ***Nom et lieu de l'Etablissement ou de la structure***	***4*** ***Année d'obtention***
Diplôme d'ingénieur en biotechnologie	Obtenu	Ecole Supérieure de Biotechnologie de Strasbourg (ESBS)	1996
Diplôme universitaire général (DEUG) de Chimie et Biochimie Mention trés bien	Obtenu	Université Joseph Fourier (Grenoble)	1992
Diplôme du Baccalauréat Sciences de la vie et de la terre Mention bien	Obtenu	Lycée Pierre Termier (Grenoble)	1990

C. Curriculum Vitae

2018 Ingénieur de recherche INSERM (IR2). Evaluation par modélisation moléculaire de la pertinence d'utiliser des mutants du domaine de tétramerisation p53 pour la synthèse d'anticorps bispécifiques. Collaboration avec le Dr. Mariel Donzeau (Biotechnologie et signalisation cellulaire, CNRS UMR7242, ESBS,). Manuscrit en préparation, 3ième auteur**, Laboratoire de biophysique chimique de la signalisation transcriptionnelle, IGBMC, Illkirch, FRANCE**.

2018 Ingénieur de recherche INSERM (IR2). Développement d'un projet de recherche intégratif en modélisation moléculaire et évolution de séquence « Homodimérisation du domaine de liaison au ligand (LBD) du récepteur nucléaire aux minéralocorticoïdes (MR)». Manuscrit en préparation, 1er auteur. **Laboratoire de biophysique chimique de la signalisation transcriptionnelle, IGBMC, Illkirch, FRANCE**.

2014-2017 Ingénieur de recherche INSERM (IR2). Développement d'un projet de recherche intégratif en modélisation moléculaire et évolution de séquence "Dimérisation alternative du domaine de liaison au ligand (LBD) du récepteur nucléaire aux glucocorticoïdes (GRα) par une interface impliquant l'hélice 9", publication en 1er auteur. **Laboratoire de biophysique chimique de la signalisation transcriptionnelle, IGBMC, Illkirch, FRANCE.**

2013 Ingénieur de recherche INSERM (IR2). Développement d'un projet de recherche intégratif en phylogénèse moléculaire, RNA-seq, et modélisation moléculaire "Co-evolution des gènes *Tex19* et *Sectm1* chez les mammifères placentaires", publication en 1er auteur correspondant. **Plate-forme bioinformatique de Strasbourg (BIPS), Laboratoire de calcul et modélisation moléculaire, IGBMC, Illkirch, FRANCE.**

2011-2012 Ingénieur de recherche INSERM (IR2). Développement d'un projet de recherche en transcriptomique cancer "Augmentation du nombre de substitutions de bases dans une population de transcrits exprimés en cancer", publication en 1er auteur correspondant. **Plate-forme bioinformatique de Strasbourg (BIPS), IGBMC, Illkirch, FRANCE.**

2010 Ingénieur de recherche INSERM (IR2). Collaboration à un projet de recherche en pharmacogénomique "Identification de molécules correctrices de la mucoviscidose par analyse de données transcriptomiques (micro-arrays)", **Laboratoire de Biochimie du Dr. D. Y. Thomas, Université McGill, Montréal, CANADA.**

2007-2010 Ingénieur de recherche INSERM (IR2). Pilote du processus expertise bioinformatique certifié ISO 9001:2000. Développement d'un outil logiciel à interface web pour le traitement haut-débit de données trancriptomiques SAGE et LongSAGE (« *Serial analysis of gene expression* »), publication en 1er auteur correspondant. **Plate-forme bioinformatique de Strasbourg (BIPS), IGBMC, Illkirch, FRANCE.**

2004-2006 Ingénieur de recherche INSERM (IR2). Développement d'un outil logiciel à interface web pour la génomique comparative "Validation de la qualité des séquences primaires de protéines prédites sur les génomes par l'analyse automatique d'alignements multiples", publication en 1er auteur correspondant. **Service informatique, IGBMC, Illkirch, FRANCE**

2000-2003 Ingénieur d'étude (IE). Développement d'un projet de recherche en phylogénèse moléculaire "*Danio rerio* est l'organisme le plus basal présentant l'enzyme PHEX dont la mutation cause le rachitisme hypophosphatémique chez l'humain", publication en 1er auteur correspondant. **Service informatique, IGBMC, Illkirch, FRANCE.**

1998 Service militaire. Escadron anti-char. Brigade Franco-Allemande. **Immendingen, Baden-Württemberg, ALLEMAGNE**

1997 Ingénieur stagiaire. Evaluation de la protéine fluorescente verte sauvage et de mutants hyperchromes comme marqueurs d'expression semi-quantitatifs chez la levure *Saccharomyces cerevisiae*. Encadrement: Dr. Eric Degryse. Amplification PCR des gènes *gag*, *pol*, *env* du virus HIV. Encadrement: Dr. Marie-Paule Kieny. **Transgène SA, Strasbourg, FRANCE**

Formations

2016 Ecole d'été internationale DYNAPEUTICS de modélisation moléculaire
25-30 septembre, Université du Pays Basque, Donastia-Saint Sébastien, ESPAGNE

2014 Modélisation moléculaire par homologie, formation continue INSERM
1-4 avril, Université Diderot, Paris, FRANCE

Diplômes

1996 Diplôme d'ingénieur en biotechnologie
Ecole Supérieure de Biotechnologie de Strasbourg (ESBS)
Université Louis Pasteur, Strasbourg, FRANCE

1993 Diplôme d'études universitaires générales de chimie et biochimie (DEUG)
Mention très bien
Université Joseph Fourier, Grenoble, FRANCE

1990 Diplôme du baccalauréat d'enseignement général mathématiques et sciences de la vie
Mention bien
Lycée Pierre Termier, Grenoble, FRANCE

D. Liste de publications et communications

Bianchetti L, Sinar D, Depenveiller C and Dejaegere A*. *Dimerization of the mineralocorticoid receptor ligand binding domain by helices 9-10 and the F-domain.* **2019**. (Manuscrit en preparation).

Vigneron M, Dietsch F, **Bianchetti L**, Dejaegere A, Nominé Y, Zuber G, Chatton B and Donzeau M*. Engineering of modular bi-specific macromolecules by self associating peptides. **2019**. (Manuscrit en préparation).

Bianchetti L, Wassmer B, Defosset A, Smertina A, Tiberti M, Stote RH and Dejaegere A*. *Alternative dimerization interfaces in the glucocorticoid receptor-α ligand binding domain*. Biochim *et* Biophys Acta. **2018.** 1862(8):1810-1825.

Bianchetti L*, Tarabay Y, Lecompte O, Stote R, Poch O, Dejaegere A and Viville S. *Phylogenetic coevolution of the cancer/testis Tex19 antigen and the Sectm1 protein which attracts monocytes to tumors*. 18ème journée scientifique de la Ligue contre le Cancer. 30 nov. **2017**, Ecole Supérieure de Biotechnologie de Strasbourg, Illkirch, France.

Bianchetti L*, Tarabay Y, Lecompte O, Stote R, Poch O, Dejaegere A and Viville S. *Tex19 and Sectm1 concordant molecular phylogenies support coevolution of both eutherian specific genes*. BMC Evolutionary Biology. **2015**. 12(15):222

Bianchetti L. *Phylogénèse et comparaison des organismes à l'ère du séquençage des génomes*. 15 avril. **2015**. Parc zoologique et botanique de Mulhouse, Lycée Louis Armand, France.

Bianchetti L. *Use of SRS (Sequence retrieval system) at the bioinformatic platform of Strasbourg*. 13th nov. **2013**. SRS special interest group meeting, Londres, Royaume Uni.

Bianchetti L*, Féderkeil R, Kieffer D and Poch O. *Increased single base substitutions in a population of transcripts expressed in cancer cells. BMC cancer* (Special issue: Bioinformatics, Network biomarkers and precision medecine). **2012**. 8(12):509.

Bianchetti L. *RNA-seq differential expression analysis of zebrafish mutants*. 18-19 oct. **2012**. Next generation sequencing user meeting, Santa Clara, Californie, Etats-Unis d'Amérique.

Zhang D, Ciciriello F, Anjos SM, Liao J, Balghi H, **Bianchetti L**, Kemmer D, Carlile GW, Robert R, Hanrahan JH et Thomas DY. *Ouabain mimics low temperature rescue of F508del-CFTR in cystic fibrosis epithelial cells.* **2011.** Poster présenté aux journées de bioinformatique de Strasbourg, France.

Bianchetti L et Thomas DY. *Cascade: a RNA-seq program increasing the identification of reads that are uniquely located on the genome and reads associated with genes.* **2011.** Poster. Human Genetics Conference. Banff, Alberta, Canada.

Alpy F, Legueux F, **Bianchetti L** and Tomasetto C. *START domain-containing proteins: a review of their role in lipid transport and exchange*. Medecine & Sciences. **2009**. 25:181-191.

Ribes V, Stutzmann F, **Bianchetti L**, Guillemot F, Dolle P and Le-Roux I. *Combinatorial signalling controls Neurogenin2 expression at the onset of spinal neurogenesis*. Developmental Biology. **2008**. 15(321):470-481.

Lagier-Tourenne C, Tazir M, Lopez LC, Quinzii CM, Drouot N, Assoum M, Busso C, Makri S, Ali-Pacha L, Benhassine T, Anheim M, Lynch D, Thibault C, Plewniak F, **Bianchetti L**, DiMauro S, Tranchant C, Poch O, Mandel JL, Barros MH, Hirano M and Koenig M. *ADCK3, an ancestral mitochondrial kinase involved in coenzyme Q biosynthesis, is mutated in a new form of recessive ataxia*. American Journal of Human Genetics. **2008**. 82(3):661-672.

Kuntz S, Kieffer E, **Bianchetti L**, Lamoureux N, Fuhrmann G and Viville S. *Tex19, a mammalian specific protein, with a restricted expression in pluripotent stem cells and germ line*. Stem Cells. **2008**. 26(3):734-744.

Bianchetti L*, Wu Y, Guérin E, Plewniak F and Poch O. *SAGETTARIUS: a program to reduce the number of tags mapped to multiple transcripts and to plan SAGE sequencing stages*. Nucleic Acids Research. **2007**. 35:18.

Lardenois A, Chalmel F, **Bianchetti L**, Sahel JA, Leveillard T, Poch O. *PromAn: an integrated knowledge-based web server dedicated to promoter analysis.* Nucleic Acids Research. **2006**. 1:(34,Web Server issue):W578-583.

Quillet R, Stoetzel C, **Bianchetti L**, Gachot-Neveu H, Barriel V, Rumpler Y, Dollfus H and Perrin-Schmitt F. *TWIST genes in Primates and head morphology.* Durham Anthropology Journal. **2005**. 12:2-3.

Bianchetti L *, Thompson JD, Lecompte O, Plewniak F and Poch O. *vALId: validation of protein sequence quality based on multiple alignment data.* Journal of Bioinformatics Computional Biology. **2005**. 3(4):929-947.

Plewniak F, **Bianchetti L**, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J, Prigent V, Ripp R, Thierry JC, Thompson J D, Wicker N and Poch O. *PipeAlign: A new toolkit for protein family analysis*. Nucleic Acids Research. **2003**. 31(13):3829-3832.

Bianchetti L*, Oudet C and Poch O. *M13 Endopeptidases: New Conserved Motifs Correlated with Structure, and simultaneous Phylogenetic Occurrence of Phex and the Bony fish*. Proteins: Structure, Function and Genetics. **2002**. 47:481-488.

** auteur correspondant*

E. Lettres de recommandation

Olivier Poch

Directeur de Recherche CNRS

Co-Directeur du CSTB

Tél. +33 (0)3 68 85 32 95

olivier.poch@unistra.fr

Strasbourg, le 24/09/2017

En ma qualité de Directeur de l'équipe *Complex Systems and Translational Bioinformatics* (http://icube-cstb.unistra.fr), responsable de la plateforme bioinformatique de Strasbourg (BIPS) à l'IGBMC de 2000 à 2012, c'est un réel plaisir pour moi d'écrire une lettre de soutien pour Monsieur Laurent BIANCHETTI et de faire connaître ma grande satisfaction et l'intérêt des travaux réalisés au sein de la plateforme BIPS ou dans le cadre de collaborations postérieures.

Au sein de BIPS, Laurent s'est toujours investi pleinement tant dans la compréhension informatique des problèmes que dans l'appropriation des aspects biologiques en faisant preuve d'une curiosité scientifique au plus haut niveau couplée à un sens aigu du service, de l'accueil et de l'écoute des besoins des utilisateurs biologistes. Sur le plan informatique, ces qualités lui ont permis non seulement, d'introduire ou maintenir divers outils de prédiction, analyse et exploitation des séquences biologiques mais aussi, de participer à des développements novateurs, notamment dans le champ de l'exploitation des données biologiques à haut débit. Dans ce dernier cadre, en participant à la réalisation de 5 programmes originaux ayant donné lieu à publication commune avec Laurent, j'ai également pu apprécier ses compétences en statistiques et sa grande rigueur scientifique, éléments essentiels dans le domaine récent de la biologie à haut débit. Sur le plan bioanalytique, sa capacité de travail et son opiniâtreté ont permis de résoudre de nombreux problèmes, notamment dans le cadre d'analyses évolutives de familles de protéines complexes. Ces efforts ont été récompensés par plusieurs publications scientifiques dans les meilleurs journaux de bioinformatique et, pour certaines familles, les solutions pertinentes et les pistes relevées par Laurent sont toujours explorées par des collaborateurs. Enfin, il est à noter que, malgré les

contraintes liées au service au sein d'une plateforme, Laurent a toujours maintenu un intérêt fervent pour la recherche, intérêt qui l'a poussé, en 2010, à demander un an de disponibilité pour découvrir de nouvelles problématiques et acquérir de nouvelles compétences au sein d'une équipe de l'Université McGill de Montréal.

A la lumière de tous ces éléments, je soutiens donc sans aucune réserve le dossier de Laurent Bianchetti pour l'obtention d'un doctorat par « validation des acquis de l'expérience », étant convaincu qu'une telle reconnaissance justifiée de ses qualités de chercheur, lui permettra d'aborder à l'avenir des aventures scientifiques de première importance.

Je reste à votre disposition pour toute question complémentaire,

Cordialement.

Dr Olivier POCH

Prof. Annick Dejaegere
Directrice de recherche
Laboratoire de calcul biologique et modélisation moléculaire
Département de biologie structurale intégrative
Institut de génétique et de biologie moléculaire et cellulaire (IGBMC)
1 rue Laurent Fries
67404 ILLKIRCH
Bas-Rhin
France

Tél : 03 68 85 47 21
annick.dejaegere@igbmc.fr

Illkirch, le 17 octobre 2017

C'est avec grand plaisir que je vous écris cette lettre de recommandation pour Mr. Laurent Bianchetti. Je connais Mr. Bianchetti de longue date, car je l'ai eu comme étudiant lors de son cursus d'Ingénieur à l'Ecole Supérieure de Biotechnologie de Strasbourg, un cursus sélectif qu'il a très bien réussi.

Cependant, j'ai vraiment commencé à interagir avec Laurent lorsqu'il a rejoint notre équipe de recherche en mars 2014. Laurent a en effet demandé à être transféré dans notre équipe suite au départ de l'équipe d'O. Poch, qui a quitté l'IGBMC en 2012.

Cette demande témoigne de la volonté de Mr. Laurent Bianchetti de s'ouvrir à des disciplines scientifiques nouvelles, en effet le coeur de l'activité de notre équipe est la modélisation moléculaire et la bioinformatique structurale, qui tout en étant complémentaires à son expertise précédente, lui ont demandé un effort d'apprentissage et d'adaptation conséquent.

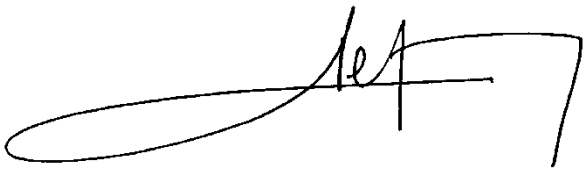
Au sein de notre équipe, Mr. Laurent Bianchetti a pris en charge un projet sur la régulation et l'assemblage du récepteur aux glucocorticoïdes (GR). GR appartient à la famille des récepteurs nucléaires, des facteurs de transcription d'une importance primordiale en santé humaine. La plupart de ces facteurs de transcription sont actifs sous forme de dimères. Dans le cas de GR, le mécanisme de dimérisation est encore très mal compris, malgré le rôle essentiel de cette dimérisation dans la régulation du récepteur. Mr. Bianchetti a pris ce projet en main de manière tout à fait autonome. Il s'est formé à l'ensemble des méthodes de modélisation moléculaire nécessaires à sa réalisation et a apporté au projet sa compétence en analyse de séquences. J'ai pu apprécier dès son arrivée dans

l'équipe les qualités scientifiques de Mr. Bianchetti et sa grande implication dans le travail. Il a apporté des informations tout à fait originales dans ce premier projet dont les résultats ont été publiés [Bianchetti L. *et al*, 2018]. Mr. Bianchetti a également publié une étude phylogénétique originale (Tex19/Sectm1) qu'il avait commencée avant de rejoindre l'équipe, en y ajoutant des informations nouvelles issues de modèles moléculaires construits dans notre laboratoire.

Mr. Bianchetti s'implique également très volontiers dans des tâches d'intérêt collectif, il a pris en charge par exemple l'organisation des séminaires communs de l'équipe.

Comparé aux autres étudiants en thèse que j'ai supervisé (13 au total) Mr. Bianchetti a sans aucun doute démontré par son travail passé les qualités de persévérance, de rigueur et d'originalité scientifique associées à un travail de thèse, et je soutiens donc sans aucune réserve son dossier pour l'obtention d'un doctorat par « validation des acquis de l'expérience ».

Je reste à votre disposition pour toute question complémentaire et je vous prie d'agréer, Madame, Monsieur, l'expression de ma parfaite considération,



Dr. Annick Dejaegere
Professeur, Université de Strasbourg

F. Analyse des acquis de l'expérience

G. Sujet de recherche

« Intégration de l'évolution pour contribuer à l'étude de la relation séquence, structure, fonction des protéines »

G.1. Introduction

L'enregistrement fossile le plus ancien connu d'une activité d'organisme vivant est daté de 3,7 milliard d'années [Nutman A.P. *et al*, 2016]. Il s'agit de stromatolithes (Figure 1), c'est-à-dire de dépôts sédimentaires produits par des micro-organismes qui précipitent des carbonates. Il en existe encore d'actives aujourd'hui à « *Shark Bay* » (Australie) [Burns B. P. *et al*, 2004], « *Highborne Cay* » (Bahamas) [Foster J.S. *et al*, 2009] et « *Laguna Negra* » (Argentine) [Mlewski E.C. *et al*, 2018]. Ces communautés de micro-organismes contiennent des procaryotes tels que des cyanobactéries (bactéries photosynthétiques) et des Archaea [Allen M. A. *et al*, 2009]. Les stromatolithes contemporaines et biologiquement actives fournissent des informations précieuses pour comprendre la physiologie des formes de vie les plus anciennes et donc l'origine de la vie sur Terre.

Figure 1: **Stromatholithe fossilisée**. Pilbara, Australie. 3,43 milliards d'années (Musée d'Histoire Naturelle de Karlsruhe, Baden-Württemberg, Allemagne). Photo Laurent Bianchetti.

La paléontologie, discipline qui étudie les fossiles, montre que la diversité des organismes a connu des phases de croissance comme celle de l'explosion cambrienne (-541 à -530 millions d'années) et des phases de crises ou extinctions de masse. La plus récente de ces extinctions est celle de la fin du Crétacé (-65 Millions d'années) qui a sonné le glas des dinosaures. Cependant, la plus désastreuse s'est produite à la fin du Permien (-250 millions d'années). Avec elle, 95% des espèces marines et 70% des espèces terrestres ont disparu [Benton M.J. et Twitchett R.J., 2003]. Cette extinction aurait été provoquée par un rejet massif de CO_2 dans l'atmosphère engendrant un effet de serre, un réchauffement climatique (+6°C) et une acidification des océans avec des conséquences dévastatrices pour la biodiversité [Benton M.J. et Twitchett R.J., 2003 ; Renne P.R. *et al*, 1995]. Si les extinctions de masse balaient bon nombre d'espèces, il est néanmoins considéré qu'elles permettraient aux espèces survivantes de jouir de l'espace et des ressources rendus disponibles par les organismes éteints (territoire, eau, nourriture, baisse de la prédation) et donc de produire un rebond évolutif [Raup D.M., 1986].

L'étude de l'évolution des espèces est une discipline scientifique dont les bases ont été posées par Carl von Linné (1707-1778), Charles Darwin (1809-1882) et Carl Woese (1928-2012), qui sont respectivement les pères de la taxonomie, des notions de sélection naturelle et d'ancêtre commun, et de l'arbre de la vie à 3 domaines (Bacterie, Archaea et Eucaryote) construit sur la séquence des ARN 16S ribosomiques. Ces pionniers ont préparé la voie à la révolution génomique du début du XXI^{ème} siècle. En effet, depuis l'invention des séquenceurs à grands segments d'ADN, *e.g.* Roche 454 et PacBio, il est possible d'obtenir la séquence complète du génome de n'importe quel organisme. De plus, des outils bioinformatiques réalisent l'assemblage des segments en chromosomes et leur annotation automatique, c'est-à-dire qu'ils localisent les gènes [Kent W.J. *et al*, 2002]. Ces technologies ont permis l'émergence de la génomique comparative qui revisite les connaissances sur l'évolution des espèces héritées de Linné, Darwin et Woese en s'appuyant sur la comparaison des séquences biologiques (protéines, ARN, ADN). Les séquences génomiques des organismes modèles (*Escherichia coli, Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditis elegans, Mus musculus, Homo sapiens*) ont été rendus accessibles dans des banques de données [Hedges S.B., 2002] et analysables par des outils bioinformatiques comme BLAST [Atschul S.F. *et al*, 1990] ou des navigateurs génomiques (UCSC). De plus, un déluge de projets de séquençages dont l'objectif est d'obtenir les génomes d'insectes, poissons, reptiles, oiseaux et mammifères est en progrés

[Genome 10K community of scientists, 2009; Bernardi G. *et al*, 2012 ; I5K community of scientists, 2013; OBrien S.J. *et al*, 2014].

Le séquençage des génomes porte la promesse de retombées énormes pour les connaissances fondamentales en biologie et des applications en biotechnologies. Sur les origines de la vie, la comparaison des séquences pourrait permettre la reconstruction phylogénétique du vivant en parcourant l'arbre de Carl Woese de « haut-en-bas », c'est-à-dire des feuilles (séquences des espèces actuelles) vers la racine (séquence de LUCA « *Last Unified Common Ancestor* ») [Pereto J., 2005]. Ainsi, la question du nombre de gènes minimum permettant la vie est posée [Glass J.I. *et al*, 2006]. Si l'annotation du génome *d'Halococcus hamelinensis*, première Archaea isolée de stromatolithes, a permis d'identifier 3.150 gènes dont 2.196 (69 %) codent pour des protéines [Gudhka R.M. *et al*, 2015], ce génome demeure néanmoins complexe. De plus, le séquençage des génomes ouvre la possibilité d'obtenir un éclairage évolutif de la séquence primaire de chaque protéine. Classiquement, le chercheur souhaite savoir comment s'organise les séquences protéiques qu'il étudie en régions conservées et non-conservées [Lecompte O. *et al*, 2001; Bianchetti L. *et al*, 2002]. Généralement, une protéine possède un jeu de quelques résidus invariants qui portent la signature de sa fonction. Dans la séquence, il y a donc un lien direct entre la contrainte de conservation qui s'exerce sur certains résidus et la fonction de la protéine. Chez les enzymes, il s'agit des résidus catalytiques. En surface des protéines, des résidus conservés peuvent servir à l'intéraction avec une protéine partenaire ou un acide-nucléique comme chez les facteurs de transcription. Le chercheur souhaite également savoir quelles sont les espèces dont le génome code pour des homologues de la protéine étudiée. Cette information est précieuse car elle permet de déterminer l'organisme ancestral commun chez qui le gène est apparu. Savoir quand les gènes apparaissent au cours de l'évolution peut aider à comprendre la mise en place des processus biologiques et leur sophistication (développement, métabolisme, homéothermie, placentation, immunité, etc …) ce qui relie une nouvelle fois la séquence à la fonction.

En biologie, la relation séquence, structure, fonction est une clé fondamentale pour comprendre la physiologie des organismes. Si la séquence se définit par l'enchaînement des acides-aminés dans le polypeptide et la structure par la position spatiale des atomes qui constituent les acides-aminés, la définition de la fonction est plus large. Il peut s'agir du rôle moléculaire ou du processus phyiologique dans lequel la protéine intervient. De plus, des informations comme la localisation cellulaire, les partenaires d'intéraction, les tissus d'expression ou même les organismes chez lesquels la protéine est présente peuvent apporter

des informations fonctionnelles. C'est pourquoi, la fonction d'une protéine se décline selon 3 critères majeurs appelés ontologie génique (GO) : fonction moléculaire, localisation cellulaire et processus biologique [Patty L. 2008]. Il est largement admis que la séquence primaire d'une protéine détermine sa structure 3D et que sa structure détermine sa fonction. Il est largement accepté également que la structure est plus conservée que la séquence [Fariselli P. *et al*, 2006]. De plus, l'étude de la relation séquence, structure, fonction a éminemment bénéficié d'une part de la résolution de structures 3D et d'autre part de la modélisation moléculaire. La détermination de structures 3D est possible grâce aux méthodes de résonance magnétique nucléaire (RMN), de cristallographie aux rayons-X ou de cryo-microscopie électronique (cryo-EM) [Berman H.M. *et al*, 2000]. Si la RMN a longtemps été appliquée à des protéines de taille inférieure à 50 kDa, des avancées récentes ont permis d'obtenir la structure de machines protéiques de plusieurs centaines de MDa [Sprangers R. *et al*, 2007]. En outre, la RMN présente l'avantage de pouvoir échantillonner plusieurs états de conformations. La cristallographie aux rayons-X s'applique à des protéines de taille plus grande (plusieurs centaines d'acides-aminés) ou des oligomères et permet d'obtenir des représentations 3D d'une résolution de 1 à quelques Å généralement. Enfin, la cryo-EM est adaptée à des études structurales de grands complexes macromoléculaires, *e.g.* ribosome eucaryote, avec des résolutions supérieures à 3 Å. La modélisation moléculaire s'appuie sur la mécanique newtonienne et sur les structures 3D pour calculer par une fonction mathématique appelée champ de force l'énergie potentielle d'une macromolécule. Cette fonction intègre un terme dit « lié » (les longueurs de liaison et les ouvertures d'angles entre atomes sont assimilées à des ressorts) et un terme « non-lié » (forces électrostatiques et forces de van der Waals qui s'exercent à distance). Sur un temps donné, *e.g.* 10 ns à 1 µs, la dynamique moléculaire permet de simuler les changements de conformations de la protéine. Enfin, des méthodes spécialisées utilisant la mécanique moléculaire et l'électrostatique des milieux continus (MM/PBSA) permettent de calculer l'énergie libre de liaison (en kcal/mol) entre 2 macromolécules en interaction [Kollman P.A. *et al*, 2000 ; Lafont V. *et al*, 2007]. Plus cette énergie libre de liaison est basse (négative) plus le complexe est stable et inversement.

Intégrer l'évolution et examiner comment certains organismes ont conquis des environnements naturels sous pression sélective peut aider à mieux comprendre la relation séquence, structure, fonction des protéines. De très beaux exemples qui illustrent cette connaissance fondamentale ont été récemment publiés au sujet d'hémoglobines dont les mutations de certains résidus trouvées chez des oiseaux vivant à plus de 4000 mètres d'altitude

sont responsables d'une affinité augmentée pour l'oxygène [Jendroszek A. *et al*, 2018]. Dans la pratique, l'étude de l'évolution des protéines nécessite la construction d'un alignement multiple de séquences. Un soin tout particulier est requis pour ces alignements multiples dont la qualité finale doit relever de la pièce d'orfèvrerie. En effet, l'alignement multiple contient un message phylogénétique, un message structural, un message fonctionnel [Lecompte O. *et al*, 2001; Plewniak F. *et al*, 2003] voire même un message d'adaptation à un environnement exerçant des contraintes spécifiques. Si la qualité de l'alignement est insuffisante, les messages biologiques sont biaisés et des erreurs peuvent se propager jusqu'aux interprétations. Il est à noter que les outils qui génèrent automatiquement des alignements et déduisent des informations (résidus conservés en interface de contact protéine-protéine, mutations corrélées, …) sans intervention d'expertise humaine pour valider la qualité de l'alignement multiple sont sujets à des erreurs. Les alignements multiples de protéines homologues, c'est-à-dire descendantes d'une protéine codée par un gène ancestral commun, montrent de façon irréfutable que des modifications de la séquence primaire se sont produites au cours du temps et de la diversification des espèces. Il peut s'agir de modifications mineures comme des substitutions, des insertions ou des délétions de quelques acides-aminés ou bien des changements majeurs comme des acquisitions ou des pertes de régions de plusieurs dizaines ou centaines d'acides-aminés [Patthy L., 2008]. Il est à noter qu'un changement mineur de séquence comme l'insertion d'un seul acide-aminé peut provoquer un changement majeur de fonction. Par exemple, cette situation a été décrite pour le récepteur nucléaire alpha aux glucocorticoïdes (GRα) dont la reconnaissance des sites de fixation sur le génome est modifiée en raison de l'insertion d'une arginine dans son domaine de liaison à l'ADN [Thomas-Chollier M. *et al*, 2003]. La perte ou l'acquisition d'une région polypeptidique implique la perte ou l'acquisition d'une région structurale. Il est peu concevable que ces changements soient neutres sur la physiologie de la protéine car ils ont été maintenus par l'évolution. Ils peuvent par exemple ajouter une interaction avec un partenaire, augmenter ou diminuer l'affinité pour un ligand, produire un ancrage dans une membrane, modifier la localisation cellulaire de la protéine, générer un site de modification post-traductionnelle, permettre d'échapper au système immunitaire d'un hôte, etc ... Pour ces raisons, il est souhaitable d'intégrer la dimension évolutive dans l'étude de la relation séquence, structure, fonction.

Je me propose de soutenir ma thèse en valorisant les résultats de mes 2 derniers projets de recherche. Je démontrerai comment l'analyse de l'évolution de séquence permet d'obtenir des informations supplémentaires sur la relation séquence, structure, fonction des protéines.

Dans le 1er projet, j'ai démontré une coévolution entre 2 gènes « *Testis expressed 19* » (Tex19) et « *Secreted and transmembrane 1* » (Sectm1). La coévolution établit un lien fonctionnel très fort entre *Tex19* et *Sectm1* alors que ces gènes interviennent dans des processus biologiques différents, régulation de l'activité des transposons et immunité, respectivement. Ce résultat pourrait présenter un intérêt en immunothérapie du cancer. En effet, *Tex19* a été identifié comme un antigène « testicule/cancer », c'est-à-dire qu'il s'exprime seulement dans le testicule de l'adulte sain et en cellule cancéreuse. Dans le 2ème projet, j'ai questionné la validité biologique de l'assemblage en homodimère du domaine de liaison au ligand (LBD) de GRα obtenu par cristallographie [Bledsoe R.K. *et al*, 2002]. Premièrement, j'ai mis en évidence que ce complexe est vraisemblablement un artefact de contact cristallin. Deuxièmement, j'ai identifié une interface alternative dont le rôle biologique paraît plus plausible [Bianchetti L. *et al*, 2018]. Le GRα est une cible thérapeutique majeure pour le traitement de l'inflammation. Or, de nombreux effets secondaires dûs à la dimérisation du GRα et l'activation de ses gènes cibles accompagnent les traitements de longue durée. Comprendre comment le LBD de GRα dimérise apporte donc une information cruciale.

G.2. Ressources & Méthodes

G.2.1. Ressources

G.2.1.1. Serveurs de calcul

Les serveurs informatiques utilisés correspondent aux équipements du pôle haute-performance de calcul (HPC) de la direction du numérique de l'Université de Strasbourg, soit 5.200 cœurs et 24 Tera octet de mémoire vive répartis sur 380 nœuds, et 500 Tera octet d'espace de stockage. Tous les processeurs ont une architecture x86_64 (famille intel 64 bits). Le système d'exploitation est un linux scientifique 6.5. Les processus de calcul sont soumis au système de file d'attente Slurm pour le partage des ressources entre utilisateurs.

G.2.1.2. Source des données biologiques

G.2.1.2.1. Information taxonomique

Le NCBI maintient une banque dédiée à la classification taxonomique des organismes pour lesquels au moins 1 séquence biologique est présente dans les banques de

données mises à disposition, soit 10% des espèces décrites sur Terre [Federhen S., 2012]. Les 3 domaines du vivant (Archaea, Bactéries, Eucaryotes) et les virus sont représentés dans les banques avec une large majorité d'eucaryotes (Figure 2).

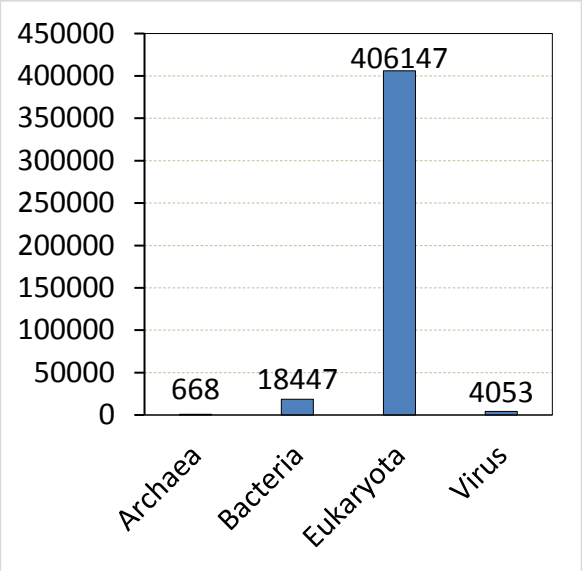


Figure 2: **Nombre d'espèces par domaine enregistrées dans la banque taxonomique du NCBI en 2018**. Pour chacune de ces espèces, au moins une séquence biologique est connue (ADN génomique, ARN, protéine).

Dans les banques de séquences, chaque enregistrement est lié par ses champs SOURCE et ORGANISME à la banque taxonomique. Le genre et l'espèce des organismes sont enregistrés en latin, *e.g. Mus musculus*, pour la souris. Il est à noter que cette banque contient également les noms d'organismes éteints pour lesquels des séquences sont disponibles, *e.g. Thylacine cynocephallus* (Tigre de Tasmanie) dont le génome mitochondrial a été séquencé [Miller W. *et al*, 2009]. La banque taxonomique du NCBI peut être interrogée interactivement à l'adresse http://www.ncbi.nlm.nih.gov/taxonomy. La banque taxonomique peut également être utile pour restreindre ou exclure les recherches de similarité de séquences dans les banques à un ou plusieurs groupes taxonomiques. Cette option est proposée dans l'interface web BLAST du NCBI à l'adresse http://blast.ncbi.nlm.nih.gov/, dans le champ « *Organism* » du formulaire de soumission.

Enfin, pour les traitements en haut-débit, un fichier plat taxonomique (360 Mb) a été téléchargé de l'institut européen de bioinformatique (EBI) depuis l'adresse ftp://ftp.ebi.ac.uk/pub/databases/taxonomy/taxonomy.dat. Pour chaque organisme et chaque rang de la classification systématique (Domaine, Règne, Embranchement, Classe, Sous-classe, Super-ordre, Ordre, Sous-ordre, Infra-ordre, Super-famille, Famille, Sous-famille, Tribue,

Sous-tribue, Genre, Espèce), ce fichier est structuré selon des identifiants parents-fils (Figure 3a). Cependant, le format de ce fichier n'est pas adapté au traitement haut-débit en raison de sa structure hierarchisée en identifiants parents-fils. J'ai donc écrit un script PERL *ad hoc* pour reformater l'information taxonomique selon une structure de tableur (Figure 3b).

(a)

```
ID: 9606
PARENT ID: 9605
RANK: species
SCIENTIFIC NAME: sapiens
//
ID: 9605
PARENT ID: 207598
RANK: genus
SCIENTIFIC NAME: Homo
//
ID: 207598
PARENT ID: 9604
RANK: subfamily
SCIENTIFIC NAME: Homininae
//
ID: 9604
PARENT ID: 314295
RANK: family
SCIENTIFIC NAME: Hominidae
//
```

(b)

Organisme	**Famille**	**Sous-famille**	**Genre**	**Espèce**
Homo_sapiens	Hominidae	Homininae	Homo	sapiens

Figure 3: **Formats d'information taxonomique**. (a) Extrait du fichier plat de la banque taxonomique fourni par l'EBI pour l'humain. (b) Reformatage en tableur généré par un script PERL *ad hoc* appliqué à la même information.

Pour les alignements de TEX19 et SECTM1, les noms et la classification de 58 mammifères ont été obtenus grâce à la banque taxonomique (Tableau 1 et 2). Pour les alignements multiples de GRα et ERα, les séquences de plus de 100 espèces du poisson à l'homme ont été collectées (information non-montrée).

G.2.1.2.2. Séquences protéiques

Quatre banques de séquences protéiques ont été utilisées: Swissprot, TrEMBL, RefSeq et GenePep [Apweiler R *et al*, 2004]. Dans Swissprot, chaque enregistrement

de séquence est contrôlé et annoté par des biologistes ce qui assure une grande qualité d'information. En novembre 2018, SwissProt contenait 560.000 enregistrements (Figure 4). Cette banque a connue une phase de croissance exponentielle jusqu'en 2010. Par la suite, le rythme des enregistrements s'est largement ralenti vraisemblablement pour favoriser la qualité des annotations par rapport à la quantité de séquences.

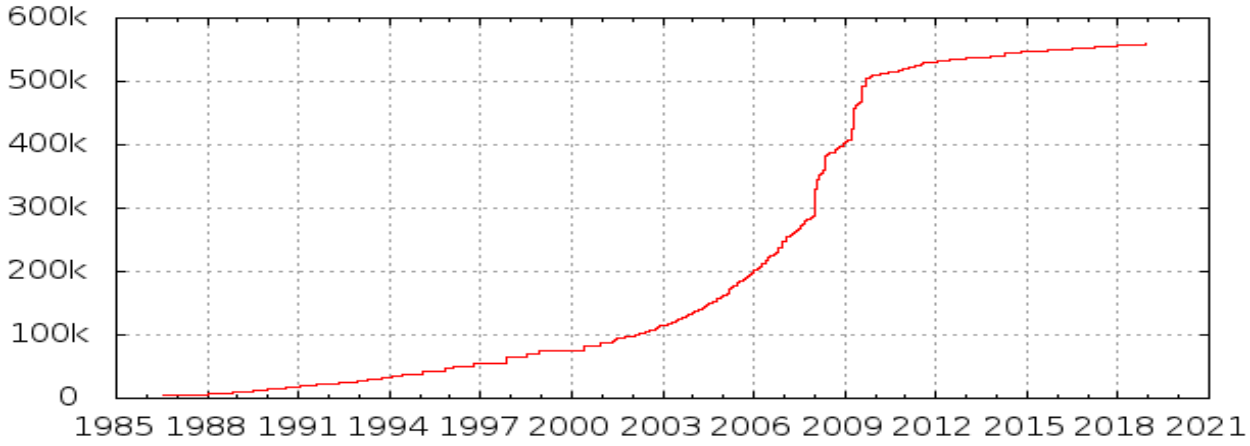


Figure 4: **Evolution du nombre d'enregistrements de séquences dans la banque Swissprot** (Source serveur Expasy, http://www.exapsy.org).

TrEMBL est une banque dans laquelle sont enregistrées les traductions et les annotations automatiques des cDNA des banques nucléotidiques Genbank/EMBL/DDBJ. En 2018, TrEMBL enregistrait 133.500.000 séquences (Figure 5). La banque a connu une croissance exponentielle jusqu'en 2015. A cette date 40 millions d'enregistrements ont été retirés puis la croissance est de nouveau repartie exponentiellement.

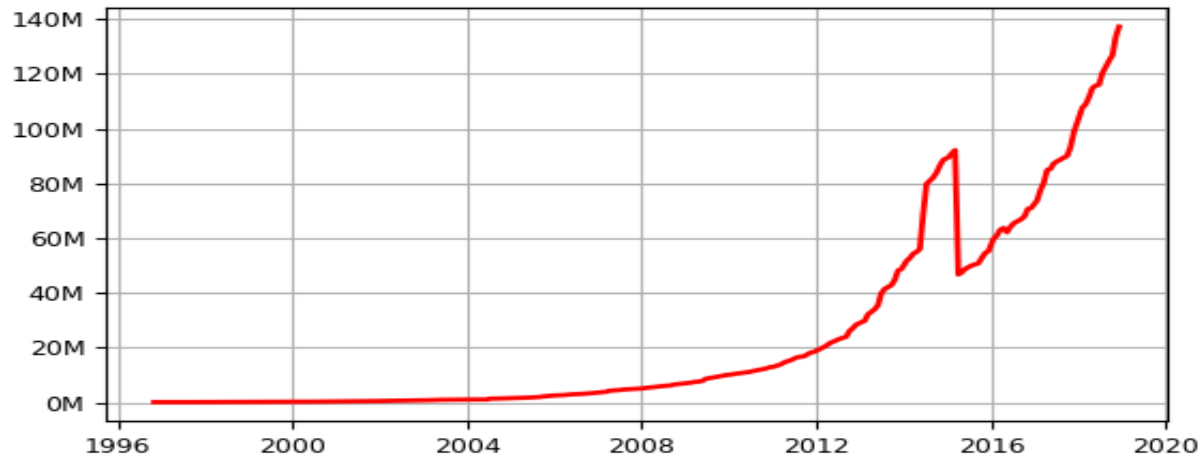


Figure 5: **Evolution du nombre d'enregistrements de séquences dans la banque TrEMBL** (Source : serveur Expasy, http://www.exapsy.org)

L'association de Swissprot et TrEMBL constitue la banque de séquence protéique universelle Uniprot [Uniprot Consortium, 2017]. Cependant, Uniprot contient des enregistrements de séquences redondantes ce qui pose un problème pour les recherches de similarité. A l'instar des administrateurs d'Uniprot qui essaient de pallier au problème de redondance, le NCBI a créé la banque RefSeq [Pruitt K.D. *et al*, 2009]. Enfin, la banque équivalente de TrEMBL maintenue par le NCBI est GenePep. En raison de l'hétérogénéité des banques protéiques, il est important de toutes les interroger pour mener des recherches de séquences exhaustives.

Les numéros d'accés des séquences protéiques collectées pour les alignements multiples de TEX19 et SECTM1 sont reportés dans les tableaux 1 et 2, respectivement. Pour les alignements multiples de GRα et ERα, les séquences de plus de 100 protéines ont été collectées (information non-montrée).

	Organisme (Genre espèce)	Protéine	Super-ordre	Ordre	Sous-ordre	Banque et numéro d'accés
1	*Homo sapiens*	Tex19	Euarchontoglires	Primates	Haplorrhini	UNIPROT:Q8NA77
2	*Pan troglodytes*	Tex19	Euarchontoglires	Primates	Haplorrhini	UNIPROT:G2HJ07
3	*Pan paniscus*	Tex19	Euarchontoglires	Primates	Haplorrhini	REFSEQ:XP_003809005
4	*Gorilla gorilla*	Tex19	Euarchontoglires	Primates	Haplorrhini	UNIPROT:G3R4A3
5	*Pongo abelii*	Tex19	Euarchontoglires	Primates	Haplorrhini	UNIPROT:H2NV59
6	*Nomascus leucogenys*	Tex19	Euarchontoglires	Primates	Haplorrhini	UNIPROT:G1SAY9
7	*Macaca fascicularis*	Tex19	Euarchontoglires	Primates	Haplorrhini	UNIPROT:G7PH33
8	*Macaca mulatta*	Tex19	Euarchontoglires	Primates	Haplorrhini	UNIPROT:F6TUT0
9	*Papio anubis*	Tex19	Euarchontoglires	Primates	Haplorrhini	REFSEQ:XP_003913662
10	*Callithrix jacchus*	Tex19	Euarchontoglires	Primates	Haplorrhini	REFSEQ:XP_002748913
11	*Saimiri boliviensis*	Tex19	Euarchontoglires	Primates	Haplorrhini	REFSEQ:XP_003931867
12	*Otolemur garnettii*	Tex19	Euarchontoglires	Primates	Strepsirrhini	REFSEQ:XP_003786582
13	*Microcebus murinus*	Tex19	Euarchontoglires	Primates	Strepsirrhini	NE
14	*Daubentonia madagascariensis*	Tex19	Euarchontoglires	Primates	Strepsirrhini	NE
15	*Tupaia chinensis*	Tex19	Euarchontoglires	Scandentia	x	GENEPEP:ELW68179
16	*Tupaia belangeri*	Tex19	Euarchontoglires	Scandentia	x	NE
17	*Cavia porcellus*	Tex19	Euarchontoglires	Rodentia	Hystricognathi	REFSEQ:XP_003464960
18	*Octodon degus*	Tex19	Euarchontoglires	Rodentia	Hystricognathi	NE
19	*Heterocephalus glaber*	Tex19	Euarchontoglires	Rodentia	Hystricognathi	UNIPROT:G5BJA3
20	*Chinchilla lanigera*	Tex19	Euarchontoglires	Rodentia	Hystricognathi	NE
21	*Mus musculus*	Tex19.1	Euarchontoglires	Rodentia	Sciurognathi	UNIPROT:Q99MV2
22	*Rattus norvegicus*	Tex19.1	Euarchontoglires	Rodentia	Sciurognathi	UNIPROT:Q5XHY3
23	*Cricetulus griseus*	Tex19.1	Euarchontoglires	Rodentia	Sciurognathi	NE
24	*Microtus ochrogaster*	Tex19.1	Euarchontoglires	Rodentia	Sciurognathi	NE
25	*Mesocricetus auratus*	Tex19.1	Euarchontoglires	Rodentia	Sciurognathi	NE
26	*Mus musculus*	Tex19.2	Euarchontoglires	Rodentia	Sciurognathi	UNIPROT:Q9D5S1
27	*Rattus norvegicus*	Tex19.2	Euarchontoglires	Rodentia	Sciurognathi	UNIPROT:D3ZXL6
28	*Cricetulus griseus*	Tex19.2	Euarchontoglires	Rodentia	Sciurognathi	NE
29	*Microtus ochrogaster*	Tex19.2	Euarchontoglires	Rodentia	Sciurognathi	NE
30	*Mesocricetus auratus*	Tex19.2	Euarchontoglires	Rodentia	Sciurognathi	NE
31	*Bos taurus*	Tex19.α	Laurasiatheria	x	Ruminantia	REFSEQ:XP_002696176
32	*Bos indicus*	Tex19.α	Laurasiatheria	x	Ruminantia	NE

(Le tableau continue sur la page suivante)

33	*Bos grunniens*	Tex19.α	Laurasiatheria	x	Ruminantia	NE
34	*Ovis aries*	Tex19.α	Laurasiatheria	x	Ruminantia	REFSEQ:XP_004013494
35	*Capra hircus*	Tex19.α	Laurasiatheria	x	Ruminantia	NE
36	*Bos taurus*	Tex19.β	Laurasiatheria	x	Ruminantia	REFSEQ:XP_005195863
37	*Bos indicus*	Tex19.β	Laurasiatheria	x	Ruminantia	NE
38	*Bos grunniens*	Tex19.β	Laurasiatheria	x	Ruminantia	NE
39	*Ovis aries*	Tex19.β	Laurasiatheria	x	Ruminantia	REFSEQ:XP_004013493
40	*Capra hircus*	Tex19.β	Laurasiatheria	x	Ruminantia	NE
41	*Sus scrofa*	Tex19	Laurasiatheria	x	x	UNIPROT:I3L8G7
42	*Tursiops truncatus*	Tex19	Laurasiatheria	Cetacea	Odontoceti	NE
43	*Orcinus orca*	Tex19	Laurasiatheria	Cetacea	Odontoceti	NE
44	*Felis catus*	Tex19	Laurasiatheria	Carnivora	Feliformia	NE
45	*Odobenus rosmarus*	Tex19	Laurasiatheria	Carnivora	Caniformia	NE
46	*Mustela putorius*	Tex19	Laurasiatheria	Carnivora	Caniformia	NE
47	*Ailuropoda melanoleuca*	Tex19	Laurasiatheria	Carnivora	Caniformia	REFSEQ:XP_002931395
48	*Leptonychotes weddelii*	Tex19	Laurasiatheria	Carnivora	Caniformia	NE
49	*Myotis lucifugus*	Tex19	Laurasiatheria	Chiroptera	Microchiroptera	NE
50	*Eptesicus fuscus*	Tex19	Laurasiatheria	Chiroptera	Microchiroptera	NE
51	*Equus caballus*	Tex19	Laurasiatheria	Perissodactyla	x	REFSEQ:XP_001914724
52	*Ceratotherium simum*	Tex19	Laurasiatheria	Perissodactyla	x	NE
53	*Erinaceus europaeus*	Tex19	Laurasiatheria	Insectivora	x	NE
54	*Atelerix albiventris*	Tex19	Laurasiatheria	Insectivora	x	NE
55	*Sorex araneus*	Tex19	Laurasiatheria	Insectivora	x	NE
56	*Chrysochloris asiatica*	Tex19	Afrotheria	x	x	NE
57	*Procavia capensis*	Tex19	Afrotheria	Hyracoidea	x	NE
58	*Loxodonta africana*	Tex19	Afrotheria	Proboscidea	x	REFSEQ:XP_003417366

Tableau 1: **Liste des séquences protéiques collectées pour la construction de l'alignement multiple de TEX19.** NE: séquence non-enregistrée dans les banques, la séquence a été obtenue par TBLASTN de la séquence-requête TEX19 du plus proche voisin sur le génome. Pour obtenir toutes les séquences au format FASTA, télécharger l'alignement à l'adresse http://figshare.com/articles/Tex19_multiple_alignment_of_complete_protein_sequences/1491363. X : le rang taxonomique n'est pas défini pour cet organisme. Certaines cellules du tableau sont en couleur pour aider la lecture des groupes taxonomiques.

	Organisme (Genre espèce)	**Protéine**	**Super-ordre**	**Ordre**	**Sous-ordre**	**Banque et numéro d'accés**
1	*Homo sapiens*	Sectm1	Euarchontoglires	Primates	Haplorrhini	UNIPROT:Q8WVN6
2	*Pan paniscus*	Sectm1	Euarchontoglires	Primates	Haplorrhini	REFSEQ:XP_003809004
3	*Pan troglodytes*	Sectm1	Euarchontoglires	Primates	Haplorrhini	UNIPROT:H2QE48
4	*Gorilla gorilla*	Sectm1	Euarchontoglires	Primates	Haplorrhini	UNIPROT:G3R187
5	*Nomascus leucogenys*	Sectm1	Euarchontoglires	Primates	Haplorrhini	UNIPROT:G1QP15
6	*Macaca mulatta*	Sectm1	Euarchontoglires	Primates	Haplorrhini	UNIPROT:G7NHH8
7	*Macaca fascicularis*	Sectm1	Euarchontoglires	Primates	Haplorrhini	NE
8	*Papio anubis*	Sectm1	Euarchontoglires	Primates	Haplorrhini	REFSEQ:XP_003913661
9	*Callithrix jacchus*	Sectm1	Euarchontoglires	Primates	Haplorrhini	REFSEQ:XP_003733133
10	*Saimiri boliviensis*	Sectm1	Euarchontoglires	Primates	Haplorrhini	REFSEQ:XP_003931868
11	*Octodon degus*	Sectm1	Euarchontoglires	Rodentia	Hystricognathi	REFSEQ:XP_004639605
12	*Chinchilla lanigera*	Sectm1	Euarchontoglires	Rodentia	Hystricognathi	NE
13	*Cavia porcellus*	Sectm1	Euarchontoglires	Rodentia	Hystricognathi	UNIPROT:H0WDS7
14	*Heterocephalus glaber*	Sectm1	Euarchontoglires	Rodentia	Hystricognathi	REFSEQ:XP_004894655
15	*Mus musculus*	Sectm1.a	Euarchontoglires	Rodentia	Sciurognathi	UNIPROT:Q921W8
16	*Rattus norvegicus*	Sectm1.a	Euarchontoglires	Rodentia	Sciurognathi	UNIPROT:Q6AYS0
17	*Mesocricetus auratus*	Sectm1.a	Euarchontoglires	Rodentia	Sciurognathi	REFSEQ:XP_005070228

(Le tableau continue sur la page suivante)

18	*Cricetulus griseus*	Sectm1.a	Euarchontoglires	Rodentia	Sciurognathi	REFSEQ:XP_003496891
19	*Microtus ochrogaster*	Sectm1.a	Euarchontoglires	Rodentia	Sciurognathi	NE
20	*Mus musculus*	Sectm1.b	Euarchontoglires	Rodentia	Sciurognathi	UNIPROT:Q9JL59
21	*Rattus norvegicus*	Sectm1.b	Euarchontoglires	Rodentia	Sciurognathi	UNIPROT:E9PTB4
22	*Mesocricetus auratus*	Sectm1.b	Euarchontoglires	Rodentia	Sciurognathi	NE
23	*Cricetulus griseus*	Sectm1.b	Euarchontoglires	Rodentia	Sciurognathi	REFSEQ:XP_003496890
24	*Microtus ochrogaster*	Sectm1.b	Euarchontoglires	Rodentia	Sciurognathi	NE
25	*Sus scrofa*	Sectm1	Laurasiatheria	x	x	NE
26	*Bos taurus*	Sectm1.α	Laurasiatheria	x	Ruminantia	UNIPROT:A6QQD6
27	*Bos grunniens*	Sectm1.α	Laurasiatheria	x	Ruminantia	NE
28	*Bos_indicus*	Sectm1.α	Laurasiatheria	x	Ruminantia	NE
29	*Ovis aries*	Sectm1.α	Laurasiatheria	x	Ruminantia	NE
30	*Capra hircus*	Sectm1.α	Laurasiatheria	x	Ruminantia	NE
31	*Bos taurus*	Sectm1.β	Laurasiatheria	x	Ruminantia	UNIPROT:A4IFS8
32	*Bos grunniens*	Sectm1.β	Laurasiatheria	x	Ruminantia	NCBI-NR:ELR44956
33	*Bos indicus*	Sectm1.β	Laurasiatheria	x	Ruminantia	NE
34	*Ovis aries*	Sectm1.β	Laurasiatheria	x	Ruminantia	REFSEQ:XP_004023415
35	*Capra hircus*	Sectm1.β	Laurasiatheria	x	Ruminantia	NE
36	*Equus caballus*	Sectm1	Laurasiatheria	Perissodactyla	x	NE
37	*Ceratotherium simum*	Sectm1	Laurasiatheria	Perissodactyla	x	REFSEQ:XP_004433025
38	*Orcinus orca*	Sectm1	Laurasiatheria	Cetacea	Odontoceti	REFSEQ:XP_004275784
39	*Tursiops truncatus*	Sectm1	Laurasiatheria	Cetacea	Odontoceti	REFSEQ:XP_004310350
40	*Felis_catus*	Sectm1	Laurasiatheria	Carnivora	Feliformia	NE
41	*Canis_lupus*	Sectm1	Laurasiatheria	Carnivora	Caniformia	NE
42	*Odobenus_rosmarus*	Sectm1	Laurasiatheria	Carnivora	Caniformia	REFSEQ:XP_004404578
43	*Mustela_putorius*	Sectm1	Laurasiatheria	Carnivora	Caniformia	REFSEQ:XP_004811174
44	*Ursus_maritimus*	Sectm1	Laurasiatheria	Carnivora	Caniformia	NE
45	*Ailuropoda_melanoleuca*	Sectm1	Laurasiatheria	Carnivora	Caniformia	UNIPROT:G1MAU9
46	*Leptonychotes_weddellii*	Sectm1	Laurasiatheria	Carnivora	Caniformia	NE
47	*Pteropus_vampyrus*	Sectm1	Laurasiatheria	Chiroptera	Megachiroptera	NE
48	*Pteropus_alecto*	Sectm1	Laurasiatheria	Chiroptera	Megachiroptera	GENEPEP:ELK12220
49	*Myotis_davidii*	Sectm1	Laurasiatheria	Chiroptera	Microchiroptera	GENEPEP:ELK38343
50	*Sorex_araneus*	Sectm1	Laurasiatheria	Insectivora	x	REFSEQ:XP_004621179
51	*Erinaceus_europeus*	Sectm1	Laurasiatheria	Insectivora	x	NE
52	*Condylura_cristata*	Sectm1	Laurasiatheria	Insectivora	x	REFSEQ:XP_004695893
53	*Chrysochloris asiatica*	Sectm1	Afrotheria	X	x	NE
54	*Procavia capensis*	Sectm1	Afrotheria	Hyracoidea	x	NE
55	*Loxodonta_africana*	Sectm1	Afrotheria	Proboscidea	x	NE
56	*Echinops_telfairi*	Sectm1	Afrotheria	x	x	REFSEQ:XP_004709269
57	*Oryctorepus_afer*	Sectm1	Afrotheria	Tubulidentata	x	NE
58	*Dasypus_novemcinctus*	Sectm1	Xenarthra	Cingulata	x	REFSEQ:XP_004478808

Tableau 2: **Liste des séquences protéiques collectées pour la construction de l'alignement multiple de SECTM1**. NE: séquence non-enregistrée dans les banques, la séquence a été obtenue par TBLASTN de la séquence-requête SECTM1 du plus proche voisin sur le génome. Pour obtenir toutes les séquences, au format FASTA, télécharger l'alignement à l'adresse http://figshare.com/articles/Sectm1_multiple_alignment_of_complete_protein_sequences/149 13634. X : le rang taxonomique n'est pas défini pour cet organisme. Certaines cellules du tableau sont en couleur pour aider la lecture des groupes taxonomiques.

G.2.1.2.3. Séquences d'ADN génomique

Plusieurs banques de séquences d'ADN génomique ont été utilisées. Pour le projet *Tex19* et *Sectm1*, c'est la division WGS (« *Whole Genome Shotgun* ») de Genbank (NCBI) qui s'est révélée la plus riche en séquences homologues. En effet, l'annotation automatique des génomes en cours de séquençage n'avait pas encore été terminée pour de nombreux mammifères [Lindblad-Toh K. *et al*, 2011]. Cette banque contient des séquences nucléotidiques en ébauche (« *draft* ») qui sont mises à disposition sans annotation. Les protéines codées par les gènes *Tex19* et *Sectm1* chez les mammifères ont été obtenues par TBLASTN (séquence-requête protéique ; banque de séquences nucléiques cibles traduites dans les 6 cadres de lecture). Des recherches de similarité de séquences ont également été menées par TBLASTN sur la banque de génomes complets maintenue au NCBI et sur le site Ensembl https://www.ensembl.org. Ces banques sont extrêmement utiles pour les projets de génomique comparative. Enfin, la banque « Génomes en ligne » (GOLD) [Mukherjee S. *et al*, 2018] a été consultée à l'adresse http://gold.jgi.doe.gov pour connaître la liste des génomes en cours de séquençage et les sites internets spécialisés où ces séquences sont accessibles, *e.g*. SkateBase [Wyffels J. *et al*, 2014]. En 2018, les statistiques de la banque GOLD montrent que les projets de séquençages de génomes bactériens sont très majoritaires (120.000) tandis que les eucaryotes comptent environ 50.000 projets (Figure 6).

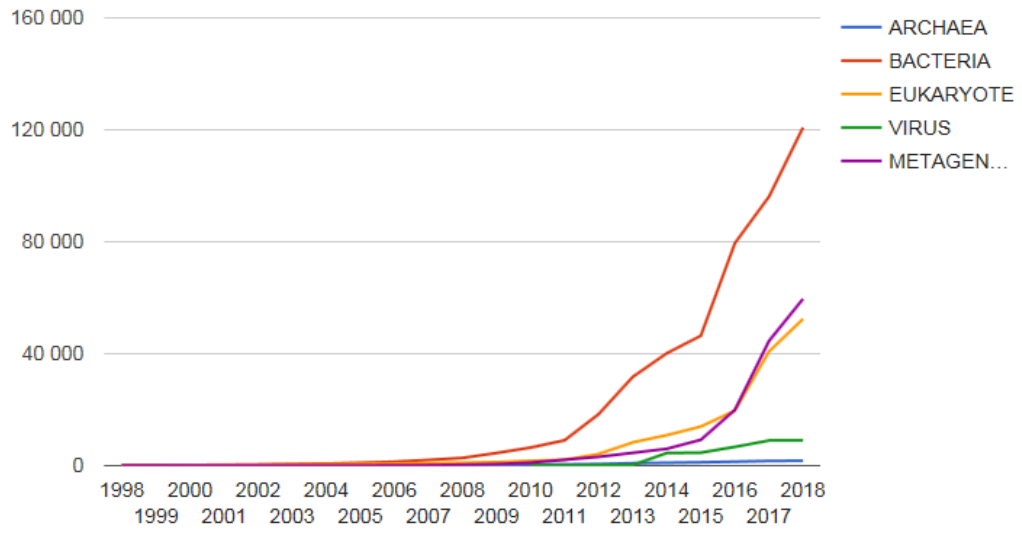


Figure 6: **Evolution des projets de séquençage de génomes par domaine taxonomique** (Source : Banque GOLD, http://gold.jgi.doe.gov)

G.2.1.2.4. Données transcriptomiques

La banque « *sequence read archive* » (SRA) [Kodama Y. *et al*, 2012] stocke et gère des données transcriptomiques (RNA-seq), cistromiques (ChIP-seq) et génomiques produites par des instruments de séquençage à haut-débit (Illumina, Roche 454, PacBio, SOLiD). Elle montre une croissance très forte depuis les 10 dernières années pour atteindre aujourd'hui le peta-byte (10^{15}) de données (Figure 7). Elle est interrogeable à l'adresse https://www.ncbi.nlm.nih.gov/sra. Les niveaux d'expression de *Tex19* et *Sectm1* dans le testicule et le placenta de différentes espèces ont été obtenus par transcriptomes RNA-seq écrits sous forme de fichiers plats (FASTQ) de séquences nucléotidiques (« *reads* »). Chaque transcriptome utilisé est reporté dans le tableau ci-dessous avec son numéro d'accés chez SRA (Tableau 3).

L'expression de *Tex19* dans les tissus adultes sains a été obtenue à l'aide de l'atlas du protéome humain (HPA) [Fagerberg L. *et al*, 2014] à l'adresse https://www.proteinatlas.org tandis que l'expression du gène en lignées cellulaires cancéreuses a été obtenue d'une part au moyen de l'encyclopédie CCLE (Institut Broad), soit 934 lignées [Barretina J. *et al*, 2012], et d'autre part au moyen de l'atlas Genentech, soit 675 lignées [Klijn C. *et al*, 2015]. Ces 2 dernières banques peuvent être interrogées à l'adresse https://www.ebi.ac.uk/gxa/home.

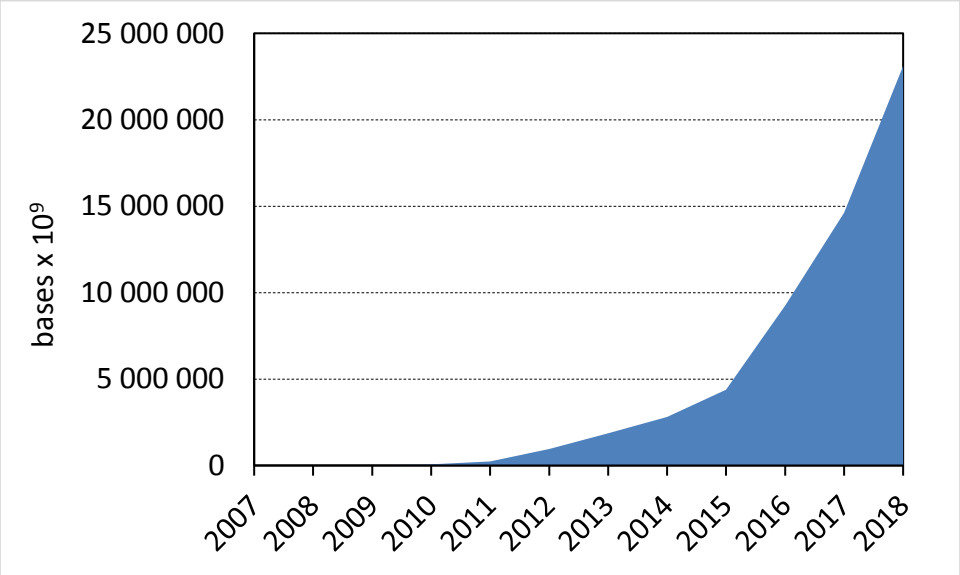


Figure 7: **Evolution du nombre de bases enregistrées dans la banque SRA**.

	Code SRA	Organisme	Tissu	Nombre de séquences	Longueur de séquence	Instrument de séquençage
1	SRR306858	*Homo sapiens*	Testicule	32.444.809	76	GaIIx
2	SRR306857	*Homo sapiens*	Testicule	9.224.054	76	GaIIx
3	SRR525100	*Homo sapiens*	Testicule	15.279.980	72	GaIIx
4	SRR306825	*Pan troglodytes*	Testicule	26.745.671	76	GaIIx
5	SRR306837	*Pan paniscus*	Testicule	16.151.902	76	GaIIx
6	SRR306810	*Gorilla gorilla*	Testicule	21.124.809	76	GaIIx
7	SRR594472	*Macaca mulatta*	Testicule	27.804.316	80	HiSeq 2000
8	SRR389103	*Macaca mulatta*	Testicule	50.362.238	90	HiSeq 2000
9	SRR594454	*Macaca mulatta*	Testicule	28.489.772	80	HiSeq 2000
10	SRR594463	*Macaca mulatta*	Testicule	115.441.819	80	HiSeq 2000
11	SRR594490	*Bos taurus*	Testicule	38.368.885	40	HiSeq 2000
12	SRR594499	*Bos taurus*	Testicule	25.780.071	90	HiSeq 2000
13	SRR594481	*Bos taurus*	Testicule	107.483.976	80	HiSeq 2000
14	SRR594401	*Mus musculus*	Testicule	109.199.938	50	HiSeq 2000
15	SRR594409	*Mus musculus*	Testicule	116.525.147	80	HiSeq 2000
16	SRR594418	*Mus musculus*	Testicule	35.789.834	40	HiSeq 2000
17	SRR306775	*Mus musculus*	Testicule	19.973.708	76	GaIIx
18	SRR594427	*Rattus norvegicus*	Testicule	115.244.483	50	HiSeq 2000
19	SRR594436	*Rattus norvegicus*	Testicule	114.820.645	75	HiSeq 2000
20	SRR594445	*Rattus norvegicus*	Testicule	40.207.033	40	HiSeq 2000
21	SRR496257	*Mus musculus*	Placenta	24.094.163	36	HiSeq 2000
22	SRR496258	*Mus musculus*	Placenta	19.787.649	36	HiSeq 2000
23	SRR1035130	*Bos taurus*	Placenta	8.793.462	49	HiSeq 2000
24	SRR1035129	*Bos taurus*	Placenta	8.495.283	49	HiSeq 2000
25	SRR638936	*Homo sapiens*	Placenta	29.852.422	72	GaIIx
26	SRR638941	*Homo sapiens*	Placenta	30.572.627	54	GaIIx
27	SRR638937	*Homo sapiens*	Placenta	25.393.875	72	GaIIx
28	SRR638939	*Homo sapiens*	Placenta	32.668.147	54	GaIIx
29	SRR638932	*Homo sapiens*	Placenta	23.328.863	72	GaIIx
30	SRR635193	*Homo sapiens*	Placenta	27.265.881	54	GaIIx

Tableau 3: **Transcriptomes RNA-seq utilisés pour la mesure des niveaux d'expression des gènes *Tex19* et *Sectm1* chez différents tissus et différentes espèces**. Les instruments de séquençage sont des ILLUMINA Genome analyzer IIx (GaIIx) ou HiSeq 2000.

Le RNA-seq SRR1296580 (100.578.707 reads, 50 bases, foie de souris adulte, ILLUMINA HiSeq 2000) a été utilisé pour vérifier la séquence de l'ARNm codant le domaine F du GRα.

G.2.1.2.5. Structures 3D

La « *protein data bank* » (PDB) est la banque qui collecte les structures 3D de macromolécules résolues par RMN, cristallographie aux rayons-X et cryo-EM [Berman H.M. *et al*, 2000]. Elle connaît une croissance soutenue depuis 1990 et enregistre aujourd'hui plus de 150.000 structures au total (Figure 8).

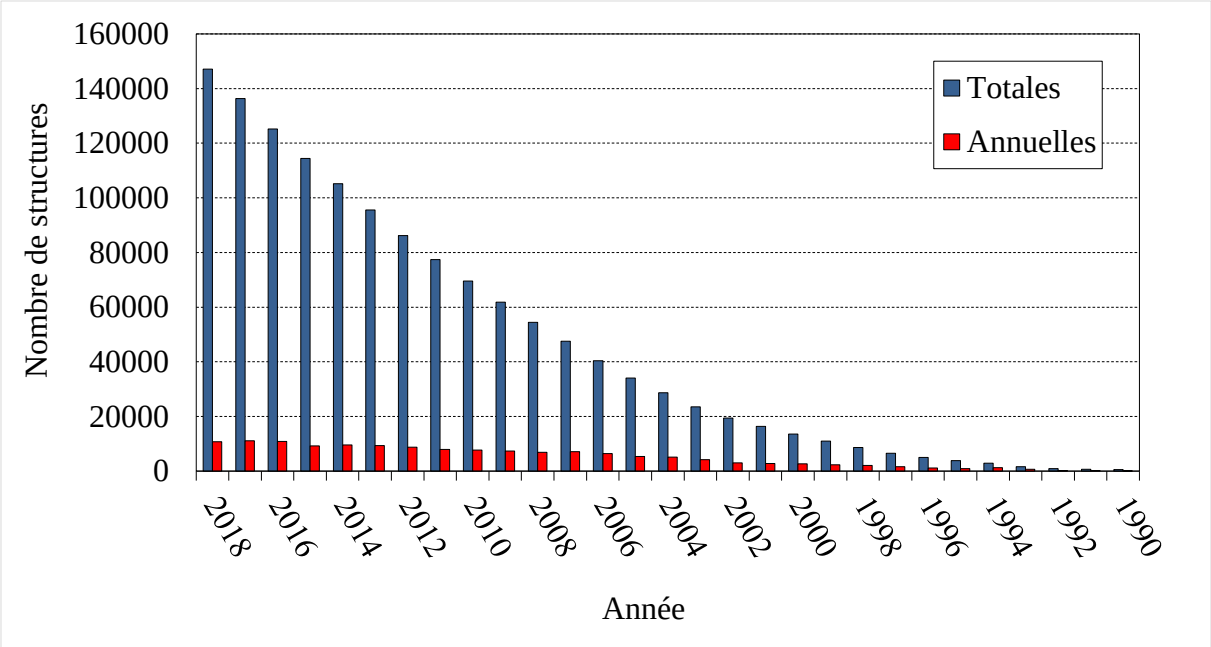


Figure 8: **Evolution du nombre de structures 3D enregistrées dans la PDB**.

Les structures de LBD de GRα ont été collectées exhaustivement et sont au nombre de 21. Les numéros d'accés PDB sont reportés dans le Tableau 4. Tous les LBDs sont en forme « holo », c'est-à-dire avec un ligand logé dans la poche hydrophobe. Ces structures présentent une hétérogénéité de ligands (agonistes, antagonistes), de résolution (1.55 à 3 Å) et sont humaines à l'exception de 3 structures de souris. Pour notre étude, les additifs de cristallisation et les coactivateurs ont été supprimés tandis que les molécules d'eau des cristaux ont été conservées. Une seule structure de LBD d'ERα a été utilisée, *i.e.* PDB:1G50 [Eiler S. *et al*, 2001]. Pour le projet *Tex19/Sectm1* un modèle moléculaire par homologie a été construit en utilisant le PDB:3SOB [Bourhis E. *et al*, 2011] comme support structural de domaine immunoglobuline (Ig).

	Code PDB	Ligand	Résolution (Å)	Auteurs et date de publication ou mise en ligne de la structure
1	5G3J	E7T*	2.4	Berger M. *et al.*, 2017
2	5G5W	R8C*	2.2	Hemmerling M. *et al.*, 2017
3	4UDC	Dexaméthasone*	2.5	Edman K. *et al.*, 2015
4	4UDD	Desisobuytyryl c.*	1.8	Edman K. *et al.*, 2015
5	4P6W	Mométasone f.*	1.95	He Y. *et al.*, 2014
6	4P6X	Cortisol**	2.5	He Y. *et al.*, 2014
7	4LSJ	$C_{25}H_{25}NO_4S$*	2.35	Carson M. *et al.*, 2014
8	4CSJ	NN7*	2.3	Edman K. *et al.*, 2014
9	4MDD	29M+	2.4	Luz J.G. and Coghlan M.J., 2014
10	3MNP	Dexaméthasone*	1.5	Seitz T. *et al.*, 2010
11	3MNE	Dexaméthasone*	1.96	Seitz T. *et al.*, 2010
12	3MNO	Dexaméthasone*	1.55	Seitz T. *et al.*, 2010
13	3K22	Alanina*	3.0	Biggadike K.B. *et al.*, 2009
14	3H52	RU486+	2.8	Schoch G.A. *et al.*, 2010
15	3K23	Prolina*	3.0	Biggadike K.B. *et al.*, 2009
16	3CLD	Fluticasone f.*	2.84	Biggadike K.B. *et al.*, 2008
17	3BQD	Deacylcortivazol*	2.5	Suino-Powell K. *et al.*, 2008
18	3E7C	GSK866*	2.15	Madauss K.P. *et al.*, 2008
19	1M2Z	Dexaméthasone*	2.5	Bledsoe R.B. *et al.*, 2002
20	1P93	Dexaméthasone*	2.7	Kauppi B. *et al.*, 2003
21	1NHZ	RU486+	2.3	Kauppi B. *et al.*, 2003

Tableau 4 : **Structures 3D collectées pour le LBD de GRα.** Toutes les structures ont été obtenues par cristallographie aux rayons –X. All structures are human except 3MNE, 3MNO and 3MNP which originate from mouse. ** ligand naturel chez l'humain, * agoniste, +antagoniste.

G.2.2. Méthodes

G.2.2.1. Analyse d'évolution de séquence

G2.2.1.1. Notions fondamentales

En phylogénèse moléculaire, il est important de distinguer l'homologie de la similarité de séquence. Je me propose de faire ici un rappel de ces notions fondamentales tant elles sont importantes et évoquées de façon récurrente dans tout le mémoire de thèse. Deux gènes sont homologues s'ils partagent une séquence ancestrale commune. Entre 2 homologues, la ressemblance des séquences est telle qu'elle ne peut pas être dûe au hasard. Au contraire, la similarité entre 2 séquences est une ressemblance qui peut être le résultat du hasard et n'implique pas une origine ancestrale commune. De plus, le concept d'homologie intègre 3 notions importantes pour comprendre l'histoire évolutive des gènes i) l'orthologie ii) la

paralogie et iii) la xénologie. Deux gènes homologues sont orthologues s'ils ont été séparés par un évènement de spéciation, c'est-à-dire les gènes sont présents sur le génome de 2 espèces d'organismes différents. Deux gènes homologues sont paralogues s'ils sont le résultat d'un évènement de duplication sur le génome d'une même espèce. Enfin, 2 gènes homologues sont xénologues par suite d'un évènement de transfert de séquence entre 2 organismes d'espèces différentes.

G.2.2.1.2. Comparaison de régions chromosomiques

Sur le génome humain, souris et rat, les loci des gènes *Tex19* et *Sectm1* ont été obtenus à l'aide du navigateur UCSC (Université de Californie de Santa Cruz) à l'adresse http://genome.ucsc.edu [Kent W.J. *et al*, 2002] et les versions des génomes utilisés sont GRCh38/hg38, GRCm38/mm10 et RGCS6.0/rn6, respectivement. Les coordonnées sur le chromosome des exons des gènes *Tex19* et *Sectm1* ont été obtenues à l'aide du *Table Browser* d'UCSC. Les longueurs des éléments génétiques ont été calculées par un script PERL écrit *ad hoc*. L'outil GSDS 2.0 [Hu B. *et al*, 2015] a été utilisé pour réaliser la représentation graphique du morcellement des exons et des introns des gènes.

G.2.2.1.3. Comparaison de séquences protéiques 2 à 2

Quatre méthodes de comparaison de séquences primaires ont été utilisées en fonction du cas à traiter. Les comparaisons des protéines complètes humaines TEX19, SECTM1, et souris Tex19.1, Tex19.2, Sectm1a et Sectm1b ont été réalisées par diagramme de points avec l'outil « *dotpath* » du logiciel EMBOSS [Rice P. *et al*, 2000] à l'adresse http://www.bioinformatics.nl/emboss-explorer. Pour obtenir les pourcentages d'identité et de similarité entre séquences, nous avons utilisé le programme « *needle* » (algorithme de Needleman & Wunsch) si les séquences étaient de longueurs sensiblement équivalentes ou l'outil « *water* » (algorithme de Smith & Waterman) dans le cas contraire. Enfin, nous avons comparé les séquences primaires des LBDs d'ERα et GRα en nous appuyant sur une superposition structurale. Les enregistrements PDB:1G50 (ERα) et PDB:1M2Z (GRα) ont été soumis à l'outil web SALIGN [Braberg H. *et al*, 2012] disponible à l'adresse http://modbase.compbio.ucsf.edu/salign/.

G.2.2.1.4. Recherche de similarité de séquence dans les banques

A partir d'une séquence-requête protéique, les recherches de similarité ont été menées soit à l'aide de l'outil BLASTP 2.8.1 (matrice de score de substitution BLOSUM62,

taille du mot = 6, Espérance seuil = 10, pénalité de création et d'extension de gap 11 et 1, respectivement) dans les banques SwissProt, Refseq, TrEMBL et GenePep, soit avec l'outil TBLASTN 2.8.1 sur les séquences génomiques (division WGS de Genbank). Toutes les requêtes ont été soumises sur l'interface https://blast.ncbi.nlm.nih.gov/Blast.cgi du NCBI ou bien le serveur Ensembl à l'adresse https://www.ensembl.org/ .

G.2.2.1.5. Recherche d'homologues distants

Les recherches d'homologues distants ont été réalisées avec l'outil PSI-BLAST [Altschul S.F. *et al*, 1997 ; Altschul S.F. & Koonin E.V., 1998] dans la banque RefSeq sur l'interface https://blast.ncbi.nlm.nih.gov/Blast.cgi du NCBI avec les paramètres par défaut (matrice de score de substitution BLOSUM62, taille du mot = 3, Espérance seuil de la 1ère recherche dans la banque = 10, Espérance seuil d'intégration de séquence dans les profiles = 0.005, pénalité de création et d'extension de gap 11 et 1, respectivement).

G.2.2.1.6. Construction d'alignements multiples

Les alignements multiples des protéines TEX19 et SECTM1 ont été construits avec l'outil PipeAlign [Plewniak F. *et al*, 2003] tandis que les alignements des LBDs de GRα et ERα ont été produits avec le programme MAFFT à l'adresse http://mafft.cbrc.jp/alignment/software [Katoh K. *et al*, 2017]. La qualité des alignements a été affinée manuellement dans l'éditeur Jalview 2.9 [Waterhouse A.M. *et al*, 2009].

G.2.2.1.7. Scores de conservation des acides aminés

Dans les alignements multiples de séquences protéiques, les scores de conservation des acides-aminés ont été calculés en utilisant l'algorithme de Livingstone & Barton [Livingstone C.D. & Barton G.J., 1993] intégré dans Jalview 2.9 [Waterhouse A.M. *et al*, 2009]. Ces scores vont de 0 pour une position très variable de l'alignement à 11 pour un résidu invariant et s'appuie sur un regroupement des acides-aminés selon leurs propriétés physico-chimiques (Figure 9).

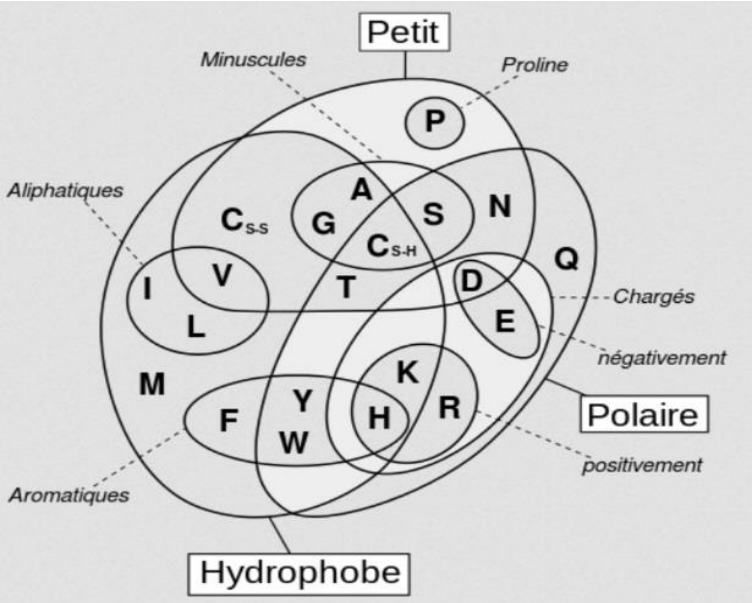


Figure 9: **Les 20 acides-aminés regroupés selon leurs propriétés physico-chimiques**. Un résidu peut appartenir à plusieurs groupes. D'après Livingstone C.D. & Barton G.J. (1993). C_{S-S} : Cystéine impliquée dans un pont disulfure ; C_{S-H} : Cystéine dont la chaîne latérale est réduite.

G.2.2.1.8. Construction de logos

Les logos sont une représentation graphique efficace de motifs conservés dans les séquences protéiques ou nucléiques [Crooks G.E. *et al*, 2004]. Pour les acides nucléiques ils sont utilisés pour définir les sites de fixation des facteurs de transcription sur l'ADN. Pour les protéines, ils servent à montrer de courtes séquences peptidiques conservées qui souvent incluent des résidus fonctionnels, tels que les résidus catalytiques des enzymes. Les logos sont obtenus à partir d'alignement multiples de séquences. Basiquement, à chaque position du logo, les symboles monolettres des acides-aminés ou des bases présents dans la colonne de l'alignement multiple de séquences sont empilées en leur attribuant une hauteur corrélée à leur fréquence. De plus, la hauteur totale des symboles empilés d'une position du logo dépend de la variété des résidus ou bases présents dans la colonne de l'alignement correspondante. Plus une position de l'alignement de séquence présente une grande variabilité de résidus, plus la somme des hauteurs des lettres du logo sera écrasée. Cette hauteur totale est donnée par la formule

$$R_i = log_2 N - (- \sum_{k=1}^{K} f_k \ log_2 \ f_k)$$

Avec N = jeu de symboles, 20 pour les acides-aminés
i = position de l'alignement reportée dans le logo
K = jeu de résidus présents dans la colonne i
f_k = fréquence du résidu k dans la colonne i

La hauteur maximale d'une position du logo pour les polypeptides est donc log_2 20 = 4,2

La hauteur H d'un symbole *k* à la position *i* est donnée par

$$H_{k,i} = R_i \times f_k$$

G.2.2.1.9. Phylogénèse moléculaire

Les arbres phylogénétiques de TEX19 et SECTM1 ont été construits avec l'outil MEGA 6 [Tamura K. *et al*, 2013] en utilisant l'algorithme du maximum de vraisemblance (« *maximum likelyhood (ML)* ») ou celui du rapprochement de voisins (« *neighbour-joining (NJ)* »). Le modèle de substitution des acides-aminés a été choisi avec l'outil ProtTest [Abascal F. *et al*, 2005; Darriba D. *et al*, 2011]. Pour tester la robustesse des noeuds des arbres, un bootstrap de 500 réplicats a été mené. Enfin, les annotations des arbres phylogénétiques ont été élaborées dans l'interface d'iTOL [Letunik I. & Bork P., 2016].

G.2.2.1.10. Coévolution de protéines

La coévolution des protéines TEX19 et SECTM1 a été étudiée par 3 méthodes. Premièrement, les copies de gènes ont été dénombrées chez 8 espèces choisies de mammifères placentaires. Ensuite, un test d'indépendance du χ^2 de Pearson avec 5% de risque d'erreur de type I a permis de décider entre 2 hypothèses : H_o « les nombres de copies de gènes *Tex19* et *Sectm1* sont indépendants » *versus* H_1 « les nombres de copies de gènes *Tex19* et *Sectm1* ne sont pas indépendants ». Deuxièmement, des vecteurs de distances entre protéines orthologues ont été calculés selon une méthode décrite par Goh C-S. *et al* (2000) et Pasoz F. & Valencia A. (2008). Les séquences humaines TEX19 et SECTM1 ont été choisies comme références. Pour les protéines des espèces représentées simultanément dans les 2 alignements multiples, les distances entre ces protéines et les séquences humaines sont calculées selon la formule :

$$d_{i,h} = S_{i,h} / P_{i,h}$$

Avec $d_{i,h}$ = distance entre la séquence d'espèce *i* et la séquence humaine *h*

$S_{i,h}$ = nombre de substitutions d'acides-aminés observées entre la séquence d'espèce *i* et la séquence humaine *h*

$P_{i,h}$ = nombre de positions considérées

Les positions avec *gaps* ont été exclues du calcul, *i.e.* pour une paire de séquences, une position est prise en compte si les 2 séquences comparées possèdent toutes les 2 un acide-aminé. Une fois les distances calculées pour chaque espèce et chaque protéine, un tableau est construit (Tableau 5). Un premier vecteur de distances d_i est obtenu pour TEX19 et un second pour les distances d'_i de SECTM1. La corrélation entre les distances d_i et d'_i a été déterminée dans le logiciel R avec la fonction cor( ). Cette approche est basée sur l'hypothèse que les taux de substitutions de 2 protéines en coévolution ne sont pas indépendants.

	Espèce 1	…	Espèce *i*	…	Espèce N
Tex19	d_1	…	d_i	…	d_n
Sectm1	d'_1	…	d'_i	…	d'_n

Tableau 5: **Vecteurs de distances entre 2 familles de protéines.** Une espèce peut être assimilée à une dimension dans un espace. Une espèce, *e.g. Homo sapiens*, est choisie comme point de référence.

Enfin, la concordance entre les 2 arbres phylogénétiques a été testée avec le programme MirrorTree [Ochoa D. & Pazos F., 2010] qui s'appuie sur toutes les espèces simultanément présentes dans les 2 arbres et le nombre de copie de gènes.

G.2.2.1.11. Test de résidus conservés à l'interface de contact

Les résidus engagés à l'interface de contact entre 2 protéines ont été déterminés par NACCESS qui calcule la variation d'exposition au solvant pour chaque résidu dans les formes monomériques et assemblées. Nous déterminons ainsi d'une part quels résidus sont enfouis à l'interface de contact et d'autre part quels résidus de surface des protéines demeurent exposés au solvant dans le complexe. Grâce à l'alignement multiple de séquences protéiques, le score de conservation de chaque résidu est connu. Pour les résidus de surface uniquement (enfouis ou non-enfouis), nous choisissons un seuil statistique de conservation égale au 3ème quartile (sur une échelle de 0 - position variable - à 11 - résidu invariant -, les LBDs présentaient un 3ème quartile égal à 10). Si un résidu de surface possède un score de conservation supérieur ou égal à 10, il est considéré comme conservé. Une table de contingence est construite qui dénombre les résidus de surface conservés ou non-conservés et enfouis ou non-enfouis (Tableau 6). Finalement, l'hypothèse de sur-représentation de résidus conservés à l'interface de contact est évaluée par un test de Fisher's Exact à 5% d'erreur de type I.

		Exposition au solvant	
		Enfouis	Non-enfouis
Conservation	Conservés	*w*	*x*
	Non-conservés	*y*	*z*

Tableau 6: **Table de contingence des résidus de surface du complexe.** Les résidus de surface sont dénombrés selon 2 catégories, *i.e.* conservation et exposition au solvant. La conservation et l'exposition au solvant peuvent prendre chacunes deux valeurs, « conservé » ou « non-conservé » et « enfouis » ou « non-enfouis » à l'interface de contact, respectivement ; w, x, y, z représentent les effectifs de classe.

G.2.2.2. Analyse d'expression

G.2.2.2.1. Alignement de « *reads* » sur transcrit

Les niveaux d'expression des ARNm de *Tex19* et *Sectm1* chez différentes espèces ont été mesurés en utilisant des transcriptomes RNA-seq enregistrés dans la banque SRA (voir Ressources). A l'aide d'un script PERL écrit *ad hoc*, les transcriptomes ont été téléchargés et convertis au format BLAST. Pour chaque expérience, le programme MegaBLAST a été utilisé pour extraire les *reads* spécifiques des ARNm de *Tex19* ou *Sectm1*. Le script PERL a ensuite dénombré ces reads et calculé le nombre de reads totaux. Finalement, les niveaux d'expression ont été déterminés dans l'unité RPKM (« *Reads per kilobase of exon and million of reads* »).

G.2.2.2.2. Alignement de « *reads* » sur génome

Afin de vérifier l'intégrité de l'ARNm codant l'extrémité C-terminale de GRα chez la souris, un transcriptome RNA-seq de foie a été téléchargé de la banque SRA. Les *reads* ont été alignés sur le génome de souris grâce au programme Bowtie 1.0.0 [Langmead B. *et al*, 2009]. Ensuite, la région codant l'extrémité C-terminale de GRα a été visualisée avec ses reads alignés dans l'outil IGV 2.3.40 (Integrative Genome Viewer) [Thorvaldsdottir H *et al*, 2013].

G.2.2.3. Analyses structurales

G.2.2.3.1. Prédiction de structures secondaires

Les structures secondaires ont été prédites à partir des séquences protéiques primaires de TEX19 et SECTM1 à l'aide de l'outil PSIPRED [McGuffin LJ *et al*, 2000 ;

Buchan D.W. *et al*, 2013]. PSIPRED utilise un profil de séquences construit par PSI-BLAST et une méthode d'apprentissage.

G.2.2.3.2. Superposition de structures

Les variations de structures entre les positions des carbones α du LBD de GRα (PDB:1M2Z) et du LBD de ERα (PDB:1G50) on été obtenues en i) superposant les 2 LBDs à l'aide de l'outil SALIGN [Braberg H. *et al*, 2012] et en calculant le RMSD (« *Root mean square deviation*») selon l'équation (1), c'est-à-dire la distance moyenne entre les atomes des 2 structures superposées.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(r_i\,(s_A) - r_i(s_B))^2}{N}} \qquad (1)$$

Avec N nombre de carbones α, r_i (S_A) et r_i (S_B) positions des carbones α dans les structures A et B, respectivement.

G.2.2.3.3. Statistiques structurales

La variété structurale des LBDs de GRα obtenus de la PDB a été évaluée à l'aide de l'outil PSS (« *Protein Structure Statistics* ») [Gaillard T. *et al*, 2013]. PSS réalise une superposition de structures et calcul le RMSF (« *Root mean square fluctuation* »), c'est-à-dire la variation moyenne de position pour chaque carbone-α le long de la chaîne principale.

G.2.2.3.4. Modélisation moléculaire par homologie

Le modèle moléculaire du domaine le plus conservé de SECTM1, *i.e.* le domaine A, a été construit par homologie avec le domaine Ig de la structure PDB:3SOB dans Modeller 9.16 [Webb B. *et al*, 2016]. Ce modèle a été validé par MolProbity [Chen V.B. *et al*, 2010] et PROSA [Wiederstein M. & Sippl M.J., 2007]. Des modèles moléculaires des LBDs de GRα et ERα de souris ont été construits à partir des structures PDB:1M2Z et PDB:1G50, respectivement.

G.2.2.3.5. Modélisation et paramètrage des ligands

Les molécules de dexaméthasone et d'estradiol qui sont respectivement des ligands de GRα et ERα ont été modélisées y compris les hydrogènes dans le programme Avogadro 1.0.3. [Hanwell M.D. et al, 2012]. Les paramètres de champ de force des ligands ont été obtenus à l'aide de l'outil PARAMCHEM [Vanommeslaeghe K. *et al*, 2010].

G.2.2.3.6. Recherche de contacts protéine-protéine

Les contacts protéine-protéine dans les cristaux de LBDs ont été déterminés à l'aide de l'outil PISA [Krissinel E. & Henrick K., 2007] et vérifiés dans PyMOL 1.7 (Schrödinger, LCC).

G.2.2.3.7. Surface de contact à l'interface protéine-protéine

L'aire des interfaces protéine-protéine a été calculée au moyen de l'outil NACCESS 2.1 [Hubbard S.J. & Thornton J.M., 1992] téléchargeable à l'adresse http://wolf.bms.umist.ac.uk. NACCESS détermine la surface de la protéine exposée au solvant en faisant rouler sur la protéine une molécule d'eau assimilée à une sphère de rayon de van der Waals 1.4 Å. Finalement, la surface de contact est obtenue en retranchant la surface exposée au solvant du complexe à la somme des surfaces exposées au solvant des 2 monomères séparés et en divisant le résultat par 2.

G.2.2.3.8. Etat de protonation des résidus titrables

Les états de protonation des résidus Asp, Glu, Lys, Arg et His ont été déterminés grâce à l'outil PDB2PQR [Dolinsky T.J. *et al*, 2004] qui intègre le programme PROPKA [Olsson M.H. *et al*, 2011]. Ces états ont été calculés à pH = 7.4.

G.2.2.3.9. Mécanique moléculaire

La mécanique moléculaire est une théorie qui sert à décrire le comportement de molécules simples (méthane, éthane, sucres, lipides, acides aminés) ou de molécules complexes comme les macromolécules biologiques (protéines, acides nucléiques). Si l'énergie potentielle d'une molécule devait être calculée rigoureusement, c'est la théorie de la mécanique quantique qui devrait être mise en oeuvre. Comme celle-

ci prend en compte les électrons, le coût du calcul augmenterait trop rapidement pour traiter de gros systèmes. La mécanique quantique est donc préférée pour les petites molécules et permet en outre de prédire les réactions chimiques (formation/rupture de liaisons covalentes). La mécanique moléculaire est une alternative avantageuse pour le calcul de l'énergie potentielle d'une macromolécule car elle s'appuie sur des modèles physico-chimiques simples pour décrire les intéractions entre atomes (voir ci-après).

G.2.2.3.10. Champ de force

L'énergie potentielle d'un système moléculaire peut être calculée en fonction des coordonnées dans l'espace de ses atomes. La fonction mathématique et les paramètres qui permettent de calculer l'énergie potentielle constituent un « champs de force ». Dans cette représentation, les atomes sont assimilés à des sphères et les liaisons covalentes à des ressorts. Le champ de force décrit les intéractions entre les atomes du système. Ces intéractions se décomposent entre atomes « liés » d'une part et atomes « non-liés » d'autre part. L'énergie potentielle totale est égale à la somme des énergies résultantes des intéractions liées et non-liées qui elles-mêmes dépendent uniquement de la position des atomes dans l'espace (Equation 1).

$$E(n) = E_{\text{lié}}(n) + E_{\text{non-lié}}(n) \qquad (1)$$

Avec n : Positions spatiales des n atomes de la molécule
E(n) : Fonction énergie potentielle qui ne dépend que de n

Les intéractions entre atomes « liés » se décomposent en 5 termes (Equation 2)

$$E_{\text{lié}} = \frac{1}{2}\sum_{liaisons} k_r\,(r - r_0)^2 + \frac{1}{2}\sum_{angles} k_\theta\,(\theta - \theta_0)^2 \qquad (2)$$

Terme de liaison — Terme de distorsion d'angle de valence

$$+\frac{1}{2}\sum_{impropres} k_\omega\,(\omega - \omega_0)^2 + \frac{1}{2}\sum_{UB} k_{UB}\,(S - S_0)^2$$

Terme des angles dièdres impropres — Terme de Urey-Bradley

$$+\frac{1}{2}\sum_{dièdres} k_\phi\,(1 + \cos(n\phi - \delta)^2$$

Terme d'énergie de torsion

Avec r_0 longueur de liaison à l'équilibre
r longueur de liaison
k_r constante de force de la liaison
θ_0 angle à l'équilibre entre 3 atomes reliés par 2 liaisons
θ angle entre 3 atomes reliés par 2 liaisons
k_θ constante de force de l'angle entre 3 atomes reliés par 2 liaisons
ω_0 angle dièdre impropre à l'équilibre
ω angle dièdre impropre
k_ω constante de déformation d'angle dièdre impropre
S_0 distance entre les atomes à l'équilibre
S distance entre les atomes
k_{UB} constante d'Urey-Bradley
ϕ angle dièdre associé à la rotation autour d'une liaison B-C entre 4 atomes consécutifs A, B, C et D
k_ϕ constante de torsion
n periodicité de la rotation (=2 pour les atomes sp2 ou 3 pour les sp3)
δ angle de phase

Le terme de tension rend compte de l'énergie potentielle dûe à une élongation ou une compression des liaisons covalentes autour d'une longueur d'équilibre. Le terme de distorsion d'angle de valence donne l'énergie potentielle pour une variation d'angle autour d'une position d'équilibre entre 2 liaisons covalentes qui relient 3 atomes. Les énergies de ces 2 termes sont décrites par des lois de Hooke.

Les intéractions entre atomes « non-liés » se décomposent en 2 termes (Equation 3)

$$E_{\text{non-lié}} = \sum_{paires} 4\varepsilon_{ij}\left\{\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right\} + \frac{1}{4\pi\varepsilon_r\,\varepsilon_o}\sum_{paires}\left(\frac{q_i q_j}{r_{ij}}\right) \quad (3)$$

Terme de van der Waals Terme électrostatique

Avec σ_{ij} coefficient relié au rayon de van der Waals pour la paire d'atomes *i* et *j*
q_i charge ponctuelle portée par l'atome *i*
ε_{ij} racine carrée du produit des valeurs absolues de l'énergie minimale de vdW entre les atomes i et j
ε_r permittivité relative qui est fonction du milieu
ε_o permittivité du vide (=1)
r_{ij} distance entre les atomes *i* et *j*

G.2.2.3.11. Minimisation d'énergie

La conformation d'une molécule est stable si son énergie potentielle est égale à un minimum d'énergie libre. Pour une structure de départ, des méthodes mathématiques permettent de trouver le minimum d'énergie le plus proche. Après placement des hydrogènes sur les structures cristallines par HBUILD [Brünger A.T. *et al*, 1988], les LBDs de GRα et ERα liés à leurs ligands ont été minimisées avec 500 étapes d'algorithme de descente la plus forte (« *steepest descent (SD)* ») en utilisant la version c37b1 du programme CHARMM [Brooks B.R. *et al*, 2009]. Les interactions électrostatiques et van der Waals ont été tronquées à 14 Å avec des fonctions « *shift* » et « *switch* », respectivement.

G.2.2.3.12. Simulation de dynamique moléculaire pour le calcul d'énergie libre de liaison entre LBDs

La dynamique moléculaire est une méthode qui permet de simuler le mouvement des atomes au cours du temps, y compris les changements conformationnels des molécules. Les positions et les vitesses des atomes sont déterminées par résolution des équations classiques du mouvement de Newton. Pour les protéines, si la structure de la séquence d'intérêt a été résolue par cristallographie aux rayons X, RMN ou cryo-EM, celle-ci est utilisée comme structure de départ pour la dynamique moléculaire. Sinon, la structure initiale est produite par modélisation moléculaire par homologie si la structure d'une séquence homologue a été résolue. On appelle trajectoire, la série de conformations par lesquelles passe la molécule au cours du temps. Selon l'équation du mouvement de Newton, la force (F) s'exerçant sur un atome *i* est égale à la masse (m) de cet atome multipliée par son accélération (a)

$$\vec{F_i} = m_i \vec{a_i} \qquad (1)$$

De plus, la force peut être exprimée comme le gradient de l'énergie potentielle E

$$\vec{F_i} = -\vec{\nabla}_i E \quad (2)$$

Finalement, les équations (1) et (2) conduisent à

$$\frac{-\partial E}{\partial x_i} = m_i \frac{d^2 x_i}{dt^2} \qquad \frac{-\partial E}{\partial y_i} = m_i \frac{d^2 y_i}{dt^2} \qquad \frac{-\partial E}{\partial z_i} = m_i \frac{d^2 z_i}{dt^2} \qquad (3)$$

Avec E énergie potentielle

x_i , y_i et z_i positions de l'atome i dans un espace cartésien à 3 dimensions

t temps

Dans un complexe, les énergies libres de liaison entre monomères dépendent des distances inter-atomiques. Pour prendre en compte les mouvements des atomes dans le calcul des énergies libre de liaison, des simulations de dynamique moléculaire (DM) sont menées. Pour chaques assemblages en homodimère des LBDs de GRα et ERα observés dans les cristaux, les simulations de DM ont été conduites avec le programme NAMD [Phillips J.C. *et al*, 2005] en utilisant le champ de force CHARMM27 [MacKerell A.D. *et al*, 1998]. Les homodimères ont été immergés dans une boîte d'eau explicite TIP3P de dimension 100 x 100 x 100 Å. Des ions chlore et sodium ont été rajoutés à la boîte d'eau à une concentration de 0,15 M. Les simulations ont été démarrées avec 2 phases de minimisation et de chauffage des molécules d'eau tandis que les positions des atomes des protéines et des ligands étaient fixes. Dans la 1ère phase, l'eau a été minimisée avec 1000 étapes d'algorithme de gradient conjugué (GC) et chauffée à 600K. Ensuite, l'eau a été minimisée par 250 étapes de GC et chauffée à 300K. Les complexes et le solvant ont été minimisés par 2000 étapes de GC et chauffés à 300K. Le système a été équilibré pendant 150 ps. Finalement, une phase de production de 10 ns de dynamique moléculaire a été menée.

G.2.2.3.13. Simulation de dynamique moléculaire pour l'étude de la flexibilité structurale du domaine F de GRα

Le protocole de simulation pour le monomère du LBD de GRα était sensiblement le même que le précédent excepté la taille de la boîte d'eau (77 x 77 x 77 Å) et la phase de production de durée 100 ns. La variation de position de chaque résidu autour d'une position moyenne de référence a été étudiée en calculant le RMSF, c'est-à-dire la racine carrée de la moyenne des fluctuations structurales le long de la séquence du LBD (Equation 1).

$$RMSF_i = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(r_i(t) - < r_i >)^2\right)} \qquad (1)$$

Avec la position moyenne de l'atome i calculée sur l'ensemble de la trajectoire donnée par (Equation 2)

$$< r_i > = \frac{1}{N} \sum_{i=1}^{N} r_i\,(t) \qquad (2)$$

Où $r_i\,(t)$ est la position de l'atome i au temps t et N le nombre total de conformations utilisées.

Le RMSF a été calculé avec le programme CHARMM [Brooks B.R. *et al*, 2009]. Les structures de la trajectoire ont été regroupées par l'outil Gromacs 5.0.4 [Pronk S. *et al*, 2013]. L'option gromos a été choisie pour le regroupement. Celle-ci utilise un seuil maximum de 0.1 nm de RMSD pour assigner 2 structures dans le même groupe. Pour chaque groupe, une structure a été sélectionnée pour obtenir une représentation des fluctuations du domaine F.

G.2.2.3.14. Energie libre de liaison (méthode MM/PBSA)

Dans une intéraction protéine-protéine, une méthode combinant la mécanique moléculaire et l'utilisation d'un solvant implicite (« *Molecular Mechanics / Poisson-Boltzmann surface area* ») permet d'identifier les acides-aminés qui contribuent favorablement ou défavorablement à la stabilité du complexe [Kollman P.A. *et al*, 2000, Lafont V. *et al*, 2007]. L'énergie libre d'association de Gibbs ($\Delta G_{association}$) entre 2 macromolécules A et B peut être décrite par un cycle thermodynamique (Figure 10).

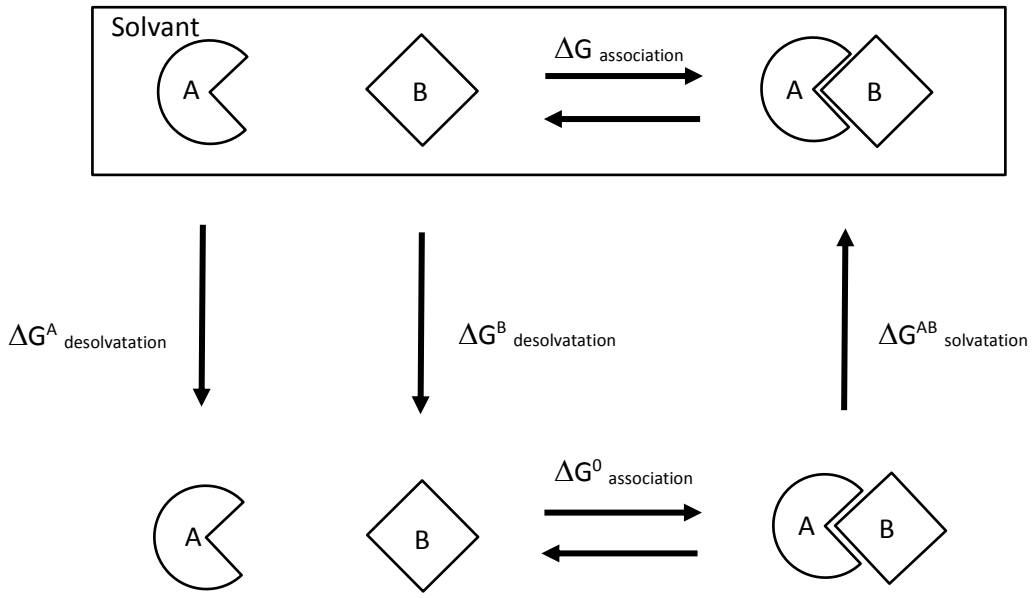


Figure 10: **Schéma du cycle thermodynamique décrivant l'association de 2 macromolécules A et B en phase gazeuse et en solvant**.

Une procédure automatique développée dans le laboratoire [Lafont V. *et al*, 2007] a été mise en œuvre pour obtenir l'énergie libre de liaison totale et la contribution favorable ou défavorable

par résidu à la formation du complexe. Lors de la formation d'un complexe, la variation d'énergie libre de liaison entre 2 protéines peut être exprimée par l'énergie libre de Gibbs (Equations 1 et 2)

$$\Delta G_{assoc} = \Delta H - T\Delta S \quad (1)$$

$$\Delta G_{assoc} \approx \Delta E_{MM} + \Delta G_{solvant} - T\Delta S \quad (2)$$

Avec : ΔE_{MM} variation d'énergie moléculaire mécanique totale

$\Delta G_{solvant}$ énergie libre de solvatation

$-T\Delta S$ variation d'entropie conformationnelle

ΔE_{MM} peut être calculée par l'expression (Equation 3)

$$\Delta E_{MM} = \Delta E_{interne} + \Delta E_{elec} + \Delta E_{vdW} \quad (3)$$

Avec : $\Delta E_{interne}$ est la variation d'énergie associée aux longueurs de liaisons, aux valeurs d'angles et angles dihèdres

ΔE_{elec} et ΔE_{vdW} représentent les termes d'énergies élèctrostatique et van der Waals, respectivement

L'énergie libre de solvatation peut être explicitée en un terme polaire (ΔG_{PB}) et un terme non-polaire (ΔG_{SA}). (Equation 4)

$$\Delta G_{solvant} = \Delta G_{PB} + \Delta G_{SA} \quad (4)$$

Dans notre analyse, l'entropie conformationnelle n'a pas été estimée. De plus, $\Delta E_{interne}$ était égal à 0 car les structures monomériques ont été générées à partir de dimères. Pour cette raison, il n'y a pas de changement conformationnel interne entre les chaînes des dimères et des monomères. A partir de la dynamique moléculaire, un échantillonnage de structures est réalisé en se basant sur l'intéraction coulombienne dans le vide. Pour les intéractions électrostatiques, une constante diélectrique de 1 et un seuil de distance de 12.5 Å avec une fonction de troncature « *shift* » ont été appliquées. Les conformations de la trajectoire ont été réparties selon 10 groupes sur la base de leur énergie électrostatiques. Pour chaque groupe, la conformation montrant l'énergie électrostatique la plus proche de la moyenne du groupe a été extraite. Finalement, l'équation 2 peut s'écrire (Equation 5)

$$\boxed{\Delta G_{assoc} = \Delta E_{elec} + \Delta E_{vdW} + \Delta G_{PB} + \Delta G_{SA}} \quad (5)$$

La contribution de la protéine et du solvant au terme électrostatique a été calculée avec le programme APBS et une grille d'espacement de 0.3 Å [Jurrus E. *et al*, 2018] tandis que les

termes de van der Waals et d'accessibilité au solvant ont été déterminés avec CHARMM [Brooks B.R. *et al*, 2009]. Dans un complexe entre 2 macromolécules, les contributions par résidu à l'énergie libre de liaison peuvent varier en fonction des distances inter-atomiques. En raison du temps de calcul, l'énergie libre de liaison d'un complexe ne peut être détermimée pour chaque conformation d'une trajectoire de DM. Un échantillonnage de conformations doit donc être effectué [Lafont V. *et al*, 2007]. Le terme électrostatique de l'énergie de liaison est calculé pour chaque conformation. Entre la valeur électrostatique la plus élevée et celle la plus basse, 10 intervalles sont créés dans lesquels sont réparties les conformations. Pour chaque intervalle, la moyenne des énergies électrostatiques est calculée. La conformation dont l'énergie se rapproche le plus de la moyenne est extraite. Finalement, l'énergie libre de liaison est calculée sur les 10 conformations ainsi selectionnées.

G.3. Résultats & Discussions

G.3.1. Co-évolution des gènes *Tex19* et *Sectm1*

G.3.1.1. Contexte biologique

Les éléments transposables (ETs) sont des séquences nucléiques répétées dans le génome qui contribuent à son évolution [Kumar C.S. *et al*, 2007]. Parmi les ETs, les rétrotransposons utilisent une séquence intermédiaire d'ARN pour se répliquer. Les séquences LINE (« *long interspersed nuclear elements* ») sont des rétrotransposons qui codent pour 2 protéines, l'ORF1 et l'ORF2 alors que les rétrotransposons ERV (rétrovirus endogènes) codent pour des protéines *gag*, *pro*, *pol*, et *env* caractéristiques des rétrovirus. Certains organismes ont développé des mécanismes pour contrôler l'activité des rétrotransposons, par exemple en méthylant l'ADN. Dans le testicule et le placenta de souris, le gène *Tex19.1* participe à la défense de l'intégrité du génome contre l'activité des rétrotransposons LINE-1 et MMERVK10C (*Mus musculus* ERV K10C) [Ollinger R. *et al*, 2008; Reichmann J. *et al*, 2013; MacLennan M. *et al*, 2017]. Les souris adultes dont le gène *Tex19.1* a été délété présentent une activation des séquences MMERVK10C et des cassures double-brins de l'ADN génomique au cours de la méiose. Chez l'humain, il n'existe qu'un seul gène *Tex19* tandis que 2 paralogues sont présents chez la souris, *Tex19.1* et *Tex19.2* [Kuntz S. *et al*, 2008]. Les gènes *Tex19* et *Sectm1* sont voisins sur le génome humain (cytobande chromosomique 17q25.3) et séparés par une région intergénique d'à peine 26 kb. *Sectm1* s'exprime dans de nombreux tissus comme le

colon, le rein, le foie et le placenta. Alors que *Sectm1* est unique chez l'humain, il existe 2 paralogues chez la souris, *Sectm1a* et *Sectm1b* [Lyman SD *et al*, 2000; Howie D. *et al*, 2013]. De plus, le voisinage des gènes *Tex19.1, Tex19.2, Sectm1a et Sectm1b* a été maintenu sur le génome de souris dans une région d'environ 100 kb (cytobande 11qE2). Enfin, les phylogénèses moléculaires de *Tex19* et *Sectm1* ne sont pas connues. Dans une précédente étude, la disponibilité de séquences TEX19 dans les banques m'avait permis de conclure à la spécificité de la protéine aux mammifères [Kuntz S. *et al*, 2008]. Cependant, les séquençages récents des génomes du monotrème (*Ornithorhynchus anatinus*) et de quelques métathériens (marsupiaux) m'ont permis de restreindre l'existence de TEX19 au groupe des euthériens (mammifères placentaires). Parce que *Tex19* et *Sectm1* sont tous deux uniques et voisins sur le génome humain et parce qu'ils ont été dupliqués et maintenus en proximité chez la souris, j'ai souhaité savoir si les 2 gènes subissent une contrainte coévolutive et, par conséquent s'il existe une relation de fonction entre eux.

G.3.1.2. Etat des connaissances sur la séquence, la structure, et la fonction des protéines TEX19 et SECTM1

Chez l'humain, le gène *Tex19* code pour une protéine TEX19 de 164 acides-aminés. Chez la souris *Tex19.1* et *Tex19.2* codent pour des protéines TEX19.1 et TEX19.2 de 351 et 317 acides-aminés, respectivement. Les structures 3D de ces protéines sont inconnues. Par immunochimie, la localisation de la protéine TEX19.1 a été rapportée dans le cytoplasme [Ollinger R. *et al*, 2008] bien qu'une étude antérieure localisait TEX19.1 et TEX19.2 dans le noyau [Kuntz S. *et al*, 2008]. Dans les cellules souches embryonnaires de souris, il a été montré récemment que TEX19.1 intéragit directement avec la protéine ORF1 du rétrotransposon LINE-1 et stimule sa poly-ubiquitylation et sa dégradation par le protéasome [MacLennan M. *et al*, 2017].

Chez l'humain, *Sectm1* code pour une protéine SECTM1 de 248 acides-aminés. De plus, SECTM1 se présente sous 2 variants d'épissage i) une protéine transmembranaire de type I (une seule hélice-α hydrophobe traverse la membrane) localisée dans l'appareil de Golgi et ii) une protéine sécrétée dans le milieu extracellulaire [Slentz-Kesler K.A. *et al*, 1998]. Au niveau structural, une faible similarité de séquence entre SECTM1 et le domaine immunoglobuline (Ig) a été rapportée [Slentz-Kesler K.A. *et al*, 1998]. Cependant, la structure 3D de SECTM1 n'a pas été résolue. Chez la souris, *Sectm1a* et *Sectm1b* codent pour des protéines SECTM1A et SECTM1B de 192 et 212 acides aminés, respectivement [Lyman SD *et al*, 2000; Howie D.

et al, 2013]. Au niveau fonctionnel, SECTM1 intéragit physiquement avec CD7, une protéine à domaine Ig exprimée par les lymphocytes T, NK et pre-B. Il est à noter que les gènes *CD7* et *Sectm1* sont voisins sur le génome humain et séparés par ~4kb. SECTM1 et CD7 modulent l'activité des lymphocytes T. Enfin, l'expression de *Sectm1* est stimulée par l'interféron-γ [Lam GK *et al*, 2005]. *Sectm1* est donc clairement impliqué dans des mécanismes immunitaires.

G.3.1.3. *Tex19* et *Sectm1* sont voisins sur le génome, uniques chez l'humain et dupliqués chez la souris et le rat

En phylogénèse moléculaire, il est parfois notable que le nombre de paralogues d'une famille de gènes soit moins élevé chez l'humain que chez d'autres espèces ce qui peut indiquer une adaptation à l'environnement. Par exemple, il existe 48 récepteurs nucléaires chez l'humain, 49 chez la souris et 270 chez *Caenorhabditis elegans* [Robinson –Rechavet M *et al*, 2001; Zhang Z. *et al*, 2004]. Dans la famille des récepteurs olfactifs, non seulement le répertoire des gènes chez l'humain est 20% inférieur à celui de la souris mais il est en plus extrêmement diversifié. En effet, 862 gènes de récepteurs olfactifs ont été dénombrés chez l'homme ce qui dénote une histoire évolutive très complexe [Gilad Y. *et al*, 2005]. Chez l'humain, le nombre de copies de *Tex19* et *Sectm1* est plus faible que chez la souris mais il est en plus réduit à son minimum, *i.e.* 1 copie chez l'humain et 2 copies chez la souris (Figure 11). Quelle fonction est assurée par un nombre de gènes si faible ? Une recherche d'orthologues sur le génome du rat a montré que le répertoire des 2 gènes est également de 2 copies. A ma connaissance, un dénombrement identique a été rapporté pour le gène de l'insuline qui existe en 1 copie chez l'humain (*Ins*) et 2 copies chez la souris (*Ins1* et *Ins2*) [Shiao M.S. *et al*, 2008]. Etant donné l'importance de l'insuline pour les mammifères, un dénombrement faible de gènes n'indique pas une fonction mineure. Des hypothèses ont été avancées pour expliquer la paralogie d'*Ins* chez la souris, comme une adaptation au régime alimentaire [Shiao M.S. *et al*, 2008]. Chez le rat, *Tex19* et *Sectm1* sont également dupliqués et se localisent sur une région synténique de la cytobande 10q32.3. Sur le génome des 3 espèces, *Tex19* et *Sectm1* sont encadrés en 5' par le gène *CD7* qui code pour un domaine Ig et en 3' par *Uts2r* (Récepteur 2 à l'urotensine). *CD7* et *Uts2r* ne sont pas dupliqués chez la souris et le rat, du moins dans cette région chromosomique. *Tex19* humain est orienté dans le même sens que ses orthologues murins *Tex19.1* alors que les gènes *Tex19.2* sont dans le sens opposé. Au contraire, la duplication de *Sectm1* chez la souris et le rat a conservé l'orientation des 2 paralogues dans le même sens que le gène humain.

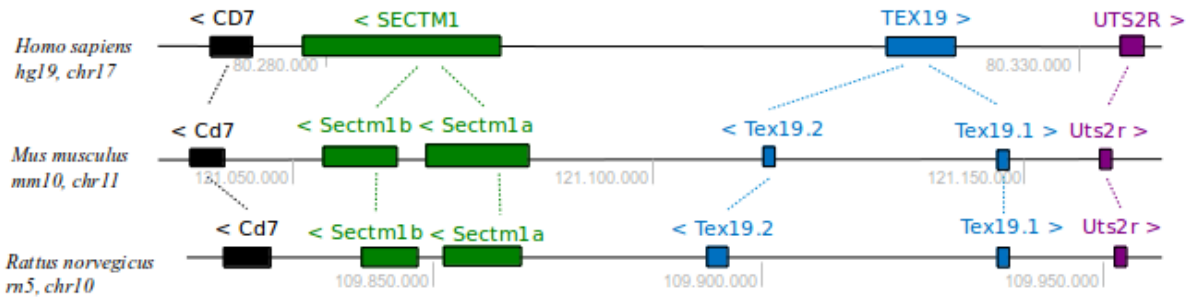


Figure 11: **Représentation schématique de la région chromosomique synténique des gènes *Tex19* et *Sectm1* chez les génomes humain, souris et rat**. Les versions des génomes sont GRCh37/hg19, GRCm38/mm10 et RGSC5.0/rn5 pour ces 3 espèces, respectivement. Les rectangles noirs, verts, bleus et violets représentent les gènes *CD7*, *Sectm1*, *Tex19* et *Uts2r*, respectivement. Les lignes en pointillés joignent les gènes orthologues. Les symboles « > » et « < » indiquent le sens de transcription des gènes. Les nombres écrits en gris indiquent la position sur le chromosome (en base).

G.3.1.4. *Tex19* et *Sectm1* ne sont pas homologues

Pour prévenir des erreurs d'analyse (noms synonymes de gènes, erreurs d'annotation des génomes), nous avons souhaité vérifier si *Tex19* et *Sectm1* appartiennent à la même famille de gènes. Pour cela, nous avons comparé par diagramme de points (« *dot plot* ») les séquences protéiques codées par les gènes *Tex19* et *Sectm1* chez l'humain et la souris. Comme contrôle, GRα et le récepteur aux minéralocorticoïdes (MR), *i.e.* 2 homologues de la super-famille des récepteurs nucléaires, ont été comparés. Le diagramme de point des domaines de liaison au ligand de GRα et MR produit de grandes diagonales qui signent l'homologie de séquence (Figure 12a). Au contraire, le diagramme de point de TEX19 et SECTM1 ne produit aucune diagonale (Figure 12b). Par conséquent, TEX19 et SECTM1 ne partagent aucune similarité de séquence et ne sont donc pas homologues. TEX19 humain présente une homologie avec son orthologue TEX19.1 de souris (Figure 12c). De même, SECTM1 humain est homologue à SECTM1A de souris (Figure 12d). Enfin, TEX19.1 et TEX19.2 d'une part (Figure 12e) et SECTM1A et SECTM1B d'autre part (Figure 12f) sont homologues bien que les séquences SECTM1A et SECTM1B aient relativement divergé (visible par diagonales au nombre de 2 et de tailles réduites).

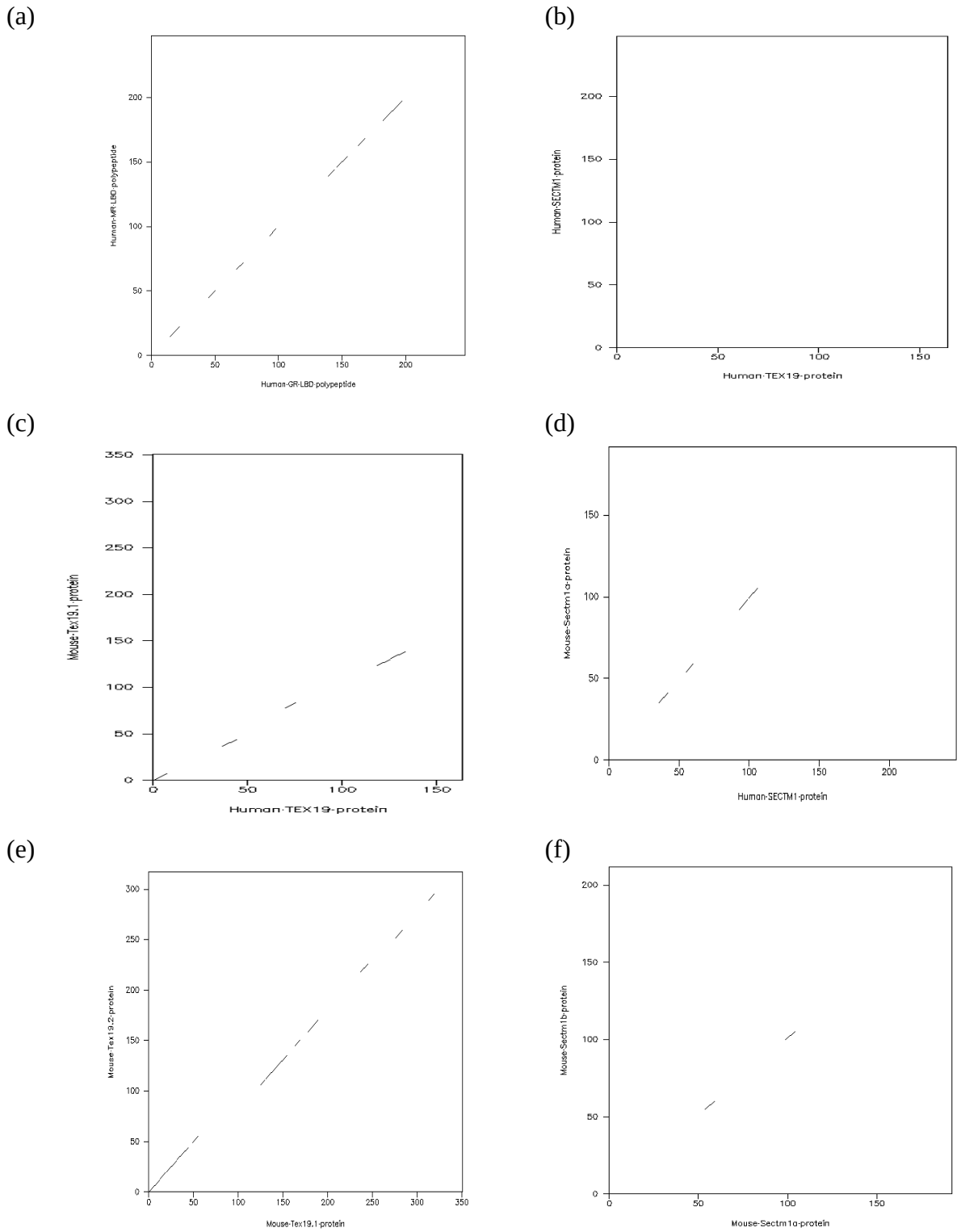


Figure 12: **Comparaison des séquences protéiques complètes de TEX19 et SECTM1 par diagramme de points.** Sur le diagramme, une diagonale de points représente un segment local d'acides-aminés identiques entre les 2 séquences. En abscisse et en ordonnée, les séquences polypeptidiques sont écrites de gauche à droite et de bas vers le haut de l'extrémité N- à C-terminale, respectivement (a) Contrôle : domaine de liaison au ligand des récepteurs nucléaires humains aux glucocorticoïdes (GR) (NP_000167, 531-777) et aux minéralocorticoïdes (MR) (NP_000892, 737-984). (b) Séquences humaines TEX19 (NP_997342) et SECTM1

(NP_002995). (c) TEX19 humain comparé à TEX19.1 de souris (NP_082878). (d) SECTM1 humain comparé à SECTM1A (NP_663348) de souris. (e) Comparaison de TEX19.1 et TEX19.2 (NP_081898) de souris. (f) Comparaison de SECTM1A et SECTM1B (NP_081183) de souris. Les chiffres sur les axes indiquent les positions des acides-aminés.

Pour affiner les comparaisons entre protéines orthologues chez l'humain et la souris, les pourcentages d'identité et de similarité des protéines ont été déterminés par algorithme global de Needleman et Wunsch (N&W) si les polypeptides étaient de longueurs sensiblement égales, ou par l'algorithme local de Smith et Waterman (S&W) dans le cas contraire (Tableau 7). TEX19.1 et TEX19.2 de souris sont sensiblement plus proches l'une de l'autre (56.1% d'identité et 64.9 % de similarité) qu'elles ne le sont de TEX19 humain. Le faible pourcentage d'identité de SECTM1A et SECTM1B de souris (34.6%) confirme la divergence observée sur le diagramme de point (Figure 12f).

	Algo.	**Identité %**	**Similarité %**	**Positions**	**Gaps %**	**Score**
TEX19 x TEX19.1	S&W	47	56.8	183	12.6	374
TEX19 x TEX19.2	S&W	53.8	62.7	158	8.9	423
TEX19.1 x TEX19.2	N&W	56.1	64.9	353	10.8	970
SECTM1 x SECTM1A	S&W	43.6	60	195	12.3	388
SECTM1 x SECTM1B	S&W	39.4	51.5	198	14.1	327
SECTM1A x SECTM1B	N&W	34.6	46.9	228	22.8	331

Tableau 7: **Comparaison 2 à 2 des séquences polypeptidiques TEX19 et SECTM1 chez l'humain et la souris**. TEX19 (164 aa) et SECTM1 (248 aa): protéines humaines. TEX19.1 (351 aa), TEX19.2 (317 aa), SECTM1A (192 aa) et SECTM1B (212 aa) : protéines de souris. S&W, algorithme local de Smith & Waterman. N&W algorithme global de Needlman et Wunsch. Matrice de score : BLOSUM62.

G.3.1.5. TEX19 et SECTM1 humains sont plus proches de TEX19.2 et SECTM1A de souris, respectivement

La souris est un modèle animal largement utilisé pour étudier des pathologies humaines et transposer chez l'homme des résultats obtenus sur des gènes de souris, à condition que les gènes orthologues de l'humain et de la souris soient suffisemment proches en séquence pour qu'ils partagent la même fonction, c'est l'orthologie 1:1. Pour *Tex19* comme pour *Sectm1*, la présence d'une seule copie chez l'humain et de 2 copies chez la souris pose immédiatement la question de savoir de quel orthologue murin, le gène humain est il le plus proche. Pour

répondre à cette question, nous avons utilisé BLASTP et appliqué la méthode de la meilleure touche (« *best hit* ») sur toutes les séquences de souris de la banque RefSeq. Par rapport à la comparaison de séquences 2 à 2 (voir G.3.1.4), cette approche introduit un indicateur supplémentaire, *i.e.* l'espérance, pour estimer la vraisemblance d'une homologie entre 2 séquences (Tableau 9). Plus l'espérance est petite, plus la ressemblance de séquence est forte et plus l'homologie est vraisemblable.

	Rang	**Couverture**	**Score**	**Espérance**
TEX19.2	1er	96%	148	6 x 10^{-44}
TEX19.1	2nd	95%	112	7 x 10^{-30}

Tableau 9: **Meilleure touche par BLASTP avec TEX19 humain comme séquence-requête sur toutes les protéines de souris de la banque RefSeq (NCBI)**. RefSeq: 58.976 séquences et 41.401.466 acides-aminés. La couverture indique le pourcentage de la séquence-requête qui s'aligne sur la séquence-cible.

Avec une espérance de 6 x 10^{-44}, TEX19.2 se pose en orthologue le plus proche de TEX19 humain alors que l'espérance de TEX19.1 est ~10^{14} fois plus grande. Pour une comparaison supplémentaire et sans équivoque par « *best hit* », nous avons utilisé la séquence de TEX19 du cochon d'Inde (*Cavia porcellus*) qui, nous le verrons, est codée par un gène unique sur son génome – comme chez l'humain - et possède 336 acides-aminés ce qui la rapproche de la longueur des séquences orthologues de souris (Tableau 10) alors que la séquence humaine, rappelons le, ne possède que 164 résidus. Avec la séquence de *Cavia porcellus* comme séquence-requête, l'espérance de TEX19.2 obtenue par BLASTP est ~3 x 10^{31} fois plus petite que celle obtenue par TEX19.1. Ce résultat consolide l'orthologie fonctionnelle entre TEX19.2 de souris et la protéine TEX19 d'une espèce dont le gène est unique sur son génome. Le protocole du *best hit* a été appliqué à SECTM1 (Tableau 11). SECTM1A obtient une espérance ~ 0.3 x 10^{17} fois plus petite que SECTM1B. SECTM1 humain est donc plus proche de SECTM1A.

	Rang	**Couverture**	**Score**	**Espérance**
TEX19.2	1er	100%	241	2 x 10^{-77}
TEX19.1	2nd	83%	160	6 x 10^{-46}

Tableau 10: **Meilleure touche par BLASTP avec le TEX19 de *Cavia porcellus* (336 acides-aminés) comme séquence-requête sur les protéines de souris de la banque RefSeq (NCBI)**.

RefSeq: 58.976 séquences et 41.401.466 acides-aminés. La couverture indique le pourcentage de la séquence- requête qui s'aligne sur la séquence-cible.

	Rang	**Couverture**	**Score**	**Expect**
SECTM1A	1er	70%	151	3×10^{-45}
SECTM1B	2nd	48%	108	1×10^{-28}

Tableau 11 : **Meilleure touche par BLASTP avec SECTM1 humain comme séquence-requête sur les protéines de souris de la banque RefSeq (NCBI)**. RefSeq: 58.976 séquences et 41.401.466 acides-aminés. La couverture indique le pourcentage de la séquence-requête qui s'aligne sur la séquence-cible.

Finalement, nous avons souhaité savoir si les structures des gènes *Tex19* et *Sectm1* permettent également de rapprocher les gènes humains de l'un des gènes de souris. Les coordonnées des gènes sur le chromosome ont été obtenues à partir du navigateur génomique UCSC et un script PERL a été écrit *ad hoc* pour calculer 6 longueurs globales d'éléments génétiques, *i.e.* i) 5' UTR (région d'exons non-traduits en 5' de la séquence codante (CDS) ii) longeur totale d'intron en 5' de la CDS iii) longueur de la séquence codante iv) longueur d'introns dans la CDS v) 3' UTR (région d'exons non-traduits en 3' de la CDS) et vi) longueur d'intron en 3' de la CDS (Figure 13). Sur la base des longueurs génétiques (Tableau 11), le coefficient de corrélation de Pearson entre gènes humains et souris a été calculé. La structure génétique de *Tex19* humain ne corrèle ni avec celle de *Tex19.1* (ρ = 0,3) ni avec celle de *Tex19.2* (ρ = 0,22) de souris. La divergence des structures génétiques est telle qu'elle ne permet pas de rapprocher le gène humain de l'un des 2 gènes de souris. En 5' du gène humain, un intron particulièrement long de 2,5 kb n'a pas d'équivalent chez les orthologues de souris. Il est à noter également que les structures des 2 gènes de souris sont fortement corrélées (ρ = 0,90) ce qui indique que les gènes ont peu divergé depuis la duplication. Pour *Sectm1*, la structure du gène humain corrèle plus avec *Sectm1a* (ρ = 0,93) que *Sectm1b* (ρ = 0,83) de souris ce qui indique que *Sectm1* humain serait plus proche de *Sectm1a* et confirme le résultat obtenu par comparaison de séquences protéiques. Inversement, les structures des gènes *Sectm1a et Sectm1b* corrèlent moins entre elles (ρ = 0,57) et montre que ceux-ci ont divergé depuis la duplication.

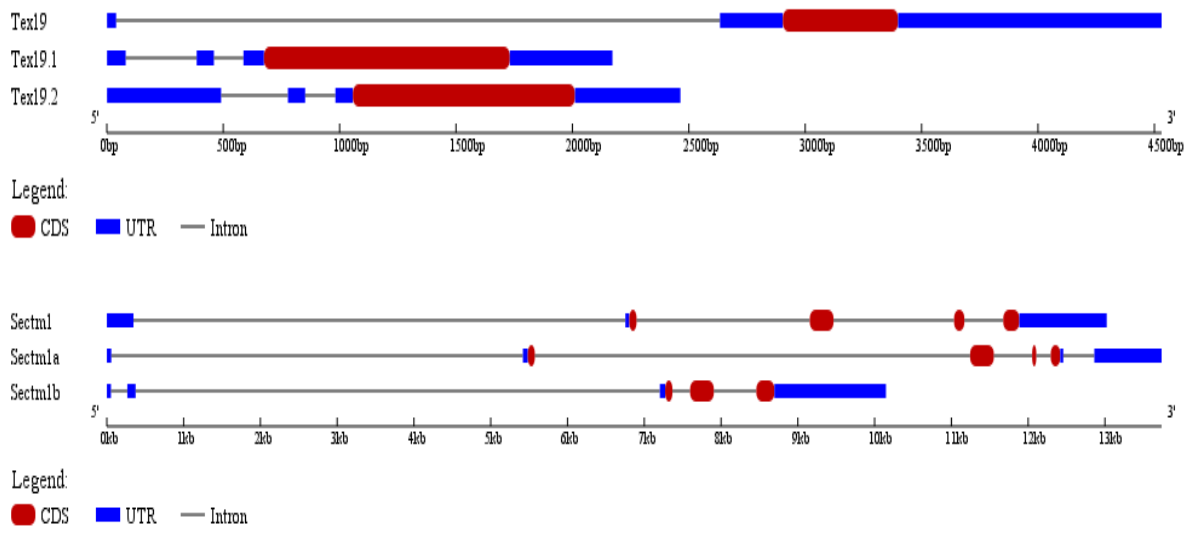


Figure 13: **Structure des gènes *Tex19* et *Sectm1* chez l'humain et la souris**. (a) Gènes *Tex19* (humain), *Tex19.1* et *Tex19.2* (souris). (b) Gènes *Sectm1* (humain), *Sectm1a* et *Sectm1b* (souris). CDS : région codante, UTR : région non traduite.

Elément génétique	***Tex19***	***Tex19.1***	***Tex19.2***	***Sectm1***	***Sectm1a***	***Sectm1b***
Exons en 5' UTR	0.31	0.243	0.640	0.398	0.122	0.231
Exons en 3' UTR	1.131	0.440	0.452	1.137	0.914	1.452
Exons de la CDS	0.495	1.056	0.954	0.747	0.579	0.639
Introns en 5' de la CDS	2.594	0.432	0.417	6.407	5.361	7.041
Introns en 3' de la CDS	0	0	0	0	0.405	0
Introns dans la CDS	0	0	0	4.333	6.356	0.784

Tableau 11 : **Longueurs des éléments génétiques des gènes *Tex19* et *Sectm1* humains et souris.** Grandeurs en kilobase (kb). 5p UTR : région non-codante en 5'. 3p UTR : région non-codante en 3'. CDS, région codante.

G.3.1.6. Scénarios évolutifs de duplication des 2 gènes

Deux scénarios évolutifs sont concevables: i) l'ancêtre commun de l'humain, de la souris et du rat ne possédait qu'une seule copie de *Tex19* et *Sectm1* et un évènement a dupliqué les 2 gènes chez l'ancêtre commun des rongeurs (Figure 14a) ou bien ii) l'ancêtre commun aux 3 espèces possédait 2 copies de chaque gène et 1 copie a été perdue chez l'humain (Figure 14b)**.** Le dénombrement singulier de ces 2 gènes a piqué ma curiosité et m'a motivé à construire les phylogénèses moléculaires des protéines codées par les 2 gènes pour savoir quel scénario évolutif serait correct.

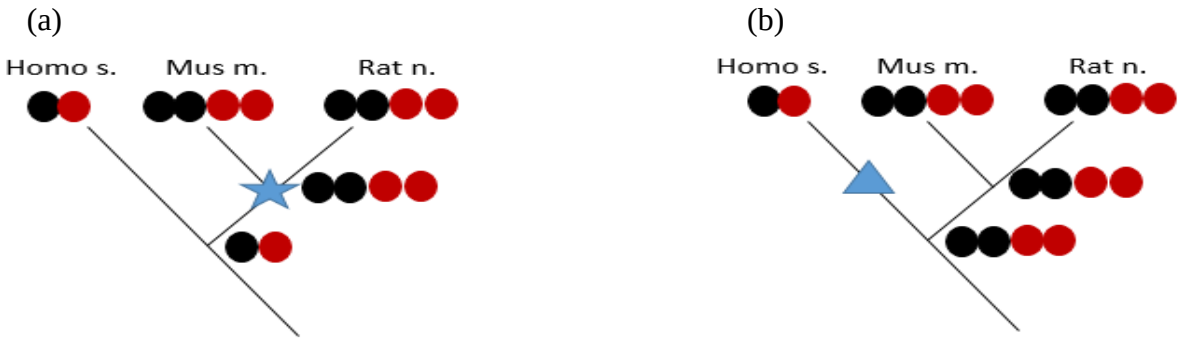


Figure 14: **Représentation schématique des scénarios évolutifs possibles pour la phylogénèse moléculaire des gènes *Tex19* et *Sectm1* chez l'humain, la souris et le rat.** Chaque disque correspond à une copie de gène. Les disques noirs et rouges représentent *Tex19* et *Sectm1*, respectivement. (a) Scénario à évènement de duplication chez l'ancêtre des rongeurs. L'étoile indique la duplication. (b) Scénario à évènement de délétion chez l'humain. Le triangle indique la délétion.

G.3.1.7. *Tex19* code pour une protéine orpheline

Afin d'obtenir des informations par homologie sur le rôle moléculaire de TEX19, nous avons mené une recherche de similarité de séquence dans les banques protéiques avec BLASTP et les paramètres par défaut. En dehors des membres de sa famille, TEX19 ne présente de similarité de séquence à aucune autre protéine connue. La structure de TEX19 n'étant pas résolue, nous avons recherché des homologies distantes avec l'outil PSI-BLAST. Au bout de 10 itérations, aucune ressemblance significative n'a pu être déterminée. Cette absence d'homologie inscrit TEX19 dans la catégorie des protéines orphelines [Tautz D. & Domazet-Loso T., 2011] et pose la question de son origine, de sa structure – il pourrait s'agir d'un nouveau repliement – et de sa fonction moléculaire.

G.3.1.8. Arguments en faveur de la coévolution des 2 gènes

G.3.1.8.1. *Tex19* et *Sectm1* sont euthériens-spécifiques

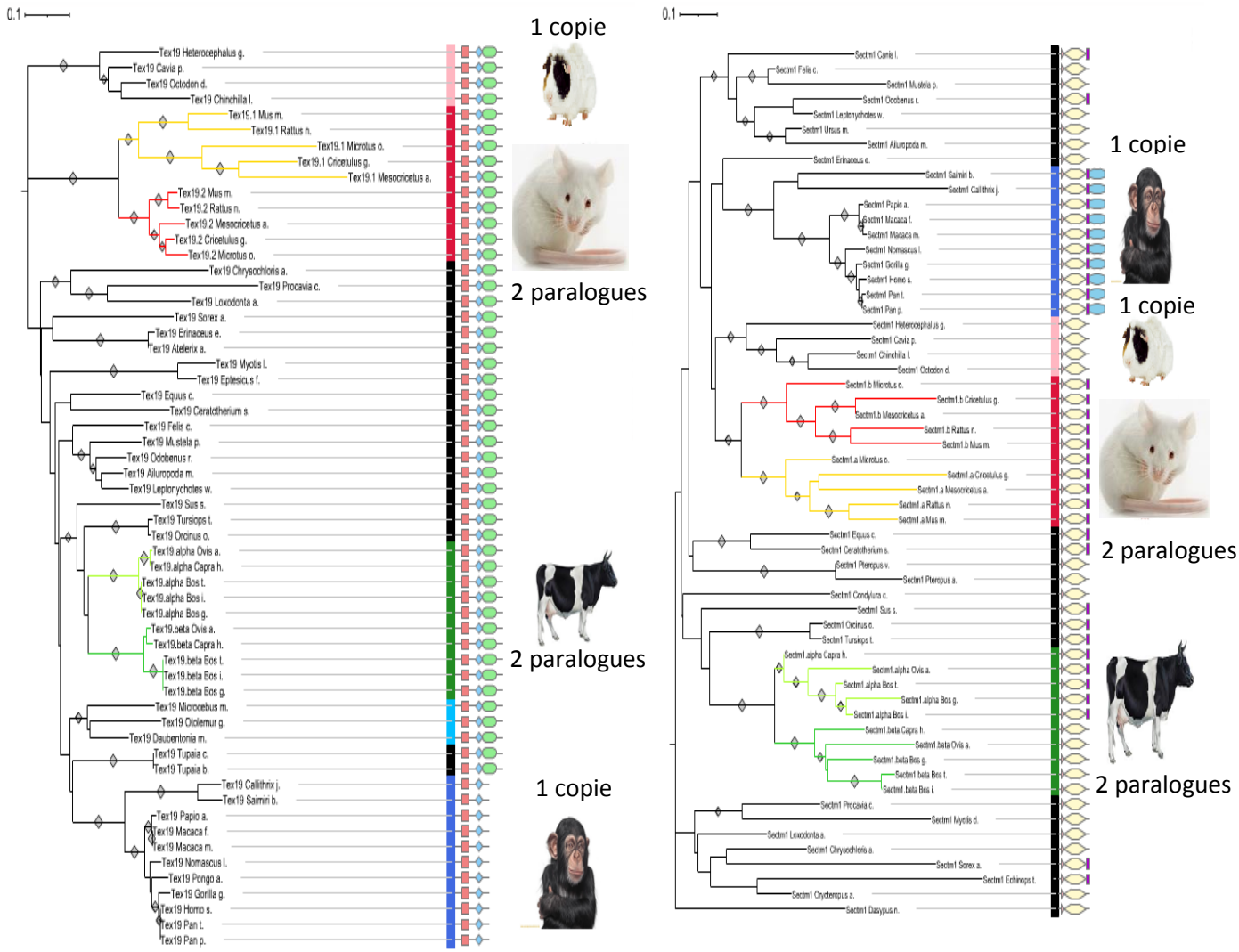
Pour retracer l'histoire évolutive de TEX19 et SECTM1, 58 homologues ont été collectés exhaustivement dans les banques pour chacune des 2 protéines et 2 alignements multiples de séquences complètes (MACS) ont été construits. Les recherches de similarité de séquences ont montré que les 2 gènes sont spécifiques des mammifères placentaires (Euthériens). En effet, ils n'existent pas chez les procaryotes, les plantes, les invertébrés, les

poissons, les reptiles, les oiseaux, l'ornithorynque (Monotrème) et les marsupiaux (Métathériens). Au contraire, ils ont pu être identifiés dans les 4 super-ordres de mammifères placentaires (*Euarchontoglires*, *Laurasiatheriens*, *Afrotheriens* et *Xenarthres*).

G.3.1.8.2. Concordance des arbres phylogénétiques

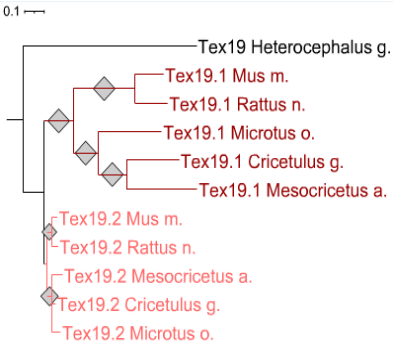
Les arbres phylogénétiques des protéines ont été construits en utilisant l'algorithme du proche voisin (« *Neighbor joining* ») et le modèle de substitution de résidus le plus adapté pour nos protéines, c'est-à-dire le Jones-Taylor-Thornton (JTT) selon l'outil ProtTest [Abascal F. *et al*, 2005 ; Darriba D. *et al*, 2011]. Ces 2 arbres montrent une concordance remarquable de topologie (Figure 15a et b).

(a) (b)

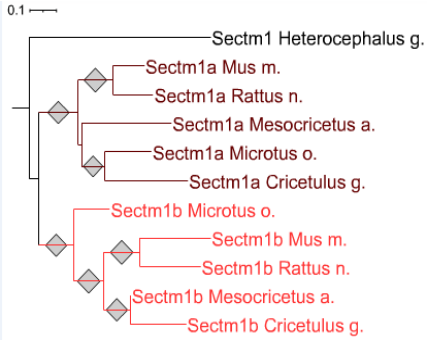


(*La figure continue sur la page suivante*)

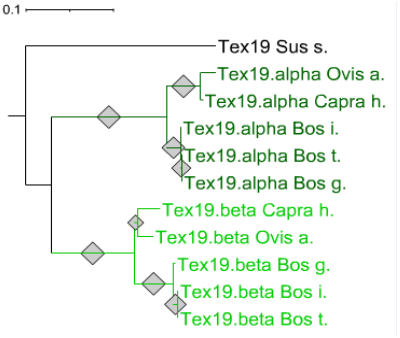
(c)



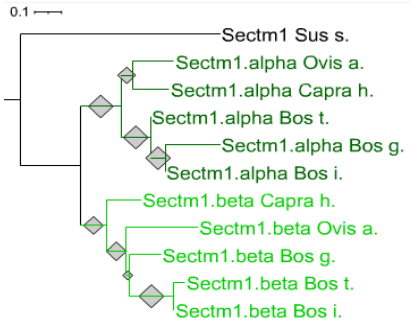
(d)



(e)



(f)



(g)

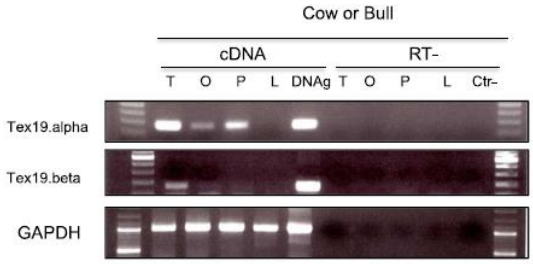


Figure 15: **Arbres phylogénétiques des protéines TEX19 et SECTM1**. Les phylogrammes ont été construits par l'algorithme du proche voisin (« *Neighbor joining* »). Les valeurs de bootstrap (test de robustesse des nœuds) supérieures à 75% sont symbolisées par des losanges gris (plus cette valeur est grande plus la topologie de l'arbre est fiable). Les bandes verticales bleues foncées, bleues claires, vertes, rouges sombres et rouges claires représentent les groupes taxonomiques des *Haplorrhini (primates non-lémuriens)*, *Strepsirrhini (primates lémuriens)*, *Bovidae*, *Sciurognathi (rongeur, e.g. souris)* et *Hystricognathi (rongeur, e.g. cochon d'Inde)*, respectivement. Les bandes verticales noires représentent n'importe quel autre *Eutheria (mammifère placentaire)*. Pour chaque séquence de l'arbre, l'organisation de la protéine en régions est schématisée (voir G.3.1.7.3 pour plus de détails). L'image de chimpanzé indique 1 seule copie de gène chez les primates, celle de la vache indique 2 copies (paralogues) chez les *Bovidae*, celle de la souris indique 2 copies chez les *Sciurognathi* et enfin celle du cochon d'Inde signifie 1 copie chez les *Hystricognathi*. (a) Phylogramme de TEX19. Les paralogues TEX19.1 et TEX19.2 chez les *Sciurognathi* sont représentés par des branches jaunes et rouges, respectivement. Les paralogues TEX19.alpha et TEX19.beta chez les *Bovidae* sont montrés avec des branches vertes claires et foncées, respectivement. (b) Phylogramme de SECTM1. Les paralogues SECTM1A et SECTM1B chez les *Sciurognathi* sont indiqués par des branches jaunes et rouges, respectivement. Les paralogues Sectm1.alpha et Sectm1.beta chez les *Bovidae*

sont coloriés en vertes claires et foncées, respectivement. Sous-arbre des séquences TEX19 (c) et SECTM1 (d) chez les *Sciurognathi*. Sous-arbre des séquences TEX19 (e) et SECTM1 (f) chez les *Bovidae*. *Heterocephalus glaber* et *Sus scrofa* ont été utilisés pour enraciner les sous-arbres des *Sciurognathi* et *des Bovidae*, respectivement). (g) RT-PCR des ARNm *Tex19.alpha* et *Tex19.beta* dans différents tissus chez la vache/taureau T:testicule, O: ovaire, P: placenta, L: foie. DNAg: ADN génomique.

Les 2 gènes se sont dupliqués chez les rongeurs du sous-ordre des *Sciurognathi* et la famille des *Bovidae* alors qu'ils ont été maintenus en 1 seule copie chez toutes les autres espèces de mammifères placentaires (Figure 15c, d, e et f). Par conséquent, l'humain n'a jamais possédé 2 gènes *Tex19* mais 1 seul (cf Figure 14a). Il est remarquable que deux évènements indépendants de duplication se soient produits, l'un chez les *Sciurognathi* et l'autre chez les *Bovidae.* Chez les *Bovidae*, nous avons nommé les 2 paralogues de *Tex19* et *Sectm1*, alpha et beta. Signalons que l'expression des paralogues *Tex19* de *Bovidae* a été confirmée expérimentalement par RT-PCR chez la vache et le taureau (Figure 15g). La concordance topologique des 2 phylogrammes suggère une co-évolution.

Pour démontrer la co-évolution, 3 méthodes ont été utilisées. La 1ère a consisté à sélectionner 8 espèces dans les 4 super-ordres d'euthériens et à dénombrer le nombre de copies des 2 gènes chez ces espèces. Une table de contingence a été construite et un test statistique du $\chi 2$ de Pearson a été mené. Une p-valeur de 0.018 a permis de rejeter l'hypothèse d'indépendance du nombre de copies des 2 gènes et donc de conclure à leur liaison. En s'appuyant sur les alignements multiples, la 2ième méthode calcule des distances (voir Ressources & Méthodes G.2.2.1.10**)** entre les séquences d'espèces co-présentes dans les 2 MACS et une séquence choisie comme référence (TEX19 et SECTM1 humain) [Goh C-S. *et al*, 2000 ; Pazos F & Valencia A, 2008] (Tableau 12). Entre les séquences des alignements de TEX19 ou SECTM1 et les séquences de référence humaines, un coefficiant de Pearson de 0.78 a été obtenu ce qui tend à montrer que les protéines codées par *Tex19* et *Sectm1* chez différentes espèces divergent des séquences humaines à des taux de substitutions corrélés.

	P.t.	*M. mu.*	*M. m.* 19.1/A	*M. m.* 19.2/B	*H. g.*	*E. c.*	*S. s.*	*L. a.*	*C.p.*
TEX19	0	0,055	0,487	0,433	0,422	0,366	0,381	0,433	0,388
SECTM1	0,032	0,181	0,588	0,668	0,544	0,602	0,619	0,478	0,5

Tableau 12: **Extrait des distances observées entre séquences de TEX19 ou SECTM1 chez différentes espèces et les orthologues de référence chez l'humain.** Ces distances ont été calculées pour 48 paires de séquences et 42 organismes. Pour la souris, les appariemments de TEX19.1 et SECTM1A d'une part et TEX19.2 et SECTM1B d'autre part sont arbritraires. *P.t. Pan troglodytes, M.mu. Macaca mulatta, M.m. Mus musculus, H.g. Heterocephalus glaber, E.c. Equus caballus, S.c. Sus scrofa, L.a. Loxodonta africana, C.p. Cavia porcellus.*

Enfin, la 3ième méthode est basée sur la comparaison statistique globale des 2 arbres phylogénétiques. Les phylogrammes ont été soumis au serveur MirrorTree [Ochoa D. & Pazos F., 2010] qui a calculé une corrélation de 0.717 et une p-valeur inférieure à 10^{-6} ce qui soutient la relation de dépendance entre les 2 histoires évolutives et donc la coévolution.

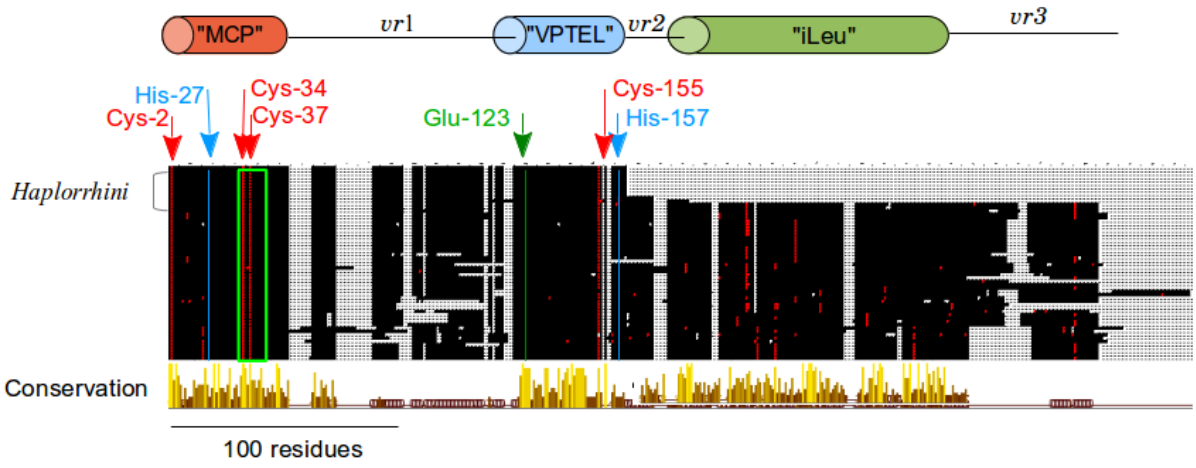
G.3.1.8.3. Insertion/délition d'une région protéique en C-ter.

Au niveau de l'architecture protéique, TEX19 est structurée en 3 régions conservées « MCP », « VPTEL » et « iLeu » (Figure 16a); les 2 premières régions sont séparées par une séquence variable. De plus TEX19 possède 4 cystéines invariantes dont 2 constituent un motif $C_{34}F[AT]CF_{38}$ (Figure 16b) présent dans la région N-terminale qui est la plus conservée. Une prédiction de structures secondaires par PSIPRED indique la présence de 2 hélices-α dans cette région (Figure 16c). Les cystéines invariantes de TEX19 sont intriguantes car la protéine est cytoplasmique. En effet, les cystéines engagées dans des ponts disulfures sont conservées et ne se forment pas dans le cytoplasme car le milieu est réducteur [Bosnjak I. *et al*, 2014]. Les 4 cystéines invariantes de TEX19 auraient donc un rôle différent, par exemple elles pourraient être impliquées dans la coordination d'un ion zinc. Selon l'outil PSIPRED, les cystéines du motif $C_{34}F[AT]CF_{38}$ seraient intégrées dans une hélice-α. La comparaison de la région N-terminale de TEX19 à la banque de domaine CDD (« *conserved domain database* ») [Marchler-Bauer A. *et al*, 2010] et l'utilisation de PSI-BLAST n'ont cependant détecté aucune homologie à des protéines connues. Il pourrait donc s'agir d'un nouveau repliement structural. Enfin, la région C-terminale « iLeu » de TEX19, riche en isoleucine, est perdue chez les primates *Haplorrhini* (non-lémuriens) alors qu'elle est présente chez toutes les autres séquences de mammifères placentaires.

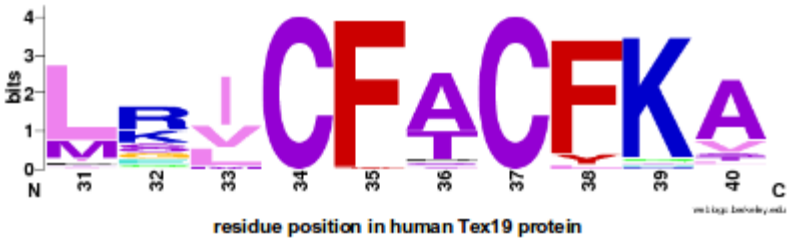
L'alignement multiple des protéines SECTM1 révèle une grande région conservée de ~100 résidus (région A), une séquence signal de sécrétion en extrémité N-terminale et une hélice-α transmembranaire en C-terminus (Figure 16d). Deux cystéines invariantes Cys38 et Cys55 sont présentes en N-terminus. Une signature très conservée a pu être mise en évidence dans la région A, *i.e.* $G_{113}xYxWxLxGxQ_{123}$ (Figure 16e) mais aucune fonction n'a pu lui être assignée. Une prédiction de structures secondaires a été réalisée avec PSIPRED (voir G.3.1.9, Figure 18b). Enfin, les primates *Haplorrhini* acquièrent une région C-terminale supplémentaire (région B). L'insertion/délétion spécifique d'une région en extrémité C-terminale de TEX19 et

SECTM1 chez les *Haplorrhini* pourrait constituer un argument supplémentaire qui signe l'inter-dépendance des 2 protéines.

(a)



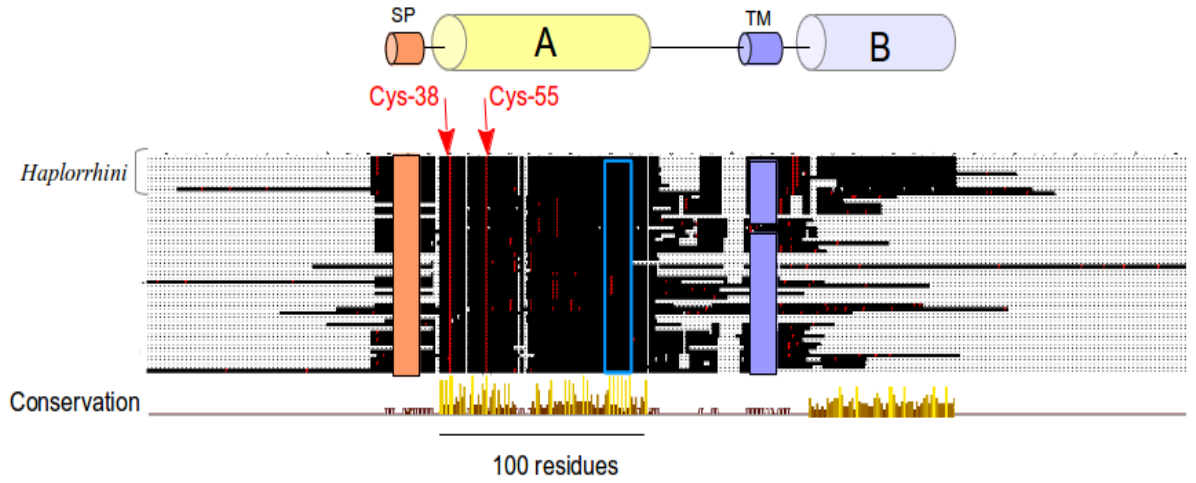
(b)



(c)

```
Conf: 988711224336725788999898763685124677458988988986500799
PSIP: CCCCCCCCCCCCCHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHCCC
Prot: MCPPVSMRYEEEGMSYLYASWMYQLQHGDQLSICFTCFKAAFLDFKDLLESEDW
      |         |         |         |         |         |
      1         10        20        30        40        50
```

(d)



(*La figure continue sur la page suivante*)

(e)

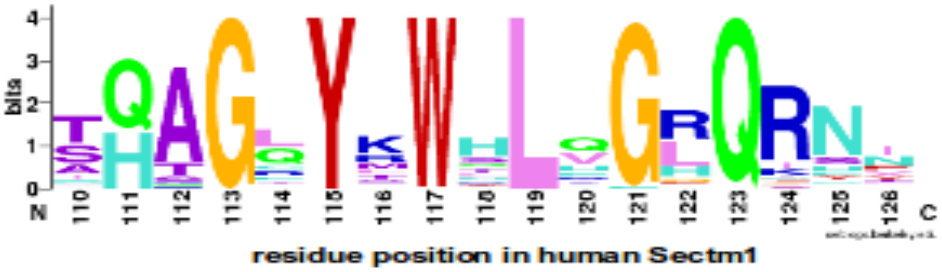


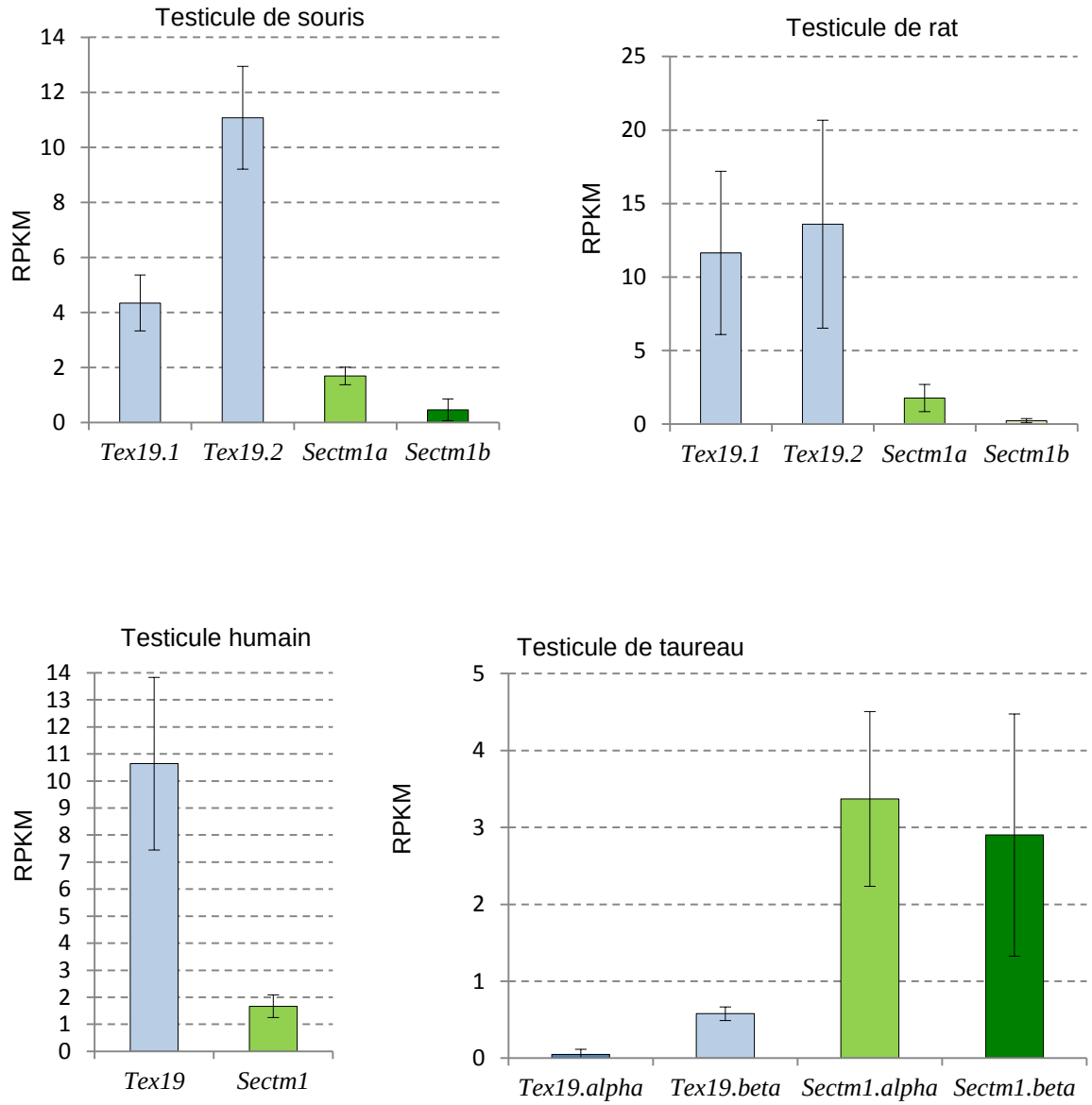
Figure 16: **Représentations schématiques des alignements multiples et de l'organisation en région des séquences protéiques complètes TEX19 et SECTM1 chez les mammifères placentaires**. Les séquences des primates *Haplorrhini* sont regroupées au sommet des alignements. En pointillés sont représentés les « *gaps* ». Les cystéines sont indiquées par des points rouges alors que tout autre résidu – sauf indication contraire - est marqué en noir. En bas des alignements, les histogrammes montrent les scores de conservation des résidus (a). TEX19: les cylindres "MCP", "VPTEL" et "iLeu" représentent des régions conservées tandis que les lignes "*vr*" sont des régions variables. Les flèches indiquent des résidus invariants et les indices se réfèrent à la sequence humaine RefSeq:NP_997342. Le cadre vert marque la position du motif (b) C_{34}F[AT]C_{37}[FY]. (c) Prédiction de structures secondaires chez TEX19. H: hélice-α, E: brin-β, C:boucle. Conf.: niveau de confiance, PSIP: prédiction PSIPRED, Mode: modèle 3D, Prot.: protein sequence. Le rectangle noir montre le motif conservé C_{34}F[AT]C_{37}[FY]$_3$. Symboles C en rouge: cysteine invariante. (d) SECTM1 : les cylindres et rectangles SP et TM dénotent un peptide signal de sécrétion et une hélice-α transmembranaire, respectivement. Le domaine A présente une similarité faible de séquence avec le domaine Ig. Le cylindre violet indique une région spécifique des primates *Haplorrhini*. Le cadre bleu indique la présence d'un motif (e) G_{113}xYxWxLxGxQ_{123} fortement conservé. Les indices des résidus se réfèrent à la séquence RefSeq:NP_002995.

G.3.1.8.4. Anti-corrélation des niveaux d'expressions

La disponibilité de données RNA-seq dans la banque SRA (« *sequence read archive* ») a permis de comparer les niveaux d'expression des gènes *Tex19* et *Sectm1* chez l'humain, la souris, le rat et la vache/taureau dans le testicule (Figure 17a) et le placenta (Figure 17b). Les courtes séquences RNA-seq (« *reads* ») spécifiques des gènes *Tex19* et *Sectm1* ont été extraites des expériences transcriptomiques par MegaBLAST en utilisant les ARN messagers –régions spécifiques de chaques homologues - comme séquences-requêtes et les niveaux d'expression ont été normalisés par calcul de RPKM (« *Reads per kilobase of exon and million of reads* »). Les niveaux d'expression de *Tex19* et *Sectm1* chez les différentes espèces sont hétérogènes. Dans le testicule humain et rongeur, *Tex19* s'exprime plus que *Sectm1* alors que chez le taureau, c'est le contraire. Pour le placenta, seules des données humaines et souris étaient disponibles (Figure 17b). Chez le placenta humain, *Sectm1* s'expriment fortement alors

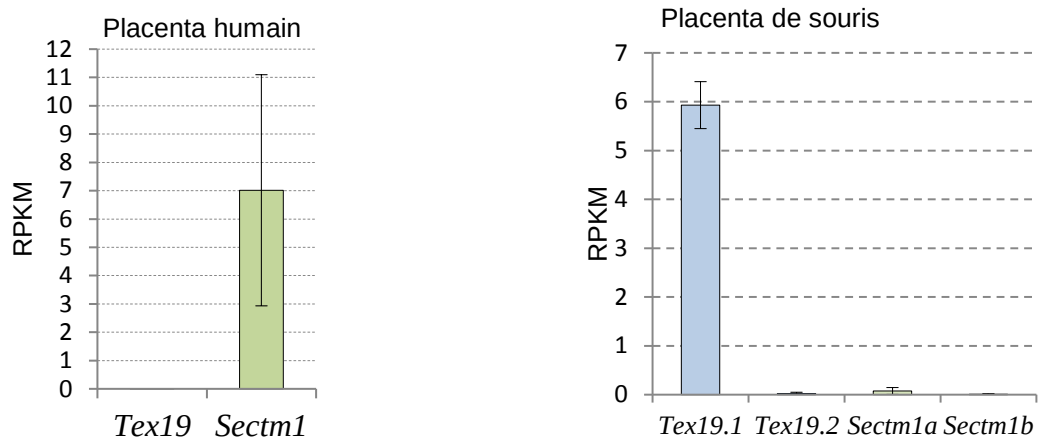
que *Tex19* est silencieux. Chez le placenta de souris, c'est le contraire, les copies de *Sectm1* sont silencieuses alors que *Tex19.1* s'expriment fortement (*Tex19.2* est silencieux). Enfin, en comparant les niveaux d'expression de *Sectm1* et *Tex19* dans le testicule chez le macaque, le gorille, le chimpanzé (*Pan troglodytes*), le bonobo (*Pan paniscus*) et l'humain, une anti-corrélation entre les niveaux d'expression des 2 gènes de -0.72 (Coefficient de Pearson) a pu être démontrée (Figure 17c).

(a)



(*La figure continue sur la page suivante*)

(b)



(c)

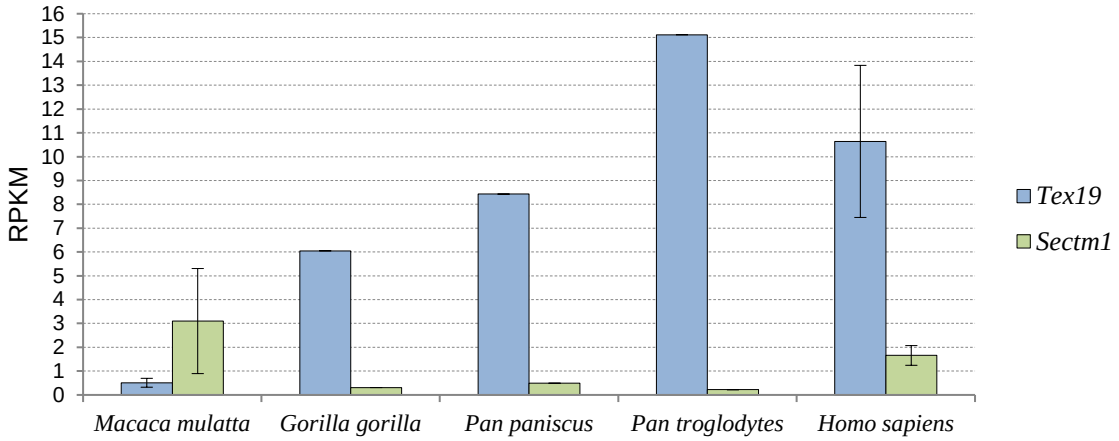


Figure 17: **Niveaux d'expression des ARN messagers de *Tex19* et *Sectm1* chez différentes espèces et tissus par RNA-seq**. Normalisation des séquences (reads) par kilobase et million (RPKM). (a) Testicule, humain (n=3), souris (n=4), rat (n=3) et taureau (n=3). (b) Placenta, humain (n=6), souris (n=3). (c) Testicule de primates, humain (n=3), *Pan troglodytes* (n=1), *Pan paniscus* (n=1), *Gorilla gorilla* (n=1), *Macaca mulatta* (n=4)

G.3.1.9. Homologie distante entre SECTM1 et le domaine Ig

Dans une étude antérieure, une similarité de séquence entre SECTM1 et le domaine Ig obtenue par BLAST avait été rapportée [Slentz-Kesler K.A. *et al*, 1998]. Entre nos mains, ce résultat n'a pas pu être reproduit (voir Discussion G.3.1.12.2). Grâce à une recherche itérative de similarité de séquence dans les banques avec l'outil PSI-BLAST, une ressemblance au domaine Ig a pu être détectée dès la 3^{ième} itération (E-value=2 x 10^{-11}) (Figure 18a). Au

niveau séquence primaire, la similarité entre SECTM1 et le domaine Ig de la protéine CD79A peut sembler faible, mais l'espérance rapportée par PSI-BLAST est significative (2×10^{-11}) et indique une homologie. De plus, une prédiction de structures secondaires par PSIPRED dans la séquence de SECTM1 a révélé une abondance de brins beta qui pourrait récapituler le repliement du domaine Ig (Figure 18b). Sur la base du résultat PSI-BLAST et la prédiction de structure secondaires, la construction d'un modèle 3D du domaine A de SECTM1 par homologie avec le domaine Ig nous a semblé pertinente (Figure 18c et d). Ce modèle a été validé par les outils ProSA [Wiederstein M. & Sippl M.J., 2007] et MolProbity [Chen V.B. *et al*, 2010]. Il est à noter que les cystéines invariantes Cys38 et Cys55 sont localisées dans le domaine Ig putatif. De façon inattendue, le pont disulfure caractéristique des domaines Ig est atypique dans le modèle de SECTM1. Au lieu de traverser le sandwich de feuillets beta (Figure 18c) comme c'est le cas dans le domaine Ig classique, le pont disulfure de SECTM1 s'établirait entre 2 cystéines du même feuillet (Figure 18d). Un pont disulfure similaire a cependant été obervé également dans le domaine Ig de la protéine CD4 qui sert de récepteur d'entrée dans les lymphocytes au virus de l'immunodéficience aquise (VIH) (Figure 18e) [Wang J. *et al*, 1990]. Enfin, le motif conservé $G_{113}xYxWxLxGxQ_{123}$ traverse le sandwich de feuillets longitudinalement (Figure 18f) ce qui pourrait servir à la transduction d'un signal.

(a)

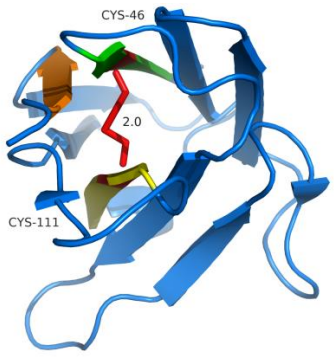
```
                                                    Cys-38
SECTM1   1    MQTCPLAFPGHVSQALGTLLFLAASLSAQNEGWDSPICTEGVVSVSWGENTVM 53
CD79A    1    MPGGPGVLQALPATIFLLFLLSAVYLG--------PGCQALWMHKVPASLMVSLGEDAHF 52

                Cys-55
SECTM1   54   SC--NISNAFSHVNIK--LRAHGQESAIFNEVAPGYFSRDGWQLQVQGGVAQLVIKGARD 109
CD79A    53   QCPHNSSNN-ANVTWWRVLHGNYTWPPEFL--GPGEDPNG-----------TLIIQNVNK 98
                Cys-54
SECTM1   110  SHAGLYMWHLVGHQRNNRQ---VTLEVSGAEPQSAPDTGFWPVPAVVTA--VFILLVALV 164
CD79A    99   SHGGIYVCRVQEGNESYQQSCGTYLRVRQPPPRPFLDMGEGTKNRIITAEGIILLFCAVV 158
                      Cys-106

SECTM1   165  MFAWYRCRCSQQRREKKFFLLE  186
CD79A    159  PGTLLLFR--KRWQNEKLGLDA  178
```

(*La figure continue sur la page suivante*)

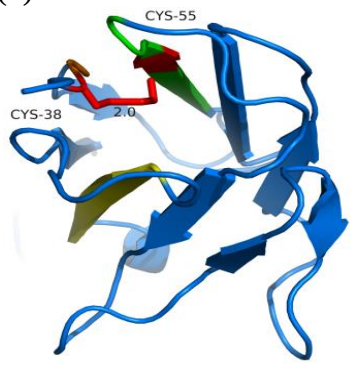
(b)

```
Conf: 9878898851357999999987430233457878985454248843499235884
PSIP: CCCCCCCCCCHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCEEEEECCCCEEEEE
Mode: ------------------------------------CCCCCCCCCEEECCCCCCEEEEE
Prot: MQTCPLAFPGHVSQALGTLLFLAASLSAQNEGWDSPICTEGVVSVSWGENTVMSC
      |         |         |         |         |         |
      1        10        20        30        40        50

Conf: 1146740269999407985124552599985889826999260879997142224
PSIP: ECCCCCCEEEEEEEECCCCCEEEEECCCCCCCCCCCEEEEEECCEEEEEECCCCCC
Mode: ECCCCCCCCCEEEECCCCCCCCEEEEECCCCCCCCCCCEEEEEECCEEEEEECCCCHH
Prot: NISNAFSHVNIKLRAHGQESAIFNEVAPGYFSRDGWQLQVQGGVAQLVIKGARDS
          |         |         |         |         |         |
         60        70        80        90       100       110

Conf: 532479896234301204899960898999999998777862122234556644 2
PSIP: CCEEEEEEECCEECCCCEEEEEECCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHH
Mode: HCEEEEEECCCCCCCCCCEEEEECCCCCCC-----------------------
Prot: HAGLYMWHLVGHQRNNRQVTLEVSGAEPQSAPDTGFWPVPAVVTAVFILLVALVM
              |         |         |         |         |
            120       130       140       150       160

Conf: 000234403555520334568741233411100036766872000259999988 9
PSIP: HHCCCCCCHHHHHHCCCCCCCCCCCHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCC
Mode: ------------------------------------------------------
Prot: FAWYRCRCSQQRREKKFFLLEPQMKVAALRAGAQQGLSRASAELWTPDSEPTPRP
          |         |         |         |         |         |
        170       180       190       200       210       220

Conf: 8542278876423346888889999999
PSIP: CCCCCCCCCCCCCCCCCCCCCCCCCCCC
Mode: ----------------------------
Prot: LALVFKPSPLGALELLSPQPLFPYAADP
              |         |
            230       240
```

(c)



(d)



(e)

(*La figure continue sur la page suivante*)

(f)

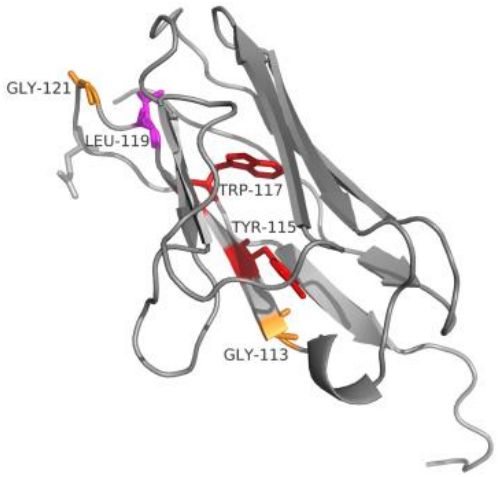


Figure 18: **Domaine Ig putatif de SECTM1**. (a) Alignement de séquences 2 à 2 produit par PSI-BLAST (3ième itération, Espérance = 2 x 10^{-11}, Score 70.1 bits) pour la comparaison SECTM1 et le domaine Ig de CD79A. Les cystéines invariantes sont sur fond rouge. Le pont disulfure du domaine Ig de CD79A est représenté en pointillés rouge. (b) Prédiction de structures secondaires chez SECTM1. H: hélice-α, E: brin-β, C:boucle. Conf.: niveau de confiance, PSIP: prédiction PSIPRED, Mode: modèle 3D, Prot.: séquence protéique. Le rectangle montre le motif conservé G_{113}xYxWxLxGxQ_{123}. Symboles C en rouge: cysteine invariante. Résidus sur fond jaune: domaine Ig putatif. (c) Représentation de la structure d'un domaine Ig classique (PDB:3SOB), le pont disulfure qui traverse le sandwich de feuillets-beta est représenté en rouge et sa longueur est indiquée en Å. Les brins beta mis en jeu pour la formation du pont disulfure dans le domaine Ig classique et le domaine Ig atypique de SECTM1 sont représentés en orange, vert et jaune. (d) Modélisation moléculaire par homologie du domaine Ig de SECTM1, le pont disulfure est prédit entre les brins beta d'une même face du sandwich. (e) Structure 3D du domaine Ig du CD4 humain obtenue par cristallographie (PDB:2NXY). Le pont disulfure atypique est indiqué en rouge. (f) Localisation du motif G_{113}xYxWxLxGxQ_{123} sur le modèle structural du domaine Ig de SECTM1.

G.3.1.10. *Sectm1* proviendrait d'une duplication du gène voisin *CD7*

Notre étude soutient que SECTM1 présente une homologie avec le domaine Ig. Il serait cependant expéditif de se satisfaire de cette information pour comprendre l'origine du gène. Pour résoudre cette question, une première recherche de similarité de séquence avec BLASTP et la séquence humaine SECTM1 comme requête a été menée sur les protéines codées par le génome humain. Cependant, cette approche n'a donné aucun résultat significatif. Une deuxième recherche de similarité de séquence a été menée sur les protéines codée par les génomes de métazoaires non-euthériens, c'est-à-dire des espèces animales chez qui *Sectm1* n'existe pas.

Parmi les 1ères séquences touchées par BLASTP dans la banque, le CD7 de *Phascolarctos cinereus* (Koala) (RefSeq:XP_020856521), *i.e.* une protéine à domaine Ig qui se lie à SECTM1 et qui est codée par le gène voisin sur le génome d'*Homo sapiens* (voir Figure 11), a produit un score de 42 pour une espérance de 0,06 (Figure 19). Sur 94 positions, la chaîne polypeptidique montre 29% d'identité et 48% de similarité. A 0,06, l'espérance est encore significative, il s'agirait donc d'une homologie. De façon frappante, SECTM1 et CD7 partagent le motif G_{113}xYxWxLxGxQ_{123} conservé chez SECTM1. Il peut donc être avancé que *Sectm1* serait apparu chez les euthériens par duplication du gène *CD7*. Cette information aurait pu être exploitée pour construire un modèle par homologie de SECTM1 en utilisant la structure 3D de CD7. Cependant, cette dernière n'a pas été résolue.

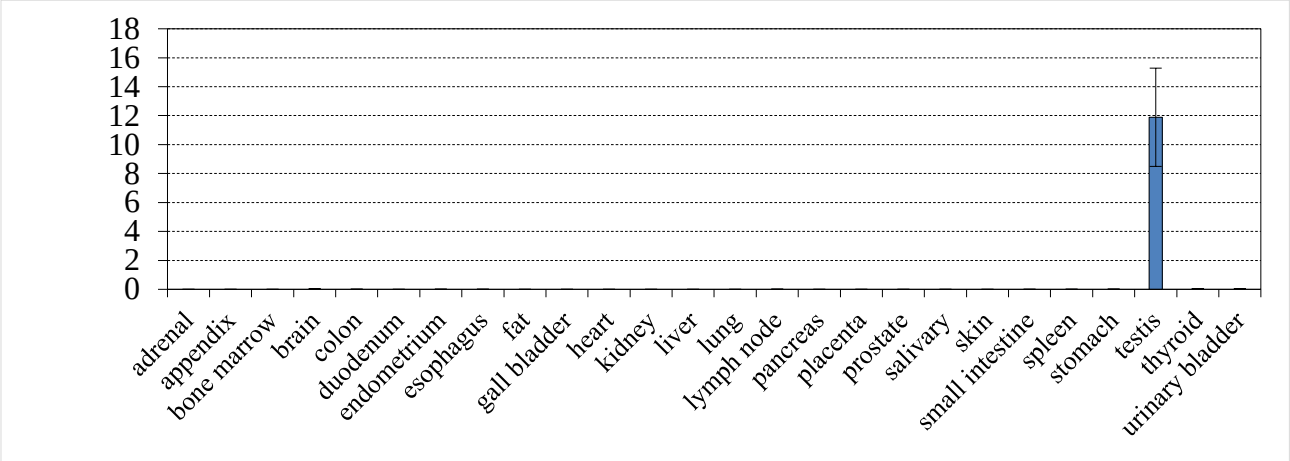
```
SECTM1  43  VSVSWGENTVMSCNISNAFSHVNIKLRAHGQ-ESAIFNEVAPGYFS-RDGWQLQVQGGVA 100
            ||    +   + |  +  |  +++  |+  + |  +|+    |    + | +|+ +   |
CD7     36  VSACRNDQVEIICKSTMDFDSIDLFFRSDKKDEELLFSMTTVGESHLQHGMRLRFKNQEA 95

SECTM1 101  QLVIKGARDSHAGLYMWHLVGHQRNNRQVTLEVS  134
             |||+  + || |+| ||| |   +|+  || ||
CD7     96  TLVIQNIQFSHCGVYRWHLSGQGSSNKFTTLTVS  129
```

Figure 19: **Comparaison de séquences polypeptidiques produite par BLASTP pour trouver l'origine de *Sectm1***. Séquence- requête: SECTM1 humain, séquence cible : CD7 du Koala (*Phascolarctos cinereus*). Motif sur fond vert *:* G_{113}xYxWxLxGxQ_{123} chez SECTM1 ; Motif sur fond jaune : polypeptide homologue chez CD7.

G.3.1.11. Spécificité « testicule/cancer » de *Tex19*

Récemment, Tex19 a attiré l'attention car il s'agirait d'un antigène « testicule/cancer » [Planells-Palop V. *et al*, 2017], c'est-à-dire qu'il s'exprime exclusivement dans le testicule chez l'adulte sain (Figure 20a) et en cellule cancéreuse (Figure 20b). Le testicule étant un tissu immuno-privilégié, des protéines antigéniques capablent de lever des réponses immunitaires dirigées contre elles s'y expriment [Fijak M. & MeinHardt A, 2006]. La spécificité d'expression testicule/cancer de *Tex19* fait de la protéine TEX19 une candidate idéale pour un vaccin cancer. Il n'est pas connu si les cellules cancéreuses exprimant TEX19 ont une activité anormale d'ETs et si l'expression de TEX19 procure un avantage au développement du cancer.

(a)



(b)

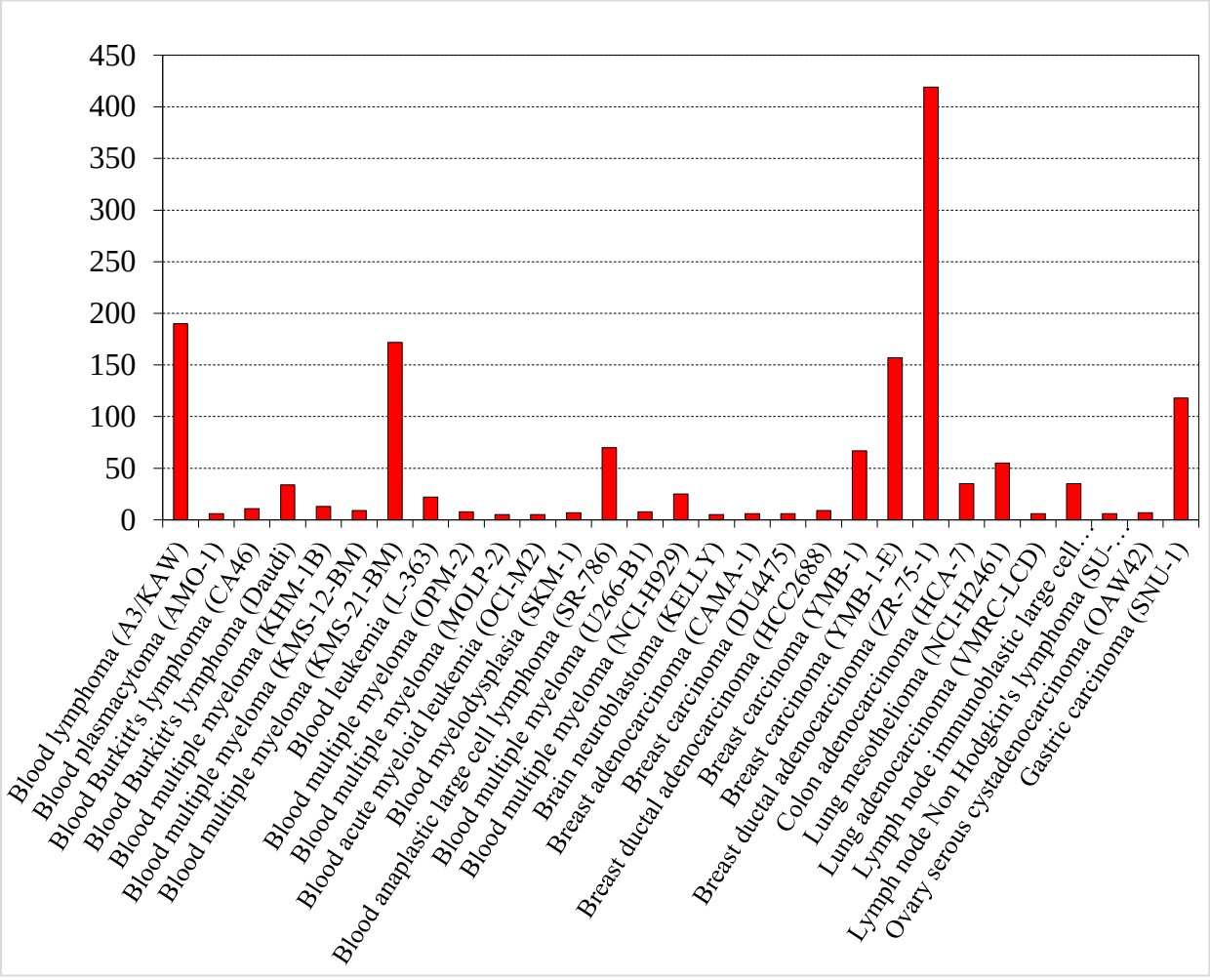


Figure 20: **Expression de l'ARNm de *Tex19* mesuré par RNA-seq chez l'humain**. Unité de l'axe vertical en RPKM. (a) Tissus sains chez l'adulte (Atlas d'expression http://ebi.ac.uk) (b) Lignées cellulaires cancéreuses (RNA-seq transcriptomes cancer Genentech)

G.3.1.12. Discussion

G.3.1.12.1. Pourquoi *Tex19* et *Sectm1* sont ils en coévolution ?

Quatre arguments soutiennent la coévolution de *Tex19* et *Sectm1* i) la spécificité des 2 gènes aux mammifères placentaires ii) la symétrie et la concordance statistique des arbres phylogénétiques iii) l'insertion/délétion simultanée de régions protéiques C-terminales chez les primates *Haplorrhini* et iv) l'anti-corrélation des niveaux d'expression dans le testicule des primates. Comprendre pourquoi *Tex19* et *Sectm1* sont en coévolution est donc une question brûlante. La coévolution est généralement observée pour des gènes codant des protéines impliquées dans un même processus biologique, par exemple une hormone et son récepteur ou une enzyme et son substrat polypeptidique. Or, *Tex19* et *Sectm1* sont impliqués dans des processus biologiques différents. *Tex19* participe au contrôle de l'activité des rétrotransposons alors que *Sectm1* est impliqué dans l'immunité. Pour réconcilier les 2 processus biologiques, il peut être avancé que *Sectm1* pourrait diriger une réponse immunitaire contre des cellules qui expriment des ETs alors que *Tex19* pourrait au contraire bloquer cette réponse dans des types cellulaires où une activité ET confèrerait un avantage sélectif. A noter, le testicule et le placenta sont décrits comme des tissus qui bénéficient d'un privilège immunitaire [Moffett A. *et al*, 2006 ; Fijak M. & Meinhardt A., 2006]. Des mécanismes empêchent le système immunitaire d'attaquer le testicule et l'embryon qui sont antigéniques.

G.3.1.12.2. Homologie distante entre SECTM1 et le domaine Ig

En génomique comparative, l'approche classique pour assigner une fonction à une nouvelle protéine obtenue par annotation d'un génome ou traduction d'un cDNA s'appuie sur l'homologie, c'est à dire une ressemblance de séquence si évidente qu'elle ne peut être dûe au hasard. C'est dans cet objectif que le programme BLAST a été développé [Altschul S.F. *et al*, 1990], c'est à dire trouver rapidement des homologies évidentes pour faciliter l'attribution de fonction à de nouvelles séquences et constituer des familles de protéines. Par annotation automatique ou expertise, la fonction est transmise – avec les précautions qu'il se doit selon le pourcentage d'identité - de la protéine de fonction connue à son homologue. Cependant, cette approche trouve sa limite quand le pourcentage d'identité devient faible, en pratique inférieur à 30%. Des homologues qui partagent moins de 30% d'identité de séquence existent néanmoins, il s'agit des homologue distants [Fariselli P. *et al*, 2006]. Les homologues distants présentent des similarités de séquences localisées aux résidus clés de la protéine, *e.g.* les résidus

catalytiques des enzymes, alors que le reste des séquences ont divergé. Si des outils tels que PSI-BLAST ont été développés pour détecter des homologies de séquences faibles [Altschul S.F. *et al*, 1997], la résolution structurale est souvent requise pour établir la preuve définitive de l'homologie. La structure étant plus conservée que la séquence, la superposition structurale de 2 homologues doit révéler une ressemblance si évidente qu'elle ne peut être dûe au hasard, y compris pour les homologues distants.

Dans le cours de notre projet *Tex19/Sectm1*, nous avons été confrontés à une propagation de fonction par homologie. *Sectm1* a été initialement cloné d'un cDNA isolé d'une lignée érythroleucémique K562 et code un polypeptide de 248 acides aminés [Slentz-Kesler K.A. *et al*, 1998]. Dans cette étude séminale, des recherches de similarité de séquences dans les banques par BLAST ont déterminé une similarité entre SECTM1 et le domaine Ig. Or, nous n'avons pas réussi à reproduire ce résultat. La recherche dans la banque Swissprot/TrEMBL n'a produit aucune similarité de séquence entre SECTM1 et le domaine Ig pour une espérance inférieure à 10. Au niveau séquence, il n'y a donc pas d'homologie évidente entre les 2 protéines. En 1998, date à laquelle la similarité de séquence entre SECTM1 et le domaine Ig a été rapportée, la banque SwissProt/TrEMBL contenait environ 1 million de séquences. Actuellement, cette banque en contient 140 millions. Or, plus une banque est grande, plus l'espérance d'atteindre des séquences similaires par hasard augmente. Il est donc vraisemblable, qu'en 1998 le domaine Ig produisait une espérance inférieure à 10 (seuil limite) par BLAST car la banque était petite. Quand nous avons mené la recherche de similarité de séquence avec SECTM1 sur la banque de 140 millions de séquences, il est probable que l'espérance produite par le domaine Ig ait été supérieure à 10, et donc BLAST l'a rejetée. Nous avons mené une recherche d'homologie distante avec PSI-BLAST. Cet outil réalise des recherches de similarité itératives dans la banque. En pratique, jusqu'à 10 itérations sont menées avant de se résoudre à l'absence d'homologie. Or, pour SECTM1 une similarité significative avec le domaine Ig (Espérance = 2×10^{-11}) a été trouvée dès la 3ème itération. C'est sur la base de cet argument que nous avons conclus à une homologie distante entre SECTM1 et le domaine Ig, et que nous avons proposé un modèle structural par homologie. Comme les domaines Ig sont associés aux protéines du système immunitaires (anticorps, protéines de présentation d'antigènes, etc …), cette fonction est transférée à SECTM1 par homologie. Néanmoins, la résolution expérimentale de la structure de SECTM1 sera nécessaire pour confirmer ce résultat.

G.3.1.12.3. Orthologie entre gènes humains et souris

Etant donné l'orthologie inégale entre le nombre de copies de gènes *Tex19* et *Sectm1* chez l'humain (1 copie) et la souris (2 copies), nous avons cherché à déterminer de quels orthologues murins, les gènes humains sont les plus proches. Par plusieurs approches, nous avons mis en évidence que *Tex19* et *Sectm1* humains sont plus proche de Tex19.2 et Sectm1a de souris, respectivement. Si la comparaison des structures des gènes était ambigüe pour *Tex19*, l'utilisation du BLASTP par « meilleure touche (*best hit*) » était sans équivoque. Nous avons insisté sur cette démonstration car une erreur semble s'être glissée dans une étude antérieure [Kuntz S. *et al*, 2008]. En se basant sur la synténie des régions génomiques, il avait été déclaré que *Tex19* humain était plus proche de *Tex19.1* car ces 2 gènes sont voisins d'*Uts2r* sur le chromosome. Etant donné la proximité de séquence que nous avons mise en évidence entre TEX19 et TEX19.2, il se peut qu'un réarrangement survenu chez l'ancêtre comun des *Sciurognathi* ait inversé la région qui intègre *Tex19.1* et *Tex19.2*.

G.3.1.12.4. Fonction de la protéine TEX19

Il arrive que la comparaison de séquence soit impuissante à propager une fonction. Il s'agit du cas où une protéine ne possède de similarité de séquence à aucune autre connue [Tautz D. et Domazet-Loso T., 2011]. Comme recours, il reste alors la construction d'alignements multiples pour identifier des résidus conservés et déterminer des signatures fonctionnelles. L'alignement de TEX19 a révélé un domaine de 50 résidus trés conservés en extrémité N-terminale dans lequel sont prédites des hélices-α et 4 cystéines invariantes dont 2 sont incluses dans une signature CFxCF. Généralement, les cystéines invariantes sont observées dans 2 situations i) si elles sont engagées dans des ponts disulfures ou ii) si elles contribuent à la coordination d'un ion métallique. Les protéines stabilisées par ponts disulfures sont sécrétées dans le milieu extracellulaire ou bien ancrées à la membrane plasmique par hélice(s) transmembranaire(s) et exposées à l'extérieur de la cellule [Bosnjak I. *et al*, 2014]. Au contraire, les protéines cytoplasmiques ne possèdent pas de ponts disulfures car le cytoplasme n'est pas un milieu oxydant. De plus, des enzymes qui catalysent la réaction d'oxydation des cystéines sont présentes dans la voie cellulaire de sécrétion. Or TEX19 est une protéine cytoplasmique. Par conséquent, ses cystéines invariantes ne forment pas de pont disulfure. Le motif CXXC de TEX19 est observé à la fois dans des enzymes à activité catalytique redox comme les thioredoxines et chez les protéines à doigt de zinc [Krishna S.S. *et al*, 2003]. Si TEX19 présentait une séquence homologue à celle d'un doigt de zinc, celle-ci aurait été identifiée par

des outils comme PSI-BLAST. La structure de TEX19 étant inconnue, toute information qui peut être arrachée sur sa fonction devient cruciale. En utilisant la phylogénèse moléculaire, j'ai montré la coévolution de *Tex19* et *Sectm1* et par conséquent la liaison fonctionnelle entre les 2 gènes. La notion de fonction protéique est trés large et intègre le rôle moléculaire, le processus biologique, l'intéraction avec un partenaire, l'appartenance à un complexe, éventuellement la localisation cellulaire [Patthy L, 2008]. Il a été montré que *Tex19* contribue au contrôle de l'activité de certains transposons et que sa délétion est associée à des cassures d'ADN double brin au cours de la méiose. Cependant, d'une part le mécanisme d'action de la protéine codée par *Tex19* est inconnu et d'autre part notre étude tend à soutenir que *Tex19* jouerait un rôle à l'interface avec des processus immunitaires. Au niveau moléculaire, la perte du domaine C-terminal de TEX19 chez les *Haplorrhini* et l'acquisition d'un domaine en C-terminus de SECTM1 chez ces mêmes espèces est intriguante et pose des questions sur le plan fonctionnel.

G.3.1.12.5. Perspectives

Etant donné que la protéine TEX19 a été identifiée comme un antigène « testicule/cancer », il n'est pas à douter qu'elle va susciter de nouvelles investigations. D'autre part, la gravité du phénotype de la délétion du gène chez la souris, *i.e.* des cassures de l'ADN-double brin à la méiose, indique une fonction cruciale pour la survie de l'espèce. La cause de la coévolution entre *Tex19* et *Sectm1* est une question brûlante et doit être élucidée. De plus, la présence de ces 2 gènes est une spécificité du génome des mammifères placentaires et pourrait être à l'origine d'un saut évolutif en raison de l'expression de *Tex19* dans les tissus reproducteurs (testicule, placenta). Clairement, la résolution des structures 3D de TEX19 – notamment le domaine « MCP » très conservé en N-terminus - et celle du complexe entre les domaines Ig de CD7 et SECTM1 apporteraient des informations précieuses pour comprendre la fonction des 2 protéines. Au laboratoire, il serait également utile de tester s'il existe une intéraction directe entre TEX19 et SECTM1. Chez les *Haplorrhini*, l'isoforme SECTM1 ancré à la membrane plasmique possède un domaine intracellulaire (domaine B, voir figure 16d) qui pourrait intéragir avec TEX19. Enfin, si l'hypothèse d'un mécanisme d'inhibition d'une réponse immunitaire impliquant TEX19 et SECTM1 s'avérait correcte, cette découverte pourrait revêtir une importance toute particulière pour comprendre comment les cellules cancéreuses s'évadent de la surveillance du système immunitaire.

G.3.2. Assemblages alternatifs en homodimères des domaines de liaison au ligand (LBD) des récepteurs stéroidiens aux estrogènes (ERα) et aux glucocorticoïdes (GRα)

G.3.2.1. Contexte biologique

Dans la cellule, les facteurs de transcription sont des protéines qui se fixent à l'ADN génomique et régulent l'expression de gènes cibles. Les récepteurs nucléaires (RNs) constituent une super-famille de facteurs de transcription présents chez les métazoaires (organismes animaux pluri-cellulaires) et dont l'apparition est antérieure à la séparation des invertébrés et des vertébrés. Les RNs jouent un rôle dans des processus physiologiques clés comme le développement, le métabolisme et la reproduction [Mangelsdorf D.J. *et al*, 1995]. Une fois activés par fixation d'un petit ligand hydrophobe, les RNs se lient spécifiquement à certaines séquences d'ADN génomique appelées « éléments de réponse » qui sont constitués de 2 hexanucléotides en répétition directe, indirecte ou inversée. Les effets pléiotropiques des RNs sont liés à leur capacité d'agir comme effecteurs allostériques entre différents signaux moléculaires, tels que le ligand, l'élément de réponse sur l'ADN, les modifications post-traductionnelles et les protéines partenaires d'intéraction comme les coactivateurs ou les corépresseurs [Weikum E.R. *et al*, 2017]. Pour activer la transcription des gènes, la fixation sur un élément de réponse et la dimérisation du récepteur permettent le recrutement de corégulateurs qui sont capables de modifier les histones par acétylation des lysines et enfin de recruter la machinerie basale de la transcription dont l'ARN polymérase II. Les RNs partagent une architecture en 2 domaines structurés, le domaine de liaison à l'ADN (DBD) et le domaine de liaison au ligand (LBD) d'environ 65 et 250 acides-aminés, respectivement. Le DBD et le LBD sont connectés par une courte région charnière non-structurée de 20 à 40 résidus. Aux extrémités N- et C-terminales existent 2 régions intrinsèquement désordonnées de longueurs variables et peu conservées. Pour se lier à l'ADN, le DBD contient 2 motifs en doigts de zinc tandis que le LBD se replie en un sandwich caractéristique de 12 hélices-α qui créent une poche hydrophobe dans laquelle se glisse le ligand (Figure 21). Chez l'humain, 48 RNs ont été dénombrés et se répartissent en 6 groupes [Robinson-Rechavi M. *et al*, 2001; Germain P. *et al*, 2006]. Dans le groupe des stéroidiens ou groupe « NR3 » selon la nomenclature de Robinson-Rechavi M. *et al*, 2001, les récepteurs sont activés par des ligands qui dérivent du cholestérol. Ce groupe se divise en 3 sous-groupes, ce sont les récepteurs aux estrogènes (NR3A), les récepteurs apparentés aux estrogènes (NR3B) et les récepteurs oxo-stéroidiens (NR3C). Chez

les récepteurs aux estrogènes, le récepteur alpha (ERα) (NR3A1) s'exprime dans des tissus comme l'ovaire, l'utérus, les glandes mammaires, la glande pituitaire, le système nerveux central, le squelette et le système cardio-vasculaire [Bookout A.L. *et al*, 2006]. ERα joue un rôle dans le développement et la maintenance des organes dans lesquels il s'exprime; en particulier il est impliqué dans la maturation du phénotype reproducteur mâle et femelle. Chez les oxo-stéroidiens, le récepteur aux glucocorticoïdes (GRα) (NR3C1) est ubiquitairement exprimé et régule des processus aussi divers que le métabolisme du glucose et des lipides, l'homéostasie, la réponse au stress, l'inflammation et les rythmes circadiens. Notons enfin qu'ERα et GRα sont des cibles thérapeutiques majeures pour traiter le cancer du sein et les maladies inflammatoires, respectivement. L'information structurale et fonctionnelle de ces récepteurs revêt donc une importance cruciale pour l'élaboration de nouvelles molécules pharmacologiques (voir G.9. article Bianchetti L. *et al*, 2018 pour une introduction plus détaillée).

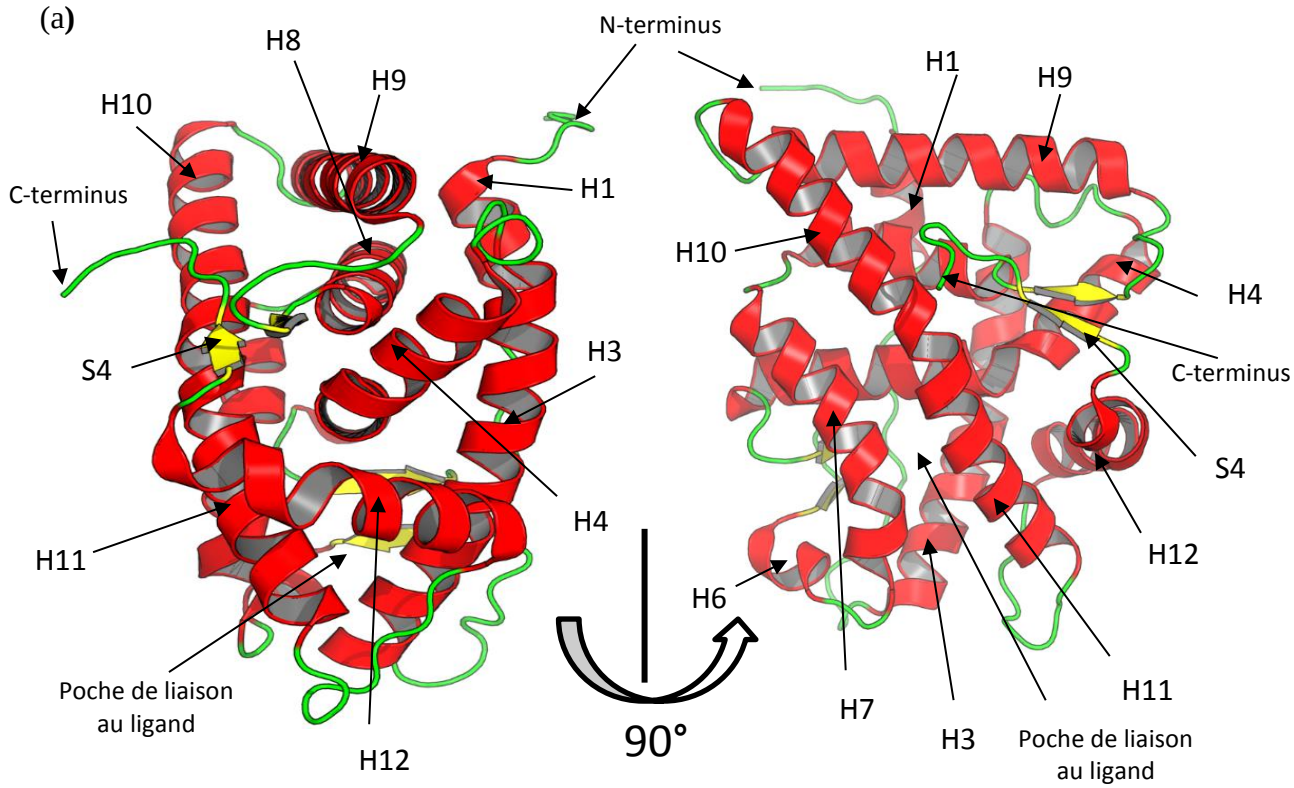


Figure 21: **Représentation schématique de la structure 3D du LBD d'un récepteur nucléaire**. La numérotation des hélices-α (H) et des brins β (S) suit la nomenclature officielle. En rouge, jaune et vert sont représentés les hélices-α les brins β et les boucles, respectivement.

G.3.2.2. Etat des connaissances sur l'assemblage en homodimère des LBDs d'ERα et GRα

L'interaction protéine-protéine est un pivot de la physiologie cellulaire. C'est un état moléculaire régit par les propriétés physico-chimiques des acides-aminés de surface des protéines qui sont amenés en contact atomique dans une interface d'interaction. Dans un complexe, la stabilité des protéines est augmentée et l'assemblage est souvent nécessaire pour activer une fonction [Marianayagam N.J. *et al*, 2004]. De plus, il a été montré que les acides-aminés qui interagissent aux interfaces de contact sont souvent conservés par l'évolution [Valdar W.S.J. & Thornton JM, 2001]. La détermination de la structure 3D des assemblages protéiques est cruciale pour comprendre comment les protéines remplissent leur fonction, pour guider l'ingénierie de molécules pharmaceutiques et comprendre le rôle des mutations dans des pathologies. La capacité des RNs à réguler la transcription des gènes dépend de leur propriété d'homo- ou d'hétérodimérisation lorsqu'ils se lient à l'ADN. En effet, les éléments de réponse sont souvent constitués de 2 motifs répétés qui fixent chacuns un DBD. Il est généralement admis que les LBDs contribuent une interface de contact plus large que les DBDs pour stabiliser le complexe. L'interface de dimérisation implique les hélices 9, 10 et 11 du LBD [Germain P. et Bourguet W., 2013]. Cet assemblage canonique - que j'appellerai le « papillon » par simplification – (voir Figure 22a) a été observé dans de nombreux cristaux de LBDs tels que l'homodimère du (ERα), l'homodimère du facteur 4 hépatocytaire (HNF4), l'hétérodimère du récepteur-γ activé du prolifératéur de peroxisome (PPARγ) et du récepteur–α au rétinoïde X (RXRα). Bien qu'ERα et GRα soient des protéines homologues qui plus est du groupe des RNs stéroidiens, les architectures des homodimères de ces 2 récepteurs sont radicalement différentes (Figure 22). Si l'homodimère du LBD d'ERα utilise la forme « papillon » [Brzozowski A.M. *et al*, 1997], GRα dimérise par la boucle située entre H1 et H3, les brins β1 et 2 et l'extrémité C-terminale de H5 [Bledsoe R.K. *et al*, 2002]. Par simplification, j'appellerai l'assemblage atypique de l'homodimère de GRα la « chauve-souris » (voir Figure 22b). Des données structurales suggèrent que le GRα ne pourrait pas former d'homodimère canonique en raison de son extrémité C-terminale qui élève un obstacle stérique [Billas I. & Moras D., 2013]. Les structures en homodimère des LBDs d'ERα et GRα ont été obtenues pas cristallographie aux rayons-X. Or, dans les cristaux de protéines, les contacts cristallins sont indiscernables des contacts biologiques [Kobe B. *et al*, 2008; Capitani G. *et al*, 2016]. Lors de l'exploration d'un cristal, les surfaces de contact les plus grandes peuvent orienter le biologiste vers l'assemblage

physiologique, mais ceci n'est en aucune façon une règle. Pour cette raison, des efforts doivent être investis pour valider les interfaces de contact dans les complexes par modélisation moléculaire (calcul d'énergie libre de liaison), analyse de conservation de séquence et expérimentation au laboratoire (mutation des résidus d'interface, expériences biophysiques comme l'ultra-centrifugation analytique). Dans ce projet, la structure de l'homodimère du LBD de GRα a été examinée au moyen de la modélisation moléculaire et de l'évolution de séquence pour en évaluer la validité. Comme point de référence, l'homodimère papillon d'ERα a été utilisé car il présente la forme canonique d'assemblage de LBDs qui est largement acceptée.

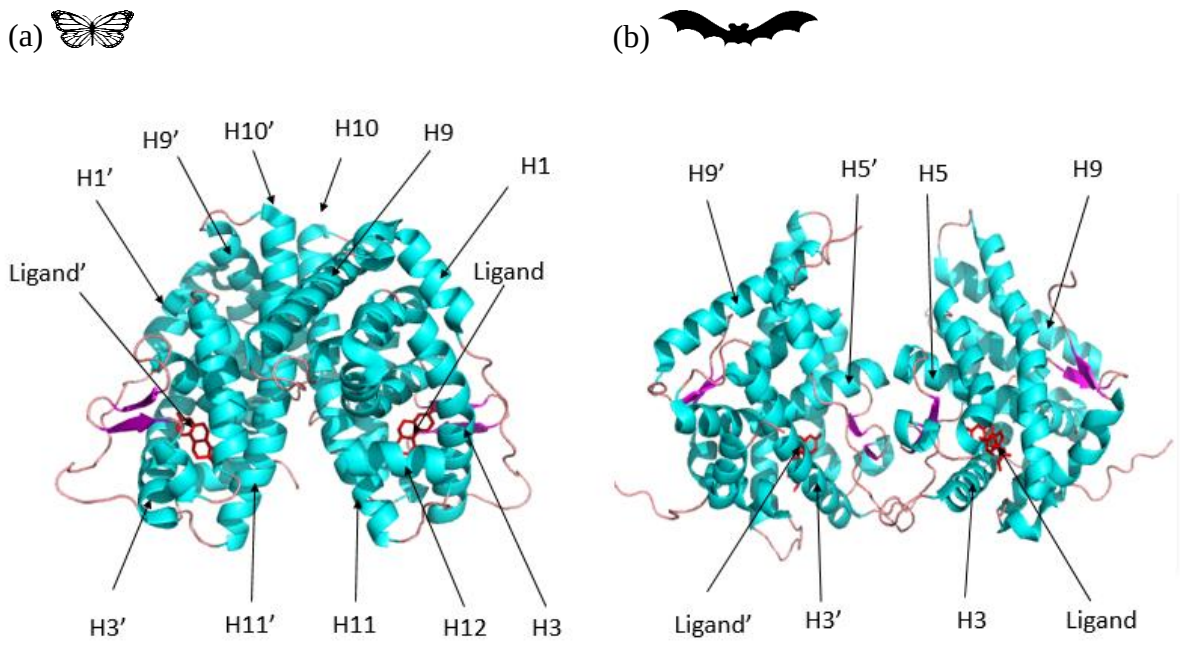


Figure 22 : **Assemblages en homodimères des LBDs d'ERα et GRα**. (a) Homodimère canonique d'ERα [Brzozowski A.M. *et al*, 1997] (le « papillon »). (b) Homodimère de GRα [Bledsoe R.K. *et al*, 2002] (la « chauve-souris »); les ligands sont en rouge. La nomenclature de certaines hélices-a est indiquée. Les apostrophes dénotent les hélices des monomères représentés à gauche.

G.3.2.3. Comparaison séquence-structure des LBDs d'ERα et GRα

Bien que les séquences humaines des LBDs d'ERα et GRα ne partagent que 26% d'identité (Figure 23a), les repliements des 2 domaines en 12 hélices-α présentent une homologie structurale frappante **(**Figure 23b**)**.

(a)

```
                      H1                          H3                              H4           H5
                 -----------                ------------------------        -----------  ------------
ERα  ----------LTADQMVSALLDAEPPMIYSEYDPSRPFSEASMMGLLTNLADRELVHMINWAKRVPGFGDLNLHDQVHLLECAWLEIILMIGLVWRSME-H
                     || |   ||  :|  || | | |    |  |  |  |       |||  |||  | | ||  ||   |  :    | |||
GRα  MVPAALPQLT---PTLVSLLEVIEPEVLYAGYDSSVPDSAWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDDQMTLLQYSWMFLMVFALGWRSYRQA
                   ----------               -------------------------        ----------- ----------------
                   H1                       H3                               H4          H5


        S1     S2   H6      H7                   H8                        H9
        ---    ---- -----   ------------------   --------------            ----------------------------
ERα  P-GKLLFAPNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSGVYT---FLSSTLKSLEEKDHIHRVLDKITDTLIHLMAKAG
         | ||| |::  :       | :     |  |:     :   ||  | |:::|| |                   :            |    || |
GRα  SGNLLCFAPDLIINEQRMTLPC-MYDQCKHMLFISTELQRLQVSYEEYLCMKTLLLLSSV--PKEGL------K--SQELFDEIRMTYIKELGKAIAKRG
        -----  --- ------   ------------------   --------------    --            -----------------------
        S1     S2  H6       H7                   H8                S3            H9


      H10-11                                      H12
      ----------------------------------          ---------
ERα  LTLQQQHRRLAQLLLILSHIRHMSNKGMEHLYNMK-C-KNVVPLYDLLLEMLDAH-----------------
         |:  |  ||  :|               :            : | | :
GRα  GNSSQNWQRFYQLTKLLDSMHDVVENLLSYCFQTFLDKSMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK
      --------------------------------------      ----------------  ---
      H10-11                                      H12               S4
```

(b)

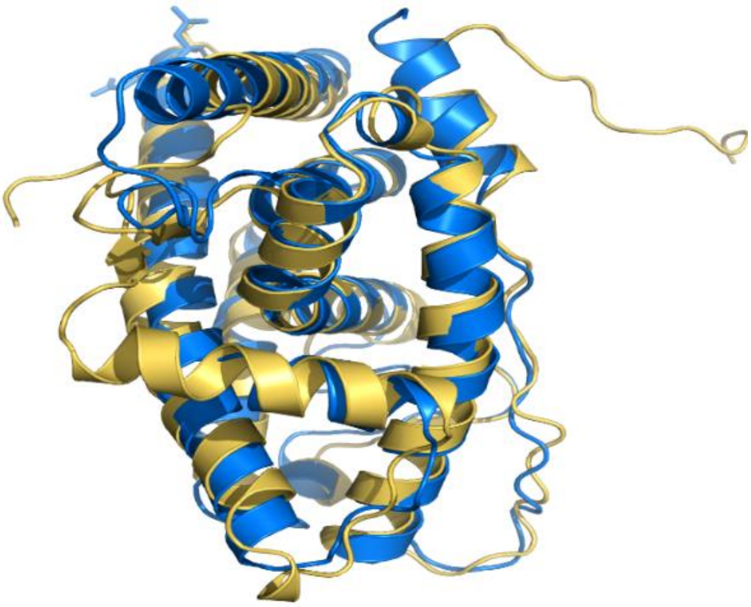


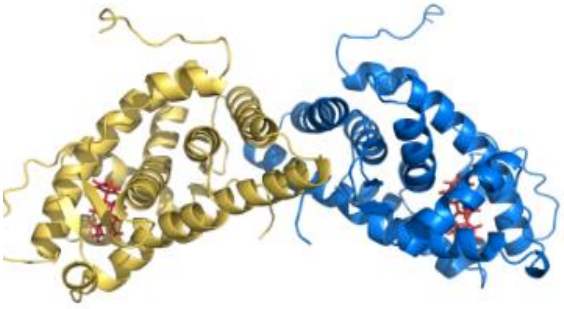
Figure 23: **Comparaison des LBDs d'ERα et GRα au niveau séquence et structure**. (a) Comparaison des séquences primaires de LBD d'ERα (Swissprot:P19785, région 314 à 551) et GRα (RefSeq:NP_032199, région 535 à 792). Les protéines ont été alignées structuralement avec SALIGN [Braberg H. *et al*, 2012]. Les hélices-α sont numérotées de 1 à 12, NB : ERα et GRα n'ont pas d'hélice n°2, les S indiquent des brins beta. (b) Superposition structurale des LBDs d'ERα (PDB:1G50) et GRα (PDB:1M2Z). RMSD:1.238 Å.

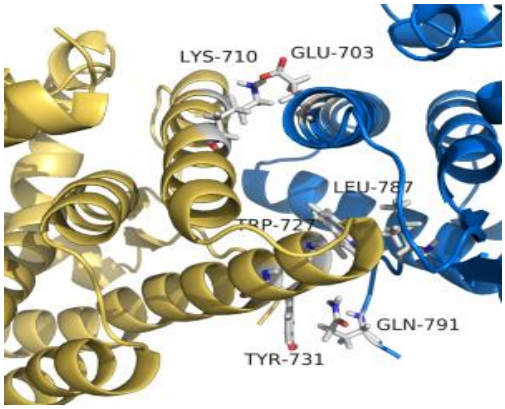
G.3.2.4. Assemblages en homodimères observés dans les cristaux

G.3.2.4.1. Contacts des LBDs de GRα

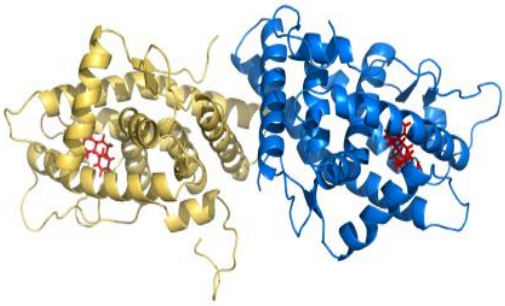
Au total, 21 enregistrements de LBD de GRα ont été obtenus exhaustivement dans la PDB. La fouille automatique des cristaux de GRα avec l'outil PISA [Krissinel E. et Henrick K., 2007] a permis d'identifier 6 assemblages différents d'homodimères (Figure 24). Nous avons donné à ces assemblages un nom basé sur les structures secondaires amenées à l'interface de contact (Figure 24). Ce sont les complexes H1, H11-H12, pH9 (pour H9 parallèles) et apH9 (pour H9 anti-parallèles), Cter, et chauve-souris. Dans les cristaux, la fréquence de ces contacts est hétérogène (Figure 25). L'assemblage « H1 » est le plus fréquent car il apparaît dans quasiment la moitié des cristaux. A l'opposé, la forme papillon (« *butterfly-like* ») n'a jamais été observée dans les cristaux de GRα. Des assemblages comme la « chauve-souris » (« *bat-like* »), le « pH9 », le « H11-H12 » et le « Cter » sont observés sporadiquement. Enfin, la forme « apH9 » est rare (1 seule occurrence).

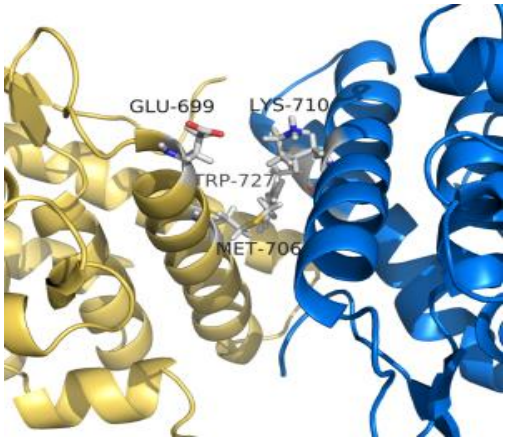
apH9 (850 Å^2)





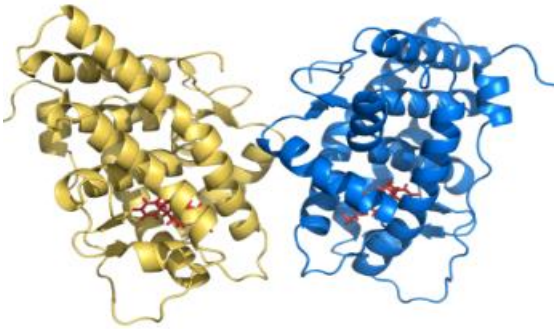
pH9
(474 Å^2)

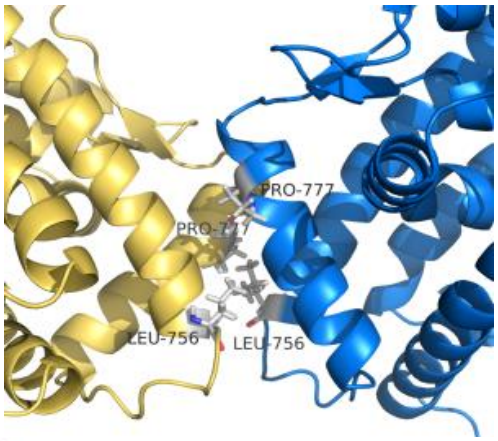




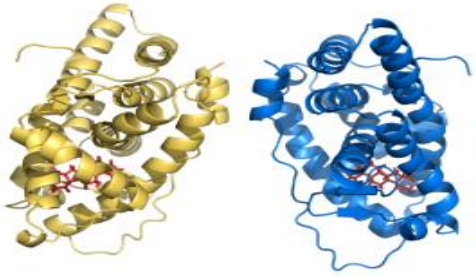
(*La figure continue sur la page suivante*)

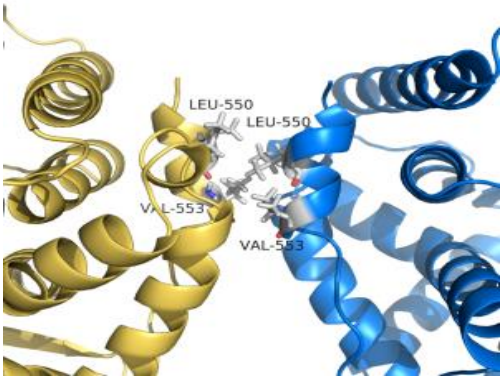
H11-H12 (436 Å^2)



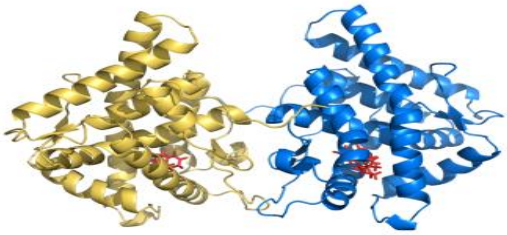


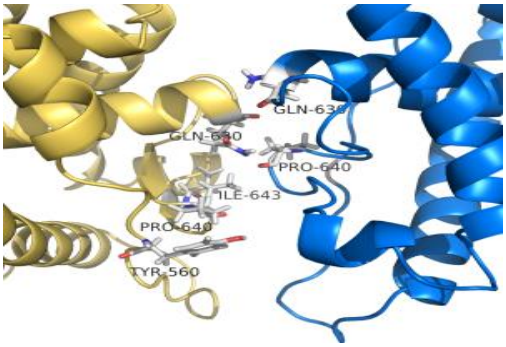
H1 (331 Å^2)



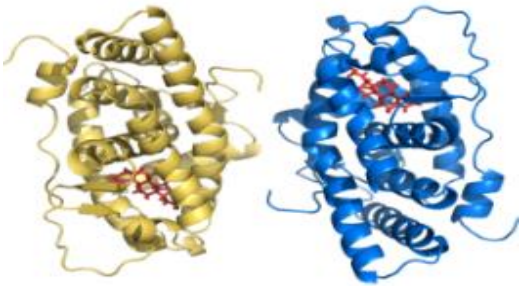


Chauve-souris
(288 Å^2)





Cter
(161 Å^2)



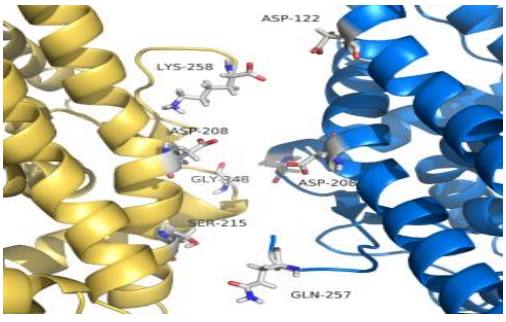


Figure 24: **Représentations structurales des assemblages en homodimères des LBDs de GRα observés dans les cristaux**. Les complexes ont été nommés en fonction des structures secondaires amenées à l'interface de dimérisation à l'exception de l'assemblage de GRα, c'est-

à-dire la chauve-souris. Les surfaces de contact sont indiquées entre parenthèses. Les ligands sont représentés en rouge. Les complexes ont été ordonnés de haut en bas par surface de contact décroissante. Pour chaque assemblage, la structure globale (à gauche) et un zoom de l'interface de contact (à droite) sont montrés. Les indices des acides-aminés se réfèrent aux séquences de souris.

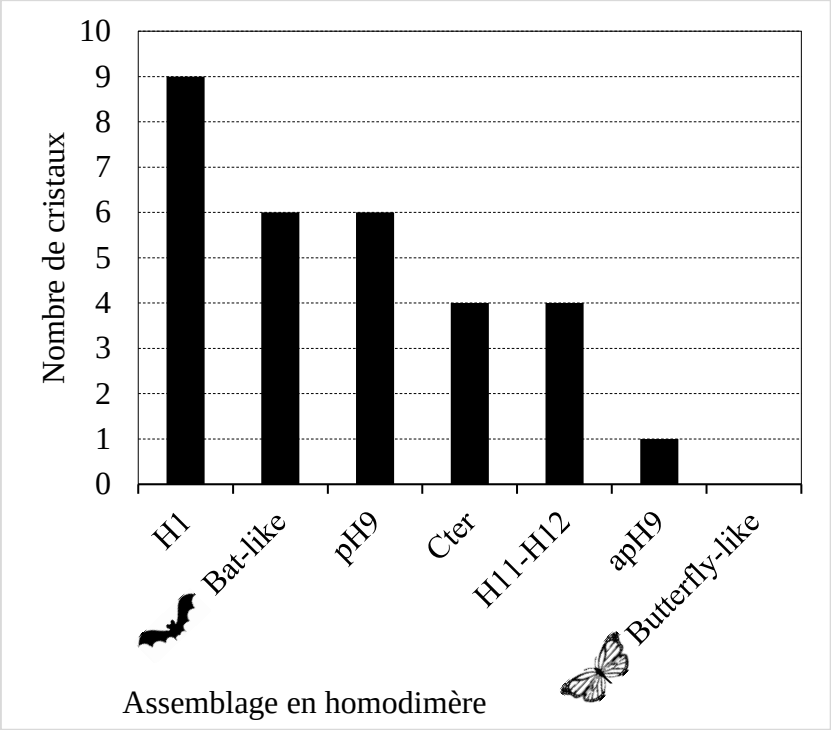
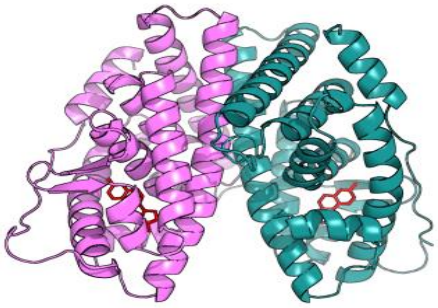


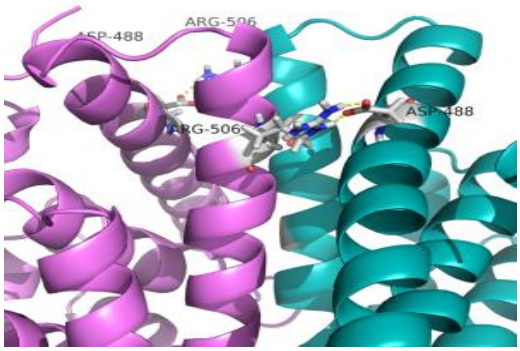
Figure 25: **Dénombrement des cristaux (enregistrements PDB) dans lesquels apparaissent les assemblages en homodimère de LBD de GRα.** Au total, 21 cristaux ont été analysés. Un assemblage peut être observé dans plusieurs cristaux et un cristal peut contenir plusieurs assemblages différents.

G.3.2.4.2. Contacts des LBDs d'ERα

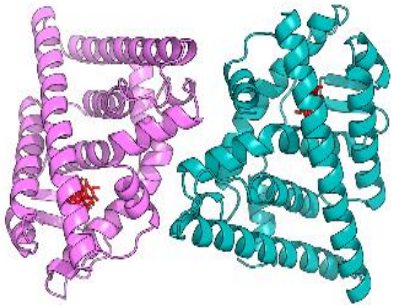
Pour ERα, l'enregistrement PDB:1G50 [Eiler S. *et al*, 2001] a été choisi pour obtenir des contacts entre LBDs. L'homodimère « papillon » a été identifié aisément entre les chaînes protéiques B et C dans le PDB:1G50 tandis que 2 contacts « H4 » et « H1-boucle-H3 » supplémentaires ont été observés dans ce cristal (Figure 26).

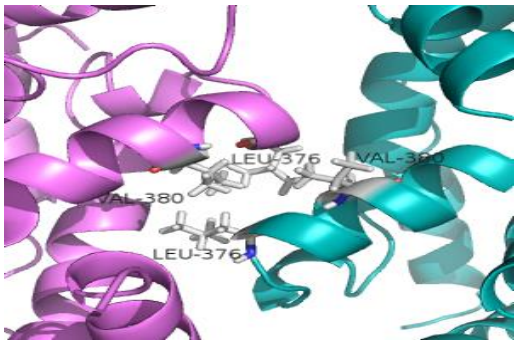
Papillon  (1494 Å^2)
Assemblage canonique



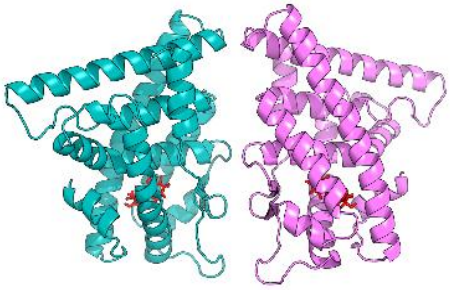


H4 (544 Å^2)





H1-Loop-H3 (342 Å^2)



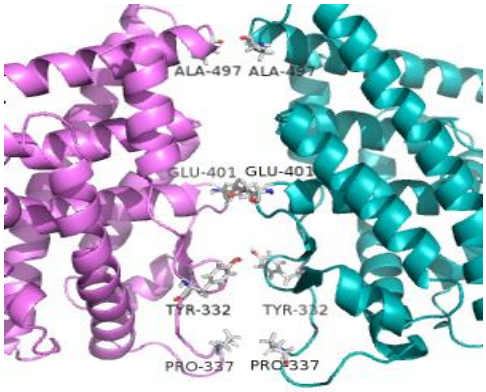


Figure 26: **Représentations structurales des contacts en homodimères des LBDs d'ERα observés dans le cristal PDB:1G50**. Les complexes ont été nommés en fonction des structures secondaires portées à l'interface de dimérisation à l'exception de la structure canonique de l'homodimère d'ERα, c'est-à-dire le papillon. Les surfaces de contact sont indiquées entre parenthèses. Les ligands sont représentés en rouge. Les complexes ont été ordonnés par surface de contact décroissante. Pour chaque assemblage, la structure globale (à gauche) et un zoom de l'interface de contact (à droite) sont montrés. Les indices des acides-aminés se réfèrent aux séquences de souris. Les monomères sont représentés en rose et cyan.

G.3.2.5. Surface de contact par assemblage

En ce qui concerne le calcul de la surface de contact entre les 2 monomères, l'outil NACCESS (Hubbard S.J. & Thornton J.M., 1992 ; http://wolf.bms.umist.ac.uk/naccess) a été appliqué aux structures moyennes issues des simulations de DM. Pour ERα, la plus grande surface de contact a été obtenue pour l'homodimère « papillon » (1.494 $Å^2$) alors que les complexes « H4 » et « H1-boucle-H3 » présentaient des surfaces de contact largement moins étendues de 544 et 342 $Å^2$, respectivement (Figure 27). Pour les homodimères GRα, la plus large surface est obtenue pour l'homodimère « apH9 » (850 $Å^2$) alors que tous les autres contacts, y compris la « chauve-souris » (288 $Å^2$), obtiennent des surfaces au moins 2 fois inférieures.

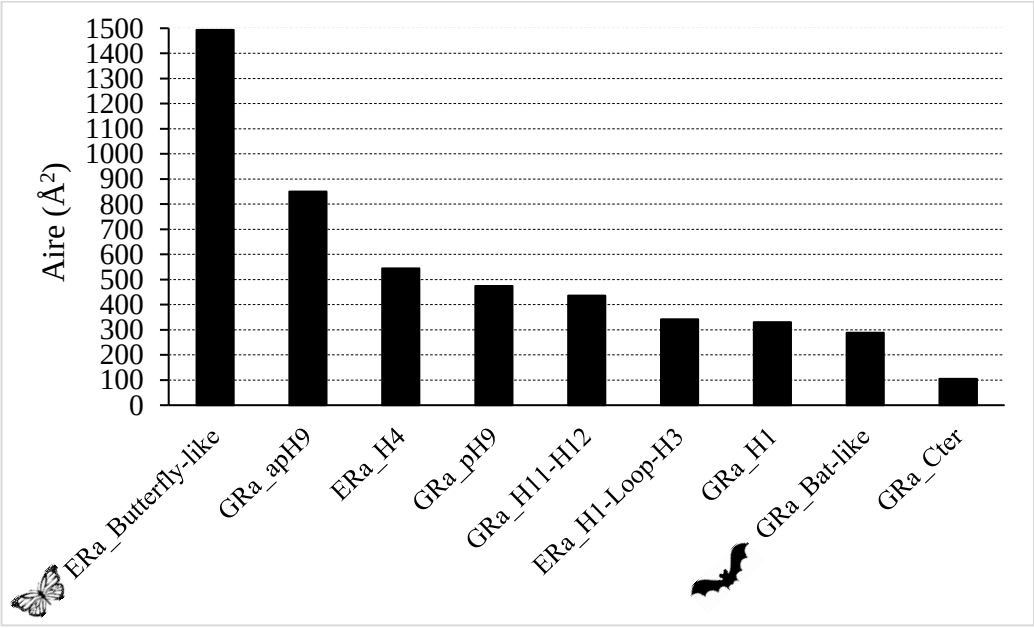


Figure 27 **: Aire de contact entre les 2 monomères dans les assemblages en homodimère des LBDs d'ERα et GRα observés dans les cristaux.**

G.3.2.6. Energie libre de liaison totale par assemblage

En modélisation moléculaire, le calcul de l'énergie libre de liaison d'un dimère est une démarche laborieuse et coûteuse en temps de calcul qui débute par une dynamique moléculaire (DM) du complexe immergé dans une boîte d'eau explicite (c'est-à-dire que les molécules d'eau sont égalements modélisées). En effet, l'homodimère est soumis à une DM car des variations structurales peuvent modifier de façon significative l'énergie libre de liaison. A l'issue de la DM, des structures représentatives sont échantillonnées sur la trajectoire et le calcul

d'énergie libre proprement dit est appliqué par décomposition de l'énergie en 3 termes, van der Waals (vdW), électrostatique (Elec) et surface polaire et non-polaire accessible au solvant (SAS). Pour chaque contact observé dans la maille cristalline, une DM de 10 ns avec le champ de force CHARMM27 [Mackerell A.D. *et al*, 1998] et le programme NAMD [Philipps G.C. *et al*, 2009] a été menée et l'énergie libre totale de formation de l'homodimère a été calculée à l'aide d'un outil développé dans le laboratoire (Dr. C. Grauffel, Dr. R.H. Stote et Dr. A. Dejaegere, voir Méthode G.2.2.3.14) [Lafont V. *et al*, 2007]. Cet outil permet d'obtenir la contribution énergétique favorable ou défavorable de chaque acide-aminé impliqué dans l'interface entre les 2 monomères. Plus cette énergie est grande (en valeur négative) plus le complexe est stable. L'homodimère papillon d'ERα a obtenu l'énergie libre de liaison négative la plus grande (-77,5 kcal/mol) alors que les contacts cristallins « H4 » et « H1-boucle-H3 » obtiennent de faibles énergies -2,5 et -9,9 kcal/mol, respectivement (Figure 28). Pour le GRα, la plus grande énergie libre de liaison est obtenue pour le complexe « apH9 » (-42 kcal/mol). Il faut noter que le ERα papillon et le GRα « apH9 » sont tous deux largement stabilisés par des forces de vdW, -130 kcal/mol et -80 kcal/mol, respectivement. Les architectures GRα « H1 » et « pH9 » obtiennent une énergie plus faible (-31 et -30,2 kcal/mol, respectivement). L'homodimère chauve-souris a présenté une énergie de liaison de -12,7 kcal/mol, donc faible.

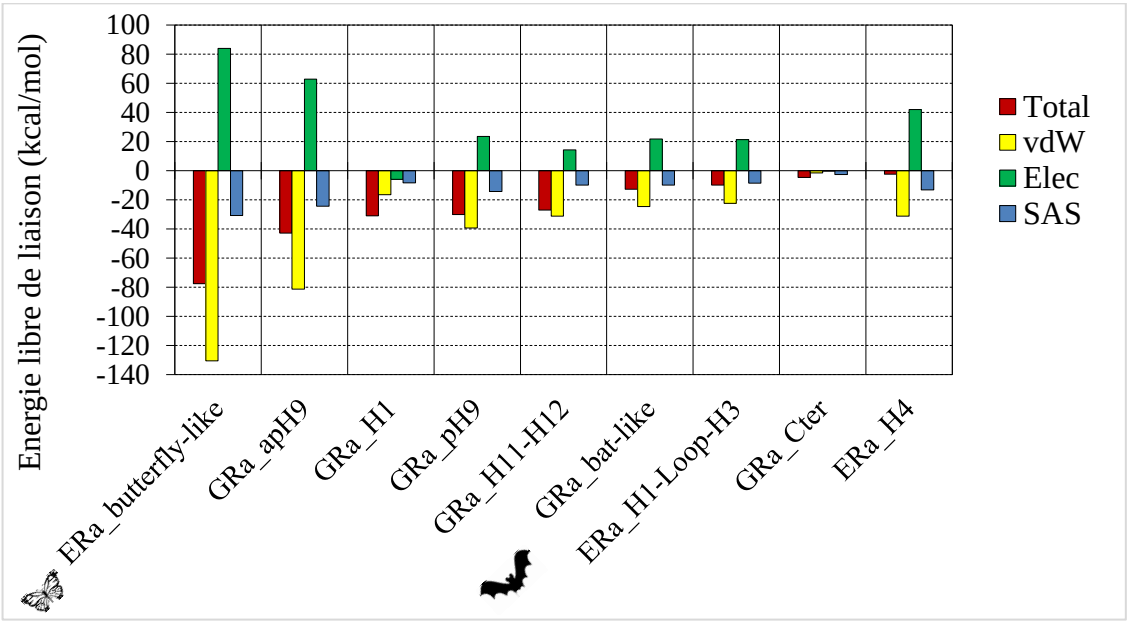


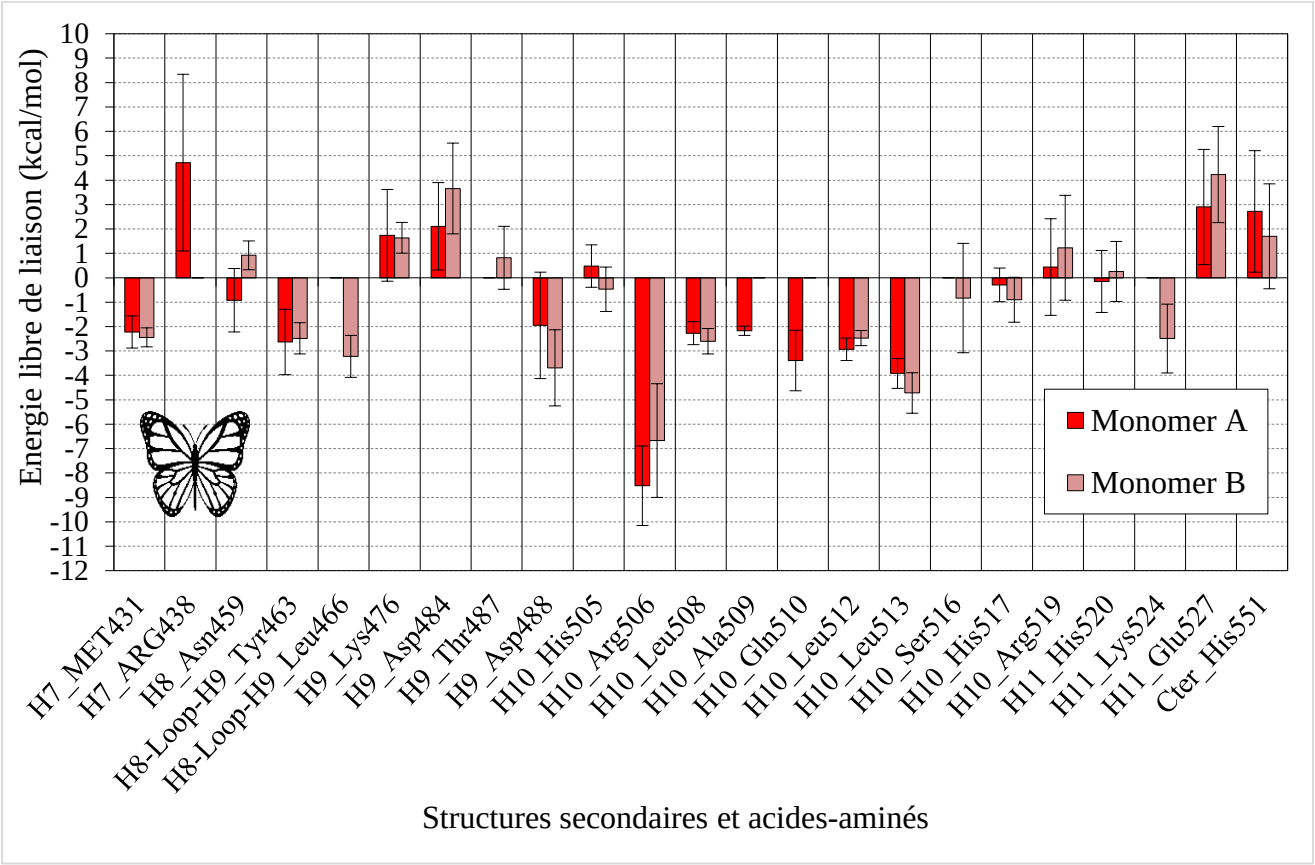
Figure 28: **Energie libre de liaison totale et décomposée (Elec, vdW, SAS) des assemblages en homodimère des LBDs d'ERα et GRα calculée par méthode MM/PBSA.** De gauche à droite, les complexes ont été rangés par énergie totale décroissante.

Tous les autres assemblages ont montré des énergies libres comprises entre -4 et -27 kcal/mol. Parmi tous les homodimères de GRα, le plus stable est donc l' « apH9 ». Bien que l'homodimère H1 soit le plus fréquent dans les mailles cristallines, son énergie libre de liaison est de l'ordre de -30 kcal/mol et présente le terme de vdW le plus faible parmi tous les assemblages analysés.

G.3.2.7. Energie libre de liaison par résidu

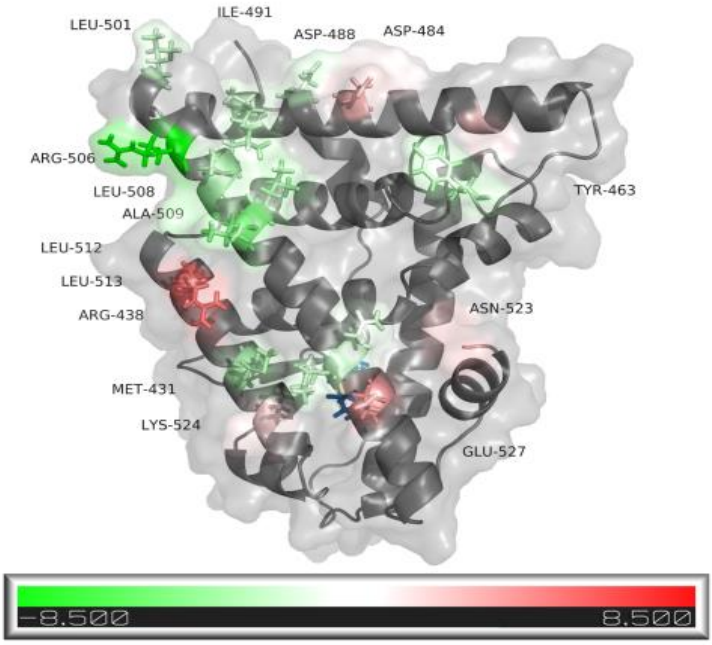
Grâce à la décomposition d'énergie libre, l'identité des acides-aminés qui contribuent à la liaison (favorablement ou défavorablement) a pu être déterminée pour chaque assemblage. Dans le cas de l'homodimère « papillon » d'ERα, l'hélice H10 joue un rôle prédominant dans la stabilisation du complexe, notamment par le résidu Arg 506 (-7 kcal/mol) et un jeu de leucines aux positions 508, 512 et 513 (-8 kcal/mol au total) (Figure 29a et b).

(a) ERα papillon

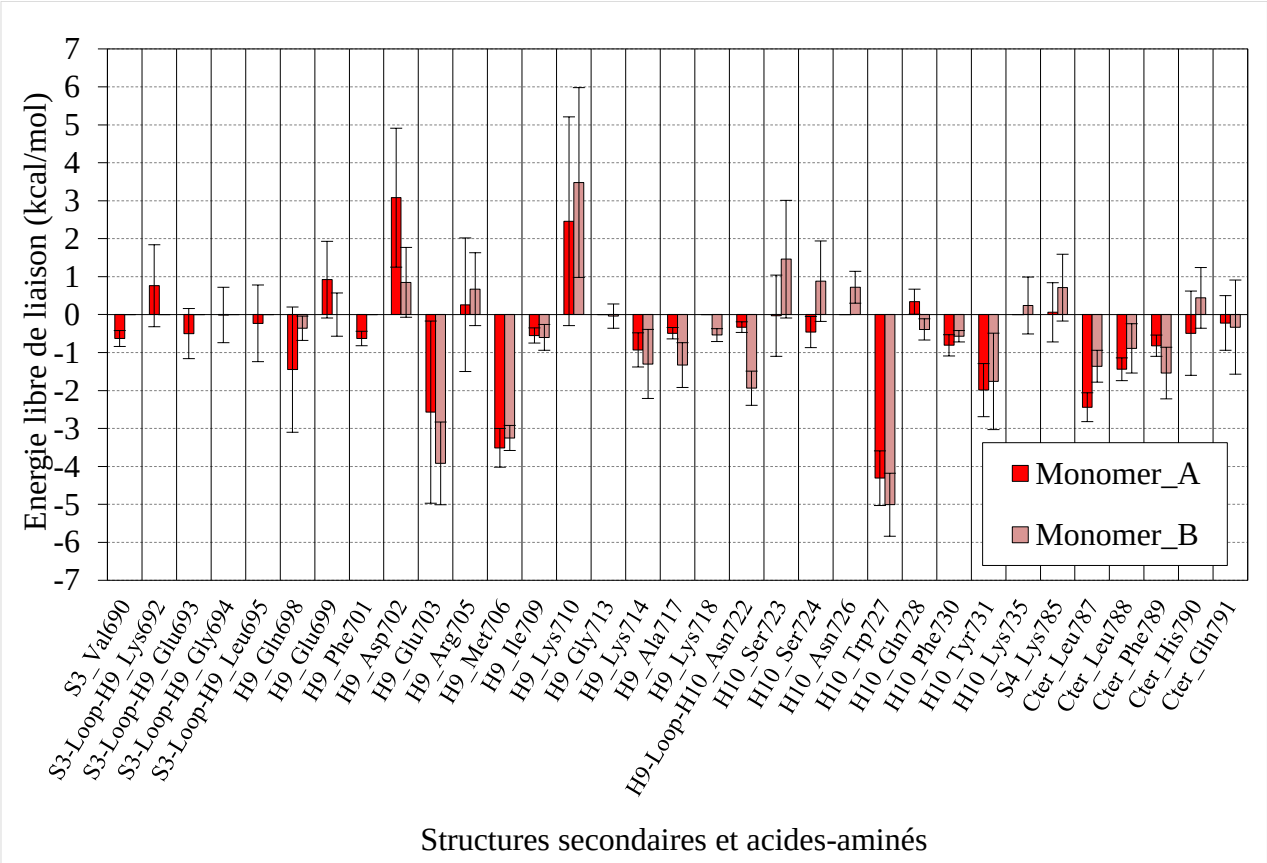


(*La figure continue sur la page suivante*)

(b) Interface de contact ERα « papillon »

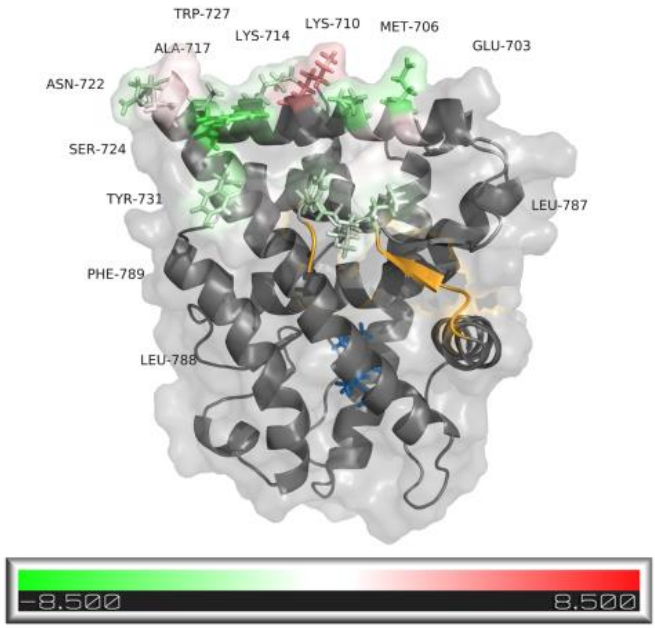


(c) GRα « apH9 »

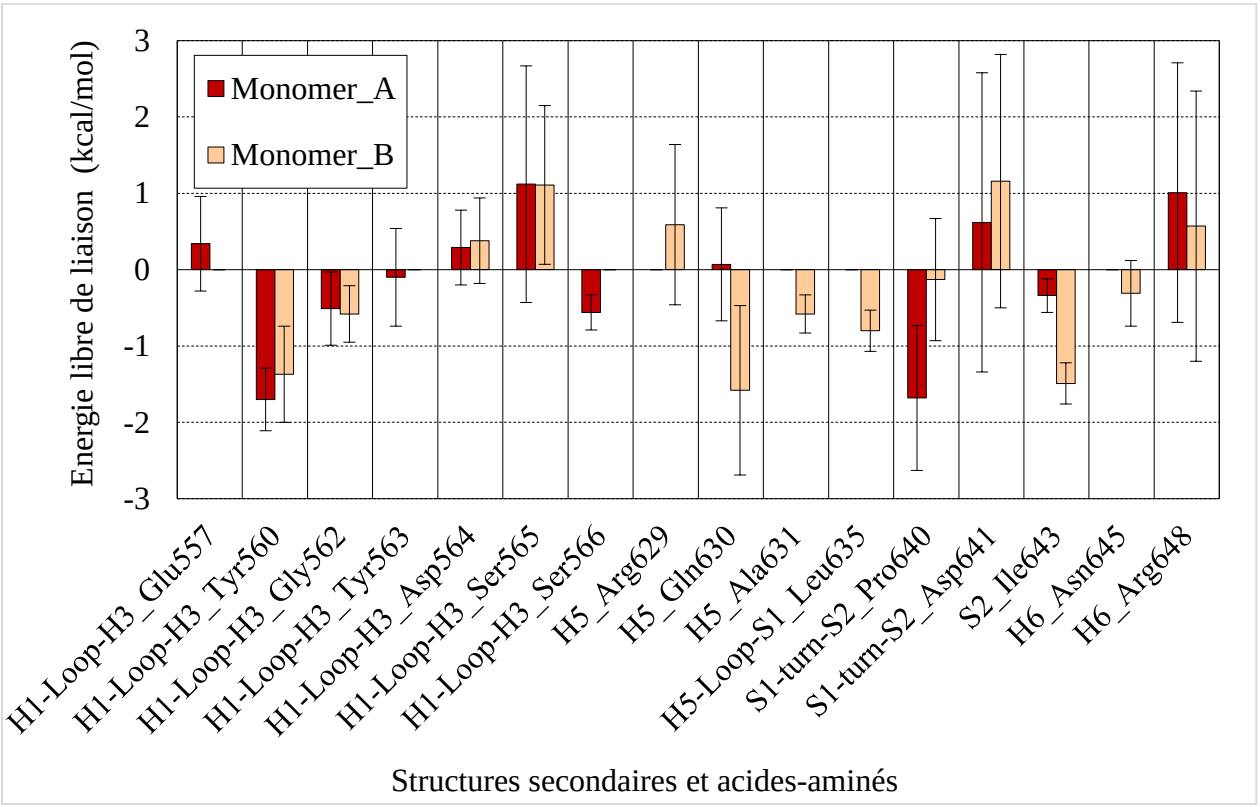


(*La figure continue sur la page suivante*)

(d) Interface de contact GRα « apH9 »



(e) GRα « chauve-souris »



(*La figure continue sur la page suivante*)

(f) Interface de contact GRα « chauve-souris »

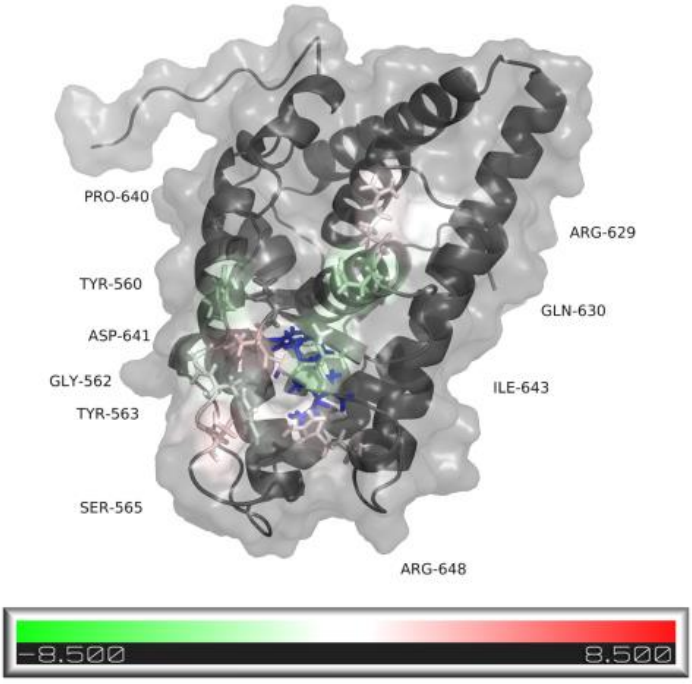


Figure 29: **Histogrammes et projection structurales de l'énergie libre de liaison totale par résidu à l'interface de dimérisation.** Les résidus qui contribuent favorablement ou défavorablement à l'énergie libre de liaison sont coloriés en vert et rouge, respectivement. Les résidus qui ne contribuent ni favorablement ni défavorablement à la liaison sont en gris. L'échelle d'énergie libre va de -8 kcal/mol (favorable) à +8 kcal/mol (défavorable). (a) Homodimère papillon ERα (b) Projection d'énergies libre sur l'interface de contact de l'un des 2 monomères de l'assemblage papillon ERα (c) Homodimère apH9 de GRα (d) Projection d'énergies libre sur l'interface de contact de l'un des 2 monomères de l'assemblage apH9 de GRα, le domaine F est en orange. (e) Homodimère GRα chauve-souris. (f) Projection d'énergies libre sur l'interface de contact de l'un des 2 monomères de l'assemblage GRα chauve-souris.

Dans l'homodimère « apH9 » de GRα, les hélices 9, 10 et le domaine F contribuent à la stabilisation (Figure 29c et d). Par exemple, un pont salin s'établit entre le Glu 703 et la Lys 710 à l'interface de dimérisation. De plus, un tryptophane de H10 (Trp 727) intriguant car largement exposé au solvant dans le LBD monomérique s'enfouit à l'interface de contact (-5 kcal/mol). Toujours dans l'homodimère apH9 de GRα il est à noter également que l'extrémité C-terminale du LBD, c'est-à-dire le domaine F, contribue à la stabilisation des 2 monomères par certains de ces acides-aminés (Leu 787, Leu 788, Phe 789). Finalement, dans l'homodimère chauve-souris de GRα, peu de résidus contribuent à la stabilité et montrent de plus des énergies de liaison faibles, *e.g.* Pro 640 (-1.5 kcal/mol) et Ile 643 (-1.5 kcal/mol) (Figure 29e et f).

G.3.2.8. Assemblages enrichis en résidus conservés à l'interface de contact.

Un test statistique de représentation de résidus conservés à l'interface a été mis en œuvre. Un alignement multiple de séquences de LBD a été construit pour ERα et GRα. Les LBDs de souris ont été utilisés comme séquences-requêtes pour une recherche de similarité par BLASTP dans les banques SwissProt, RefSeq, GenePep et TrEMBL avec les paramètres par défaut. Toutes les séquences annotées ERα ou GRα ont été collectées et alignées avec l'outil MAFFT [Katoh K. *et al*, 2002]. Les alignements ont ensuite été affinés dans Jalview [Waterhouse A.M. *et al*, 2009] et un score de conservation pour chaque acide-aminé a été obtenu par la méthode de Livingstone et Barton [Livingstone C.D. & Barton G.J., 1993]. NACCESS a permis d'identifier les acides-aminés exposés au solvant dans les monomères mais enfouis à l'interface de contact dans les homodimères. Dans l'outil R, un test de Fisher's exact avec 5% d'erreur de type I et appliqué à chaque homodimère a permis de tester la sur-représentation de résidus conservés à l'interface de dimérisation. Dans le cas d'ERα, l'homodimère papillon obtient une p-valeur largement significative ($5{,}56 \times 10^{-5}$) alors que le contact cristallin H1-boucle-H3 n'est pas significatif (0.57) (Figure 30).

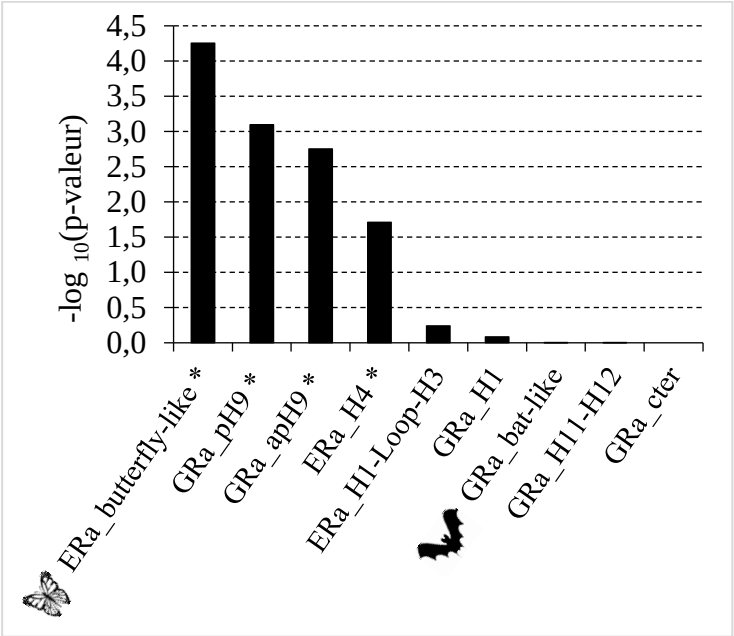


Figure 30: **Test statistique de sur-représentation de résidus conservés à l'interface de contact.** Un test unilatéral de Fisher's exact a été utilisé. Les homodimères avec une p-valeur inférieure à 0,05, c'est à dire $-\log_{10}$ (p-valeur) > 1,3, présentent une interface de contact sur-représentée en résidus conservés par rapport au reste de la surface des LBDs et sont marqués avec un astérisque.

Le contact cristallin H4 s'est au contraire révélé significatif. Par hasard, cet assemblage s'établit à l'interface de contact du LBD avec son cofacteur polypeptidique [Shiau A.K. *et al*, 1998], il est donc vraisemblable que cette surface présente une sur-représentation de résidus conservés. Pour le GRα, les complexes pH9 (p-valeur = 8 x 10^{-4}) et apH9 (p-valeur = 1.8 x 10^{-3}) ont des interfaces de contacts sur-représentées en résidus conservés. Enfin, la chauve-souris n'a pas de surface de contact sur-représentée en résidus conservés (p-valeur = 0.99).

G.3.2.9. Extrémité C-terminale/domaine F de GRα

G.3.2.9.1. Obstacle stérique à l'homodimère canonique

Il est important de noter que l'homodimère canonique de LBD, c'est-à-dire le « papillon », n'a jamais été observé dans les cristaux de PR [Williams S.P. & Sigler P.B., 1998], AR [Nadal M. *et al*, 2017], GRα et MR [Billas I. & Moras D., 2013]. Dans un modèle moléculaire d'homodimère papillon de GRα (Figure 31), l'interface de contact montre un clash stérique entre les extrémités C-terminales (domaines F) et les hélices H10 des 2 monomères. Chez GRα, comme chez MR, PR et AR, il n'est donc pas exclu que le domaine F fasse obstacle à l'homodimérisation canonique.

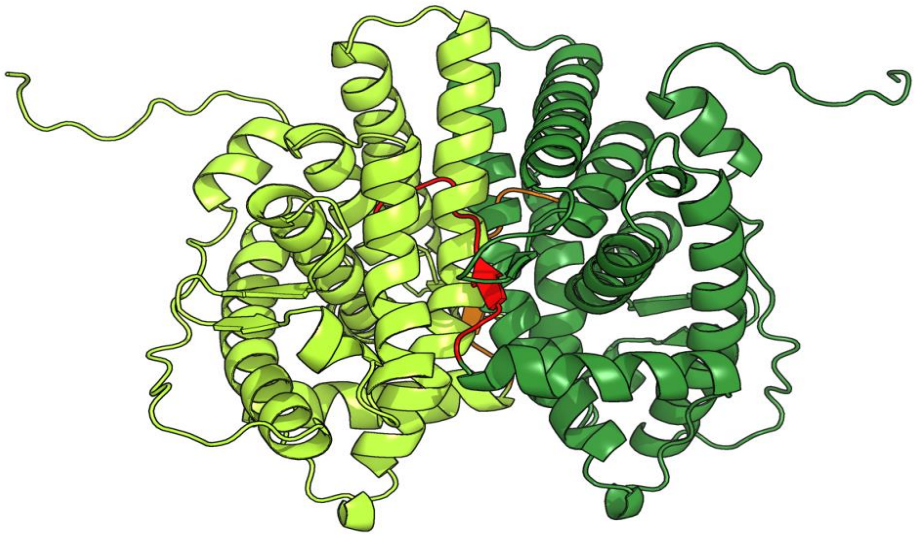


Figure 31: **Modélisation moléculaire d'un homodimère canonique du LBD de GRα.** L'homodimère d'ERα a été utilisé comme structure de référence pour la construction du modèle. Les 2 monomères sont représentés en vert clair et vert foncé. Les domaines F sont en rouge et orange. Les domaines F forment un obstacle stérique à la formation du complexe par les hélices H10.

G.3.2.9.2. Rigidité structurale de l'extrémité C-terminale

L'extrémité C-terminale du LBD de GRα qui ne mesure qu'une dizaine d'acides-aminés est un court brin β suivi d'une boucle. Or, il est généralement admis que les boucles sont flexibles. Pour tester cette hypothèse, une simulation de DM de 100 ns du LBD a été menée (PDB-ID:1M2Z). L'extrémité N-terminale du LBD montre de très grandes variations structurales (Figure 32a et b). Certaines boucles du LBD fluctuent considérablement, comme celle entre H1 et H3 (RMSF=3 Å) ou bien celle entre H5 et S1 (RMSF = 2,5 Å). Au contraire, excepté ses 2 derniers résidus, le domaine F présente une faible fluctuation structurale au cours de la DM (Figure 32a et b). Ce résultat est corroboré par des B-facteurs faibles dans cette région (un indicateur de vibration) (Figure 32a). De plus, cette séquence est fermement maintenue au contact du LBD par des ponts hydrogènes (information non-montrée) et un court feuillet β.

(a) (b)

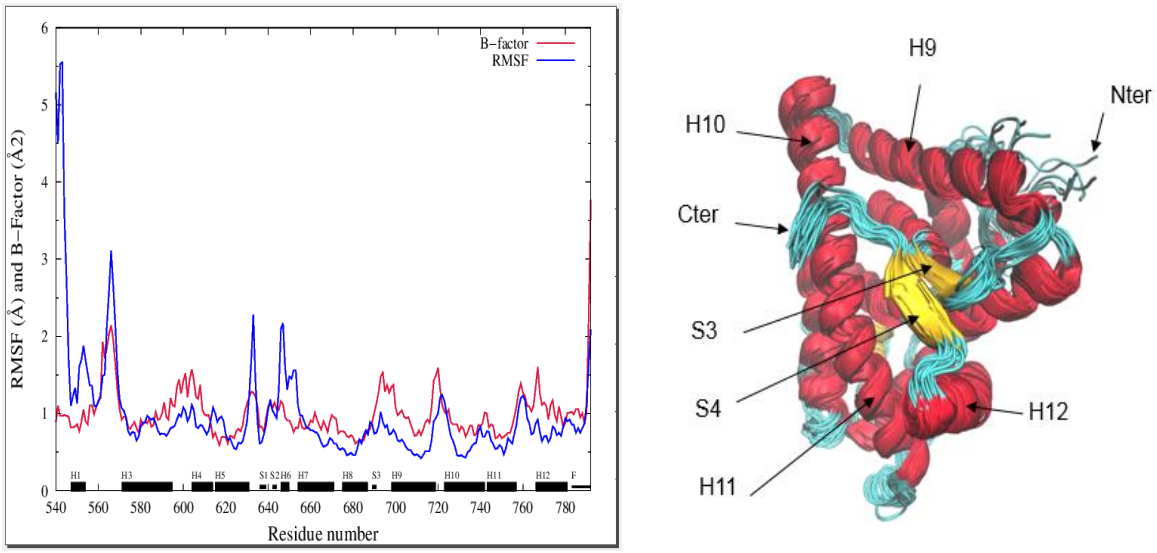


Figure 32: **Variations structurales du LBD de GRα mesurées au cours d'une simulation de DM de 100 ns.** (a) Variation du RMSF et du B-facteur le long de la séquence. Le RMSF est calculé sur la dynamique moléculaire tandis que le B-facteur est normalisé sur toutes les structures cristallographiques disponibles du LBD de GRα. (b) Echantillon représentatif de structures extraites de la DM et superposition des structures.

G.3.2.9.3. Intégrité de l'ARNm codant le domaine F

Afin d'écarter l'hypothèse qu'une édition de l'ARN messager codant le GRα pourrait supprimer cette extrémité C-terminale qui fait barrage à la formation de l'homodimère « papillon », j'ai vérifié l'intégrité de la séquence des ARN messagers en analysant les

séquences (« *reads* ») d'un RNA-seq de foie de souris, c'est-à-dire un tissu dans lequel le GRα s'exprime [Bookout A.L. *et al*, 2006]. Après alignement sur le génome de 100 millions de séquences, 22 séquences se sont localisées sur l'extrémité 3' de l'ARN messager codant le C-terminus de GRα. Aucun des reads ne présentait de substitution (Figure 33). Le domaine F est donc à priori exprimé de façon intègre dans le foie de souris.

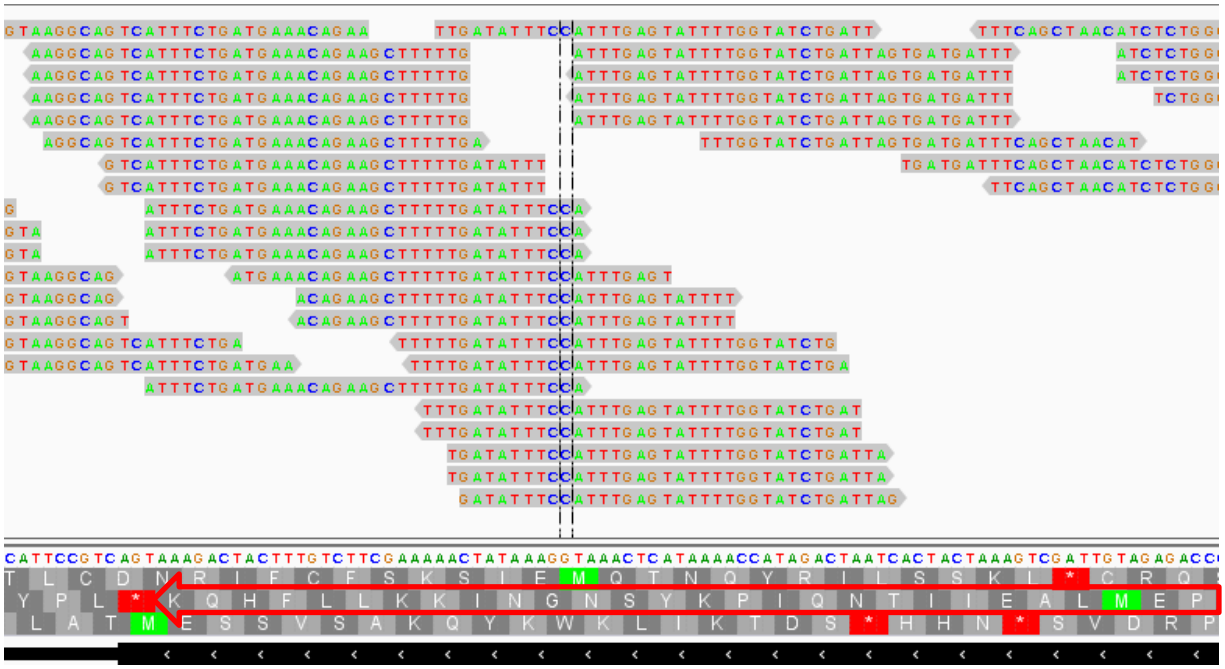


Figure 33**: Alignement de reads RNA-seq sur l'extrémité de l'ARNm codant le domaine F de GRα.** La séquence protéique du domaine F est encadrée d'une flèche de bordure rouge et écrite de droite à gauche du Nter au Cter. La séquence nucléotidique du génome codant l'extrémité Cter est la ligne écrite juste au dessus des polypeptides.

G.3.2.10. Mutations dans le LBD de GRα et transactivation

Les mutations du GRα sont à l'origine d'une pathologie appelée syndrome de Chrousos qui rend le récepteur insensible à son ligand naturel, le cortisol. Des mutations qui affectent la séquence du LBD et qui modifient à des degrés divers l'activité transcriptionnelle du récepteur ont été rapportés [Nicolaides NC *et al*, 2014]. Nous avons déterminé sur la structure du LBD de GRα la position de ces mutations (Figure 34). Les mutations affectent les hélices 3, 7, 10, 11, 12 et le domaine F. Aucun des résidus qui contribuent à l'interface « chauve-souris » n'est impliqué dans une mutation provoquant la maladie. Au contraire, la Leu 773 chez l'humain, *i.e.* Leu 788 chez la souris, qui contribue à la stabilité de l'homodimère « apH9 » (voir Figure 29c) était mutée chez un patient [Charmandari E. *et al*, 2005]. De plus, une délétion

de 2 bases dans le codon Leu 773 provoque un décalage de cadre de lecture et l'ajout de 19 résidus supplémentaires à l'extrémité C-terminale du récepteur ce qui abolit l'activité de transactivation [McMahon S.K. *et al*, 2010].

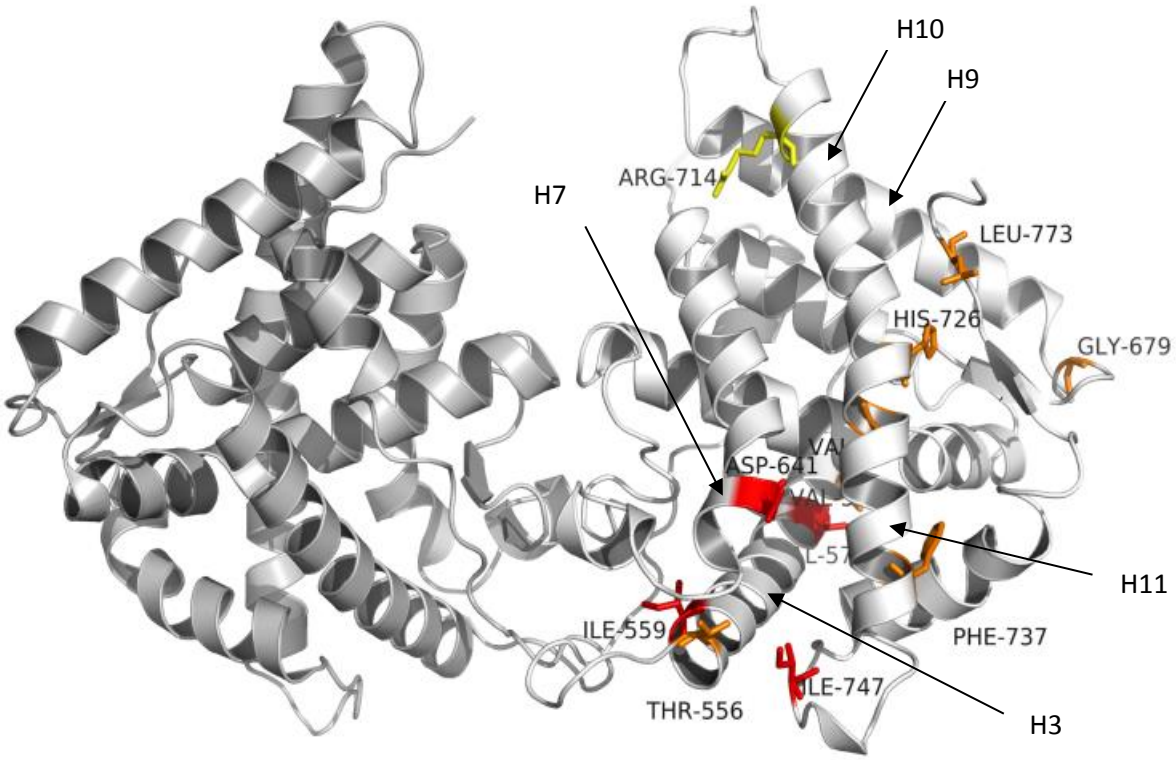


Figure 34**: Spectre de mutations dans le LBD de GRα et sévérité transactivationnelle.** Les mutations sont rapportées sur un LBD de l'homodimère « chauve-souris ». Rouge: grave, orange: moyen, jaune: faible. Les indices de résidus se réfèrent à la séquence humaine (Refseq:NP_000167). NB: ajouter 15 à l'indice de résidu de la séquence humaine pour obtenir l'indice d'acide-aminé dans la séquence de souris (Refseq:NP_032199) telle que les résidus sont rapportés dans les figures 30e et f.

G.3.2.11 Discussion

G.3.2.11.1. Assemblages alternatifs de LBD homologues

Dans la super-famille des RNs, la fonction de transactivation, c'est à dire la capacité à induire la transcription de gènes cibles en se fixant à l'ADN, est subordonnée à la fonction de dimérisation du récepteur, qu'il s'agisse d'homo- ou d'hétérodimères. L'assemblage canonique des LBDs qui met en jeu l'interface constituée par les hélices 9,10 et 11 a été démontré par des expériences *in vitro* et *in vivo* [Germain P. et Bourguet W., 2013; Jiang G. *et al*, 1997]. Dans le cas du ERα, nous avons calculé que cette interface est large (1494 $Å^2$), stabilisée par une grande énergie libre de liaison (-77 kcal/mol) et caractérisée par une sur-représentation de résidus conservés (p-valeur = 5,56 x 10^{-5}). Les résidus de la séquence d'ERα

qui construisent cette interface de contact présentent donc toutes les caractéristiques qui marquent une interface biologique. Phylogénétiquement, le GRα appartient au groupe des RNs oxo-stéroidiens, qui comprend le récepteur à la progestérone (PR), le récepteur aux androgènes (AR) – cible majeure du traitement du cancer de la prostate – et le récepteur aux minéralocorticoïdes (MR). Cependant, pour GRα de même que pour les 3 autres récepteurs oxo-stéroidiens, il est devenu de plus en plus évident que l'oligomérisation de leurs LBDs déroge à l'architecture de l'assemblage canonique. Premièrement, aucun des 21 cristaux de LBDs de GRα que nous avons analysés n'a livré de contact canonique. Deuxièmement, les oxo-stéroidiens ont été rapportés avec une diversité d'interface de dimérisation [Bledsoe R.K. *et al*, 2002; Nadal M. *et al*, 2017; Williams S.P. & Sigler P.B. 1998]. Or, la superposition structurale des LBDs d'ERα et GRα ne laisse aucun doute, ce sont des homologues. Si les fonctions de transactivation et de dimérisation du récepteur sont partagées par ERα et GRα, la propagation de l'interface de dimérisation canonique d'ERα à GRα pose problème. En effet, chez les récepteurs oxo-stéroidiens, l'interface canonique ne trouve aucun support expérimental. Pour comprendre cette divergence, nous avons analysé méticuleusement la structure du LBD de GRα et les contacts entre LBDs observés dans les mailles de ses cristaux obtenus par cristallographie aux rayons X. La divergence entre ERα et GRα est tout d'abord visible au niveau de leur séquence, les 2 récepteurs ne partagent que 26% d'identité. Ensuite, une divergence existe au niveau de leurs domaines F. Alors qu'ERα possède un domaine F peu conservé d'environ 40 résidus dont le repliement n'est pas visible sur les structures cristallographiques – car délété pour l'expression du cDNA codant le LBD [Brzozowski A.M. *et al*, 1997; Eiler S. *et al*, 2001], le domaine F de GRα mesure seulement 10 résidus et se replie en un court brin beta (S4) stabilisé par le brin beta S3 et suivi d'une boucle qui fait obstruction à l'assemblage canonique. Chez GRα, tous les cristaux de LBD possèdent cette extension qui adopte une conformation amarrée au LBD. De plus, une simulation de dynamique moléculaire de 100 ns a montré que ce domaine F est peu flexible. S'il y a consensus sur l'obstacle que représente le domaine F- chez les récepteurs oxo-stéroidiens [Williams S.P. & Sigler P.B., 1998; Bledsoe R.K. *et al*, 2002; Billas I. & Moras D., 2013; Nadal M. *et al*, 2017], il n'y en a pas au contraire sur l'interface alternative d'homodimérisation. Pour GRα, l'interface rapportée par R.K. Bledsoe *et al* (2002) met en jeu la boucle entre H1 et H3, le feuillet constitué des brins beta S1 et S2 et l'extrémité C-terminale de H5. Chez AR, l'interface de Nadal M. *et al* (2017) évoque la chauve-souris de GRα, mais elle enfouit une surface de contact plus large. Chez PR, le contact s'établit

avec H11 et H12 [Williams S.P. & Sigler P.B. 1998]. Enfin, si des contacts peuvent être observés dans les cristaux de MR, leur validité biologique n'a pas été discutée [Bledsoe R.K. *et al*, 2005]. Malgré l'homologie structurale frappante qui existe chez les LBDs de RNs, il semblerait que les oxo-stéroidiens aient évolué une barrière à la dimérisation canonique. Sur l'arbre phylogénétique des RNs, cette barrière les distingue de leurs cousins, les récepteurs aux estrogènes (ER et ERR). Or les LBDs des oxo-stéroidiens partagent une identité de séquence plus forte qu'elle ne l'est entre eux et les ERs. La diversité des interfaces de dimérisation rapportées ches les oxo-stéroidiens pose donc également une question au regard de leur homologie. En utilisant la modélisation moléculaire et l'analyse d'évolution de séquence, les interfaces de dimérisation observées dans les cristaux de LBDs de GRα ont été questionnées [Bianchetti L. *et al*, 2018]. Nous avons montré que le contact de Bledsoe R.K. *et al* (2002) – la chauve-souris - ne présente ni une bonne énergie libre de liaison, ni une sur-représentation de résidus conservés (p-valeur = 0,99) ni une large surface de contact (288 $Å^2$). Au contraire, un assemblage alternatif mettant en contact des résidus appartenants aux hélices 9, 10 et au domaine F et que nous avons nommé "apH9" a montré des caractéristiques d'interface biologique (surface 850 $Å^2$, -43 kcal/mol, p-valeur = 1.8 x 10^{-3}). Une interface similaire a été trouvée dans les cristaux de MR (Bianchetti L. *et al*, manuscrit en préparation).

G.3.2.11.2. Validation expérimentale de l'assemblage apH9

Notre approche combinant modélisation moléculaire et analyse d'évolution de séquence nous a permis de conclure que l'assemblage de l'homodimère de LBD de GRα actuellement admis, c'est à dire la chauve-souris, est vraisemblablement un artefact de contact cristallin. En effet, la surface de contact et l'énergie libre de liaison sont faibles et l'interface de contact n'est pas sur-représentée en résidus conservés. Au contraire, un assemblage alternatif que nous avons identifié dans la maille cristalline révèle une interface qui semble mieux stabiliser l'homodimère et faire plus sens évolutivement. Il s'agit de l'homodimère apH9 pour lequel une large surface de contact, une forte énergie libre de liaison (négative) et une sur-représentation significative de résidus conservés à l'interface ont été trouvées. Pour valider cette interface, des expériences de mutagénèse dirigée sont requises (collaboration en cours). La décomposition d'énergie libre par résidu offre un outil puissant pour proposer aux biologistes des résidus à muter de façon ciblée afin de déstabiliser le dimère putatif. Cependant, en raison de l'allostérie des RNs, la mutation d'acides-aminés peut avoir des conséquences sur la dimérisation quand bien même ces résidus ne seraient pas présents à l'interface.

Pour valider expérimentalement l'interface de dimérisation du LBD de GRα, l'idéal serait de questionner les interfaces de contact putatives sans avoir recours à des mutations. Une méthode élégante consiste à produire des peptides qui miment les séquences mises en jeu à l'interface de dimérisation. Cette approche s'est révélée fructueuse pour valider l'homodimère d'ERα [Yudt M.R. et Koide S., 2001]. De plus, nous avons mis en évidence que le domaine F participe à la formation de l'assemblage apH9 du GRα en s'enfouissant à l'interface de contact. Or, une autre approche pour valider cette structure consisterait à neutraliser le domaine F par un anticorps pour empêcher la formation de l'homodimère. Des résultats obtenus dans une étude antérieure [Weigel N.L. *et al*, 1992] sur le LBD de l'homologue PR tendent à montrer que le domaine F est alternativement accessible ou inaccessible à un anticorps spécifique C-262 selon la présence de ligand antagoniste (RU486) ou agoniste (progestérone), respectivement. Les auteurs concluent que la fixation du ligand agoniste au LBD induit un changement de conformation qui dérobe l'epitope du domaine F à l'anticorps. Or ce changement de conformation pourrait être en fait une dimérisation du LBD qui enfouit le domaine F à l'interface. Il faut noter que notre résultat pourrait bénéficier d'une portée plus large que le seul GRα. Pour PR et AR, deux structures d'homodimères de LBD ont été rapportées dans la littérature. L'homodimère de PR [Williams S.P. & Sigler P.B., 1998] rappelle le contact cristallin de GRα que nous avons nommé H11-H12 alors que l'homodimère de LBD d'AR [Nadal M. *et al*, 2017] évoque la chauve-souris. Pour MR, aucun homodimère n'a été décrit bien que la structure du LBD ait été résolue par cristallographie [Bledsoe R.K. *et al*, 2005]. Pour comprendre comment les récepteurs oxo-stéroïdiens dimérisent et induisent la transcription de leurs gènes cibles, il sera crucial d'identifier l'assemblage du LBD de MR (*manuscrit en préparation*) et de valider les interfaces de contacts des homodimères de PR et AR.

G.3.2.11.3. Perspectives

Sur le plan clinique, le GRα est une cible thérapeutique majeure pour le traitement de l'inflammation. Cette activité repose sur un dialogue moléculaire (« *cross-talk* ») entre GRα et les facteurs de transcription AP-1 et NFκB qui est à la base de la transrépression opérée par le récepteur. Selon un modèle appelé « ancrage », le GRα intéragit de façon directe avec AP-1 ou NFκB ce qui empêche les 2 récepteurs d'activer des gènes de l'inflammation dont ils régulent l'expression. Ce mécanisme ne requière pas la dimérisation du récepteur [Ratman D. *et al*, 2013] au contraire de l'activation des gènes sous le contrôle direct du GRα.

Chez les patients, les traitements de longue durée aux glucocorticoïdes sont responsables d'effets secondaires graves, comme diabète ou ostéoporose, et seraient dû à l'activation des gènes cibles du récepteur. Empêcher la dimérisation du récepteur de façon ciblée pourrait donc permettre de favoriser l'activité anti-inflammatoire du GRα. Certains ligands du GRα appelés « dissociés » ont été développés dans ce sens [Robertson S. *et al*, 2010]. Par conséquent, l'élucidation de l'interface correcte de dimérisation revêt toute son importance. Signalons enfin que la structure des complexes entre GRα et AP-1 ou NFκB n'est pas connue ce qui fait barrière à la compréhension du mécanisme de transrépression.

Au niveau structural, notre étude a montré l'existence d'une interface de contact impliquant la H9 du LBD et présentant les caractéristiques d'une intéraction biologique dans l'assemblage en homodimère. Certains contacts que nous avons écartés mériteraient peut-être un réexamen, comme le complexe « H1 » qui avait obtenu une bonne énergie libre de liaison sur structure cristallographique avec l'outil PISA. En effet, des publications récentes indiquent que le récepteur serait capable de tétramérisation en se fixant à l'ADN et que celle-ci dépendrait de la présence du LBD [Presman D.M. *et al*, 2016 ; Presman D.M. & Hager G.L., 2017]. En l'absence de structure 3D pour le tétramère, des tentatives de modélisation peuvent être effectuées. Cette découverte ajoute un niveau de complexité supplémentaire à l'oligomérisation du récepteur et pose la question de son rôle physiologique.

G.4. Conclusion

L'étude de la relation séquence, structure, fonction des protéines est une clé de voûte de la biologie. En m'aidant de l'éclairage apporté par l'évolution, j'ai soutenu ma thèse en décrivant 2 projets qui ont contribué à l'étude de cette relation pour 4 protéines, *i.e.* TEX19, SECTM1, ERα, et GRα. Dans les prochaines années, l'étude de TEX19 et SECTM1 pourraient gagner de l'importance pour déchiffrer l'interface de communication entre les processus biologiques de la reproduction et de l'immunité chez les euthériens. Quant à ERα et GRα se sont des cibles thérapeutiques de première importance pour le traitement de pathologies humaines (cancer du sein, inflammation, ostéoporose, ...), la dimérisation du GRα et l'activation transcriptionelle de ses gènes cibles étant la cause des effets secondaires des glucocorticoïdes utilisés en clinique.

Dans le premier projet, une coévolution entre 2 gènes non-homologues, *Tex19* et *Sectm1*, a été mise en évidence. Sans la construction des alignements multiples, nous n'aurions pas découvert que ces 2 gènes sont liés par l'évolution et que, par conséquent, il existe entre

eux une relation de fonction. Nous avons suggéré que la protéine TEX19 pourrait bloquer une réponse immunitaire dirigée par SECTM1 contre des types cellulaires dans lesquels existe une activité des transposons. Dans le 2ième projet, la structure 3D de l'homodimère de LBD du GRα a été revisitée. Nous avons établi que l'architecture actuellement acceptée [Bledsoe R.K. *et al*, 2002] serait vraisemblablement un artefact de contact cristallin alors qu'un homodimère de structure alternative amenant à l'interface des 2 monomères l'hélice 9 du LBD aurait plus de chance d'être biologique. Pour parvenir à cette conclusion, nous avons mis en œuvre des calculs d'énergie libre de liaison et un test statistique de sur-représentation de résidus conservés à l'interface d'assemblage. L'identification des résidus conservés a nécessité la construction d'un alignement multiple de séquences. L'éclairage évolutif contribué par l'alignement apporte un argument de poids pour distinguer les interfaces biologiques des artefacts de contacts cristallins. Certains outils comme EPPIC ont d'ailleurs été développés pour automatiser cette procédure [Duarte J.M. *et al*, 2012]. Si l'architecture alternative apH9 de l'homodimère de GRα était confirmée expérimentalement, la biologie structurale du récepteur, voire celle des oxo-stéroidiens, devrait être reconsidérée.

Connaître la séquence, la structure et la fonction du protéome humain sera un défi du XXIème siècle. Selon les dernières estimations, le génome humain coderait pour 20.000 polypeptides [Southan C., 2017 ; Kim M.-S. *et al*, 2014]. Cependant, ce nombre fluctue encore en raison de la détection des petits cadres ouverts de lecture qui codent pour des polypeptides de longueur inférieure à 100 résidus [Southan C., 2017]. Pour chaque protéine codée par les 20.000 gènes du génome humain, la séquence, la structure, et la fonction sont cruciales à obtenir. Cependant, ces 3 informations s'obtiennent avec des degrés de difficulté et des pas différents. Premièrement, la mise à disposition des séquences chez des organismes non-modèles a longtemps été un facteur limitant des études phylogénétiques. Or, grâce à la rapidité de séquençage des génomes, cette disponibilité a connu une accélération fulgurante. Bien qu'il y ait maintenant pléthore de séquences et d'espèces dans les banques de données, le phylogénéticien veut toujours plus d'espèces pour comprendre comment s'opère la sélection des séquences polypeptidiques pour la conquête des milieux naturels, *e.g.* les hémoglobines à forte affinité pour l'oxygène chez les espèces qui vivent à très haute altitude [Natarajan C. *et al*, 2018]. De plus, la construction d'un alignement multiple de qualité et l'obtention d'un profil de conservation des résidus, c'est-à-dire exploitables pour la recherche, constitue toujours un travail laborieux et semé de pièges (hétérogénéité des banques de séquences, redondance des séquences, erreurs de prédictions de séquences –indel-, extensions N- et C-terminales, erreurs

d'annotation, complexité taxonomique, synonymes d'espèces et de noms de gènes, paralogies, isoformes, variants d'épissage, etc …). Deuxièmement, si les méthodes de résolution structurale (RMN, cristallographie, cryo-EM) peuvent déterminer le repliement de nombreuses protéines ou domaines, elles se heurtent encore à l'écueil des régions intrinsèquement désordonnées. Par exemple, le GRα possède une région N-terminale de 420 résidus dont la résolution structurale est réfractaire. Cette région est cependant nécessaire pour l'activité de contrôle d'expression des gènes cibles du récepteur. Signalons qu'il n'y a à priori pas d'obstacle à étudier la conservation des régions intrinsèquement désordonnées par l'alignement multiple de séquences homologues à condition que celles-ci soient de bonne qualité [Bianchetti L. *et al*, 2005]. L'abondance de séquences provenant d'espèces proches pourrait aider cette analyse quand bien même ces régions montreraient une variabilité entre grands groupes taxonomiques. Cette approche apporterait une information dont on aurait tort de se priver. De plus, certaines études en cryo-EM menées sur des macro-complexes transcriptionnels impliquant ERα montrent que la forme globale des régions intrinsèquement désordonnées est visible à faible résolution, *i.e.* ~25 Å [Yi P. *et al*, 2015]. Enfin, certaines molécules pharmaceutiques ciblent les régions désordonnées N-terminales des RNs pour produire un effet physiologique [Banuelos C.A. *et al*, 2016] ce qui prouve toute leur importance. Troisièmement, dans la littérature, il est difficile d'obtenir une estimation globale sur le pourcentage des 20.000 gènes humains dont la fonction est connue tant la notion de fonction est déclinable à différents niveaux (rôle moléculaire, processus biologique, localisation cellulaire, intéraction avec un partenaire, oligomérisation, …). De plus, de nombreux gènes humains ne sont connus que par leur sur-expression en cancer mais leur fonction est en fait inconnue. Fin 2018, la banque SwissProt répertoriait 20.413 enregistrements de séquences protéiques humaines (voir https://www.uniprot.org). C'est sur cette information que s'appuie l'estimation du nombre de gènes chez l'humain [Southan C., 2017]. Si l'élucidation de la fonction de chaque protéine prendra encore de nombreuses décennies, la génomique aura permis de tendre vers la connaissance complète des séquences primaires polypeptidiques au bémol près de certaines erreurs classiques (localisation du début de traduction incorrecte, petits polypeptides indétectables, etc …). Signalons que des approches protéomiques par spectrométrie de masse et séquençage direct de polypeptides ont été menées pour affiner l'annotation du génome humain sur une échelle globale, tissu par tissu et pour une résolution de polypeptides allant jusqu'à 6 acides-aminés [Kim M.-S. *et al*, 2014]. De plus, les variants d'épissage apportent un niveau supplémentaire de complexité au protéome et reposent à chaque isoforme la question séquence, structure, fonction, *e.g.* le variant GRβ dont les hélices

11, 12 du LBD et le domaine F sont remplacés par une séquence alternative de 15 résidus inhibe la transcription des gènes cibles du GRα selon un mécanisme inconnu [Min J. *et al*, 2018]. Pour connaître les variants d'épissage, un instrument de séquençage à haut-débit produisant de longues lectures a longtemps fait défaut. Quand bien même des approches comme le séquençage d'EST (Etiquettes de transcrit) et HTC (cDNA à haut-débit) [Kawai J. *et al*, 2001] ont apporté une information massive de séquences transcrites dans Genbank, celles-ci étaient soit parcellaires et de mauvaise qualité (rétention d'intron, taux d'erreur élevés de base) soit au coût de séquençage très élevé (Séquenceur Roche 454) [Hampton M.H. *et al*, 2011]. Prochainement, la mise sur le marché d'un séquenceur PacBio dédié au séquençage d'ARNm en haut-débit est attendue pour résoudre le problème de longueur, qualité des reads et coût de séquençage.

Nous l'avons vu, l'étude de l'évolution apporte des informations cruciales pour comprendre la relation séquence, structure, fonction des protéines. Cependant, comme toute science, elle possède ses difficultés méthodologiques et ses limites. Premièrement, elle utilise des données de séquences protéiques de qualité trés hétérogènes (prédictions par homologie, prédictions *ab initio*, traduction de cDNA clonés) obtenues par traduction bioinformatique de séquences nucléotidiques elles-mêmes de qualités trés hétérogènes (ébauches de génomes – *drafts* -, séquences génomiques terminées, cDNA clonés, cDNA haut-débits). Deuxièmement, le transfert de fonction d'une protéine connue à un homologue tel qu'il est pratiqué par des outils automatiques demande beaucoup de vérifications pour transformer une prédiction en connaissance. Au niveau séquence, cette attribution de fonction est d'autant plus difficile à réaliser s'il y a homologie distante. La disponibilité de la structure est alors un atout majeur. Troisièmement, tandis que l'homologie indique une origine commune et des caractéristiques moléculaires partagées, des substitutions de résidus peuvent modifier significativement les fonctions de 2 protéines homologues. Dans notre étude de l'évolution des gènes *Tex19*, *Sectm1*, ERα et GRα, nous avons été confrontés aux difficultés et aux limites de l'analyse d'évolution de séquence. Ces 4 protéines sont codées par des génomes de métazoaires. Or, chez les eucaryotes, les régions intergéniques longues et les introns sont à l'origine de toutes sortes d'erreurs bioinformatiques de prédiction de polypeptides (fusion ou troncature de séquences, traduction de régions non-codantes, extensions ou délétions en extrémités N- et C-terminales etc …) [Mathé C. *et al*, 2002; Bianchetti L. *et al*, 2005]. Pour pallier à ces problèmes, la méthode d'annotation de génome la plus largement utilisée repose sur la comparaison de séquence et l'homologie. Une banque de polypeptides de référence, si possible des cDNA clonés au

laboratoire, servent à déterminer la localisation des gènes sur les génomes et prédire correctement les séquences protéiques homologues par comparaison de séquences. Il faut bien voir que la disponibilité d'un cDNA constitue un avantage majeur pour la recherche d'homologues et la qualité des séquences prédites qui serviront à construire les alignements multiples. Pour étudier l'évolution des protéines codées par les gènes *Tex19*, *Sectm1*, ERα et GRα, une recherche exhaustive des séquences polypeptidiques homologues prédites sur les génomes a été mise en oeuvre. Tracer au mieux l'histoire évolutive d'une protéine et connaître au mieux l'articulation de ses régions conservées et variables requièrent une collecte complète des homologues dans les banques dont les enregistrements augmentent rapidement (voir Ressources et Méthodes). L'étendue taxonomique et la complexité de la famille de protéine à étudier (paralogies) sont corrélées à la difficulté de rassembler les homologues disponibles dans les banques et maintenir cet ensemble à jour. Pour la collecte des protéines TEX19 et SECTM1, notre étude phylogénétique a été facilitée par la spécificité aux mammifères placentaires de ces 2 gènes ce qui a réduit l'étendue taxonomique de la recherche d'homologues. Cependant, elle a été compliquée par le fait que les prédictions de séquences protéiques n'étaient pas encore disponibles dans les banques. En effet, nos recherches de similarité de séquences TEX19 et SECTM1 ont été menées avant même que les génomes des euthériens n'aient été annotés par les outils bioinformatiques du NCBI et de l'EBI. Nous avons profité du fait que les ébauches de séquences génomiques étaient disponibles dans les banques nucléotidiques pour chercher les cadres de lecture ouverts des 2 gènes. Pour *Tex19*, la recherche de similarité de séquence sur les génomes a été facilitée d'une part par la disponibilité de cDNA humain et souris [Wang P.J. *et al*, 2001; Kuntz S. *et al*, 2008] et d'autre part par le fait que toute la séquence codante était portée par 1 seul exon. Pour *Sectm1*, la disponibilité d'un cDNA humain [Slentz-Kesler K.A. *et al*, 1998] et la conservation de la protéine (domaine Ig) a aidé la prédiction de la séquence des orthologues en dépit d'un morcèlement de la séquence codante sur 4 exons. De plus, les outils bioinformatiques du NCBI avaient rendu disponibles certaines prédictions de polypeptides mais avec des problèmes d'extensions en N- et C-terminus comme il apparaît sur le schéma de l'alignement de SECTM1 (voir Figure 16d). Pour ERα et GRα, ces 2 gènes existent des poissons à l'homme [Menuet A. *et al*, 2002; Baker M.E. *et al*, 2013] ce qui a rendu la collecte d'homologues plus laborieuse car taxonomiquement plus étendue que celle de TEX19 et SECTM1. Les produits polypeptidiques ERα et GRα issus de l'annotation bioinformatique des génomes avait été rendus disponibles pour environ 100 espèces (poissons, oiseaux, reptiles, mammifères). En utilisant les séquences protéiques traduites des cDNA

d'ERα et GRα humains et souris comme références, nous avons constaté que les LBDs prédits à partir des génomes d'organismes non-modèles étaient de bonne qualité. Au premier abord, cette qualité de séquence prédite peut surprendre car la structure des gènes qui codent les RNs est morcelée sur des centaines de kb et pose un problème majeur pour la prédiction bioinformatique de polypeptide dès qu'il s'agit d'un organisme pour lequel des cDNA ne sont pas disponibles. Signalons que la structure du gène GR humain se compose de 10 exons qui s'étendent sur 110 kb et que la séquence codant le LBD de la forme alpha est portée par 4 exons [Zhou J. et Cidlowski J.A., 2005]. Chez la souris, la structure du gène ERα s'étend sur plus de 200 kb et le LBD est codé par 5 exons espacés entre eux de 4 à 57kb [Swope D.L. *et al*, 2002]. Etant donné la complexité génomique d'ERα et GRα et la qualité des séquences de LBDs que nous avons manipulées, il ne fait aucun doute que ces polypeptides aient été prédits par des outils basés sur l'homologie de séquence. La qualité de séquence prédite des LBDs au travers des 100 espèces analysées est également au crédit de la conservation du LBD. Il en aurait été autrement si l'objet de notre étude avait été le domaine N-terminal (variabilité de séquence, morcèlement des exons, longueur de 180 et 420 acides-aminés chez ERα et GRα respectivement). En conclusion, la simplicité du gène *Tex19* et la disponibilité de cDNA pour les 4 protéines TEX19, SECTM1, ERα et GRα ont constitué des atouts majeurs pour la qualité de nos études d'évolution de séquences.

Un champ fondamental de l'étude de l'évolution est de comprendre comment apparaissent de nouvelles séquences, de nouvelles structures et de nouvelles fonctions. Nous l'avons vu, les méthodes classiques de recherche de similarité de séquences et de reconstruction phylogénétique de famille de protéines sont toutes basées sur la comparaison de séquences homologues qui dérivent d'un ancêtre commun. Par conséquent, ces méthodes ne peuvent détecter que ce qui est déjà connu. Il est largement admis que l'expansion de familles de gènes codants des protéines homologues est le résultat d'évènements de duplication. Les mécanismes de duplication sont de 2 types et mettent en jeu soit l'ADN génomique soit un ARN comme intermédiaire [Patthy L., 2008]. Dans le premier cas, la duplication de segments chromosomiques peut concerner une région mineure comme un exon ou un gène, ou bien une région majeure comme une bande ou un bras de chromosome. De plus, des duplications de génomes entiers ont été rapportées comme chez les poissons téléostéens [Glasauer S.M. et Neuhauss S.C., 2014]. Ces évènements sont attribués à des partages ou "*crossing-over*" inégaux du matériel génétique pendant la méiose et ont donc lieu dans les organes reproducteurs, testicule et ovaire chez les animaux. Dans le deuxième cas, appelé rétroposition ou

rétroduplication, un ARN produit par la transcription d'un gène parent est rétro-transcrit – ce qui nécessite la présence d'une transcriptase inverse – et le produit de rétro-transcription s'intègre dans le génome. Pour que cette duplication produise un gène fonctionnel, une région régulatrice de transcription en 5' du rétro-transcrit est requise [Kaessmann H., 2010]. Notons que les expériences de cistromique (ChIP-seq) ont montré que les sites de fixation de facteurs de transcription sur le génome se comptent en dizaines de milliers alors que les sites physiologiques seraient une minorité. Cette abondance de sites de fixation pourrait servir la transcription de gènes rétroposés si le hasard jouait un rôle dans la mise en proximité des 2 éléments génétiques. Enfin, il est considéré que l'apparition *de novo* de gènes codants des protéines fonctionnelles à partir d'un ADN de séquence aléatoire par association de triplets de nucléotides codants est extrêmement faible [Patthy L, 2008]. Néanmoins, le séquençage et l'annotation des génomes ont montré que 10 à 30% des gènes ne présentent de similarité de séquence à aucun autre connu si bien que les protéines codées pourraient effectivement être apparues *de novo* [Bornberg-Bauer E. *et al*, 2015]. Au cours de cette thèse, la question de la création génétique *de novo* s'est posée pour la protéine TEX19. En effet, il n'a pas été possible d'assigner une fonction à TEX19 par comparaison de séquence. Que ce soit par BLAST, PSI-BLAST (10 itérations) ou d'autres méthodes de recherche d'homologie distantes, aucune similarité de séquence n'a pu être détectée. De plus, des recherches de similarité plus poussées en utilisant les produits de traduction du brin réverse du cDNA codant la protéine TEX19 sont demeurées sans résultat. *Tex19* est un gène orphelin [Tautz D. et Domazet-Loso T., 2011]. D'une part se pose l'intriguante question de son origine, d'autre part se pose naturellement la question de sa fonction. Chez l'adulte humain mâle, l'expression de *Tex19* est spécifique au testicule ce qui n'est pas sans rappeler que cette caractéristique est partagée avec des gènes formés dans cet organe et dont les protéines codées sont antigéniques en raison de leur nouveauté [Kaessmann H., 2010]. Se pourrait il que *Tex19* soit un gène apparu *de novo* d'une séquence aléatoire d'ADN chez l'ancêtre des euthériens ? Il faut remarquer que la séquence de son domaine protéique le plus conservé est courte, à peine 50 résidus ce qui peut plaider en faveur de cette hypothèse. Une hypothèse alternative serait que *Tex19* ait été créé par duplication d'un gène ancestral suivie d'une divergence de séquence si grande qu'il aurait perdu toute ressemblance avec le gène parent. Puisque rien du gène parent ne subsiste, ce scénario rapprocherait l'origine de *Tex19* à la création *de novo*. En outre, la divergence de séquence semble être un processus lent. En effet, la traduction bioinformatique en 3' du cDNA humain codant TEX19 restitue la séquence de la région C-terminale perdue chez les *Haplorrhini*

(information non-montrée). Quand bien même la séparation des *Haplorrhini* et des *Strepshirrini* est estimée à 74 millions d'années [Pozzi L. *et al*, 2014], la similarité des régions C-terminales de TEX19 entre les séquences de ces 2 groupes taxonomiques est encore visible. En d'autres termes, 74 millions d'années n'ont pas effacé l'homologie de séquence entre le C-terminus non-traduit de TEX19 chez les *Haplorrhini* et le C-terminus de TEX19 des *Strepshirrini*. Un exemple connu d'homologie distante entre 2 séquences protéiques ayant fortement divergées est celui de la thermolysine bactérienne (PDB:1HYT) et de la néprilysine humaine (PDB:1DMT) [Bianchetti L. *et al*, 2002]. Bien que ces 2 enzymes catalysent la même réaction chimique, leur pourcentage d'identité est si faible que seuls les résidus des signatures catalytiques s'alignent. Cependant, la comparaison structurale des 2 protéines montre que les résidus catalytiques et certaines structures secondaires se superposent parfaitement ce qui confirme l'homologie. Malgré les milliards d'années d'évolution qui séparent la bactérie de l'homme, il reste encore trace de l'homologie entre les 2 séquences. Si l'homologie de séquence et de structure a persisté entre la thermolysine bactérienne et la néprilysine humaine pendant des milliards d'année, *Tex19* au contraire n'est homologue à rien. Ce dernier argument serait donc en faveur d'une création *de novo*, de séquence, structure et fonction.

G.5. Remerciements

J’adresse particulièrement ma gratitude au Dr. Olivier Poch et au Prof. Annick Dejaegere pour avoir tous deux soutenu ma candidature au doctorat par VAE. De plus, je remercie d’une part le Dr. O. Poch pour les projets menés tous les deux en génomique comparative et d’autre part le Prof. A. Dejaegere et le Dr. R.H. Stote pour les projets réalisés ensemble en modélisation moléculaire. Merci à tous les 3 pour leur patience, leur pédagogie, leur supervision et leur sens du partage de la connaissance. J’adresse mes remerciements à mes collègues, les Drs. Yasmine Chebaro, Isabelle Lebars, Katia Zanier, Christelle Thibault-Carpentier et Corinne Di trani-Zimmerman pour leurs encouragements continus à soutenir une thèse doctorale. Enfin, je souhaite témoigner ma reconnaissance à mon employeur, l’INSERM, qui m’a accompagné positivement dans ma démarche VAE.

G.6. Contributions

G.6.1. Projet Tex19/Sectm1

Laurent Bianchetti	Question biologique Bibliographie Planification de projet Analyses bioinformatiques Comparaison de séquences (dotplot, 2x2, régions chromosomiques) Recherche de similarité de séquence dans les banques Construction d'alignement multiple Phylogénèse moléculaire et annotation des arbres Tests statistiques dans R Construction de modèles moléculaires par homologie Analyse RNA-seq Scripting PERL Interprétation des résultats Rédaction manuscrit & figures Soumission article Lettre à l'éditeur Réponse aux critiques et éditions de l'article Présentation d'un poster (18ème Journée de la Ligue)
Prof. Annick Dejaegere & Dr.Roland Stote	Expertise en modélisation moléculaire Contrôle-qualité des analyses, résultats et interprétations Discussions Lecture/Correction du manuscrit Aide à l'écriture des lettres aux éditeurs
Dr. Olivier Poch & Prof. Odile Lecompte	Expertise en évolution de séquence Contrôle-qualité des analyses, résultats et interprétations Discussions Lecture/Correction du manuscrit
Prof. Stéphane Viville & Dr. Yara Tarabay	Expertise biologique sur Tex19 Contrôle-qualité des analyses, résultats et interprétations Discussions RT-PCR chez la vache et le taureau Lecture/Correction du manuscrit

G.6.2. Projet GRα/ERα

Laurent Bianchetti	Question biologique Bibliographie Planification de projet Analyses bioinformatiques Recherche de contacts protéine-protéine Minimisation d'énergie avec CHARMM Dynamiques moléculaires (DM) 10 ns avec NAMD Energie libre de liaison MM/PBSA, Recherche de similarité de séquence dans les banques Construction d'alignement multiple Tests statistiques dans R Graphiques sous Excel et Gnuplot Scripting PERL Encadrement de A. Smertina, A. Defosset et B. Wassmer Rédaction d'une ébauche de manuscrit Construction de figures
Prof. Annick Dejaegere & Dr. Roland Stote	Question biologique Supervision globale du projet Bibliographie Contrôle-qualité des analyses, résultats et interprétations Formation de L. Bianchetti à la modélisation moléculaire Supervision de A. Smertina, A. Defosset et B. Wassmer et M. Tiberti Lecture/Correction/Ecriture du manuscrit Lettre à l'éditeur et réponse aux critiques Conférence et poster (Lyon 2017 intéraction protéine-protéine, Crête EMBO sept 2018)
Marion Tiberti	Simulation de DM de 10 ns du LBD d' ERα
Anna Smertina	Minimisation d'énergie du GRα chauve-souris MM/PBSA sur structure minimisée
Audrey Defosset	Alignement multiple du LBD d'ERα
Bianca Wassmer	Simulation de DM 100 ns du LBD de GRα

G.7. Références bibliographiques

Abascal F, Zardoya R and Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. **2005**; 21(9): 2104-2105.

Allen MA, Goh F, Burns BP and Neilan BA. Bacterial, archaeal and eukaryotic diversity of smooth and pustular microbial mat communities in the hypersaline lagoon of Shark Bay. Geobiology. **2009**; 7: 82-96.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local alignment search tool. J. Mol. Biol. **1990**; 215(3): 403-410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. **1997**; 25(17): 3389-3402.

Altschul SF and **Koonin EV**. Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. Trends Biochem. Sci. **1998**; 23(11): 444-447.

Apweiler R, Bairoch A, Wu CH. Protein sequence databases. Current Opinion in Chemical Biology. **2004**; 8:76-80.

Baker ME, Funder JW and Kattoula SR. Evolution of hormone selectivity in glucocorticoid and mineralocorticoid receptors. Journal of Steroid Biochemistry & Molecular Biology. **2013**; 137: 57-70.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S *et al*. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. **2012**; 483(7391): 603-607.

Banuelos CA, Tavakoli I, Tien AH, Caley DP, Mawji NR, Liz Z *et al*. Sintokamide A is a novel antagonist of androgen receptor that uniquely binds activation function-1 in its amino-terminal domain. J. Biol. Chem. **2016**; 291(42): 22231-22243.

Benton MJ and **Twitchett RJ**. How to kill (almost) all life: the end-Permian extinction event. Trends in Ecology and Evolution. **2003**; 18(7): 358-365.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H *et al*. The protein data bank. Nucleic Acids Research. **2000**; 28(1): 235-242.

Bernardi G, Wiley EO, Mansour H, Miller MR, Orti G, Haussler D *et al*. The fishes of genome 10K. Marine Genomics. **2012**; 7: 3-6.

Bianchetti L, Oudet C and Poch O. M13 endopeptidases: New conserved motifs correlated with structure, and simultaneous phylogenetic occurence of Phex and the bony fish. Proteins: Structure, Function and Genetics. **2002**; 47: 481-488.

Bianchetti L, Thompson JD, Lecompte O, Plewniak F and Poch O. vALId: validation of protein sequence quality based on multiple alignment data. Journal of Bioinformatics and Computational Biolology. **2005**; 3(4): 929-947.

Bianchetti L, Kieffer D, Féderkeil R and Poch O. Increased frequency of single base substitutions in a population of transcripts expressed in cancer cells. BMC Cancer. **2012**; 12(509): 1-13.

Bianchetti L, Tarabay Y, Lecompte O, Stote R, Poch O, Dejaegere A and Viville S. Tex19 and Sectm1 concordant molecular phylogenies support coevolution of both eutherian specific genes. BMC Evolutionary Biology. **2015**; 15(222): 1-15.

Bianchetti L, Wassmer B, Defosset A, Smertina A, Tiberti M, Stote RH and Dejaegere A. Alternative dimerization interfaces in the glucocorticoid receptor-α ligand binding domain. Biochim. Biophys. Acta General Subjects. **2018**. 1862(8): 1810-1825.

Billas I and **Moras D**. Allosteric controls of nuclear receptor function in the regulation of transcription. J. Mol. Biol. **2013**; 425(13):2317-2329.

Bledsoe RK, Montana VG, Stanley TB, Delves CJ, Apolito CJ, McKee DD *et al*. Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition. **Cell**. 2002; 111(1): 93-105.

Bledsoe RK, Madauss KP, Holt JA, Apolito CJ, Lambert MH, Pearce KH *et al*. A ligand-mediated hydrogen bond network required for the activation of the mineralocorticoid receptor. The Journal of Biological Chemistry. **2005**; 280: 31283-31293.

Bookout AL, Jeong Y, Downes M, Yu RT, Evans RM and Mangelsdorf DJ. Anatomical profiling of nuclear receptor expression reveals a hierarchical transcriptional network. Cell. **2006**; 126: 789-799.

Bornberg-Bauer E, Schmitz J and Heberlein M. Emergence of de novo proteins from 'dark genomic matter' by 'grow slow and moult'. Biochem. Soc. Trans. **2015**; 43(5):867-873.

Bosnjak I, Bojovic V, Segvic-Bubic T and Bielen A. Occurrence of protein disulfide bonds in different domains of life: a comparison of proteins from the Protein Data Bank. Protein Eng. Des. Sel. **2014**; 27(3): 65-72.

Bougarne N, Paumelle R, Caron S, Hennuyer N, Mansouri R, Gervois P *et al*. PPARα blocks glucocorticoid receptor α-mediated transactivation but cooperates with the activated glucocorticoid receptor-α for transrepression on NF-κB. PNAS. **2009**; 106(18): 7397-7402.

Bourhis E, Wang W, Tam C, Hwang J, Zhang Y, Spittler D *et al*. Wnt antagonists bind through a short peptide to the first beta-propeller domain LRP5/6. Structure. **2011**; 19:1433-1442.

Braberg H, Webb BM, Tijioe E, Pieper U, Sali A, Madhusudhan MS. SALIGN: a web server for alignment of multiple protein sequences and structures. Bioinformatics. **2012**; 28(15):2072-2073.

Brokx RD, Bolewska-Pedyczak E and Gariépy J. A stable human p53 heterotetramer based on constructive charge interactions within the tetramerization domain. The Journal of Biological Chemistry. **2003**; 278(4): 2327-2332.

Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B *et al*. CHARMM: the biomolecular simulation program. J. Comput Chem. **2009**; 30(10):1545-614.

Brünger AT, Karplus M. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. Proteins. **1988**; 4(2):148-156.

Brzozowski AM, Pike AC, Dauter Z, Hubbard RE, Bonn T, Engström O *et al*. Molecular basis of agonism and antagonism in the oestrogen receptor. **1997**; 389(6652): 753-758.

Buchan DW, Minneci F, Nugent TC, Bryson K and Jones DT. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Research. **2013**; 41(Web server issue): W349-357.

Burns BP, Goh F, Allen M and Neilan BA. Microbial diversity of extant stromatolites in the hypersaline marine environment of Shark Bay, Australia. Environ Microbiol. **2004**; 6(10): 1096-1101.

Capitani G, Duarte JM, Baskaran K, Bliven S and Somody JC. Understanding the fabric of protein crystals: computational classification of biological interfaces and crystals contacts. Bioinformatics. **2016**; 32(4):481-489.

Charmandari E, Raji A, Kino T, Ichijo T, Tiulpakov A, Zachman K *et al*. A novel point mutation in the ligand-binding domain (LBD) of the human glucocorticoid receptor (hGR) causing generalized glucocorticoid resistance: the importance of the C-terminus of hGR LBD in confering transactivational activity. J. Clin. Endocrinol. Metab. **2005**; 90(6): 3696-3705.

Chen VB, Arendal WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ *et al*. Mol-Probity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica. **2010**; D66:12-21.

Crooks GE, Hon G, Chandonia JM and Brenner S. WebLogo: A sequence logo generator. Genome Research. **2004**; 14: 1188-1190.

Darriba D, Taboada GL, Doallo R and Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. **2011**; 27(8): 1164-1165.

Robertson S, Allie-Reid F, Vanden Berghe W, Visser K, Binder A, Africander D *et al*. Abrogation of glucocorticoid receptor dimerization correlates with dissociated glucorticoid behavior of compound a. J. Biol. Chem. 2010; 285(11): 8061-8075.

Dolinksy TJ, Nielsen JE, McCammon JA and Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Research. **2004**; 32(Web server issue): W665-667.

Drubin DG. Any jackass can trash a manuscript, but it takes good scholarship to create one (how *MBoC* promotes civil and constructive peer review). Molecular Biology of the Cell. **2011**; 22: 525-527.

Duarte JM, Srebniak A, Schärer MA and Capitani G. Protein interface classification by evolutionary analysis. BMC Bioinformatics. **2012**; 13(334): 1-6.

Eiler S, Gangloff M, Duclaud S, Moras D and Ruff M. Overexpression, purification and crystal structure of native ERa LBD. Protein Expr. Purif. **2001**; 22: 165-173.

Fagerberg L, Hallstro BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J *et al*. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Molecular & Cellular Proteomics. **2014**; 13(2): 397-406.

Fariselli P, Rossi I, Capriotti E and Casadio R. The WWWH of remote homolog detection: the state of the art. Briefings in bioinformatics. **2006**; 8(2): 78-87.

Federhen S. The NCBI taxonomy database. Nucleic Acids Research. **2012**; 40: Database issue: D136-D143.

Fijak M and **Meinhardt A**. The testis in immune privilege. Immunol Rev. **2006**; 213: 66-81.

Foster JS, Green SJ, Ahrendt SR, Golubic S, Reid RP, Hetherington KL and Bebout L. Molecular and morphological characterization of cyanobacterial diversity in the stromatolites of Highborne Cay, Bahamas. The ISME Journal. **2009**; 3: 573-587.

Fuentes-Prior P, Rojas A, Hagler AT and Estebanez-Perpina E. Diversity of quaternary structures regulates nuclear receptor activities. Trends Biochem Sci. **2018**; 18:1-5.

Gaillard T, Schwarz BB, Chebaro Y, Stote RH and Dejaegere A. Protein structural statistics with PSS. J Chem. Inf. Model. **2013**; 53(9): 2471-2482.

Genome 10K community of scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. Journal of Heredity. **2009**; 100(6): 659-674.

Germain P and **Bourguet W**. Dimerization of nuclear receptors. Methods in Cell Biology. **2013**; 117: 21-41.

Germain P, Staels B, Dacquet C, Spedding M and Laudet V. Overview of nomenclature of nuclear receptors. Pharmacological reviews. **2006**; 58:685-704.

Gilad Y, Man O and Glusman G. A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Research. **2005**; 15: 224-230.

Glasauer SM and Neuhauss SC. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol. Genet. Genomics. **2014**; 289(6): 1045-1060.

Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M *et al*. Essential genes of a minimal bacterium. PNAS. **2006**; 103(2):425-430.

Goh C-S, Bogan AA, Joachimiak M, Walther D and Cohen FE. Co-evolution of proteins with their interaction partners. J.Mol.Biol. **2000**; 299: 283-293.

Gudhka RK, Neilan BA and Burns BP. Adaptation, ecology, and evolution of the halophilic stromatolite archaeon Halococcus hamelinensis inferred through genome analyses. Archaea. **2015**; 241608: 1-11.

Hampton M, Melvin RG, Kendall AH, Kirkpatrick BR, Peterson N and Andrews MT. Deep sequencing the transcriptome reveals seasonal adaptative mechanisms in a hibernating mammal. Plos One. **2011**; 6(10): e27021.

Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E and Hutchison GR. Avogadro : an advanced semantic chemical editor, visualization, and analysis platform. Journal of Cheminformatics. **2012**; 4:17.

Hedges SB. The origin and evolution of model organisms. Nat Rev Genet. **2002**; 3: 838-849.

Hinz M and **Scheidereit C**. The IkB kinase complex in NFkB regulation and beyond. EMBO reports. **2013**; 1-15.

Howie D, Garcia Rueda H, Brown MH and Waldmann H. Secreted and transmembrane 1A is a novel co-stimulatory ligand. Plos One. **2013**; 8(9):e73610.

Hu B, Jin J, Guo AN, Zhang H, Luo J and Gao G. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics. **2015**; 31(8): 1296–1297.

Hubbard SJ and **Thornton JM**. NACCESS: Atomic solvent accessible area calculations. *Unpublished*. **1992**; http://wolf.bms.umist.ac.uk/naccess/

Hudson WH, Vera IMS, Nwachukwu JC, Weikum ER, Herbst AG, Yang Q, Bain DL, Nettles KW, Kojetin DJ and Ortlund EA. Cryptic glucocorticoid receptor-binding sites pervade genomic NF-kB response elements. Nature Communication. **2018**; 9(1): 1337

I5K community of scientists. The i5K initiative: advancing arthropod genomics for knowledge, Human Health, Agriculture, and the Environment. Journal of Heredity. **2013**; 104(5): 595-600.

Jeffrey PD, Gorina S and Pavletich NP. Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7 angstroms. Science. **1995**; 267:1498-1502.

Jendroszek A, Malte H, Overgaard CB, Beedholm K, Natarajan C, Weber RE, Storz JF and Fago A. Allosteric mechanisms underlying the adaptative increase in hemoglobin-oxygen affinity of the bar-headed goose. Journal of Experimental Biology. **2018**; 221:1-10.

Jiang G, Lee U and Sladek FM. Proposed mechanism for the stabilization of nuclear receptor DNA binding via protein dimerization. Mol. Cell. Biol. **1997**; 17(11): 6546-6554.

Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE *et al*. Improvements to the APBS biomolecular solvation software suite. Protein Sci. **2018**; 27(1):112-128.

Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Research. **2010**; 20:1313-1326.

Katoh K, Misawa K, Kuma K and Miyata T. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics. **2017**; 1-7.

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y *et al*. Functional annotation of a full-length mouse cDNA collection. Nature. **2001**; 409:685-690.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM *et al*. The human genome browser at UCSC. Genome Research. **2002**; 12(6): 996-1006.

Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R et al. A draft map of the human proteome. Nature. **2014**; 0:1-7.

Klijn C, Durinck S, Stawiski EW, Peter M Haverty PM, Jiang Z, Liu H *et al*. A comprehensive transcriptional portrait of human cancer cell lines. Nature Biotechnology. **2015**; 33(3): 306-312.

Kobe B, Guncar G, Buchholz R, Huber T, Bohumil M, Cowieson N *et al*. Crystallography and protein-protein interactions: biological interfaces and crystal contacts. Biochemical Society Transactions. **2008**; 36(6): 1438-1441.

Kodama Y, Shumway M, Leinonen R, International Nucleotide Database Collection. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Research. **2012**; 40(Database issue): D54-D56.

Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L *et al*. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc. Chem. Res. **2000**; 33: 889-897.

Krishna SS, Majumdar I and Grishin NV. Structural classification of zinc finger: survey and summary. Nucleic Acids Research. **2003**; 31(2):532-550.

Krissinel E and **Henrick K**. Inference of macromolecular assemblies from crystalline state. J Mol Biol. **2007**; 372: 774-797.

Kumar CS, Qureshi SF, Ali A, Satyanarayana ML, Rangaraju A, Venkateshwari A and Nallari P. Hidden magicians of genome evolution. Indian journal of medical research. **2013**; 137(6): 1052-1060.

Kuntz S, Kieffer E, Bianchetti L, Lamoureux N, Furhmann G and Viville S. Tex19, a mammalian-specific protein with a restricted expression in pluripotent stem cells and germ line.Stem cells. **2008**; 26(3): 734-744.

Lafont V, Schaefer M, Stote RH, Altschuh D and Dejaegere A. Protein-Protein recognition and interaction hot spots in an antigen-antibody complex: free energy decomposition identifies efficient amino-acids. Bioinformatics. **2007**; 67: 418-434.

Lam GK, Liao HX, Xue Y, Alam SM, Scearce RM, Kaufman RE, et al. Expression of the CD7 ligand K12 in human thymic epithelial cells: regulation by IFN-gamma. J. Clin. Immunol. **2005**; 25(1): 41-49.

Langmead B, Trapnell C, Pop M and Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. **2009**; 10:R25:1-10.

Lecompte O, Thompson JD, Plewniak F, Thierry JC and Poch O. Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene. **2001**; 270:17-30.

Letunik I and **Bork P**. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research. **2016**; 44(W1): W242-245.

Lindblad –Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S *et al*. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. **2011**; 478(7370): 476-482.

Livingstone CD and **Barton GJ**. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation.CABIOS. **1993**; 9(6): 745-756.

Lyman SD, Escobar S, Rousseau AM, Armstrong A and Fanslow WC. Identification of CD7 as a cognate of the human K12 (SECTM1) protein. J. Biol. Chem. **2000**; 275(5):3431-3437.

MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ *et al*. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem B. **1998**; 102(18): 3586-3616.

MacLennan M, Garcia-Canadas M, Reichmann J, Khazina E, Wagner G, Playfoot CJ *et al*. Mobilization of LINE-1 retrotransposons is restricted by Tex19.1 in mouse embryonic stem cells. eLIFE. **2017**; 6:e26152.

Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schütz G, Umesono K et al. The nuclear receptor superfamily: the second decade. Cell. **1995**; 83:835-839.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C *et al.* CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res. **2011**; 39: D225-229.

Marianayagam NJ, Sunde M and Matthews JM. The power of two: protein dimerization in biology. TRENDS in Biochemical Sciences. **2004**; 29(11): 618-625.

Mathé C, Sagot MF, Schiex T and Rouzé P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research. **2002**; 30(19):4103-4117.

McGuffin LJ, Bryson K and Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. **2000**; 16(4): 404-405.

McMahon SK, Pretorius CJ, Ungerer JPJ, Salmon NJ, Conwell LS, Pearen MA and Batch JA. Neonatal complete generalized glucocorticoid resistance and growth hormone deficiency caused by a novel homozygous mutation in helix 12 of the ligand binding domain of the glucocorticoid receptor (NR3C1). J. Clin. Endocrinol. Metab. **2010**; 95(1): 297-302.

Menuet A, Pellegrini E, Anglade I, Blaise O, Laudet V, Kah O and Pakdel F. Molecular characterization of three estrogen receptor forms in zebrafish: binding characteristics, transactivation properties, and tissue distributions. Bio. Reprod. **2002**; 66(6): 1881-92.

Mighell AJ, Smith NR, Robinson PA and Markham AF. Vertebrate pseudogenes. FEBS Letters. **2000**; 468:109-114.

Miller W, Drautz DI, Janecka JE, Lesk AM, Ratan A, Tomsho L *et al*. The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). Genome Research. **2009**; 19: 213-220.

Min J, Perera L, Krahn JM, Jewell CM, Moon AF, Cidlowski JA, Pedersen L. Probing dominant negative behavior of glucocorticoid receptor b through a hybrid structural and biochemical approach. Molecualr and Cellular Biology. **2018**; 3(8):e000453-17.

Mlewski EC, Pisapia C, Gomez F, Lecourt L, Rueda ES, Benzerara K *et al*. Characterization of pustular mats and related Rivularia-Rich laminations in Oncoids from the Laguna Negra lake (Argentina). Frontiers in Microbiology. **2018**; 9(996): 1-23.

Moffett A and **Loke C**. Immunology of placentation in eutherian mammals. Nat. Rev. Immunol. **2006**; 6(8): 584-594.

Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Katta HY, Mojica A *et al*. Genomes OnLine Database (GOLD) v.7: updates and new features. Nucleic Acids Research. **2018**; 1-11.

Nadal M, Prekovic S, Gallastegui N, Helsen C, Abella M, Zielinska K *et al*. Structure of the homodimeric androgen receptor ligand-binding domain. Nature communications. **2017**; 8(14388): 1-14.

Natarajan C, Jendroszek A, Kumar A, Weber RE, Tame RH, Fago A and Storz JF. Molecular basis of hemoglobin adaptation in the high-flying bar-headed goose. PLOS Genetics. **2018**; 1-19.

Nicolaides NC, Charmandari E, Chrousos GP and Kino T. Recent advances in the molecular mechanisms determining tissue sensitivity to glucocorticoids: novel mutations, circadian rhythm and ligand-induced repression of the human glucocorticoid receptor. BMC Endocr. Disord. **2014**; 14(71):1-12.

Nutman AP, Bennett VC, Friend CRL, Van Kranendonk MJ and Chivas AR. Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. Nature. **2016**; 537: 535-539.

OBrien SJ, Haussler D and Ryder O. The birds of genome 10K. GigaScience. **2014**; 3(32):1-4.

Ochoa D and **Pazos F**. Studying the co-evolution of proteins families with the Mirrortree web server. Bioinformatics. **2010**; 26(10): 1370-1371.

Ollinger R, Childs AJ, Burgess HM, Speed RM, Lundegaard PR, Reynolds N *et al.* Deletion of the pluripotency-associated Tex19.1 gene causes activation of endogenous retroviruses and defective spermatogenesis in mice. Plos genetics. **2008**; 4:e1000199.

Olsson MH, Sondergaard CR, Rostkowski M and **Jensen JH**. PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. J. Chem. Theory Comput. **2011**; 7(2): 525-537.

Patthy L. Protein evolution. Blackwell Publishing. Second edition. **2008**; 1-374

Pazos F and **Valencia A**. Protein co-evolution, co-adaptation and interactions. EMBO Journal. **2008**; 27(20): 2648-2655.

Pereto J. Controversies on the origin of life. International Microbiology. **2005**; 8:23-31.

Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E *et al*. Scalable molecular dynamics with NAMD. Journal of Computational Chemistry. **2005**; 26: 1781-1802.

Planells-Palop V, Hazazi A, Feichtinger J, Jezkova J, Thallinger G, Alsiwiehri NO *et al.* Human germ/stem cell-specific TEX19 influences cancer cell proliferation and cancer prognosis. Mol Cancer. **2017** ; 16(84): 1-18.

Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, *et al.* PipeAlign: a new toolkit for protein family analysis. Nucleic Acids Res. **2003**; 31(13): 3829-3832.

Pozzi L, Hodgson JA, Burrell AS, Sterner KN, Raaum RL and Disotell TR. Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. Mol. Phylogenet. Evol. **2014**; 75:165-183.

Pressman DM, Ganguly S, Schiltz RL, Johnson TA, Karpova TS and Hager GL. DNA binding triggers tetramerization of the glucocorticoid receptor in live cells. PNAS. **2016**; 113(29):8236-8241.

Pressman DM and **Hager GL**. More than meets the dimer: What is the quaternary structure of the glucocorticoid receptor ? Transcription. **2017**; 8(1):32-39.

Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P, Apostolov R *et al*. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. **2013**; 29(7):845-854.

Pruitt KD, Tatusova T, Klimke W and Maglott DR. NCBI Reference Sequences: current status, policy and new initiativesNucleic Acids Research. **2009**; Database issue; 37: D32-D36.

Ratman D, Vanden Berghe W, Dejager L, Libert C, Tavernier J, Beck IM and De Bosscher K. How glucocorticoid receptors modulate the activity of other transcription factors: A scope beyond tethering. Molecular and Cellular Endocrinology. **2013**; 380:41-54.

Raup DM. Biological extinction in Earth history. Science. **1986**; 231(4745): 1528-1533.

Reichmann J, Reddington JP, Best D, Read D, Ollinger R, Meehan RR and Adams IR. The genome-defence gene Tex19.1 suppresses LINE-1 retrotransposons in the placenta and prevents intra-uterine growth retardation in mice. Human molecular genetics. **2013**; 22(9): 1791-1806.

Renne PR, Zichao Z, Richards MA, Black MT and Basu AR. Synchrony and causal relations between permian-triassic boundary crises and siberian flood volcanism. Science. **1995**; 269: 1413-1416.

Rice P, Longden I and Bleasby A. The european molecular biology open software suite. Trends in Genetics. **2000**; 16(6): 276-277.

Robinson-Rechavet M, Carpentier AS, Duffraisse M and Laudet V. How many nuclear hormone receptors are there in the human genome ? Trends in Genetics. **2001**; 17(10):554.

Shiao MS, Liao BY, Long M and Yu HT. Adaptative evolution of the insulin two-gene system in mouse. Genetics. **2008**; 178: 1683-1691.

Shiau AK, Barstad D, Loria PM, Cheng L, Kushner PJ, Agard DA and Greene GL. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell. **1998**; 95:927-937.

Slentz-Kesler KA, Hale LP and Kaufman RE. Identification and characterization of K12 (SECTM1), a novel human gene that encodes a Golgi-associated protein with transmembrane and secreted isoforms. Genomics. **1998**; 47(3): 327-340.

Southan C. Last rolls of the yoyo: Assessing the human canonical protein count. F1000Research. **2017**; 6(448): 1-23.

Sprangers R, Velyvis A and Kay LE. Solution NMR of supramolecular complexes: providing new insights into function. Nature Methods. **2007**; 4(9): 697-703.

Swope DL, Harrell JC, Mahato D and Korach KS. Genomic structure and identification of a truncated variant message of the mouse estrogen receptor α gene. Gene. **2002**; 294:239-247.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M and Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution. **2011**; 28(10): 2731-2739.

Tarabay Y, Kieffer E, Teletin M, Celebi C, Van Monfoort A, Zamudio N *et al*. The mammalian-specific Tex19.1 gene plays an essential role in spermatogenesis and placenta-supported development. Hum Reprod. **2013**; 28(8): 2201-14.

Tautz D and **Domazet-Loso T**. The evolutionary origin of orphan genes. Nature Reviews Genetics. **2011**; 12(10): 692-702.

Thomas-Chollier M, Watson LC, Cooper SB, Pufall MA, Liu JS, Borzym K *et al*. A naturally occuring insertion of a single amino acid rewires transcriptional regulation by glucocorticoid receptor isoforms. Proc Natl Acad Sci USA. **2013**; 110(44): 17826-17831.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG. The Clustal_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research. **1997**; 25(4): 4876-4882.

Thorvaldsdottir H, Robinson JT and Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. **2013**; 14(2): 178-192.

UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Research. **2016**; 45: Database issue: D158-D169.

Valdar WSJ and **Thornton JM**. Conservation helps to identify biologically relevant crystal contacts. J Mol Biol. **2001**; 313: 399-416.

Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J *et al*. CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force-field. J. Comput. Chem. **2010**; 31(4):671-690.

Wang J, Yan Y, Garrett TPJ, Liu J, Rodgers DW, Garlick RL *et al*. Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. Nature. **1990**; 348: 411-418.

Waterhouse AM, Procter JB, Martin DMA, Clamp M and Barton GJ. Jalview version 2: a multiple sequence alignment editor and analysis workbench. Bioinformatics. **2009**; 25(9): 1189-1191.

Webb B and **Sali A**. Comparative protein structure modelling using Modeller. Current Protocols in Bioinformatics. John Wiley & Sons, Inc. **2016**; 5.6.1-5.6.37.

Weigel NL, Beck AC, Estes PA, Prendergast P, Altmann M, Christensen K and Edwards DP. Ligands induce conformational changes in the carboxyl-terminus of progesterone receptors which are detected by a site—directed antipeptide monoclonal antibody. Mol. Endo. **1992**; 6(10):1585-1597.

Weikum ER, De Vera IMS, Nwachukwu JC, Hudson WH, Nettles KW, Kojetin DJ and Ortlund EA. Tethering not required: the glucocorticoid receptor binds directly to activator protein-1 recognition motifs to repress inflammatory genes. Nucleic Acids Research. **2017**; 45(14):8596-8608.

Weikum ER, Knuesel MT, Ortlund EA and Yamamoto KR. Glucocorticoid receptor control of transcription: precision and plasticity via allostery. Nat. rev. Mol. Cell. Biol. **2017**; 18(3):159-174.

Wiederstein M and **Sippl MJ**. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. **2007**; 35 (Web server Issue): W407-W410.

Williams SP and **Sigler PB**. Atomic structure of progesterone complexed with its receptor. Nature. **1998**; 393: 392-395.

Wyffels J, King BL, Vincent J, Chen C, Wu CH and Polson SW. SkateBase, an elasmobranch genome project and collection of molecular resources for chondrichthyan fishes. F1000 Research. **2014**; 3(191):1-24.

Yi P, Wang Z, feng Q, Pintille GD, Foulds CE, Lanz RB et al. Structure of a biologically active estrogen receptor-coactivator complex on DNA. Molecular Cell. **2015**; 57:1047-1058.

Yudt MR and **Koide S**. Preventing estrogen receptor action with dimer-interface peptides. Steroids. **2001**; 66: 549-558.

Zhang Z, Burch PE, Cooney AJ, Lanz RB, Pereira FA, Wu J *et al*. Genomic analysis of the nuclear receptor family: new insights into structure, function, regulation, and evolution from the rat genome. Genome Research. **2004**; 14:580-590.

Zhou J and **Cidlowski JA**. The human glucocorticoïd receptor: one gene, multiple proteins and diverse responses. Steroids. **2005**; 70(5-7):407-417.

G.8. Glossaire

AR : Récepteur nucléaire aux androgènes.

BLAST: « *Basic Local Alignment Search Tool* ». Outil de recherche de similarité de séquence dans les banques.

CCLE: « *Cancer cell line encyclopedia* ». Ressource bioinformatique d'expression de gènes dans des lignées de cellules cancéreuses et mise à disposition par l'institut Broad (MIT & Harvard). [Barretina J. *et al*, 2012]. http://portals.broadinstitute.org

ChIP-seq: Méthode de cistromique permettant d'obtenir par immunoprécipitation et séquençage à haut-débit tous les sites de fixation d'une protéine se liant à l'ADN génomique.

EBI : Institut européen de bioinformatique. Serveur de banques de données de séquences biologiques et d'outils dédiés à leur analyse. http://www.ebi.ac.uk

ERα : Récepteur alpha aux estrogènes.

GOLD: « *Genomes on-line database* ». Banque de données qui fait le point sur les projets de génomes en cours de séquençage. https://gold.jgi.doe.gov

GR: Récepteur nucléaire aux glucocorticoïdes.

GRα : Forme d'épissage la plus longue du récepteur nucléaire aux glucocorticoïdes.

GRE : Elément de réponse sur l'ADN génomique au récepteur aux glucocorticoïdes

HPA: atlas des protéines humaines [Fagerberg L. *et al*, 2014]. http://proteinatlas.org

Ig: Domaine protéique immunoglobuline caractérisé par un sandwich de feuillets beta.

MM/PBSA: « Molecular mechanics Poisson-Boltzmann surface area ». Méthode de thermodynamiqque permettant de calculer l'énergie libre de liaison de 2 macromolécules en interaction. Plus cette énergie (kcal/mol) est basse (négative) plus l'assemblage en oligomère est stable.

MR : Récepteur nucléaire aux minéralocorticoïdes.

LBD : « *Ligand binding domain* ». Domaine de liaison au ligand des récepteurs nucléaires.

MACS : Alignement multiple de séquences protéiques complètes.

NCBI : « *National Center for Biotechnology Information* ». Centre national américain de l'information pour les biotechnlogies. Serveur de banques de données biologiques et de programmes accessible à l'adresse http://www.ncbi.nlm.nih.gov

RMSD: « *Root mean square deviation* ». Mesure globale de la distance moyenne entre les atomes de 2 structures superposées, par exemple entre les LBDs de ERα et GRα. Au cours

d'une simulation de dynamique moléculaire, cette grandeur se représente par un graphe avec la distance moyenne en ordonnée et le temps en abscisse.

RMSF: « *Root mean square fluctuation* ». Variation moyenne de la position de chaque acide-aminé dans une structure, soit au cours du temps d'une simulation de dynamique moléculaire soit dans un jeu de structures superposées. Par exemple, RMSF de la chaîne polypeptidique du LBD de GRα au cours de la dynamique de 100 ns ou dans le jeu des 21 structures obtenues de la PDB. Cette grandeur se représente par un graphe avec la variation de position en ordonnée et la séquence protéique en abscisse.

RNA-seq: Méthode de transcriptomique qui permet par séquençage à haut débit d'obtenir la séquence des ARNm exprimés dans un tissu cellulaire et leurs niveaux d'expression.

PERL: « *Practical expression report language* ». Langage de script interprêté largement utilisé en bioinformatique car adapté au traitement des chaînes de caractères.

PDB : « Protein data bank ». Banque de donnée qui collecte les informations structurales de macromolécules obtenues par cristallographie aux rayons X ou résonance magnétique nucléaire.

PPARα : Récepteur nucléaire alpha associé au proliferateur de peroxisome

PR : Récepteur nucléaire à la progestérone.

PSI-BLAST: **«** *Position specific iterated BLAST* ». Programme de la famille BLAST qui mène une recherche de similarité de séquence itérative dans une banque de donnée par construction d'un profile position-spécifique en lieu et place d'une séquence-requête, sauf pour la 1ère étape de recherche.

Read: Courte séquence nucléotique de quelques dizaines de bases produite en masse par un séquenceur à haut-débit sur des projets de transcriptomique ou cistromique.

RefSeq: Banque de donnée de séquences protéiques et nucléiques de référence maintenue par le NCBI

RN: Récepteur nucléaire.

RPKM : Unité de grandeur quantitative pour le niveau d'expression d'un gène selon la méthode du RNA-seq. « *Read per kilobase of exon and million* ». Il s'agit du nombre de reads qui se sont alignés sur la séquence d'un gène divisés par le nombre de kilobase du gène et ramenés à 1 million de reads totaux séquencés.

SECTM1: Protéine « *Secreted and transmembrane 1* ».

SRA: **«** *Sequence read archive* ». Banque de donnée qui collecte les informations de séquençage haut-débit d'expériences de transcripromique (RNA-seq) et cistromique (ChIP-seq).

TBLASTN: Programme de la famille BLAST qui reçoit une protéine comme séquence -requête et traduit dans les six cadres de lecture les séquences nucléiques d'une banque avant de comparer chaque polypeptide ainsi produit à la séquence-requête.

TEX19: Protéine *« Testis expressed 19* ».

UCSC: « *University of California of Santa Cruz* ». Université qui met à disposition un navigateur de séquences génomiques annotées à l'adresse http://genome.ucsc.edu.

vdW: Forces de van der Waals. Il s'agit d'intéractions faibles qui s'exercent à très courtes distances entre atomes par le biais de dipôles induits.

WGS : « *Whole genome shotgun* ». Segments d'ADN génomiques éparses et en ébauche qui précèdent la mise à disposition d'un génome complet de haute qualité.

G.9. Copies d'articles

Bianchetti L, Tarabay Y, Lecompte O, Stote R, Poch O, Dejaegere A and Viville S. *Tex19 and Sectm1 concordant molecular phylogenies support coevolution of both eutherian specific genes*. BMC Evolutionary Biology. **2015**. 12(15):222.

Bianchetti L, Wassmer B, Defosset A, Smertina A, Tiberti M, Stote RH and Dejaegere A. *Alternative dimerization interfaces in the Glucocorticoid Receptor-α Ligand Binding Domain*. Biochim Biophys Acta. **2018.** 1862(8):1810-1825.

Université de Strasbourg

Laurent BIANCHETTI

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

Intégration de l'évolution pour contribuer à l'étude de la relation séquence, structure, fonction des protéines

Résumé

Intégrer l'évolution peut aider à comprendre la relation séquence, structure, fonction des protéines. Dans un 1er projet, j'ai utilisé la phylogénèse moléculaire pour montrer que les gènes « *Testis expressed 19* » (Tex19) et « *Secreted and transmembrane 1* » (Sectm1) coévoluent. Bien que Tex19 et Sectm1 interviennent dans des processus biologiques différents, régulation des transposons et immunité respectivement, la coévolution établit entre eux un lien fonctionnel très fort. Comme *Tex19* ne s'exprime que dans le testicule de l'adulte sain et en cellule cancéreuse, ce résultat pourrait présenter un intérêt en immunothérapie du cancer. Dans un 2nd projet, je me suis appuyé sur des calculs de modélisation moléculaire et sur l'analyse d'évolution de séquence pour interroger la validité de la structure de l'homodimère du domaine de liaison au ligand (LBD) du récepteur α aux glucocorticoïdes (GRα) [Bledsoe R.K. *et al*, 2002]. Premièrement, ce complexe serait vraisemblablement un artefact de contact cristallin. Deuxièmement, j'ai identifié un assemblage alternatif présentant les caractéristiques moléculaires d'une interface de contact biologique.

Mots clés : Evolution, Séquence, Structure, Fonction, Protéine, Phylogénèse moléculaire, Tex19, Sectm1, Récepteurs nucléaires, Dimérisation, Dynamique moléculaire

Résumé en anglais

Evolution can provide valuable information to understand protein sequence, structure and function relationship. In a first project, I used molecular phylogeny to show the coevolution of « *Testis expressed 19* » (Tex19) and « *Secreted and transmembrane 1* » (Sectm1) genes. Although Tex19 and Sectm1 are involved in different biological pathways, *i.e.* transposon regulation and immunity respectively, coevolution supports a strong functional relationship between both genes. Since Tex19 is expressed only in adult healthy testis and cancer cells, this result may be useful for cancer immunotherapy. In a second project, I used molecular modelling and sequence evolution analysis to question the validity of the glucocorticoïd receptor α (GRα) ligand binding domain (LBD) homodimeric assembly [Bledsoe R.K. *et al*, 2002]. First, this complex is likely a crystallization artefact. Second, I have identified an alternative assembly that presents the molecular characteristics of a biological contact interface.

Keywords: Evolution, Sequence, Structure, Function, Protein, Molecular phylogeny, Tex19, Sectm1, Nuclear receptors, Dimerization, Molecular dynamic