Deep learning pour l'analyse génomique des mélanomes canins

Camille Kergal

AGROCAMPUS OUEST

☐ **CFR Angers**

☒ **CFR Rennes**

Année universitaire : 2018 - 2019 Spécialité : Data Science pour la Biologie Spécialisation (et option éventuelle) :	**Mémoire de fin d'études** ☐ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☒ d'un autre établissement (étudiant arrivé en M2)

Deep learning pour l'analyse génomique des mélanomes canins

Par : Camille KERGAL

Soutenu à Rennes le 5 septembre 2019

Devant le jury composé de :

Président :

Maître de stage : Christophe Hitte

Enseignant référent : Marie-Pierre Etienne

Autres membres du jury (Nom, Qualité)

Remerciements

Je souhaite tout d'abord remercier mes deux encadrants, Christophe Hitte et Thomas Derrien, de m'avoir accueillie au sein de l'équipe Génétique du Chien de l'IGDR et de m'avoir accordé leur confiance pour la réalisation de ce projet. J'ai apprécié la qualité de leur encadrement.

Je remercie également l'ensemble des membres de l'équipe Génétique du chien et Sacha Schutz pour le temps qu'ils ont consacré à mon apprentissage de la génomique et pour la confiance qu'ils m'ont accordé pour répondre à certaines de leurs problématiques.

Enfin, je remercie l'équipe pédagogique du département statistique de l'Agrocampus Ouest qui m'a encadrée pendant cette année d'étude.

Table des matières

Chapitre 1 : Introduction

1.1. Introduction générale

L'Institut de Génétique et Développement de Rennes (IGDR) est composé de 185 personnes réparties au sein de 21 équipes de recherche et de 7 plateaux techniques et services communs. Les projets de recherche s'articulent autour de la conception d'approches innovantes et pluridisciplinaires pour développer une meilleure compréhension quantitative et dynamique des maladies génétiques et du développement de la cellule à l'organisme. Un des axes de recherche concerne les maladies génétiques et parmi celles-ci, le cancer qui fait l'objet de nombreux projets de recherche au sein de l'IGDR. Les cancers sont responsables de 16% des décès dans le monde (OMS), avec des facteurs de risques multiples, qui peuvent être externes, environnementaux en étant liés à l'alimentation, la consommation d'alcool ou de tabac, l'exposition aux rayons ultraviolet ou même à certaines substances polluantes. D'autres facteurs sont relatifs à la composante génétique de l'individu comme la présence de mutations dans le génome qui vont prédisposer aux maladies complexes telles que les cancers. On considère aujourd'hui que plus de 2000 gènes sont susceptibles d'être impliqués dans l'apparition ou le développement d'un cancer (Ilie et al, 2014).

1.2. Les cancers de type mélanome

L'affection étudiée dans le cadre de mon projet de Master 2 est le mélanome, un cancer de la peau, dont il existe deux principaux types. Le premier type est le mélanome cutané qui représente 90% des mélanomes. La fréquence de ce cancer est élevée, l'OMS recense 132000 nouveaux cas dans le monde chaque année, ce qui favorise son étude. Il est reporté qu'entre 65 à 95% des cas de mélanome cutané sont causés par l'exposition au rayonnement ultraviolet (UV) (Centre de lutte contre le cancer Léon Bernard) et dépendent principalement de l'interaction de cette exposition aux UV et le phototype de peau, c'est à dire la classification des catégories de peaux vis-à-vis de leur sensibilité au rayonnement UV. Ainsi, 6 phototypes (1 à 6) sont identifiés, correspondant à 6 types de peaux. Plus le phototype est faible, plus le sujet a besoin de se protéger du soleil. En France, en 2017, un mélanome cutané a été détecté chez 15400 patients dont 52% de femmes et représentait 3,9% des cancers (Institut national du Cancer, 2017). Le taux de mortalité chez ces patients est de 11,6% (Institut national du Cancer, 2017). Ce type de cancer de la peau est très agressif et voit aussi son incidence en constante augmentation depuis 1975 (Siegel et al, 2017).

Le second type de mélanome est le mélanome muqueux. Il comprend les tumeurs qui se développent sur les parties du corps peu ou non exposées au rayonnement ultraviolet. Les mélanomes muqueux surviennent au niveau de plusieurs localisations dont la cavité orale, l'anus, la conjonctive ou encore la muqueuse génitale féminine, des localisations qui ne sont pas liées à l'exposition aux UV. Le mélanome muqueux représente 0,03% de l'ensemble des cancers chez l'Homme (Bennani et al, 2013) ce qui en fait une tumeur rare. Du fait de cette rareté, il est encore peu étudié et souvent diagnostiqué à un stade tardif, souvent métastatique. Le pronostic est particulièrement sombre pour les mélanomes muqueux, qui sont particulièrement agressifs et résistants aux traitements et le taux de survie à 5 ans est inférieur à 25% car les options de traitement sont limitées.

La faible prévalence du mélanome muqueux rend l'utilisation de modèles animaux cruciaux pour améliorer la connaissance et la compréhension de la maladie et pour prédire la réponse à certains traitements. Un modèle animal correspond à une espèce présentant une pathologie similaire à une pathologie humaine qui va donc servir de schéma, de modèle pour l'étude de cette affection. A titre d'exemple, l'invention du vaccin contre la rage a été permise grâce au modèle du lapin (Hicks et al, 2012) et la compréhension du traitement de l'information spatiale par le cerveau vient du modèle du rat (Hafting et al, 2005). Pour le cancer, la quête de modèles plus proches du fonctionnement de l'Homme et plus soucieux du bien-être animal mène aujourd'hui à l'exploitation de tumeurs développées naturellement par nos animaux de compagnie. Dans le cas du mélanome muqueux, les modèles canins, équins et félins sont très pertinents.

1.3. Le modèle canin

C'est auprès de l'équipe « Génétique du chien » à l'IGDR que j'ai réalisé mon stage et découvert les avantages liés à l'étude du modèle canin. Entre tous les mammifères, l'espèce canine est celle où nous observons la plus forte variabilité morphologique entre les 349 races qui la composent (Fédération Cynologique Internationale). Un exemple frappant est celui du poids d'un chihuahua et celui d'un Saint-Bernard qui varie de 2 kg à 90 kg, ces deux races appartenant bien à la même espèce, le chien domestique (*Canis lupus familiaris*). Les chiens que nous connaissons sont issus de la domestication du loup gris qui remonte à environ 15000 ans (Galibert et al, 2011). C'est par des pratiques intensives de sélection qui impliquent des croisements au sein de populations fermées ou à partir de quelques reproducteurs que l'Homme a pu créer d'aussi nombreuses races qui se différencient fortement.

Néanmoins, la création des races canines a entrainé des effets négatifs sur la santé des animaux. En effet au sein d'une race canine, il y a une forte homogénéité des phénotypes et de la génétique. Ainsi une race constitue un véritable isolat génétique qui peut entraîner de multiples prédispositions aux maladies génétiques. Par comparaison, on considère qu'une maladie est très fréquente chez l'Homme dès lors qu'elle touche 0,2% de la population or, il est commun d'observer une maladie atteignant 1 à 10% des chiens d'une même race (Giger et al, 2006). Cependant, cette sélection et cette structuration de l'espèce canine en races bien distinctes permet de faciliter l'identification des altérations génétiques qui prédisposent le développement de maladies car la population d'une race canine est très homogène à la fois du point de vue phénotype (aspect, morphologie, pelage) et génétique. Plus de 730 maladies génétiques sont répertoriées chez le chien (Online Mendelian Inheritance in Animals) dont approximativement 360 sont homologues aux maladies humaines (Shearin et al, 2010). La compréhension de l'origine et du développement de certaines maladies humaines peut donc être enrichie avec les études génétiques qui utilisent le modèle canin.

Il se trouve également que les chiens partagent souvent le même environnement que leur propriétaire. Dans le cas d'une exposition à une substance cancérigène, le délai de développement d'une tumeur est plus rapide chez le chien que chez l'Homme et confirme donc sa place de modèle et son rôle de sentinelle en étant le 1er atteint. Par exemple, la détection d'un mésothéliome, un cancer affectant le revêtement des poumons, chez le chien indique la présence d'amiante dans l'environnement et permet de prévenir l'apparition ou d'avancer le diagnostic du même cancer chez son propriétaire (Glickman et al, 1983).

Par ailleurs, l'espérance de vie des chiens est relativement courte, et dépendante de la race mais sa médiane se situe à 10 ans et 4 mois (Lewis et al, 2018), ce qui permet ainsi d'étudier plus facilement et plus rapidement l'effet des traitements (Porrello et al, 2004). Aussi, l'effet d'une thérapie sur une tumeur est fréquemment la même chez le chien que chez l'Homme (Gordon et al, 2010). Ainsi, des essais cliniques chez le chien peuvent permettre d'anticiper et donc mieux préparer ceux de l'Homme en informant sur l'interaction ou la toxicité de traitements sur l'organisme. De plus, le domaine de la recherche et développement pharmaceutique est contraint d'utiliser un modèle d'étude *in-vivo* non humain avant tout essai sur l'Homme (Déclaration d'Helsinki)

Le mélanome muqueux représente 3,15% de la totalité des cancers canins et est le type de mélanome le plus présent chez cette espèce. Parmi les races canines, certaines sont fortement prédisposées, telles que les caniches ainsi que le Labrador et le golden retriever (Cadieu et al, 2014). D'importantes similarités entre les mélanomes muqueux canins et humains sont rapportées. Les mélanomes canins se développent sur les mêmes sites anatomiques que les humains et sont analogues d'un point de vue cytologique (les cellules qui composent les tumeurs) et histologique (la morphologie des tumeurs) (Gillard et al, 2013). L'incidence de la pathologie permet d'analyser plus de cas que chez l'Homme et donc de mieux comprendre, mesurer et prédire ses effets ainsi que d'identifier les événements qui permettent son développement, tout cela au double bénéfice de la médecine humaine et vétérinaire.

1.4. Machine learning et génomique

La génomique est une discipline récente de la biologie qui étudie le fonctionnement d'un organisme, d'une fonction, d'un cancer, etc. à partir des données génétiques à l'échelle globale du génome. Elle implique l'analyse de données en grande dimension ou « big data ». Au sein de mon équipe d'accueil l'équipe « Génétique du chien », c'est auprès des bioinformaticiens que mon apport en apprentissage automatique ou machine learning avait sa place.. Pour pouvoir exploiter les données génétiques de l'ADN, nous devons d'abord le séquencer. La séquence obtenue est une suite d'élément qu'on appelle les nucléotides. Nous comptons quatre nucléotides différents, l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T). Par exemple, la séquence du génome humain est composé de 3,2 milliards de nucléotides et celle du génome canin de 2,4 milliards. La méthode permettant d'obtenir ces séquences a commencé à voir le jour dans les années 1970 mais le premier séquençage complet du génome humain a nécessité une dizaine d'année de travail et s'est achevé en 2004 (International Human Genome Sequencing Consortium) pour un investissement de 2,7 milliards de dollars. Aujourd'hui les nouvelles techniques de séquençage haut-débit (NGS ou Next Generation Sequencing) sont capables de séquencer 16 génomes humains en trois jours pour un coût inférieur à 1000 euros chacun (Sheridan, 2014). Ces avancées technologiques ont grandement favorisé la prolifération des données de séquences disponibles et ont donc permis au machine learning de trouver sa place dans les projets de génomique. La composition de l'ADN guide le comportement des gènes et ces derniers gouvernent notre apparence, notre état de santé et plus globalement qui nous sommes. L'intérêt d'utiliser des approches de machine learning en génomique est donc d'interpréter la structure, la fonction et le sens de toutes ces informations dans le but de découvrir le lien entre le génotype et un phénotype.

Depuis quelques années, nous pouvons constater le progrès exponentiel des approches de deep learning (Zou et al, 2018). Cet essor est soutenu par le bond de la puissance de calcul informatique qui permet de traiter l'information et par la communauté grandissante autour de

l'intelligence artificielle. Dans la plupart des domaines d'application génomique de l'apprentissage automatique, le deep learning dépasse en performance les autres procédures de machine learning (SVM, RandomForest). De plus, elle gagne en efficacité en traitant des grands volumes de données, son emploi dans les problématiques de génomique semble donc particulièrement pertinent.

Nous sommes aujourd'hui capables de connaître la composition de l'ADN de chaque espèce en détaillant aussi bien sa séquence nucléotidique (l'ADN qui compose le génome) que l'expression des gènes associée (l'ARN qui compose le transcriptome). Ces connaissances trouvent de nombreuses applications comme l'analyse de l'évolution des espèces, la médecine reproductive ou la médecine légale. Pour obtenir le séquençage d'un génome, il est tout d'abord nécessaire de prélever un échantillon sanguin, de tissu, de cheveux ou de salive qui va contenir de l'ADN. Nous allons ensuite isoler cet ADN des tissus à l'aide de détergents et de solvants organiques puis d'une phase de décantation ou de centrifugation. Un génome entier représentant plusieurs milliards de nucléotides, il est nécessaire d'allier une approche biologique à celle de la bioinformatique pour le séquencer. Le premier point consiste à couper l'ADN en fragments à l'aide d'enzymes de restriction. Ces courts fragments sont ensuite séquencés de manière aléatoire et il s'agit ensuite de mettre en oeuvre les outils bioinformatiques pour assembler dans le bon ordre les séquences en repérant les suites communes et ainsi connaître l'enchaînement nucléotidique complet qui compose notre génome.

Quelque soit le tissu ou les types cellulaires étudiés au sein d'un même individu, le génome sera toujours composé de la même suite de nucléotides. Ce qui va différencier un tissu d'un autre sera essentiellement le niveau de l'expression des gènes, ou transcriptome, associé. C'est ce type de données que j'ai utilisé dans le cadre de mon stage et ces informations ont été recueillies par une technologie de séquençage appelée RNA-seq ou séquençage complet du transcriptome. L'ARN est la partie transcrite de l'ADN et est séquencé en courtes séquences d'environ 100 nucléotides que l'on appelle des reads et qui se chevauchent entre elles. L'objectif est d'aligner ces reads et d'évaluer leur abondance le long du génome correspondant aux positions des gènes afin d'obtenir une mesure quantitative de l'expression des gènes, on parle de profil d'expression pour un tissu donné.

Des outils reposant sur un algorithme de deep learning sont déjà développés pour étudier et faire parler l'ADN. Par exemple, DeepVariant (Poplin et al, 2018), développé par une équipe de Google, permet d'identifier les variations, les mutations, d'un génome individuel à l'échelle du nucléotide. Sa spécificité vient du fait qu'il traite l'alignement des reads sur un génome de référence comme une image en accordant une valeur à chaque position comme si c'était un pixel, et que sa performance est bien supérieure aux outils utilisés jusqu'alors et dont le fonctionnement ne repose pas sur du deep learning. De même, certains outils visent à expliquer et prédire l'effet des variations nucléotidiques sur l'expression des gènes en utilisant les réseaux de neurones convolutif (cf. Matériel et méthodes). En effet, même si nous pouvons identifier quelles variations génèrent un trait physique ou une maladie, nous ne connaissons pas les liens qui les unissent. Ce pan de la génétique reste encore à explorer et c'est dans ce sens que se dirige la recherche.

Chapitre 2 : Matériel et méthodes

2.1. Données génomiques

2.1.1. Le séquençage de l'ADN et l'expression des gènes

L'ADN d'une cellule est constitué d'une succession des nucléotides (A, C, G, T) reliés les uns aux autres grâce à des liaisons chimiques. Cet enchaînement de nucléotides est organisé en unités fonctionnelles, les gènes, qui occupent un emplacement déterminé dans les génomes. Chaque gène permet de produire des molécules d'ARNs qui peuvent être codant et non codant pour des protéines. Les ARNs non-codant pour des protéines assurent des fonctions de régulation des gènes régissant le fonctionnement des cellules de l'organisme. Toutes les cellules d'un individu possèdent les mêmes gènes, le même ADN, mais seuls certains gènes sont exprimés ou autrement dit, transcrit en ARNs, dans une cellule donnée, selon le type cellulaire concerné (neurone, mélanocyte, etc.), le stade de développement de l'organisme, le sexe, etc. On désigne par le terme de transcriptome l'expression de l'ensemble des gènes d'une cellule ou d'un tissu à un stade de développement donné.

Dans le cadre de mon stage, j'ai utilisé plusieurs type de données d'expression des gènes. J'ai tout d'abord travaillé avec des données d'expression issues d'une méthode CAGE (Cap Analysis of Gene Expression), différente du RNA-seq puisqu'elle cible seulement la partie amont (30-50 nucléotides) des gènes. Dans cette analyse, trois tissus (l'aorte, l'artère et la valve pulmonaire) ont été séquencés par CAGE chez l'Homme via le consortium FANTOM (Bertin et al, 2017). Ces données étaient disponibles en téléchargement, liées à l'utilisation d'un outil que je détaille par la suite. Chaque région du génome est associée à un pic qui correspond au nombre de reads provenant du séquençage du transcriptome de cette même région (Figure 1). L'expression des gènes de cette région est représentée par le pic associé. Un gène peut ne pas être exprimé dans un tissu ou dans une condition particulière et être très exprimé dans un autre tissu ou une autre condition.

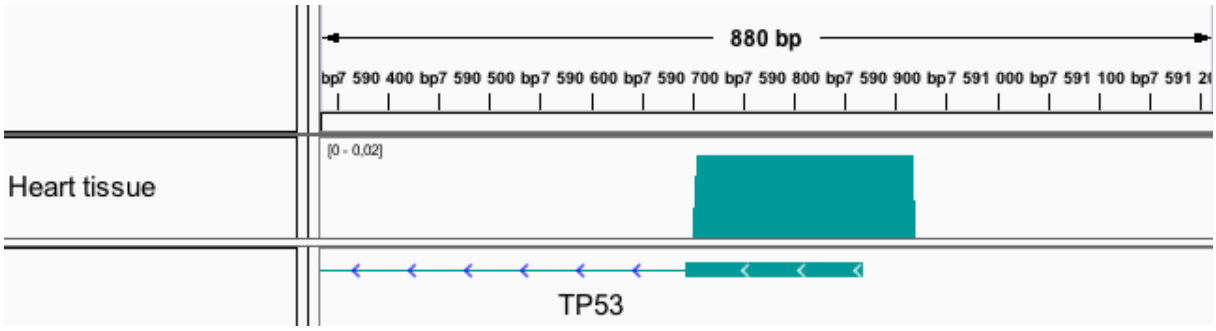

Figure 1. Représentation de l'expression génique grâce au logiciel IGV. Nous visualisons ici l'expression du gène TP53 d'un tissu provenant de l'aorte d'un Homme. La valeur de l'expression se trouve entre 0 et 0,02 sur cette partie du génome composée de 880 nucléotides dans le début du gène.

2.1.2. Le mélanome muqueux du chien

Nous disposons au sein du laboratoire de 39 échantillons transcriptomiques de tumeurs buccales de mélanomes muqueux canins et de 12 échantillons contrôles (muqueuse buccale saine) qui ont été séquencés par RNA-seq. Les tissus qui ont servi à leur élaboration ont tous été re-

cueillis selon un protocole rigoureux par des vétérinaires partenaires de l'équipe Génétique du chien. Ces données de RNA-seq du projet de chiens atteints du mélanome muqueux sont analysées par rapport au génome canin de référence pour évaluer leur expression. Dans le cas du chien, cette référence correspond au génome d'une femelle boxer connue dans son intégralité car séquencée entièrement dès 2005 (Lindblad-Toh et al, 2005).

2.2. Le programme Basenji

L'outil appelé Basenji publié récemment (Kelley et al, 2018) permet de prédire l'expression des gènes chez l'Homme grâce à une approche de deep learning, utilisant le réseau de neurones convolutif ou CNN (convolutional neural networks). Basenji est configuré pour prendre en entrée des données d'expression de type CAGE, afin d'entraîner un modèle de réseau de neurones convolutif pour ensuite prédire l'expression des gènes à partir de la séquence d'ADN fournie. L'outil est disponible en téléchargement libre sur la plateforme Github (https://github.com/calico/basenji) et est proposé en langage Python par le laboratoire Calico.

Les données de séquençage étudiées sont tout d'abord transformées pour pouvoir être traitées par une approche de deep learning. Ainsi, 17217 séquences d'une taille de 131072 nucléotides ($= 2^{17}$) sont formées depuis les nucléotides du génome de référence. Dans cette étape, nous excluons les régions du génome dont la séquence n'est pas encore de haute qualité. Chaque séquence est représentée par 4 lignes de variables binaires représentant la présence ou l'absence d'un A, C, G ou T à chaque position de nucléotide. Par exemple, le nucléotide A devient le vecteur (1, 0, 0, 0) et ce nouveau format correspond à une conversion en un format appelé 'One-hot coded'. Ces séquences sont les variables explicatives associées aux valeurs d'expression des gènes, qui correspondent elles aux variables à expliquer.

2.2.1. Réseau de neurones convolutif

Un réseau de neurones convolutif est une approche de deep learning déjà connue pour le traitement de données d'imagerie. Basenji utilise cette stratégie pour prédire l'expression des gènes en s'appuyant sur deux étapes principales. La première étape est la convolution. L'objectif de cette phase est de chercher les caractéristiques globales qui définissent chaque jeu de données de séquence de chaque individu. Elle se décompose elle-même en plusieurs points successifs. Il s'agit dans un premier temps de caractériser l'information apportée par les données, région par région, à travers des filtres dont on choisit les dimensions. Pour l'analyse des données génomiques, ces filtres sont des matrices poids-position, ou PWM (Position Weight Matrix).

L'outil Basenji réalise une première convolution, c'est-à-dire un produit scalaire, entre une PWM et chaque région du génome, en utilisant 128 PWM ($= 2^{7}$). C'est à dire que l'ensemble des régions va être analysé par 128 matrices différentes et se verra donc résumé de 128 manières différentes (Figure 2). Le résultat est une nouvelle matrice où un point correspond à la convolution d'une PWM avec une région, à laquelle on applique la fonction d'activation ReLU (pour rectified linear unit) qui consiste à rendre nulles les valeurs négatives et à garder telles qu'elles les autres valeurs. L'algorithme cherche ensuite à réduire la dimension grâce à l'étape de pooling (Figure 2) qui consiste à résumer l'information de plusieurs neurones voisins en une seule information. Cette agrégation peut être la moyenne de ces neurones ou encore le maximum. Dans l'apprentissage de ce processus, ce sont directement les poids des ma-

trices de convolution que l'on cherche à optimiser dans le but d'extraire des séquences d'entrée l'information la plus pertinente possible.

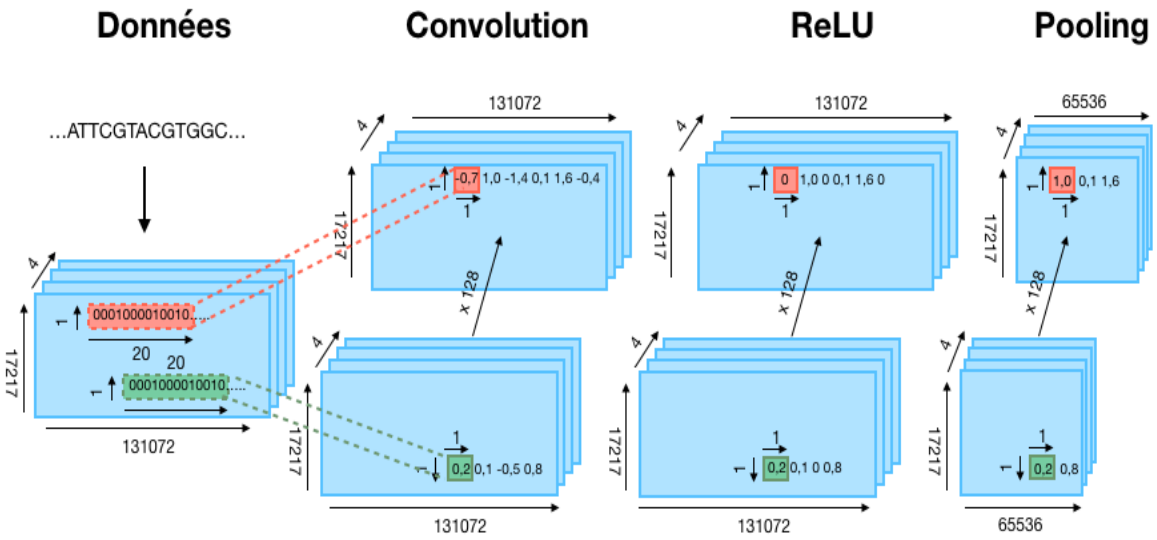

Figure 2. Première couche de convolution de Basenji. L'ADN codé en ACGT subit en premier lieu une transformation que l'on appelle One-hot coded. Un nucléotide est alors codé par un vecteur de quatre éléments dont un vaut 1 et les autres 0 et l'ADN est découpé en 17217 séquences de 131072 (soit 2^{17}) nucléotides. Nous avons donc ici une structure de données à dimension 3. Ensuite, cet objet est convolué par fragments de 20 unités avec 128 PWM et un pas de 1. Cette transformation abouti à un nouvel objet de dimension 131072*17217*4*128 où chaque point est un produit scalaire. Certains sont négatifs, la fonction ReLU les élève à la valeur 0. L'étape de pooling vise à garder uniquement le maximum des points deux à deux et l'information de 131072 éléments est divisée par 2.

Ces opérations de convolution et de pooling peuvent être répétées, avec modification ou non des paramètres (nombre de filtre, taille de pooling..) aboutissant à la fin de ces opérations à un vecteur de bien plus petite dimension mais toujours très représentatif des différentes régions du jeu de données d'entrée. Ainsi, il devient plus simple d'analyser ces données par un nouveau réseau de neurones. Basenji réalise au total 5 enchaînements de ce type avant de passer à la deuxième étape du réseau de neurones convolutif.

Cette deuxième étape consiste à utiliser le résultat de l'étape de convolution par le moyen d'un réseau de neurones classique. Cette partie se décompose encore une fois en plusieurs couches. La première couche est la couche d'entrée. Cette couche est composée d'autant de neurones qu'il y a de caractéristiques pour décrire chaque individu du jeu de données. Pour les données relatives à mon projet, leur dimension à l'issu de la convolution est de 17217*1024*4 au lieu de 17217*131072*4 avant la convolution. Un individu devient alors caractérisé par 1024 vecteurs de 4 éléments ce qui implique que cette première couche soit constituée de 1024 neurones. Chacun des neurones de cette première couche est relié à chacun des neurones de la seconde couche, qui peut être une couche cachée ou la couche de sortie, par des poids.

Lors de la première itération de l'algorithme, ces poids sont initialisés aléatoirement. Le nombre de couche entre la couche d'entrée et la couche de sortie fait partie des hyperparamètres du réseau. Nous choisissons le nombre de neurones car cette fois, les couches ne dépendent pas de la structure des données. Cependant, elles sont toutes entièrement reliées par

des poids. La dernière couche, la couche de sortie, est celle qui détermine la valeur de prédiction. Pour Basenji, cette couche donne 1024 niveaux d'expression pour chaque séquence issue de la convolution. Cela signifie que pour une séquence d'entrée d'une taille de 131072 nucléotides, le réseau est capable de prédire un niveau d'expression tous les 128 nucléotides.

Pour pouvoir réaliser son entraînement et s'améliorer dans les prédictions, l'algorithme a besoin de créer deux jeux de données à partir des données initiales. Un jeu d'entraînement et un jeu de validation. Dans le cadre de Basenji, l'auteur préconise l'utilisation de 80% des séquences nucléotidiques et de leur niveau d'expression associée pour le jeu d'entraînement et 10% pour le jeu de validation. Ce qui permet à l'algorithme d'améliorer ces prédictions à mesure qu'il analyse les individus du jeu d'entraînement est le principe de rétropropagation. Pour chaque séquence analysée dans la phase d'entraînement, l'algorithme assigne une prédiction. Lorsque le réseau a analysé l'ensemble des individus, il compare les prédictions qu'il a réalisé avec les valeurs réelles. On dit qu'il a réalisé une 'époque' et il en établit une fonction de coût mesurant la distance entre la prédiction et la cible de chaque individu. Pour l'époque suivante, le réseau va modifier chacun des poids qui relient les neurones entre eux mais aussi les valeurs composant les PWM en commençant par la dernière couche du réseau pour aller vers la première couche de convolution.

L'objectif est de diminuer le résultat de cette fonction de coût. En appliquant le modèle formé à partir de ces paramètres au jeu de validation, on obtient un autre résultat de fonction de coût. Si celui-ci diminue, alors le modèle inhérent à cette époque est considéré comme meilleur que le précédent. Une difficulté lors de cette recherche de minimum est que l'algorithme stagne vers un minimum local et ait besoin de la réalisation de nombreuses époques pour pouvoir en sortir. Un hyperparamètre capable de maîtriser ce phénomène est le taux d'apprentissage. Il représente la distance entre deux valeurs à tester (Figure 3).

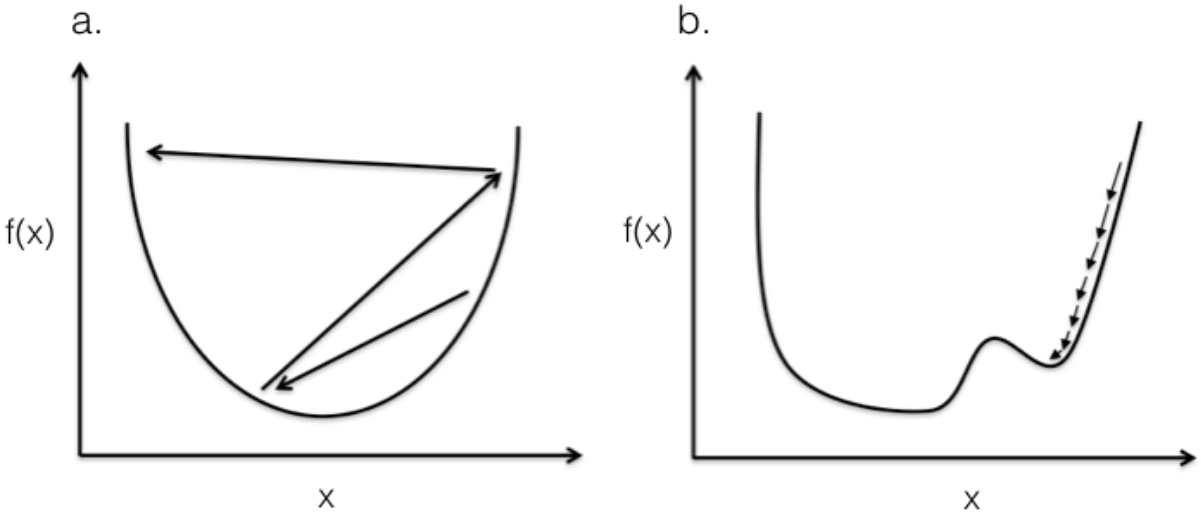

Figure 3. Incidence de la valeur du taux d'apprentissage sur la recherche de minimum. (a) Fonction dotée d'un minimum global avec un taux d'apprentissage élevé. L'algorithme évalue le résultat de f(x) pour des valeurs de x trop éloignées, ne lui permettant pas d'atteindre le minimum. (b) Fonction dotée d'un minimum global et un autre local avec un taux d'apprentissage faible. L'algorithme évalue le résultat de f(x) pour des valeurs de x trop proches, provoquant un bloquage autour du minimum local et nécessitant un grand nombre d'époque pour en sortir et trouver le minimum global.

L'enjeu lors du choix du taux d'apprentissage d'un réseau de neurones est de trouver celui qui permettra de trouver le minimum global de la fonction de coût sans avoir à réaliser un trop grand nombre d'époque. Dans le cas d'un réseau de neurones complètement entraîné, il doit être capable de prédire exactement la valeur du jeu d'entraînement. En revanche, il est préférable de choisir un modèle qui ne s'ajuste pas trop aux données d'entrainement pour pouvoir prédire le plus précisément possible les valeurs du jeu de validation. Il existe un hyperparamètre capable d'éviter le sur-ajustement, le 'dropout'. Il correspond à un taux de neurones qui se trouvent ignorés durant la phase d'entrainement permettant au modèle de ne pas s'accorder complètement aux données d'entrainement mais de n'en retenir que les principales caractéristiques.

Une fois que le nombre d'époque choisie est atteint, que le modèle permet d'obtenir une fonction de coût suffisamment faible sur le jeu de validation ou encore que cette même fonction de coût n'est plus possible à minimiser, l'algorithme s'arrête. L'ensemble des poids est alors figé et permet d'obtenir le modèle final.

Chapitre 3 : Résultats

3.1. Utilisation du programme Basenji

3.1.1. Prise en main de Basenji avec le tutoriel

Mon premier objectif lors de ce projet de master a été de mettre en place et de rendre opérationnel l'outil Basenji. Il est composé de nombreux scripts codés en langage Python et repose principalement sur la bibliothèque TensorFlow, spécifique au deep learning. J'ai téléchargé Basenji à partir de la plateforme Github ainsi que le tutoriel proposé par l'auteur (Kelley et al, 2018) pour faciliter la compréhension de l'outil et des démarches nécessaires à son fonctionnement. Par ailleurs, l'auteur précise que Basenji est encore en développement et n'est donc pas encore développé dans sa version optimale.

J'ai utilisé le tutoriel qui consiste à construire un modèle de deep learning, capable de prédire l'expression des gènes chez l'Homme à partir de la séquence d'ADN et donc sans produire les données expérimentales de transcriptome. J'ai utilisé les trois jeux de données mis à disposition pour l'expérimentation de l'outil. Il s'agit de données d'expression d'échantillons humains de l'aorte, l'artère et la valve pulmonaire obtenues selon la méthode de séquençage CAGE dans un format appelé BigWig où une valeur d'expression est associée à chaque nucléotide du génome de référence.

La première étape consiste à convertir le format des données pour qu'elles puissent être utilisables par la suite. Pour cela, les fichiers BigWig sont divisés pour obtenir d'une part, un fichier répertoriant les positions de début et de fin de séquences de 131072 nucléotides et aussi la distribution de chacune de ces séquences entre le jeu d'entraînement, de validation ou de test dans une proportion de 80%, 10% et 10 % respectivement. D'autre part, les données d'expression de ces séquences sont fournies au format HDF (Hierarchical Data Format), spécifiques aux grandes quantités de données. Ces transformations sont nécessaires pour permettre une dernière conversion au format TFRecord, propre à la bibliothèque TensorFlow, qui permet d'associer les séquences et leurs expressions au génome de référence de l'Homme.

Durant mon stage, j'ai installé et utilisé la bibliothèque TensorFlow v.1.13 sur mon poste de travail.

J'ai ensuite réalisé l'étape d'entraînement du modèle qui utilise le jeu d'entraînement et le jeu de validation, formés précédemment. Avant l'entraînement, il est également impératif de préciser quels hyperparamètres feront la structure du réseau de neurones convolutif. Dans le tutoriel, l'auteur en propose une sélection précise permettant un résultat optimal sur ces données d'exemple. J'ai utilisé le nombre de couche de convolution, leur taille et les autres hyperparamètres du réseau recommandés durant la phase d'entraînement.

Lorsque le modèle à analysé et proposé une prédiction de l'expression de l'ensemble des données d'entraînement, il a réalisé la première époque et en a retenu un modèle sous-jacent. La qualité de ce modèle est mesurée en fonction de la valeur d'une fonction qui est, chez Basenji, la fonction de coût de Poisson. Lorsqu'un modèle propose une fonction de coût minimale, je réalise 25 époques supplémentaires pour rechercher un modèle encore meilleur. S'il n'y a pas de meilleur modèle dans ces 25 nouvelles époques, l'algorithme s'arrête. Si un modèle est meilleur, 25 époques sont à nouveau réalisées.

Une fois l'entraînement du modèle terminé, il s'agit de le tester. Cette phase est réalisable grâce au 10% restant du jeu de données initial. Basenji permet de mesurer l'efficacité du modèle en déterminant la corrélation de Pearson entre l'expression prédite par le modèle et celle mesurée expérimentalement en appliquant auparavant un log2 à ces données. J'ai représenté graphiquement la corrélation entre la prédiction du modèle et la réalité sous forme de nuage de points pour un tissu donné (Figure 4). Le modèle étant entraîné sur la base de trois échantillons différents (aorte, artère, valve pulmonaire), j'ai pu évaluer la capacité de prédiction sur ces trois tissus également à l'aide des trois nuages de points obtenus (Tableau 1).

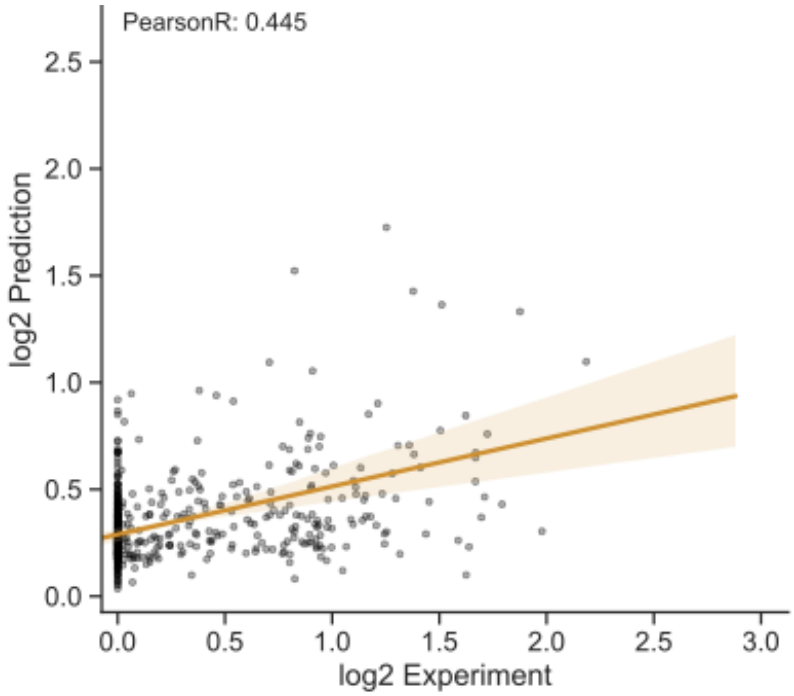

Figure 4. Graphique mesurant l'efficacité du modèle. Il met en opposition le log2 de la valeur de l'expression génique de la valve pulmonaire mesurée expérimentalement (abscisses) et de celle prédite par le modèle de réseau de neurones convolutif (ordonnées). Ici, 3628 séquences de 131072 nucléotides sont testées. Pour chaque séquence, l'outil est capable de prédire l'expression de 960 points mais l'auteur à fait le choix de ne représenter qu'un point sur huit. Nous avons donc ici 435360 points re-

présentés. La corrélation de Pearson présentée (0,445) est calculée sur ces 435360 points et non pas sur l'intégralité des points testés et prédits.

	Aorte	Artère	Valve pulmonaire
Pearson log2	0,59367	0,69871	0,52504
Pearson	0,55444	0,66016	0,52504

Tableau 1. Mesure de performance de l'intégralité des données testées par le modèle de prédiction de l'expression dans les trois tissus testés : aorte, artère, valve pulmonaire. La première ligne correspond à la corrélation de Pearson sur les données transformées par un log2. La deuxième ligne correspond à la corrélation de Pearson sur les données brutes.

3.1.2. Détection d'une anomalie dans l'échantillonnage

Durant cette phase d'exploration de l'outil, j'ai identifié une anomalie dans le script lors la répartition des données entre le jeu d'entraînement, de validation et de test. En effet, les données de validation et de test se trouvaient pour 100% et 90% respectivement dans le jeu d'entraînement. Cet évènement conduisait donc à un phénomène de sur-ajustement inévitable et invisible, aussi bien dans la phase d'entraînement que de test. J'ai donc modifié le script pour assurer le bon échantillonnage des données et soumis ma modification à l'auteur de Basenji qui a confirmé mon hypothèse (https://github.com/calico/basenji/issues/38).

Suite à cette modification, j'ai donc recommencé les analyses avec les données humaines disponibles en réalisant un nouveau modèle de prédiction. Comme attendu, les statistiques de corrélation présentaient une prédiction moins bonne qu'avec le premier modèle, et cela pour les trois tissus testés (Tableau 2).

	Aorte	Artère	Valve pulmonaire
Pearson log2	0,47494	0,59753	0,444
Pearson	0,51179	0,51234	0,45816

Tableau 2. Résultats suivant le même principe que le tableau 1, inhérent au modèle de prédiction réalisé suite aux modifications de l'outil. Nous pouvons donc voir que les corrélations sont plus faibles avec ce nouveau modèle.

3.2. Application au modèle canin

L'objectif final de mon projet de recherche est l'analyse génomique des mélanomes canins afin de déterminer les variations d'expression des gènes dans la condition tumorale, qui peuvent être des marqueurs de diagnostic ou des cibles pour un traitement. La seconde étape de mon stage était donc de transposer les modèles de deep learning réalisées chez l'Homme au modèle canin.

3.2.1. Transformation des données

Basenji a été configuré pour analyser des données humaines. La première difficulté que j'ai rencontrées lors de la transposition de l'outil a été le format des données. En effet, les données humaines utilisées lors de ma prise en main de l'outil ne viennent pas du même consortium que les données canines que je souhaitais utiliser. Cela se traduit par des formats d'annotation différents dans les fichiers d'expression qui bloquaient le processus d'échantillonnage. J'ai donc dû modifier et adapter le script pour qu'il puisse prendre en compte ces nouvelles annotations mais aussi le fait que le génome du chien n'est pas de même taille ni de même structure que celui de l'Homme. Le génome canin est composé de 39 chromosomes et celui de l'Homme 23 et chacun d'entre eux ne fait pas la même taille nucléotidique.

3.2.2. Réalisation du modèle sur les données génétiques canines

Suite à ces nouvelles modifications, j'ai donc pu réaliser un premier modèle de prédiction de l'expression des gènes chez le chien. De la même manière que pour le tutoriel proposé chez l'Homme, j'ai utilisé trois fichiers d'expression différents. Ces derniers correspondent à des prélèvements de de tissu de la muqueuse de la truffe chez trois labradors. Tout comme le modèle humain, j'ai réparti 80% de ces données pour le jeu d'entraînement, 10% pour le jeu de validation et 10% pour le jeu de test.

Basenji est configuré pour arrêter l'entraînement du modèle dès lors que la fonction de coût ne diminue plus sur le jeu de validation après 25 époques. Ce modèle s'est entraîné durant 152 époques à raison d'un temps de calcul d'en moyenne 11 minutes par époque. Cela indique qu'un tel modèle nécessite en entraînement de 27,8 heures. J'ai ensuite testé ce modèle canin de la même manière que le modèle humain, grâce au jeu de test.

	Chien 1	**Chien 2**	**Chien 3**
Pearson log2	0,39748	0,40991	0,37616
Pearson	0,26397	0,27595	0,26685

Tableau 3. Résultats statistiques de l'efficacité du modèle canin entraîné sur des échantillons de truffes provenant de trois chiens différents.

Nous pouvons observer une baisse de performance de prédiction (Tableau 3) par rapport aux performances chez l'Homme. Ainsi par rapport au modèle humain (Moyenne du Pearson log2 = 0,50549, moyenne du Pearson = 0,49409), le modèle est moins performant (Moyenne du Pearson log2 0,3945, moyenne du Pearson = 0,26892) mais ces résultats démontrent qu'il est possible d'appliquer la méthode sur les données génomiques canines.

3.2.3. Recherche de la configuration optimale

Après avoir mis en place et évalué Basenji pour la prédiction de l'expression des gènes chez le chien, mon objectif était d'améliorer cette capacité de prédiction. Ma première piste pour réaliser cet objectif était de déterminer le comportement d'un modèle en fonction des données dont il s'est servi pour s'entraîner. En effet, il est possible de préciser à Basenji le taux d'utili-

sation des données c'est-à-dire que l'utilisateur peut décider d'en garder seulement un certain pourcentage afin d'accélérer le processus. Le tutoriel propose par défaut d'utiliser 10% des données. J'ai fait le choix de comparer cette manière de faire avec l'utilisation de la totalité des séquences car mon objectif est dans un premier temps d'améliorer la capacité de prédiction du modèle sans prendre en compte le coût du temps de calcul. De même, il est possible de proposer au modèle d'utiliser le nombre de tissus que l'on souhaite pour l'entraînement. J'ai ici aussi fait le choix de tester l'apprentissage sur 3 échantillons (3 tissus de muqueuse de truffe issus de 3 individus) d'une part et sur 1 échantillon (1 tissu truffe d'un individu) d'autre part. Ces hypothèses m'ont amené à comparer les performances de quatre modèles différents.

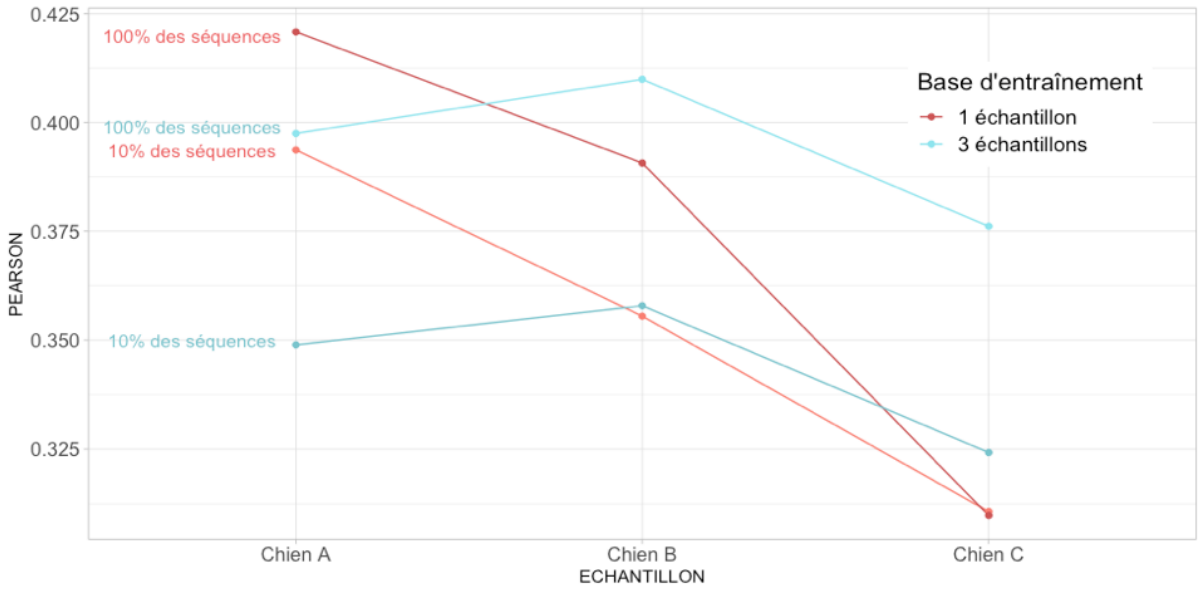

Figure 5. Comparaison des modèles canins. Ce graphique permet de comparer les valeurs de corrélation (en ordonnée) entre l'expression prédite par 4 modèles et l'expression mesurée expérimentalement de 3 tissus de muqueuse de truffe provenant de 3 individus (en abscisse). Les corrélations des prédictions des modèles entraînés sur 1 échantillon sont en rouge et celles des modèles entraînés sur 3 échantillons sont en bleu. Pour ces deux bases d'entraînement, j'ai réalisé un modèle avec 10% des séquences et un autre avec 100% des séquences.

Les résultats indiquent que prendre en compte 'intégralité des séquences nucléotidiques (100%) pour l'élaboration du modèle améliore ses performances (Figure 5). En moyenne, les modèles construits sur l'utilisation de 10% des séquences ont une corrélation de 0,348 et ceux construits sur l'utilisation de 100% des séquences ont une corrélation de 0,384. Les résultats montrent aussi des performances plus stables via l'utilisation de 3 échantillons dans la construction du modèles et en moyenne une légère augmentation de la corrélation (0,37 pour 3 échantillons et 0,36 pour 1 échantillon).

3.3. Application au mélanome canin

3.3.1. Réalisation des modèles

Après avoir adapté l'outil à l'étude des données canines et étudié la capacité de prédiction de plusieurs modèles, j'ai entrepris l'analyse des données du mélanome muqueux du chien. Pour

cela, j'ai réalisé plusieurs modèles, deux modèles entraînés sur des données tumorales et deux modèles entraînés sur des données issues de tissus sains (ou contrôles) pour lesquels nous disposons d'échantillons. Dans le cas tumoral comme dans le cas contrôle, j'ai fait le choix d'entraîner un modèle avec un échantillon et un modèle avec 10 échantillons. En effet, les différences de prédiction entre un modèle entraîné sur 3 échantillons et un modèle entraîné sur 1 échantillon ne permettent pas de trancher facilement en faveur de l'un ou l'autre.

Pour la réalisation du premier modèle, j'ai utilisé les données d'expression d'un échantillon provenant de tissu muqueux correspondant à la tumeur d'un caniche atteint du mélanome muqueux. De la même manière, 80% des séquences de 131072 nucléotides et leurs niveaux d'expression associés sont utilisés pour l'entraînement et 10% pour la validation. J'ai modifié le script d'entraînement du modèle pour pouvoir choisir à quel moment je pouvais interrompre l'algorithme. Pour ce premier modèle, j'ai interrompu l'algorithme au bout de 145 époques qui ont duré en moyenne 12 minutes chacune mais le modèle qui a été conservé (celui avec la plus faible fonction de coût sur le jeu de validation) est celui obtenu lors de la 46ème époque. J'ai réalisé un deuxième modèle en suivant le même échantillonnage des séquences nucléotidiques et en utilisant les données d'expression issues de l'analyse de 10 tumeurs de 10 chiens (6 caniches et 4 golden retrievers). J'ai interrompu l'algorithme après la 145ème époque et le modèle conservé est celui de la 55ème époque.

Deux autres modèles ont été entraînés à partir de données d'expression provenant d'échantillons contrôles, c'est à dire d'une partie de tissu de muqueuse sain des chiens atteints du mélanome muqueux. Un modèle est basé sur l'apprentissage de l'échantillon d'un caniche, l'autre modèle est entraîné à partir de 10 échantillons (correspondant à 7 caniches et 3 golden retrievers). Les algorithmes ont été interrompus après 145 époques, le premier entraînement a retenu la modèle issu de la 20ème époque et le deuxième celui de la 40ème époque.

3.3.2. Test des modèles

Après avoir créé ces modèles, j'ai testé leur performance de prédiction sur des nouvelles données issues d'individus dont les données n'ont pas été utilisées dans la construction des modèles. J'avais à disposition 10 nouveaux échantillons de mélanomes muqueux provenant de 5 golden retrievers, 4 Labradors et 1 caniche. Je n'avais à disposition qu'un échantillon contrôle de mélanome muqueux de caniche et un autre de golden retriever. J'ai donc testé les modèles sur 5 autres prélèvements sains, sans lien avec le mélanome, à savoir des échantillons de truffes de trois labradors (ceux avec lesquels j'ai cherché une configuration optimale), un échantillon du coeur d'un beagle et un dernier échantillon provenant du cerveau d'un berger belge afin d'avoir plusieurs éléments de comparaison. L'auteur de Basenji mesure l'efficacité des modèles en déterminant la corrélation du log2 des expressions prédites avec celles mesurées expérimentalement (Tableau 4). Cette corrélation n'étant pas la même lorsque les données ne sont pas transformées, j'ai choisi de déterminer également la corrélation des valeurs non transformées en log2 (Tableau 5).

	Nb d'individus pour l'entraînment	Test sur données contrôles	Test sur données tumorales	MOYENNE
Entraînement sur données contrôles	**1**	0,35	0,302	0,326
	10	0,376	0,336	0,356
Entraînement sur données tumorales	**1**	0,291	0,25	0,27
	10	0,329	0,28	0,304
Moyenne		0,336	0,336	

Tableau 4. Comparaison de la corrélation entre le log2 de l'expression prédite par les quatre modèles et le log2 de l'expression mesurée expérimentalement sur les échantillons contrôles et les échantillons tumoraux. Par exemple, la première cellule (0.35) correspond à la moyenne des corrélations calculée depuis le log2 des prédictions de l'expression de 7 échantillons contrôles du modèle entraîné sur 1 échantillon contrôle et le log2 de l'expression des 7 échantillons contrôles mesurée expérimentalement. La dernière cellule (0.28) correspond à la moyenne des corrélations calculée depuis le log2 des prédictions de l'expression 10 échantillons tumoraux du modèle entraîné sur 10 échantillons tumoraux et le log2 de l'expression des 10 échantillons tumoraux mesurée expérimentalement.

Les modèles entraînés sur les échantillons contrôles proposent de meilleures corrélation (0.326 et 0.356) que ceux entraînés sur des données tumorales (0.27 et 0.304) (Tableau 4). De même, dans les deux cas, un entraînement sur 10 échantillons contrôles plutôt qu'un seul est meilleur (Pearson : 0,356 versus 0,326) ainsi que pour les données tumorales (Pearson : 0,304 versus 0,27). Par ailleurs, nous déterminons que la prédiction de l'expression en condition tumorale est supérieure avec les modèles contrôles (0,302 et 0,336) qu'avec les modèles tumoraux (0,25 et 0,28). Ces différences dans la qualité de prédiction des modèles est moins claire lorsque l'on observe le comportement de la corrélation calculée sur les valeurs non transformées en log2. En effet, on observe globalement les mêmes tendances, mais les écarts de prédiction d'un modèle à l'autre sont plus resserrés (Tableau 5).

	Nb d'individus pour l'entraînment	Test sur données contrôles	Test sur données tumorales	MOYENNE
Entraînement sur données contrôles	**1**	0,247	0,225	0,236
	10	0,254	0,228	0,241
Entraînement sur données tumorales	**1**	0,225	0,204	0,215
	10	0,246	0,217	0,232
Moyenne		0,243	0,219	

Tableau 5. Comparaison de la corrélation entre l'expression prédite par les quatre modèles et l'expression mesurée expérimentalement sur les échantillons contrôles et les échantillons tumoraux sans transformation en log2.

3.3.3. Analyse des résultats

Pour la création des 4 modèles, j'ai fait le choix de réaliser plus d'époque que l'auteur ne le préconise pour pouvoir observer et comprendre le comportement de l'algorithme. Pour cela, j'ai utilisé comme critère d'évaluation la valeur de la fonction de coût pour le jeu de validation et le jeu d'entraînement à chaque époque réalisée dans l'apprentissage des quatre modèles. Les hyperparamètres liés à ces réseaux de neurones sont identiques pour les quatre cas (Figure 6).

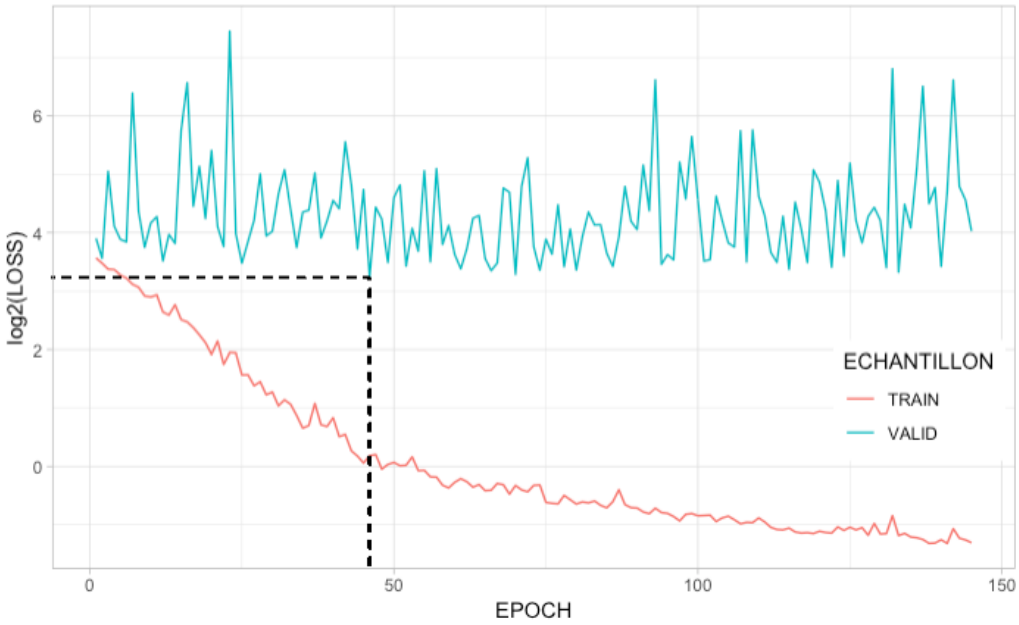

Figure 6. Evolution de la fonction de coût. Elle est calculée sur 145 époques (en abscisse) pour le jeu d'entraînement et le jeu de validation lors de l'entraînement du modèle basé sur un échantillon tumoral. La valeur de la fonction de coût (en ordonnée) est transformée au log2 pour faciliter la représentation.

Pour le jeu d'entraînement, la valeur brute de la fonction de coût est maximale à la première époque et vaut 11,83. Elle décroît progressivement jusque la fin de l'entraînement du modèle pour atteindre 0,40 à la 145ème époque. Pour le jeu de validation, la fonction de coût ne converge pas et oscille entre 9,39, minimum atteint à la 46ème époque et 175,43, maximum enregistré à la 24ème époque. Les fonctions de coût des trois autres modèles suivent la même tendance pour les deux jeux de données. Le réseau de neurones convolutif s'ajuste très bien aux données d'entraînement et permet à la fonction de coût de converger vers zéro (Figure 6). En revanche, pour le jeu de validation, la valeur de la fonction de coût est bien plus variable d'une époque à l'autre et ne converge pas vers un minimum global. Ces résultats mènent à réfléchir sur la structure du réseau et aux hyperparamètres associés, et nécessiteront davantage d'investigations.

J'ai utilisé les quatre modèles pour prédire l'expression des gènes associée aux échantillons tests (n=17 : 10 échantillons tests tumoraux et 7 échantillons contrôles) et réalisé plusieurs analyses pour comparer leur efficacité sous un autre angle. En premier lieu, j'ai regardé si le nombre d'échantillon qui compose le modèle avait un effet sur les mesures de corrélation. Les

modèles spécifiques au mélanome muqueux présentent une meilleure corrélation entre la cible et la prédiction lorsque le modèle repose sur l'apprentissage de dix échantillons. Cette différence est significative lorsque la corrélation est celle des valeurs transformées par log2 (ANOVA, p-value < 0,01) mais ne l'est pas sur les valeurs brutes (p-value = 0,061). J'ai réalisé les mêmes analyses en évaluant l'effet du type d'échantillon sur les modèles, à savoir tumoral ou contrôle. Cette fois, il apparaît de manière significative pour les deux indices de corrélation que les modèles entraînés sur l'expression de tissus tumoraux prédisent moins bien que les modèles entraînés sur les tissus contrôles (p-value < 0,001 dans les deux cas).

Enfin, j'ai analysé l'effet du modèle en lui même sur les corrélations calculées. Il est hautement significatif (p-value < 0,0001) sur la corrélation résultant d'un calcul sur les données transformées au log2 et en comparant les effets deux à deux (Figure 7). Il apparaît que le modèle entraîné sur 10 tissus contrôles est significativement meilleur que tous les autres modèles et que celui entraîné sur 1 échantillon tumoral est significativement moins bon que tous les autres. Le modèle entraîné sur 10 échantillons tumoraux et celui entraîné sur 1 échantillon contrôle sont les seuls modèles ne proposant pas de différence significative sur la corrélation entre leurs prédictions et les valeurs attendues transformées au log2.

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = V4 ~ MODELE, data = perf_mod)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
control10 - control1 == 0  0.004914   0.007518   0.654  0.91393
tumor1 - control1 == 0    -0.021324   0.007518  -2.836  0.03062 *
tumor10 - control1 == 0   -0.004834   0.007518  -0.643  0.91763
tumor1 - control10 == 0   -0.026238   0.007518  -3.490  0.00468 **
tumor10 - control10 == 0  -0.009747   0.007518  -1.296  0.56858
tumor10 - tumor1 == 0      0.016491   0.007518   2.193  0.13594
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Figure 7. Test de Tukey sur la corrélation des données log2. Analyse de l'effet des modèles deux à deux avec un test de Tukey sur la corrélation entre les valeurs prédites et attendues transformées au log2.

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = V5 ~ MODELE, data = perf_mod)

Linear Hypotheses:
                          Estimate Std. Error t value Pr(>|t|)
control10 - control1 == 0  0.03129    0.01002   3.123  0.01414 *
tumor1 - control1 == 0    -0.05472    0.01002  -5.461  < 0.001 ***
tumor10 - control1 == 0   -0.02143    0.01002  -2.139  0.15178
tumor1 - control10 == 0   -0.08601    0.01002  -8.584  < 0.001 ***
tumor10 - control10 == 0  -0.05272    0.01002  -5.262  < 0.001 ***
tumor10 - tumor1 == 0      0.03328    0.01002   3.322  0.00795 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Figure 8. Test de Tukey sur la corrélation des données brutes. Analyse de l'effet des modèles deux à deux sur la corrélation entre les valeurs prédites et attendues avec un test de Tukey.

Les analyses de la corrélation calculée sur les données brutes sont très différentes. En effet, même si on constate un effet significatif du modèle, en comparant l'effet des modèles deux à deux, seuls les tests impliquant le modèle entraîné sur un échantillon tumoral sont significatifs. Cela met en évidence le fait que ce modèle prédit particulièrement moins bien que les autres, mais ces derniers ne se démarquent pas entre eux.

Chapitre 4 : Conclusion et perspectives

Le projet de recherche de mon stage de Master 2 m'a tout d'abord amenée à découvrir la génétique, une discipline plutôt récente de la biologie et qui génère d'importantes quantités de données via les approches de séquençage haut-débit. Etant tout à fait novice dans le domaine, j'ai acquis les bases et les principaux concepts avec mes encadrants ainsi qu'avec l'ensemble des biologistes de l'équipe de Génétique du chien. J'ai également pu me rendre compte de l'intérêt du modèle animal et de la pertinence de celui du chien dans l'étude de certaines maladies dont les cancers et plus particulièrement, le mélanome de type muqueux. En effet, le mélanome muqueux est un cancer rare chez l'Homme (<1% de tous les cas de mélanome) dont l'investigation est un challenge important en recherche fondamentale et pour le développement de traitements. L'oncologie comparée, permet d'examiner entre plusieurs espèces le risque de cancer et le développement de tumeurs homologues. Elle est particulièrement pertinente pour l'étude des tumeurs humaines rares qui sont fréquentes chez le chien. L'analyse des cancers d'origine naturelle chez le chien constitue un modèle approprié pour faire avancer la compréhension, le diagnostic et la gestion du cancer chez l'Homme. La force du modèle canin se fonde sur sa structuration en races, véritables isolats génétiques au sein desquels une faible diversité génétique se reflète par de grandes régions du génome en déséquilibre de liaison et des effets fondateurs forts responsables des maladies homologues aux maladies humaines.

Le projet de recherche initié dans mon Master2 est d'utiliser et d'optimiser un outil de deep learning pour prédire l'expression des gènes et pour identifier les altérations génétiques responsables de la dérégulation de l'expression des gènes liée au développement des mélanomes chez le chien. Ces données permettront de rechercher les altérations et dérégulations les plus pertinentes dans les données génétiques chez l'Homme. Cet apprentissage des questions biologiques allait de paire avec l'autre aspect de mon stage qui était la prise en main et l'utilisation de l'outil Basenji. Pour parvenir à maîtriser l'outil, j'ai dû acquérir les grands principes des méthodes inhérentes au deep learning mais aussi travailler à comprendre le fonctionnement de l'algorithme sous-jacent. Je me suis familiarisée davantage avec le langage Python pour cerner la complexité des nombreux scripts qui composaient l'outil. C'est durant cette période d'apprentissage que j'ai pu déceler une anomalie dans l'écriture d'un script et ainsi pu améliorer la capacité de généralisation des modèles issus de l'utilisation de Basenji. J'ai ensuite adapté l'outil à l'analyse des données canines, en modifiant certains scripts. J'ai pu atteindre cet objectif mais les prédictions se trouvaient moins justes que celles que l'on pouvait obtenir chez l'Homme. Connaissant alors mieux l'outil, j'ai pu faire des choix quant à la composition des données et le taux d'utilisation des séquences permettant d'améliorer légèrement les prédictions de l'expression des gènes du chien.

Ce projet me permet de définir plusieurs pistes à examiner pour améliorer les prédictions d'expression des gènes. En effet, les différences dans la qualité de prédiction entre l'Homme et le chien peuvent tout d'abord s'expliquer par le type de données utilisées dans l'entraînement du modèle de deep learning. Basenji a été créé et optimisé pour prédire l'expression des gènes humains en se basant sur l'apprentissage de données de séquençage de type CAGE. Celles ci diffèrent des données RNA-seq canines que j'ai utilisées dans le cadre de mon stage, dans le sens où le CAGE annote l'expression en début de chaque gène alors que le RNA-seq couvre la totalité du gène. Cette dernière technologie engendre donc du bruit à traiter par l'algorithme dont la structure a été développé pour l'analyse d'informations moins abondantes et répondant à une logique de construction le long du génome. Une première possibilité pour optimiser la prédiction des données canines serait alors d'utiliser des données CAGE.

Basenji est configuré pour considérer un génome de référence et d'y associer l'expression d'un ou plusieurs individus ou tissus d'intérêt. Dans ce cas de figure, 80% du génome de référence et des expression associés sont utilisées dans le jeu d'entraînement. Cette façon de procéder semble évidente lorsque l'on dispose de peu d'échantillon. En revanche, il semble qu'il serait plus pertinent et efficace de prendre en compte plusieurs individus en incluant le génome de chacun d'entre eux avec leurs expressions géniques associées. Cela engendrerait un temps de calcul plus conséquent, mais de cette manière, nous pourrions utiliser 80% d'individus dans leur intégralité pour l'entraînement, 10% pour le jeu de validation et 10% pour le jeu test. En effet, les différences d'expression des gènes sont dues à des variations nucléotidiques pouvant varier d'un individu à l'autre. Ainsi, un modèle auquel on présente une nouvelle séquence du génome dont l'expression est à prédire serait plus performant.

Les algorithmes de deep learning sont réputés pour leur bonne assimilation des données hétérogènes. Nous envisageons d'introduire des données complémentaires aux données génétiques dans l'établissement d'un modèle de prédiction. Plusieurs données cliniques sont disponibles comme la taille, le poids, la race du chien ainsi que des informations relatives à la tumeur comme le stade de développement. La mise en place d'un tel projet demanderait une refonte complète de la méthode d'assimilation des données de l'outil Basenji mais dans le cadre de la poursuite du projet, nous proposons l'analyse intégrative des données hétérogènes (cliniques, génétiques et histologiques) des mélanomes buccaux canins comme modèle d'étude des mélanomes muqueux chez l'Homme. Les différents modèles de prédiction que

j'ai réalisés, aussi bien chez l'Homme que chez le chien, sont basés sur un jeu d'hyperparamètres communs, celui que propose l'auteur de Basenji. Il explique dans une publication (Kelley et al, 2018) avoir défini ces hyperparamètres grâce à une méthode d'optimisation bayésienne via la bibliothèque GPyOpt (https://github.com/SheffieldML/GPyOpt). L'optimisation des hyperparamètres représente un projet de recherche à part entière et mettre en place une méthode d'automatisation de leur découverte permettrait d'améliorer très nettement l'adéquation entre les valeurs prédites par un modèle et les valeurs attendues.

L'analyse des données de mélanomes muqueux canins a pour objectif d'améliorer la médecine vétérinaire mais aussi la médecine humaine. La très faible prévalence de la maladie chez l'Homme complique son analyse, d'où l'intérêt du modèle canin. C'est avec l'idée d'utiliser l'abondance de données chez une espèce pour mieux étudier un phénotype commun chez une autre espèce que David Kelley a très récemment proposé d'utiliser Basenji en utilisant cette fois ci les données de l'Homme et de la souris pour réaliser un seul modèle de prédiction (Kelley, 2019). Dans le cadre de la poursuite du projet, nous envisageons l'analyse intégrative des données des mélanomes muqueux de plusieurs espèces (Homme, chien) comme pour développer un modèle de prédiction performant des données canines.

Références bibliographiques

Bennani, el Fatemi, Erraghay, Mobakir, Ameurtess, Souuaf, 2013 « The primary melanoma of the female genital tract: report of three cases and review of literature » Pan Afr. Med. J. 16:58.

Bertin, 2017 « Linking FANTOM5 CAGE peaks to annotations with CAGEscan » *Sci. Data* 4:170147

Cadieu, Brito, Gillard, Abadie, Vergier, Guillory, Devauchelle, Degorce, Lagoutte, Hédan, M-D. Galibert, F. Galibert, André, 2014 « Analyse comparée des mélanomes chez le chien et l'homme » Bulletin de l'Académie vétérinaire de France, 167, N°3, pp. 213-220

Galibert, Quignon, Hitte, André, 2011 « Toward understanding dog evolutionary and domestication history » Comptes Rendus Biologies 334, N°3, 190-196.

Giger, Sargan, and McNiel, 2006 « Breed-Specific Hereditary Diseases and Genetic Screening ». Cold Spring Harbor Laboratory Press, 249–89.

Glickman, Domanski, 1983 « Mesothelioma in pet dogs associated with exposure of their owners to asbestos. » Environ Res 32(2): 305-313.

Gordon and Khanna, 2010 « Modeling opportunities in comparative oncology for drug development. » ILAR J 51(3): 214-220

Hafting, Fyhn, Molden and Moser, 2005 « Microstructure of a spatial map in the entireinal cortex » Nature, 436, 801-806

Hicks, Fooks ans Johnson, 2012 « Developments in rabies vaccines » Clin Exp Immunol, 2012, 169 (3): 199-204

Ilie, Long, Hofman, Lespinet, Bordone, Washetine, Gavric-Tanga, and Hofman. 2014. « Les Méthodes de Séquençage de « nouvelle Génération » (NGS) et Le Cancer Broncho-Pulmonaire: Principales Technologies, Applications et Limites Actuelles En Pathologie » *Revue Francophone Des Laboratoires* 2014 (458): 51–58

Kelley, Reshef, Bileschi, Belanger, McLean, Snoek, 2018 « Sequential regulatory activity prediction across chromosomes with convolutional neural networks » Genome Research 23(5):739-750

Kelley, 2019 « Cross-species regulatory sequence activity prediction » bioRxiv 660563

Lewis, Wiles, Llewellyn-Zaidi, Evans and O'Neill, 2018 « Longevity and mortality in Kennel Club registered dog breeds in the UK in 2014 » Canine Genetics and Epidemiology 5, 10.

Lindblad-Toh, Wade, Lander, 2005 « Génome sequence, comparative analysis and haplotype structure of the domestic dog ». Nature 438, 803-819

Poplin, Chang, Alexander, Schwartz, Colthurst, Ku, Newburger, Dijamco, Nguyen, Afshar, Gross, Dorfman, McLean, DePristo, 2018 « A universal SNP and small-indel variant caller using deep neural network » Nature Biotechnologie 36, 983-987.

Porrello, Cardelli, 2004 « Pet models in cancer research: general principles. » Exp Clin Cancer Res 23(2): 181-193

Shearin and Ostrander, 2010 « Leading the way: Canine models of genomics and disease » Disease Models and Mechanisms 3 (1-2): 27-34

Sheridan, 2014 « Illumina claims $1,000 genome win » Nat Biotechnol. 32(6): 507

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal, 2017 « Cancer Statistics, 2017. » *CA: A Cancer Journal for Clinicians* 67 (1): 7–30.

Zou, Huss, Abid, Mohammadi, Torkamani and Telenti, 2018 « A primer on deep learning in genomics » Nature Genetics 51: 12-18

	Diplôme : Master Spécialité : Sciences des données pour la biologie Spécialisation / option : Enseignant référent : Marie-Pierre Etienne
Auteur(s) : Camille Kergal Date de naissance* : 10/10/1995	Organisme d'accueil : IGDR Adresse : 2, avenue du Professeur Léon Bernard 35000 Rennes Maître de stage : Christophe Hitte
Nb pages : 20 Annexe(s) : 0	
Année de soutenance : 2019	

Titre français : Deep learning pour l'analyse génomique des mélanomes canins

Titre anglais : Deep learning for genomic analysis of canine melanomas

Résumé (1600 caractères maximum) :

Le mélanome muqueux est un cancer rare et agressif chez l'Homme. Sa faible prévalence freine son étude et sa compréhension et nécessite de s'orienter vers un modèle animal. Le chien représente un bon modèle car il développe naturellement des mélanomes muqueux avec une forte prévalence et de fortes similitudes cliniques avec l'Homme. Le projet de recherche que j'ai réalisé consiste en l'analyse génomique du mélanome muqueux canin avec l'objectif de développer la méthode pour prédire l'expression des gènes impliqués dans la tumorigénèse. Pour cela, j'ai mis en place un outil bioinformatique permettant de prédire l'expression des gènes chez l'Homme grâce à un algorithme de deep learning. Un premier résultat a été de transposer cet outil à l'analyse du génome canin. Puis, j'ai testé et déterminé la performance de prédiction de l'expression des gènes canins par tests de corrélation (Pearson r=0.4). La dernière étape du projet vise à améliorer la capacité de prédiction des modèles de deep learning réalisés grâce à l'outil en recherchant les hyperparamètres optimaux du réseau de neurones et en élargissant l'échantillon d'apprentissage du modèle.

Abstract (1600 caractères maximum) :

Mucosal melanoma is a rare and aggressive cancer in humans. Its low prevalence hinders its study and understanding thus requires moving towards an animal model. The dog represents a good model because it naturally develops mucosal melanomas with a high prevalence and strong clinical similarities with humans. The research project I carried out consists of genomic analysis of canine mucosal melanoma with the objective of developing the method to predict the expression of genes involved in tumorigenesis. For this, I set up a bioinformatic tool to predict the expression of genes in humans through a deep learning algorithm. A first result was to transpose this tool to the analysis of the canine genome. Then, I tested and determined the predictive performance of canine gene expression by correlation tests (Pearson r = 0.4). The final step of the project aims to improve the predictive ability of the deep learning models realized through the tool by searching for the optimal hyperparameters of the neural network and widening the learning sample of the model.

Mots-clés : deep learning, génomique, cancer

Key Words: deep learning, genomics, cancer

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*