EPIPREDICT

-

Apprentissage de modèles d'association entre épigénome et transcriptome

Master 2 Ingénierie statistique

Santa Kirezi

Référent universitaire :
Aymeric Stamm

Encadrants :
David Causeur
Gaël Le Trionnaire
Nadia Ponts

29 septembre 2023

Remerciements

J'aimerais remercier mes trois encadrants David Causeur, Gaël Le Trionnaire et Nadia Ponts de m'avoir accordé leur confiance et m'avoir permis de travailler sur le projet EPIPREDICT. Leur accompagnement et leurs encouragements ont été essentiels au bon déroulé de ce stage.

Je tiens également à remercier le corps enseignant du Master Ingénierie Statistique de Nantes Université pour la formation de qualité que j'y ai suivi.

Je tiens enfin à remercier les membres de l'unité pédagogique de mathématiques appliquées de l'Institut Agro Rennes - Angers pour leur accueil et leur contribution à un cadre de travail agréable et enrichissant.

Table des matières

Chapitre 1

Introduction

Les exigences de durabilité et le dérèglement climatique exposent le secteur agricole à de multiples défis. La politique scientifique de l'INRAE vise à générer des connaissances éclairées pour faire face à ces enjeux de manière argumentée. Ainsi la mission de mon stage s'inscrit dans un des métaprogrammes mis en place par l'INRAE pour favoriser les approches pluridisciplinaires de ces questions.

Plus précisément, le projet EPIPREDICT (UR MyCSA, Bordeaux, UMR IGEPP et UMR IRMAR, Rennes) s'intéresse à l'adaptation à court terme des bioagresseurs à des stress thermiques répétés, par des mécanismes épigénétiques. Les bioagresseurs englobent divers organismes vivants qui causent des dommages aux cultures, tels que les insectes nuisibles, les agents phytopathogènes et les mauvaises herbes [1]. Dans ce contexte, de nombreuses interrogations surgissent. Par exemple, comment un bioagresseur réagira-t-il à une future vague de chaleur ? S'adaptera-t-il en développant une résistance, ou au contraire, sera-t-il fortement impacté par les températures élevées, perdant ainsi son statut de menace ?

Ce rapport présente les travaux que j'ai effectué lors de mon stage de deuxième année du Master en Ingénierie Statistique de l'Université de Nantes. Je commencerai par exposer le contexte du projet EPIPREDICT et la structure d'accueil du stage. Les données ayant servi de support à ma mission de stage étant complexes, issues de méthodes de séquençage pour lesquelles pas ou peu de méthodes d'analyse existent, je décrirai leur format et les différentes étapes de pré-traitement mises en œuvre. Ensuite, je présenterai quelques hypothèses de recherche formulées à l'issue de quelques premières analyses exploratoires. Enfin, j'aborderai la question de l'apprentissage d'un modèle d'association entre épigénome te transcriptome.

1.1 Contexte du sujet

Le projet EPIPREDICT, développé par l'INRAE, aborde ces questions en se penchant sur deux bioagresseurs dotés de caractéristiques particulières de reproduction clonale et de remarquables capacités de résilience :

— Le puceron du pois, *Acyrthosiphon pisum*, qui cause des dégâts importants sur différentes légumineuses et démontre une plasticité phénotypique exceptionnelle en réponse à son environnement.
— Le champignon phytopathogène filamenteux, *Fusarium graminearum*, qui est responsable d'épidémies de fusariose de l'épi de blé à l'échelle mondiale. Il produit des mycotoxines et possède une capacité d'adaptation rapide aux changements environnementaux.

Pour mieux appréhender l'adaptation de ces organismes à leur environnement, le projet EPIPREDICT se focalise sur les variations épigénétiques. Ces changements héritables dans l'expression des gènes, sans altération de leur séquence, peuvent survenir en réponse aux fluctuations des conditions environnementales à court terme. Ils permettent aux organismes vivants de développer de nouveaux phénotypes afin de survivre et de prospérer.

Le code épigénétique régissant ces variations est complexe et composé de diverses marques. Son étude fait appel à des méthodes de séquençage haut-débit, générant d'abondantes données hétérogènes. Cependant, les méthodes actuelles d'analyse ne garantissent pas toujours une compréhension approfondie de ces données. L'objectif du projet EPIPREDICT est donc de concevoir des approches statistiques et mathématiques novatrices pour identifier les éléments expliquant les variations de l'expression génique. En fin de compte, décoder la manière dont les gènes réagissent à l'environnement pourrait servir de modèle décisionnel pour développer des agroécosystèmes résilients et économiquement viables.

Ainsi, EPIPREDICT est un projet pluridisciplinaire coordonné par les trois encadrants de mon stage :

— Nadia Ponts, chercheuse au sein de l'unité de recherche INRAE MycSA à Bordeaux, analyse les événements moléculaires conduisant à l'accumulation de toxines de Fusarium dans les grains de blé ou de maïs.
— Gaël Le Trionnaire, chercheur à l'unité mixte de recherche INRAE-Institut Agro-Université Rennes 1 IGEPP à Rennes, se concentre sur la compréhension des mécanismes moléculaires régulant la plasticité du mode de reproduction des pucerons.
— David Causeur, enseignant-chercheur à l'unité mixte de recherche CNRS IRMAR à Rennes, se consacre aux méthodologies statistiques pour diverses applications biologiques, notamment la gestion de la dépendance dans l'inférence statistique en grande dimension.

Le projet bénéficie d'une expertise diversifiée, chaque collaborateur apportant sa propre connaissance en méthodes mathématiques et en mécanismes épigénétiques à l'échelle moléculaire pour les deux espèces étudiées.

1.2 Structure d'accueil

Mon expérience de stage s'est déroulée à Rennes, au sein de l'unité pédagogique de Mathématiques Appliquées de l'Institut Agro Rennes-Angers. L'Institut Agro Rennes-Angers, avec l'Institut Agro Dijon et l'Institut Agro Montpellier, constitue l'une des trois écoles de l'Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement, également connu sous le nom d' « Institut Agro ». Créé en janvier 2020 par la fusion de plusieurs écoles nationales supérieures agronomiques déjà existantes, cet établissement public à caractère scientifique, culturel et professionnel (EPSCP) est sous la responsabilité principale du ministère de l'Agriculture et de l'Alimentation, tout en relevant de la tutelle pédagogique du ministère de l'Enseignement Supérieur.

Auparavant connu sous le nom d'Agrocampus Ouest, l'Institut Agro Rennes-Angers puise ses racines dans la ferme-école de Grand Jouan à Nozay, fondée en 1830 par Jules Rieffel et Charles Haetjens. L'établissement forme des ingénieurs spécialisés dans des domaines tels que l'agroalimentaire, l'agronomie, le paysage et l'horticulture, et joue un rôle de soutien auprès des établissements d'enseignement technique agricole.

Situé à Rennes, le campus de l'Institut Agro Rennes-Angers constitue le noyau central d'un pôle d'enseignement et de recherche agronomique multidisciplinaire influent. En partenariat avec INRAE depuis plus de 50 ans, les enseignants-chercheurs, les chercheurs INRAE, les acteurs économiques et les étudiants profitent d'installations expérimentales avancées et d'équipements de pointe. Cette collaboration tire parti des approches complémentaires de chacun et a contribué à faire du campus de Rennes l'un des principaux sites français pour la formation, la recherche et le développement en agronomie.

L'Unité Pédagogique de Mathématiques Appliquées, sous la direction de David Causeur, est composée de sept membres et est intégrée à l'équipe statistique de l'IRMAR. Les thèmes de recherche privilégiés au sein de cette unité portent sur des domaines tels que l'apprentissage de données en grande dimension, l'analyse de données génomiques, écologiques et de signaux biophysiques, l'analyse de données à haut débit, l'analyse de données multidimensionnelles, les données incomplètes et la sensométrie.

Chapitre 2

Format des données

2.1 Prérequis biologiques

Il convient tout d'abord de rappeler quelques notions de biologie moléculaire indispensables à la compréhension de la démarche d'analyse mise en œuvre dans la suite.

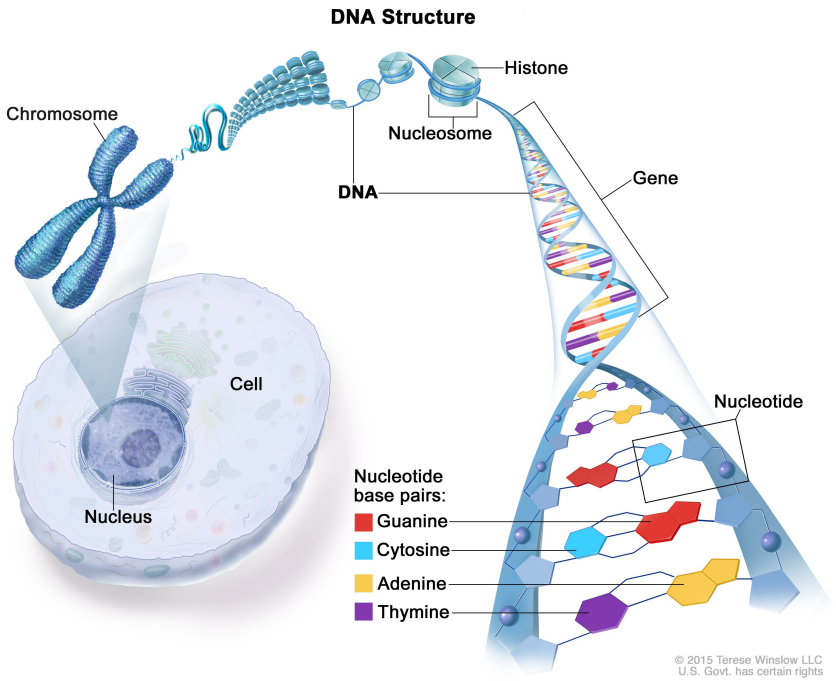


Figure 2.1 – *Illustration de la structure et de l'organisation de l'ADN*

L'**ADN** (acide désoxyribonucléique) est une molécule présente dans presque toutes les cellules eucaryotes, contenant l'ensemble des informations génétiques d'un organisme vivant. L'ADN est transmis lors de la reproduction et joue un rôle crucial dans l'hérédité. On qualifie l'ADN de bicaténaire car il est constitué de deux brins de polymères antiparallèles enroulés en une double hélice. Les briques fondamentales de ces brins sont les nucléotides, composés d'un sucre (désoxyribose), d'un groupe phosphate et d'une base azotée. Les bases azotées sont l'adénine (A), la thymine (T), la cytosine (C) et la guanine

(G). Des liaisons hydrogènes entre leurs bases azotées (A avec T et C avec G) associent les deux brins d'ADN. C'est l'enchainement spécifique des nucléotides qui forme le code génétique, organisé notamment en gènes (figure 2.1).

Les **gènes** sont des segments d'ADN contenant les instructions nécessaires à la synthèse de molécules (ARN puis éventuellement protéines) régulant les fonctions cellulaires. Cependant, les gènes ne deviennent actifs qu'après leur expression. Ils possèdent des parties codantes (exons) et non codantes (introns). Seuls les exons participent donc à la synthèse des protéines. Les gènes sont séparés par des régions intergéniques qui, pour la plupart, jouent un rôle dans la régulation de leur expression.

L'expression des gènes s'effectue en plusieurs étapes, la première étant la **transcription**. Durant cette phase, les séquences géniques sont copiées en molécules d'ARN par une enzyme nommée ARN polymérase. Les **ARN** (acides ribonucléiques) partagent de nombreuses similitudes avec l'ADN, mais présentent quelques différences majeures :

— L'ARN contient du ribose tandis que l'ADN contient du désoxyribose.
— La thymine de l'ADN est remplacée par l'uracile dans l'ARN, ayant la même capacité d'appariement avec l'adénine.
— L'ARN est généralement monocaténaire, c'est-à-dire à simple brin dans les cellules, tandis que l'ADN se compose de deux brins complémentaires formant une double hélice.
— Les ARN cellulaires sont plus courts que l'ADN du génome, avec des longueurs allant de quelques dizaines à quelques milliers de nucléotides, comparé à quelques millions à quelques milliards de nucléotides pour l'ADN.

Pendant la transcription, une enzyme, l'hélicase, sépare les deux brins de l'ADN, permettant à l'ARN polymérase d'agir. Celle-ci identifie et se lie à une région spécifique en amont de la région codante du gène, appelée site promoteur. Une fois cette phase d'initiation achevée, l'ARN polymérase copie la séquence du brin complémentaire et antiparallèle jusqu'au site terminateur, où l'ARN nouvellement synthétisé se détache.

Certains ARN, les ARN messagers (ARNm), sont traduits en protéines tandis que d'autres remplissent des rôles régulateurs ou structurels au sein de la cellule (ARN de transfert, ARN ribosomiques, etc.). Avant la traduction, l'ARNm doit subir une maturation post-transcriptionnelle. Étant généré sous forme de pré-ARNm incluant la séquence complète du gène, introns compris, il est ensuite soumis à une étape d'épissage : un complexe nucléoprotéique reconnaît les introns et les élimine. Les ARN messagers portent l'information génétique nécessaire à la synthèse d'une protéine et exprimée par des triplets de nucléotides appelés **codons**, chacun correspondant à un acide aminé. Certains codons, dits "codons STOP", ne correspondent pas à un acide aminé et signalent la fin de la traduction de l'ARN en protéine. Un seul gène peut donner naissance à plusieurs protéines différentes grâce au processus d'épissage alternatif des introns, qui permet de supprimer certains exons de l'ARN et générer ainsi différentes protéines.

Le **transcriptome** représente l'ensemble des ARN produits par la transcription, et son analyse permet d'identifier les gènes exprimés ainsi que leurs niveaux d'expression. La transcription est le niveau de contrôle principal de l'expression des gènes, et dépend

principalement de deux variables :

— La présence de régulateurs transcriptionnels tels que les facteurs de transcription
— L'état de condensation de l'ADN, qui détermine la capacité des régulateurs transcriptionnels à moduler l'expression génique.

Dans les organismes eucaryotes tels que les animaux, les champignons et les plantes, l'ADN se situe principalement dans le noyau cellulaire sous forme de **chromatine**. Cette structure (figure 2.1) englobe l'ADN et des protéines (histones et non-histones). Elle peut être très compactée, on parle alors d'hétérochromatine ou plus relâchée, c'est l'euchromatine. Les **chromosomes** constituent la forme la plus condensée de la chromatine. Avant toute action de la machinerie protéique, la chromatine doit être décompactée pour permettre l'accès à l'ADN.

Toutes les cellules d'un organisme partagent le même génome, mais certains gènes ne sont exprimés que dans des cellules spécifiques, à certaines étapes de la vie de l'organisme ou sous certaines conditions. La régulation de l'expression génétique est donc un mécanisme fondamental pour la différenciation cellulaire, la morphogenèse et l'adaptabilité d'un organisme face à des stimuli externes.

Un rôle important est joué par les **mécanismes épigénétiques**, qui interviennent à chaque étape de l'interprétation de l'ADN. Ce sont des mécanismes réversibles, transmissibles, adaptatifs qui n'altèrent pas la séquence des nucléotides de l'ADN. Ils consistent plutôt en des modifications chimiques de l'ADN et des protéines associées, régulant par exemple l'enroulement plus ou moins étroit de l'ADN autour des histones pour former des **nucléosomes**. La présence et la densité des nucléosomes est un facteur clé de la compaction de la chromatine et donc de la transcription. Leurs positions peuvent être cartographiées à l'aide de méthodes de séquençage à haut débit. L'**épigénome** est l'état épigénétique d'une cellule ; son étude a mis en évidence comment des éléments comme le comportement ou les facteurs environnementaux, peuvent influencer l'expression génique par adaptation biologique. Elle explique également comment certains traits peuvent être acquis, transmis ou même perdus d'une génération à l'autre.

Le **séquençage** est la méthode bio-informatique permettant de déterminer numériquement l'ordre des composants d'une macromolécule. On peut ainsi séquencer des ADN, des ARN et des protéines. En utilisant le séquençage d'ARN, on peut identifier et quantifier le transcriptome d'un organisme. Les techniques de séquençage à haut débit fournissent des données détaillées sur le transcriptome et l'épigénome d'un organisme.

2.2 Présentation des donnés

L'objectif du stage est d'ordre méthodologique. En effet, l'étude du rôle des mécanismes épigénétiques dans l'adaptation étant considérée sous l'angle de la recherche de motifs d'association entre épigénome et transcriptome, la mission consiste à poser les bases d'une démarche d'analyse statistique des relations entre données épigénomiques et données d'expressions de gènes. Pour concevoir et évaluer de telles démarches, on dispose de données mesurées par comptages de lectures de séquençage sur *Fusarium graminearum* PH-1

cultivé in vitro sans autre microorganisme ou souche.

Le dispositif expérimental complet intègre des individus soumis à différents stress thermiques, afin que les phénomènes d'adaptation puissent être observés. Toutefois, dans le cadre de ce stage, on se limitera à l'étude de trois réplicats biologiques, observés dans les mêmes conditions. Pour ces trois réplicats, les données épigénomiques sont obtenues par séquençage MAINE (micrococcal assisted isolation of nucleosomal elements) et les données transcriptomiques par séquençage ARN . Pour chaque gène, on dispose donc de données décrivant l'accessibilité de la chromatine sous forme d'un signal numérique le long du génome, signal épigénomique, et de données décrivant l'intensité de transcription, également sous forme de signal le long du génome. Notons qu'il est impossible d'associer un signal épigénomique pour un réplicat à un signal transcriptomique. Une des étapes de l'analyse consistera donc à réduire les signaux épigénomiques et transcriptomiques observées pour trois réplicats biologiques à un seul signal, débarrassé de la variabilité entre réplicats.

Dans les six tableaux :
— il y a 14145 lignes, chaque ligne correspondant à un gène
— chaque colonne correspond à une position de nucléotide
— les 800 premières colonnes sont les 800 positions en amont du codon ATG d'un gène, position de départ de toutes les séquences codantes eucaryotes
— les 800 dernières colonnes sont les 800 positions en aval du codon stop d'un gène, position de fin de toutes les séquences codantes eucaryotes.

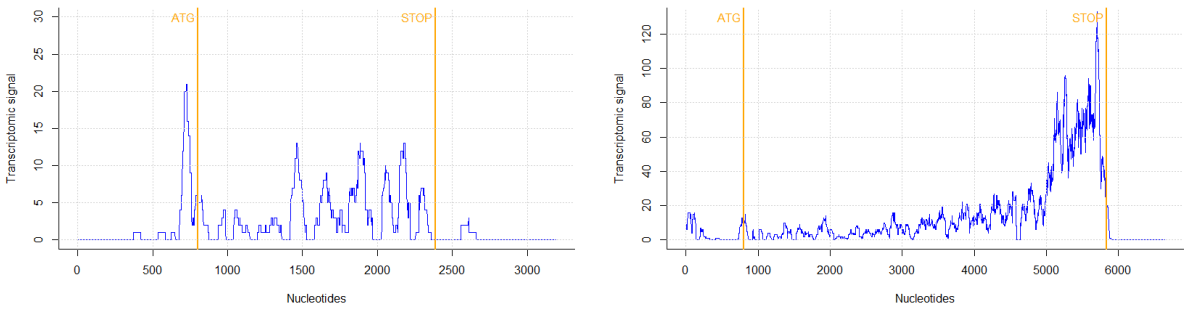


Figure 2.2 – *Exemples de signaux bruts transcriptomiques marqués des position de début et de fin de la région codante, pour deux gènes de longueurs différentes (3190 nucléotides à gauche et 6640 à droite)*

Le nombre de colonnes de ces six tableaux est très variable d'une ligne à une autre. En effet ce dernier dépend de la longueur du gène dans la zone codante entre les positions ATG et stop (figure 2.2). La structure génique (longueur, exons/introns, *etc.*) est également variable et décrite dans un septième tableau de données.

2.3 Pré-traitement des donnés

Le format des données brutes ne permet pas l'analyse des relations entre signaux épigénomiques et transcriptomiques par les méthodes statistiques usuelles.

D'une part, comme évoqué plus haut, pour chaque gène, il est impossible d'associer le signal épigénomique d'un réplicat au signal transcriptomique qui lui correspondrait. Une étape de prétraitement des données consiste donc à synthétiser des signaux mesurés sur trois réplicats biologiques. Cette étape s'appuyant sur l'analyse de la variabilité biologique observée pour chaque gène entre les réplicats, elle donnera lieu à la proposition d'une méthode pour identifier des données extrêmes.

D'autre part, le cadre usuel permettant une étude d'association entre des variables explicatives (ici le signal épigénomique) et des variables à expliquer (ici le signal transcriptomique) suppose que les mêmes variables soient mesurées pour tous les individus statistiques (ici les gènes). Or, comme évoqué plus haut, les signaux bruts étant mesurés en les positions nucléotidiques le long du génome, chaque gène a une longueur différente ou, en termes statistiques, les variables mesurées pour chaque individu ne sont pas les mêmes. Ce point sera également discuté dans la suite.

2.3.1 Gestion des réplicats biologiques et détection d'outliers

La variabilité entre réplicats biologiques est très différente selon les gènes. La figure 2.3 montre les signaux épigénomiques et transcriptomiques pour les trois réplicats de deux gènes, l'un pour lequel la variabilité biologique est importante, l'autre au contraire pour lequel cette variabilité est faible.

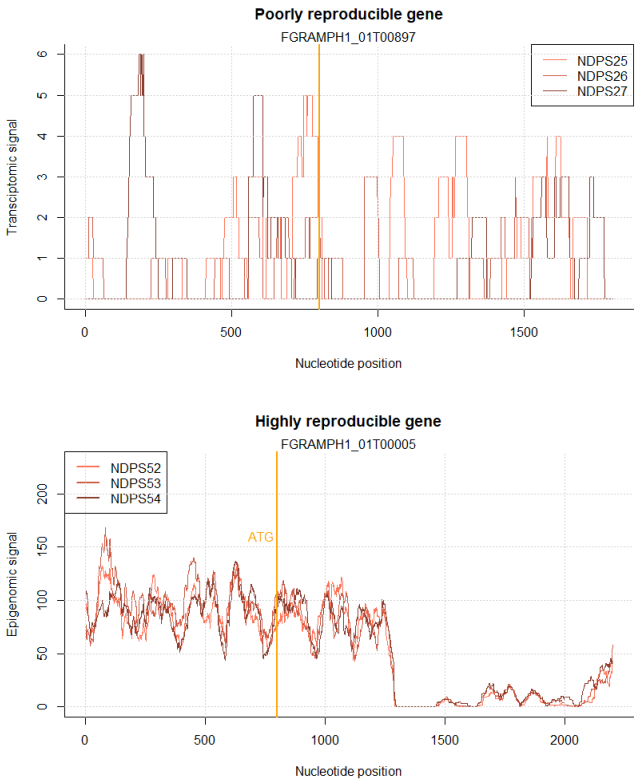


Figure 2.3 – *Superposition des trois réplicats de signaux transcriptomiques d'un gène très variable (haut) et épigénomiques d'un gène peu variable (bas)*

Afin que le calcul d'un signal moyen, synthétique des trois répétitions, soit représentatif d'une réalité commune aux réplicats, une règle permettant d'identifier des données faisant l'objet d'une variabilité biologique anormalement grande est adoptée. Si l'on veut adopter une règle d'évaluation du niveau de variabilité biologique commune à tous les gènes, l'écart-type mesuré le long du génome est un indicateur insuffisant. En effet, selon le niveau moyen d'un signal, un même écart-type peut signifier une grande ou au contraire une faible variabilité. On propose donc de fonder la règle d'identification sur le coefficient de variation(CV), à savoir le rapport entre écart-type et moyenne, mesuré le long du génome.

Plus la valeur de ce coefficient est élevée, plus la dispersion autour de la moyenne est grande. On considère qu'un coefficient de variation supérieur à 1 indique une grande variance relative. Enfin, le coefficient de variation médian est retenu pour obtenir un indicateur de variabilité synthétique pour chaque gène et chaque type de signal. La figure 2.4 présente un histogramme des médianes de ces coefficients de variation.

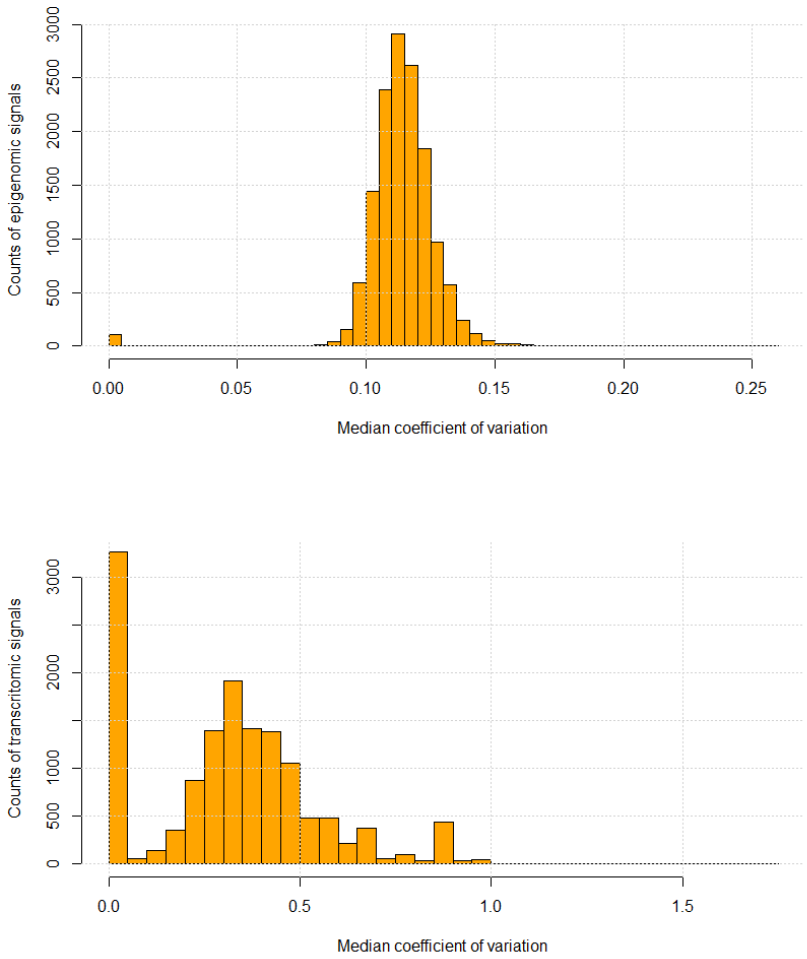


Figure 2.4 – *Distribution des coefficients de variation médians des signaux épigénomiques (haut) et transcriptomiques (bas) de tous les gènes de F.graminearum*

En observant la distribution des coefficients de variation médians des gènes, on peut remarquer que les signaux épigénomiques sont beaucoup moins variables que les signaux transcriptomiques.

On peut identifier 46 gènes ayant des signaux transcriptomiques fortement variables (CV>1) alors que leurs signaux épigénomiques le sont très peu. La figure 2.5 montre un histogramme des coefficients de variation médians des signaux transcriptomiques après suppression de ces 46 gènes.

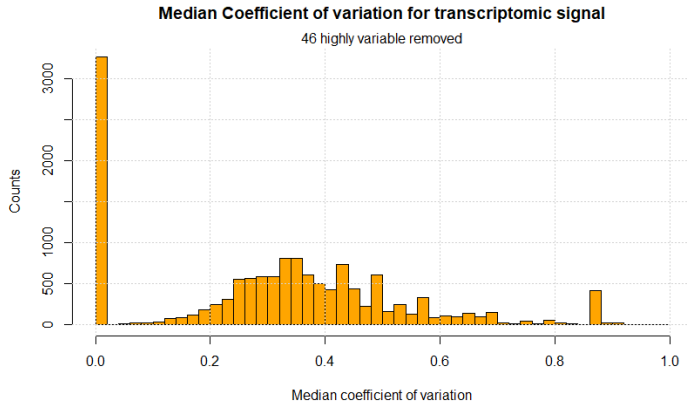


Figure 2.5 – *Distribution des coefficients de variation des signaux transcriptomiques sans les gènes extrêmement variables*

Cette analyse de la variabilité biologique entre réplicats m'a permis d'identifier des situations jugées pathologiques, de signaux épigénomiques égaux à zéro sur de longs intervalles, voire sur l'ensemble du gène (voir figure 2.6). Ces situations seront expliquées et gérées ultérieurement.

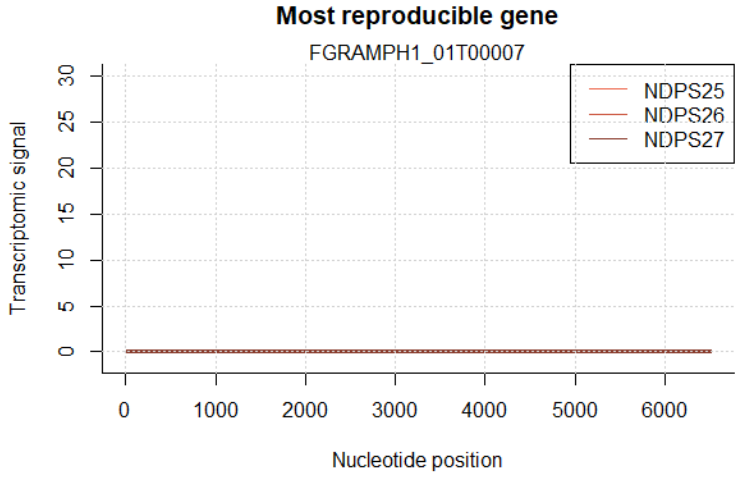


Figure 2.6 – *Distribution des coefficients de variation des signaux transcriptomiques sans les gènes extrêmement variables*

Les 6 tableaux ont alors été agrégés par moyenne des 3 réplicats en deux tableaux, les signaux moyens épigénomiques et les signaux moyens transcriptomiques. La faible dispersion des signaux épigénomiques fait que cette moyenne est en général représentative des 3 réplicats (figure 2.7).

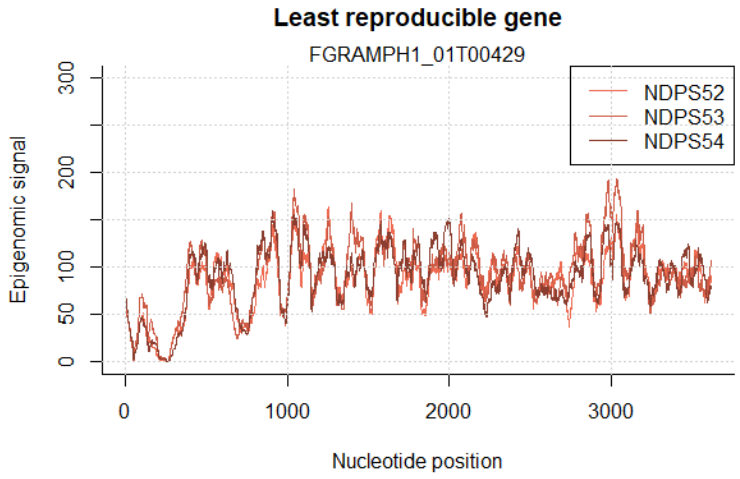


Figure 2.7 – *Illustration de la faible variabilité des signaux épigénomiques*

2.3.2 Distances intergéniques

Avec 14145 gènes répartis sur 36 millions de nucléotides, le génome de *Fusarium graminearum* est très condensé. Pour comparaison, l'homme a 26517 gènes répartis sur 3400 millions de nucléotides. Or les signaux de chaque gène ont été enregistrés sur 800 nucléotides avant et après leurs séquences codantes. Il faut donc vérifier si ces 800 nucléotides ne contiennent pas une partie ou la totalité d'un autre gène.

Pour cela, il faut vérifier que les régions intergéniques de *Fusarium*, qui peuvent avoir des longueurs très variables, sont plus longues que 800 nucléotides. Le fichier décrivant la structure des gènes donne accès aux positions de début et de fin des régions codantes des gènes sur les chromosomes. Ainsi, il est possible de calculer toutes les distances intergéniques du génome du fusarium et faire les remarques suivantes :

— Certaines distances sont négatives. En effet certains gènes se chevauchent selon diverses configurations [2] : deux gènes peuvent être superposés, un gène peut être inclus dans l'intron d'un autre gène, etc.

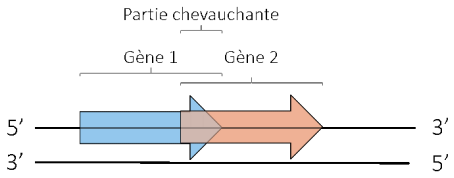


Figure 2.8 – *Superposition de deux gènes*

— Ces distances sont très variables, entre 1 et 66000 nucléotides, cependant 75% de ces distances se situent entre 400 et 1500 nucléotides. La distance intergénique médiane est de 790. Pour environ 50% des gènes, une partie plus ou moins grande de la séquence codante d'un autre gène se situe donc sur les 800 nucléotides en amont ou en aval de leur séquence codante.

Afin de limiter les interférences entre signaux de gènes trop proches, on a fait le choix de tronquer les signaux, passant de 800 nucléotides à 400 nucléotides en amont et en aval des séquences codantes. De plus, les 20% des gènes distants de moins de 400 nucléotides, ont été retirés de l'étude.

2.3.3 Signaux épigénomiques nuls

Certains gènes dans la base de données ont des signaux épigénomiques totalement ou en partie nuls (figure 2.9).

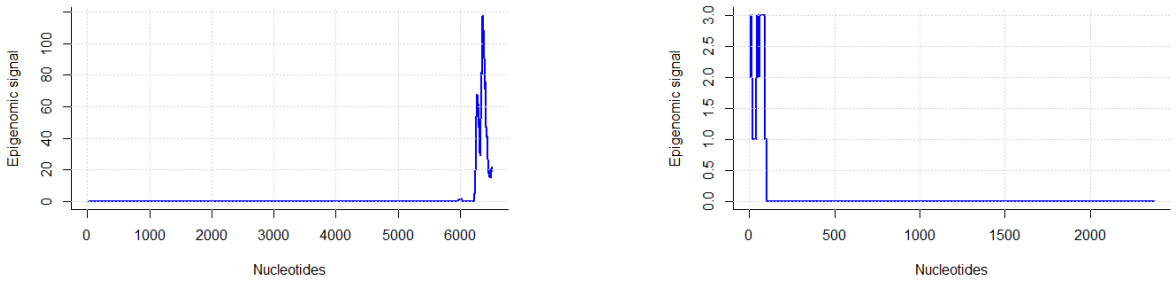


Figure 2.9 – *Deux exemples de signaux épigénomiques nuls étendus*

On n'a pas gardé ceux dont les signaux étaient complètement nuls car il n'y a pas de sens à expliquer un signal transcriptomique donné par un signal épigénomique nul. De plus, un signal épigénomique nul peut indiquer l'absence de ce gène chez l'individu souche cultivé.

Les gènes dont les signaux épigénomiques étaient en partie nuls, soit égaux à zéro sur des intervalles de 100 nucléotides ou plus, n'ont pas non plus été gardés. En effet, ces intervalles de signal nul peuvent indiquer que le gène correspondant a subi une insertion ou une délétion. Par cette démarche, environ 200 gènes ont été supprimés des données épigénomiques et transcriptomiques.

2.3.4 Uniformisation des longueurs de gènes

Comme évoqué plus haut, étudier l'association entre les signaux épigénomiques et transcriptomiques se confronte au fait que la longueur des signaux est variable d'un gène à l'autre. Cette variabilité ne porte que sur la partie codante, la longueur des parties en amont et en aval ayant été fixées à 400 nucléotides.

Une approche possible pour créer un tableau de données avec un même nombre de variables pour chaque individu (gène) consiste à synthétiser chaque signal par des indi-

cateurs invariants du support physique du gène : longueur du signal, moyenne du signal, écart-type du signal, moyenne de la dérivée du signal, *etc.* Cette méthode suppose que l'on ait des a priori biologiques permettant d'orienter le choix d'indicateurs pertinents. On lui préfèrera une autre approche dans la suite, consistant à conserver l'information brute portée par chaque signal en les contraignant à avoir un support commun à tous les gènes.

L'interpolation est une opération mathématique permettant de remplacer une courbe ou une fonction par une autre courbe (ou fonction) plus simple, mais qui coïncide avec la première en un nombre fini de points donnés au départ. C'est donc un outil de choix pour exécuter l'approche souhaitée. La longueur de ce support commun a été fixée de manière arbitraire à 2000 positions, 75% des gènes ayant une longueur comprise entre 900 et 7300 nucléotides. Ainsi par exemple, pour chaque gène, la position 20 correspond donc à un écart par rapport au démarrage de le séquence codante de 1% de la longueur totale du gène. La mdr j'ai pas tout compris

Pour chaque gène, le support des signaux a donc été défini comme une grille commune allant de 1 à 2000. Or, les valeurs des signaux étant observées sur une grille plus ou moins dense (selon le gène considéré) que la grille commune avec 2000 nœuds, on ne dispose en général pas des valeurs des signaux aux positions de la grille commune. Pour reconstruire ces valeurs, les points correspondant à des valeurs observées des signaux autour d'une valeur de la nouvelle grille commune ont été joints par un segment de droite et la valeur en ordonnée sur cette droite au point de la grille commune a été retenue comme une approximation du signal en cette position (fonction `approx` sous R). ????

Afin d'évaluer l'écart entre les signaux bruts et interpolés, on calcule pour chaque gène et chaque type de signal un coefficient de corrélation au carré (R^2) entre les signaux bruts et reconstitués sur leur support original à partir des valeurs interpolées. Pour les deux types de signaux, ces mesures varient entre 0.91 et 1 avec des médianes à 0.99, indiquant une très bonne qualité.

Sur la figure 2.10 on peut voir les signaux d'expressions interpolé et original sur une même grille afin de pouvoir visualiser la qualité de l'interpolation pour deux gènes en particulier.

Les parties en amont et aval des séquences codantes n'ayant pas été modifiées, les tableaux de signaux uniformisés ont donc 2800 colonnes (400+2000+400).

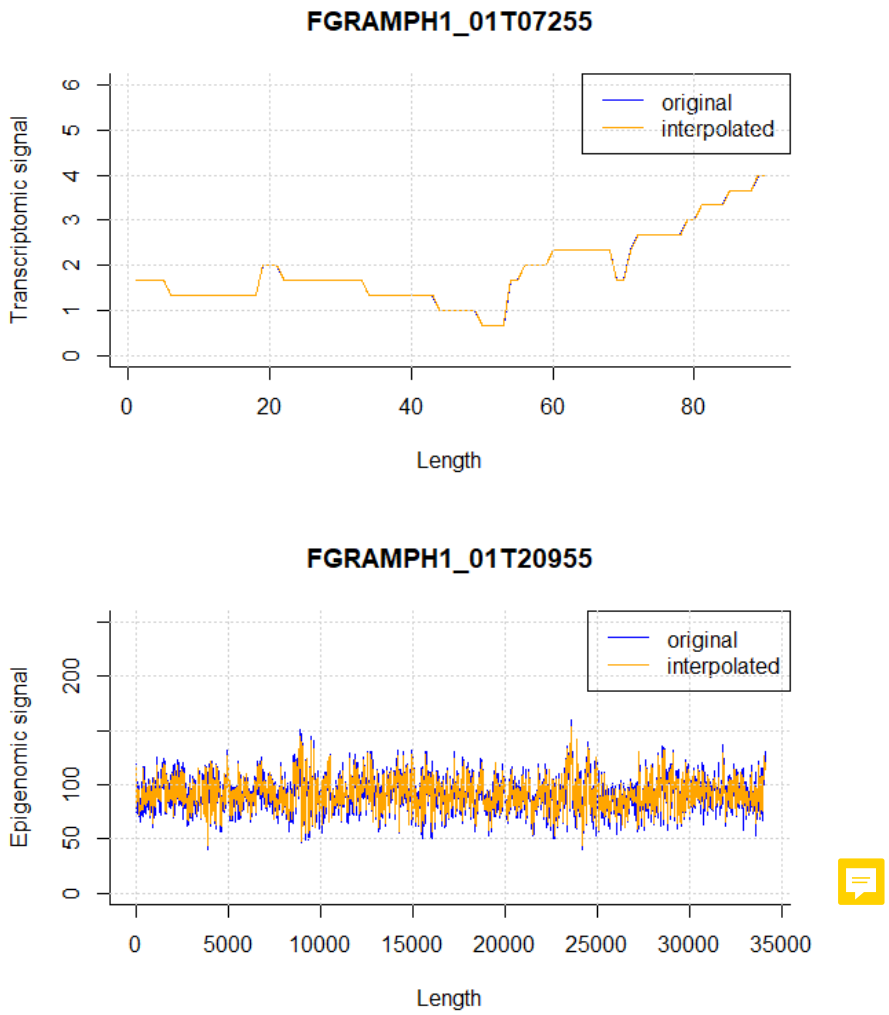


Figure 2.10 – *Superposition des signaux originaux et de signaux reconstruits à partir de l'interpolation*

2.3.5 Réduction de dimension par lissage B-spline

Le prétraitement fait précédemment a permis dans une certaine mesure de réduire les dimensions initiales des données. La régularité des signaux laisse penser qu'il est possible de réduire encore la dimension des profils épigénomiques et transcriptomiques par des méthodes de lissage.

Cette réduction vise à faciliter la tâche d'entrainement d'un modèle associant signaux épigénomiques et transcriptomiques. L'objet de ce lissage est bien de réduire le nombre de variables mais dans une approche conservatrice, garantissant un faible écart entre les signaux lissés et non-lissés. Il existe de nombreuses méthodes empiriques de lissage, telles que les moyennes mobiles, le lissage exponentiel simple ou double, etc. Les splines de lissage sont des fonctions polynomiales définie par morceaux faisant partie de ces méthodes empiriques.

On propose d'approcher chaque signal par une fonction spline cubique. Une telle fonction est définie comme un polynôme de degré 3 sur chaque intervalle d'une partition de la plage d'observation du signal (de 1 à 2000), la régularité du signal lissé étant garantie par des contraintes de continuité, dérivabilité et même continuité de la dérivée en chaque point de contact entre deux intervalles successifs de la partition. La régularité du signal lissé est d'autant plus grande que le nombre d'intervalles de la partition est faible. Dans le cas présent, on a choisi un nombre suffisamment grand d'intervalles dans la partition, de telle sorte que le carré du coefficient de corrélation entre valeurs lissées et non-lissées soit supérieur à 0.99 pour 96% des gènes.

En pratique, l'espace des fonctions splines cubiques associées à une partition étant un espace vectoriel, on peut reconstruire chaque fonction spline par combinaison linéaire de fonctions appelées B-splines, constituant une base de cet espace. La figure 2.11 représente sept B-splines (courbes en couleurs) recouvrant un intervalle allant de 0 à 1. Sur cette Figure, la courbe en noire est une fonction spline, que l'on peut obtenir comme combinaison linéaire des sept B-splines.

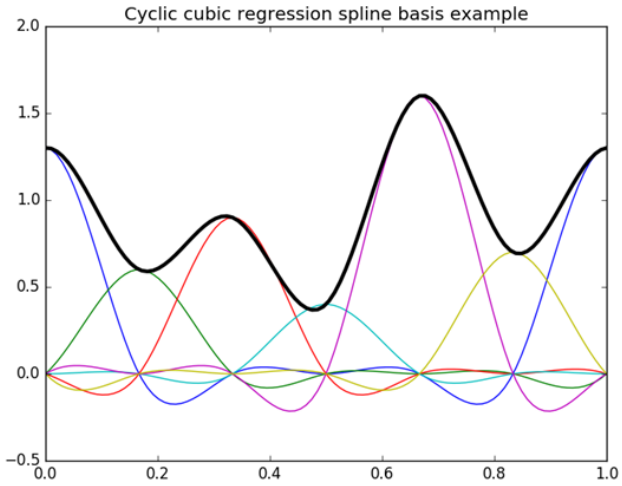


Figure 2.11 – *Illustration de la faible variabilité des signaux épigénomiques*

Les supports de chaque B-spline sont des unions d'intervalles contigus de la partition, les supports de B-splines successives se recouvrant partiellement. Ainsi, ces supports couvrent de manière ordonnée toute la partition et la dimension de l'espace (le nombre de B-splines) est directement lié au nombre de classes de la partition et donc à la régularité de la fonction spline. L'implémentation des B-splines est possible sous R grace au package `splines`.

Chaque signal lissé pouvant être reconstitué à partir de la connaissance des coefficients associés à chaque B-spline dans cette combinaison linéaire, on appellera signal comprimé la suite ordonnée des coefficients des B-splines pour la fonction spline approchant au mieux le signal observé. Ainsi, avec 700 coefficients de B-splines pour les signaux épigénomiques et 500 pour les signaux transcriptomiques, on obtient des approximations des signaux bruts par les signaux lissés associés à des carrés de coefficients de corrélation supérieurs à 0.99 pour 96 % des gènes.

Chapitre 3

Analyse exploratoire des données

3.1 Analyse des corrélations spatiales

La figure 3.1 est une représentation en image de la matrice de corrélation entre les valeurs successives du signal épigénomique comprimé. A chaque valeur de corrélation est associé un niveau de couleur entre bleu (-1) et rouge (1). Les corrélations représentées ici peuvent donc être vues comme distribuées spatialement car les variables dont on mesure les corrélations sont ordonnées selon leur position sur le génome. La Figure révèle une forte structuration des corrélations spatiales en trois blocs, correspondant aux trois parties du gène (amont, codante et aval). Au sein de chaque bloc, particulièrement celui de la partie codante, on observe une forte auto-corrélation. Par ailleurs, une corrélation négative apparait en fin de partie amont entre le bloc amont et le bloc codant et en fin de partie codante avec le bloc aval.

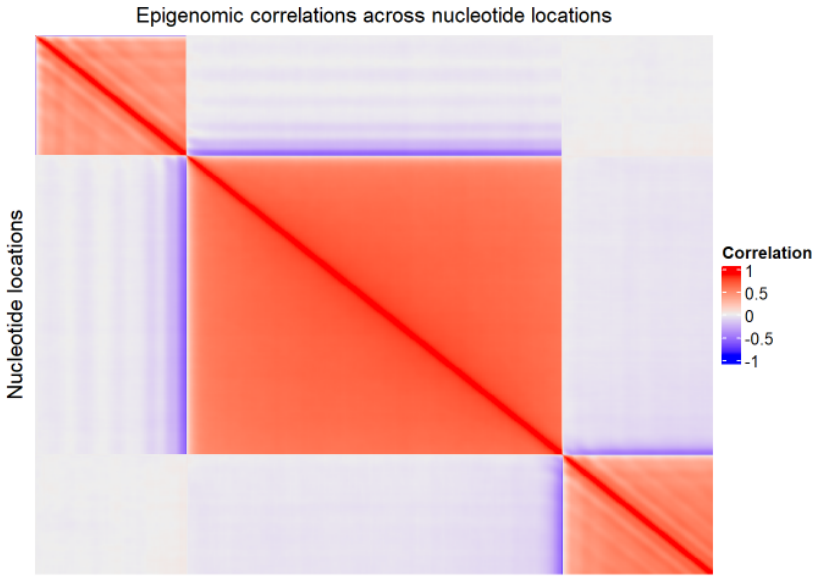


Figure 3.1 – *Représentation de la matrice des corrélations spatiales épigénomiques*

La matrice de corrélation du signal transcriptomique comprimé représentée en 3.2 ne présente que la partie codante du gène. C'est en effet sur cette partie uniquement qu'est

mesurée le niveau d'expression. La matrice de corrélation comprend cette fois des valeurs toutes positives. On peut y remarquer une forte corrélation entre positions proches qui décroît au fur et à mesure que des positions sont éloignées.

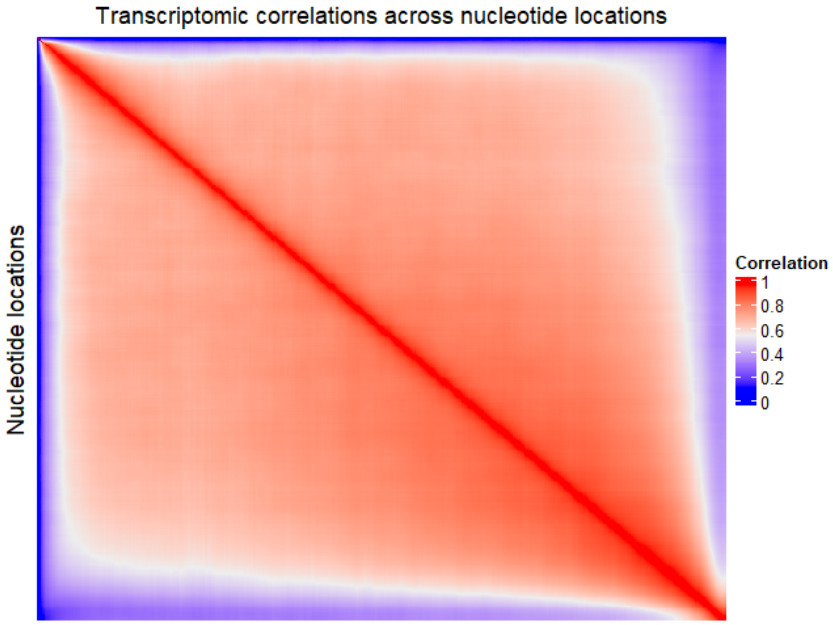


Figure 3.2 – *Représentation de la matrice des corrélations spatiales transcriptomiques*

La matrice de corrélation entre les signaux comprimés épigénomiques et transcriptomiques (uniquement pour la partie codante) représentée en 3.3 comprend des valeurs variant entre -0.4 et 0.4. Cette figure fait ressortir un bloc central de corrélations positives entre les signaux comprimés épigénomiques et transcriptomiques s'étalant sur les régions codantes. Ce bloc est encadré de deux marges de corrélations négatives ; les plus fortes apparaissant au niveau des positions en amont de la région codante et les plus faibles en aval de celle-ci.

Les structures de corrélation observées ci-dessus correspondent à une réalité biologique : le lien entre la déplétion des nucléosomes et le niveau d'expression d'un gène[3]. Sur un gène, des portions dépourvues de nucléosomes indiquent une accessibilité de l'ADN indispensable à la transcription[4]. De plus des études ont montré que les régions où ces déplétions de nucléosomes sont les plus importantes et influent le plus sur l'expression sont les sites promoteurs, positionnés en amont des codons ATG. Ces régions sont aussi présentes dans une moindre mesure au niveau des codons STOP[5]. Ainsi une forte déplétion de nucléosomes en amont et en aval de la région codante d'un gène favorise un fort niveau d'expression. Au contraire, une déplétion faible, voire une réplétion de nucléosomes indique un gène peu ou pas exprimé.

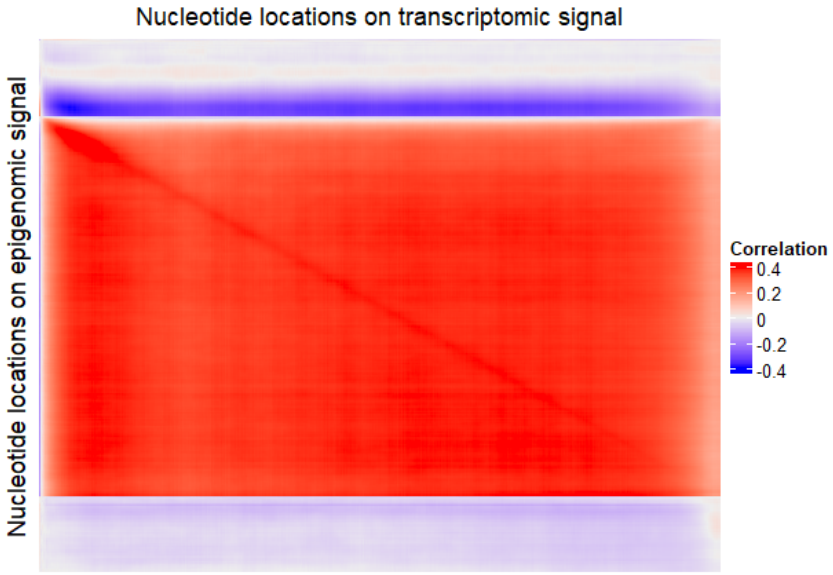


Figure 3.3 – *Représentation de la matrice des corrélations spatiales entre l'épigénome et le transcriptome*

3.2 Mise en évidence des classes

Dans le prolongement des observations précédentes, l'objectif de cette partie est d'identifier des signaux épigénomiques présentant des motifs qui seraient caractéristiques de fort niveau d'expression. Par exemple, en lien avec l'étude précédente portant sur les corrélations spatiales entre signaux, on peut s'attendre à identifier une classe de gènes dont les signaux épigénomiques ont un pic négatif large juste avant la partie codante, cette classe étant normalement caractérisée par un fort niveau d'expression.

La classification de données de courbes fait l'objet de nombreux développements de la part d'équipe de recherche en statistique. En effet, la forte corrélation entre les mesures des courbes sur une grille discrétisée, le grand nombre de points de mesure, et plus généralement le caractère fonctionnel de ces données donnent à la tâche de classification une difficulté plus importante que pour des données plus classiques.

Dans la suite, on choisit de procéder en deux étapes : d'abord une réduction des données par analyse en composantes principales, en poursuivant le même principe de réduction conservative que celui utilisé pour la réduction par l'approximation spline, et une classification ascendante hiérarchique des gènes par une méthode standard à partir des premières composantes principales. Une telle succession de méthodes (on parle de classification spectrale) est prévue dans le package FactoMineR de R, qui a été utilisé dans le cas présent.

3.2.1 Analyse en composantes principales (ACP)

L'analyse en composantes principales est un des outils de réduction de dimension de données les plus populaires. Elle consiste à une projeter les données dans un sous espace de dimension plus petite. Elle transforme les variables originelles qui peuvent être corrélées

en nouveau système de variables orthogonales et donc décorrélées : les composantes principales. Ces composantes sont ordonnées, la première étant celle qui restitue le plus la variance des données. Typiquement, cette méthode est utilisée en réduisant la dimension à 2 ou 3 composantes principales permettant une visualisation claire et pertinente des données et de leur structure.

Ici, il a cependant été décidé de garder autant de composantes principales que nécessaires pour garder beaucoup de variance expliquée par celles-ci. Ainsi pour garder au moins 95% de variance expliquée, 80 composantes ont été gardées pour les coefficients épigénomiques et 40 composantes pour les coefficients transcriptomiques. Ces composantes principales ont ensuite été utilisées comme des variables pour établir une classification des gènes.

3.2.2 Classification

La classification est une méthode d'analyse des données non supervisée permettant de regrouper des objets (individus ou variables) en classes. Les objets au sein d'un groupe doivent être le plus semblable possible tandis que d'un groupe à l'autre ils doivent être le plus distincts possible. C'est une méthode qui permet une analyse descriptive des données en révélant des structures sous-jacentes. La classification ascendante hiérarchique (CAH) est l'une des nombreuses méthodes de classification. A l'aide d'une matrice de distances qui mesure les dissimilarités entre les objets à classifier, elle regroupe ces derniers itérativement en classes de plus en plus grande, construisant ainsi un arbre de classification ou dendrogramme. C'est alors le niveau de coupe du dendrogramme qui détermine le nombre de classes choisi (figure 3.4)

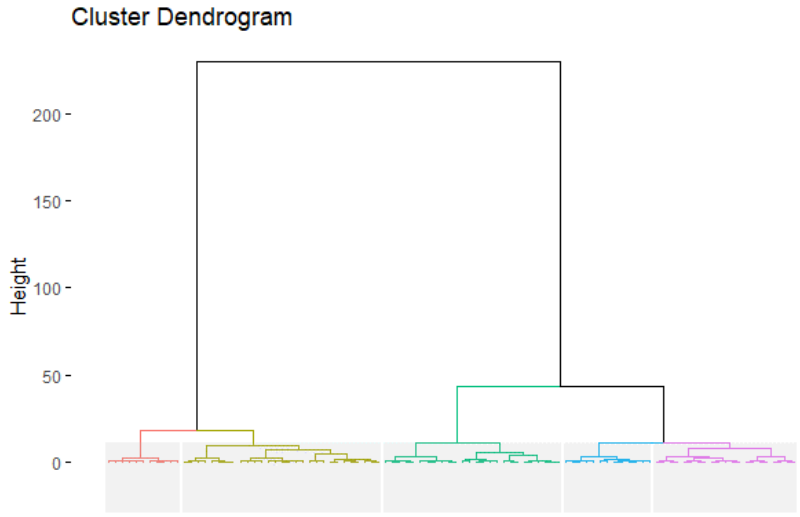


Figure 3.4 – *Exemple de dendrogramme pour une classification en cinq classes*

La CAH a été appliquée sur les composantes principales des signaux épigénomiques compressés à la suite d'un pré-regroupement en 100 classes par la méthode des k-means. Ce pré-regroupement vise à faciliter la classification hiérarchique en évitant le calcul d'une matrice de distances entre 14000 gènes. Cette classification mixte a permis de traiter

le nombre important de gènes à classifier rapidement et de consolider la classification obtenue.

Il existe plusieurs règles permettant de déterminer le nombre de classes optimal à choisir, basés sur différents critères : critère des silhouettes, critère du coude, etc. Ces méthodes ont été mises en œuvre mais les classifications obtenues à partir des nombres de classes indiqués n'étaient pas suffisamment interprétables. La figure 3.5 représente les signaux moyens épigénomiques et transcriptomiques au sein des classes pour un choix de trois et de dix classes.

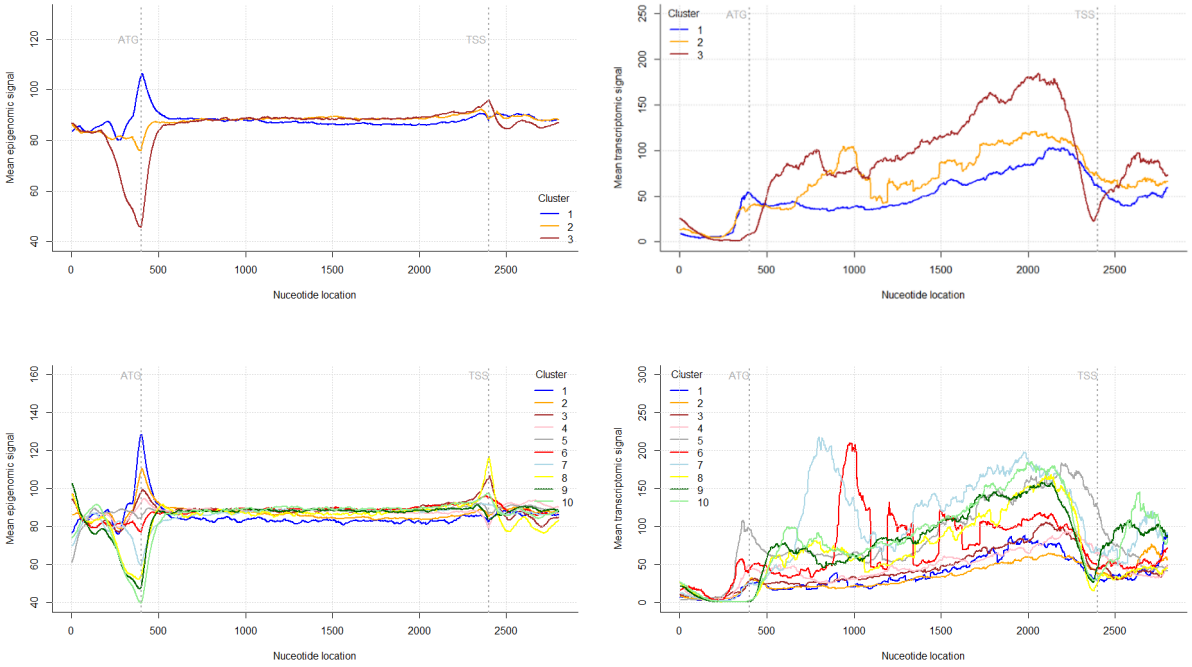


Figure 3.5 – *Signaux épigénomiques (à droite) et transcriptomiques (à gauche) moyens pour 3 (haut) et 10 (bas) classes de gènes*

L'interprétabilité des signaux a finalement été privilégiée pour retenir cinq classes. Ce nombre permet de distinguer des profils épigénomiques et transcriptomiques moyens suffisaments divers, distincts et interprétables. En effet, les profils moyens que l'on peut observer à la figure 3.6 correspondent à une certaine réalité biologique déjà visible avec les corrélations des signaux comprimés.

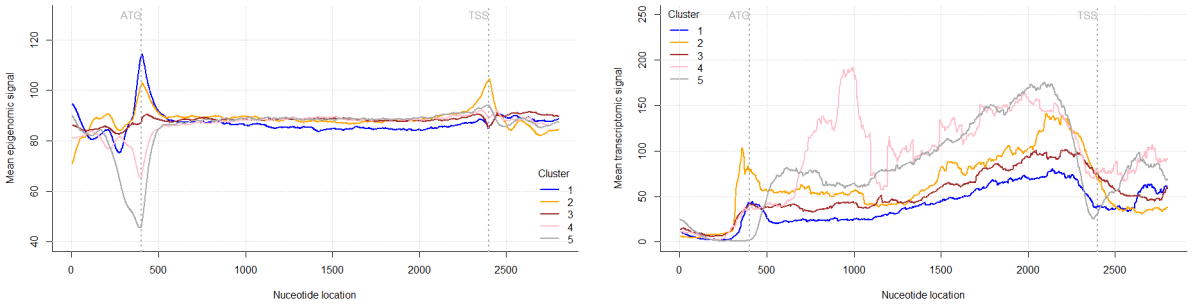


Figure 3.6 – *Signaux épigénomiques (à droite) et transcriptomiques (à gauche) moyens pour 5 classes de gènes*

On peut faire les constats généraux suivants sur les cinq classes de gènes établies

- Classe 1 : le signal épigénomique moyen présente un pic positif de grande amplitude au niveau du codon ATG, indiquant une forte présence de nucléosomes correspondant au profil transcriptomique moyen le plus faible.
- Classe 2 : leur profil épigénomique moyen présente un pic positif un peu moins important et un profil transcriptomique moyen
- Classe 3 : le profil épigénomique moyen est assez monotone tout au long des gènes avec un profil transcriptomique moyen assez faible
- Classe 4 : leur profil épigénomique moyen présente un pic négatif, indiquant une déplétion de nucléosomes, et un profil transcriptomique fort
- Classe 5 : leur profil épigénomique moyen présente un pic négatif de grande amplitude et le profil transcriptomique moyen le plus fort

Chapitre 4

Apprentissage d'un modèle de prédiction du transcriptome par l'épigénome

La problématique de l'apprentissage des données d'expression par les signaux épigénomiques peut s'envisager de différentes manières. Le point de vue usuel consiste à réduire le signal transcriptomique en une seule valeur, considérée comme le niveau d'expression du gène. La mesure la plus classique est la somme des valeurs du signal le long de la partie codante du gène. Se posent alors les questions de normalisation de cette valeur, notamment pour prendre en compte les variations de longueurs des gènes. Ces questions de normalisation ne seront pas discutées dans ce rapport.

En choisissant cette option, le problème devient celui de la prédiction d'une valeur réelle, un scalaire, par un signal (*function-to-scalar regression*), le signal épigénomique. Toutefois, une même valeur scalaire d'expression peut masquer différents motifs du signal transcriptomique : une répartition plus ou moins régulière de l'expression le long de la partie codante, une répartition concentrée autour du démarrage de la partie codante, autour de sa partie finale, *etc.* A l'appui de cette remarque, il est intéressant d'observer que la somme des valeurs du signal transcriptomique brut (comptages en nombres entiers) ne prend qu'environ 5000 valeurs différentes sur l'ensemble des 14000 gènes.

Une autre approche consiste donc à tenter de prédire le signal transcriptomique lui-même à partir du signal épigénomique (*function-to-function regression*). Ainsi, on espère d'une part profiter des corrélations spatiales entre les deux types de signaux mesurées sur un même support physique et d'autre part reconstituer des niveaux d'expression à partir des signaux transcriptomiques prédits .

Enfin, dans une approche plus exploratoire et moins ambitieuse de classification, on peut aussi réduire la tâche à la prédiction de classes de signaux transcriptomiques, correspondant à des motifs de signaux ou des niveaux d'expression différents. Cette approche de classification n'ayant pas été réalisée au cours du stage, elle ne sera pas développée dans la suite.

4.1 Méthodes classiques d'apprentissage statistique

Dans le but de réduire les temps de calcul et de ne pas dépasser les capacités de mémoire du matériel informatique à disposition, les méthodes décrites ci-dessous ont été appliquées à des composantes, moins nombreuses, des signaux comprimés.

Au niveau épigénomique, il est important de pouvoir récupérer une grande partie de l'information contenue dans les signaux, un nombre assez élevé de composantes principales obtenues par ACP a donc été choisi. En l'occurrence, les 200 composantes principales épigénomiques retenues cumulent plus 98% de variance expliquée des données.

En contraste, au niveau transcriptomique les composantes ont été obtenues par la méthode de régression PLS (Partial Least Squares) afin de favoriser dans leur construction la corrélation avec les niveaux d'expression des gènes. Le package R `plsr` permet d'implémenter cette méthode et d'obtenir par validation croisée le nombre optimal de composantes PLS maximisant cette corrélation.

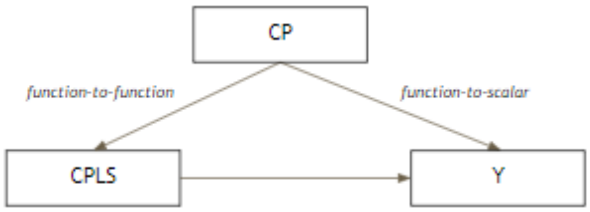


Figure 4.1 – *Schéma de modélisation de l'épigénome à partir du transcriptome*

La stratégie de modélisation décrite au début de ce chapitre peut alors être résumée dans la figure 4.1 où :

— CP est la matrice de design de dimension (10807,200) formée par les composantes principales épigénomiques
— Y est la variable cible déterminée par les niveaux d'expression des 10807 gènes
— $CPLS$ est la réponse multivariée de dimension (10807, 6) constituée des 6 composantes PLS transcriptomiques optimales.

4.1.1 Méthodes de régression pénalisée

Parmi les méthodes d'estimation de modèles de régression en grande dimension, la pénalisation du critère des moindres carrés, plus généralement de la vraisemblance, fait partie des approches les plus populaires, en particulier dans le domaine de la chimiométrie où les données sont souvent fonctionnelles (issues de méthodes de spectroscopie).

Le principe général de ces méthodes d'estimation est de minimiser un critère que l'on peut voir comme la classique somme des carrés des écarts entre les valeurs observées de la variable à expliquer et le modèle (le score linéaire), à laquelle s'ajoute un terme de pénalisation pouvant prendre plusieurs formes.

Les formes les plus connues sont un terme scalaire positif, généralement appelé paramètre de pénalité, multiplié par la somme des carrés des coefficients du score linéaire ou la somme des valeurs absolues de ces coefficients[6].

Dans le premier cas, la méthode d'estimation s'appelle régression Ridge et dans le second cas la régression Lasso . Les sommes des carrés ou des valeurs absolues des coefficients du score linéaire peuvent être vues comme des indicateurs de la complexité du modèle. Ainsi, en pratique, le paramètre de pénalité est utilisé pour donner plus ou moins d'importance à la réduction de la complexité du modèle, au regard de la qualité d'ajustement mesurée par la somme des carrés des écarts entre les valeurs observées de la variable à expliquer et le modèle. Si ce terme vaut 0, alors minimiser le critère pénalisé revient à minimiser le critère des moindres carrés, sans contrainte particulière sur la complexité du modèle.

Dans le cas de variables explicatives fortement corrélées ou caractérisées par un grand nombre de variables, cette solution conduit souvent à des estimations à très forte variance, induisant une mauvaise performance de prédiction, voire n'est numériquement pas implémentable. Augmenter le terme de pénalité réduit la complexité du modèle et la variance des estimations mais introduit un biais dans l'estimation. On peut choisir la valeur du paramètre de pénalité réalisant le meilleur compromis entre biais et variance d'estimation en cherchant à minimiser la variance de l'erreur de prédiction estimée par le MSEP (Mean Squared Error of Prediction) dans un dispositif de validation croisée. Cette méthode est implémentée dans le package R `glmnet`, la méthode de validation croisée par défaut étant à 10 segments.

Il est intéressant de noter que les méthodes Ridge et Lasso ont des propriétés différentes, ce qui justifie souvent que leur performance de prédiction soient comparées. La méthode Ridge conduit à une estimation dite rétrécie des coefficients du score linéaire, ce rétrécissement étant généralement réparti de manière homogène sur l'ensemble des coefficients. La méthode Lasso conduit également à une estimation rétrécie des coefficients, mais ce rétrécissement consiste à annuler certains coefficients et donc à ne conserver que certains coefficients non-nuls. Le nombre de ces coefficients non-nuls, et par conséquent le nombre de variables explicatives conservées dans le modèle, diminue lorsque le paramètre de pénalité augmente. Cette dernière propriété de la méthode Lasso concourt à sa popularité car elle permet la sélection de variables explicatives, contrairement à la méthode Ridge. Hélas, la sélection de variables explicatives par la méthode Lasso est souvent jugée instable, sensible à des modifications mineures des données.

La méthode dite Elastic Net offre une solution à ce problème d'instabilité en pénalisant le critère des moindres carrés par une combinaison linéaire de la pénalité Ridge et de la pénalité Lasso. Cette méthode introduit un second hyperparamètre permettant de contrôler le poids respectif de la pénalité Lasso et la pénalité Ridge. Le plus souvent, en pratique, ce poids est fixé par l'utilisateur. Par exemple, un choix populaire en sélection génomique est celui consistant à donner un poids de 10% à la pénalité Ridge et donc de 90% à la pénalité Lasso.

De très nombreuses variantes de régression pénalisée existent. Seule la méthode appelée Adaptive Lasso a été introduite dans l'étude comparative qui suit. Dans cette méthode, chaque coefficient est pénalisé de manière individuelle, les paramètres de pénalité étant mis à jour dans un processus itératif. L'objectif de cette variante d'estimation pénalisée est de donner plus d'importance à certaines zones du gène dans la prédiction de l'expression.

4.1.2 Régression Random Forest

Le Random Forest ou forêt aléatoire est un ensemble d'arbres de décision entrainés sur des sous-échantillons aléatoires des données d'entrainement (méthode de bagging). Chaque arbre de décision produit une estimation et c'est l'agglomération des résultats qui donne la prédiction finale. La régression par Random Forest consiste schématiquement à calculer la moyenne des prévisions obtenues par l'ensemble des estimations des arbres décisionnels de la forêt aléatoire.

Cette méthode a été effectuée grâce au package `randomForest` de R. Cependant, malgré la réduction de dimension des données faite en amont, l'optimisation des paramètres de cette fonction (nombre et profondeur des arbres, taille minimale des feuilles terminales, etc.) par validation croisée automatique s'est révélée très lente. Il a donc été fait une validation 10-fold de paramètres choisis et optimisés progressivement.

4.1.3 Régression SVM (Support Vector Machines)

Les SVM sont des classifieurs linéaires dont le but est de tracer l'hyperplan qui maximise la marge entre deux groupes de données. Lorsque les données ne sont pas linéairement séparables, des fonctions à noyaux permettent de projeter les données dans un espace de dimension supérieure où il est plus probable que les données soient linéairement séparables.

Les algorithmes de SVM sont applicables à des tâches régression [7] avec l'objectif de minimiser la norme du vecteur w orthogonal à l'hyperplan sous la contrainte que la distance entre les valeurs prédites et les valeurs d'entrainement soit inférieure à un ϵ fixé . La fonction svm du package `e1071` de R a servi à réaliser cette régression en considérant un noyau polynomial. Pour les mêmes raisons que la régression Random Forest, une validation 10-Fold a été effectuée pour des ϵ choisis et optimisés progressivement.

4.2 Apprentissage profond

L'apprentissage profond est un domaine de l'apprentissage statistique particulièrement utile pour faire des prédictions précises et révéler des structures sous-jacentes entre des variables sans a priori établis[8]. Dans ce domaine, des réseaux de neurones artificiels inspirés du fonctionnement du cerveau humain, sont constitués de couches successives (figure 4.2) de neurones artificiels interconnectés et associées à des poids différents.

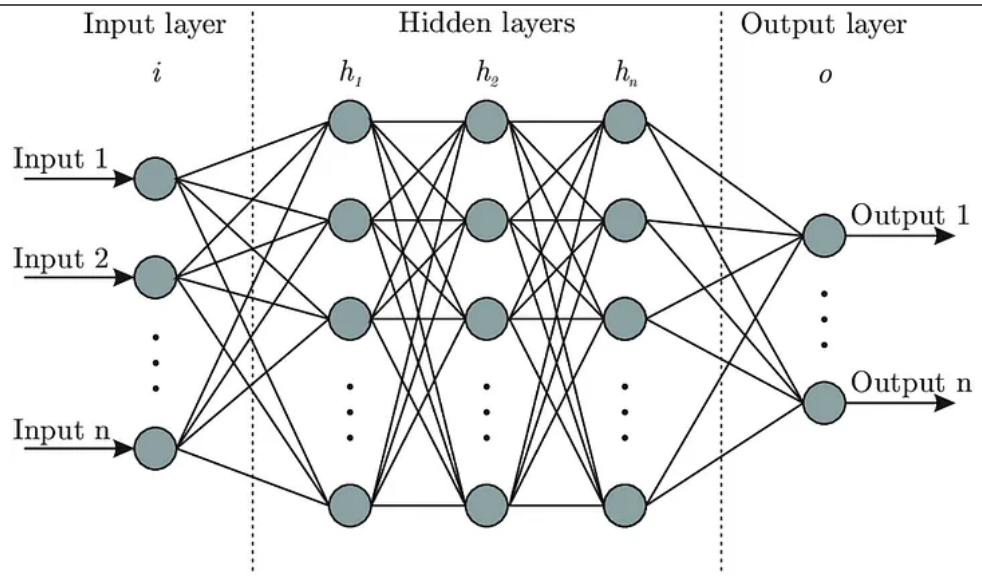


Figure 4.2 – *Illustration d'une architecture de réseaux de neurones*

Dans un réseau de neurones artificiels, les données subissent des transformations successives, linéaires ou non, à chaque couche dans le but de minimiser une fonction de coût qui mesure l'erreur entre les valeurs prédites par le réseau et les valeurs d'entraînement.

Les réseaux de neurones artificiels peuvent être construits de manière très diverse selon le type ou le nombre de couches utilisées, le nombre de neurones dans chaque couche, *etc.* Le choix d'une architecture appropriée au problème à traiter est donc une première étape cruciale. Sous les recommandations de Romain Tavenard, enseignant-chercheur à l'Université de Rennes 2 et expert en intelligence artificielle appliquées aux séries temporelles, deux types de couches de neurones semblent être indispensables au réseau de neurones qui résoudra la problématique qui nous intéresse.

→ Les couches de convolution 1D.

Ce sont des couches utilisées dans le traitement de données unidimensionnelles comme les électrocardiogrammes, les pistes audio et les séries temporelles en général[9]. Elles appliquent plusieurs filtres sur une fenêtre glissant le long des signaux avec l'objectif de détecter un ensemble de motifs caractéristiques et de produire des cartographies des motif détectés.

→ Couches d'auto-attention multi-têtes.

Leur habilité à gérer des données séquentielles en fait un outil privilégié pour traiter des données séquentielles comme des textes à des fins de traduction.
L'auto-attention est une implémentation des mécanismes d'attention qui sont des systèmes d'encodage-décodage améliorés. En effet, un système classique d'encodage-décodage prend en entrée une donnée séquentielle, la compresse dans un vecteur de contexte de taille fixe dont les éléments seront décodés un à un. Le vecteur de contexte ne garde que les informations les plus importantes de la séquence d'entrée.

Si la séquence est trop longue il est donc possible que le vecteur ne soit pas assez grand et « oublie » des informations essentielles[10]. Les mécanismes d'attention rajoutent alors une couche d'informations récupérées pendant l'encodage et associées à des poids qui varient selon leur importance.
L'opération d'attention qui attribue des poids aux informations de l'encodage exécutée plusieurs fois constitue plusieurs « têtes d'attention » permettant la parallélisation des calculs et un apprentissage plus robuste[11].

Un réseau de neurone provisoire implémentant ces deux types de couches a été implémenté à l'aide des librairies `Keras` et `KerasNLP` sous Python. L'architecture définitive du réseau souhaité dépendra de la structure des données qui lui seront fournies en entrée. Il a alors été proposé de concevoir des simulations de signaux épigénomiques et transcriptomiques simplifiés afin d'effectuer des premiers essais. A ce stade du projet, la démarche de simulation est encore en construction.

Chapitre 5

Résultats

5.1 Évaluation des approches de régression

En toute rigueur, une procédure de validation croisée devrait être mise en oeuvre ici pour tenir compte du risque de sur-ajustement dans l'évaluation de la performance de prédiction. Toutefois, dans le cas présent, la forte dépendance entre les individus, liée au réseau de régulation des gènes, ne permet pas de garantir les conditions idéales dans lesquelles les propriétés de la validation croisée sont en général étudiées. Par ailleurs, le choix des hyperparamètres des méthodes est lui le résultat d'une procédure de validation croisée, ce qui limite le risque de sur-ajustement.

Bien que l'épigénome soit un facteur important de régulation de l'expression génique, il n'est pas le seul. Ce travail n'a donc pas pour objectif de pouvoir prédire parfaitement les niveaux d'expression des gènes. Par conséquent, la corrélation au carré entre les valeurs prédites et les valeurs d'entrainement , notée R^2, est le score de prédiction utilisé pour juger la qualité des modèles testés. Comprise entre 0 et 1, cette mesure présente notamment l'avantage d'être rapidement interprétable.

Pour les deux approches de régression, *function-to-scalar* et *function-to-function*, postulées précédemment, ces scores sont retranscrits dans les tableaux 5.1 et 5.2.

Modèle	R^2
Ridge	0.248
Lasso	0.246
Adaptive Lasso	0.246
Elastic Net	0.247
Random Forest	0.312
SVM	0.268

Table 5.1 – *Scores R^2 entre les niveaux d'expression prédits par l'approche function-to-scalar et les niveaux d'expression d'entrainement*

Modèle	$R^2_{CPLS_1}$	$R^2_{CPLS_2}$	$R^2_{CPLS_3}$	$R^2_{CPLS_4}$	$R^2_{CPLS_5}$	$R^2_{CPLS_6}$
Ridge	0.42	0.10	0.06	0.08	0.06	0.06
Lasso	0.41	0.09	0.05	0.08	0.05	0.05
Adaptive Lasso	0.41	0.10	0.05	0.08	0.06	0.06
Elastic Net	0.42	0.10	0.05	0.08	0.06	0.06
Random Forest	0.47	0.15	0.11	0.09	0.06	0.04
SVM	0.45	0.10	0.09	0.06	0.05	0.03

Table 5.2 – *Scores R^2 entre les composantes PLS prédites par l'approche function-to-function et les composantes PLS d'entrainement*

On peut voir dans le tableau 5.1 qu'aucun des modèles testés avec l'approche *function-to-scalar* ne dépasse un score R^2 de 0.5 indiquant ainsi une assez mauvaise performance prédictive du niveau d'expression. De manière similaire les scores de l'approche *function-to-function* ne sont pas assez élevés pour conclure à un modèle performant de prédiction des composantes PLS transcriptomiques.

Toutefois, la première colonne du tableau 5.2 montre que la prédiction de la première composante PLS transcriptomique est systématiquement meilleure que celle des niveaux d'expression, démontrant ainsi l'intérêt de l'approche function-to-function à travers le gain d'information apporté.

Partant de ce constat, la dernière étape de la stratégie de modélisation énoncée précédemment, consistant à reconstruire des niveaux d'expression à partir des composantes PLS transcriptomiques mieux prédites, a été mise en place. Les plus fiables prédictions de composantes PLS transcriptomiques obtenues par la régression Random Forest deviennent alors des variables prédictives.

5.2 Reconstitution du niveau d'expression des gènes

Étant donné que les composantes PLS d'entrainement que l'on notera $CPLS_{train}$, ont été construites pour maximiser la corrélation linéaire avec le niveau d'expression des gènes, la modélisation de la relation entre ces deux quantités permet de connaître le niveau maximal de qualité auquel on peut prétendre avec un modèle prenant en entrée les composantes PLS prédites, notées $CPLS_{pred}$.

Trois modèles ont été ajustés sur $CPLS_{train}$ et le niveau d'expression des gènes :

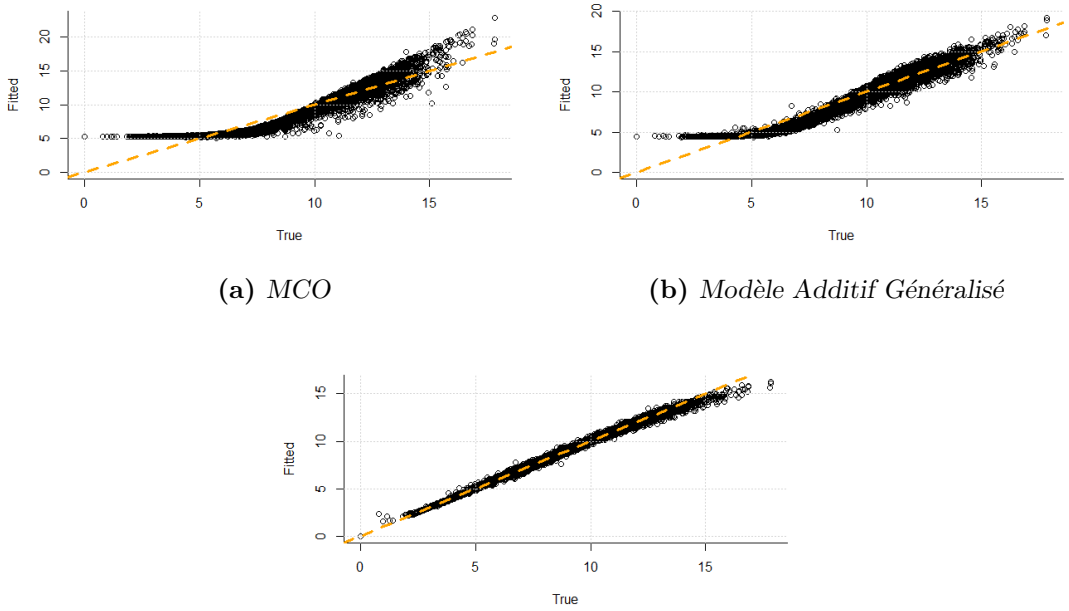
- Modèle linéaire avec estimation par moindres carrés ordinaires (MCO) réalisée en même temps que la construction des composantes PLS dans la fonction `plsr` du package `pls`.
- Modèle additif généralisé (ou GAM) avec des fonctions de lissage spline implémenté dans le package `gam` de R
- Modèle de régression par Random Forest

Le tableau 5.3 montre que les scores d'ajustement R^2 des trois modèles sont tous proches de 1, indiquant des ajustements de bonne qualité, la régression par Random Forest ayant l'ajustement le plus précis.

Modèle	MCO	MAG	Random Forest
R^2_{train}	0.79	0.86	0.99

Table 5.3 – *Scores R^2 d'ajustement du niveau d'expression à partir sur $CPLS_{train}$*

La figure 5.1 représente les valeurs d'expression ajustées par chaque modèle en fonction des véritables niveaux d'expression. On peut notamment y voir que le point faible des modèles linéaires réside dans la prédiction des niveaux d'expression les plus faibles.



(a) *MCO*

(b) *Modèle Additif Généralisé*

(c) *Random Forest*

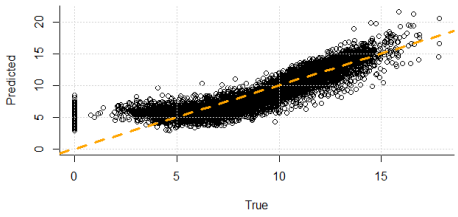
Figure 5.1 – *Valeurs d'expression ajustées par trois types de modèles sur $CPLS_{train}$ en fonction de leurs valeurs d'entrainement*

Les performances de prédiction de ces mêmes modèles ajustés cette fois-ci sur $CPLS_{pred}$ sont retranscrites dans le tableau 5.4. Malgré la perte de qualité due aux composantes PLS moins bien prédites, on retrouve scores élevés de prédiction, la régression par Random Forest se démarque également par sa performance supérieure.

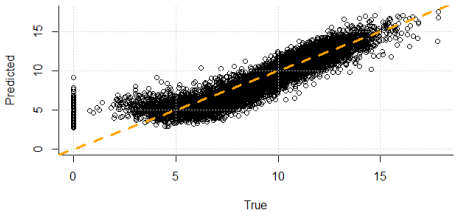
Modèle	MCO	MAG	Random Forest
R^2_{pred}	0.75	0.81	0.89

Table 5.4 – *Scores R^2 de prédiction du niveau d'expression à partir de $CPLS_{pred}$*

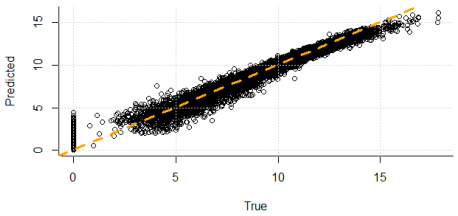
La figure 5.2, représentant pour chaque modèle les niveaux d'expression prédits par $CPLS_{pred}$ en fonction des niveaux d'expression réels, illustre la qualité de prédiction ainsi que la difficulté à prédire les plus bas niveaux d'expression.



(a) *MCO*



(b) *Modèle Additif Généralisé*



(c) *Random Forest*

Figure 5.2 – *Valeurs d'expression prédites par trois types de modèles sur $CPLS_{pred}$ en fonction de leurs valeurs d'entrainement*

Chapitre 6

Conclusion et appréciation du stage

La recherche de modèles d'association entre l'épigénome et le transcriptome de *Fusarium graminearum* est soumise à des obstacles de plusieurs types. D'une part, il est important d'appliquer un prétraitement aux mesures brutes de signaux épigénomiques et transcriptomiques afin d'avoir des données exploitables statistiquement. D'autre part, ces mesures sont stockées dans des jeux de données dont la dimension doit être réduite pour en faciliter l'analyse. Une partie importante de ce stage a donc consisté à développer la démarche de prétraitement décrite dans ce rapport de manière à ce qu'elle puisse être implémentable sur des données génomiques recueillies en conditions de cultures modifiées ou des données appartenant à une autre espèce comme le puceron du pois *Acyrthosiphon pisum* mentionné en introduction.

Par la suite, les analyses exploratoires effectuées sur les signaux ont confirmé et dépeint le lien, bien connu en biologie moléculaire, entre l'épigénome et le transcriptome. En particulier le rôle que la déplétion en nucléosomes et le relâchement de la chromatine joue sur le niveau d'expression des gènes a été illustré.

Enfin, les travaux de modélisation réalisés ont permis en premier lieu d'écarter l'approche directe de prédiction du niveau d'expression des gènes par le signal épigénomique en faveur d'une approche indirecte passant par la prédiction du signal transcriptomique. Dans un deuxième temps, une approche par apprentissage profond a été entamée et l'architecture du réseau de neurones sera développée en passant par la construction d'un jeu de données "jouet" dans la suite du projet.

Pour conclure, ce stage de fin d'études a été une expérience très enrichissante. J'y ai découvert le monde de la recherche en assistant notamment à différents congrès et séminaires et cela m'a permis d'ouvrir mes horizons. A travers mon encadrement, j'ai également beaucoup appris que ce soit au niveau de la méthodologie, des mathématiques ou des compétences techniques. En outre, le sujet de stage lui-même était épanouissant car il s'inscrit dans ma motivation première à poursuivre des études en statistiques appliquées : aider à la résolution de problèmes du monde réel grâce aux outils mathématiques et statistiques.

Bibliographie

[1] "Site web du ministère de la culture." Journal officiel du 07/09/2018.

[2] V. Laudet and E. Bonnlye, "Les gènes chevauchants," *médecine/sciences*, 1994.

[3] S. Bloyer and R. Koszul, "Nucleosome patterns in four plant pathogenic fungi with contrasted genome structures," *Peer Community In Genomics*, 07 2022.

[4] A. K. Singh and F. Mueller-Planitz, "Nucleosome positioning and spacing : From mechanism to function," *Journal of Molecular Biology*, vol. 433, no. 6, p. 166847, 2021. Diving into Chromatin across Space and Time.

[5] X. Fan, Z. Moqtaderi, Y. Jin, Y. Zhang, S. Liu, and K. Struhl, "Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 17945–50, 10 2010.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *Linear Methods for Regression*, pp. 43–99. New York, NY : Springer New York, 2009.

[7] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, (Cambridge, MA, USA), p. 155–161, MIT Press, 1996.

[8] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.

[9] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications : A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.