Apprentissage statistique pour la prédiction des interactions chromosomiques

Océane Cassan
Encadrée par Laurent Bréhélin, Charles-Henri Lecellier, Sophie Lèbre

Université Claude Bernard Lyon 1
INSA de Lyon, France

Résumé L'identité d'une cellule est déterminée par le profil d'expression de ses gènes. Les repliements dynamiques de la molécule d'ADN à l'intérieur du noyau permettent de mettre en contact différentes régions régulatrices, afin de moduler la fixation des protéines initiant la transcription des gènes. Les mécanismes régissant ces contacts ne sont encore que partiellement compris, ainsi que les nombreux paramètres biologiques les gouvernant. Notamment, l'influence de la séquence ADN sur ces repliements est assez peu étudiée. L'objectif de ce stage est donc de déterminer si les séquences permettent à elles seules de prédire les contacts entre les régions régulatrices, dans un contexte d'apprentissage automatique.
La construction de jeux de données adaptés basés sur des expériences biologiques est proposée, ainsi qu'une méthode d'extraction de *features* par segmentation des séquences. Différents types de modèles de classification supervisée sont entraînés sur ces jeux de données, et comparés. Avec la méthodologie appliquée, les résultats semblent montrer que les variables liées à la séquence ne portent pas toute l'information recherchée. De plus, les séquences environnant les régions régulatrices semble plus informatives que les régions régulatrices elles mêmes pour prédire leur mise en contact.

Mots-clés: Interactions chromosomiques, apprentissage automatique, classification supervisée, génomique, extraction de features

Abstract. The identity of a cell is determined by the expression patterns of its genes. Inside the nucleus, the DNA molecule is dynamically folded so regulatory regions can be in contact. Those physical interactions enable transcription factors to bind to DNA, thus controling the initiation of genes transcription. The mecanisms of such contacts are still far from being fully understood, as well as the biological parameters driving them. In literature, very few projects focus on DNA sequence as a direct predictor variable. The objective of this internship is to determine if sequence alone is enough to predict the distal interactions between regulatory regions, in a statistical learning framework.
We propose here methods to build apropriate datasets based on biological experiments, and to extract features using a sequence segmentation approach. Different types of models for supervised classification are trained and compared on those datasets.
With our current methods, the results show that the sequence variables of regulating regions do not seem to carry enough information to accurately predict DNA folding. However, our feature extraction method revealed that the neighborhood of regulatory regions is more informative to predict contacts than the regions themselves, and lead to acceptable classification.

Keywords: Chromosomic interactions, machine learning, supervised classification, genomics, feature extraction

1 Contexte

1.1 Organisme d'accueil

Le LIRMM est le laboratoire de recherche en Informatique, Robotique et Microélectronique de Montpellier. Il s'agit d'une unité mixte dépendant du Centre National de la Recherche Scientifique ainsi que de l'Université de Montpellier. Les recherches menées au LIRMM portent sur les domaines des sciences et technologies de l'information, de la communication et des systèmes. Plus précisément, ce stage ce déroule au sein l'équipe MAB : Méthodes et Algorithmes pour la Bioinformatique. Elle a pour objectif l'analyse de données génomiques et post-génomiques par les outils numériques et mathématiques adaptés. Le sujet du projet s'inscrit dans une thématique proche des recherches menées dans l'équipe par plusieurs de ses membres. Ceux-ci ont déjà travaillé à prédire des aspects de la régulation à partir de la séquence ADN. En utilisant l'expertise de l'équipe en biologie, en analyse de séquences et en apprentissage automatique, ce stage vise à explorer une nouvelle piste du domaine. Son originalité réside dans la nature des prédictions à réaliser, qui portent sur les repliements de la molécule d'ADN dans le cadre de la régulation.

1.2 Cadre biologique

Toutes les cellules de notre organisme possèdent, à quelques nucléotides près, le même code génétique. L'existence de différents tissus ou types cellulaires est permise par une grande diversité des profils d'expression des gènes dans chacune de nos cellules. Les gènes dits actifs, ou exprimés, définissent fondamentalement une cellule au travers des protéines qui la composent, de ses fonctions métaboliques, ou de ses capacités d'adaptation à l'environnement.

Le processus biologique qui module l'expression des gènes est la transcription. Il s'agit d'un mécanisme au cours duquel un complexe protéique, l'ARN polymérase, vient se fixer à la molécule d'ADN, l'ouvre, et synthétise un ARN messager contenant la séquence d'un gène précis. Cet ARN messager sera ensuite exporté à l'extérieur du noyau et traduit en protéine pour exercer sa fonction dans la cellule.

Le recrutement de la polymérase se fait via les facteurs de transcription, des protéines venant se lier à des régions situées en amont d'un gène, des régions dites promotrices. Les régions promotrices, ou promoteurs, ne codent pas d'information comme le font les gènes, mais permettent de réguler leur transcription via leur séquence et leur accessibilité en accueillant les protéines initiant la transcription. Des régions régulatrices plus distantes des gènes, appelées les enhancers, peuvent également être impliquées dans leur régulation. Dans ce cas, leur contact avec un promoteur via des repliements de la molécule d'ADN s'avère indispensable à la transcription, et permet de créer des patterns d'expression génétique plus complexes, flexibles et robustes. Chez les vertébrés, ces régions distantes sont connues pour interagir de manière dynamique avec les promoteurs pour constituer des profils d'expression tissu-spécifiques [12].

Ces interactions longue-distance sont rendues possibles par les repliements tri-dimensionnels de la chromatine. La chromatine est composée de l'ADN enroulé autour de protéines, appelées histones, qui lui confèrent sa structure et définissent son degré de compaction. Des forces physiques antagonistes, entre répulsion et attraction moléculaires [12], régissent la spatialisation de la chromatine. Ainsi, les enhancers peuvent s'affranchir de distances chromosomiques de l'ordre du Mégabase pour communiquer leur information régulatrice à un promoteur, comme illustré en Figure 1. Ces contacts se produisent le plus souvent au sein d'un même chromosome, mais des rapprochements inter-chromosomiques sont également observés. D'une manière plus large, le repliement de la chromatine met en contact de nombreuses régions régulatrices actives simultanément dans des compartiments fonctionnels appelés les TADS : Topologically Associated Domains.

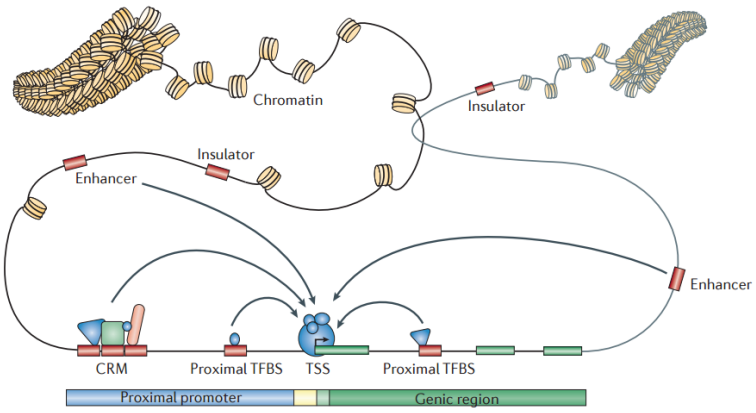


Fig. 1: Vue schématique d'un gène et de ses régions régulatrices, issu d'un fragment de chromatine déroulée. Le promoteur est en bleu et ses enhancers en rouge. Le contact se fait précisément au niveau du TSS (Transcription Starting Site), et amorce la transcription de la région génique en aval.[10]

Parmi ce concert de mouvements gouvernés par les voies métaboliques et adaptatives de la cellule et admettant une certaine stochasticité, identifier quels enhancers régulent quels gènes dans un type cellulaire particulier demeure une question difficile. Beaucoup de travaux publiés tendent à montrer que des variables épigénétiques, c'est à dire liées à l'environnement de l'ADN, à sa conformation, ou à des modifications chimiques apportées aux histones peuvent être informatives, mais peu s'intéressent à la séquence nucléotidique directement. L'objectif de ce stage est de savoir s'il est possible de prédire l'interaction entre deux régions chromosomiques impliquées dans la régulation en se basant sur leur séquence uniquement.

1.3 Sources de données pour les interactions chromosomiques

Des consortiums internationaux fournissent aujourd'hui une cartographie assez complète de l'ensemble des enhancers et promoteurs du génome humain. Par exemple le consortium FANTOM (Fonctional ANonTation Of the Mammalian genome) créé en 2000 a pu les répertorier en utilisant une méthode basée sur la mesure des ARN d'expression (e-ARN) de ces régions. En effet, lors de la transcription des gènes, les enhancers et les promoteurs génèrent des transcrits. Ces courts transcripts "capés" sont capturés par la technique CAGE (Cap Analysis Gene Expression). Ces e-ARN ont des caractéristiques différentes chez les enhancers et les promoteurs, ce qui permet de les différencier. Après alignement, leur localisation sur les chromosomes est répertoriée [16] [17]. La base de données FANTOM contient environ 40 000 enhancers, et 95% des promoteurs des gènes annotés comme codant pour des protéines à ce jour. Ce travail a été fait sur plus de 500 types cellulaires différents chez l'homme, et l'atlas constitué a été utilisé tout au long de ce stage pour définir et nommer les régions chromatiniennes étudiées. La nomenclature utilisée pour un promoteur sera par exemple `chr10:7890..7925,+`, signifiant qu'il est situé sur le chromosome 10, commence à la base 7890 et se termine à la base 7925. Après la virgule, le signe + ou - informe sur le brin d'ADN, direct ou complémentaire, sur lequel se situe le promoteur. Pour un enhancer, les identifiants seront de la forme `chr12:43690..43877`. La signification reste la même, à la différence d'absence d'information sur le sens du brin, car les e-ARN CAGE caractéristiques des enhancers sont bidirectionnels.

Une fois la liste des régions régulatrices établie, la question est d'obtenir l'information de leur rapprochement dans l'espace. Pour cela, on distingue différentes approches.

Les techniques expérimentales de détection de contacts Des techniques d'identification des contacts chromatiniens appelés Chromosome Conformation Capture Analysis (3C) permettent de constituer de telles bases de données. Parmi elles, notre attention s'est portée sur deux types de capture :

- Les expériences de Hi-C. Les portions génomiques mises en contact sont liées par des *linkers*, des petites séquences d'ADN se chaînant à tout fragment d'ADN proche de lui. Les linkers se lient aux régions régulatrices en contact, puis s'en suit une phase de dégradation de l'ADN par une enzyme de restriction, qui détruit la molécule d'ADN non liée. On obtient, à l'issu de la coupe des *linkers*, des paires de portions d'ADN issues de deux fragments spatialement proches dans le noyau.
- Les expériences de ChIA-PET. Ces expériences sont très proches des expériences de Hi-C. La technique de Hi-C capture les interactions se produisant sur tout le génome sans restrictions, tandis que les données de ChIA-PET filtrent les contacts faisant intervenir une protéine d'intérêt. Ainsi, nous obtenons les paires de régions en contact médié par une certaine protéine, par exemple un facteur de transcription précis, ou l'ARN polymérase.

Ces techniques garantissent des jeux de données riches et suffisamment résolutifs pour traiter notre problématique. Elles restent néanmoins lourdes et difficiles à mettre en place sur de nombreux types cellulaires. C'est ce qui a motivé d'autres méthodes de constitution de paires de régions chromosomiques, généralisables à un grand nombre de types cellulaires.

L'association de régions suivant une corrélation d'expression Dans la publication de leur Atlas des enhancers [16], le consortium FANTOM réalise des associations entre enhancers et promoteurs, en utilisant les mesures d'expression obtenues par la méthode CAGE dans leurs nombreux types cellulaires. Une valeur d'expression est un nombre de *reads* d'e-ARNs séquencés pour cette région dans le type cellulaire considéré.

Cette démarche de constitution de paires part du postulat qu'un enhancer qui régule un promoteur doit avoir un profil d'expression similaire à celui du promoteur dans l'ensemble des tissus analysés. Pour chaque couple enhancer-promoteur de leurs atlas, les corrélations sont calculées deux à deux entre les profils d'expression. Chaque mesure de corrélation fournit une p-valeur renseignant sur la significativité de la corrélation, et donc de l'appariement de l'enhancer et du promoteur considérés. Cette procédure impliquant des tests multiples, une correction a été appliquée dans le calcul des p-values, la méthode False Discovery Rate. Cette méthode permet de contrôler le nombre de faux positifs dans l'ensemble des tests. Au final, 340150 paires enhancer-promoteur significatives sont identifiées.

Bien qu'elle ne nécessite pas d'expérimentations coûteuses comme celles des 3C (outre la technique CAGE), cette méthode perd l'information de l'appariement enhancer-promoteur dans un type cellulaire précis. En effet, les corrélations sont calculées sur tous les types cellulaires, donnant en résultat un score d'appariement global.

1.4 Formalisation du problème d'apprentissage statistique

Les jeux de données précédemment décrits constituent la matière première de l'apprentissage supervisé que nous souhaitons mettre en place. Les modèles d'apprentissage automatiques envisagés doivent permettre, pour une paire de régions chromosomiques en entrée, de prédire si ces régions vont se retrouver en contact dans l'espace du noyau cellulaire ou non. On définit donc par "exemple", ou "élément", une paire de régions régulatrices, du type enhancer-promoteur, ou promoteur-promoteur. La classe associée à cet élément est 0 ou 1, suivant que les régions sont en contact ou non. Cette information est donnée par les jeux de données FANTOM ou les expériences de ChIA-PET.

Ces jeux de données permettent d'entraîner un modèle prédictif avec une *ground truth* pour un ensemble fourni de régions chromosomiques, mais également de tester ce modèle sur un ensemble réservé à l'évaluation. En effet, une propriété fondamentale de l'apprentissage statistique est la capacité de distinguer des patterns sur un ensemble d'apprentissage, pour ensuite les généraliser à des données inconnues.

Étant donnée la nature tissu-spécifique de ces interactions, notamment entre enhancer et promoteurs, l'objectif est de construire les modèles de manière spécifique à un type cellulaire. Différentes variables en *input* de ces modèles caractérisant la séquence des régions à prédire seront explorées. De plus, plusieurs méthodes de *machine learning* seront étudiées afin de répondre au mieux à la problématique traitée.

1.5 État de l'art

Les interactions entre régions régulatrices ont déjà été étudiées dans la littérature, et il existe plusieurs travaux visant à les prédire. En sont présentés ici quelques uns, ayant nourri la réflexion préalable à ce projet.

Certaines méthodes publiées ne font pas appel à du *machine learning*. Par exemple, ABC model [6] définit un score de contact comme le produit de l'expression de l'enhancer par la fréquence des contacts entre l'enhancer et le promoteur dans des types cellulaires connus. Cette fréquence des contacts est donnée par des expériences 3C, et l'expression par les *peaks* CAGE pour de nombreux types cellulaires. Plus ce score est élevé, plus il est probable que l'enhancer et le promoteur soient en contact. Cependant, cette méthode reste simple et ne permet pas de retranscrire toutes les subtilités des données.

Un article souvent cité pour l'utilisation de modèles de *machine learning* dans la prédiction des interactions enhancer-gènes est l'article présentant le modèle TargetFinder [19]. Les variables de leurs modèles sont liées à des mesures d'ouverture de la chromatine, à la méthylation [1] de l'ADN, à l'expression, à la fixation de facteurs de transcription et de protéines structurales, ou à la modification des histones. Ces variables sont observées en dynamique dans les types cellulaires, et traduisent leur paysage épigénétique, mais sans s'intéresser directement à la séquence. Les modèles de *machine learning* proposés prennent en entrée l'ensemble de ces *features* pour chaque élément d'une paire (enhancer, promoteur, et fenêtre les séparant), pour faire une prédiction. Parmi les modèles testés, la méthode ensembliste à base de *Boosted Trees* a été plus performante qu'un *linear SVM* ou qu'un arbre de décision seul. La méthode de quantification de l'importance des variables est celle implémentée dans ce le package `scikit-learn`, par Hastie et Al [9]. Les auteurs parviennent à discriminer les paires en interaction, labelisées grâce à des données de Hi-C, et atteignent une performance de 83% sur la moyenne des types cellulaires. Ce chiffre est cependant à nuancer car un travail [20] a été publié en 2018 afin de corriger une erreur de cette publication. En effet, les modèles présentés sont basés sur des variables structurelles et épigénétiques décrivant les enhancers, les promoteurs, et la portion d'ADN les séparant. Lorsque N enhancers consécutifs interagissent avec le même promoteur, sont formées N paires distinctes. Or, ces paires partagent des caractéristiques au niveau du promoteur et de la région les séparant, appelée la

[1] La méthylation est une altération chimique des bases de l'ADN, via l'ajout d'un groupement méthyle. Elle est connue pour jouer un rôle clé dans la régulation des gènes.

fenêtre. Si deux paires se chevauchent, même si elles font intervenir des promoteurs ou enhancers différents, elles peuvent partager certaines *features* via leur fenêtre. Si le jeu de test comporte des paires partageant des *features* avec des paires rencontrées lors de l'apprentissage, les capacités de généralisation se voient faussées par une sur-spécificité au jeu d'entraînement. Les auteurs de cet article correctif suggèrent de classer les paires par position chromosomique afin que les paires contenant des éléments en commun se retrouvent uniquement dans le jeu d'apprentissage ou de test, et non dans les deux simultanément comme le ferait une cross-validation aléatoire. Les nouvelles valeurs de score f1 utilisées pour quantifier la performance de la classification de TargetFinder après cette correction sont montrées en annexe 1, Supp. figure .1, et attestent que les prédictions ne sont en réalité pas meilleures que le hasard. Ce phénomène de biais de *features* partagées sera donc pris en compte dans l'élaboration des jeux d'apprentissage et de test du projet.

D'autres méthodes, comme celle intitulée RIPPLE [13], visent à prédire les interactions dans tous les types cellulaires. Celle-ci entraîne d'abord des classifieurs spécifiques aux types cellulaires un utilisant les marques des histones, les sites de fixation de facteurs de transcription, et les niveaux d'e-ARN. Les variables pour l'enhancer et le promoteur sont ensuite combinées en en faisant le produit, l'addition, ou la corrélation. Ces modèles permettent la sélection de variables, et les variables sélectionnées sont utilisées pour nourrir un *Random Forest* plus général, prédisant les interactions enhancer-promoteur de manière globale, quel que soit le type cellulaire.

À l'inverse, les auteurs de la procédure JEME [4] commencent par combiner l'information commune à tous les types cellulaires. Le principe est de sélectionner des variables informatives sur l'ensemble des types cellulaires grâce à des régressions logistiques pénalisées, puis de les utiliser dans des *Random Forests* propres au type cellulaire. Encore une fois, les variables utilisées sont des mesures d'accessibilité de l'ADN, les *peaks* d'e-RNA, et la distance entre l'enhancer et le promoteur, mais pas la séquence elle même.

Le *deep learning* a récemment fait son apparition dans la litérature concernant les interactions enhancer-promoteur. Les auteurs de la méthode SPEID [15] ont été les premiers a s'intéresser à la séquence uniquement pour prédire ces contacts. Leur modèle est un réseau de neurones composé de plusieurs couches de convolutions, suivies d'un réseau récurrent Long Short Term Memory (LSTM). Cette approche permet l'extraction de *features* par convolution des séquences fournies en entrée, (celles de l'enhancer et du promoteur), puis combine ces *features* dans une architecture récurrente, adaptée à la nature séquentielle des données. Cette méthode semble permettre la prédiction des interactions, mais ne laisse pas la liberté de choisir et analyser les variables informatives. En effet, les convolutions et *max-pooling* qui en découlent sont peu interprétables.

Les auteurs de cette dernière méthode ont également, un an auparavant, développé une approche qui s'avère la plus proche de nos objectifs : PEP [21]. Son module PEP-Motif a pour principe d'extraire des scores d'occurrence de motifs (courtes séquences d'ADN) connus pour recruter des facteurs de tran-

scription, et de les utiliser comme variables prédictives d'un modèle de Gradient Tree Boosting. Les résultats présentés sont très encourageants, mais s'appuient cependant sur le même jeu de données que TargetFinder, lequel présentait un biais de surévaluation de ses performances. La validité de ces très bons résultats est donc remise en question par les travaux présentés en référence [20].

Au terme de cet état de l'art, il est envisagé de s'inspirer des modèles rencontrés, tels que la régression logistique, les méthodes ensemblistes, ou le *deep learning*, mais en utilisant des variables d'une nature différente. Contrairement à la grande majorité des articles rencontrés, nous souhaitons utiliser des variables prédictives uniquement liées aux séquences, et non aux signaux épigénétiques et structuraux liés à molécule d'ADN. Nous cherchons à construire des modèles plus intelligibles que les modèles de *deep learning* qui restent difficilement interprétables, et à vérifier les résultats des quelques méthodes ([21] et [15]) qui suggèrent que la séquence à elle seule peut être informative.

2 Méthodes

2.1 Problèmes de classification étudiés

Au cours de ce stage, nous avons souhaité prédire deux types de contacts : les contacts enhancer-promoteur et les contacts promoteur-promoteur. Ces prédictions sont faites en se basant à la fois sur la composition nucléotidique des régions chromosomiques et sur leurs scores PWM, variables décrites en section 2.2. Dans un premier temps la question de la construction d'un jeu d'apprentissage équilibré se pose.

Sélection des exemples positifs et négatifs

Dans les jeux de données mentionnés en section 1.3, les exemples de paires enhancer-promoteur ou promoteur-promoteur en interaction sont clairement identifiables. Cependant, dans un contexte d'apprentissage automatique, il nous faut également choisir un ensemble de paires (dites négatives) qui ne sont pas en contact. Il s'agit du *background*, les paires contre lesquelles nous souhaitons faire ressortir un signal. Le choix du *background* est crucial car il ne fera pas apparaître les mêmes informations prédictives suivant sa composition. Suivant la nature des données, les exemple négatifs ont été construits différemment.

Paires pré-constituées par le consortium FANTOM : Dans leur article [16], les auteurs de l'atlas des promoteurs et des enhancers du génome humain constituent des associassions enhancers-promoteurs basées sur une mesure d'expression de ces régions parmi toutes leurs librairies [2]. Il en résulte un ensemble de paires valables sur un ensemble d'environ 600 types cellulaires. Or, étant donnée la tissu spécificité de l'action des enhancers, nous souhaitons créer un jeu de données par

[2] Une librairie désigne un type cellulaire.

type cellulaire. Les données d'expression des enhancers et des promoteurs FANTOM sont accessibles sous format matriciel. Ces matrices contiennent la valeur d'expression de chaque promoteurs et enhancers dans chacune des librairies.

Dans une première approche, nous sommes partis du principe que pour un type cellulaire précis, les paires enhancers-promoteurs positives seront les paires FANTOM pré-constituées par corrélation qui ont à la fois leur enhancer et leur promoteur actifs dans ce tissu. Pour définir cette notion d'activité, nous sélectionnons les enhancers et promoteurs dont le niveau d'expression est strictement supérieur à 0. Pour les paires négatives, nous sélectionnons les paires dont le promoteur est actif et l'enhancer inactif pour traduire le cas où un promoteur est exprimé, mais régulé par un autre enhancer. Une visualisation de ces ensembles est montrée en figure 2.

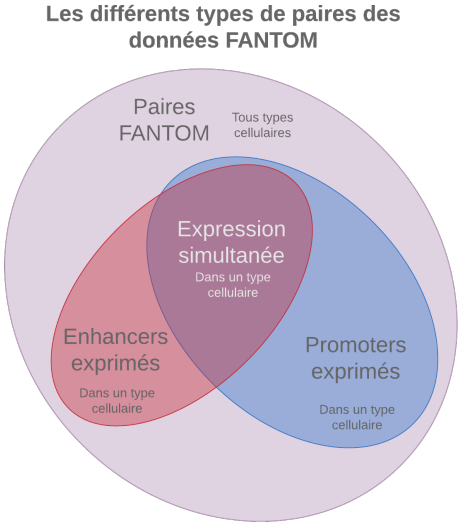


Fig. 2: Diagramme des paires constituées par le consortium FANTOM intersectées avec les paires dont l'enhancer et/ou le promoteur sont actifs dans un type cellulaire précis.

Pour éviter tout biais, il est important de contrôler l'expression des promoteurs dans la classe positive et la classe négative. En effet, les promoteurs peu exprimés risquent d'être sur-représentés dans la classe négative, ce qui introduirait un facteur de confusion dans l'étude. Au moment de constituer les paires négatives, des paires avec des mesures d'expression similaires à celles de la classe positives sont échantillonnées. Les mesures d'expression pouvant contenir des valeurs extrêmement fortes en petit nombre, c'est le logarithme de l'expression qui a été utilisé pour cet échantillonage. Cela permet de resserrer la distribution

et de réduire la proportion d'*outliers*, facilitant la création d'une distribution proche de celle de la classe positive.

Le problème ainsi défini revient à prédire les paires avec enhancers et promoteurs actifs, contre celles dont seul le promoteur est actif. De fait, le problème se réduit dans ce cas à un problème de classification de l'activité des enhancers. Nous avons donc imaginé une seconde problématique. Les paires positives considérées sont ici les paires FANTOM dont l'enhancer et le promoteur sont simultanément soit actifs, soit inactifs. De même, les paires négatives sont les paires dans lesquelles uniquement l'un des membres de la paire est actif. Une comparaison et une interprétation biologique de ces deux problématiques sera faite dans la section résultats. Comme précédemment, un contrôle sur l'expression est effectué afin de considérer des promoteurs ayant des activités similaires dans les deux classes.

Paires obtenues par ChIA-PET : Ces données permettent de disposer de paires enhancer-promoteur ou promoteur-promoteur positives, en fournissant les enhancers et promoteurs en contact. Commr pour les données FANTOM cependant, ces données ne portent pas directement d'informations relatives aux paires négatives, car elles contiennent uniquement les régions impliquées dans un contact. Une méthode pour obtenir ces régions est de prendre les paires positives et de les randomiser. Un désappariement est alors effectué, pour créer des paires fictives labéllisées comme n'interagissant pas. Pour les paires enhancer-promoteur, pour se rapprocher de la nature des paires FANTOM dont la distance enhancer-promoteur n'excède pas 500 kbases, ce même contrôle a été fait. Les paires shufflisées n'ont donc été gardées que si elles satisfaisaient cette contraintes, permettant des paires plus réalistes.

Constitution du jeu d'apprentissage et de test

Le jeu de données utilisé en *machine learning* doit être séparé en un ensemble d'entraînement, et un ensemble de validation. Pour retranscrire au mieux les capacités de généralisation du modèle, ce jeu de validation doit être complètement indépendant du jeu d'apprentissage. En effet, si des exemples entièrement ou en partie identiques, sont retrouvés à la fois dans le jeu d'apprentissage et dans le jeu de test, le modèle ne ferait que re-mobiliser avec succès des prédictions vues lors du training, montrant des performances artificiellement hautes.

La nature des données utilisées dans ce projet rend cette problématique assez importante et plus délicate à traiter. Comme montré dans la référence [20], des éléments partageant certaines *features*, même s'ils ne sont pas exactement identiques, peuvent conduire à une sur-estimation de l'AUC. En effet, comme les éléments à classifier sont des paires, il est possible de retrouver un même enhancer ou promoteur dans plusieurs paires. Une précaution à prendre est donc de regrouper les paires contenant un élément en commun soit dans les jeu de test, soit dans le jeu d'apprentissage uniquement.

De la même manière, si des régions se chevauchent, il est possible que celles ci partagent des caractéristiques et biaisent l'évaluation du modèle. Comme pour

les paires identiques, les paires chevauchantes seront donc contrôlées de manière à n'apparaître que lors de l'apprentissage ou de l'évaluation.

2.2 Représentation des exemples

Différentes variables peuvent être considérées lorsque nous cherchons à caractériser une séquence d'acides nucléiques. L'ADN se compose d'un alphabet de 4 nucléotides, A, C, G et T, respectivement Adénine, Cytosine, Guanine et Thymine. Les proportions de ces bases et leur arrangement en motifs précis sont connus pour être liés à la nature des régions génomiques ainsi qu'à leur fonction régulatrice [18] [2]. La taille des séquences considérées variera suivant les tests réalisés. Elle est, dans les modèles initiaux, de 1000 paires de bases (bp), centrée sur le milieu de la région régulatrice.

La composition en nucléotides : La proportion des nucléotides, également appelées bases, constitue une première information à capter. Au delà d'une seule base, l'information de la fréquence de plusieurs bases consécutives est à considérer, comme la fréquence de dinucléotides, trinucléotides ou de quadrinucléotides. Plus généralement, la notion de k-mer se réfère à un mot de k nucléotides. Au total, à l'instar des variables utilisées dans la référence [18], 12 variables de composition nucléotidique seront utilisées dans les modèles : les taux de AT et CG, représentant la fréquence des nucléotides A et T sur les deux brins d'ADN, et de C et G, ainsi que les taux de 10 dinucléotides. Les détails sur le choix de ces 10 2-mers sur les 16 possibles avec un alphabet de 4 lettres sont présentés en annexe 2.

Les motifs dénommés PWM : Les facteurs de transcription reconnaissent des k-mers particuliers, appelés motifs, nécessaires à leur fixation. Un facteur de transcription peut reconnaître des motifs qui diffèrent de quelques bases. Il existe environ 1600 facteurs de transcription connus chez l'homme. Les expériences de Chip-Seq permettent d'établir quels sont les motifs sur lesquels se fixent préférentiellement chaque facteur. Lors d'une telle d'expérience, les séquences fixées par le facteur étudié sont protégées d'une immunoprécipitation de la molécule l'ADN, puis séquencées. Cela permet d'identifier tous les k-mers fixés par un facteur de transcription, qui sont ensuite résumés dans modèle probabiliste particulier, le PWM (Position Weight Matrix). Il existe des bases de données de ces motifs comme la base Jaspar [14] qui en répertorie 662. sur les 1600 environ connus chez l'homme. Il est possible de les visualiser sous forme de logos comme montré en figure 3.

Des logiciels développés et optimisés pour la recherche d'occurrences sont ensuite utilisés pour repérer ces motifs dans un set de séquences d'intérêt. Un score PWM est obtenu en calculant l'adéquation entre les k-mers de la séquence et les PWM d'intérêt. À l'issue de ce processus, il est possible d'attribuer aux séquences désirées un score pour chaque facteur de transcription, représentant les chances de présence d'un motif pour lequel ce facteur a une forte affinité. Ces scores constitueront un second type de variables explorées au cours de ce stage.

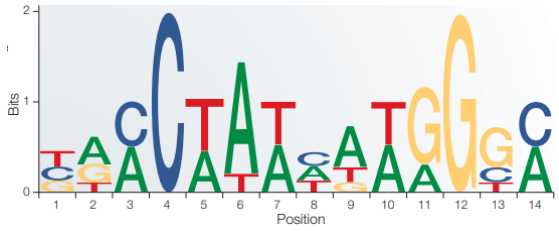


Fig. 3: Exemple de motif PWM. La taille des lettres des nucléotides varie suivant la probabilité de le trouver en position donnée par l'abscisse.

Ajout de variables d'interaction

Le jeu de données final contient autant de paires positives, c'est à dire dont l'enhancer et le promoteur sont en contact, que de paires négatives. Chaque élément d'une paire est caractérisé par sa composition nucléotidique et les scores de 662 PWM. Des jeux de données ont également été construits avec uniquement l'une de ces deux variables pour en tester l'utilité.

Les modèles considérés sont soit de nature linéaire, comme la régression logistique pénalisée, ou non linéaires, comme les méthodes ensemblistes. Il est possible que les variables de l'enhancer et du promoteur combinées de manière linéaire ne suffisent pas à expliquer l'appariement de ces régions. Contrairement aux modèles linéaires, les modèles non linéaires sont supposés pouvoir potentiellement capter des interactions intéressantes pour la prédiction. Mais il est également possible que même les modèles non linéaires aient du mal à discerner clairement cette interaction. Il pourrait alors s'avérer utile de leur fournir une variable représentant directement les deux composantes de la paire. Plusieurs fonctions on été envisagées, traduisant différentes hypothèses biologiques. Tout d'abord, le produit des variables de l'enhancer et du promoteur a été testé, le produit étant généralement utilisé dans les modèles linéaires avec interaction. Puis, pour traduire une similarité des compositions des séquences, nous avons choisi la valeur absolue de la différence des deux variables, ou la valeur absolue du logarithme de leur rapport. Un minimum de ces deux valeurs pourrait donner l'équivalent d'un "et" logique, un maximum représenterait un "ou" logique pour les variables de compositions en nucléotides ou scores PWM. Ces variables sont nommées "variables d'interaction".

Au terme de ces manipulations, nous obtenons des données de travail, se présentant comme en figure 4. Les variables d'interaction seront ajoutées ou non pour en tester l'influence sur les modèles.

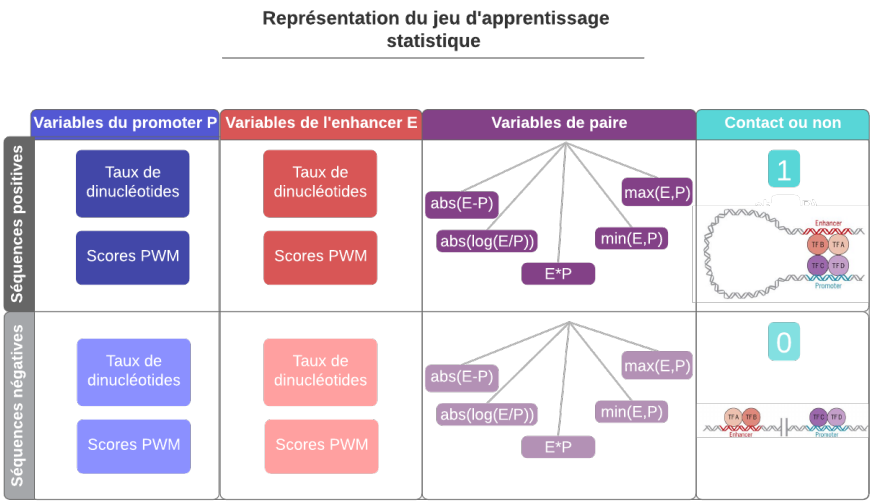


Fig. 4: Structure du jeu de données utilisé pour construire les modèles de classification. Images : Developmental enhancers and chromosome topology, Furlong et Al, 2018 [7]

2.3 Méthodes de classification

Pour quantifier le niveau d'information porté par les séquences pour prédire les interactions chromosomiques, des méthodes de classification de nature différente ont été envisagés et comparés.

La régression logistique pénalisée

La régression linéaire est une méthode statistique permettant de trouver une relation linéaire entre une variable réponse Y et un ensemble de variables prédictives X. Son écriture matricielle est la suivante :

$$Y = X\beta + \epsilon, \tag{1}$$

avec β le vecteur de coefficients, et ϵ l'erreur aléatoire, dite résiduelle. Lorsque l'on ajuste un modèle linéaire, le vecteur de coefficients est estimé de manière minimiser l'erreur quadratique $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, avec y_i la valeur à prédire, et $\hat{y}_i$ la valeur prédite par le modèle.

Les coefficients estimés permettent de réaliser des prédictions sur la variable réponse. Cependant, cela reste insuffisant pour identifier les variables les plus importantes pour la classification. Cette question est particulièrement importante quand la dimension de l'espace des variables est grande, nécessitant d'identifier les variables les plus informatives.

Pour cette raison, il est possible d'appliquer une méthode pénalisant la régression classique. Lors de l'estimation des coefficients, une contrainte supplémentaire est ajoutée sur le vecteur de coefficients, dont la norme ne doit pas excéder une certaine constante λ, appelée pénalité ou paramètre de tuning. Suivant la norme choisie, on parle de différentes méthode de régularisation. S'il s'agit de la norme 1, c'est à dire la somme en valeur absolue des coefficients, on parle de la méthode du Lasso, pour Least Absolute Shrinkage and Selection Operator. Lorsque la norme 2 est employée, c'est à dire la somme du carré des coefficients on parle de la pénalisation Ridge. Une combinaison de ces deux normes conduit à la régularisation Elastic Net. Plus le paramètre de tuning est stringeant, plus le modèle est contraint et doit diminuer la valeur des coefficients qui lui sont le moins utiles pour minimiser l'erreur. La régularisation que nous avons choisie est la méthode du Lasso, en raison de sa capacité à fixer la valeur des coefficients les moins informatifs à 0, comme illustré sur la figure 5.

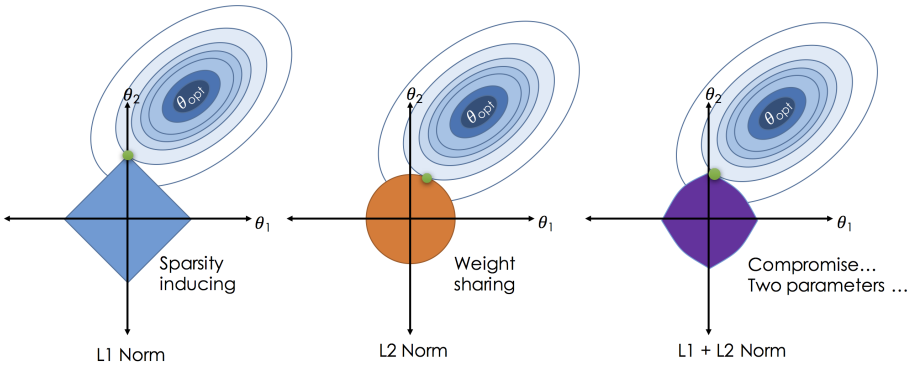


Fig. 5: Représentation des différents types de régularisation dans une espace de coefficients à deux dimensions. La zone θ_{opt} représente les valeurs de coefficients minimisant l'erreur prédictive du modèle. Lorsque l'on ajoute une régularisation, les coefficients estimés seront à l'intersection entre cette zone optimale et la boule de la norme mathématique associée à la pénalité choisie. La norme L1, pour le Lasso, crée une boule carrée, favorisant le placement des coefficients sur l'un des axes du plan, le mettant à 0. La méthode Ridge utilisant la norme 2 minimise ces coefficients sans les fixer à 0, et Elastic Net trouve un compromis entre les deux. Source : The Biais Variancec Readeoff and regularization, slides by Joseph E. Gonzales.

La fonction à optimiser lors de l'estimation des coefficients se formalise donc

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \qquad (2)$$

avec n le nombre d'observations utilisées pour ajuster le modèle, et p le nombre de variables explicatrices. L'argmin de cette fonction, c'est à dire le vecteur de coefficients β, est obtenu par des méthodes d'optimisation telles que la descente de gradient.

Pour choisir la valeur de λ, une plage de valeurs est généralement testée. Des modèles sont ajustés sur un jeu d'entraînement, pour différentes valeurs de λ, et testés pour chaque valeur de λ sur un jeu de test indépendant. La valeur choisie est celle qui minimise l'erreur sur le jeu de test. Lorsque le jeu de données n'est pas suffisamment grand, il vaut mieux procéder par k-fold cross validation. Cette technique consiste à découper le jeu de données en k partitions. k-1 partitions sont utilisées pour entrainer le modèle, et la dernière permet de le tester. Cette démarche est répétée k fois, de manière à ce que chaque partition ait servi une fois de test. Pour chaque jeu de test, on choisit la valeur optimale de λ. En regardant les valeurs de λ sur les k jeux de test, nous avons une valeur de λ qui maximise les performances calculée sur l'ensemble du jeu d'apprentissage.

Une régression linéaire pénalisée réalise des prédictions quantitatives. Dans notre contexte, un modèle de classification permettant de prédire une sortie binaire (interaction ou non des régions chromosomiques présentées) est requis. C'est donc la régression logistique du modèle linéaire généralisé que nous avons utilisée. Plutôt que d'estimer la valeur Y directement, le logarithme de l'odds d'interaction, également nommée fonction logit, est approximé par une combinaison linéaire des variables. La régression logistique se formalise ainsi :

$$\text{logit}(p_{\text{interaction}}) = \ln(\frac{p_{\text{interaction}}}{1 - p_{\text{interaction}}}) = X\beta + \epsilon, \tag{3}$$

avec $p_{\text{interaction}}$ la probabilité de contact entre deux régions régulatrices.

Le principe de régularisation peut tout aussi bien y être appliqué afin de sélectionner les variables les plus importantes. Des implémentations de régressions logistiques pénalisées incluant la validation croisée sont disponibles en R dans le package `glmnet` [5], ou en python dans la librarie `scikt learn` [11]. C'est la version du langage R qui a été majoritairement utilisée au cours du stage, avec la fonction la fonction `cv.glmnet` et 10 folds pour la validation croisée.

Les Random Forests

Un arbre de décision est une structure prédictive construite à partir de grandes quantités de données. Les variables permettant de discriminer au mieux la variable réponse sont repérées à partir de mesures d'entropie ou de l'indice de Gini suivant la nature de l'arbre. Chaque noeud représente une condition sur une variable. Plus une variable explique la séparation de la réponse entre les classes à prédire, plus elle est placée en haut de l'arbre. Au bas de l'arbre, se trouvent les feuilles représentant la décision du modèle. Il s'agit de modèles très interprétables, car ils peuvent être vus comme une série de requêtes et de tests à propos des données aboutissant à une prédiction.

Un arbre de décision seul peut fournir des résultats satisfaisants, mais est connu pour souffrir de problèmes de stabilité et d'*overfitting*. La combinaison de plusieurs arbres permet une performance accrue dans la classification, tout en réduisant considérablement les risques d'*overfitting*. Le principe de cette méthode dite ensembliste est de créer un ensemble d'arbres qui, pris séparément, sont des *weak learners*[3], mais dont les décisions collectives sont significativement meilleures. Les Random Forests sont introduites pour la première fois en 2001, avec la publication de Breiman et Al. [3].

Pour parvenir à un meilleur pouvoir de classification, les arbres de la forêt doivent pouvoir rendre des décisions complémentaires, dues à des différences structurelles. Deux techniques sont employées dans cette optique :

- Le *bootstrapping* : les différents arbres sont entraînés sur des jeux de données échantillonnés avec remise dans le jeu de données initial. Ainsi, certaines observations seront dupliquées, d'autres seront absentes suivant les jeux d'entraînement, causant des choix de variables discriminantes différents entre les arbres.
- À chaque noeud, l'arbre de décision ne pourra utiliser qu'un nombre réduit de variables tirées parmi la liste totale des variables.

Ces opérations associées à l'*averaging* des prédictions de tous les arbres, permettent de réduire le bruit présent dans les données. Il est à noter que l'aspect hautement interprétable des arbres est perdu lorsqu'on utilise des forêts. Cependant, il est également possible d'accéder assez facilement à l'importance des variables utilisées pour la classification d'un modèle de *Random Forests*. L'une des méthodes principales repose sur les *Out Of the Bag examples*, les éléments ayant été écartés de l'apprentissage lors du *bootstrapping*. Ces éléments n'ayant pas participé à l'ajustement des modèles peuvent être utilisés pour tester l'importance des variables. Cette importance est donnée par la différence d'*accuracy* entre le modèle sur les données de test, et le jeu de test dans lequel la variable d'intérêt a été randomisée. Cette perte de performance pour une variable est calculée pour tous les arbres, et permet de quantifier en moyenne l'apport de cette variable dans les capacités de classification d'une Random Forest. C'est cette technique qui est utilisée par les `RandomForestClassifier` de `scikit-learn`.

Les *Random Forests* utilisés dans le cadre du stage utilisent le critère de Gini pour quantifier la pureté d'un noeud, et contiennent 200 individus. Cette valeur de 200 individus a été obtenue au terme d'une GridSerachCV optimisant le critère d'AUC[4] du modèle.

Le Stochastic Gradient Boosting

L'intuition du *boosting* est de corriger de manière itérative des modèles prédictifs, en accordant une importance plus forte aux éléments les moins bien

[3] Un *weak learner* est un modèle de *machine learning* effectuant des prédictions meilleures que le hasard, même légèrement.

[4] Ce critère est décrit en section 2.4

estimés. Cette méthode peut être appliquée à des arbres de décisions et entre, tout comme les *Random Forests*, dans la catégorie des méthodes ensemblistes. La différence est que les différents prédicteurs ne sont pas entraînés de manière indépendante, mais séquentielle.

Le *Gradient Boosting* correspond au fait d'ajuster un nouveau modèle, à chaque itération de l'algorithme, aux résidus du modèle précédent. En effet, si les prédictions ne sont pas satisfaisantes, on peut supposer qu'un *pattern* reste à trouver dans l'erreur résiduelle du modèle. On note F_m le modèle appris à l'itération m, et $F_m(x)$ la prédiction de ce modèle pour un exemple x vectoriel. Les pseudo-résidus sont calculés comme la dérivée négative de l'erreur par rapport aux prédictions du modèle, notés h_{m+1}. Ces résidus servent de jeu d'entraînement pour le modèle suivant : la valeur de γ est trouvée de manière à minimiser l'erreur du nouveau modèle $F_{m+1}(x) = F_m(x) + \gamma h_{m+1}(x)$. Ce modèle est ensuite ajouté à l'ensemble des modèles et servira de base pour l'itération suivante. La décision du modèle de *Gradient Boosting* sera la moyenne des M modèles ainsi obtenus.

Le *Stochastic Gradient Boosting* ajoute une composante aléatoire à l'algorithme précédant. Le principe est de n'utiliser qu'une fraction du jeu de données total tirée aléatoirement pour entraîner les modèles $F_m(x)$ au début de chaque itération. Cet ajout permet un gain en robustesse par rapport au *Gradient Boosting*. Cependant, les méthodes additives telles que le *Gradient Boosting* sont connues pour être un peu plus sensibles à l'*overfitting*[5] que les méthodes indépendantes comme les *Random Forests*.

Perceptron multi-couches

Un réseau de neurones est une technique de *deep learning* usuellement utilisée en classification supervisée. Un réseau se compose d'unités élémentaires excitables, les neurones. Disposées en couches et reliées entre elles, en l'occurrence de manière *fully connected*, cette structure permet l'approximation de fonctions. Chaque connexion est caractérisée par un poids, et ces poids sont appris de manière à minimiser l'erreur entre les prédictions du réseau et ce qui est attendu. Un apprentissage se compose de deux phases, itérativement répétées. Tout d'abord, la propagation de l'activité est faite à partir d'une donnée présentée au réseau. La prédiction du réseau est comparée à la sortie attendue, fournissant une valeur d'erreur. Une rétro-propagation de cette erreur est alors effectuée, corrigeant les poids en fonction de l'erreur faite. Un réseau de neurones a de multiples hyper-paramètres, comme le nombre de couches, le nombre de neurones par couches, la fonction d'activation des neurones, ou la fonction d'erreur. La méthode de correction des poids, qui lie la fonction d'erreur aux ajustements faits aux connexions de modèle, est dite *optimizer*. Un perceptron multicouche est utilisé dans la seconde partie du projet, dans un optique d'exploration et

[5] L'*overfitting* correspond à un sur-spécificité d'un modèle par rapport à son jeu d'apprentissage, et par conséquent à une incapacité de généralisation correcte sur de nouvelles données.

de validation simple. Ne représentant pas un axe principal de ce stage, l'accent ne sera donc pas mis sur sa paramétrisation ni sur la recherche d'une structure de réseau plus complexe. Le perceptron utilisé est le perceptron `MLP` du package sci-kit learn. Il est utilisé avec la fonction d'activation RELU, l'*optimizer* Adam et 100 neurones par couches. Le nombre de couches est appris afin de maximiser l'AUC.

2.4 Évaluation des performances d'un modèle

Lors de l'évaluation d'un modèle de classification, il est d'usage de comparer le vecteur de prédictions obtenu sur un jeu de test aux labels connus, aussi dits *ground truth*. La comparaison de ces deux vecteurs peut être faite de différentes manières afin de renseigner sur les performances du modèle. Une approche simple est de calculer la proportion d'éléments correctement classifiés, donnant une mesure de l'*accuracy*. Il est également possible de compter le nombre de vrais/-faux négatifs et positifs, donnant des mesures de précision et de rappel pour ces classes.

Durant ce stage, nous avons choisi l'utilisation du critère d'Area Under the Curve (AUC). L'AUC mesure l'aire sous la courbe ROC. Cette courbe est obtenue en traçant les taux de vrais positifs en fonction des taux de faux positifs pour différentes valeurs seuils binarisant les prédictions quantitatives[6]. Un modèle donnant une AUC à 0.5 est comparable à une classification au hasard, et plus l'AUC se rapproche de 1, plus le modèle est performant. Alors que l'*accuracy* pourrait être haute pour un classifieur prédisant invariablement la classe la plus fréquente sur un jeu de données déséquilibré, l'AUC s'affranchit de se problème en se concentrant sur le pouvoir séparateur d'un modèle par rapport à des prédiction aléatoires. Un exemple de courbe ROC et de l'AUC associée sont visibles en figure12.

2.5 Extraction de *features*

Les variables utilisées pour faire des prédictions dans les modèles précédents sont des variables relatives à la totalité des séquences régulatrices considérées, promoteurs et enhancers. Il est légitime de se demander si une extraction de *features* plus précise et permettant de ne s'intéresser à des compositions de nucléotides que sur certaines zones des séquences régulatrices pourrait être informatif.

DExTER

[6] Cela correspond à tracer la sensibilité du modèle en fonction de un moins sa spécificité. D'autres courbes existent et font apparaître la précision en fonction du rappel : leur AUC permet de définir un critère de performance de classification encore différent.

Principe DExTER (Domain Exploration To Explain gene Regulation) est un algorithme développé par Christophe Menichelli au cours de sa thèse dans l'équipe MAB. Sa création a été motivée par l'intuition que, dans une séquence, seulement certaines portions peuvent contenir l'information recherchée pour expliquer la régulation des gènes. L'objectif est donc d'extraire les régions dont la composition en nucléotides s'avère être la plus corrélée à une variable réponse étudiée.

Dans l'article [2], il a été montré que la segmentation des régions promotrices s'avérait porteuse d'information pour la prédiction de la régulation de l'expression. Cette segmentation, basée sur la connaissance à priori de différents régions impliquées, était cependant manuelle, et la nécessité d'un outil pouvant la réaliser de manière automatique s'est imposée. Dans un premier temps, DExTER a servi à prédire l'expression de gènes étant donnée la séquence de leurs promoteurs, centrée autour du TSS. Ainsi, en fournissant un fichier de séquences promotrices de gènes, associé à un fichier contenant le niveau d'expression de chaque gène, DExTER est capable de fournir un ensemble de variables explicatrices de cette expression.

DExTER entre dans la catégorie des méthodes d'extraction de *features*. Une exploration des séquences est conduite de manière à déterminer quelles sont les paires (compositions nucléotidiques, régions de séquences) les plus prédictives pour une réponse donnée, ici l'expression. Une variable, est donc représentée par un tuple (fréquence de k-mer, région), dénotés $D_{k,r}$. L'objectif est d'identifier un ensemble de variables dont les valeurs sont corrélées à l'expression. Pour identifier les meilleures régions associées à un k-mer, on se base sur la structure en treillis sous-jacente à l'ensemble des sous-régions.

Un treillis est un objet mathématique associé à un ordre pour lequel chaque paire d'élément admet une borne supérieure et inférieure. Ici chaque élément est une région de la séquence et l'ordre considéré est l'inclusion. La borne supérieure d'une paire de région est la plus petite région qui contienne les 2 éléments de la paire. Formellement, nous considérons un demi-treillis puisqu'il n'y a pas de borne inférieure. Nous utilisons ce treillis des régions comme un support pour calculer ensuite les pourcentages de k-mer dans chaque région. Deux lignes partant respectivement de régions A et B s'intersectent en un noeud qui contient la proportion du k-mer dans la plus petite région qui inclut A et B. Plus on monte dans le treillis, plus on se rapproche de la composition globale du k-mer dans la séquence entière. La figure 6 montre un exemple de cette structure.

Procédure d'exploration Nous souhaitons trouver des paires (taux de k-mer,région) corrélées à l'expression des gènes. Cette corrélation s'écrit

$$Cor = Spearman\big(D_{k,r}, Y\big)$$

avec $D_{k,r} = (d_{k,r,1}, ..., d_{k,r,i}, ..., d_{k,r,N})$ le vecteur des taux du k-mer k sur la région r pour les N séquences, et $Y = (y_1, ..., y_i, ..., y_N)$ le vecteur d'expression des N séquences. Cette mesure de corrélation est explicitée en figure 7.

Cependant, tester tous les k-mer de taille maximale K, sur toutes les régions qu'il est possible de créer en découpant une séquence en R portions constitue un espace extrêmement grand à explorer. En l'occurrence, la complexité d'un

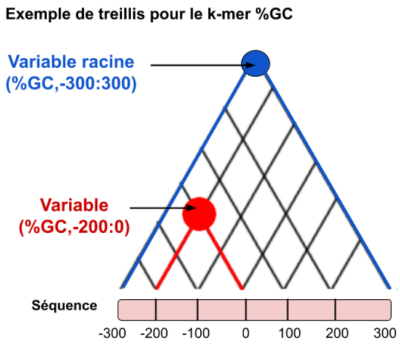


Fig. 6: Exemple schématique d'un treillis pour le dinucléotide GC sur une séquence de 601 bases.

tel algorithme d'exploration peut être calculée. Si l'on choisit de découper une séquence en R régions, il existe $\frac{R^2}{2}$ régions à explorer (nombre de noeuds d'un demi treillis de résolution R). Pour chaque région, il faut tester tous les k-mers possibles, des dinucléotides jusqu'aux K-mers. Cela nous donne un nombre de $4^2 + ... + 4^K = \Sigma_{k=2}^{K} 4^k = \frac{4^{K+1}-4^2}{4-1}$ (formule d'une série géométrique) k-mers. Enfin, cette exploration doit être faite pour chacune des N séquences des données, aboutissant une complexité finale en $\mathrm{O}(NR^2 4^K)$.

Cette complexité étant beaucoup trop élevée, on utilise une heuristique itérative qui fonctionne en 2 étapes que l'on répète tant que l'on observe une amélioration de la corrélation:

- Étant donné un k-mer on cherche les meilleures régions. Cette exploration est faite sur la base du treillis,
- On essaye d'étendre ce k-mer à un k+1-mer sur la même région.

La procédure d'exploration consiste à parcourir les variables notées $D_{k,r}$ représentant le k-mer k sur la région r de toutes les séquences, et d'exhiber celles qui se montrent les plus corrélées au vecteur réponse Y. La liste L des variables à explorer est initialisée aux dinucléotides sur une région r correspondant à l'ensemble de la séquence. Au cours de la procédure, des variables sont ajoutées à la liste à explorer en rajoutant un nucléotide au k-mer courant, ou en ajustant la région de ce k-mer.

On désigne par "parent" une variable qui donne naissance à une nouvelle variable. Cette nouvelle variable peut soit être une variable de k-mer identique à son parent, mais sur une région différente, soit une variable de k+1 mer sur la même région que son parent. Ces phases sont alternées afin de conduire l'exploration de manière équilibrée. Lorsque l'on vient d'ajouter un nucléotide à une variable, l'étape suivante est de ré-effectuer une segmentation pour trouver la région, plus restreinte ou plus large, qui maximise la corrélation de ce k-mer avec Y.

Pour une région donnée, il s'agit d'une corrélation de Spearman entre le vecteur de taux de k-mer sur cette région et la variable réponse, cela pour l'ensemble des séquences fournies.

Sont retenus et ajoutés à la liste des variables validés L_{val} les variables ayant une corrélation à Y supérieure à celle de leur parent, et qui vérifient une condition de contrôle. Cette condition de contrôle $\text{Cor}(D_{k,r}, Y) > \text{Cor}(D_{k-1,r}, Y)$ assure qu'on a bien une augmentation de corrélation pour ce variable par rapport à une variable avec un $k-1$-mer différent de son parent, testant ainsi si l'effet est bien dû au nouveau k-mer et pas à un k-mer plus petit qu'il contient.

Ce contrôle de l'augmentation de corrélations est paramétrable par l'utilisateur, qui peut choisir une augmentation absolue d'une certaine valeur, ou bien un *increase ratio* de son choix. Le pseudo-code de cette exploration des variables est donné par l'Algorithme 1 en annexe .

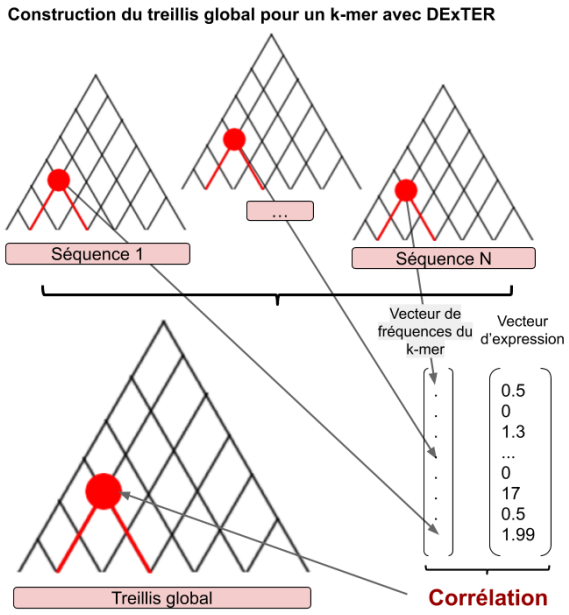


Fig. 7: Calcul de la corrélation d'une variable avec un vecteur d'expression Y. Le vecteur des taux de k-mer pour toutes les séquences et une région r donnée est obtenu à partir des treillis des fréquences de chaque séquence (treillis du haut). Ces treillis individuels sont tels que celui présenté en figure 6. Un treillis global des corrélations est ensuite construit en calculant le score de corrélation entre le vecteur de taux du k-mer considéré sur la région r de toutes les séquences, et Y. Il permet de visualiser la corrélation d'un k-mer à la variable réponse sur l'ensemble des séquences à disposition.

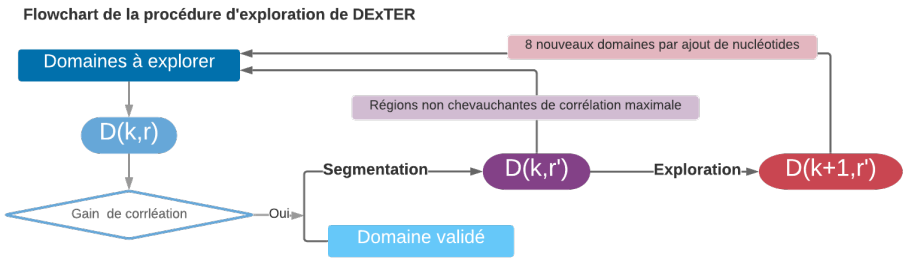


Fig. 8: Schéma représentant le flow d'information au cours de l'exploration des variables par DExTER. Les variables sont testés en termes de corrélation à Y, et s'ils présentent meilleure corrélation que celle de leurs parents, ils sont étendus ou segmentés pour créer de nouveaux variables à tester.

DExTRA

Nous avons souhaité adapter la méthode DExTER à la problématique de proximité spatiale des régions régulatrices. L'objectif est de pouvoir extraire des variables expliquant au mieux les contacts entre régions régulatrices dans le noyau d'une cellule, plutôt que l'expression des gènes comme le fait la version originale.

Cette adaptation de DExTER, nommée DExTRA (Domain Exploration To explain chromosomic Regions Associations) a donc pour objectif de traiter des paires de régions régulatrices pour en prédire l'interaction dûe au repliement spatial de la molécule d'ADN. Pour cela, des données contenant un ensemble de paires et les contacts associés sont données à l'algorithme en tant que vecteur à prédire Y.

Jeu de données Les données utilisées ici sont les données issues de ChIA-PET. La liste des séquences régulatrices impliquées dans des interactions chromosomiques a été faite, puis une matrice répertoriant le nombre de contacts entre ces séquences a été construite. Cette matrice, extrêmement *sparse*, se remplit essentiellement autour de sa diagonale, mettant en contact des régions déjà proches dans l'ordre chromosomique. Des contacts plus distants sont également visibles, de manière symétrique autour de cette diagonale. Un ensemble de paires peut donc être constitué, en échantillonnant des couple de séquences régulatrices avec une distribution en *counts*, c'est à dire nombre de contacts, la plus uniforme possible. Nous sélectionnons également à l'aide de cette matrice la même quantité de paires de séquences ayant une valeur nulle de *counts*, afin de représenter les paires de *background*. Enfin, ces *counts* sont binarisés, pour associer à une paire la valeur 1 s'il existe des interactions spatiales entre les deux promoteurs, et 0 sinon. Pour la problématique de la prédiction des interactions promoteur-promoteur, 8269 paires constituent ce vecteur.

Adaptation de l'algorithme DExTER Le principe d'exploration des séquences, leur segmentation et la validation des variables informatives sont conservés par rapport à DExTER. Dans DExTRA, la différence intervient dans le critère à maximiser pour sélectionner ou non une variable lors de cette exploration. Alors que DExTER mesure une corrélation entre les fréquences de k-mer d'une liste de séquences et les expressions relatives à chacune de ces séquences lors de la construction d'un treillis global, DExTRA utilise un score différent. Ce score doit renseigner sur le niveau d'information apporté par les fréquences en k-mer des deux éléments de la paire quant à leur mise en contact ou non.

La liste de paires obtenues après le traitement des données de ChIA-Pet est donc fournie à DExTRA. Pour chaque variable $D_{k,r}$ rencontrée lors de l'exploration, on crée un vecteur calculé comme la valeur absolue de la différence du taux du k-mer k calculé dans les 2 séquences. Cela fournit un vecteur caractérisant les paires de séquences régulatrices par une grandeur représentant leur similarité en termes de fréquence de k sur la région r. Ce vecteur doit ensuite être comparé à la variable réponse, l'interaction ou non des promoteurs de la paire, modélisé par Y. Y étant binaire, une mesure de corrélation comme le faisait DExTER n'est pas la plus pertinente. C'est donc une mesure d'AUC sous la courbe ROC qui est utilisée. Cette mesure, usuellement utilisée pour évaluer la qualité des prédictions quantitatives d'un modèle par rapport au vecteur à deux classes attendu, peut également être utilisée ici. Cela revient à estimer la pertinence d'un modèle à une seule variable, ici la différence en valeur absolue des taux de k sur r, pour prédire la variable réponse.

Par conséquent, la procédure de construction d'un treillis de score global est modifiée, telle que montrée en figure 9.

Classification À l'issue de la méthode DExTRA, nous obtenons un ensemble variables dont les valeurs dans les paires d'entraînement optimisent la prédiction des contacts entre séquences. Le finalité de cette procédure d'extraction de *features* est de les combiner dans un modèle, qui sélectionnera les plus importantes, afin de prédire au mieux les interactions chromosomiques. Les modèles décrits en section 2.3 sont alors envisagés en utilisant deux types de variables explicatives, présentées en figure 10 :

- 1. La valeur absolue de la différence des variables des deux séquences
- 2. Chaque variable dans les deux séquences

3 Résultats

3.1 Prédiction des interactions enhancer-promoteur sans extraction de *features*

Cette section rend compte des résultats liés aux méthodes de la section 2.1, visant à prédire les interactions de paires enhancers-promoteurs. Ces prédictions

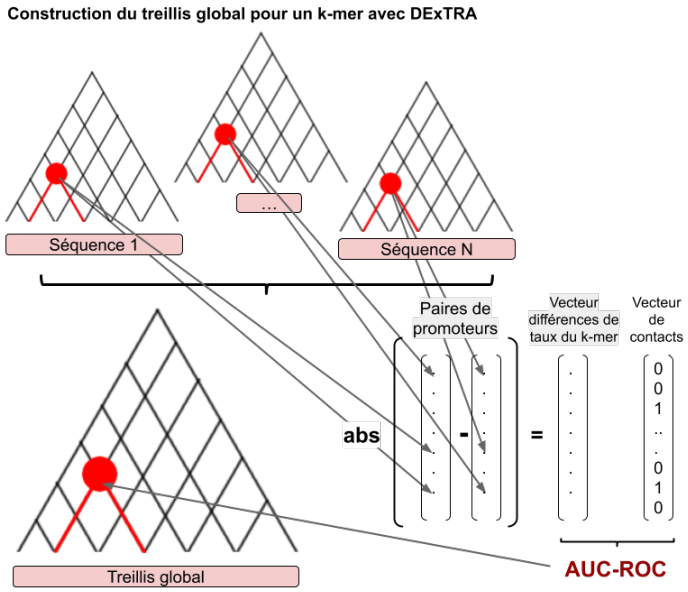


Fig. 9: Calcul de l'AUC d'une variable avec un vecteur de contacts Y. Le vecteur des taux de k-mer pour toutes les séquences et une région r donnée est obtenu à partir des treillis de chaque séquence individuelles. Ces treillis individuels sont tels que celui présenté en figure 6). Un treillis global est ainsi construit comportant le score d'AUC entre le vecteur de différence en valeur absolue des taux du k-mer considéré sur la région r de toutes les séquences, et Y. Il permet de visualiser le score prédictif d'un k-mer par rapport à la variable réponse sur l'ensemble des paires positives ou négatives formées.

sont faites sur les compositions en nucléotides et les motifs PWM sur l'ensemble des séquences, sans segmentation, et sur le jeu de données de la figure 4.

Validation sur un jeu de données artificiel

Afin de s'assurer de la validité des modèles développés, nous avons souhaité les tester sur des jeux de données fictifs pour la problématique enhancer-promoteur. Les taux du score PWM d'un certain facteur de transcription, ATF1, ont été modifiés de manière à ce que les scores de l'enhancer et du promoteur soient proches dans les paires positives, et inchangés dans les paires négatives. La modélisation est la suivante : pour une proportion p de paires positives, la valeur de score du promoteur est tirée suivant une loi normale centrée en la valeur de l'enhancer, et d'écart type 0.1. Pour les autres paires positives, et pour les paires

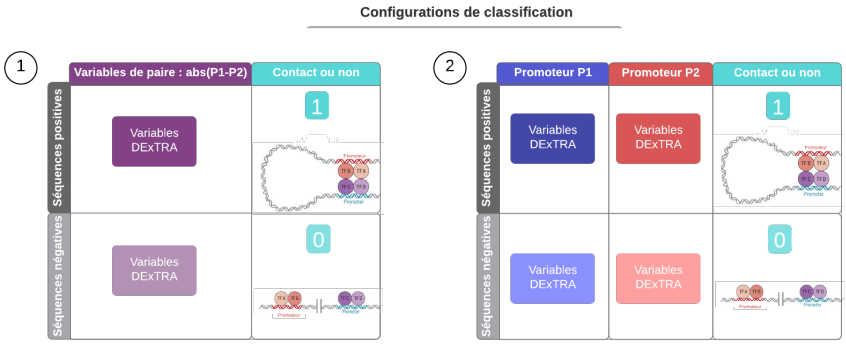


Fig. 10: Différents jeux de données utilisés pour entraîner et tester les modèles.

négatives, le score du PWM dans le promoteur n'est pas modifié. Les modèles de classification sont alors construits, incluant les taux de PWM et dinucléotides pour l'enhancer et le promoteur, ainsi que le *logratio* de ces deux valeurs. Avec une probabilité $p = 0.9$, cette composition similaire est bien visible est captée par les modèles, comme le montre la Supp. figure .4 en annexe 5, graphe de sélection des variables lors de la régression logistique pénalisée.

Les autres modèles classent également cette variable d'interaction relative à ATF1 comme la plus importante. Les AUCs de classification sont alors extrêmement hautes, excédant 90%, car l'information insufflée au dataset pourrait être plus subtile (en faisant notamment varier p). Nous en concluons que lorsque les données comportant de l'information discriminant les paires en contact des autres, les classifieurs sont capables de la détecter.

Performances et variables importantes sur les 2 problèmes FANTOM

Tout d'abord, nous avons considéré deux manières de construire les paires positives et négatives, formant deux problèmes différents, décrits en 2.1. Pour étudier ces deux cas, 300 modèles de classification ont été entraînés, chacun sur un type cellulaire représenté dans les données FANTOM. Ces modèles sont construits en prenant comme variables les compositions nucléotidiques de l'enhancer et du promoteur.

Le problème consistant à prédire les paires dont l'enhancer est actif contre celles dont l'enhancer est inactif s'est avéré plus évident à prédire en classification. Le second problème, prédisant l'expression simultanée contre seulement l'un des membres exprimés dans la paire, présente des AUC inférieures. Nous constatons que ces résultats sont cohérents suivant les modèles utilisés : qu'il s'agisse des la régression logistique pénalisée (Lasso), des *Random Forests*, ou du *Stochastic Gradient Boosting*, les distributions d'AUC sont comparables pour

un problème donné. La figure 11 montre les valeurs d'AUC pour 300 types cellulaires, les différents modèles utilisés ainsi que le problème traité.

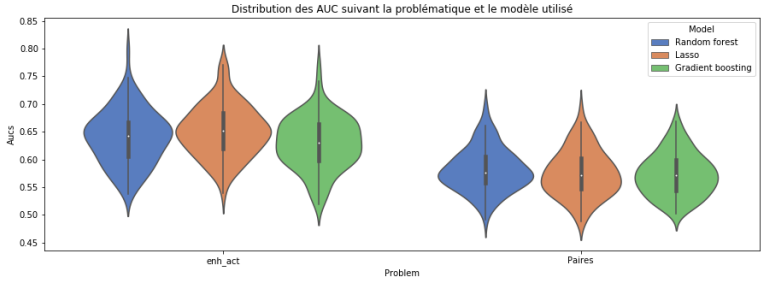


Fig. 11: Valeurs d'AUC pour deux problèmes de prédiction de paires enhancer-promoteur. À gauche : enhancer actif ; à droite : expression simultanée.

Cette différence de performances entre les problèmes s'explique par le fait que le premier problème revient à prédire l'activité de l'enhancer, ce qui peut se faire en utilisant simplement la séquence de celui-ci. Les AUCs du second problème, légèrement supérieures à 0.5, suggèrent que les prédictions de tels contacts ne peuvent pas être faites sur la bases des compositions nucléotidiques uniquement. Nous étudions par la suite ce problème dans un type cellulaire précis.

La construction du jeu d'apprentissage précédemment décrit permet l'étude des importances relatives des variables de l'enhancer et du promoteur, des variables de fréquences nucléotidiques aux PWM, de l'ajout de variables d'interaction, en comparant les performance de prédiction des contacts. La construction d'un modèle avec la composition en nucléotides a donc été comparée avec un modèle utilisant la composition en nucléotides et les scores PWM, à la fois pour l'enhancer et le promoteur. Ce modèle est créé pour classifier le problème d'expression simultanée des paires FANTOM, et n'utilise pas de variables d'interactions. Les figures 12 et 13 montrent les résultats de telles classifications, prenant en compte ou non les PWM.

Les AUC sont faibles, légèrement supérieures à un classifieur aléatoire. L'ajout des scores PWM ne semble pas informatif par rapport à l'appariement de ces séquences régulatrices, pouvant même conduire à une perte de performances causée par un grand nombre de variables non utiles.

Nous avons également construit des modèles utilisant uniquement les enhancers ou les promoteurs, pour déterminer si les variables de l'un ou de l'autre peuvent être prépondérantes dans l'information. La table 1 référence les valeurs d'AUC et montre que ce sont majoritairement les variables de l'enhancer qui portent le faible signal prédictif.

Performances et validation sur les données de ChIA-PET

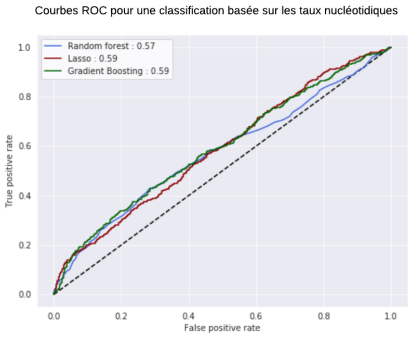


Fig. 12: Courbes ROC et AUC pour un modèle entraîné sur les variables de taux de nucléotides des enhancers et des promoteurs

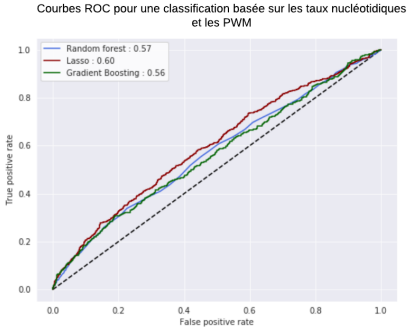


Fig. 13: Courbes ROC et AUC pour un modèle entraîné sur les variables de taux de nucléotides et les scores PWM des enhancers et des promoteurs

Variables nucléotidiques pour :	Lasso	Random Forest	Gradient Boosting
Enhancer et promoteur	0.59	0.57	0.59
Enhancer	0.59	0.56	0.56
Promoteur	0.50	0.50	0.52

Table 1: Valeurs d'AUC pour les différents classifieurs et différents membres des paires injectés dans le modèle.

Nous avons ensuite appliqué ce modèle aux données de ChIA-PET. Dans ce cas, les paires négatives sont les paires positives randomisées. Nous avons donc construit des modèles avec les variables simples (colonnes rouge et bleue de la figure 4), puis ces mêmes modèles avec, en plus, des variables d'interaction (toutes les colonnes de la figure 4). Les résultats sur le modèle de base, avec les variables de nucléotides pour l'enhancer et le promoteur confirment ceux obtenus avec les paires FANTOM, comme le montre la Supp. figure .2, avec des AUCs entre 0.53 et 0.54. L'ajout des variables d'interaction au jeu de données, montré en annexe 4, Supp. figure .3 n'a pas permis d'amélioration des prédictions, avec des AUCs également comprises entre 0.53 et 0.54.

Les autres fonctions utilisées pour créer des variables d'interactions se sont montrées tout aussi peu informatives sur cette problématique. Sur cette partie, apparaît une réelle concordance des résultats entre les données FANTOM et les données de ChIA-PET : qu'il s'agisse d'un appariement inféré par corrélations, ou des contacts exhibés par expériences biologiques, il semblerait que les variables de séquence étudiées ne permettent pas de prédire le repliement de la chromatine.

3.2 Prédiction des interactions promoteur-promoteur avec extraction de *features*

Dans cette section, sont présentés les résultats obtenus avec la méthode DExTRA, présentée en section 2.5, sur la problématique promoteur-promoteur. A la suite d'une exploration par fractionnement des séquences, des variables associant un k-mer à une région de la séquence sont retournés, et sensés expliquer au mieux individuellement les appariements promoteur-promoteur observés dans les données de ChIA-PET. Ces variables sont ensuite injectées dans un modèle global pour prédire les contacts en combinant toutes ces variables. Nous avons dans un premier temps pensé à des régions chromosomiques de taille similaire à celles de la partie précédente, c'est à dire $\pm 500bp$ autour du centre du promoteur. Nous avons cependant souhaité élargir ces régions à $\pm 10000bp$, afin d'être sûr de ne manquer aucune information, laissant à DExTRA la liberté de choisir des domaines plus restreints au besoin.

Les résultats sont relativement surprenants. En effet, lors de l'exploration des domaines, DExTRA génère des treillis globaux avec l'AUC des k-mers par rapport à la réponse sur l'ensemble des séquences, comme celui montré en figure14. Plus les AUCs d'une région sont hautes, plus cette région explique les contacts chromatiniens. Ici, plutôt que de montrer des AUC fortes sur la région promotrice, que l'on supposait jusqu'ici informative sur les contacts, c'est son voisinage qui semble explicatif. L'AUC maximale est obtenue au sommet du treillis, c'est à dire la région correspondant à la séquence complète $\pm 10000bp$, car le sommet combine l'information du voisinage amont et aval du promoteur. Une zone "morte", avec peu d'informations apportées sur la réponse à prédire est située au centre de la séquence, à l'endroit du promoteur.

Une étude des treillis obtenus avec les différent di-nucléotides montre que ce *pattern* ne se retrouve pas avec tous les k-mers, mais seulement avec ceux qui traduisent le taux de GC (proportion de nucléotides G ou C dans les séquences, à ne pas confondre avec le dinucléotide GC). On parle de *GC content*. Les dinucléotides impliqués dans le *GC content* et exhibant ce *pattern* sont en l'occurrence AT, TA, GC, CG, TT, AA, CC et GG. Les k-mer mélangeant purines (A ou T) et pyrimidines (C ou G) sont à l'origine de treillis uniformes. Nous pouvons donc en conclure que c'est probablement l'information des TADs qui est captée par l'extraction de *features* : ces domaines regroupant les gènes actifs simultanément sont connus pour être fortement liés aux isochores, de longues régions de l'ADN, d'au moins 300 kbp, uniformes en termes de *GC content*. On note que certains k-mers de taille plus élevée sont également sélectionnés, toujours sur l'ensemble de la séquence complète. L'ensemble des variables construites par DExTRA est visible dans le graphe d'exploration de la Supp. figure .5 de l'annexe 6.

Ces variables sont par la suite utilisées comme variables explicatives dans l'approche Lasso, Random Forest et Gradient Boosting. Les exemples d'apprentissage sont les paires utilisées lors de l'exploration de DExTRA pour mesurer les AUC, et les exemples de test sont celles n'ayant aucun promoteur en commun avec les paires de *training*. Les deux types de variables explicatives mentionnés

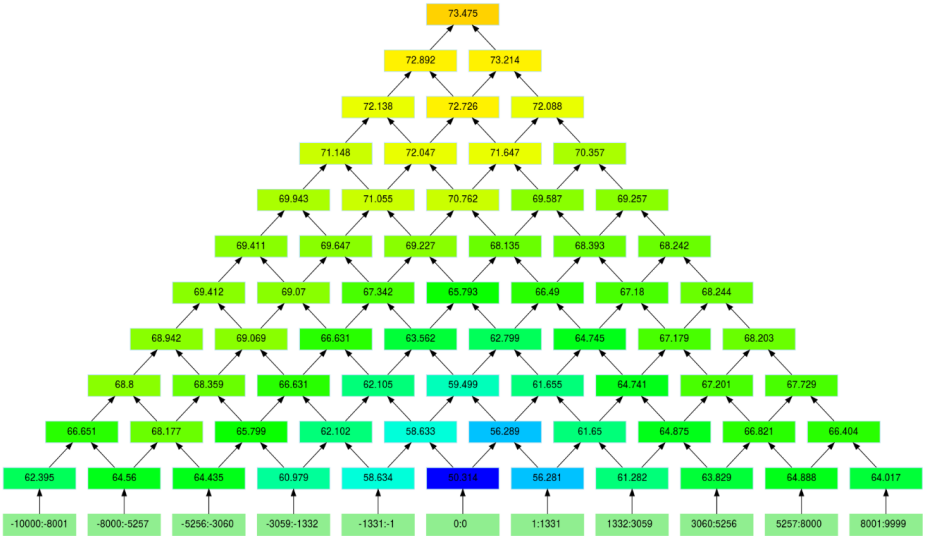


Fig. 14: Treillis global d'AUC pour le dinucléotide AT obtenu lors de l'exploration DExTRA. Les couleurs les plus chaudes indiquent un valeur d'AUC plus forte pour le k-mer sur le noeud considéré.

en figure 10 de la section 2.5 (variables simples et variables d'interaction calculées par la valeur absolue de la différence des fréquences dans les deux séquences) sont testées, et fournissent les résultats de la table 2. Les courbes ROC de chacune des configurations se retrouvent en Suup. figure .7 de l'annexe 7.

	Lasso	Random Forest	Gradient Boosting	Perceptron multicouche
(1)	0.629	0.799	0.774	0.760
(2)	0.737	0.772	0.744	0.743

Table 2: Valeurs d'AUC pour les différents classifieurs et les configurations de variables testées.
(1) : Variables simples, (2) : Variables d'interaction

Les variables sélectionnées par chacun de ces modèles concordent avec les variables montrant la plus forte AUC dans le graphe d'exploration, c'est à dire les k-mer traduisant le *GC content*. La Supp. figure .6 de l'annexe 7 montre l'ordre d'importance des variables pour les *Random Forests* pour chacune des configurations.

Ces deux configurations de variables explicatives éclairent le comportement et la nature des modèles en jeu. En effet, lorsque l'on donne à la régression

logistique les taux de k-mer pour chacun des promoteurs de la paire sans en faire la différence en valeur absolue, les performances sont clairement inférieures à celles des modèles non linéaires. L'AUC est de 0.631 pour la régression, et entre 0.76 et 0.799 pour les autres classifieurs. Lorsque la variable de différence en valeur absolue des taux de k-mers est directement fournie comme dans la seconde configuration, la régression fait tout aussi bien que les autres modèles, montrant ainsi que le terme d'interaction comporte un pouvoir prédictif élevé. Un autre résultat est la capacité des *Random Forests* à combiner les variables de taux de k-mers pour chaque membre de la paire, dans la configuration 1, de manière plus informative que celle donnée par la seconde. L'AUC avec les variables pour chaque promoteur est en effet de 0.799 dans la configuration 1, contre 0.774 dans la seconde. On peut en déduire qu'il existe une fonction d'interaction différente de la valeur absolue de la différence des variables, qui est identifiée par les *Random Forests*, et qui permet une meilleure classification des paires.

4 Perspectives et discussions

Au cours de ce stage, il a été montré que la prédiction de la mise en contact de régions génomiques régulant la transcription est un mécanisme encore complexe à prédire. Afin de construire une approche correcte, il est nécessaire de s'inspirer de la littérature tout en prenant conscience des erreurs de design expérimental que l'on peut y trouver, et de les inclure à une nouvelle approche.

En considérant la région stricte des promoteurs et des enhancers comme en section 3.1, les résultats tendent à prouver que des variables de séquence uniquement ne permettraient pas d'expliquer le repliement chromatinien, tout du moins avec les variables et modèles utilisés ici.

Peu de résultats positifs sont obtenus en se basant sur les séquences des enhancers et des promoteurs, mais des constats émergent lorsque nous élargissons la fenêtre d'étude au voisinage chromosomique de ces régions, et que nous prédisons des contacts entre promoteurs comme présenté en section 3.2.
Des régions sélectionnées à posteriori grâce à une extraction de *features* indiquent finalement que la composition nucléotidique du voisinage des promoteurs a un effet sur leurs interactions plus longues distances. Ces résultats, obtenus sur la problématique promoteur-promoteur, pourraient sans doute s'appliquer à la problématique promoteur-enhancer.

Il est important de noter que les difficultés de prédiction de ces interactions peuvent avoir différentes origines. La précision et la validité des expériences de ChIA-PET ne sont pas parfaitement établies, et certaines expériences indiquent qu'il y aurait un nombre important de faux positifs dans les contacts détectés. Il a été montré que dans certains cas chez la drosophile, l'expression des gènes ne dépend pas de la conformation tri-dimensionnelle de l'ADN [8]. De même, des chercheurs ont observé que chez l'homme, certains enhancers n'ont pas besoin de venir en contact avec un promoteur pour réguler le gène associé. (c'est le cas du gène SOX2 par exemple [1]). Il reste beaucoup de pistes d'améliorations

possibles pour ce projet, notamment quant à la méthode DExTRA. Une autre méthode de construction de treillis serait envisageable. En effet, la structure actuelle se base sur les taux de k-mers entre deux points A et B d'une séquence, mais une variable associée aux bornes A et B pourrait également représenter toute la séquence excluant le segment AB. Ainsi, nous pourrions utiliser des critère d'AUC sur l'amont et l'aval d'une région simultanément. Cela pourrait peut-être permettre d'obtenir des AUCs plus élevées parce que s'affranchissant des zones centrales peu informatives observées.
Une question intéressante serait de distinguer les contacts intra-chromosomiques des contacts inter-chromosomiques dans les expériences. Nous pourrions voir si leurs prédictions se font aussi bien les unes que les autres, ou si les patterns qui les dirigent sont différents.
Les modèles construits pourraient probablement largement bénéficier de l'ajout de variables épigénétiques relatives à la conformation de l'ADN, sa forme, sa méthylation. Cependant, cela sort du cadre de l'hypothèse de recherche que nous cherchons à vérifier. D'une manière générale, et comme cela apparaît très majoritairement dans la littérature, la séquence génétique et son environnement ne sont pas indépendantes. La compréhension actuelle de ces phénomènes reste floue, et suggère une action réciproque de la séquence et de son environnement structurel. De la même manière, l'expression des gènes semble influencée par la séquence, et l'influencer en retour.

Différentes propriétés des modèles de *machine learning* utilisés ont pu être mises en lumière, notamment au niveau de leur capacité à détecter les interactions de variables les plus adéquates. Ici, les *Random Forests* se sont montrés particulièrement adaptés. L'utilisation du *deep learning* ne s'est pas faite à son plein potentiel : d'autres approches basées sur des réseaux de convolution auraient pu être testées également. En effet, ces réseaux de neurones sont capables de construire eux-mêmes des *features* importantes à partir des séquences directement. Cependant, ces *features* sont ensuite difficiles à extraire pour une interprétation biologique, ce qui n'est pas satisfaisant dans notre cas.

Le travail effectué dans ce stage est à ce stade encore très exploratoire, donnant naissance à de nombreuses nouvelles réflexions, tant au niveau des méthodes d'extraction de *features*, que sur les techniques de classification. L'apprentissage statistique, de part sa capacité à identifier des *patterns* parmi d'incroyables quantités de données, permet indéniablement d'éclairer les mécanismes moléculaires permettant notre développement, et notre adaptation au monde extérieur.

References

1. Alexander, J.M., Guan, J., Li, B., Maliskova, L., Song, M., Shen, Y., Huang, B., Lomvardas, S., Weiner, O.D.: Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. eLife 8, e41769 (May 2019), https://doi.org/10.7554/eLife.41769
2. Bessière, C., Taha, M., Petitprez, F., Vandel, J., Marin, J.M., Bréhélin, L., Lèbre, S., Lecellier, C.H.: Probing instructions for expression regulation in gene nucleotide compositions. PLoS Computational Biology 14(1) (Jan 2018), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5766238/
3. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
4. Cao, Q., Anyansi, C., Hu, X., Xu, L., Xiong, L., Tang, W., Mok, M.T.S., Cheng, C., Fan, X., Gerstein, M., Cheng, A.S.L., Yip, K.Y.: Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nature Genetics 49(10), 1428–1436 (Sep 2017), http://www.nature.com/doifinder/10.1038/ng.3950
5. Friedman, J.H., Hastie, T., Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33(1), 1–22 (Feb 2010), https://www.jstatsoft.org/index.php/jss/article/view/v033i01
6. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Patwardhan, T.A., Nguyen, T.H., Kane, M., Doughty, B., Perez, E.M., Durand, N.C., Stamenova, E.K., Lieberman Aiden, E., Lander, E.S., Engreitz, J.M.: Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations. bioRxiv (Jan 2019), http://biorxiv.org/lookup/doi/10.1101/529990
7. Furlong, E.E.M., Levine, M.: Developmental enhancers and chromosome topology. Science 361(6409), 1341–1345 (Sep 2018), https://science.sciencemag.org/content/361/6409/1341
8. Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R.R., Korbel, J.O., Furlong, E.E.M.: Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nature Genetics p. 1 (Jul 2019), https://www.nature.com/articles/s41588-019-0462-3
9. Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27(2), 83–85 (2005)
10. Lenhard, B., Sandelin, A., Carninci, P.: Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature Reviews. Genetics 13(4), 233–245 (Mar 2012)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
12. Robson, M.I., Ringel, A.R., Mundlos, S.: Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3d. Molecular Cell 74(6), 1110–1122 (Jun 2019), http://www.sciencedirect.com/science/article/pii/S1097276519304046
13. Roy, S., Siahpirani, A.F., Chasman, D., Knaack, S., Ay, F., Stewart, R., Wilson, M., Sridharan, R.: A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Research 43(18), 8694–8712 (Oct 2015), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4605315/

14. Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W., Lenhard, B.: JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research 32(suppl_1), D91–D94 (Jan 2004), https://academic.oup.com/nar/article/32/suppl_1/D91/2505159
15. Singh, S., Yang, Y., Poczos, B., Ma, J.: Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks. bioRxiv (Feb 2018), http://biorxiv.org/lookup/doi/10.1101/085241
16. The FANTOM Consortium, Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., Jørgensen, M., Andersen, P.R., Bertin, N., Rackham, O., Burroughs, A.M., Baillie, J.K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C.J., Meehan, T.F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C.O., Heutink, P., Hume, D.A., Jensen, T.H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A.R.R., Carninci, P., Rehli, M., Sandelin, A.: An atlas of active enhancers across human cell types and tissues. Nature 507(7493), 455–461 (Mar 2014), http://www.nature.com/articles/nature12787
17. The FANTOM Consortium and the RIKEN PMI and Clst (dgt), Forrest, A.R.R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., Andersson, R., Mungall, C.J., Meehan, T.F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y.A., Plessy, C., Vitezic, M., Severin, J., Semple, C.A., Ishizu, Y., Young, R.S., Francescatto, M., Alam, I., Albanese, D., Altschuler, G.M., Arakawa, T., Archer, J.A.C., Arner, P., Babina, M., Rennie, S., Balwierz, P.J., Beckhouse, A.G., Pradhan-Bhatt, S., Blake, J.A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Maxwell Burroughs, A., Califano, A., Cannistraci, C.V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H.C., Dalla, E., Davis, C.A., Detmar, M., Diehl, A.D., Dohi, T., Drabløs, F., Edge, A.S.B., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M.C., Faulkner, G.J., Favorov, A.V., Fisher, M.E., Frith, M.C., Fujita, R., Fukuda, S., Furlanello, C., Furuno, M., Furusawa, J.i., Geijtenbeek, T.B., Gibson, A.P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T.J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K.J., Ho Sui, S.J., Hofmann, O.M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B.R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A.S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y.I., Kawashima, T., Kempfle, J.S., Kenna, T.J., Kere, J., Khachigian, L.M., Kitamura, T., Peter Klinken, S., Knox, A.J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A.T., Laros, J.F.J., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-sim, A., Manabe, R.i., Mar, J.C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D.A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C.L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Nozaki, T., Ogishima, S., Ohkura, N., Ohmiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D.A., Pain, A., Passier, R., Patrikakis, M., Persson, H.,

Piazza, S., Prendergast, J.G.D., Rackham, O.J.L., Ramilowski, J.A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M.B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Satoh, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E.A., Schulze-Tanzil, G.G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J.W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R.K., 't Hoen, P.A.C., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyoda, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L.M., Verardo, R., Vijayan, D., Vorontsov, I.E., Wasserman, W.W., Watanabe, S., Wells, C.A., Winteringham, L.N., Wolvetang, E., Wood, E.J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S.E., Zhang, P.G., Zhao, X., Zucchelli, S., Summers, K.M., Suzuki, H., Daub, C.O., Kawai, J., Heutink, P., Hide, W., Freeman, T.C., Lenhard, B., Bajic, V.B., Taylor, M.S., Makeev, V.J., Sandelin, A., Hume, D.A., Carninci, P., Hayashizaki, Y.: A promoter-level mammalian expression atlas. Nature 507(7493), 462–470 (Mar 2014), https://www.nature.com/articles/nature13182

18. Vandel, J., Cassan, O., Lèbre, S., Lecellier, C.H., Bréhélin, L.: Probing transcription factor combinatorics in different promoter classes and in enhancers. BMC Genomics 20(1), 103 (Feb 2019), https://doi.org/10.1186/s12864-018-5408-0
19. Whalen, S., Truty, R.M., Pollard, K.S.: Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nature Genetics 48(5), 488–496 (May 2016), http://www.nature.com/articles/ng.3539
20. Xi, W., Beer, M.A.: Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. PLOS Computational Biology 14(12), e1006625 (Dec 2018), https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006625
21. Yang, Y., Zhang, R., Singh, S., Ma, J.: Exploiting sequence-based features for predicting enhancer–promoter interactions. Bioinformatics 33(14), i252–i260 (2017)

Remerciements

Il me tient à coeur de commencer en remerciant toutes les personnes ayant fait de ce stage une expérience riche.

Tout d'abord, je remercie sincèrement Laurent Bréhélin pour la qualité de son encadrement et sa grande disponibilité. Il s'est montré d'une aide incontestable pour guider notre projet à chacune de ses étapes, de la construction de la problématique jusqu'à la rédaction de ce rapport. Deux stages à ses côtés ont été d'idéales premières expériences dans le monde de la recherche.

Je souhaite également remercier Charles-Henri Lecellier pour les précieux conseils offerts sur le cadre biologique du stage, ainsi que pour son intérêt communicatif envers les subtilités de la génétique.

Merci à Sophie Lèbre, dont les apports en statistique ont éclairé plus d'un aspect de mon travail. Je me réjouis de notre collaboration plus étroite dans un futur proche.

Je suis très reconnaissante à Christophe Menichelli pour avoir partagé avec moi son excellent travail de thèse, et m'avoir permis de prendre part à son extension.

Enfin, je remercie Raphaël Romero, dont le travail sur des thématiques proches a permis un échange fructueux d'informations.

Les fréquentes réunions avec ces personnes ont permis un suivi précis, personnalisé, aussi apprécié qu'utile au projet. Je leur souhaite à tous une très bonne continuation au sein de l'équipe, dans laquelle évoluer pendant 5 mois fut un réel plaisir.

Enfin, je remercie Sergio Peignier, m'ayant encadrée du côté de l'INSA, et Sam Meyer qui a accepté d'être le second membre de mon jury.

"Nature composes some of her loveliest poems for the microscope and the telescope." Theodore Roszak, Where the Wasteland Ends, 1972

Annexes

1. Valeurs de scores pour de la méthode TargetFinder après correction du biais

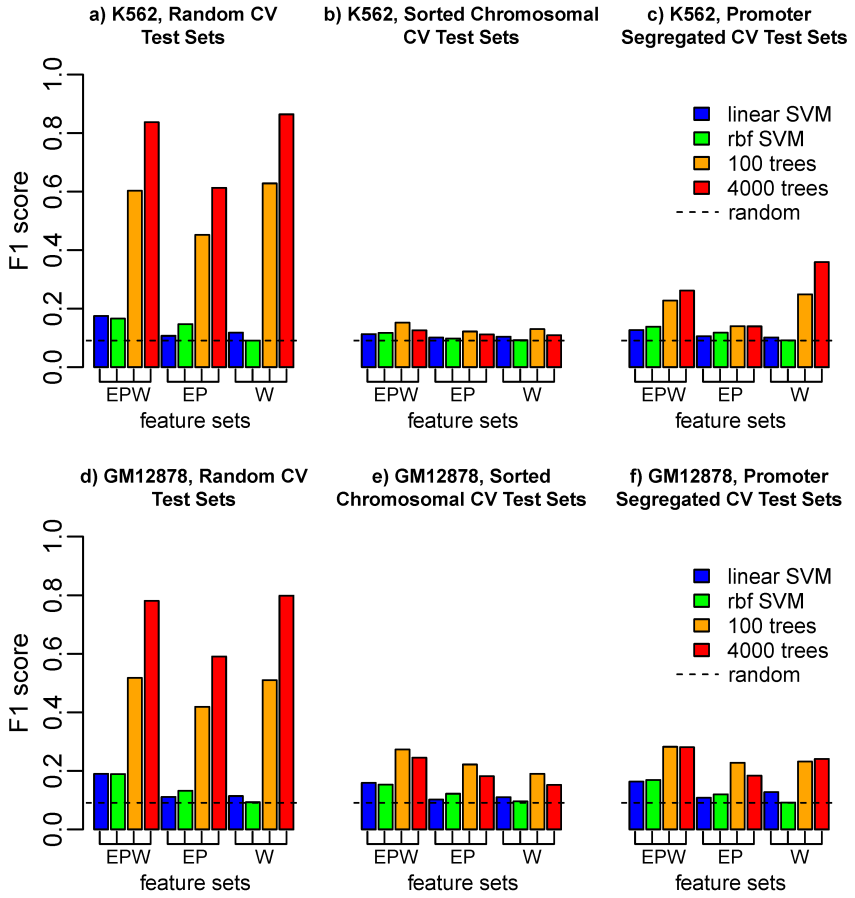


Fig. .1: Différence entre les résultats publiés dans TargetFinder à gauche, et la méthode sur un jeu de données de Test issu d'un tri par ordre chromosomique au centre. Nous constatons que les prédictions correctement évaluées ne valent pas plus que le hasard.

2. Calcul des variables de fréquences nucléotidiques

La molécule d'ADN est une molécule double brin. La séquence d'un brin est dite complémentaire de l'autre brin : on passe de l'une à l'autre en remplaçant tous les A par des T, les T par des A, les C par des G, les G par des C. Le brin complémentaire est lu dans le sens inverse du brin direct.

Les 16 dinucléotides qu'il est possible de former avec 4 bases n'apparaissent pas tous directement dans les variables utilisées pour caractériser nos séquences, qui sont au nombre de 10. En effet, comme nous considérons une séquence en double brin, compter la fréquence d'un dinucléotide sur les deux brins revient à compter la fréquence de ce dinuclétodide ainsi que celle de son reverse-complément sur un seul brin. Par exemple le taux de AA correspond à la fréquence des AA sur un brin, à laquelle on ajoute la fréquence des TT sur ce même brin. Ainsi, il n'est pas utile de considérer à la fois le taux de AA et de TT dans le modèle, ces informations étant contenues dans une seule de ces deux variables. Un cas particulier est à prendre en compte avec les k-mer pairs. En effet, dans le cas où un k-mer est également son *reverse-complement* (e.g AT, CG, ATAT, CCGG...), sa fréquence en double brin est directement donnée par sa fréquence en simple brin.

Les variables utilisées pour construire les modèles sont donc 6 dinucléotides comprenant aussi leur reverse-complément dans le cas normal, et les 4 dinucléotides considérés comme des cas particuliers (AT, TA, CG et GC). A cela nous rajoutons le AT content et le GC content, qui sont eux des taux de nucléotides, et non des dinucléotides, pour arriver à un total de 12 variables nucléotidiques.

3. Pseudocode de l'exploration DExTER

Algorithm 1 Procédure d'exploration de DExTER

Entrée:
Y : vecteur à prédire
Séquences au format fasta
Initialisation :
$L \leftarrow$ la liste des dinucléotides sur l'ensemble de la séquence
$L_{val} \leftarrow \emptyset$

tant que $L \neq \emptyset$ **faire**
  $D_{k,r} \leftarrow$ variable prise dans L
  si $\text{Cor}(D_{k,r}, Y) > \text{Cor}(\text{Parent}(D_{k,r}), Y)$ et $\text{Cor}(D_{k,r}, Y) > \text{Cor}(D_{k-1,r}, Y)$ **alors**
    $L_{val} \leftarrow L_{val} + D_{k,r}$
    si k est égal au k-mer de $\text{Parent}(D_{k,r})$ **alors**
      {La variable $D_{k,r}$ est issue d'une phase de segmentation, ses enfants seront les k+1-mers sur la même région r}
      Children $\leftarrow$ 8 nouvelles variables $D_{k+1,r}$ enfants de $D_{k,r}$ créées en ajoutant un nucléotide en fin ou en début du k-mer
      $L \leftarrow L+$ Children
    sinon
      {la variable $D_{k,r}$ est issue d'une phase d'exploration, on a ajouté un nucléotide à son parent. Ses enfants seront les $D_{k,r'}$, variables aux régions optimales pour ce k-mer}
      R $\leftarrow$ Liste des régions triée par ordre décroissant de corrélation pour le k-mer k
      Children $\leftarrow \emptyset$
      pour r' in R **faire**
        si r' n'intersecte pas une région de Children et $\text{Cor}(D_{k,r'}, Y) \geq \text{Cor}(D_{k,r}, Y)$ **alors**
          Children $\leftarrow$ Children + $D_{k,r'}$
        fin si
      fin pour
      $L \leftarrow L+$ Children
    fin si
  fin si
fin tant que
return L_{val}

4. Prédiction des contacts de ChIA-PET enhancers-promoteurs dans K562 : ajout des variables d'interaction

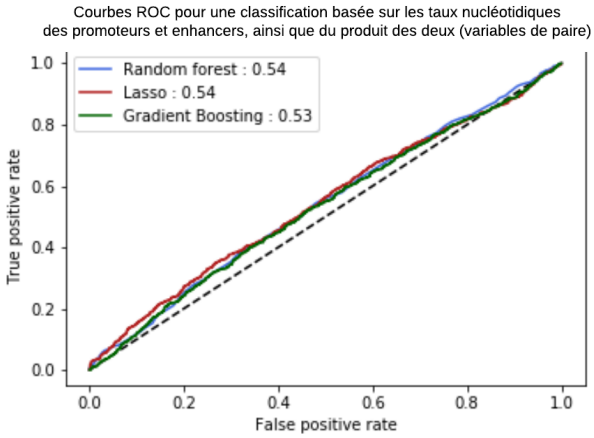


Fig. .2: Courbes ROC et AUCs pour des modèles entraînés sur les variables de taux de nucléotides des enhancers et des promoteurs sur les données de ChIA-PET

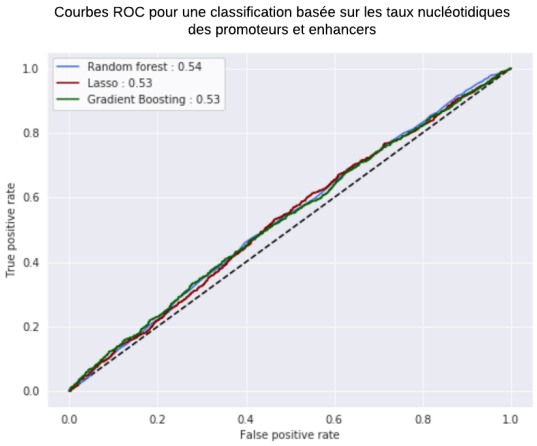


Fig. .3: Courbes ROC et AUCs pour des modèles entraînés sur les variables de taux de nucléotides des enhancers et des promoteurs, ainsi que sur les variables d'interactions : le produit du taux de chaque membre de la paire, sur les données de ChIA-PET.

5. Sélection de variable sur jeu de données fictif

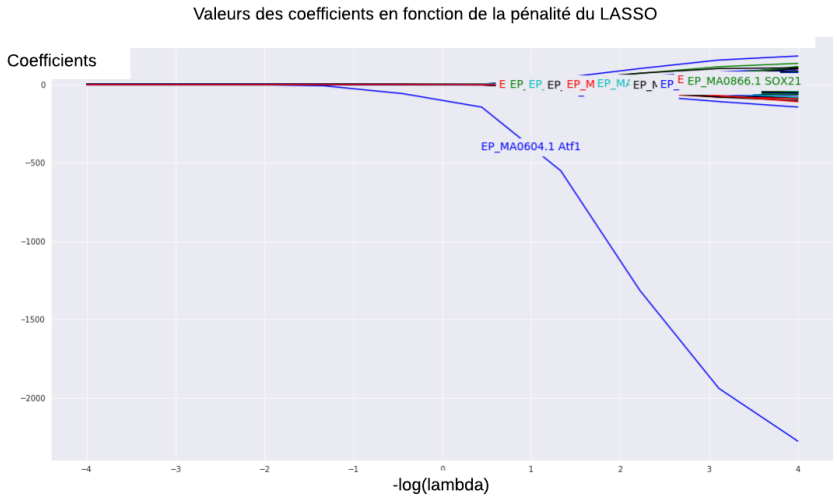


Fig. .4: Graphe de sélection de variables lors de la régularisation : lorsque la pénalité est diminuée, la valeur du coefficient relatif à ATF1 est bien celle qui prend en premier une valeur non nulle.

6. Graphe d'exploration DExTRA

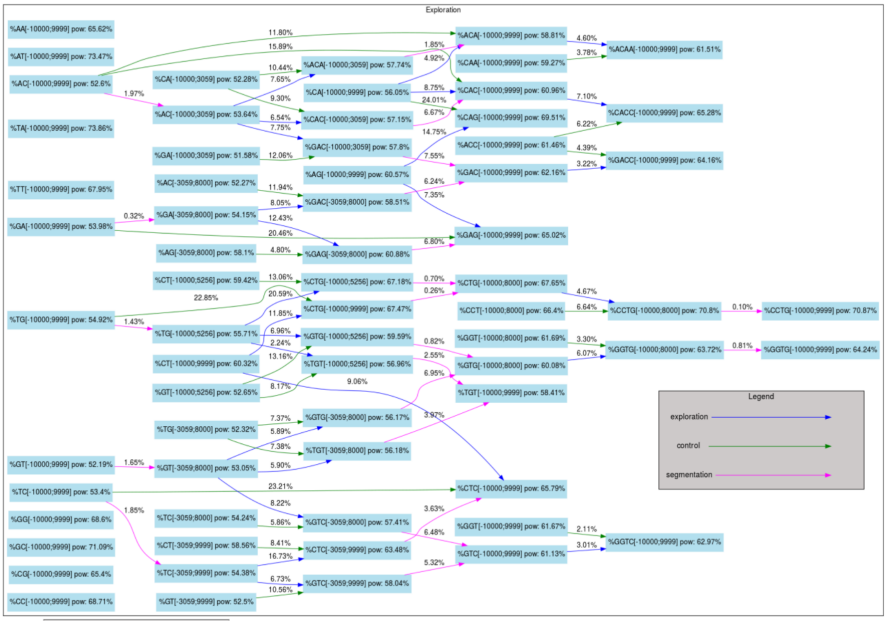
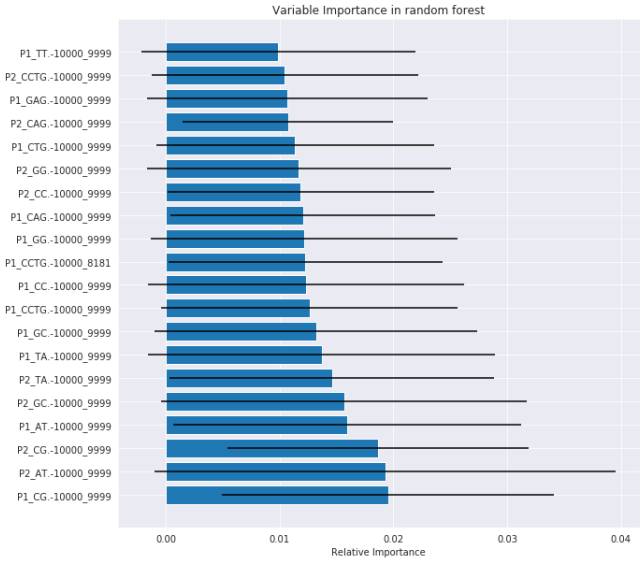
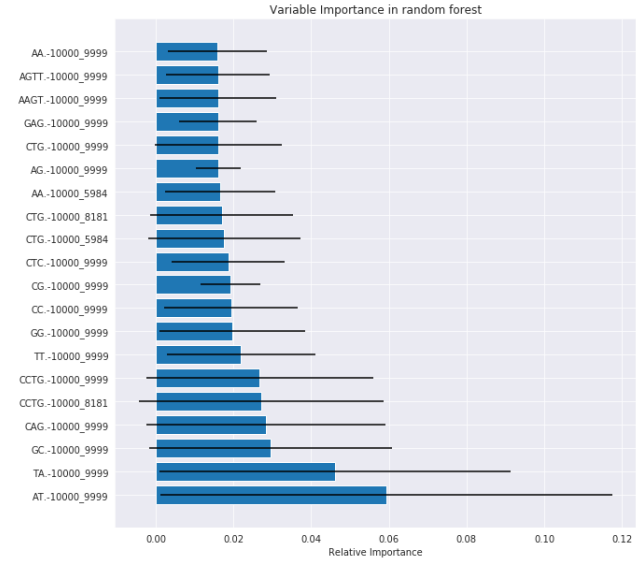


Fig. .5: Graphe d'exploration issu de la procédure DExTRA. Chaque noeud représente un domaine exploré par ajout de nucléotide (flèches bleues) ou refractionné, segmenté (flèches roses). Les flèches vertes représentent les conditions de contrôle mentionnées dans le pseudo-code, au moment de décider de la validation d'un domaine.

7. Importance des variables dans les Random Forests pour les deux configurations de variables issues de l'extraction de *features*, et les courbes ROC correspondantes pour tous les modèles

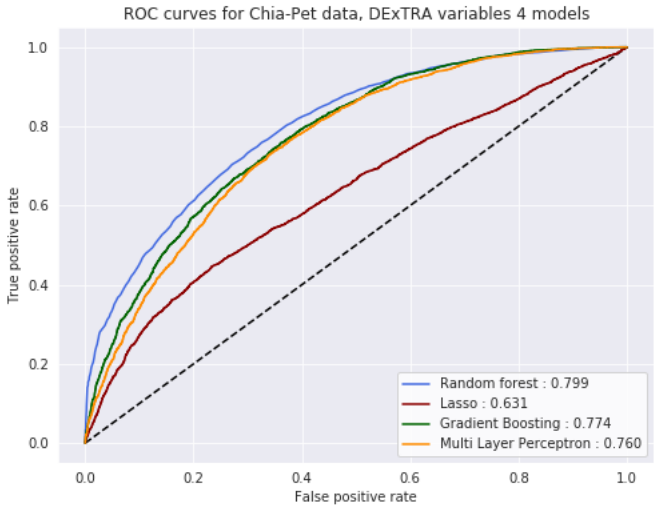


[1]

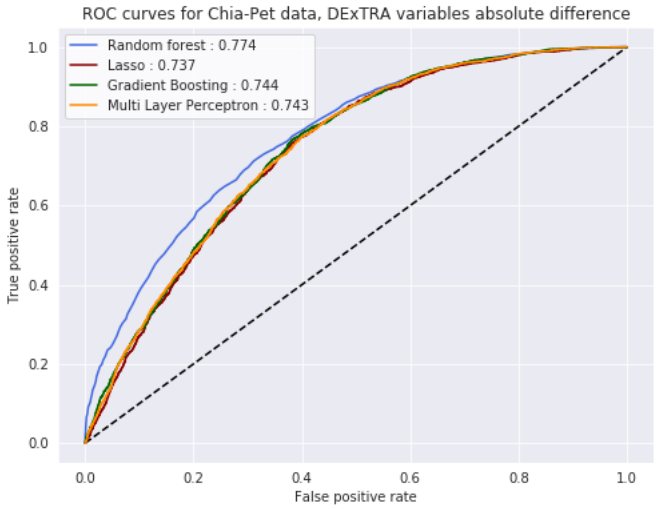


[2]

Fig. .6: 10 premières variables classées par importance par les *Random Forests* entraînés.
(1) : Taux pour chaque promoteur P1 et P2 de la paire. Nous voyons que lorsque la variable d'un promoteur P1 est sélectionnée, cette même variable chez P2 apparaît non loin dans le classement, confirmant leur combinaison par le modèle.
(2) : Différence en valeur absolue des taux de P1 et P2.



[1]



[2]

Fig. .7: Courbes ROC pour les configurations 1 et 2
(1) : Taux pour chaque promoteur de la paire
(2) : Différence en valeur absolue des deux taux