École nationale de la statistique et de l'analyse de l'information



Groupe 17

Rapport de projet statistique

OMICS : L'ADN peut-il permettre de récupérer une erreur de laborantin ?

Élèves :
Corentin CALAS
Jules LÉNÉ
Styve LAM

Tuteur :
Guillaume DESACHY

Remerciements

Nous tenons à remercier dans un premier temps toute l'équipe pédagogique de l'ENSAI pour sa disponibilité, en particulier Madame Brigitte Gelein qui a su nous assister et répondre à nos nombreuses questions.

Nous souhaitons également remercier très chaleureusement notre tuteur, Monsieur Guillaume Desachy qui lors de ce projet a assuré un suivi régulier et précieux. Très disponible et impliqué, ses conseils nous ont souvent éclairés et offert de nouvelles perspectives de recherches.

Dans de telles conditions, cela a été un réel plaisir de travailler sur ce projet.

Table des matières

1 Introduction **8**
1.1 Contexte 8
1.2 Problématique et démarche 9

2 Première prise en main des données **10**
2.1 Brève introduction aux concepts génétiques 10
2.2 Des données abondantes 12

3 Préparer les données à la prédiction : un contrôle-qualité nécessaire **14**
3.1 Quelques données aberrantes 14
3.2 Des valeurs manquantes : le sexe est-il explicatif ? 14
3.3 Notre prédiction va-t-elle être perturbée par des individus atypiques ? 15
3.4 Recherche d'un seuil 17
3.5 Une dernière vérification 18

4 Une modélisation adaptée à un grand nombre de données **19**
4.1 Les limites du modèle linéaire simple 19
4.1.1 Réduction de l'amplitude des variables 20
4.1.2 Réduction du nombre de variables 20
4.1.3 Elastic Net : Un modèle stable et parcimonieux 20
4.1.4 Comment mesurer la qualité de notre prédiction ? 22
4.2 Premiers modèles et comparaisons 23
4.2.1 Premier modèle : l'elastic net appliqué aux données seuillées à 22 23
4.2.2 Deuxième modèle : l'elastic net appliqué aux données seuillées à 27 . . . 24
4.2.3 Comparaisons de M1 et de M2 24
4.3 Amélioration de la sélection et réduction du temps de calcul : tests multiples sur les variables 24

5 Est-il possible d'écarter un des deux types de données ? **26**

6 Retour sur le modèle final **29**
6.1 Quelles sont les variables les plus importantes ? 29
6.2 Où sont situées les variations génétiques ? 32
6.3 Quelles variations génétiques pour quel cancer ? 34

7 Dernière étape : notre modèle à l'épreuve de l'échantillon de validation **35**

8 Conclusion **36**

9 Annexe **38**

Table des figures

1 La molécule d'ADN, en forme de double-hélice 10
2 Phénomène du SNP 11
3 Phénomène du CNV 11
4 Comparaison de la répartition du sexe en fonction du type de cancer, CCLE, 2012 14
5 Analyse de la lignée cellulaire 115 sur différents plans factoriels, CCLE, 2012. . . 16
6 Nombre de variations génétiques en fonction du nombre de lignées cellulaires, CCLE, 2012 17
7 Résumé de notre processus de contrôle qualité 18
8 Grille initiale des hyperparamètres α, λ du modèle elastic net appliqué aux variations génétiques présentes chez au moins 22 lignées cellulaires, CCLE 2012 . . . 23
9 Grille agrandie des hyperparamètres α, λ du modèle elastic net appliqué aux variations génétiques présentes chez au moins 22 lignées cellulaires, CCLE 2012 . 24
10 Processus de sélection successifs des variables du quatrième modèle 25
11 Comparaisons des ACP d'ensemble des données (en haut) et de la table CN seule (en bas) 26
12 Comparaisons de la précision entre les modèles complet, CN et mutations, CCLE 2012 27
13 Chemin de régularisation du modèle prédictif M4 finalement retenu, CCLE, 2012 29
14 Les 20 variables dont les coefficients sont les plus importants dans le modèle prédictif M4 finalement retenu, CCLE, 2012 30
15 Types et nombre de variations génétiques par type de cancer, CCLE, 2012 . . . 34
16 Capture d'écran de notre application Shiny globale : les différents onglets pour chaque étape du projet 38
17 Options de notre application Shiny "ACP interactives" 39
18 Capture d'écran des outils à disposition pour notre application Shiny d'ACP interactives 40
19 Capture d'écran de notre application Shiny d'aide à la détermination de seuil . . 41
20 Agrandissement de notre application Shiny d'aide à la détermination de seuil . . 41
21 Capture d'écran de notre application d'aide à l'optimisation des hyperparamètres 42
22 Capture d'écran de notre application de comparaisons de modèles prédictifs . . . 43

Table des tableaux

1 Extrait de la table CN, CCLE, 2012 12
2 Extrait de la table *mutations* qui liste l'ensemble des SNP étudiés chez les lignées cellulaires, CCLE, 2012 13
3 Tableau de contingence du sexe et du type de cancer, CCLE, 2012 15
4 Récapitulatif et comparaison de performances des modèles prédictifs, CCLE, 2012 25
5 Comparaisons du nombre de variables d'entrée et gardées par le modèle entre les modèles complet, CN et mutations, après une préselection des variables, CCLE, 2012 28
6 Valeurs des 7 coefficients les plus importants dans le modèle prédictif M4 finalement retenu, CCLE, 2012 31
7 Répartition des cancers selon la délétion des gènes TARP et de ADAM6, CCLE, 2012 31
8 Répartition des cancers selon la présence de SNP aux locus KEAP1.MC et RB1 32
9 Répartition des différentes variations génétiques retenues par le modèle M4 sur les différents chromosomes, CCLE 2012 et UCSC 2018 33
10 Matrice de confusion sur l'échantillon de validation. 35

Liste des abréviations

ACP (p. 15) Analyse en Composantes Principales.

ADN (p. 10) Acide Désoxyribonucléique, molécule en double-hélice située dans les cellules vivantes, portant les informations génétiques, composée d'un enchaînement de nucléotides.

AUROC (p. 22) Area Under ROC curve, *i.e.* aire sous la courbe ROC. C'est une mesure de la performance d'un classifieur binaire.

CCLE (p. 12) Cancer Cell Line Encyclopedia. Projet de plusieurs laboratoires et a pour ambition de recenser le profil génomique de l'ensemble des cancers humains.

CNV (p. 10) Copy Number Variation. C'est une variation génétique qui peut être une duplication ou une délétion d'une portion d'ADN.

LASSO (p. 22) Least Absolute Shrinkage and Selection Operator. C'est une méthode de contraction des coefficients d'une régression.

Omics (p. 7) Fait référence aux technologies relatives à la génomique, la protéomique, métabolomique... Ces domaines, pour certains très récents, étudie le fonctionnement des systèmes organiques à tous les niveaux biologiques.

ROC (p. 22) Receiver operating characteristic, *i.e.* Fonction d'efficacité du récepteur. Renseigne sur la spécificité et la sensibilité d'un classifieur binaire.

SNP (p. 10) Single Nucleotide Polymorphism, *i.e.* Polymorphisme Nucléotidique Simple. C'est une variation génétique située sur un nucléotide.

Abstract

During the last few years, we have witnessed the rising of precision medecine. Until now, most medical treatments have been conceived considering the average patient. This has led to very succesful results for some patients but also more nuanced ones for others. On the other hand, precision medecine is a personalized approach which proposes the customization of healthcare, medical decision and treatments by taking into account people's genes and environments. Precision medecine has already proved its efficiency with innovative treatments adapted to specific characteristics such as the genetic profile of the patient [1]. Therefore, fields of study such as genomics, proteomics or metabolomics (which are refered to by the term "Omics") are essential to the improvement of treating diseases such as cancer.

Our statistical project falls within this context. By studying a database of 361 cell lines, with either lung cancer or lymphatic cancer, and for each, 48 269 genetic variations, our goal was to predict cancer type based on the cell lines' genetic profile. There were two types of genetic variation at our disposal : copy number variation, which can be a DNA segment duplication or deletion, and Single Nucleotide Polymorphism (SNP), a genetic variation in a single nucleotide[2].

In order to do that, we had access to 80% of the cell lines' cancer type. By analyzing this training data set, we have managed to select the relevant variables among all the genetic variations. Finally, our predictive model used 69 variables, and achieved a very satisfying accuracy of 92% on the validation set.

The first conclusion of our study is that, out of all the variables, only a small proportion are relevant to characterize cancer type (our final model uses only 0.14% of all the variables available). Moreover, selecting such a limited number of variables allowed us to analyze them in detail and find out their role in cancer type explanation.

We have also identified the most important genes in cancer type explanation[3] and listed all the relevant genes' positions on the different chromosomes. We found out that 70% of the explanatory variables are located on the 5th and the 18th chromosome. In addition, 60% of the genetic variation identified by our model as relevant are gene duplications[4].

Finally, an additional goal of our project was to study the possibility to construct a predictive model with information on only one type of genetic variation (copy number variation or SNP), which would lead to reduce costs for biologists. Unfortunately, our conclusion is that a strong and robust predictive model for cancer type needs information on both copy number variations and SNPs, neither of them can be ignored or the precision severely decreases.

[1] `https://www.ncbi.nlm.nih.gov/pubmed/16618717]`

[2] Nucleotides form the basic structural unit of DNA.

[3] In a case of predicting lung cancer against lymphatic cancer.

[4] 33% are gene deletions, 7% are "SNPs".

1 Introduction

1.1 Contexte

Le cancer est la première cause de mortalité en France (149 500 en 2015) et ne cesse de gagner du terrain (385 000 nouveaux cas la même année). La lutte contre le cancer est au centre des problématiques de santé publique et mobilise les médecins et chercheurs du monde entier.

Les progrès de la recherche ont permis de mieux comprendre les mécanismes biologiques à l'origine du développement et de la progression des cancers. On sait ainsi aujourd'hui que chaque tumeur est différente et possède des caractéristiques qui lui sont propres. Une meilleure compréhension de ces dernières permet de mettre en œuvre de nouveaux traitements qui ne reposent plus seulement sur la localisation de la tumeur, mais qui sont adaptés à ses caractéristiques moléculaires. C'est ce que l'on appelle la médecine de précision, ou encore médecine personnalisée, qui propose un traitement adapté et personnalisé au patient basé sur l'étude de son profil génétique, de son environnement ou encore de son mode de vie.

De plus, on assiste durant ces dernières années à des évolutions technologiques majeures. L'évolution soutenue des capacités de stockage de données et des puissances de calcul ont permis le développement d'outils d'analyse de l'ADN de plus en plus performants. En 1990 débute le *projet génome humain* dont la mission était d'établir le séquençage complet de l'ADN du génome humain. Depuis, il devient progressivement possible d'établir un profil de plus en plus complet de chaque tumeur, et ce, pour un nombre important de patients. On désigne par « omics » cet ensemble de nouvelles technologies faisant référence à « genomics, transcriptomics, proteomics, metabonomics... ». Les « omics » offrent des clefs de compréhension précieuses sur de nombreuses maladies et ouvrent une nouvelle ère de défi à la génomique, marquée par des nouveaux outils de diagnostic et l'identification de nouveaux traitements et médicaments.

C'est dans ce cadre que se situe notre projet statistique. Nous avons à disposition les profils « omics » de nombreuses lignées cellulaires et souhaitons les caractériser. Une lignée cellulaire est un amas stable de cellules homogène de cellules capables disposant d'une capacité illimitée de division. En général, cela est dû au fait qu'elles sont cancéreuses. C'est le cas ici : les lignées cellulaires étudiées sont soit atteintes d'un cancer pulmonaire, soit d'un cancer lymphatique. Le but de notre projet a donc été d'identifier les caractéristiques génétiques associées à ces deux types de cancer et de prédire le type de cancer à partir du profil omic.

1.2 Problématique et démarche

Nous disposons du profil génétique de 361 lignées cellulaires. Ces lignées cellulaires sont toutes atteintes d'un cancer, soit lymphatique, soit pulmonaire. Pour 80% d'entre elles, le type de leur cancer est connu ; l'objectif de notre projet est de prédire celui des 20% restant à partir de leur profil génétique.

Faisant face à un problème de prédiction, nous avons établi en premier lieu un contrôle-qualité afin d'écarter les données qui diminueraient la qualité de notre prédiction (individus aberrants, variables non explicatives...). Suite à cela, nous avons procédé à plusieurs modélisations que nous avons optimisées afin d'obtenir la meilleure qualité de prédiction. Nous avons mesuré la qualité de nos modèles par validation-croisée sur un échantillon d'entraînement, puis par validation sur un échantillon de test (les 20% dont le type de cancer est inconnu). Enfin, disposant d'informations de deux types de variations génétiques, nous avons étudié la possibilité de s'affranchir des informations d'une de ces deux variations, tout en gardant une bonne qualité de prédiction.

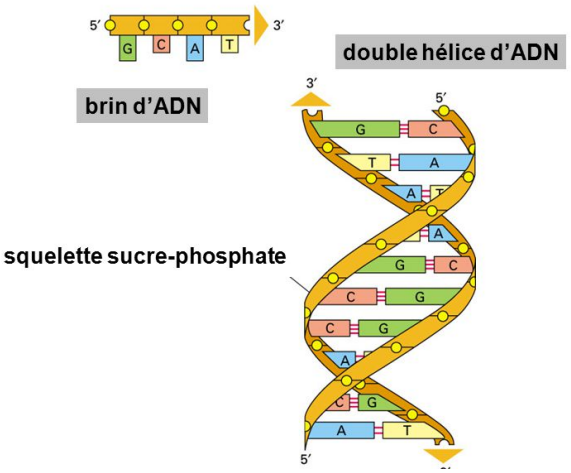

Figure 1: La molécule d'ADN, en forme de double-hélice

2 Première prise en main des données

2.1 Brève introduction aux concepts génétiques

Chez tous les êtres vivants, chaque cellule contient de l'ADN (*acide désoxyribonucléique*) qui est le support de l'information génétique. L'ADN d'une cellule est réparti en plusieurs chromosomes ; chez l'Homme, il y a 23 paires de chromosomes.

La molécule d'ADN est formée de deux brins complémentaires composés chacun d'un enchaînement de nucléotides. Un nucléotide est composé d'un groupement phosphate, d'un sucre (le désoxyribose) et d'une base azotée. Chaque nucléotide peut être identifié par sa base azotée. Il y a quatre types de bases azotées : adénine, cytosine, guanine et thymine, et donc autant de types de nucléotides. De plus, ces bases sont complémentaires deux à deux, l'adénine avec la thymine, la cytosine avec la guanine. Ainsi, les deux brins de l'ADN qui sont complémentaires obéissent à ces associations : l'adénine d'un brin est toujours en face d'une thymine de l'autre brin, de même pour la cytosine et la guanine. Par exemple, la séquence ATCG d'un brin sera liée par des liaisons hydrogène à la séquence TAGC du brin homologue.

Le cycle de vie d'une cellule est marqué par la division cellulaire qui, à partir d'une cellule mère, donne deux cellules filles identiques.

Lors de cette division cellulaire, des variations peuvent survenir, provoquant ainsi le développement de cellules cancéreuses. Parmi ces variations, on en dénote deux grands types : les polymorphismes nucléotidiques simples, souvent abrégé SNP (*Single Nucleotide Polymorphism*) et les variations du nombre de copies d'ADN, souvent abrégé CNV (*Copy Number Variation*).

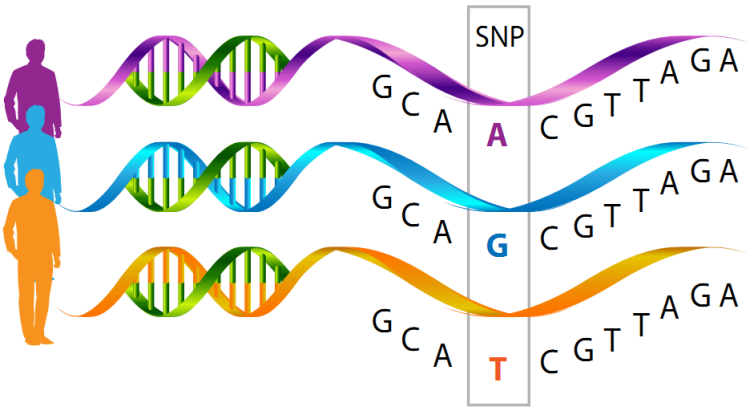

Figure 2: Phénomène du SNP

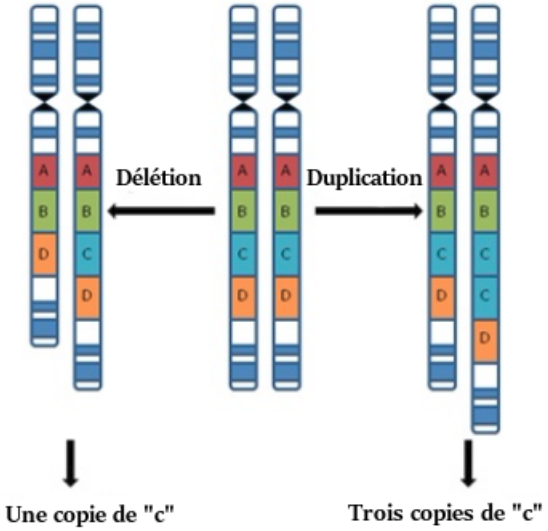

Figure 3: Phénomène du CNV

Les SNP sont une variation d'un nucléotide de l'ADN à un locus[5] donné présente dans plus de 1% de la population.

Les variations du nombre de copies d'ADN, quant à elles, peuvent soit être une délétion soit une duplication d'un segment de l'ADN. On parle alors de sous-expression ou de sur-expression d'un gêne.

On peut voir par exemple sur l'illustration ci-dessus que la portion de gène C peut soit disparaître sur un des deux chromosomes (à gauche) ou bien se dédoubler (à droite) lors de la transmission du patrimoine génétique.

Notre travail a été, à partir des informations de ces deux grands types de variations génétiques sur une lignée cellulaire (présence de CNV ou de SNP), de prédire le type de son cancer (pulmonaire ou lymphatique).

[5]Emplacement dans l'ADN.

2.2 Des données abondantes

Nos données proviennent du projet Cancer Cell Line Encyclopedia (CCLE[6]) dont le but est de rassembler et d'analyser des données cellulaires afin de lutter contre le cancer. Les données du CCLE sont disponibles en libre accès.

Le CCLE naît de l'initiative conjointe des laboratoires *Broad Institute*, *Novartis Institutes for Biomedical Research* et du *Genomics Institute of the Novartis Research Foundation*. Sa première publication en 2012 regroupait les données de 947 lignées cellulaires humaines, qui sont cultivées et qui se développent indéfiniment en laboratoire. Ayant pour ambition de représenter la diversité des cancers humains, le CCLE regroupe des données aussi bien de cancers répandus que de cancers rares. La publication de ces données ainsi que son exploitation a mené à d'importants progrès dans la recherche scientifique et médicale.

Dans la suite du rapport, nous utiliserons les abréviations présentées dans la section précédente: **CN** pour les variations de copie d'ADN, **SNP** pour les polymorphismes nucléotidiques simples.

Nous disposions ainsi de trois tables de données :

- *Info* qui contient la liste des lignées cellulaires, leur sexe, ainsi que le cancer qui leur est associé.
- *Mutations* qui contient la liste des lignées cellulaires, ainsi que l'ensemble de leurs SNP.
- *CN* qui contient la liste des lignées cellulaires, ainsi que l'ensemble de leurs CN.

Les tables Mutations et CN comportent respectivement 1638 et 46 630 variables. Outre le numéro de la lignée cellulaire, toutes les variables sont binaires et valent 1 lorsque la variation génétique est présente, 0 dans le cas contraire.

Pour la table CN, on dispose de deux variables par gène, une pour la délétion et une pour la duplication de celui-ci. Ainsi, on dispose d'environ 46 000 variables pour 23 000 locus.

Lignée cellulaire	A1BG-AS1-DEL	A1BG-DEL	A2LD1-DEL	...	A1BG-AS1-AMP	...
1	1	1	0	...	0	...
2	0	0	0	...	1	...
...	...	...	...	...	...	...

Table 1: Extrait de la table CN, CCLE, 2012

Les variables de cette table sont sous la forme suivante : *nom_du_gene-X*, avec X qui vaut **DEL** lorsqu'il s'agit d'une délétion, et **AMP** pour une duplication.

[6] `https://portals.broadinstitute.org/ccle`

Pour la table *mutations*, on dispose cette fois-ci de 1638 variables, qui correspondent chacune à un locus étudié.

Lignée cellulaire	AAK1_MUT	AATK_MUT	ABCA3_MUT	...
1	0	0	0	...
2	0	1	0	...
...	...	...	...	...

Table 2: Extrait de la table *mutations* qui liste l'ensemble des SNP étudiés chez les lignées cellulaires, CCLE, 2012

3 Préparer les données à la prédiction : un contrôle-qualité nécessaire

Avant de procéder à la phase de modélisation, une première étape de contrôle qualité est nécessaire afin d'identifier par exemple des données incohérentes ou manquantes. De plus, nos données comportent un très grand nombre de variables ($\simeq$ 48 000), ce qui nous a conduit à produire un critère de sélection afin de réduire le nombre de variables de travail. La production de ce critère a été au cœur de notre contrôle qualité.

3.1 Quelques données aberrantes

Nous avons débuté notre contrôle qualité par une analyse descriptive des tables. Cela nous a permis d'identifier une variable ne contenant aucune valeur (NA pour toutes les lignées cellulaires) dans la table mutation, une ligne contenant les valeurs 1000 et 1001 dans les deux tables CN et mutation, ce qui n'est pas possible. Nous avons donc supprimé ces deux lignées cellulaires de nos tables.

En dehors de ces deux variations, les données génétiques ne comportaient pas de données manquantes.

3.2 Des valeurs manquantes : le sexe est-il explicatif ?

En dehors de données génétiques, nous disposons du sexe pour les lignées cellulaires étudiées. Néanmoins, cette variable présente des valeurs manquantes. Nous allons donc étudier son importance afin d'éventuellement l'écarter.

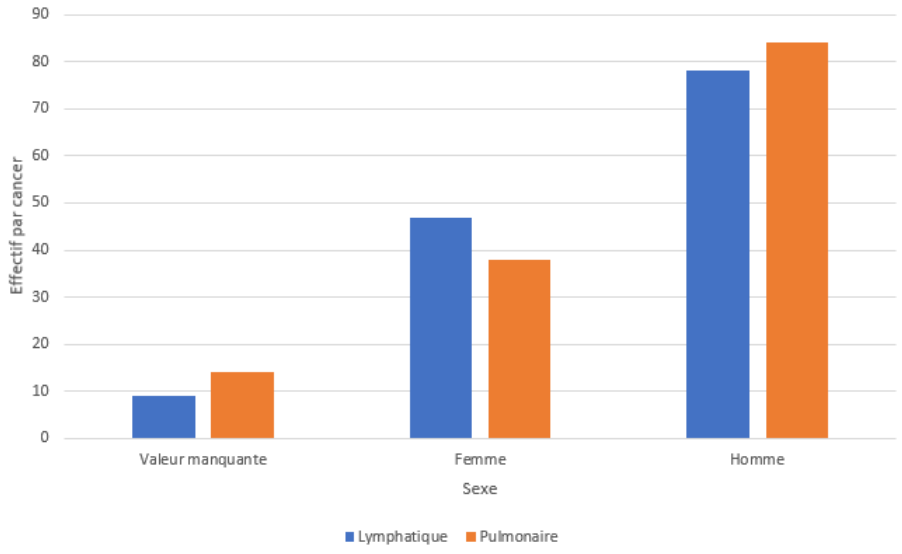

Figure 4: Comparaison de la répartition du sexe en fonction du type de cancer, CCLE, 2012

Au vu de ces deux histogrammes, le sexe ne semble pas avoir de lien avec le type de cancer. Confirmons cela avec un test.

Nous avons ainsi établi le tableau de contingence suivant :

	Lymphatique	Pulmonaire
Valeur manquante	9	14
Femme	47	38
Homme	78	84

Table 3: Tableau de contingence du sexe et du type de cancer, CCLE, 2012

Afin de déterminer le lien possible entre le sexe et le type de cancer, établissons le test suivant :

- H0 : Indépendance entre le sexe et le type de cancer
- H1 : Dépendance entre le sexe et le type de cancer

Sous l'hypothèse H0, la distance entre le tableau ci-dessus (*valeurs observées*) et le tableau des effectifs théoriques (en cas d'indépendance) suit une loi du Chi2 à 2 degrés de liberté. Ce dernier est défini par la formule :

degré de liberté = (Nb de modalités en ligne-1)(Nb de modalités en colonne-1) = 2

On fixe le seuil de décision à 5% qui est associé à une valeur du Chi2 de **5.99**. On obtient ici un Chi2 calculé de $\mathbf{2.25} < 5.99$ pour une *p-value* de $0.32 > 5\%$. On ne peut donc pas rejeter H0, soit l'indépendance entre le sexe et le type de cancer.

Disposant de 48 268 autres variables et l'influence du sexe n'étant pas significative pour prédire le type de cancer, nous avons décidé de nous affranchir de cette variable.

3.3 Notre prédiction va-t-elle être perturbée par des individus atypiques ?

En vue de construire un modèle prédictif, il nous faut écarter les individus trop atypiques. En effet, la présence de ce type de lignées cellulaires dans nos données d'entraînement biaiserait notre prédiction et rendrait la construction d'un modèle beaucoup plus difficile. De même, un individu atypique à prédire sera mal expliqué par le modèle. Nous devons donc les exclure. Pour cela, nous avons procédé à plusieurs Analyses en Composantes Principales (ACP) afin de déterminer les individus qui se démarquent sous plusieurs composantes.

Ce travail de contrôle qualité nécessitant de produire et d'analyser de nombreuses ACP (2 tables à contrôler, 6 dimensions par table), nous avons réalisé une application **Shiny** facilitant l'étude de nos ACPs.

Une documentation détaillée de cette application se trouve dans la section **A** de l'annexe.

Sur l'exemple ci-dessous, on remarque par exemple que la lignée cellulaire 115 est atypique selon les axes principaux 1, 3 et 5 par rapport au jeu de données de la table *mutations*.



Figure 5: Analyse de la lignée cellulaire 115 sur différents plans factoriels, CCLE, 2012.

Finalement, nous avons identifié 8 lignées cellulaires atypiques dans la table *mutations* et aucune dans la table *CN*, ce qui représente 2% du jeu de données. Parmi ces 8 lignées cellulaires, 3 comportent un cancer pulmonaire, 3 un cancer lymphatique et 2 inconnus. Cela respecte globalement les proportions du jeu de données entier. Ces individus étant peu nombreux et présentant des caractéristiques pouvant diminuer la qualité de notre modèle, nous avons fait le choix de les exclure.

3.4 Recherche d'un seuil

Les tables Mutations et CN comportent respectivement 1638 et 46631 variables. Néanmoins, certaines variations génétiques ne sont présentes que dans un nombre restreint de lignées cellulaires et n'apportent donc pas assez d'informations pour être pertinentes quant à la construction de notre modèle. Il nous a donc fallu déterminer une règle de décision pour réduire le nombre de variables à utiliser. Pour cela, nous avons décidé d'utiliser un seuil et de ne prendre en compte que les modifications génétiques présentes chez une proportion significative de la population. En effet, une modification génétique peu présente dans les données d'entraînement ne nous apporterait que peu d'informations pour notre modèle ; une modification génétique peu présente dans les données à prédire serait difficile à exploiter à cause de sa rareté et biaiserait notre prédiction. La détermination de ce seuil a été au centre de notre contrôle qualité.

Pour cela, nous avons développé une application interactive (R Shiny) qui permet de répondre à la question suivante : *Combien de variables sont présentes chez au moins X lignées cellulaires* avec X en entrée, sous forme d'un graphique :

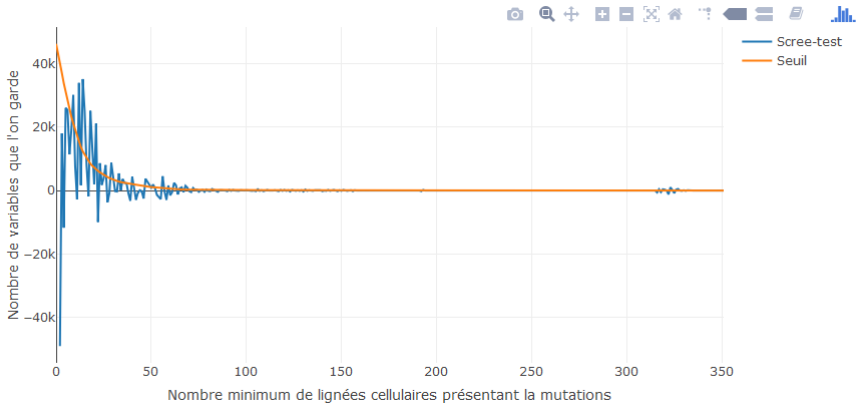

Figure 6: Nombre de variations génétiques en fonction du nombre de lignées cellulaires, CCLE, 2012

Note de lecture : Le point indiqué ci-dessus signifie qu'en ne gardant que les variables exprimées dans au moins 20 lignées cellulaires, on aurait 7149 variables.

Une documentation détaillée de cette application se trouve dans la section B de l'annexe.

La courbe orange représente le nombre de variables qui présentent une variation génétique en fonction du nombre de lignées cellulaires. Afin d'identifier un seuil, nous avons cherché un

décrochage dans la décroissance de cette courbe. Le critère du coude ne pouvant être appliqué ici de manière précise, nous avons décidé d'utiliser sa forme mathématisée : le critère du Scree-Test. Ce critère revient à calculer les différences secondes entre le nombre de variables gardées et à s'arrêter lorsque celles-ci sont négatives. La courbe bleue représente l'ensemble de ces différences.

Bien que l'on obtienne des différences négatives pour des faibles valeurs en abscisse, on ne retient pas ces seuils. En effet, ces derniers ne sont pas assez restrictifs. Par exemple, le premier seuil proposé garde encore 33 854 variables. Finalement, en étudiant les régions qui gardent moins de 10 000 variables, le Scree-Test nous suggère de prendre pour seuils 22 ou 27, ce qui correspond à garder respectivement 6184 et 4266 variables. Cela correspond aux variables présentes chez au moins 6% et 7% de notre échantillon.

Nous avons choisi de garder ces deux seuils. Cela nous permettra de tester les modèles associés et de choisir le meilleur. On pourra par exemple savoir si l'on a été trop restrictif ou pas assez sur le seuil.

3.5 Une dernière vérification

Afin de conclure cette première phase de contrôle qualité, il a été nécessaire à nouveau d'identifier les potentielles lignées cellulaires atypiques qui pourraient diminuer la qualité de notre prédiction. Finalement, aucun individu n'est apparu comme atypique suite à l'application du seuil sur les nouvelles ACPs : il n'y a pas d'individus significativement atypique sur une dimension et il n'y a pas d'individus qui se détachent du groupe sur plus d'une dimension.

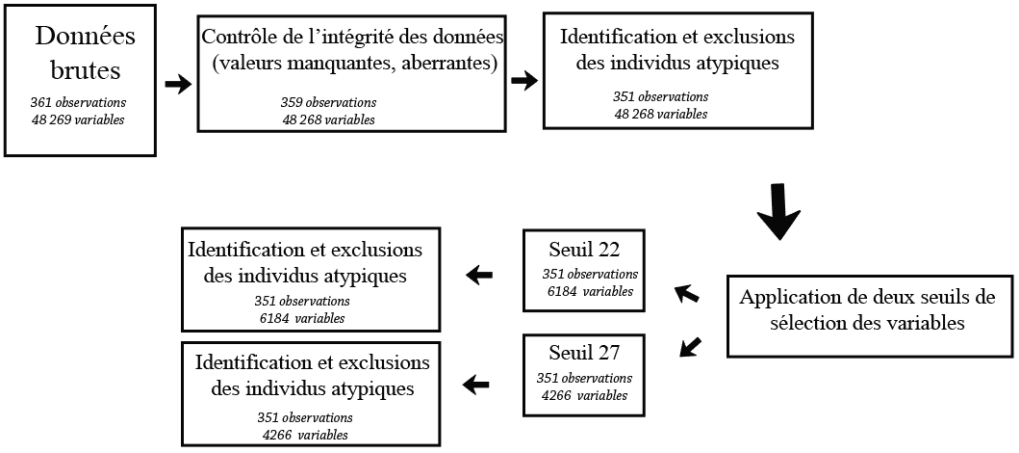

Figure 7: Résumé de notre processus de contrôle qualité

4 Une modélisation adaptée à un grand nombre de données

Malgré l'application d'un seuil, nos données contiennent encore un grand nombre de variables (toujours supérieur au nombre de lignées cellulaires). Cette contrainte nous a poussés à choisir un modèle linéaire capable de traiter un grand nombre de données. Notre choix s'est porté sur le modèle **elastic net**[7].

4.1 Les limites du modèle linéaire simple

Dans le modèle linéaire classique, on modélise notre variable de réponse Y (le type de cancer) de la manière suivante, sous sa forme vectorielle :

$$Y = X\beta + \epsilon$$

qui se décline ligne à ligne (individu par individu) en

$$Y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

avec :
Y_i le type de cancer de la lignée cellulaire i,
$x_{i,j}$ la valeur de la variable explicative j pour l'individu i
ϵ_i le terme d'erreur (bruit) pour l'individu i

La nature de nos données ne nous permet pas d'utiliser ce modèle tel quel. En effet, il est nécessaire de vérifier qu'il n'y a pas de corrélations entre les différentes variables. Or, cela n'est pas possible pour un si grand nombre de variables. Par conséquent, on risque de ne pas obtenir une solution unique et d'avoir une grande variabilité dans les coefficients du modèle. De plus, le nombre de variables étant beaucoup plus grand que celui des lignées cellulaires, on a un risque de sur-apprentissage. Mathématiquement, cela signifie que la matrice X n'est pas de rang colonne plein et X^tX n'est pas inversible.

Pour limiter le sur-apprentissage, on peut utiliser une technique, la régularisation, qui consiste à contrôler simultanément l'erreur du modèle sur le jeu d'entraînement et la complexité du modèle. Pour cela, on rajoute un hyperparamètre λ et un régularisateur $g(\beta)$ dans notre modèle qui permettra de pénaliser la complexité et le nombre de variables du modèle. Il faut alors minimiser :

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left(\underbrace{||Y - X\beta||_2^2}_{erreur} + \underbrace{\lambda g(\beta)}_{regularisation} \right)$$

[7]Les explications suivantes sont inspirées du MOOC (cours en ligne) disponible à l'adresse suivante : https://openclassrooms.com/courses/entrainez-un-modele-predictif-lineaire.

4.1.1 Réduction de l'amplitude des variables

Afin d'éviter le sur-apprentissage, on utilise souvent[8] $g(\beta) = ||\beta||_2^2$

où $||x||_2^2 = \sum_i^n x_i^2$

Autrement dit, on cherche à résoudre pour $\lambda \in \mathbb{R}_+$:

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left(\underset{erreur}{||Y - X\beta||_2^2} + \underset{complexite}{\lambda ||\beta||_2^2} \right)$$

Cette régression permet ainsi d'éviter le sur-apprentissage en restreignant l'amplitude des coefficients. Elle a un effet de sélection groupée : si deux variables sont très corrélées, la régression leur attribuera le même coefficient.

4.1.2 Réduction du nombre de variables

La méthode ci-dessus nous permet de réduire l'amplitude des coefficients d'une régression linéaire et d'éviter le sur-apprentissage. Cependant, elle ne permet pas de sélectionner certains coefficients. On utilise alors une autre méthode[9] : certaines variables auront un coefficient égal à zéro et ne feront plus partie du modèle, qui en sera d'autant simplifié.

Pour cela, on remplace le terme de régularisation par $g(\beta) = ||\beta||_1$ avec $||x||_1 = \sum_i^p |x_i|$; on passe de la norme L_2 de β à sa norme L_1,

ce qui revient à résoudre pour $\lambda \in \mathbb{R}_+$:

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left(||Y - X\beta||_2^2 + \lambda ||\beta||_1 \right)$$

En utilisant la norme L_1, on va réduire encore une fois les valeurs de β_j, mais cette fois-ci la réduction sera tellement importante que certaines seront nulles, réduisant ainsi le nombre de variables.

4.1.3 Elastic Net : Un modèle stable et parcimonieux

En résumé, l'utilisation du premier régulateur (*Ridge*) permet d'obtenir un modèle stable en combinant les effets des variables explicatives très corrélées. Néanmoins, elle ne permet pas de réduire le nombre de variables. Le deuxième régulateur (*LASSO*), quant à lui, permet de sélectionner les variables pertinentes. Néanmoins, si deux variables sont très corrélées, elle choisit arbitrairement l'une des deux et fixe son coefficient à 0.

Afin de s'affranchir des inconvénients et de profiter de ces deux méthodes, nous avons décidé d'utiliser la régression Elastic Net, qui utilise deux hyperparamètres. On cherche à minimiser :

[8] Cette méthode est appelée *Regression Ridge*

[9] Méthode dite LASSO: Least Absolute Shrinkage.

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left(||Y - X\beta||_2^2 + \underbrace{\lambda}_{intensite} \underbrace{(\alpha||\beta||_1 + (1-\alpha)||\beta||_2^2)}_{penalite} \right)$$

Cette méthode permet d'obtenir une solution moins parcimonieuse que le LASSO, mais plus stable et dans laquelle toutes les variables corrélées pertinentes pour la prédiction sont sélectionnées et reçoivent un poids identique.

Dans ce modèle, le paramètre α est un paramètre mixte qui donne plus ou moins de poids aux normes L1 et L2. $\alpha = 0$ correspond au modèle ridge ; $\alpha = 1$ au modèle LASSO. Le paramètre λ quant à lui multiplie la pénalité et correspond donc à son intensité.

Néanmoins, puisqu'il faut sélectionner à présent un couple optimal d'hyperparamètres, cette méthode nécessite davantage de ressources computationnelles.

4.1.4 Comment mesurer la qualité de notre prédiction ?

Afin de mesurer la qualité de notre modèle, nous avons utilisé la validation croisée sur notre échantillon d'entraînement (80% des lignées dont les cancers sont connus). Nous avons mesuré la qualité de nos modèles avec le pourcentage de bien classés et le Kappa de Cohen.

Principe de la validation croisée : On divise l'échantillon original en k échantillons, puis on sélectionne un des *k* échantillons comme ensemble de validation et les *k-1* autres échantillons constitueront l'ensemble d'apprentissage. On calcule comme dans la première méthode le score de performance (ici le taux de bien classés), puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les *k-1* échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi *k* fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des *k* erreurs est enfin calculée pour estimer l'erreur de prédiction.

De plus, nous avons choisi d'utiliser une validation croisée dite répétée. Afin de s'affranchir de l'aléatoire dû au tirage des k échantillons, nous répétons plusieurs fois la validation croisée afin d'avoir un résultat plus robuste. Le nombre de répétitions est limité par son coût en ressources computationnelles. Nous avons effectué 100 répétitions par validation croisée.

Kappa de Cohen : Le taux de bien classés et de la courbe ROC[10] n'étaient pas assez discriminants pour choisir les paramètres de notre modèle (beaucoup donnaient des résultats similaires). Ainsi, nous avons utilisé une deuxième métrique, le Kappa de Cohen. Celui-ci est également plus robuste face aux distributions inégales de modalités. Il varie entre 0 et 1 et plus il est proche de 1, meilleure est la qualité.

Une fois les paramètres de nos modèles fixés, nous avons calculé l'aire sous la courbe ROC (AUROC)[11] : plus cette aire est importante (*proche de 1*), meilleure sera la qualité.

Puisque l'on n'a pas de préférence de prédiction entre les deux types de cancer (on veut autant prédire l'un que l'autre), nous allons considérer identiquement la spécificité et la sensibilité. De plus, nous n'allons pas différencier les coûts d'erreur de prédiction.

Enfin, nous validerons notre modèle grâce à un échantillon test constitué des 20% des lignées cellulaires dont le type de cancer est inconnu.

[10] Receiver Operating Characteristic : fonction d'efficacité du récepteur.
[11] Area Under ROC curve.

4.2 Premiers modèles et comparaisons

4.2.1 Premier modèle : l'elastic net appliqué aux données seuillées à 22

Rappel Les données seuillées à 22 correspondent aux variables présentes chez au moins 22 lignées cellulaires (soit environ 6% de notre population).

Afin de déterminer le meilleur couple α, λ, nous avons représenté les différents couples possibles sur une grille.

À nouveau, nous avons conçu une application Shiny pour nous aider dans la détermination des paramètres optimaux. La documentation de celle-ci se trouve dans la section C de l'annexe.

La grille ci-dessous a été obtenue en effectuant 100 validations croisées sur les données d'entraînements. Pour chaque couple (α, λ), on obtient donc 100 taux de bien classés différents. La taille du carré représente la variabilité des taux de bien classés obtenus : plus il est grand, plus la variance des taux de bien classés obtenus est élevée (donc moins le résultat est fiable). Au contraire, plus le carré est petit, plus le taux de bien classés associé (égal à la moyenne des taux de bien classés obtenus) est fiable.

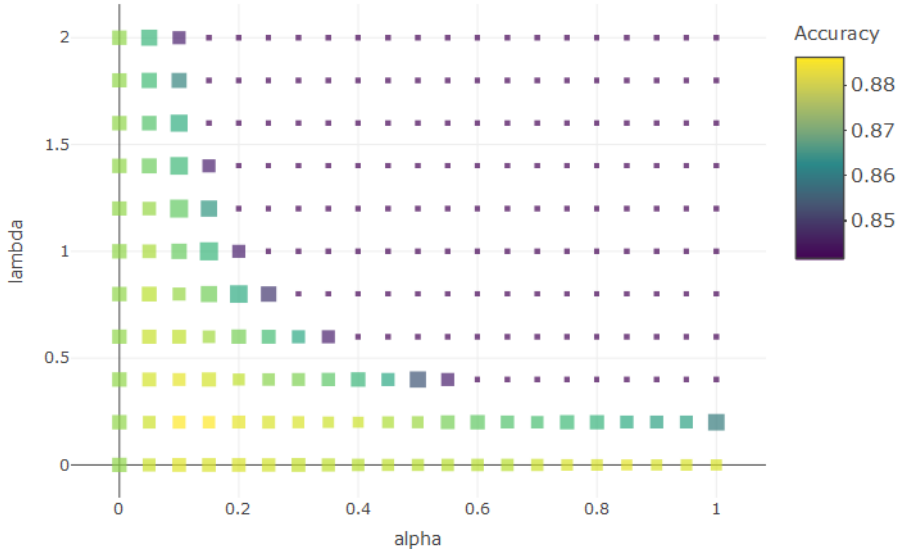

Figure 8: Grille initiale des hyperparamètres α, λ du modèle elastic net appliqué aux variations génétiques présentes chez au moins 22 lignées cellulaires, CCLE 2012

Afin d'obtenir une meilleure précision, nous avons reproduit cette grille à nouveau en diminuant le pas et en la centrant sur les zones plus intéressantes.

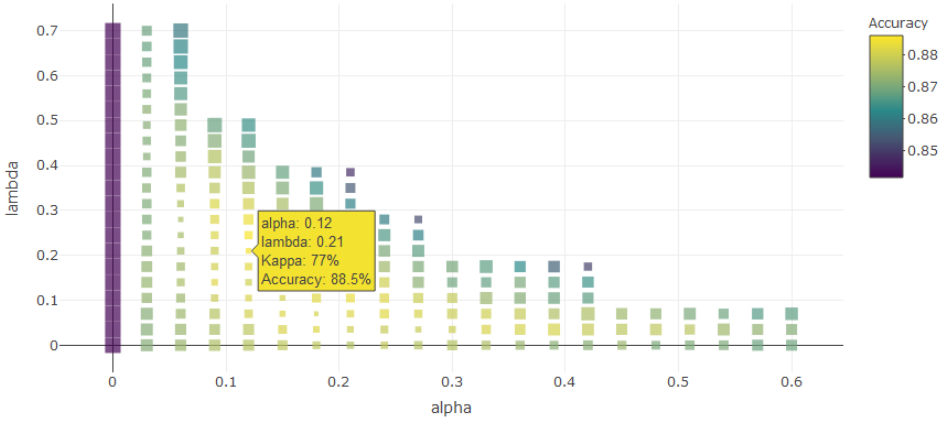

Figure 9: Grille agrandie des hyperparamètres α, λ du modèle elastic net appliqué aux variations génétiques présentes chez au moins 22 lignées cellulaires, CCLE 2012

Finalement, le couple (α, λ) tel que $\alpha = 0.12$ et $\lambda = 0.21$ donne un taux de bien classés optimisé de 88% et un Kappa de 77%. Suite à l'application du modèle, on ne garde plus que 536 variables sur les 6184 variables d'entrées. Nous appellerons ce modèle $M1$.

4.2.2 Deuxième modèle : l'elastic net appliqué aux données seuillées à 27

En utilisant la même méthodologie, nous avons déterminé que le couple $(\alpha, \lambda) = (0.20, 0.20)$ donne un taux de bien classés optimisé de 87%. Suite à l'application du modèle, on ne garde plus que 293 variables sur les 4266 d'entrées. Nous appellerons ce modèle $M2$.

4.2.3 Comparaisons de M1 et de M2

Le taux de bien classés est très similaire entre $M1$ et $M2$ (respectivement 88% et 87%). Néanmoins, le modèle $M2$ comprend significativement moins de variables (55% moins). De plus, en étudiant ces variables, nous avons remarqué que le modèle M2 était imbriqué dans le modèle M1, à l'exception d'une seule variable. Ainsi, les variables présentes dans $M1$ mais pas dans $M2$ n'améliorent pas la qualité de la prédiction et ne sont pas pertinentes dans le modèle. Le principe de parcimonie nous pousse ainsi à préférer le modèle avec le moins de variables, soit $M2$.

4.3 Amélioration de la sélection et réduction du temps de calcul : tests multiples sur les variables

Dans le but d'obtenir une meilleure interprétabilité et de réduire le temps de calcul de nos modèles, nous avons adopté une autre stratégie de sélection de variables en plus du seuil. Afin de sélectionner les variables pertinentes dans notre modèle, nous avons conduit une analyse univariée pour chaque variable disponible (après seuil) en implémentant pour chacune d'entre

elles une modélisation (régression linéaire simple) qui cherche à expliquer le type de cancer par la variable OMIC. Ainsi, nous avons obtenu autant de modèles (et donc de tests) que de variables. Nous avons alors fait face à un problème de multiplicité des tests, que nous avons corrigé avec la méthode dite de Bonferroni[12].

La méthode dite de Bonferroni : Pour se prémunir contre l'augmentation du risque de première espèce dans les tests multiples, la méthode de Bonferroni propose de majorer le risque réel pour l'ensemble des tests par le niveau de test choisi (en général 5%), en divisant pour cela le niveau de chacun des tests effectuées par le nombre total de tests.

Enfin, nous avons effectué un modèle elastic net en n'intégrant que les variables OMICs qui ressortent significatives (après correction de la multiplicité des tests).

Cette démarche nous a permis d'accélérer les temps de calcul du modèle elastic net, et également de réduire considérablement le nombre de variables dans notre modèle.

Ainsi, en appliquant cette méthodologie à notre table seuillée à 22, on garde 254 variables en entrée. Puis, suite à la sélection de variables de l'elastic net, on ne garde plus que 102 variables au total pour une qualité de prédiction de 87% : ce résultat est extrêmement intéressant. Nous appellerons ce modèle $M3$.



Figure 10: Processus de sélection successifs des variables du quatrième modèle

De même, cette méthodologie appliquée à notre table seuillée à 27 ($M4$)produit les résultats suivants : 87% de taux de bien classés pour 69 variables. De même que pour $M1$ et $M2$, on a que $M4$ est imbriqué dans $M3$, à l'exception d'une variable (la même que pour $M1$ et $M2$). De plus, on obtient avec $M3$ le modèle le plus parcimonieux tout en gardant une très bonne qualité de prédiction. C'est donc celui-ci que nous choisirons de garder; ses hyperparamètres optimaux sont $(\alpha, \lambda) = (0.59, 0.01)$. Pour confirmer ce choix, nous avons calculé son AUC pour lequel on obtient 95%, ce qui est un très bon résultat.

	M1 (seuil 22)	M2 (seuil 27)	M3 (22, Bonferroni)	M4 (27, Bonferroni)
Nb de var d'entrée	6184	4266	254	241
Nb de var gardées	536	293	102	69
Taux de bien classés	88%	87%	87%	87%

Table 4: Récapitulatif et comparaison de performances des modèles prédictifs, CCLE, 2012

[12]L'explication suivante est inspirée du cours suivant : https://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-m-modmixt8-bonfer.pdf

5 Est-il possible d'écarter un des deux types de données ?

Lors de la phase de contrôle qualité, nous avons produit une ACP sur l'ensemble des données, mais également uniquement sur la table CN :



Figure 11: Comparaisons des ACP d'ensemble des données (en haut) et de la table CN seule (en bas)

Ces deux graphiques se ressemblant beaucoup, nous avons eu l'intuition qu'un des deux types de données (CN ou mutations) pouvait ne pas être pertinent dans la prédiction.

Pour confirmer ou infirmer cette intuition, nous avons produit un modèle elastic net (avec la méthodologie précédente) sur les données suivantes :

1. Ensemble des données (CN et mutations), que l'on appellera "modèle complet".
2. La table CN seulement, que l'on appellera "modèle CN".
3. La table mutations seulement, que l'on appellera "modèle mutations".

Si l'on continue à garder de bons résultats de prédiction avec seulement l'une des deux tables, cela signifiera que l'on peut s'affranchir de l'autre.

Ayant identifié le modèle seuillé à 27 joint à la méthode de Bonferroni comme étant le meilleur modèle (241 variables d'entrée pour 69 variables gardées par le modèle), c'est celui-ci que nous allons utiliser pour comparer les différents modèles.

Pour comparer ces différents modèles, nous allons effectuer 100 validations croisées sur chacun d'entre eux, puis comparer la distribution des précisions obtenues. Les résultats obtenus

sont listés dans le graphique suivant :

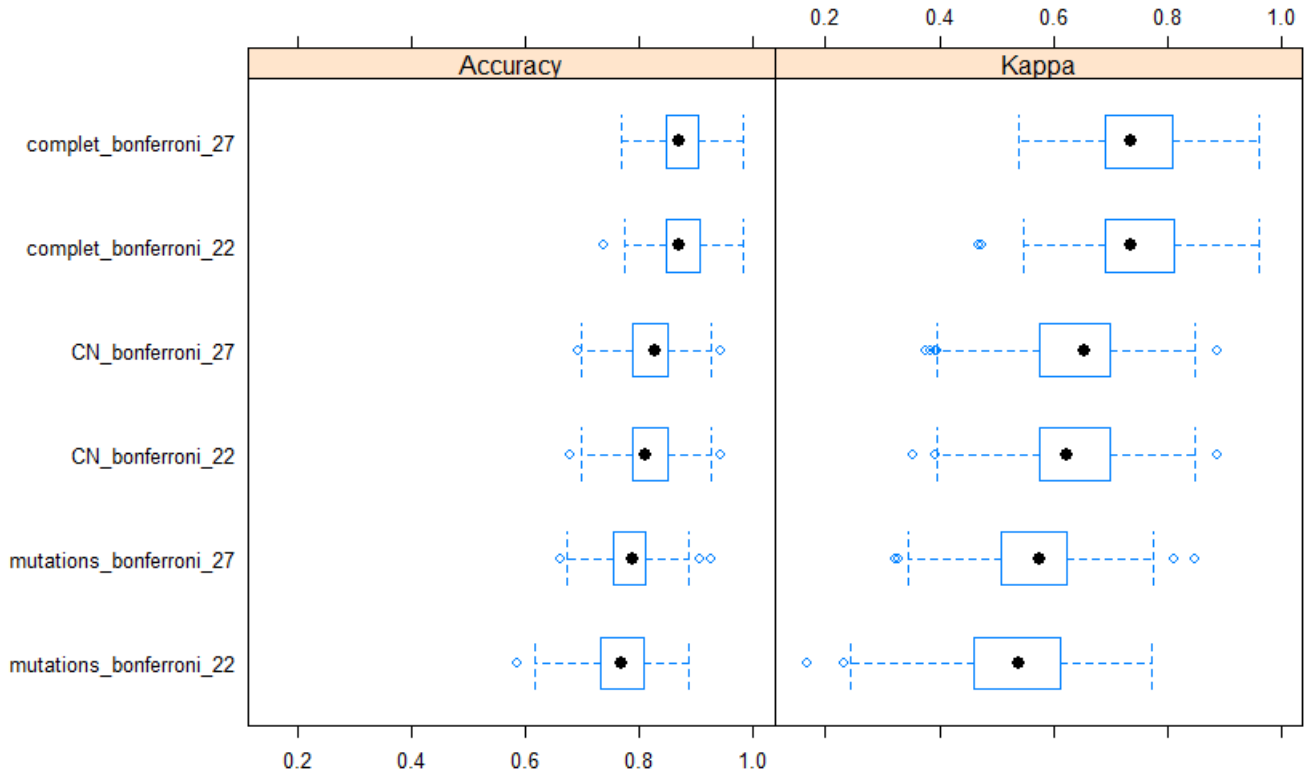

Figure 12: Comparaisons de la précision entre les modèles complet, CN et mutations, CCLE 2012

Nous avons également créé une application Shiny pour cette étape afin de faciliter la comparaison des modèles. Une documentation détaillée de celle-ci se trouve dans la partie D de l'annexe.

En comparant ces différents modèles, on arrive à une précision médiane de 82% pour CN et 76% pour mutations. Bien que la qualité de la prédiction soit bonne, elle ne l'est pas autant qu'en prenant en compte l'ensemble des variables. La différence est encore plus accentuée lorsque l'on considère le Kappa de Cohen.

	Complet, 27, Bonferroni	CN, 27, Bonferroni	Mutations, 27, Bonferroni
Nombre var avant	241	236	11
Nombre var après	69	32	11
Taux de bien classés	87%	82%	78%

Table 5: Comparaisons du nombre de variables d'entrée et gardées par le modèle entre les modèles complet, CN et mutations, après une préselection des variables, CCLE, 2012

Note de lecture : Complet, 27, Bonferroni, signifie que le modèle correspond au jeu de données complet, à partir duquel nous n'avons gardé que les variables présentes chez au moins 27 lignées cellulaires, enfin on applique la méthode de Bonferroni au jeu de données obtenu

Néanmoins, travailler sur seulement l'une des tables permet d'avoir une meilleure sélection des variables. On s'aperçoit également que le modèle garde l'ensemble des variables d'entrées pour la table Mutations. Cela n'est pas étonnant car il n'y a que 11 variables en entrée.

Finalement, il ne nous semble pas possible d'écarter l'une des deux variations génétiques dans la construction d'un modèle prédictif précis et robuste. Néanmoins, il est intéressant de produire les modèles seulement sur les jeux de données des tables Mutations et CN car il permet d'identifier un petit nombre de variables explicatives. En particulier, Mutations ne garde que 11 variables dans ce modèle, ce qui est propice à une analyse détaillée. Enfin, la table CN seule propose de meilleurs résultats que la table Mutations seule, pour un nombre raisonnable de variables. À choisir, nous privilégierons la table CN par rapport à la table Mutations, dans l'optique de créer un modèle prédictif.

6 Retour sur le modèle final

Finalement, le modèle choisi est $M4$ qui garde 69 variables. Parmi ces 69 variations génétiques, 5 sont des SNP et 64 sont des CN. Parmi ces CN, 41 sont des duplications et 23 des délétions. Ainsi, les duplications de gènes représentent 60% de nos variables et semblent jouer un rôle majeur dans la prédiction du type de cancer.

On remarque aussi qu'il n'y a que 5 SNP retenus. En étudiant les modèles M1, M2 et M3, nous nous sommes aperçus que c'était aussi le cas pour ces modèles et que les 5 SNP retenus sont toujours les mêmes. Cela nous pousse à penser que ces 5 SNPs semblent ici jouer un rôle majeur dans la prédiction du type de cancer.

6.1 Quelles sont les variables les plus importantes ?

Étudions de plus près les coefficients de notre modèle. Pour cela traçons le chemin de régularisation de notre modèle :

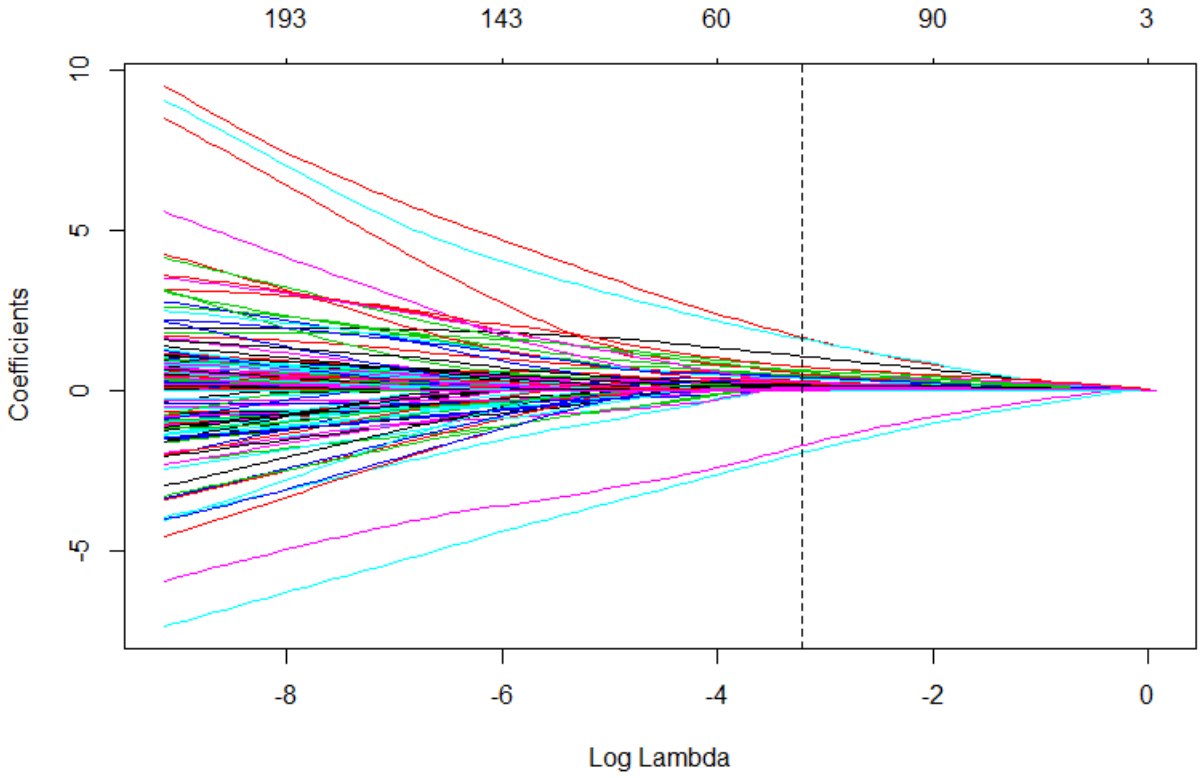

Figure 13: Chemin de régularisation du modèle prédictif M4 finalement retenu, CCLE, 2012

En fixant le paramètre de mixage α sur une valeur optimisée, nous faisons varier sur le graphique ci-dessus la valeur de λ. Chaque trait représente le coefficient d'une variable selon

λ (ou $log\lambda$). On peut voir qu'à droite du graphique, les coefficients sont tous nuls, et au fur et à mesure que l'on se déplace vers la gauche, on assiste à un décollement des variables (les coefficients associés ne sont plus nuls). L'axe horizontal du haut représente le nombre de variables dont le coefficient est non nul (qui ont donc décollé). On voit que 4 variables décollent de manière significative assez rapidement, et les deux premières très tôt dans le modèle. On remarque également que ces deux dernières atteignent des valeurs de coefficients plus importantes (en valeur absolue).

Représentons ces variables à présent par ordre d'importance :

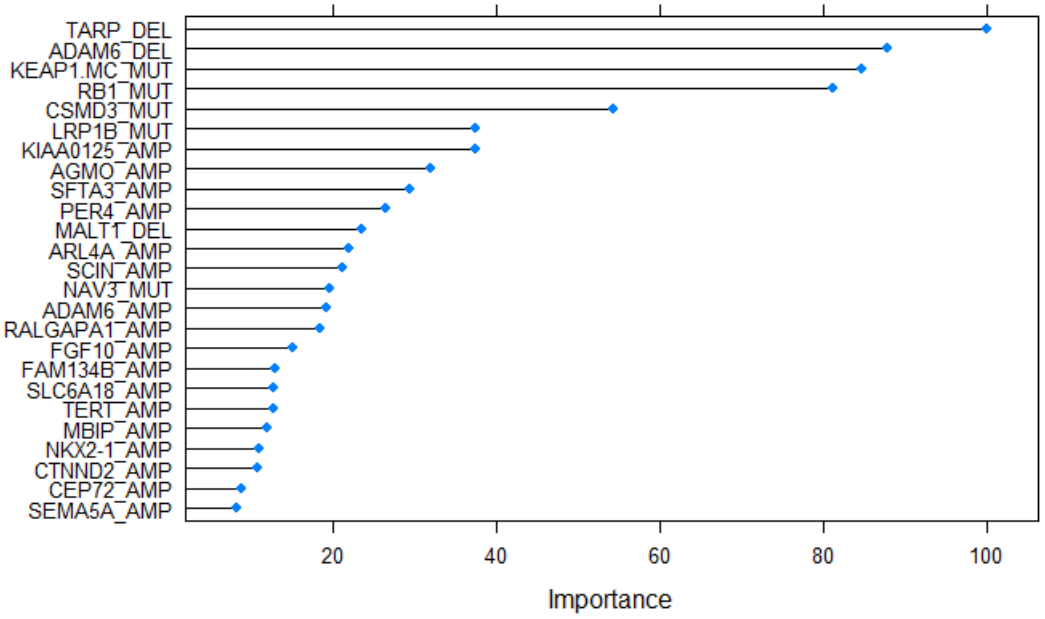

Figure 14: Les 20 variables dont les coefficients sont les plus importants dans le modèle prédictif M4 finalement retenu, CCLE, 2012

On définit l'importance d'une variable par sa statistique de Student. Plus cette valeur est importante, plus la p-valeur associée est faible, et plus on rejette l'hypothèse de nullité du coefficient ; donc plus le coefficient est pertinent.

Ce graphique confirme ce que l'on avait vu sur le chemin de régularisation, avec quatre variables qui se détachent.

En appliquant un critère du coude au graphique des *t-test*, on peut choisir de retenir 4, 5 ou 7 coefficients à analyser. On remarque cependant que 4 SNPs sont contenus dans les 7 premiers coefficients, c'est pourquoi ce sont ces derniers que nous retenons dans notre analyse.

Variation génétique	Coefficients	Signe
TARP_DEL	-1.97	-
ADAM6_DEL	-1.73	-
KEAP1.MC_MUT	1.67	+
RB1_MUT	1.60	+
CSMD3_MUT	1.07	+
LRP1B_MUT	0.74	+
KIAA0125_AMP	0.74	+

Table 6: Valeurs des 7 coefficients les plus importants dans le modèle prédictif M4 finalement retenu, CCLE, 2012

Encore une fois, ce tableau est en accord avec le chemin de régularisation, qui mettait en évidence quatre variables importantes avec deux coefficients négatifs plus grands en valeur absolue que les autres.

Puisque notre variable de réponse est binaire (0 ou 1), le signe du coefficient a son importance. Regardons de plus près la répartition des cancers selon la présence des variations génétiques ci-dessus.

Délétion de TARP	Lymphatique	Pulmonaire
Absence	101	134
Présence	28	0

Délétion de ADAM6	Lymphatique	Pulmonaire
Absence	105	134
Présence	24	0

Table 7: Répartition des cancers selon la délétion des gènes TARP et de ADAM6, CCLE, 2012

Pour les coefficients négatifs On remarque que lorsqu'une délétion du gène TARP est présente, la lignée cellulaire présente systématiquement une cancer lymphatique. On peut voir des résultats similaires pour le gène ADAM6.

SNP sur KEAP1.MC	Lymphatique	Pulmonaire
Absence	129	107
Présence	0	27

SNP sur RB1	Lymphatique	Pulmonaire
Absence	128	109
Présence	1	25

Table 8: Répartition des cancers selon la présence de SNP aux locus KEAP1.MC et RB1

Pour les coefficients positifs On remarque des résultats inversés. Lorsqu'il y a présence d'un SNP sur les locus KEAP1.MC et RB1, le cancer associé est systématiquement (sauf pour une lignée) pulmonaire. On retrouve cette tendance pour les autres coefficients, mais la répartition est moins distincte.

Finalement, on en déduit que les variables aux coefficients négatifs sont associées au cancer lymphatique tandis que les variables aux coefficients positifs sont associés au cancer pulmonaire. De plus, les seuls coefficients négatifs dans notre modèle sont ceux associées à la délétion des gènes TARP et ADAM6. Ce sont aussi les plus importants. Ainsi, l'explication du cancer lymphatique est concentrée en grande partie dans ces deux gènes, qui représentent peu de variables.

En faisant des recherches dans la littérature existante, nous avons trouvé que le gène TARP[13] est connu des biologistes et est associé aux cancers du sein et de la prostate. Quant au gène ADAM6[14], il fait partie de la famille de gènes *ADAM*, et est connu pour jouer un rôle dans le développement de tumeurs malignes[15].

6.2 Où sont situées les variations génétiques ?

Nous avons également étudié la localisation des CN et des SNP retenus par notre modèle grâce à l'outil en ligne proposé par l'Université de Californie à Santa Cruz (UCSC)[16]. Les résultats sont listés dans le tableau ci-dessous :

[13] *TCR gamma alternate reading frame protein.*
[14] *A disintegrin and metalloproteinases.*
[15] Cancéreuses.
[16] `https://genome.ucsc.edu`

Chromosome	Délétion	Duplication	SNP	Total
2	0	0	1	1
5	0	27	0	**27**
7	1	7	0	8
8	0	0	1	1
12	0	0	1	1
13	0	0	1	1
14	1	6	0	7
18	21	1	0	**22**
19	0	0	1	1

Table 9: Répartition des différentes variations génétiques retenues par le modèle M4 sur les différents chromosomes, CCLE 2012 et UCSC 2018

On peut ainsi voir que 70% de nos variables sont concentrées sur les chromosomes 5 et 18. On remarque également que l'ensemble des variations génétiques sur le chromosome 5 sont des duplications, tandis que toutes celles du chromosome 18, à l'exception d'une, sont des délétions de gène. Nous n'avons pas trouvé de littérature intéressante concernant le lien entre les chromosomes 5 et 18 et les cancers lymphatiques et pulmonaires.

6.3 Quelles variations génétiques pour quel cancer ?

Nous avons vu que notre modèle retenait une majorité de duplication de gènes (60% des variables). Étudions la répartition des variations génétiques selon le type de cancer.

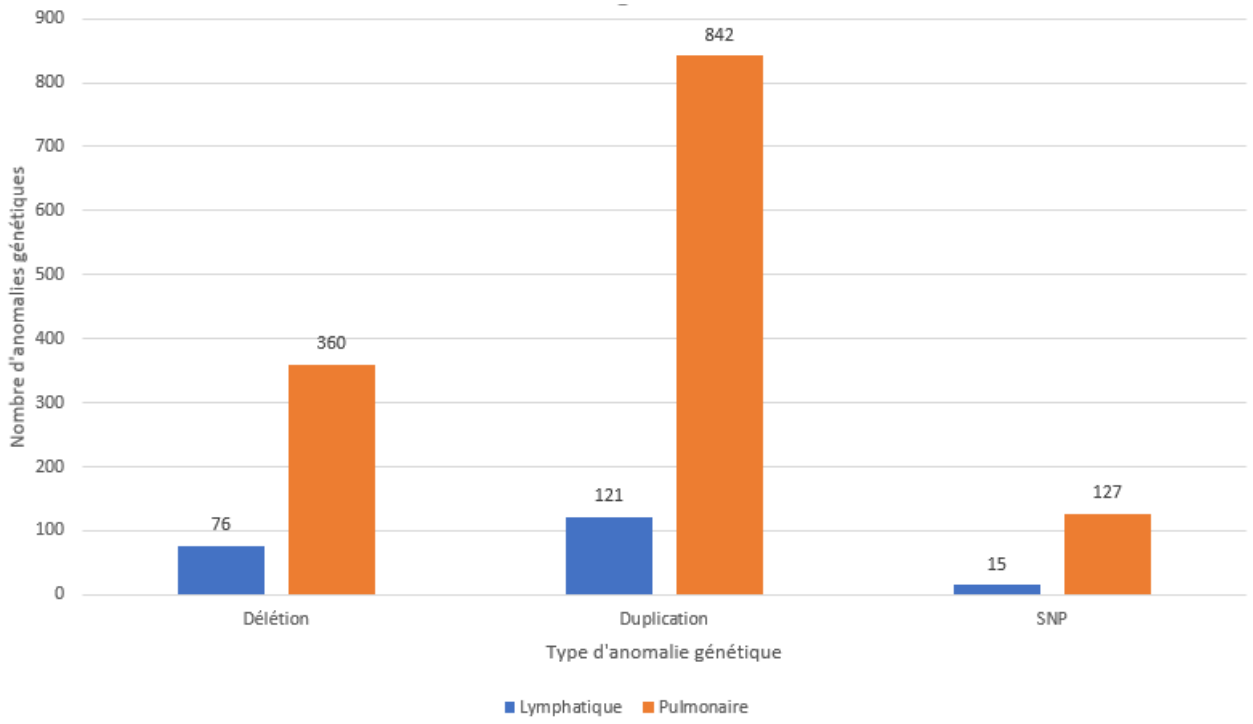

Figure 15: Types et nombre de variations génétiques par type de cancer, CCLE, 2012

Note de lecture : Parmi les 69 variations génétiques retenus par le modèle M4, on compte 842 duplications de gènes au total dans l'ensemble des lignées cellulaires atteintes par un cancer pulmonaire.

Au vu des effectifs des variations génétiques selon le type de cancer, nous pouvons voir que les lignées cellulaires atteintes de cancer lymphatique présentent moins de variations génétiques (212 contre 1329, soit 6 fois moins) que celles atteintes d'un cancer pulmonaire. Ainsi, la présence d'un grand nombre de variations génétiques dans une lignée cellulaire suggère un cancer pulmonaire. De plus, la duplication semble être la variation génétique la plus répandue parmi les lignées cellulaires atteintes d'un cancer : cela est dû aux variables finalement retenues par le modèle (41 duplications sur 69 variations génétiques). On remarque enfin que les lignées cellulaires atteintes d'un cancer pulmonaire présentent au total dix fois plus de SNP que celles atteintes d'un cancer lymphatique.

7 Dernière étape : notre modèle à l'épreuve de l'échantillon de validation

Pour conclure notre étude et valider notre modèle, nous avons appliqué notre modèle aux lignées cellulaires dont les types de cancer sont inconnus. Nous avons ensuite comparé nos prédictions aux types réels de cancer[17]. En voici la matrice de confusion :

	Lymphatique (Réel)	Pulmonaire (Réel)
Lymphatique (Prédit)	41	5
Pulmonaire (Prédit)	2	40

Table 10: Matrice de confusion sur l'échantillon de validation.

Note de lecture : 41 lignées avaient pour type réel de cancer "lymphatique" et ont été prédites "lymphatique".

Nous obtenons ainsi un résultat de 92% de bons classés, ce qui est très satisfaisant.

Nous nous sommes intéressés au cas des individus mal classés. On remarque que parmi ces derniers, les cancers pulmonaires conduisent à deux fois plus d'erreur que les cancers lymphatiques. Cela peut être dû au fait que l'explication du cancer lymphatique est concentrée dans deux variables, ce qui donne une règle de prédiction plus simple pour le modèle.

En étudiant les 7 variables les plus importantes de notre modèle parmi les 7 individus mal classés, nous n'avons pas identifié d'incohérence dans la prédiction. Cela signifie que l'erreur est due aux 62 autres variables qui composent notre modèle. On peut également supposer que ces individus sont atypiques et ne rentrent donc pas dans notre modèle de prédiction.

[17] fournis à la fin de notre projet par notre tuteur

8 Conclusion

Le projet Cancer Cell Line Encyclopedia (CCLE[18]) a pour but de récolter, regrouper et mettre à disposition le plus grand nombre de données possibles au sujet de lignées cellulaires atteintes de cancer. Ainsi, chacun est libre de disposer de ces données et de conduire ses recherches afin de contribuer à la recherche scientifique et médicale. En ce sens, nous sommes très heureux que notre projet statistique de deuxième année se soit basé sur les données du projet CCLE.

À partir de 361 lignées cellulaires initiales (toutes atteintes d'un cancer soit lymphatique, soit pulmonaire) nous avons pu, suite à une préparation rigoureuse des données, produire et comparer plusieurs modèles prédictifs. Même si l'on obtient malgré tout des résultats relativement satisfaisants en ne considérant qu'un type de variation génétique, il ressort de ces comparaisons une nécessité de considérer à la fois les variations du nombre de copies ainsi que les polymorphismes de nucléotides simples afin de construire un modèle précis et robuste.

Une préparation des données complète via un seuil et des tests multiples couplée à un modèle adapté à la sélection de variables nous a permis de ne retenir que 69 variables sur les 48 269 initiales, soit seulement 0.14% du jeu de données de départ pour une précision finale de 92% sur le jeu de données à prédire. Cela est évidemment très satisfaisant, mais signifie surtout que la grande majorité du jeu de données complet constitue du bruit statistique, qu'il faut donc filtrer.

Enfin, nous avons pu nous faire une meilleure idée sur les mécanismes expliquant les deux types de cancer. D'abord la prépondérance des duplications par rapport aux autres variations génétiques, mais également la concentration des facteurs explicatifs du cancer lymphatique sur deux variations génétiques seulement. Ce constat se généralise puisque les lignées cellulaires atteintes de cancer pulmonaire présentent en moyenne six fois plus d'anomalies génétiques que celles atteintes du cancer lymphatique. Enfin, plus des deux tiers des variations génétiques explicatives retenues par notre modèle sont situées sur les chromosomes 5 et 18. Ce sont vers eux qu'ira notre attention suite au rendu de ce rapport, avant d'étudier à nouveau et plus en détails, les différentes variations génétiques identifiées par notre modèle.

[18] `https://portals.broadinstitute.org/ccle`

Bibliographie

- CHAOUCHE, Yannis MOOC. *Entraînez un modèle prédictif linéaire*, 2018 https://openclassrooms.com/courses/entrainez-un-modele-predictif-lineaire
- https://portals.broadinstitute.org/ccle
- https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modmixt8-bonfer.pdf

9 Annexe

Dans cette annexe se trouve la documentation détaillée de l'ensemble de nos applications Shiny. Ces dernières sont en fait regroupées en une seule application, à partir de laquelle il est possible de choisir la partie du projet que l'on souhaite travailler.

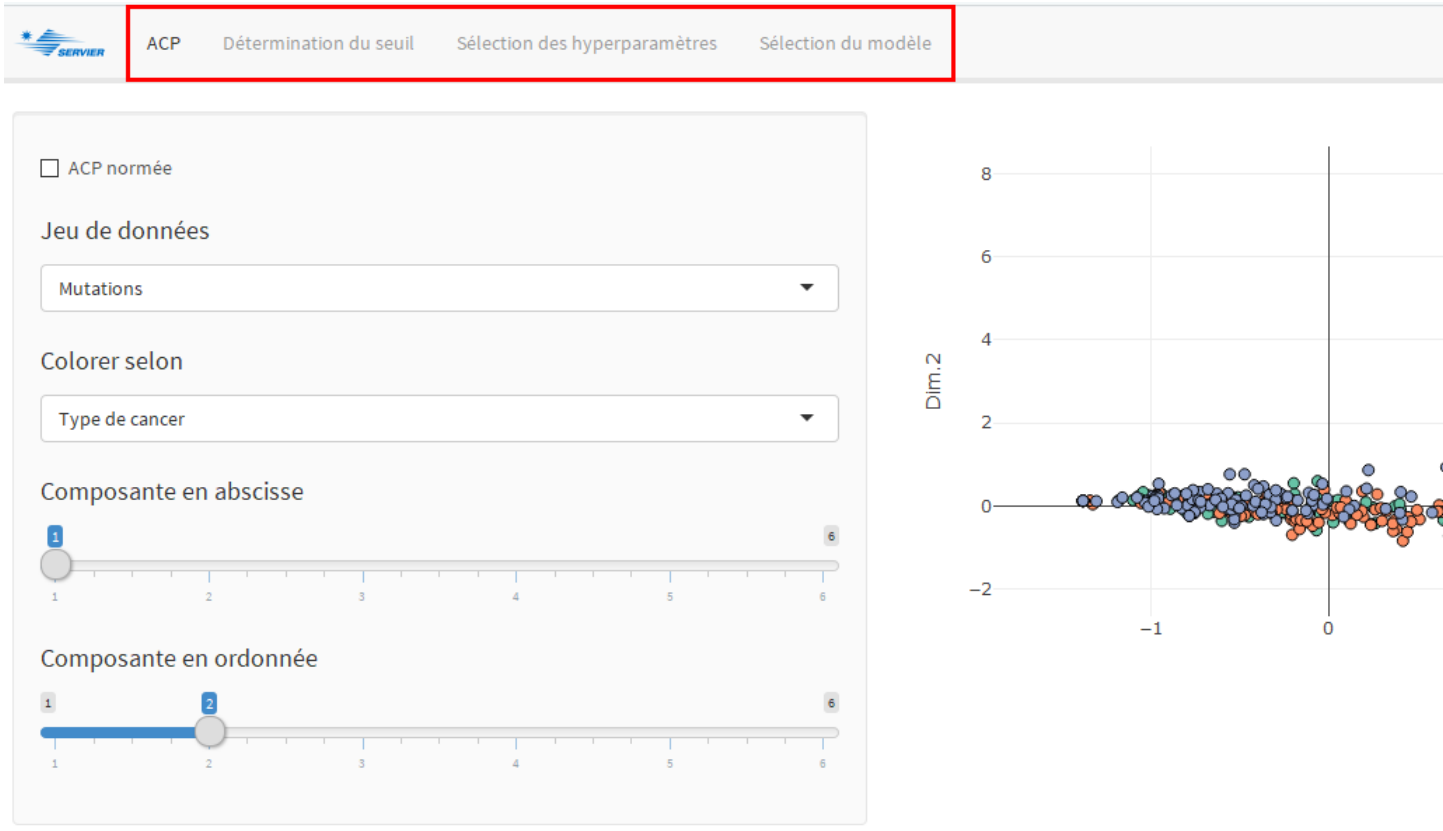

Figure 16: Capture d'écran de notre application Shiny globale : les différents onglets pour chaque étape du projet

A. Étudier des ACP de manière interactive

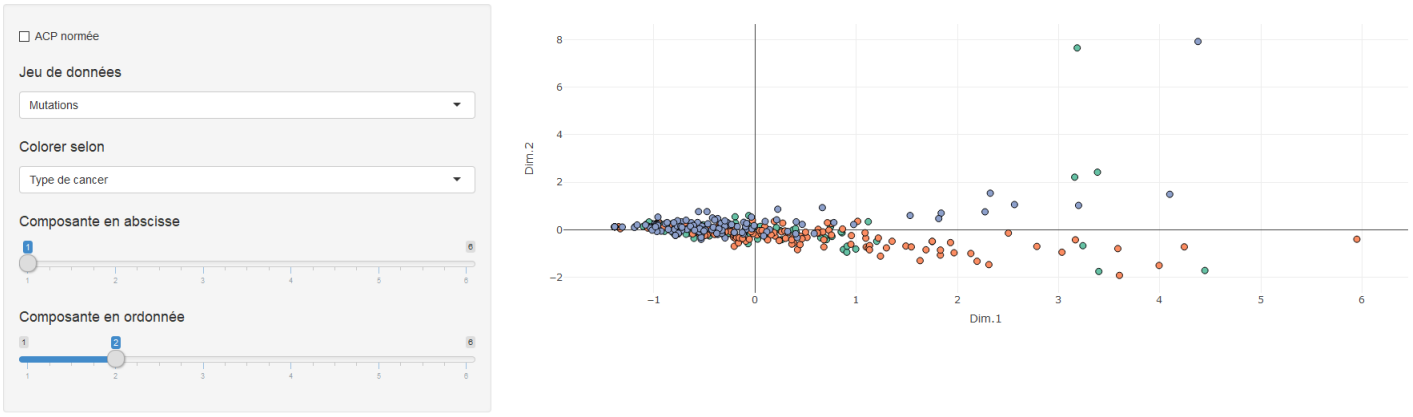

Figure 17: Options de notre application Shiny "ACP interactives"

Le menu à gauche permet de :

- Choisir de normer l'ACP.
- Choisir le jeu de données (la table Mutation, la table CN, ou la table complète).
- Colorer les points selon le type de cancer ou le sexe.
- Choisir la composante principale représentée en abscisse.
- Choisir la composante principale représentée en ordonnée.

Lorsque l'on place sa souris sur un point, on a accès à :

- Ses coordonnées sur les deux axes choisis.
- Son sexe ou son type de cancer selon l'option choisie.
- Le numéro de la lignée cellulaire.

Il est également possible d'interagir en cliquant directement sur le graphique, ou à l'aide du menu ci-dessous :



Figure 18: Capture d'écran des outils à disposition pour notre application Shiny d'ACP interactives

Cette boîte à outil est située en haut à droite de l'application. Les boutons permettent (de la gauche vers la droite) :

1. D'enregistrer le graphique.
2. D'agrandir une zone (zoom), possible également en formant un rectangle en cliquant avec la souris directement sur le graphique.
3. De se déplacer sur le graphique (en cas de zoom important par exemple).
4. De former une zone d'intérêt à main levée.
5. De zoomer d'un cran.
6. De dézoomer d'un cran.
7. De revenir à l'agrandissement initiale et de recentrer le graphique.

B. Établissement d'un seuil

Cette application dispose des mêmes outils que l'application d'ACP interactives. L'utilisation de l'outil zoom nous a permis de déterminer rapidement notre seuil visuellement.

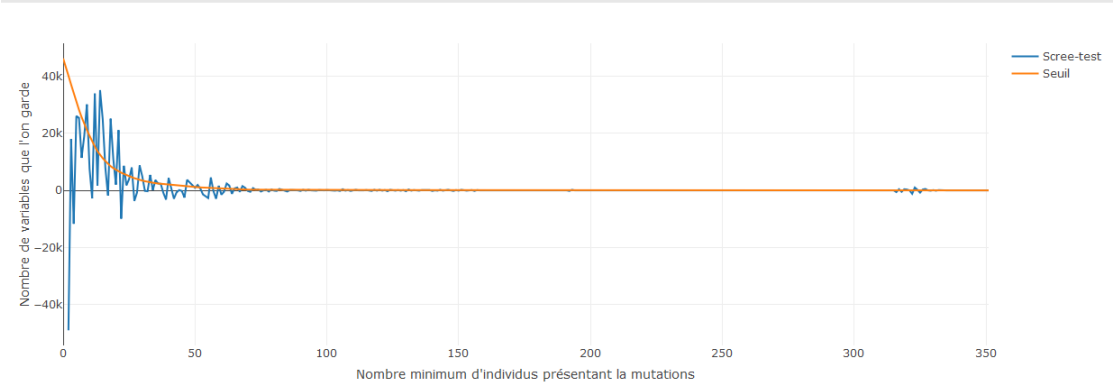

Figure 19: Capture d'écran de notre application Shiny d'aide à la détermination de seuil

On a également accès, pour chaque valeur du scree-test, au seuil correspondant et au nombre de variables gardées.

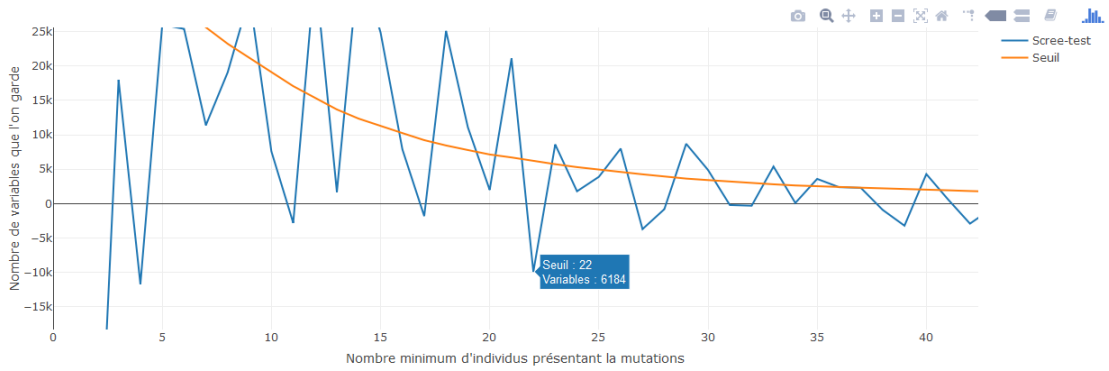

Figure 20: Agrandissement de notre application Shiny d'aide à la détermination de seuil

C. Optimisation des hyperparamètres

Cette application nous a permis de visualiser successivement plusieurs grilles d'hyperparamètres et de les parcourir, facilitant ainsi la sélection des hyperparamètres optimaux.

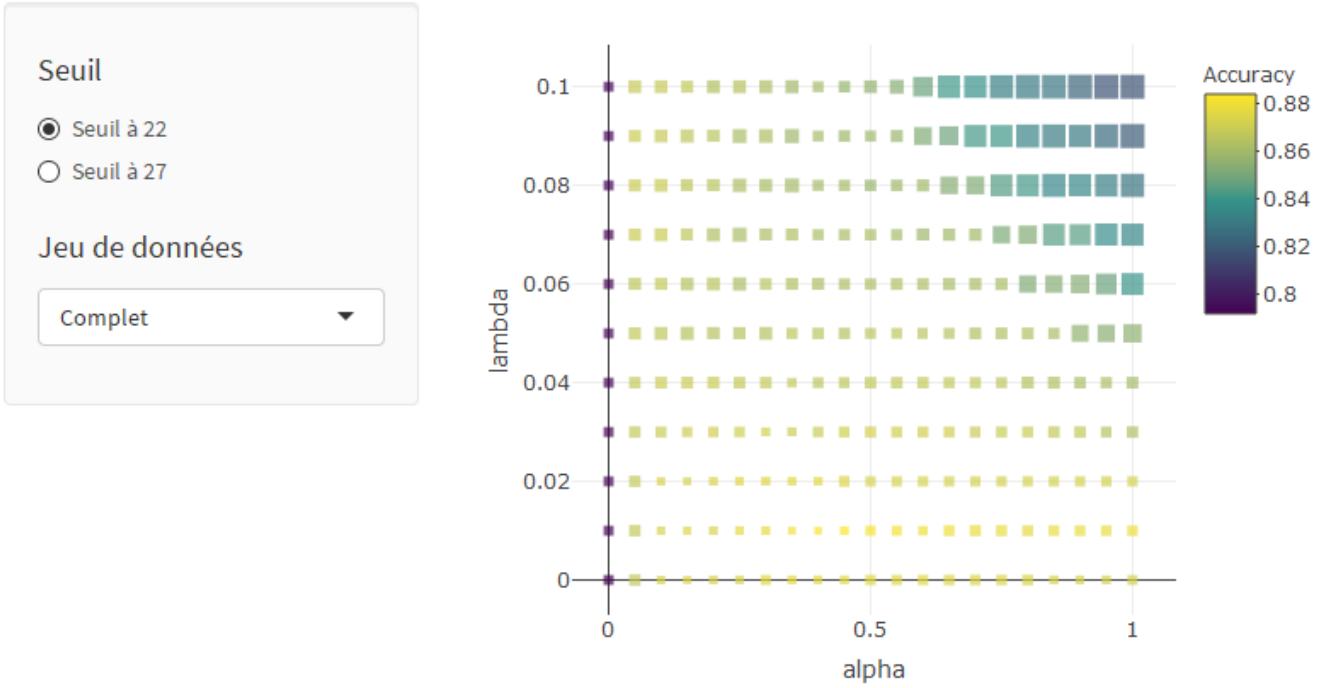

Figure 21: Capture d'écran de notre application d'aide à l'optimisation des hyperparamètres

On peut via le menu :

- Sélectionner le seuil à appliquer.
- Sélectionner le jeu de donnée du modèle à construire.

Lorsque l'on place sa souris sur un carré, on a accès aux hyperparamètres correspondant, ainsi qu'à son taux de bien classés et son Kappa.

D. Comparaisons de précision de modèles prédictifs

Cette application nous permet de comparer les précisions des différents modèles. L'utilisateur peut choisir la métrique du modèle (taux de bien classés ou Kappa).

La procédure utilisée est la suivante : 100 validations croisées sont réalisés sur chacun des modèles, à chaque répétition on obtient une valeur soit de taux de bien classés, soit de Kappa, puis la distribution des précisions obtenues est représentée par une boîte à moustache. Lorsque l'on place sa souris sur l'une des boîtes, on obtient la valeur du premier et du troisième quartile, de la médiane, du premier et du neuvième décile, ainsi que celle des éventuelles valeurs extrêmes.

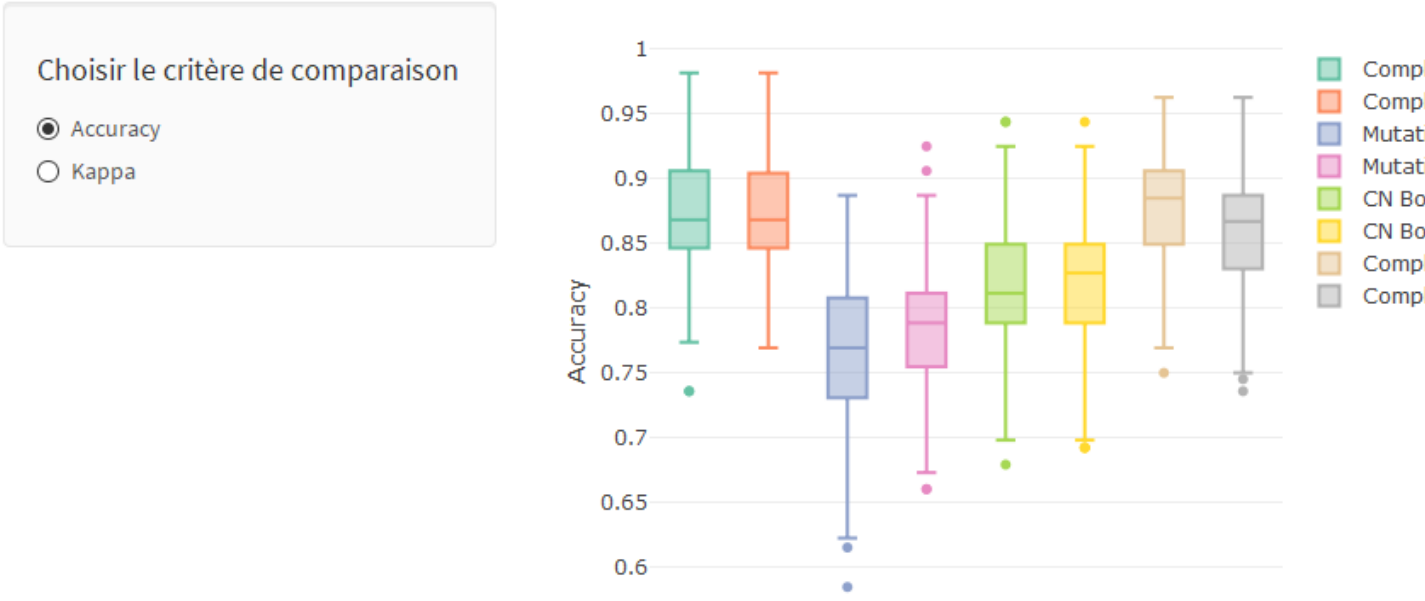

Figure 22: Capture d'écran de notre application de comparaisons de modèles prédictifs