PROJET TER PAR L'ÉQUIPE 5

RAPPORT INTERMÉDIAIRE 1 - QUESTIONS DE RECHERCHE

Thomas AYRIVIÉ, Mehdi BELKHITER, Jamila CHERKAOUI, Dina EL HIJJAWI, Magatte LO.

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Décembre 2023

Soumis comme contribution partielle
pour le cours TER TV15MI - TV25MI

Déclaration de non plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation:

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature : ***Thomas AYRIVIÉ, n°22000580***, Date : ***15/10/2023***.

Signature : ***Mehdi BELKHITER, n°21813356***, Date : ***15/10/2023***.

Signature : ***Jamila CHERKAOUI, n°22309204***, Date : ***15/10/2023***.

Signature : ***Dina EL HIJJAWI, n°22310171***, Date : ***15/10/2023***.

Signature : ***Magatte LO, n°22311161***, Date : ***15/10/2023***.

Préface

Dans le cadre de notre master 1 Miashs, les étudiants réalisent un **projet de Travaux d'Études et de Recherche** (TER) en lien avec un commanditaire. Ce présent document est un rapport intermédiaire dans lequel nous présenterons la définition de notre projet et nos questions de recherches regroupées dans quatre parties générales.

Sujet 6 : « Identification de voies de signalisation activées par des traitements de cancer du sein ».

Encadrement pédagogique :
Mme Sophie Lèbre, sophie.lebre@univ-montp3.fr (Institut Montpelliérain Alexander Grothendieck) avec Mathilde Robin (Institut de Recherche en Cancérologie de Montpellier, LIRMM Laboratoire d'informatique, de robotique et de microélectronique de Montpellier), Charles Lecellier (LIRMM) et Laurent Bréhélin (LIRMM)."

Composition du groupe :
Dina EL HIJJAWI, n°22310171, dina.el-hijjawi@etu.univ-montp3.fr. Coordinatrice.
Thomas AYRIVIÉ, n°22000580, thomas.ayrivie@etu.univ-montp3.fr.
Mehdi BELKHITER, n°21813356, mehidi.belkhiter@etu.univ-montp3.fr.
Jamila CHERKAOUI, n°22309204, jamila.cherkaoui@etu.univ-montp3.fr.
Magatte LO, n°22311161, magatte.lo@etu.univ-montp3.fr.

Table des matières

Chapitre 1

Contexte

1.1 Notre commanditaire

Le commanditaire de projet est une collaboration avec Mathilde Robin qui est ingénieure à l'IRCM ainsi que Laurent Bréhélin (chercheur CNRS au LIRMM), Charles Lecellier (chercheur à l'IGMM) et Sophie Lèbre est maître de conférences au sein du département Mathématiques et Informatique Appliquées (MIAp) de l'Université Paul Valéry Montpellier 3, membre de l'équipe Probabilité et Statistiques de l'Institut Montpelliérain Alexander Grothendieck (IMAG), co-responsable du Master MIASHS (Mathématiques et Informatique pour les Sciences Humaines et Sociales), master en alternance sur les 2 années M1-M2.

Le sujet du cancer en général fait partie du travail de recherche de Mme Lebre. Pour elle, l'augmentation du nombre de patients atteints de cancer, et pas seulement de cancer du sein, est préoccupante, ce sujet devrait donc intéresser tous ceux qui travaillent dans le domaine de la recherche ou en dehors.

Au-delà de cet aspect biologique, elle s'intéresse également à la condition difficile statistique à traiter (avec de fortes corrélations). Son travail lui permet de comparer différentes méthodes pour la sélection des variables quand la corrélation est élevée et proposer de nouvelles méthodes innovantes.

1.2 Les actualités

Selon les Nations Unies, le cancer est la deuxième cause de décès dans le monde entier et a fait 9,6 millions de morts en 2018, soit un décès sur six[1]. La recherche sur le cancer revêt donc une importance capitale pour les chercheurs et la société. Elle vise à réduire la mortalité en développant de nouvelles méthodes de dépistage et de traitement, améliore la qualité de vie des patients, permet une meilleure compréhension des causes sous-jacentes du cancer, stimule l'innovation médicale, réduit les coûts des soins de santé, lutte contre les disparités en matière de santé, favorise l'innovation et aide à prévenir les cancers évitables. En somme, elle contribue de manière significative à la santé, au bien-être et à l'économie, faisant d'elle une priorité majeure pour la société.

Le cancer du sein représente un défi de santé mondial majeur, touchant un grand nombre de personnes à l'échelle internationale. Au sein de LIRMM, les chercheurs mènent des recherches sur les différents types de cancer mais le jeu de données que nous devons exploiter dans ce projet est spécifiquement axé sur le cancer du sein.

[1] ONU, Cancer. https://www.who.int/fr/health-topics/cancer

Cette décision est motivée par la fréquence élevée de cette maladie, avec un nombre alarmant de 61 214 nouveaux cas enregistrés en 2023, selon les données de l'Institut national du cancer[2].

À l'heure d'écriture de ce rapport intermédiaire, nous sommes au mois d'octobre. Le mois d'octobre est devenu un mois de sensibilisation au cancer du sein à l'échelle mondiale, marqué par des campagnes emblématiques telles qu'Octobre Rose[3]. Ces initiatives de sensibilisation comprennent des récoltes de dons visant à soutenir la recherche sur le cancer du sein, à sensibiliser le public et à fournir un soutien aux patients et à leurs familles.

1.3 Choix du sujet

Il se trouve donc que le fait de travailler sur une petite partie de ce grand domaine d'étude peut intéresser davantage certains d'entre nous, les étudiants qui y travaillent, et les motiver à poursuivre la recherche à l'avenir, en s'ajoutant à la multitude de chercheurs qui tentent d'approfondir leur compréhension du cancer et de la manière de le traiter.

[2]Institut National du Cancer, le cancer du sein. https://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Epidemiologie-des-cancers/Les-cancers-les-plus-frequents/Cancer-du-sein

[3]Ars Bretagne, Octobre Rose. https://www.bretagne.ars.sante.fr/octobre-rose-un-mois-pour-sensibiliser-au-depistage-du-cancer-du-sein-0

Chapitre 2

Définition du sujet

Le sujet de recherche se concentre sur l'identification des voies de signalisation activées en réponse aux traitements du cancer du sein. Au fil du temps, les cellules cancéreuses ont la capacité de s'adapter et de développer des mécanismes de résistance, y compris face à la chimiothérapie. Comprendre ces mécanismes est essentiel pour améliorer les stratégies de traitement.

Pour ce faire, nous utilisons une analyse approfondie des séquences génétiques afin de déterminer quelles combinaisons de nucléotides obtiennent les scores les plus élevés dans les modèles de prédiction de la classe Y. Ces scores reflètent la similitude ou la pertinence d'une séquence donnée par rapport à un modèle de référence.

Préalablement, un modèle de régression a été développé pour expliquer l'activité des gènes en utilisant uniquement la séquence d'ADN[1]. Cependant, nous visons maintenant à mettre en place un nouveau modèle de classification, pour qu'on puisse construire un modèle capable de prédire quels gènes sont différentiellement exprimés (soit actifs soit inhibés) en réponse à un traitement donné. Les détails seront abordés ultérieurement au cours du projet.

[1] Hsiaowang Chen, Series GSE130787 de NCBI.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130787

Chapitre 3

Données et traitements

Pour mener à bien notre recherche, nous utilisons la base de données Jaspar[1], une ressource spécialisée dans les matrices de séquences d'ADN liées aux sites de liaison des facteurs de transcription. Ces matrices, également appelées matrices de poids de position (PWM), offrent une représentation numérique des motifs de liaison des facteurs de transcription qui sont des protéines codées par des gènes ADN qui sont responsables de l'activation ou l'inhibition d'autres gènes. Jaspar fournit des informations détaillées sur la fréquence de chaque base (A, C, G, T)[2] à chaque position au sein de ces motifs de liaison. Cette base de données est essentielle pour développer des modèles plus précis dans le but de mieux comprendre la régulation génique et les réponses cellulaires aux traitements du cancer du sein et elle est aussi essentielle pour identifier des sites de liaisons potentiels au sein des séquences et construire un ensemble de variables explicatives à partir des données existantes afin d'améliorer la performance et la capacité prédictive du modèle.

En complément, nous utilisons pareillement la base de données Series GSE130787 GEO[3] (Gene Expression Omnibus) qui est chargée dans RStudio grâce à GEOquery[4]. GEO stocke une grande variété de données d'expression génique provenant de diverses expériences et technologies, incluant les puces à ADN et les séquences ARN. Les chercheurs déposent leurs données d'expression génique dans GEO, permettant ainsi à d'autres scientifiques d'y accéder, de les partager et de les analyser. Cette base de données facilite l'exploration et la comparaison des profils d'expression génique dans différents contextes biologiques et expérimentaux.

Jaspar nous aide à analyser les motifs de liaison des facteurs de transcription, tandis que Geoquery nous permet d'accéder à une vaste gamme de données d'expression génique pour une exploration approfondie des mécanismes biologiques et des réponses aux traitements.

[1] Wasserman Lab. http://www.cmmt.ubc.ca/wasserman-lab/

[2] Les bases azotées symbolisées par A, C, G, T revoient dans le mémoire à A pour Adénine, C pour Cytosine, G pour Guanine, T pour Thymine.

[3] Hsiaowang Chen, Series GSE130787 de NCBI.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130787

[4] Davis S, Meltzer P (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." Bioinformatics, 14, 1846–1847. https://bioconductor.org/packages/release/bioc/html/GEOquery.html

Chapitre 4

Méthodologie

Le but principal est de déterminer les variables, ou motifs, qui jouent un rôle discriminant entre divers gènes, les classifiant en catégories telles qu' actifs, inactifs et inhibés. L'objectif du projet est aussi de décrypter les facteurs de transcription impliqués dans le développement du cancer du sein . Nous cherchons également à approfondir notre compréhension du fonctionnement du traitement, en mettant l'accent sur les cellules qui réagissent le mieux à ce traitement et à identifier les individus pour lesquels ce traitement est le plus efficace. Sur le plan statistique, notre démarche implique de comparer différentes méthodes statistiques pour la sélection de variables, surtout lorsque deux variables sont corrélées. Cela vise à élucider les nuances entre ces méthodes et à déterminer la plus adaptée pour notre analyse, assurant ainsi la précision de nos conclusions dans cette étude complexe.

Pour constituer un jeu de données, nous utilisons Fimo de la suite Meme qui sert à rechercher les occurrences individuelles de motifs dans les séquences. Les données Jaspar y sont intégrées avec les données de séquences des régions promotrices des gènes[1] au format fasta. La sortie de ce logiciel calcule les scores maximum (sous la forme de p-valeur).

Grâce au package Tidyverse de RStudio, ces données sont ensuite regroupées dans une matrice qui croise les gènes de chaque séquence et leurs motifs avec des p-valeurs associées pour chaque croisement.

Une fois cette matrice créée, nous pourrons faire une jointure avec les données GEOquery NCBI (Patientes suivies en cancer du sein). En choisissant correctement des valeurs seuils et grâce au logarithme de 10 du Fold Change nous pourrons construire les classes Y (up, down, neutre).

À partir de là, nous serons dans un problème de classification. Nous devrons rechercher des groupes de variables corrélées pour trouver un modèle explicatif qui utilise la structure de corrélation des variables. Pour cela, nous utiliserons le package glmnet et le package mlgl[2], modèle Lasso[3] et elastinet. Nous approfondirons les

[1] LÈBRE Sophie, ROBIN Mathilde. https://upvdrive.univ-montp3.fr/s/kKRzYDzzcXe32dX

[2] Grimonprez, Q., Blanck, S., Celisse, A., and Marot, G. (2023). MLGL: An R Package Implementing Correlated Variable Selection by Hierarchical Clustering and Group-Lasso. Journal of Statistical Software, 106(3), 1–33. https://doi.org/10.18637/jss.v106.i03 3 .Meinshausen N. and Bühlmann P (2010) Stability selection Journal of the Royal Statistical Society: Series B (Statistical Methodology) https://doi.org/10.1111/j.1467-9868.2010.00740.x

[3] Tibshirani R. (Lecture, Spring 2017) Sparsity, the Lasso, and Friends. https://upvdrive.univ-montp3.fr/s/jakYqikj52WZ7Ci

recherches avec les forêts aléatoires dans Random Forest dans R[4] et en indiquant les forces et faiblesses du modèle Lasso. Pour un modèle de classification, la précision, le rappel, le F1-score, et l'aire sous la courbe ROC (AUC) peuvent être utilisés pour évaluer la performance.

Nos variables sont les motifs de séquence (représentés par motif_alt_id) : Ils décrivent les motifs d'ADN qui sont reconnus par les facteurs de transcription. Les p-valeurs associées à chaque motif : Elles donnent une mesure de la signification statistique de l'occurrence d'un motif dans une séquence d'ADN. Les identifiants de séquence (sequence_name): Ils identifient de manière unique chaque séquence d'ADN analysée.

En parallèle de celà, nous allons faire une analyse exploratoire des données (Analyse des composantes principales, visualisation de la structure des données, qualification des variables...). Et nous allons utiliser les packages RStudio avec de petits jeux de données pour comprendre leur utilisation correcte.

Pour une analyse d'exploration (comme l'ACP), la variance expliquée par chaque composante principale peut être utilisée comme métrique. D'autres métriques comme la régression sera également abordée et utiliser et si on a du temps l'objectif sera la mise en place d'un modèle de machine learning (pour prédire pour les prochains patients).

[4]Genuer, R., and Poggi, J. (2020). Random Forests with R. Dans Use R ! https://doi.org/10.1007/978-3-030-56485-8

Annexe : Questions de recherche plus spécifiques

Questions de recherche plus spécifiques qui peuvent être résolues en utilisant des nombres et du texte.

1. **Combien (régression) ?** Quelle est l'importance relative de chaque motif dans la régulation de l'expression génique ? *Quelle est l'importance relative de chaque motif dans la régulation de l'expression génique ?*
2. **Quelle catégorie (classification) ?** Un motif donné est-il associé à une augmentation, une diminution ou aucune modification de l'expression génique ? *Un motif donné est-il associé à une augmentation, une diminution ou aucune modification de l'expression génique ?*
3. **Quel groupe (clustering) ?** Existe-t-il des groupes de motifs qui ont des comportements similaires en termes d'effets sur l'expression génique ? *Existe-t-il des groupes de motifs qui ont des comportements similaires en termes d'effets sur l'expression génique ?*
4. **Est-ce anormal (détection d'anomalies) ?** Y a-t-il des motifs qui se comportent de manière atypique par rapport à d'autres motifs similaires ? *Y a-t-il des motifs qui se comportent de manière atypique par rapport à d'autres motifs similaires ?*
5. **Quelle option choisir (recommandation) ?** Quels motifs devraient être ciblés pour des études plus approfondies en fonction de leur pertinence pour la régulation de l'expression génique ? *Quels motifs devraient être ciblés pour des études plus approfondies en fonction de leur pertinence pour la régulation de l'expression génique ?*