PROJET TER PAR L'ÉQUIPE 5

PROJET TER PAR L'ÉQUIPE 5 RENDU INTERMÉDIAIRE 4 : REVUE DE LA LITTÉRATURE

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 : ***06/12/2023***.

Signature : ***Mehdi BELKHITER, n°21813356***, Date : ***06/12/2023***.

Signature : ***Jamila CHERKAOUI, n°22309204***, Date : ***06/12/2023***.

Signature : ***Dina EL HIJJAWI, n°22310171***, Date : ***06/12/2023***.

Signature : ***Magatte LO, n°22311161***, Date : ***06/12/2023***.

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 ainsi que la revue de la littérature.

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 global du projet

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. 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. 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.

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. 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.

Chapitre 2

Liste de mots clés

Cancer du sein : Un cancer signifie la présence de cellules anormales qui se multiplient de façon incontrôlée. Dans le cas du cancer du sein, les cellules peuvent rester dans le sein ou se répandre dans le corps par les vaisseaux sanguins ou lymphatiques. La plupart du temps, la progression d'un cancer du sein prend plusieurs mois et même quelques années.Le cancer du sein est le plus diagnostiqué chez les femmes dans le monde. Une femme sur 9 sera atteinte de ce cancer au cours de sa vie et 1 femme sur 27 en mourra(source : passeport santé)

Base de données Jaspar : une base de données qui répertorie les motifs spécifiques de séquences d'ADN auxquels se lient les protéines. Elle permet de comprendre comment certaines protéines interagissent avec l'ADN en identifiant les séquences génétiques qu'elles ciblent. Cela aide à décoder le langage génétique et à mieux comprendre la régulation des gènes.

Base de données GEO (Gene Expression Omnibus) : une base de données (GEO) qui stocke une variété de données d'expression génique, facilitant l'exploration et la comparaison des profils d'expression génique dans différents contextes biologiques et expérimentaux.

Fichier FASTA : Un format de fichier texte couramment utilisé pour représenter des séquences d'acides nucléiques (ADN) ou des séquences de protéines. Chaque fichier FASTA est composé d'une ligne d'en-tête suivie de la séquence correspondante.

MEME (Fimo) : Fimo est une référence à Find Individual Motif Occurrences dans le contexte du logiciel MEME Suite. Fimo est utilisé pour rechercher des occurrences de motifs particuliers dans une séquence génomique.

Génes : unité d'information héréditaire située sur l'ADN (acide désoxyribonucléique) d'un organisme. Les gènes contiennent les instructions nécessaires à la synthèse des protéines, qui sont essentielles à la structure, à la fonction et à la régulation des cellules et des organismes.

ADN : Molécule biologique qui porte l'information génétique dans les cellules vivantes. Il se présente sous la forme d'une double hélice, composée de deux brins enroulés autour l'un de l'autre. Chaque brin est constitué d'une séquence spécifique de quatre bases azotées.

Bases azotées (A.C.G.T) : Adénine (A), Cytosine (C), Guanine (G), Thymine (T) : Les bases azotées sont des composés organiques qui constituent les unités de base de l'information génétique dans l'ADN.

Matrice de Position de Poids (PWM) : Représentation mathématiques des motifs de liaison des facteurs de transcription.

P-valeurs : Mesures statistiques utilisées dans l'analyse pour évaluer la significativité des motifs et des séquences ADN.

Classes Up, Down, Neutre : Classes utilisées pour indiquer une augmentation de l'expression génique, l'absence ou la diminution.

Logarithme 10 du Fold Change : Une mesure de la différence relative entre deux conditions expérimentales, souvent utilisée pour évaluer l'expression génique différentielle. Elle subit une transformation logarithmique pour stabiliser les variations et faciliter l'interprétation des changements relatifs.

Jointure : Opération dans laquelle des données de deux sources sont combinées en fonction de certains critères, mais qui peut entraîner la suppression de certaines données.

Analyse exploratoire des données : Approche statistique visant à découvrir des tendances, des motifs et des relations dans les données, souvent utilisée pour comprendre la structure sous-jacente des données.

Dont l'ACP (Analyse des composantes principales) : Technique d'analyse multivariée qui transforme les données originales en un nouvel ensemble de variables non corrélées appelées composantes principales, souvent utilisée pour réduire la dimensionnalité des données.

Chapitre 3

Littérature utilisée

Livres

Genuer Robin et al. Random Forests with R. 1st ed. 2020. Cham: Springer International Publishing, 2020. Print. Lien : https://github.com/DinaLH/-TER-2023-2024-Trai
blob/main/Bibliographie/Random_Forests_with_R_Robin_GenuerJean-Michel... _Z-Library.pdf

Thèses

Bessiere, Chloé. Etude des éléments régulateurs de l'expression des gènes chez l'humain. Montpellier : 2018. Université de Montpellier : thèse de doctorat, Biologie Santé, sous la direction de Lecellier, Charles-Henri . Disponible sur https://ged.biu-montpellier.fr/florabium/jsp/nnt.jsp?nnt=2018MONTT099. Les scores de motifs sont décrits dans les équations (1.1) et (1.2). Lien : https://github.com/DinaLH/-TER-2023-2024-Traitements-de-Cancer-du-Sein/blob/main/Bibliographie/73253_BESSIERE_2018_archivage.pdf

Documents

Ryan Tibshirani, Larry Wasserman. Sparsity, the Lasso, and Friends Statistical Machine Learning, Spring 2017. Lien : https://github.com/DinaLH/-TER-2023-2024-Traitements-de-Cancer-du-Sein/blob/main/Bibliographie/Tibshirani%2C%20Ryan%20-%20Sparsity%2C%20the%20Lasso%2C%20and%20Friends.pdf

Diaporama

Laurent Bréhélin, Sophie Lèbre, Charles Lecellier. Statistical modeling and inference to identify DNA sequence elements involved in transcription regulation., Présentation donnée à la Journée Statistique et Sciences de la Santé, Lille 27 Juin 2022. Lien : https://github.com/DinaLH/-TER-2023-2024-Traitements-de-Cancer-du-S
blob/main/Bibliographie/presentation_projet.pdf

Articles

Quentin Grimonprez, Samuel Blanck, Alain Celisse, Guillemette Marot. MLGL: An R Package Implementing Correlated Variable Selection by Hierarchical Clustering and Group-Lasso. Journal of Statistical Software. March 2023, Volume 106, Issue 3. Lien : https://github.com/DinaLH/-TER-2023-2024-Traitements-de-Cancer-du-S
blob/main/Bibliographie/MLGL%20An%20R%20Package%20Implementing%20Correlated%20Variable%20Selection.pdf

Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. Nature reviews genectics, april 2004, volume 5. Lien : https://github.com/DinaLH/-TER-2023-2024-Traitements-de-Cancer-du-Sein/blob/main/Bibliographie/Wasserman_Sandelin_2004.pdf

Chloé Bessière, May Taha, Florent Petitprez, Jimmy Vandel, Jean-Michel Marin, Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier. Probing instructions for expression regulation in gene nucleotide compositions. Plos Computational Biology, 2 janvier 2018. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005921âŃĚLien : https://github.com/DinaLH/-TER-2023-2024-blob/main/Bibliographie/Bessie_re%20et%20al.%20-%202018%20-%20Probing%20instructions%20for%20expression%20regulation%20in%20.pdf