**Data Proposal: Baptiste Rivoirard, Amine Imarraine, Paul Cloarec—Pioffet**

Le cancer du sein est une maladie caractérisée par la croissance incontrôlée de cellules mammaires qui forment alors des tumeurs. Le cancer du sein est le cancer le plus fréquent des cancers féminins, il représente 33 % des cancers chez la femme. Ce cancer est associé des mutations spécifiques qui influent sur le caractère invasif du cancer et son traitement1,2. Identifier les interactions entre variables cliniques et mutations génétiques permet une meilleure compréhension des mécanismes moléculaires conduisant à la croissance des tumeurs.

Nous étudierons un sous ensemble du dataset METABRIC (Breast Cancer Gene Expression Dataset) qui est public et regroupe des données cliniques et génétiques (expression génique et mutation de certains gènes). Plus précisément, nous souhaitons intégrer à la fois des données cliniques (par exemple : overall\_survival\_months, overall\_survival, death\_of\_cancer, etc.) et des informations sur l’évolution des tumeurs (telles que tumor\_stage, tumor\_size, etc.), ainsi que des données sur les mutations génétiques impliquées dans le cancer du sein, selon la littérature (telles que BRCA1/2, TP53, et d’autres). Il existe de nombreuses études utilisant ce dataset ainsi que des notebooks d’analyses pour la classification des sous-groupes de patients du cancer du sein3.

Dans notre projet nous chercherons à identifier les liens entre variables cliniques et les mutations génétiques sous forme de réseaux biologiques et voir comment ses derniers influencent la survie et l’évolution des tumeurs. Autrement dit nous chercherons à déterminer les éléments cliniques et moléculaires prépondérants dans les bons ou mauvais pronostiques de cancer et in fine proposer une liste de biomarqueurs utiles et pertinents pour diagnostiquer le cancer et son type.

Pour cela, nous nous choisirons d’abord les features à inclure dans notre analyse grâce à la littérature (i.e. les variables cliniques que la communauté scientifique sait informative, les mutations dans les gènes connus pour être lié au cycle cellulaire et la progression tumorale4). Nous réaliserons ensuite une étape de nettoyage des données, a cette étape nous pourront éventuellement imputer des données s’il y a des drop-outs ou des valeurs manquantes et que nous jugeons l’imputation nécessaire. Ensuite viendra la normalisation des données de certaines variables continues (comme tumor\_size) avec des méthodes telles que la standardisation. Il nous faudra ensuite reconstruire le réseau à l’aide d’algorithmes comme MIIC (en mode standard) car il permet la reconstruction des réseaux pondérés et dirigés à partir de jeux de données mixtes 5. Enfin, la dernière étape consistera en une analyse de notre réseau, de ses caractéristiques (comme sa modularité) et de celles des nœuds le composant (leur centralité, leur transitivité…) et en une confrontation des résultats obtenus et des conclusions qui en ont été tirées avec la littérature.

Ce projet nous semble réalisable car, bien que le dataset soit grand, nous comptons prendre un sous-ensemble avec un nombre limité de features. La majorité des features a déjà été sélectionnée, mais nous comptons tester plusieurs tailles de sous-dataset afin d’ajuster la taille de manière à obtenir un maximum d’informations et de liens intéressants entre les features dans notre réseau tout en préservant un équilibre avec la puissance de calcul et la lisibilité du graph.   Nous espérons retrouver des résultats intéressants, car des articles ont déjà associé des mutations de certains gènes tels que BRCA avec un mauvais pronostic du cancer pour certaines populations 6.

**Bibliographie :**

1. Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. Ann Oncol. 2015 Jul;26(7):1291-9. doi: 10.1093/annonc/mdv022. Epub 2015 Jan 20. PMID: 25605744; PMCID: PMC4478970.
2. Wunderle M, Gass P, Häberle L, Flesch VM, Rauh C, Bani MR, Hack CC, Schrauder MG, Jud SM, Emons J, Erber R, Ekici AB, Hoyer J, Vasileiou G, Kraus C, Reis A, Hartmann A, Lux MP, Beckmann MW, Fasching PA, Hein A. BRCA mutations and their influence on pathological complete response and prognosis in a clinical cohort of neoadjuvantly treated breast cancer patients. Breast Cancer Res Treat. 2018 Aug;171(1):85-94. doi: 10.1007/s10549-018-4797-8. Epub 2018 May 3. PMID: 29725888.
3. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S; METABRIC Group; Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, Aparicio S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012 Apr 18;486(7403):346-52. doi: 10.1038/nature10983. PMID: 22522925; PMCID: PMC3440846.
4. Feng Y, Spezia M, Huang S, Yuan C, Zeng Z, Zhang L, Ji X, Liu W, Huang B, Luo W, Liu B, Lei Y, Du S, Vuppalapati A, Luu HH, Haydon RC, He TC, Ren G. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes Dis. 2018 May 12;5(2):77-106. doi: 10.1016/j.gendis.2018.05.001. PMID: 30258937; PMCID: PMC6147049.
5. Cabeli, V. *et al.* (2020) ‘Learning clinical networks from medical records based on information estimates in mixed-type data’, *PLOS Computational Biology*, 16(5). doi: 10.1371/journal.pcbi.1007866.
6. Tinterri C, Di Maria Grimaldi S, Sagona A, Barbieri E, Darwish S, Bottini A, Canavese G, Gentile D. Comparison of Long-Term Oncological Results in Young Women with Breast Cancer between BRCA-Mutation Carriers Versus Non-Carriers: How Tumor and Genetic Risk Factors Influence the Clinical Prognosis. Cancers (Basel). 2023 Aug 19;15(16):4177. doi: 10.3390/cancers15164177. PMID: 37627205; PMCID: PMC10452863.