Building Mutational Signatures in Cancer using Deep Bayesian Neural Networks

A Deep Dive into Cancer Mutational Signatures

Jan Alfonso Busker

*352905 - j.a.busker@st.hanze.nl*

*Hanzehogeschool Groningen*

*Missing in abstract:*

* *Most important results*
* *Conclusion and future vision*

Abstract

Cancer, characterized by uncontrolled cell growth, accumulate mutations. The mutations that occur in the context of cancer development are a result of exposure to various DNA-damaging processes and accumulate throughout life. The sources of these DNA-damaging processes include both endogenous and exogenous factors. These genetic variations result in unique "mutational signatures" within the DNA sequence. These mutational signatures provide valuable information about the causes of the genetic variations. Additionally, they offer insights into the degree of cellular exposure to specific mutagenic events. By looking at the direct context of the mutations, we can infer the way these mutations have arisen.

This project aims to refine the statistical model and the current representation of mutations by building mutational signatures of cancer using deep bayesian neural networks. Additionally, there is a plan to expand the representation to capture more context. Increasing the context involves subdividing mutations as seen in fig 1. By looking at an extra nucleotide on each side. The aim of this division is to potentially reveal recurring contextual imprints associated with the nucleotides on the left and right of the mutations. This will be achieved by using Latent Dirichlet Allocation (LDA).

The methodology includes curating and quality controlling variant calling samples, replicating mutational signatures using Non-negative Matrix Factorization (NMF), and determining priors for the expanded mutation representation.

The significance of this project lies in advancing our understanding of cancer mutational signatures, which has implications for cancer diagnosis, treatment, and prevention.

Abbreviations

Latent Dirichlet Allocation (LDA)

Non-negative Matrix Factorization (NMF)

Organisation

This assignment is given by UMCG. The Department of Epidemiology is a major driving force in initiating and conducting life course research and is instrumental to the clinical research within the UMCG’s main theme of Healthy Ageing. The research group ‘Medical statistics and decision making’ is part of the Department of Epidemiology. They focus on developing methods for statistical modeling in clinical and epidemiological studies and analyzing large cohort data. And develop decision analysis techniques to support benefit-risk assessments of medicines and medical decision-making. External guidance is provided by Prof. G.A. Lunter and Dr. Hylke C. Donker.

Introduction

Cancer is a genetic illness characterized by uncontrolled cell proliferation. Cancer cells develop numerous abilities to support this expansion, mostly through genomic alterations. Throughout life, these mutations accumulate as a result of exposure to DNA-damaging mechanisms from both endogenous and external causes [1-3].

It is becoming increasingly clear that a wide range of mutational pathways contribute to the mutation landscape of cancer. Cancer genomes contain somatic mutations acquired during the normal cell cycle as well as those triggered by cancer-related aberrations of DNA maintenance machinery such as mismatch repair or by carcinogenic exposures such as tobacco smoking, ultraviolet light, and replication stress. Each of these processes frequently results in a particular pattern of changes, known as the mutational signature [4,5].

The study of mutational changes, specifically the modification of a single letter within the surrounding context, has been the focus of numerous studies. The utilization of this methodology has yielded a meticulously selected collection of mutational signatures as provided by COSMIC [6]. Several approaches have been developed to determine the signatures and infer the attributions.

MutationalPatterns is an R/Bioconductor tool for analyzing single and double base substitutions, as well as small insertions and deletions. It additionally allows the analysis of regional mutation spectra and the discovery of strand asymmetry occurrences [7]. And decompTumor2Sig is an R package that can decompose a single tumor genome into a series of Alexandrov- or Shiraishi-type signatures. These signatures represent distinct patterns of somatic mutations in cancer genomes [6]. This enables quantification of the contribution of several mutational processes to the somatic mutations found in a certain tumor [8].

Despite the advancements in understanding mutational signatures and the development of tools to analyze them. There are still several gaps in current knowledge. Most studies have only looked at mutations that include a single letter of context to the left and right. This has hampered the ability to distinguish between other DNA damage sources., Such as ABOPEC3A and ABOPEC3B which require at least one more context letter [9]. These enzymes plays a role in DNA editing and mutation processes, with implications in immune defense and cancer development [9]. Large sets of tumor samples are required for the initial discovery and definition of mutational signatures, which are not always available [8]. Some packages show signature overfitting when determining the contribution of known patterns to a sample, resulting in a disproportionate amount of signatures being assigned. It also only allows for the analysis of spectra for mutations in the entire genome, making studying the role of specific genomic elements challenging [7].

Researchers would be able to better distinguish between different mutation origins if they considered additional context around mutations. This could lead to a better understanding of the mutational processes that contribute to cancer. In addition, techniques for mutational signature analysis should be improved. Addressing present techniques' weaknesses such as overfitting and the inability to study specific genetic components. This would significantly improve their accuracy and usefulness. This could allow for more in-depth and advanced studies of mutational phenomena.

Objective

The previously stated aim will be accomplished of improving the model in order to evaluate mutational signatures that are hierarchically linked, as well as augmenting the context size by incorporating an additional context letter.

The study will involve the curation and quality control of samples and mutations, the replication of mutational signatures through the utilization of NMF, and the subsequent comparison of the outcomes with signatures derived from bayesian neural networks consisting of one and two layers.

Furthermore, the study aims to establish prior probabilities for the expanded mutation representation and compare the identified signatures of higher dimensions with the existing 96 feature representation.

Theory

Mutational signatures are a critical aspect of cancer research, providing a physiological readout of a cancer's biological history [10]. They are the imprints left on the genome by numerous endogenous and exogenous mutagenesis events that have happened throughout a human being's lifetime. These processes can lead to base substitutions, insertions and deletions, or structural modifications, each of which leaves a distinct pattern or signature [10]. SBS7a is a mutational signature with a specific pattern that is easy to identify. C>T mutations and TT dinucleotides on both flanks characterize it [13]. This characteristic is linked to skin malignancies in sun-exposed locations and is most likely caused by UV radiation exposure [12]. Mutational signatures have proven to be a valuable resource in understanding cancer treatment and prevention. For instance, they have been instrumental in studying the molecular processes of DNA damage, DNA repair, and DNA replication [11]. For example, several mutational signatures, such as SBS4 and SBS92, are linked to the same etiology, as are SBS1 and SBS5, both clock-like signatures [4].

*Fig 1: This formula expresses the summation over all possible combinations of nucleotides on the left (x) and right (y) of the mutation A[C>A]A. The approximation of the original mutation A[C>A]A by considering all possible nucleotide combinations in the expanded context.*

The context of mutations is crucial in understanding mutational signatures. Increasing the context size is a strategy used to capture complex phenomena. Current single base substitution signatures are based on one flanking nucleotide left and right of the substitution, but this is not enough context to discriminate [9]. As seen in fig 1,this subdivision ensures a more detailed representation, considering the nucleotides on the left and right of the mutations.

References

[1] Hanahan, D., & Weinberg, R. (2000). "The Hallmarks of Cancer." Cell, 100(1), 57-70.

[2] Hanahan, D., & Weinberg, R. (2011). "Hallmarks of Cancer: The Next Generation." Cell, 144(5), 646-674.

[3] Hanahan, D. (2022). "Hallmarks of Cancer: New Dimensions." Cancer Discovery, 12(1), 31-46.

[4] Alexandrov, L., et al. Stratton, M. R. (2020). "The repertoire of mutational signatures in human cancer." Nature, 578(7793), 94-101.

[5] Huang, X., Wojtowicz, D., & Przytycka, T. M. (2018). "Detecting presence of mutational signatures in cancer with confidence." Bioinformatics, 34(2), 330-337.

[6] COSMIC. (n.d.). Wellcome Sanger Institute. https://cancer.sanger.ac.uk/cosmic.

[7] Kim, Y., Leiserson, M. D. M., Moorjani, P., Sharan, R., Wojtowicz, D., & Przytycka, T. M. (2021). "Mutational Signatures: From Methods to Mechanisms." Annual Review of Biomedical Data Science, 4, 189-206.

[8] Krüger, S., & Piro, R. M. (2019). "decompTumor2Sig: identification of mutational signatures active in individual tumors." BMC Bioinformatics, 20(4), 152.

[9] Chan, K., et al. (2015). "An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers." Nature Genetics, 47(9), 1067-1072.

[10] Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). "Mechanisms underlying mutational signatures in human cancers." Nature Reviews. Genetics, 15(9), 585-598.

[11] Steele, C. D., Pillay, N., & Alexandrov, L. B. (2022). "An overview of mutational and copy number signatures in human cancer." The Journal of Pathology, 257(4), 454-465.

[12] Jialong, L., Jing, B., Tao, P., Haozhe, Z., Yueying, G., Jing, G., Qi, X., Juan, X., Yongsheng, L., & Xia, L. (2022). "Clinical and genomic characterization of mutational signatures across human cancers." International Journal of Cancer, 152(8), 1613-1629.

[13] Machado, H. E., etal. (2022). "Diverse mutational landscapes in human lymphocytes." Nature, 608(7924), 724-732.