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)

Single Based Substitution (SBS)

Variant Call Format (VCF)

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.

I. 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 play 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 number 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.

II. 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.

III. 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) plus the right (y) of the probability of mutation A[C>A]A of SBS1. The approximation of the original probability of mutation A[C>A]A of SBS1by 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 SBS signatures are based on max. two 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.

IV. Materials

The method of analysis is based on an in-depth knowledge of mutational signatures, drawing insights from the COSMIC Mutational Signatures reference collection. The GenomeSigInfer tool (<https://github.com/AlfonsoJan/GenomeSigInfer>) was developed for processing and analysis. DeepBayesMutSig (<https://github.com/AlfonsoJan/DeepBayesMutSig>) has detailed analysis and results. All of the libraries and tools used are listed in appendix table 1.

The study "The repertoire of mutational signatures in human cancer" [4] produced Variant Call Format (VCF) files that include both somatic and germline variant calls. VCF files are tab-delimited text files that contain meta-information lines, a header line, and then data lines each containing information about a position in the genome. The study examined 2,658 complete cancer genomes and their corresponding normal tissues from 38 different types of tumors, as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG). The resulting datasets can be accessed on Synapse at the following link: ([https://www.synapse.org/#!Synapse:syn11801870](https://www.synapse.org/%23!Synapse:syn11801870)). The VCF files underwent processing to generate Single Based Substitution (SBS) files.

SBS are a type of mutation where one nucleotide is replaced by another in the DNA sequence. SBS files as ‘datasets’ containing detailed count data for single base substitutions within the genomic sequences of the mutation in the VCF file. For quality assurance and validation, the ‘WGS\_Other.96.csv’ and ‘WGS\_PCAWG.96.csv’ played a pivotal role as a benchmark dataset, by ensuring that it’s output of the GenomeSigInfer tool precisely mirrored the 2 \*96.csv files.

*Fig 2: This formula the expresses the SBS files of non-negative integers with dimension K × G, where K is the number of mutation types and G is the number of samples.*

The COSMIC mutational signatures are a reference collection that provides the mutational characteristics of each signature. Where each row represents the likelihood for each mutation signature. And each column is the likelihood of each signature mutation [6]. Each row or column adds up to 1.

V. Methods

The SigProfiler framework has been systematically designed to analyze mutational signatures in genomic data. The methodology is based on the use of NMF, which is essential in the decomposition of high-dimensional mutational matrices into non-negative basis matrices [4]. The framework minimizes a generalized Kullback-Leibler divergence that is strictly bound for non-negativity in each iteration of NMF [4]. This procedure reveals unique mutational signatures and their contributions to each sample [4]. Furthermore, SigProfiler employs a hierarchical de novo extraction technique, which improves the precision of its analysis [4]. This strategic approach allows for a thorough examination of mutational patterns, allowing for the detection of small alterations in the genomic landscape.

*Fig 3: This formula expresses 6 possible pyrimidine single nucleotide variants.*

V.I Creating SBS files

Current SBS files are based on maximum two flanking nucleotide left and right of the substitution. *GenomeSigInfer* can create SBS files with increase context using a function called: *‘GenomeSigInfer.sbs.SBSMatrixGenerator()’*

1. The VCF files are filtered. They are filtered on a given reference genome and if the mutation is one of the 6 possible pyrimidine variants (fig.3).
2. Parse the VCF files and group them by chromosome. And create a SBS file for all the mutation in the VCF file, with the max context.
3. Compress the SBS matrix one context down, until the context is three (fig 1). Each matrix was saved as parquet file.

V.II Analysis mutational signatures

V.II.I Optimal NMF parameters

To determine the optimal NMF initialization and beta loss parameters. Utilizing the ‘96’ context file as input, all combinations of NMF initialization and beta loss parameters were ran. And decomposed using *‘SigProfilerAssignment’.*

The cosine similarity metrics on the decomposition results was determined. To assess the impact of these parameter choices on the accuracy and reliability of the mutational signature extraction process.

V.II.II Run NMF

With the optimal NMF initialization and beta loss parameters identified. The next step was to run NMF on the SBS files. Before initiating the NMF algorithm, crucial preprocessing steps were implemented.

1. A cutoff value based on a 96 percentile of the data was computed. To establish a threshold for data normalization.
2. The data was given random noise using multinomial randomness, contributing to the robustness of the subsequent analysis.
3. Normalizing the data by calculating the sum of each column and adjusting small values for stability

By utilizing a random seed, the Preprocessing class ensures that the introduction of variability through multinomial randomness is controlled and reproducible. Researchers can confidently share their code, knowing that others can replicate their findings by using the same random seed. Following this the NMF algorithm could be initiated on the preprocessed data.

V.II.III Decompose

Following the completion of the NMF process on the normalized genomic data, the linear sum assignment approach was used to determine optimal column assignments between each NMF result and the COSMIC dataframe. This phase was critical for assigning the correct mutational signature to each row in the NMF dataframe, ensuring exact mutational signature alignment across multiple contexts.

Following this alignment, the Jensen Shannon Distance and Cosine Similarity metrics were generated to quantify the dissimilarity and similarity, respectively, between the 96-context mutational signatures and those produced from larger contexts. These measures were useful in giving a quantitative assessment of the interactions between various genomic contexts, putting insight on the complex differences in the mutational landscape.

V.III Signature plots

Barplots were created for each decomposed matrix in different genomic contexts to visually show the correlations and trends identified by comparing mutational signatures. These visualizations presented a thorough and context-specific view, revealing the distribution of mutational types within each genomic context.

This strategy was expanded in the event of additional context files by constructing customized bars for each situation. Bars were color-coded in these enhanced visualizations to effectively reflect the proportion contribution of each nucleotide. This improved graphical depiction enabled a more sophisticated view of the complex mutational landscape, highlighting particular nucleotide contributions within various genomic settings.

V.IV Cluster

In pursuit of simplifying the representation of the genomic context, a method was constructed based on the nucleotide. There are also letters that represent ambiguity which are used when more than one kind of nucleotide could occur at that position [15].

A way to cluster the existing 9-context structure into a more condensed format, resulting in the same number of features as the original 7-context. By employing a clustering approach that combines the 2 nucleotides on the far left and far right into one of the following categories:

* "W" (Weak) or "S" (Strong)
* "M" (aMino) or "K" (Keto)
* "R" (puRine) or "Y" (pYrimidine)

It effectively reduces the dimensionality of the context while still preserving essential structural information.

*Fig 4: This formula expresses all the possibilities for the 7-context mutation. 4x4x4x6x4x4x4*=*24576 total combinations.*

*Fig 5: This formula expresses all the possibilities for the clustered 9-context mutation based on strength. 2x2x4x4x6x4x4x2x2*=*24576 total combinations.*

A similar perplexity between the two structures would indicate that the clustering effectively retains the essential information present in the original context, providing a more streamlined yet equally informative representation.

Perplexity (fig 6) provides a quantitative measure of how well the NMF model captures the complex patterns within genomic data.

*Fig 6: This formula expresses how ‘per word perplexity’ is calculated.*

V.V Meta signatures

Meta signatures are constructed by combining the topics of different chains to its centroid by repeatedly solving the optimal transport problem for the Jensen-Shannon distance (JSD) using the Hungarian algorithm until the centroid converged in terms of silhouette score [16]. The multinomial belief network is a bayesian deep belief network that uses multinomial-distributed variables as output [16]. The generative model is given by:

*Fig 7: This formula expresses the generative model of multinomial belief networks that uses Dirichlet samples.*

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.

[14] J.A. Busker. GenomeSigInfer. (2024). GenomeSigInfer.

[15] Nomenclature Committee of the International Union of Biochemistry (NC-IUB). (1986). Proceedings of the National Academy of Sciences of the United States of America, 83(1), 4-8.

[16] H. C. Donker, D. Neijzen, G. A. Lunter. (2023). " Multinomial belief networks." Cornell University.