Building mutational signatures in cancer using deep Bayesian neural nets

Background

To a large extent, cancer is a disease of the genome [1-3]. Cancer is characterised by uncontrolled growth of cells. To sustain this level of growth, the cancer cells have acquired several capabilities such as evading cell death and the construction of vascular infrastructure. Most of these capabilities are, primarily, acquired by mutation of the genome. These mutations have accrued by exposure to DNA-damaging processes coming from within (endogeneous) and from outside (exogeneous) throughout the course of life. It turns out that when these damaging processes create mutations, they leave an imprint on the genome [4]. These "*genomic scars*" arise because various chemicals preferentially bind to specific nucleotide regions of the DNA. By looking at the direct context of the mutations, you can thereby "infer" the way these mutations were created (aetiology).

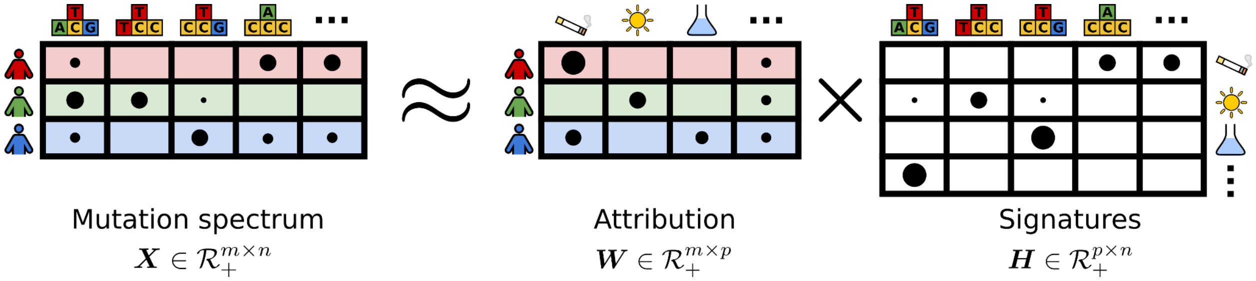
T
ACG
T
TCC
T
CCG
A
CCC
...
≈
×
Mutation spectrum
$\boldsymbol{X} \in \mathcal{R}_{+}^{m \times n}$
Attribution
$\boldsymbol{W} \in \mathcal{R}_{+}^{m \times p}$
Signatures
$\boldsymbol{H} \in \mathcal{R}_{+}^{p \times n}$

Fig. 1: Non-negative matrix factorisation of point mutations with flanking context into mutational signatures and their attributions. Single base substitutions with left and right flanking context are indicated by pyramids. Example aetiologies are indicated by cigarette smoke, sun, and Erlenmeyer.

In addition you can also quantify how much the cancer cells (and its descendents) have been exposed to specific processes. The trick is to count the number of mutations per context, and then to factorise this matrix using non-negative matrix factorisation. For single base substitutions (where only one DNA letter changes) this is illustrated in Fig. 1. In addition to singlets (point mutations), you can also look at doublets (pairs of substitutions) and small insertion-deletions [5]. (Or even copy number alterations and structural variants!)

Project

Up to now, most researchers have looked at the changing mutation including a single letter of context to the left and right [4,6]. This has resulted in a curated set of mutational signatures by COSMIC, which can be viewed online. In addition, several methods have been developed to determine the signatures and infer the attributions [7]. The goal of this project is to refine: (1) the statistical model and (2) the current representation of the mutations.

1. Refined model: hierarchical signatures

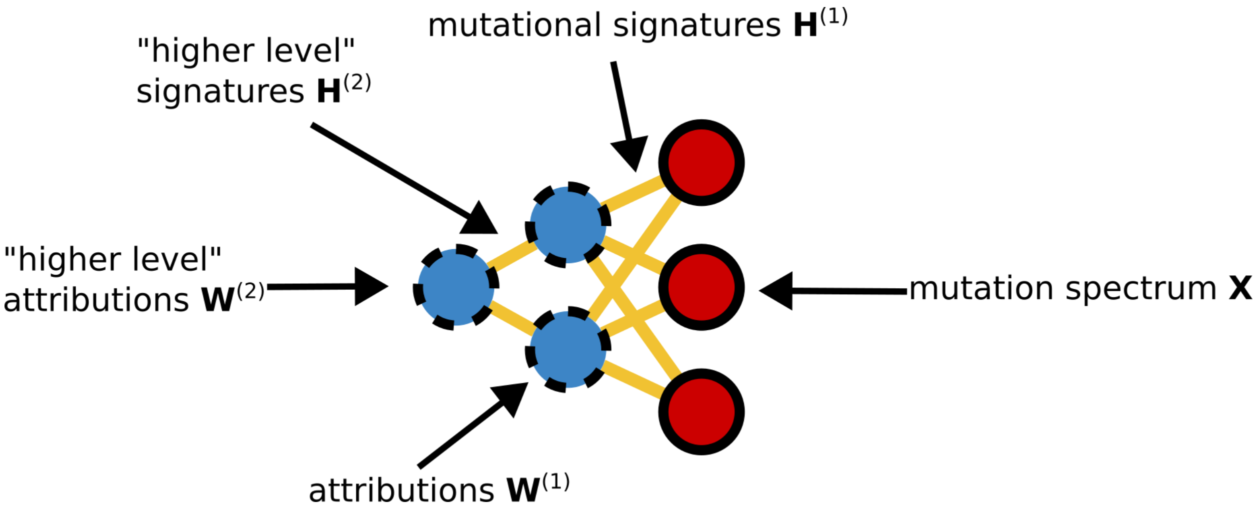
mutational signatures $\mathbf{H}^{(1)}$
"higher level" signatures $\mathbf{H}^{(2)}$
"higher level" attributions $\mathbf{W}^{(2)}$
mutation spectrum $\mathbf{X}$
attributions $\mathbf{W}^{(1)}$

Fig. 2: Non-negative matrix factorisation using a two-layer Bayesian neural net.

Interpretation of mutational signatures is currently mostly based on *post hoc* analysis. That is, after constructing the signatures (see, e.g., Ref. [4]). Moreover, several signatures are linked to the same aetiology. For example, for single base substitutions (SBS) signatures:

- SBS4 and SBS92 are both linked to tobacco smoking.
- SBS1 and SBS5 are both clock-like signatures.

We have recently implemented a Bayesian neural network for count data [2]. We are currently wrapping up a similar Bayesian neural net for categorical data. These methods are available in a ready-to-use Python software package to analyse hierchically related mutational signatures (Fig. 2) In theory, this method should be able to relate these aforementioned pairs of signatures from first principles.

2. Refined representation: more context

Current single base substitution signatures are based on one flanking nucleotide left and right of the substitution. However, this is not enough context to discriminate, for example, DNA damage generated by ABOPEC3A from ABOPEC3B. This requires at least one extra context letter [9]. In addition, we hypothesise that due to stocheometric hindering of the DNA helix and winding of DNA across nucleosomes, there are chemical bridging mechanisms at periodicities corresponding to these loops. In turn, it is expected that this leaves contextual imprints at the corresponding periodicities.

By increasing the context size, we should be able to capture these phenomena. Some work in this direction, using latent dirichlet allocation (LDA) has already been started by Shiraishi *et al.* [10].

Alternatively and complementary to this, it is also possible to increase the space of the dimensions of the mutations by dividing mutations further according to DNA annotations. Earlier work in this direction was done by Vohringer [11], which split mutations according to transcription/replication strand, nucleosome position and epigenetic environment.

Analysis plan:

- Curate and quality control the samples and mutations similar to Ref. [12]. In that work, they use data from ICGC and the cancer genome atlas (TCGA), which is publicly available.
    - We will try to get access to additional sources of data from Hartwig Medical Foundation (HMF) and the Pancancer Analysis of Whole Genomes (PCAWG) to increase the dataset.
- Replicate the mutational signatures using non-negative matrix factorisation.
- Compare with signatures from 1-layer and 2-layer Bayesian neural net.
- Determine priors for the expanded mutation representation (e.g., based on existing COSMIC signatures). Contrast determined signatures of larger dimensions with the current 96 feature representation.

References

[1] Hanahan, Douglas, and Robert A. Weinberg. "The hallmarks of cancer." cell 100.1 (2000): 57-70.
[2] Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of cancer: the next generation." cell 144.5 (2011): 646-674.
[3] Hanahan, Douglas. "Hallmarks of cancer: new dimensions." Cancer discovery 12.1 (2022): 31-46.
[4] Alexandrov, Ludmil B., et al. "The repertoire of mutational signatures in human cancer." Nature 578.7793 (2020): 94-101.
[5] Bergstrom, Erik N., et al. "SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events." BMC genomics 20.1 (2019): 1-12.
[6] Degasperi, Andrea, et al. "Substitution mutational signatures in whole-genome–sequenced cancers in the UK population." Science 376.6591 (2022): abl9283.
[7] Kim, Yoo-Ah, et al. "Mutational signatures: From methods to mechanisms." Annual Review of Biomedical Data Science 4 (2021): 189-206.
[8] Zhou, Mingyuan, Yulai Cong, and Bo Chen. "Augmentable gamma belief networks." The Journal of Machine Learning Research 17.1 (2016): 5656-5699.
[9] Chan, Kin, et al. "An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers." Nature genetics 47.9 (2015): 1067-1072.
[10] Shiraishi, Yuichi, et al. "A simple model-based approach to inferring and visualizing cancer mutation signatures." PLoS genetics 11.12 (2015): e1005657.
[11] Vöhringer, Harald, et al. "Learning mutational signatures and their multidimensional genomic properties with TensorSignatures." Nature communications 12.1 (2021): 3628.

[12] Alexandrov, Ludmil B., et al. "Signatures of mutational processes in human cancer." Nature 500.7463 (2013): 415-421.