Multimodal Cell-Free DNA Embeddings are Informative for Early Cancer Detection

Abstract

Cell-free DNA is a promising biomarker for early cancer detection, as it circulates in the blood and can be extracted non-invasively. However, methods of analysing the genetic and epigenetic patterns present in cell-free DNA are outdated, and fail to fully capture the wealth of biological information contained within these molecules. We present a Transformer based deep learning model that combines the three distinct modalities contained within cell-free DNA: epigenetic information in the form of DNA methylation patterns, genetic sequence, and cell-free DNA fragment length. After training on publicly available data, we demonstrate our model can accurately distinguish liver cancer patients using cell-free DNA samples alone. We demonstrate model generalisability by accurate classification of liver cancer patients from entirely distinct patient cohorts. Finally, we show that the vector embeddings of cell-free DNA learnt by this multimodal deep-learning model are biologically informative, and may help shed light on the origins and aetiology of this elusive bio-molecule.

1 Introduction

DNA is released from somatic cells undergoing apoptosis, and circulates in the peripheral bloodstream, contributing to what is known as the cell-free DNA (cfDNA) pool. Cell-free DNA molecules carry three distinct information modalities, all of which vary with cancer status: genomic sequence, a methylation pattern and the fragment length (as in Figure 1). This molecule can be extracted non-invasively from blood plasma, following routine blood draws. In cancer, cell cycle dysregulation leads to increased rates of cell turnover, which causes downstream fluctuations to tumour-specific cfDNA concentrations detected in the blood.

These changes are subtle, but if reliably detected they provide a route to early and non-invasive cancer diagnosis. Methylation sequencing methods are needed to detect these tissue-specific changes in cfDNA release, because genomic sequence is identical in all somatic cells whereas DNA methylation is tissue-specific. Recently, methods have been developed to simultaneously capture methylation, sequence and fragment length information of cfDNA with high fidelity and depth of coverage, resulting in a rich dataset describing the genome-wide cell-free DNA state of cancer patients.

Current clinical methods of analysing cfDNA are outdated in three important ways. Firstly, they were designed for low quality, low-depth genomic sequencing data, so tend to aggregate information across all cfDNA fragments at each genomic locus, disregarding the unique information each individual bio molecule may contain. Secondly, classification models of cfDNA have relied on manual feature extraction at known risk loci, which limits research to current genomic hypotheses and scales poorly. Automated feature extraction methods are more suitable to explore this relatively unknown bio molecule and its role in cancer aetiology at scale.

Finally, there is no way current way to combine the distinct data modalities of cfDNA sequence, methylation and fragment length. Clinical biomarker tests that rely on cfDNA tend to focus on just one of these modalities. In doing so, they neglect not only the additional information that could

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

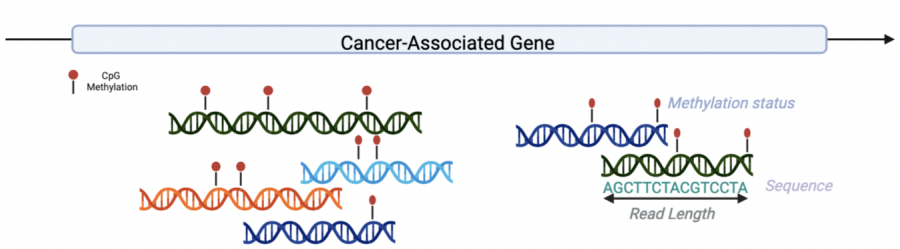

Figure 1: A schematic of cell-free DNA fragments found in peripheral blood, aligned to a gene of interest. These fragments contain three distinct modalities of information: DNA sequence, methylation pattern (shown here on CpGs) and their variable fragment length.

be gleaned from this molecule, but also the potential to better understand a key question in cfDNA biogenesis: what is the interplay between genomic sequence, methylation status and fragment length? Are certain sequence motifs prone more to aberrant methylation during tumourigenesis, and are more highly methylated fragments typically shorter? As a result, despite garnering considerable interest as an early cancer biomarker, the dynamics and mechanisms of cfDNA fragmentation and release into the blood are poorly understood.

Here we present a Transformer model that classifies each individual cfDNA fragment, incorporating methylation state, genomic sequence and fragment length simultaneously into a single, multimodal classifier. We use this model to accurately distinguish liver cancer patients from healthy controls across patient cohorts.

2 Results

We encode each individual cfDNA molecule as a variable length vector, where numbers 1-4 represent each of the possible nucleotide bases at each position. Methylated cytosine is encoded as a 5, to distinguish it from unmethylated cytosine. These encoded molecules are then passed to a Transformer deep learning model (standard architecture (Vaswani *et al* 2017; Tay *et al* 2022) with a binary classification final layer, whose task is to predict fragment origin: Healthy or Cancer. This model was trained on public data, which consisted of billions of individual reads from either Healthy cfDNA samples or Hepatocellular Carcinoma tumour samples.

For any given cancer patient or healthy control, each individual cfDNA fragment is assigned a probability of originating from cancer, as shown in Figure 2. As shown, the probability distribution over all fragments is largely similar between healthy and cancer cfDNA samples. This is to be expected: cell-free DNA is thought to be released from most healthy somatic tissues, with the majority contribution coming from healthy lymphocytes. In cancer patients, only a small subset of cfDNA fragments actually originate from the tumour, and even this circulating tumour DNA fraction varies considerably with cancer stage and cancer type. A small bump in the probability distribution around P=0.9 is seen in cancer cfDNA samples (red) but not in healthy cfDNA samples (green), and can be attributed to cfDNA originating from the tumour.

We aggregate fragment scores to develop a patient-level risk score, which we then use to classify patients. This model is still in active development, and we are currently seeking to improve model architecture and hyperparameter selection. Patient cfDNA data collection is also ongoing, but we have thus far evaluated the model on the following three datasets:

1. The held-out test set of unseen cfDNA samples, but taken from the same cohort as the training dataset. All patients in the test cohort of this dataset (n=24) are correctly classified.
2. A separate publicly available dataset of cfDNA samples from liver cancer patients and healthy controls (n=8). All of these patients are also correctly classified by the model, demonstrating its generalisability to new patient cohorts.
3. A third dataset of cfDNA samples from liver cancer patients and healthy controls generated in-house (n=53). A subset of patients in this cohort have early stage liver cancer, and are

incorrectly classified as healthy by our model, most likely due to insufficient cfDNA material originating from cancer. Taking these misclassifications into account, we observe an overall AUC of 0.81 for this patient cohort, and late-stage disease is still detected.

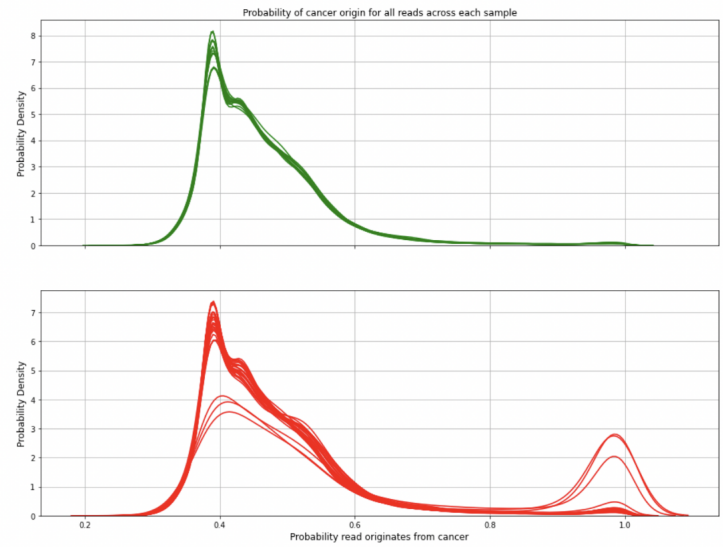

Figure 2: The distribution of cfDNA fragment scores across healthy controls (green) and liver cancer patients (red) in the test set. Each line represents the distribution over all scores for a single patient's cfDNA sample, where the score for a single fragment is the probability that fragment originated from cancer. This data is from patient cohort 1 and 2.

By obtaining classifications for each individual fragment in a patient's cfDNA population, we can begin to untangle the relationship between methylation state, genomic sequence and fragmentation length of individual cfDNA fragments, and how this affects the predicted origin of the fragment. For example in Figure 3, we can see that shorter fragments are assigned a lower cancer score on average. We can also see that as cfDNA fragment methylation rate decreases, the assigned probability that fragment originates from cancer increases.

Population-level cancer screening tests must be non-invasive, and liquid biopsy is the most promising pan-cancer non-invasive test. As cfDNA-based liquid biopsy diagnostics become clinically adopted, we desperately need new methods to make sense of the richly informative cfDNA molecules that circulate in our blood. This method details the initial results of an ongoing data collection and modelling effort towards this end.

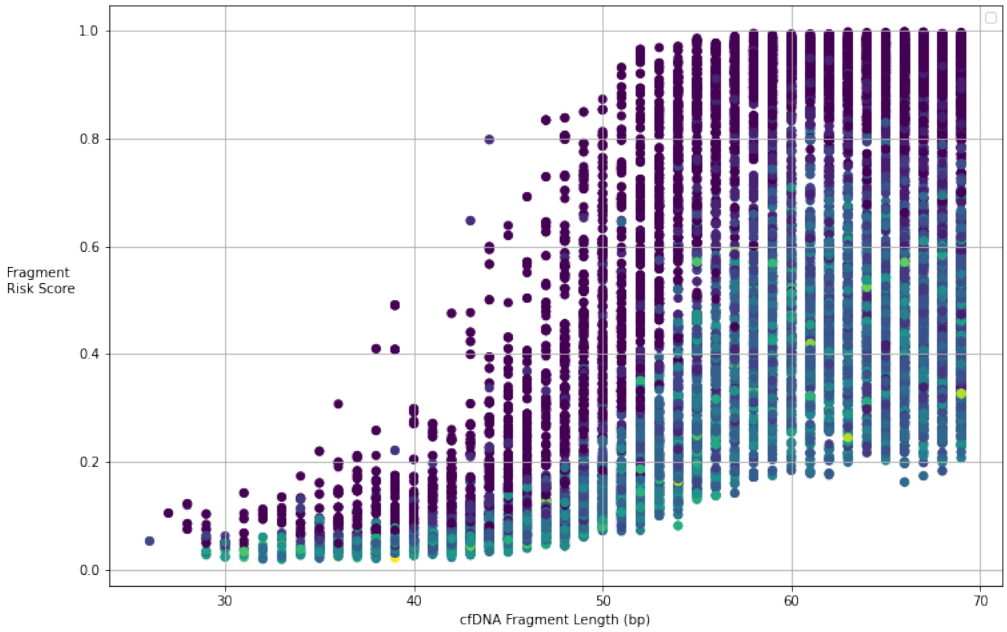

Figure 3: The distribution of cfDNA fragment risk scores stratified by fragment length, for a single liver cancer patient (downsampled for legibility). Methylation state (normalised by fragment length) is represented by marker color, where a lighter color means higher methylation.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient Transformers: A Survey. ACM Comput.Surv. (apr 2022). Just Accepted.