OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Aptamer Candidate Development through ctDNA Fragments

Ananya Goli
ananya.goli007@gmail.com

ABSTRACT

Cancer deaths are largely attributable to late diagnosis and lack of recurrence surveillance, especially seen in low and middle income countries. Cell-free DNA (cfDNA) in plasma provides a minimally invasive window into tumor biology, making it a promising biomarker for cancer detection and monitoring due to distinctive end motif signatures. I hypothesize that ctDNA fragment-end motifs can be translated into stable aptamer candidates with higher binding specificity with suitable properties for a biosensor. Public cfDNA whole-genome sequencing for lung cancer and healthy plasma datasets were analyzed. K-mer frequency analysis and motif enrichment scanning was conducted, and custom motifs which were canonical transcription factor binding sites such as TATAAA, CCAAT, and GC-rich sequences were quantified between cancer and healthy datasets, with statistical validation to ensure accuracy. From these enriched motifs, aptamer candidates were designed and evaluated for GC content, melting temperature, and secondary structure stability using Biopython and RNAfold. Motif enrichment analysis identified 1 match for TATAAA in ctDNA, 11 matches for CCAAT (compared to 5 in healthy), and 6 matches for GC-rich motifs (compared to 3 in healthy). These motifs revealed five potential aptamer candidates with stability. Of these, secondary structure prediction for aptamer 1and 2 formed a weak stem-loop, with aptamer 1 also forming a tertiary structure. These findings suggest that CCAAT and GC-rich motifs are selectively enriched in cancer cfDNA. The modeling of aptamer 1 based on the enriched motifs demonstrates that fragmentomic signals can be translated into aptamer scaffolds with realistic folding potential.

INTRODUCTION

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 alone [1]. The World Health Organization estimates that by 2040 there will be 29.5 million new cancer diagnoses and 16.5 million cancer-related deaths globally [2]. 70% of cancer deaths in the world in 2020 are from low- and middle-income countries [3]. This is largely attributable to two factors: late diagnosis and lack of recurrence surveillance due to inaccessibility of effective treatment [3].

As explained in Figure 1, Cell-free DNA (cfDNA) are short fragments of DNA that are released into bloodstreams from cells undergoing apoptosis or necrosis. These DNA fragments circulate freely in blood

plasma and are typically derived from normal tissues. Similarly, the DNA that sheds from the tumor cells rather than normal tissue and circulates into the blood stream are called circulating tumor DNA (ctDNA), which is especially present in lung cancer [4].

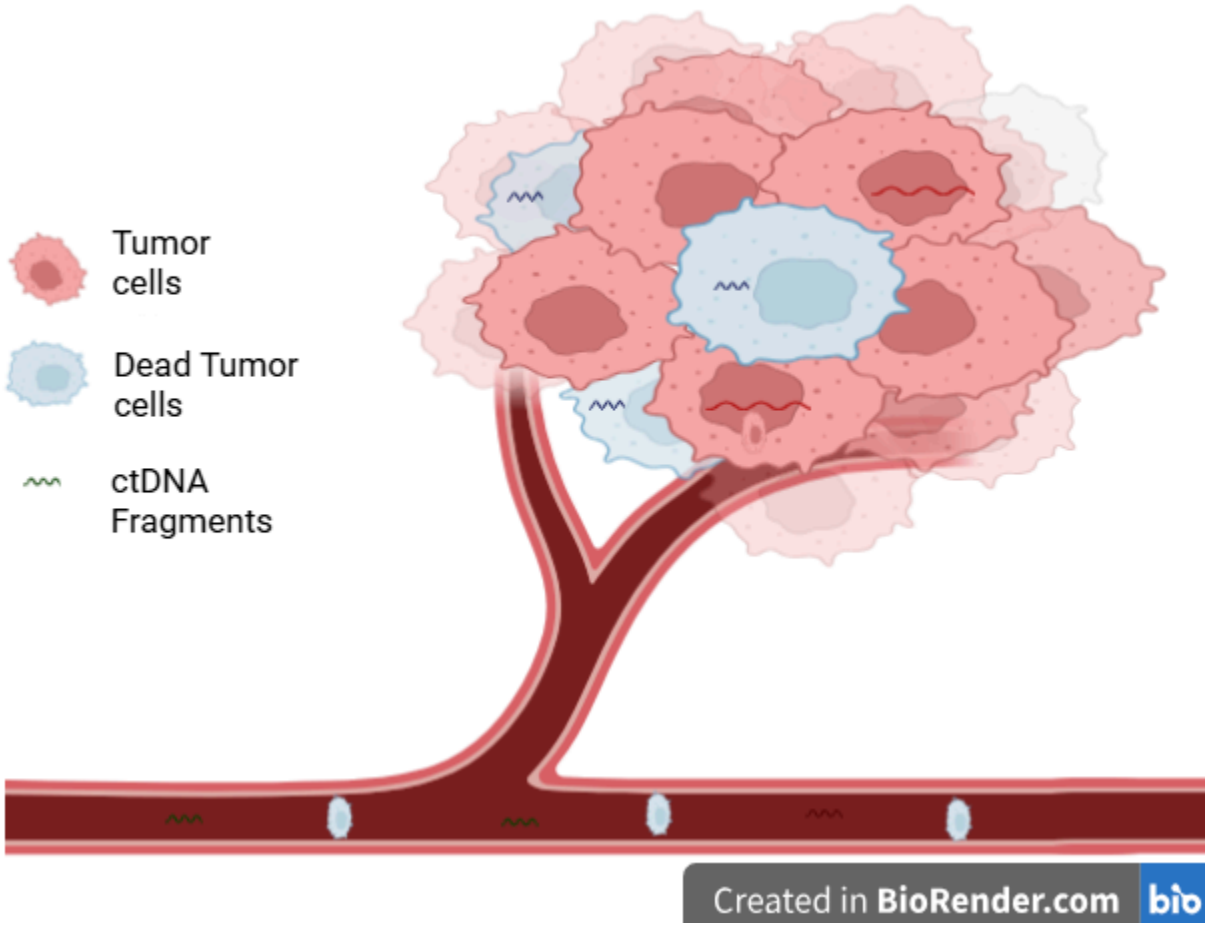

Figure 1. Figurative representation of ctDNA fragments shedding from a tumor and passing into blood stream made from biorender.com

Studies have shown that ctDNA fragments are shorter and display distinct DNA patterns at the ends of the fragment, or fragment-end motifs, compared to healthy cfDNA [5]. These fragment end patterns in ctDNA versus cfDNA can be distinguishable by sequence motifs, which are short recurring patterns, and k-mers, which are substrings of specific length [5].

In recent years, ctDNA, the tumor derived form of cfDNA, has shown promise as a non-invasive biomarker for cancer detection and surveillance, which can improve accessibility, early detection, and personalization. However, traditional ctDNA detection methods have low sensitivity, are expensive, and remain less effective in low–tumor-fraction settings [6][7]. Despite the promise of ctDNA as a

non-invasive biomarker, and fragment-end motif enrichment already being observed in ctDNA, we still cannot translate these fragments into functional aptamers, which are short, structured oligonucleotides as depicted in Figure 2, for a solution that has high sensitivity This limits the effective use of ctDNA in lung cancer detection.

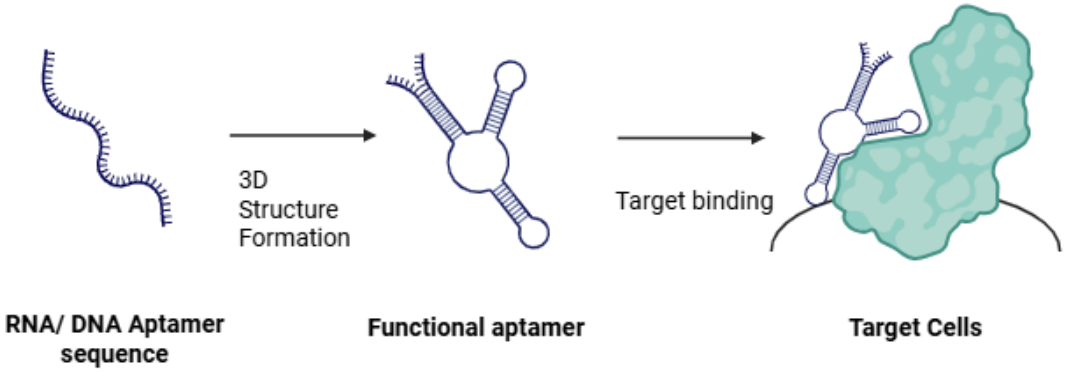

Figure 2. Figurative representation of aptamer binding with RNA/DNA sequence with its target molecule made from biorender.com

Previous studies describe how mutation-based ctDNA detection is limited when tumor fraction is low because mutations are rare and require deep sequencing [8]. ctDNA often represents less than 1% of total cfDNA, which makes mutation detection difficult even with ultra-deep sequencing [9].

Recent fragmentomic studies have identified that the ctDNA motifs have shorter lengths typically in the 120–140 bp range, compared to ~166 bp modal size in healthy cfDNA, have distinct positioning of start and end sites relative to nucleosome dyads [5] [8]. Recent studies have also asserted that the enrichment for specific fragment-end sequence motifs is near cleavage sites [10], [11]. It is further demonstrated that combining these fragmentomic signals improves ctDNA enrichment beyond size-selection alone, suggesting that sequence-level features may be directly exploited for selective molecular capture, highlighting the potential of end-motif signals as discriminators [10]. By shifting away solely from mutation-based detection and toward fragmentomic signatures, recent studies have asserted ctDNA's full potential for cancer detection [10], [11].

Aptamers have the potential to offer a low-cost implementation of ctDNA for lung cancer detection with high binding specificity, making this more preferable than current ctDNA-based technology [12]. If aptamers can be designed through the end patterns of ctDNA, they could selectively capture this tumour derived DNA from samples. This research aims to accurately identify tumor-enriched fragment-end motifs in cfDNA and with this, and design aptamer candidates that target these motifs for a new form of cancer detection. This can enable more affordable, sensitive, and accessible cancer diagnostics.

As previously mentioned, the two factors of late diagnosis and lack of recurrence surveillance contribute heavily to the fact that 70% of cancer cases occur in low and middle income countries. Economic barriers pose a major obstacle to timely care [13]. Patients from poor countries often go to health facilities only at advanced stages of their disease. Specialized medical services are predominantly located in urban centers, while rural and remote populations are left with limited access to care without systematic follow up [14]. Creating aptamers for therapy against cancer works towards improving these factors. Developing aptamers for cancer therapy can enable the creation of low-cost diagnostic and monitoring tools, helping to reduce the burden on patients by the current healthcare system and its socioeconomic challenges.

The research questions are two fold - can enriched fragment-end motifs of ctDNA be translated into stable aptamer candidates and if computationally designed aptamers display the needed stability and properties for biosensor development? I hypothesize that the tumor-enriched cfDNA, or ctDNA, fragment-end motifs can be computationally translated into stable aptamer candidates with higher binding specificity with suitable properties for a biosensor.

This study enhances previous fragmentomic research by utilizing a computational pipeline that first extracts cfDNA fragment ends, then identifies tumor enriched k-mers and motifs, and finally designs aptamer candidates while assessing their stability and folding. This process connects computational biology and translational diagnostics by combining enriched ctDNA ends with aptamer modeling. The expected outcomes of this study is identification of fragment-end sequences significant in ctDNA compared to healthy cfDNA, computational design of aptamer candidates from these tumor-enriched fragment ends and generation of ~~high-sensitivity aptamer candidates~~ potential preliminary aptamer scaffolds suitable for cancer detection.

METHODS

The work is organized into five steps as laid out in Figure 3 below.

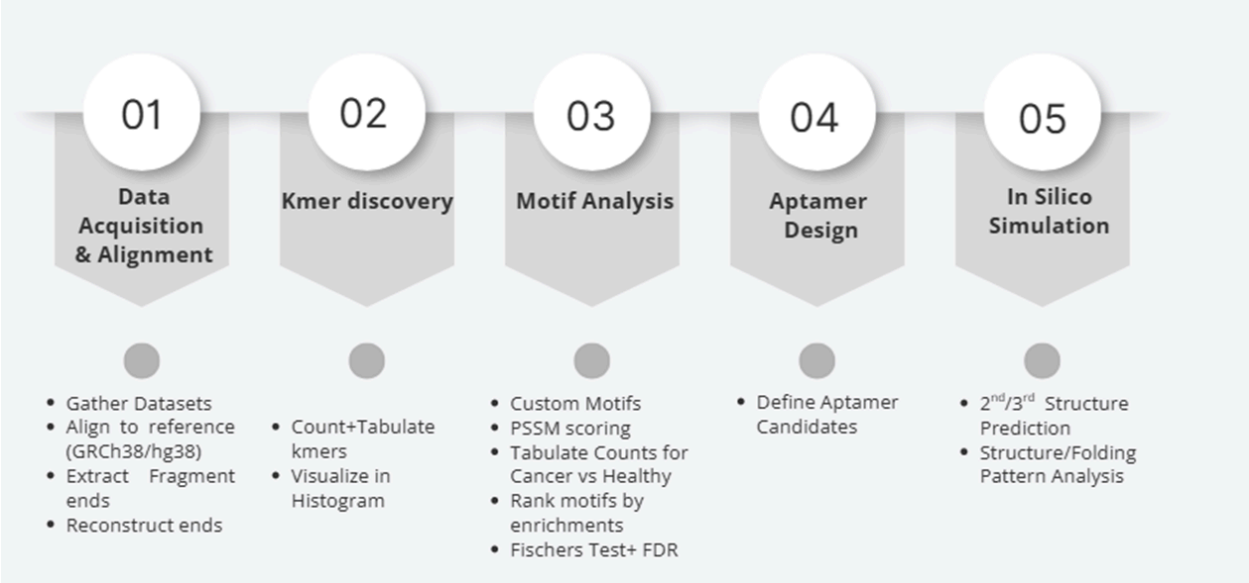

Figure 3.explains the five steps involved in the methods.

The five objectives that I set out to achieve as laid out in the five steps above are to extract cfDNA fragment ends, identify highly seen fragment ends, or k-mers, find tumor enriched motifs, design aptamer candidates, and assess stability and folding.

The datasets for whole-genome sequencing datasets of plasma cell-free DNA were taken from the European Nucleotide Archive (ENA) and the Sequence Read Archive (SRA). Two datasets were selected to simulate a cancerous environment and a healthy, control environment: SRR8298030, from an individual with lung cancer, and SRR8298035, from a healthy individual. The human reference genome used for aligning was GRCh38/hg38 (Ensembl release 110), accounting for compatibility with current genome mapping.

Raw FASTQ files were processed using a standardized alignment workflow. Reads were aligned to the reference genome using BWA-MEM (v0.7.17), a widely adopted algorithm for short-read mapping. Resulting BAM files were sorted and indexed with Samtools (v1.10) to facilitate downstream analysis. Quality control revealed 125,696 total reads in the cancer sample, of which 98 mapped successfully (0.08%). Although mapping rates were low, reflecting the repetitive and fragmented nature of cfDNA, fragment ends were nonetheless recovered and proved sufficient for motif analysis.

Fragment start and end coordinates were reconstructed from paired-end alignments using the pysam library. End sequences of 10 bp length were extracted and exported to FASTA format for motif discovery. K-mer frequencies were calculated using Jellyfish (k=10) to capture longer sequence contexts, and custom Python scripts were applied for shorter k-mers (k=6) to highlight enriched motifs.

Motif scanning was performed using Biopython's motifs module. Position-specific scoring matrices (PSSM) were applied to both cancer and healthy fragment-end sequences. Candidate motifs included canonical transcription factor binding sites such as TATAAA and CCAAT, as well as GC-rich sequences, E-box (CACGTG), and SP1 (GGGCGG). Enrichment was quantified by comparing motif counts between

cancer and healthy datasets, with statistical testing performed using Fisher's exact test and Benjamini–Hochberg false discovery rate (FDR) correction. Aptamer candidates selected have p values of less than 0.5, indicating statistical significance. Please see the full supplemental file here.

Synthetic validation was conducted to confirm binding potential for motifs absent in direct counts, specifically E-box and SP1, which achieved high binding scores despite not appearing in raw enrichment tables.

From the enriched motifs, five aptamer sequences were designed to serve as candidate molecular probes. Each aptamer was evaluated for GC content and melting temperature using Biopython's MeltingTemp module, providing insight into thermodynamic stability. Secondary structure prediction was performed using RNAfold (ViennaRNA), generating dot-bracket notation and minimum free energy values. Aptamer 1 and Aptamer 2 exhibited weak stem-loop structures with slightly negative free energy values, while Aptamers 3–5 remained linear with neutral energy values. Aptamer 1 was further modeled in RNAComposer, producing a compact tertiary structure consistent with secondary structure predictions and serving as proof of concept for aptamer folding in three dimensions.

The computation work is performed in Google Colab using Python and BioPython and the tools that were utilized are BWA for aligning sequences to the reference genome, Samtools for manipulating BAM files and performing quality control, Jellyfish for k-mer counting, Matplotlib for plotting k-mer and motif distributions, Biopython's motifs module for motifs construction, Pandas for tabular data manipulation, Scipy.stats and statsmodels for statistical analysis, and ViennaRNA (RNAfold) for predicting aptamer secondary structures and calculating free energy values.

RESULTS

Quality control confirmed expected fragmentomic differences between cancer and healthy cfDNA. The cancer sample contained shorter fragments and exhibited low mapping rates, consistent with tumor-derived cfDNA characteristics.

The analysis of the resulting K-mers showed a highly skewed frequency distribution, with most occurring infrequently and strong enrichment occurring in a small subset.. The most frequent k-mers were homopolymeric GC-rich sequences such as GGGGGG (1117 counts) and CCCCCC (1131 counts), followed by AT-rich motifs like TTTTTT (220 counts) and AAAAAA (305 counts). Further, GC-rich variants like GGGGGA, GGGGGC, and GGGGCG were also enriched, each appearing more than 10 times. This pattern was validated by plotting the k-mer counts on a histogram as seen in Figure 3, which showed a distribution skewed right on a logarithmic scale, aligned with selective motif enrichment in cfDNA fragment ends.

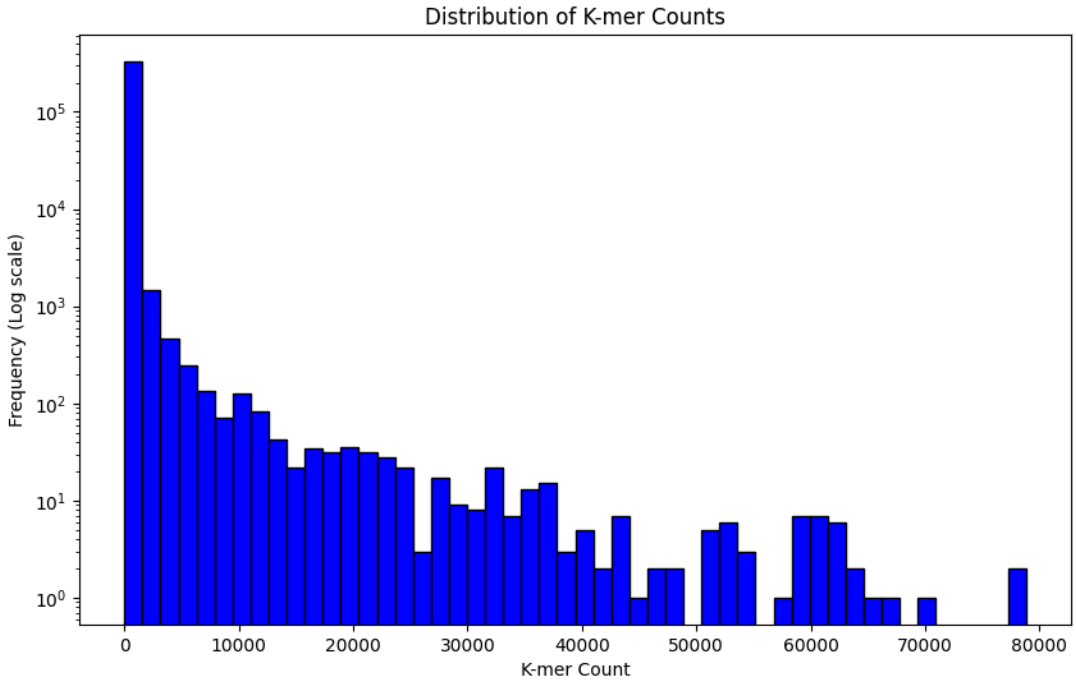

Figure 4. Distribution of K-mer counts

The analysis of enriched motifs found 1 match for TATAAA in cfDNA 11 matches for CCAAT (compared to 5 in healthy), and 6 matches for GC-rich motifs (compared to 3 in healthy). However, despite high binding potential confirmed by synthetic validation (10.83), no direct matches were found in E-box or SP1 in ctDNA

As shown in Table 1, canonical motifs such as TATAAA were rare, while CCAAT boxes were more frequent in cancer fragments. These findings suggest that CCAAT and GC-rich motifs are selectively enriched in cancer cfDNA, while E-box and SP1 motifs, although absent in direct counts, remain structurally compatible as confirmed by synthetic validation.

Motif	**Cancer Reads**	**Healthy Reads**
TATAAA	**1**	**0**
CCAAT	**11**	**5**
GC-rich	**6**	**3**
E-box	**0**	**0**
SP1	**0**	**0**

Table 1. Enrichment of fragment-end motifs in cancer and healthy cfDNA samples

From the enriched motifs, five aptamer candidates were designed. Aptamers derived from GC-rich motifs exhibited higher GC content (70–90%) and melting temperatures (34–38 °C), indicating greater thermodynamic stability. Secondary structure prediction revealed that Aptamers 1 (GCUUCUUUGC) and 2 (GCGAGAGAGC) formed weak stem-loop structures with slightly negative free energy values (–0.1 and –0.3 kcal/mol), while Aptamers 3–5 (GCCCCCCUGC, GCCCCCUCGC, GCCCCUCCGC) remained linear with free energy values of 0.0 kcal/mol as depicted in Figures 4 - 8. This free energy measures stability of the 3-D folded structure, and affinity to the target molecule. A negative free energy is desirable and indicates stronger stability.

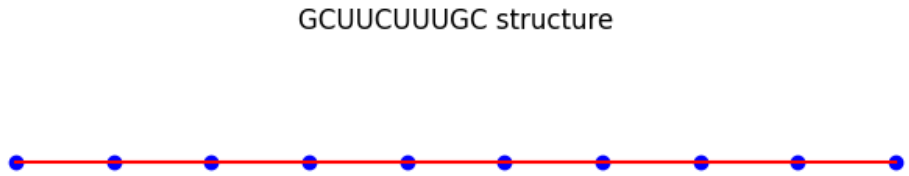

Figure 5. Secondary structure of Aptamer 1

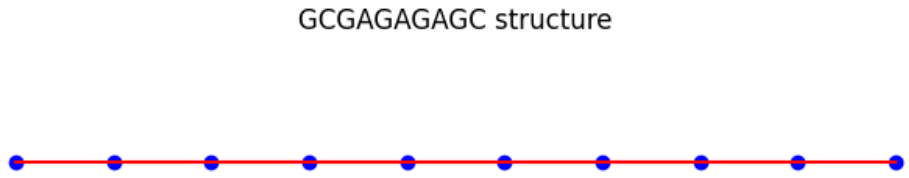

Figure 6. Secondary structure of Aptamer 1 and Aptamer 2 (stem-loop folding)

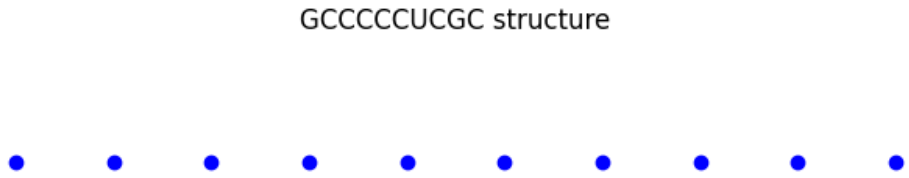

Figure 7. Secondary structure of Aptamer 3 (linear)

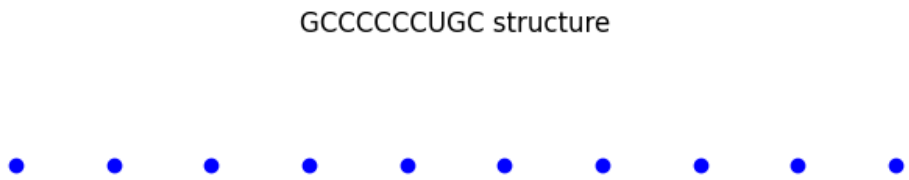

Figure 8. Secondary structure of Aptamer 4 (linear)

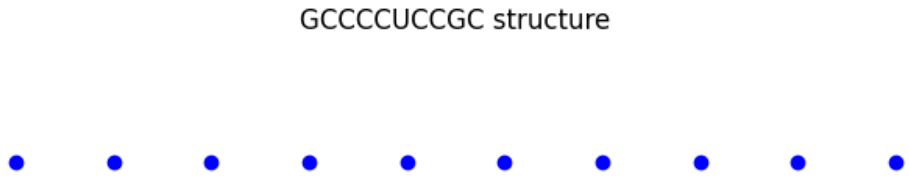

Figure 9. Secondary structure of Aptamer 5 (linear)

As shown in Table 2, free energy values ranged from −0.3 to 0 kcal/mol, consistent with limited structural stability. Aptamer 1 was successfully modeled in RNAComposer, producing a compact tertiary structure consistent with the predicted folding as depicted in Figure 9.

Sequence	Structure	Free Energy (kcal/mol)
GCUUCUUUGC	**((.....))**	**-0.1**
GCGAGAGAG C	**((.....))**	**-0.3**
GCCCCCCUGC	**.....**	**0**
GCCCCCUCGC	**.....**	**0**
GCCCCUCCGC	**.....**	**0**

Table 2. Predicted secondary structures and free energy values of aptamer candidates

Figure 10. Tertiary structure model of Aptamer 1 generated using RNAComposer

These findings suggest that while GC-rich aptamers are generally stable, only a subset form secondary structures suitable for tertiary modeling, highlighting Aptamer 1 as the most promising candidate for downstream validation. As shown in Figure 4–8 it illustrates the predicted folding behavior of all five aptamer candidates, ranging from weak stem-loop structures to fully linear conformations. Figures 5–8, Aptamers 3–5 remained linear, while Aptamers 1 and 2 exhibited modest folding.

DISCUSSION

This study demonstrates the feasibility of connecting fragmentomic motif discovery with aptamer design through a reproducible in silico pipeline. By focusing on cfDNA fragment ends, we were able to identify sequence motifs that distinguish cancer samples from healthy controls. The enrichment of GC-rich motifs and CCAAT boxes in cancer cfDNA is consistent with prior fragmentomic studies [10], [11], reinforcing the biological relevance of these signals. Such motifs may represent underlying chromatin accessibility and transcription factor binding patterns in tumor cells, and may provide insight into cfDNA fragmentation processes.

The aptamer candidates designed from such motifs had moderate stability, with GC content of 50-90% and melting temperatures of 30-38 °C. Such values indicate that the aptamers are not extremely stable, but are stable enough to act as proof-of-concept molecules. While the aptamers designed as part of this proof-of-concept computational study show moderate secondary structure stability, it should be noted that

these candidates are preliminary and would require further optimization and experimental validation before functional use.

Secondary structure prediction revealed that most aptamers remained linear, which may limit their binding potential, but Aptamer 1 formed a weak stem-loop and was successfully modeled in RNAComposer. This modeling of tertiary structure is a crucial validation step, as it shows that the fragmentomic signals can be decoded into aptamer structures with a viable folding ability.

One of the major advantages of this pipeline is that it is reproducible. Every step, from the retrieval of raw data to aptamer modeling, is recorded and automated, so that the findings can be reproduced or applied to other datasets. However, it is important to note that there are certain limitations to this pipeline The reported mapping rate of each data set of 0.08% is very low as ctDNA and cfDNA can be challenging to map due to its repetitive and fragmented nature. I acknowledge that a small number of mapped reads may limit statistical strength. Involving larger datasets to increment mapping rate as part of a future study can improve confidence of the findings.

In addition, the found aptamer candidates are preliminary and would require further optimization and experimental validation before functional use. Future work should expand motif discovery across larger cohorts and integrate experimental validation of aptamer binding. Together, these results establish a pipeline that connects computational biology and translational diagnostics by demonstrating how fragmentomic signals can be transformed into aptamer candidates. This approach offers both biological insight into tumor-specific cfDNA features and practical outputs that can be extended to biosensor development. In the long term, such pipelines may contribute to non-invasive cancer diagnostics by enabling selective molecular capture of ctDNA fragments, thereby enhancing sensitivity and specificity in clinical assays.

CONCLUSION

We successfully implemented a computational pipeline that links fragment-end motif discovery with aptamer design, demonstrating a reproducible workflow that bridges fragmentomic analysis and translational diagnostics. By analyzing cfDNA fragments from cancer and healthy plasma samples, we identified enrichment of GC-rich motifs and CCAAT boxes, consistent with prior fragmentomic studies. Aptamer sequences derived from these motifs exhibited moderate stability, as reflected in GC content and melting temperature analysis, while secondary structure prediction confirmed weak folding for select candidates. Importantly, Aptamer 1 was successfully modeled in RNAComposer, providing proof of concept for tertiary structure generation and validating the potential of this approach.

The output of the pipeline, which are motif enrichment tables, aptamer libraries, stability data, and structural models, offers not only biological understanding of tumor-specific cfDNA characteristics but also the necessary information that can be developed further for the purpose of biosensor design. In addition to the reproducibility, the pipeline demonstrates capability of computational biology to produce concrete results that can aid in the innovation of diagnostic tools.

Future work will include the optimization of aptamers to enhance folding and binding, extension of motif discovery to larger and more varied cfDNA datasets, and inclusion of experimental validation to confirm aptamer efficacy in vitro. Additional computational simulations of capture efficiency will also evaluate diagnostic potential, laying the path for assays that could potentially enhance sensitivity and specificity in cancer diagnosis

BIBLIOGRAPHY

1. World Health Organization: WHO. (2025, February 3). *Cancer*. https://www.who.int/news-room/fact-sheets/detail/cancer
2. Shah, S. C., Kayamba, V., Peek, R. M., & Heimburger, D. (2019). Cancer control in Low- and Middle-Income countries: Is it time to consider screening? *Journal of Global Oncology*, *5*(5), 1–8. https://doi.org/10.1200/jgo.18.00200
3. The global cancer burden. (n.d.-b). American Cancer Society. https://www.cancer.org/about-us/our-global-health-work/global-cancer-burden.html
4. Sánchez-Herrero, E., Serna-Blasco, R., De Lope, L. R., González-Rumayor, V., Romero, A., & Provencio, M. (2022). Circulating Tumor DNA as a Cancer Biomarker: An Overview of Biological Features and Factors That may Impact on ctDNA Analysis. *Frontiers in Oncology*, *12*, 943253. https://doi.org/10.3389/fonc.2022.943253
5. Underhill, H. R., Kitzman, J. O., Hellwig, S., Welker, N. C., Daza, R., Baker, D. N., Gligorich, K. M., Rostomily, R. C., Bronner, M. P., & Shendure, J. (2016). Fragment length of circulating tumor DNA. *PLoS Genetics*, *12*(7), e1006162. https://doi.org/10.1371/journal.pgen.1006162
6. Rui, M., Wang, Y., & You, J. H. S. (2025). Health Economic Evaluations of Circulating Tumor DNA Testing for Cancer Screening: Systematic review. *Cancer Medicine*, *14*(3), e70641. https://doi.org/10.1002/cam4.70641
7. Yi, X., Ma, J., Guan, Y., Chen, R., Yang, L., & Xia, X. (2017). The feasibility of using mutation detection in ctDNA to assess tumor dynamics. *International Journal of Cancer*, *140*(12), 2642–2647. https://doi.org/10.1002/ijc.30620
8. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., & Shendure, J. (2016). Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*, *164*(1–2), 57–68. https://doi.org/10.1016/j.cell.2015.11.050
9. Staff, L. I. T. L., Staff, L. I. T. L., & Staff, L. I. T. L. (2025, January 24). *Cell-Free DNA (cfDNA) vs. Circulating Tumor DNA (ctDNA) Explained*. Life in the Lab. https://www.thermofisher.com/blog/life-in-the-lab/cfdna-vs-ctdna/
10. Markus, Havell & Chandrananda, Dineika & Moore, Elizabeth & Mouliere, Florent & Morris, James & Brenton, James & Smith, Christopher & Rosenfeld, Nitzan. (2022). "Refined Characterization of Circulating Tumor DNA Through Biological Feature Integration." *Scientific Reports*, vol. 12, no. 1, Feb. 2022, p. 1928, doi:10.1038/s41598-022-05606-z.
11. Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., Jensen, S. Ø., Medina, J. E., Hruban, C., White, J. R., Palsgrove, D. N., Niknafs, N., Anagnostou, V., Forde, P., Naidoo, J., Marrone, K., Brahmer, J., Woodward, B. D., Husain, H., Velculescu, V. E. (2019). "Genome-wide

Cell-free DNA Fragmentation in Patients With Cancer." *Nature*, vol. 570, no. 7761, May 2019, pp. 385–89, doi:10.1038/s41586-019-1272-6.

12. *Aptamers*. (2026). Aptagen. https://www.aptagen.com/
13. *Health and Income: Understanding the Health-Poverty Trap*. (2026, January 11). City of Bridgeport. https://www.bridgeportct.gov/news/health-and-income-understanding-health-poverty-trap
14. Anjos, G. R. D., Machado, G. F., De Barros, C. P., De Andrade, V. P., De Barros Maciel, R. M., & Cunha, L. L. (2025). Cancer survivorship in low- and middle-income countries: challenges, needs, and emerging support strategies. *Frontiers in Public Health*, *13*, 1601483. https://doi.org/10.3389/fpubh.2025.1601483