Detecting Circulating Tumor DNA Sequences

Victor Song [*]

Abstract

The liquid biopsy procedure to screen for early-stage cancers and monitor treatment responses involves the detection of cancer biomarkers in bodily fluids such as blood and urine. Recent advances in liquid biopsy techniques have involved detecting the presence of tumor-derived cell-free DNA (circulating tumor DNA, ctDNA) in a simple blood test. This detection procedure is minimally invasive and provides an attractive and reliable alternative to tissue biopsy. One of the major challenges in ctDNA analysis lies in its relatively low concentration and the difficulty in detecting ctDNA from the background cell-free DNA fragments derived from normal cells (cfDNA). Despite some recent progress, accurate detection methods remain elusive. In this paper, I develop two types of probabilistic classifiers for distinguishing ctDNA from cfDNA. The performance of the proposed classifiers is evaluated and measured by the receiver operating characteristic (ROC) curve. Its accuracy is demonstrated by the area under the ROC curve. Both types of classifiers are easy to compute and fairly accurate with the potential to become relatively cheap and applicable tools for the early detection of cancers.

Key Words: Logit classifier, ROC curve, Likelihood Ratio Classifier, Negative binomial distribution

1. Introduction

There has been a growing interest in the use of “liquid biopsy” technique to screen for early-stage cancers and monitor treatment responses. Liquid biopsies involve the detection of cancer biomarkers in bodily fluids such as blood and urine [1]. Recent advances in liquid biopsy techniques have involved detecting the presence of tumor-derived cell-free DNA (circulating tumor DNA, ctDNA) in a simple blood test. This detection procedure is minimally invasive and provides an attractive and reliable alternative to tissue biopsy, which has limited success and associated complications. One of the major challenges in ctDNA analysis lies in its relatively low concentration and the difficulty in detecting ctDNA from the background cell-free DNA fragments derived from normal cells (cfDNA) [2, 3].

One way to distinguish ctDNA from cfDNA relies on identifying cancer hot-spot mutations in ctDNA and much progress has been made in this area, facilitated by the advance in PCR (polymerase chain reaction, e.g. digital PCR) and specialized next generation sequencing techniques (e.g. targeted sequencing). Nevertheless, there are several disadvantages associated with these assay-based approaches. First, point mutations are overall sparse and often unevenly distributed in human genome. Second, they are relatively expensive and time-consuming. Third, prior knowledge of the cancer mutation profile is often instrumental in achieving the desired sensitivity [1].

In this paper, using publicly available ctDNA sequencing data, I have developed two types of probabilistic classifiers to distinguish cancerous ctDNA from normal cfDNA. They are based on features such as fragment length and sequence content of the DNA fragment with the potential to incorporate additional predictors as more distinguishing properties of ctDNA become available. The performance of the proposed classifiers is evaluated and

[*] Texas Academy of Mathematics and Science at the University of North Texas, 1155 Union Circle, Denton, TX 76203

measured by the receiver operating characteristic (ROC) curve. Its accuracy is demonstrated by the area under the ROC curve (AUC). In addition, my classifiers do not rely on cancer hot-spot mutation detection and are generalizable to any form of cancer requiring only training data to produce an initial model and bypassing the need for costly sequencing techniques. Consequently they have the potential to become relatively cheap and applicable tools for the early detection of cancers, monitoring recurrence, and evaluating responses to therapy.

1.1 DNA Sequence Data

I obtained the raw sequence data used in this paper from Sequence Read Archive (SRA) website hosted by the National Center for Biotechnology Information [4]. The sequence data from two sequence files (SRA accession code: SRP040228) were used to build statistical models and test my classifiers. One file contains the ctDNA from the cell lines of human non-small cell lung cancer. The other file contains the cfDNA from a healthy control. Since the sequences downloaded from SRA are raw sequencing reads, I used the Barrows-Wheeler Aligner [5] and mapped the reads against a reference genome. The reference genome is hg19 downloaded from the genome browser at University of California Santa Cruz [6]. The alignment files were further processed and merged using software from SAMTools [7] and bedtools [8] to produce the DNA fragments with information regarding their chromosomal position, length, and mutations. Over 192 thousand ctDNA and 720 thousand cfDNA fragments were produced after the alignment and processing. These sequences were used in my subsequent model building and classifier development.

1.2 Features of DNA Sequences

In order to build a statistical model to classify ctDNA, I needed to find features capable of distinguishing ctDNA from cfDNA. Based on the current research findings on ctDNA, the length of the DNA sequences is a potentially important feature. In addition, there are other biological meaningful features that may be relevant in classifying ctDNA. For example, the GC-content of the DNA sequences. The four nucleobases found in DNA sequences are adenine (A), cytosine (C), guanine (G) and thymine (T). The bases are covalently linked together in a chain through the sugars and phosphates forming the backbone of DNA double helix. Hydrogen bonds between the bases hold the two chains together. GC-content of a DNA fragment is the percentage of bases that are either guanine or cytosine. GC-content not only varies considerably among genomes from different species, it can also differ greatly within the same genome. Higher GC-content is usually associated with increased thermo-stability by virtue of base stacking and the presence of a triple hydrogen bond between GC base pairs, compared with a double hydrogen bond between AT base pairs. Regions with higher GC content also tend to have higher relative gene density than regions with lower GC content. Therefore, variations in GC-content could be used to potentially distinguish ctDNA from cfDNA. The GC-content of DNA fragments is expressed as a percentage using the following standard formula: GC-content= (G+C)/(A+T+G+C).

2. Probabilistic Classifiers

2.1 Exploratory Data Analysis

To investigate which features are important in distinguishing ctDNA from cfDNA, I started some exploratory data analysis of the potential features of GC-content and the length of the DNA sequences. For both ctDNA and cfDNA datasets, I computed some numerical summary measures for each of these features as given in Table 1 and Table 2.

Table 1: GC Content

DNA Sequences	Min	1st Quartile	Median	Mean	3rd Quartile	Max
ctDNA	0.1089	0.4059	0.4828	0.4771	0.5484	0.7742
cfDNA	0.0707	0.4083	0.4800	0.4787	0.5461	0.8911

Table 2: Fragment Lengths

DNA Sequences	Min	1st Quartile	Median	Mean	3rd Quartile	Max
ctDNA	88	93	101	130.2	146	12520
cfDNA	94	150	166	170.8	180	11320

Table 1 shows that numerical measures such as sample means, medians, first quartiles, and third quartiles for the GC-content variable are almost identical for both ctDNA and cfDNA datasets. These results suggest that any procedure based on the summary measures alone will not be effective in detecting the difference between ctDNA and cfDNA. On the other hand, as seen in Table 2, there are significant differences between ctDNA fragment lengths and cfDNA fragment lengths. The mean length of ctDNA and cfDNA is about 130 bp (base pairs) and 171 bp, respectively. These numbers are consistent with existing findings in the literature [2, 9]. Furthermore, Table 2 shows that the lengths of ctDNA sequences are much shorter than the lengths of cfDNA across all measures of quartiles, not just the mean/median lengths.

2.2 Logistic Regression

Logistic regression is one of the most widely used statistical analysis and inferential tools to explore and model the relationship between a categorical response variable and predictor variables. My response variable is categorical with two categories coded via a class indicator variable taking two discrete values with 1 representing ctDNA and 0 representing cfDNA. My predictor variables are GC-content and fragment lengths of the DNA sequences. To predict whether a given DNA sequence is ctDNA or cfDNA, I used the logistic regression to model the posterior probabilities of the two classes (ctDNA and cfDNA) via a linear function in my feature variables. The model is specified in terms of the logit transformation or the log-odds: the logarithmic ratio of the probability that the tested sequence is ctDNA to the probability that it is cfDNA given the feature variables. More specifically, to examine and quantify which of my feature variables are statistically important for classifying ctDNA, I used the R base package and ran the logistics regression of my class indicator variables on the linear combination of the GC-content and fragment length feature variables. However, this direct logistic regression approach caused numerical failures leading to the estimated probability either exactly equal to 0 or 1. Upon close examination, I found out that the failures were caused by some of excessively outlying fragment lengths as shown by the boxplots and histograms in Figure 1 and Figure 2

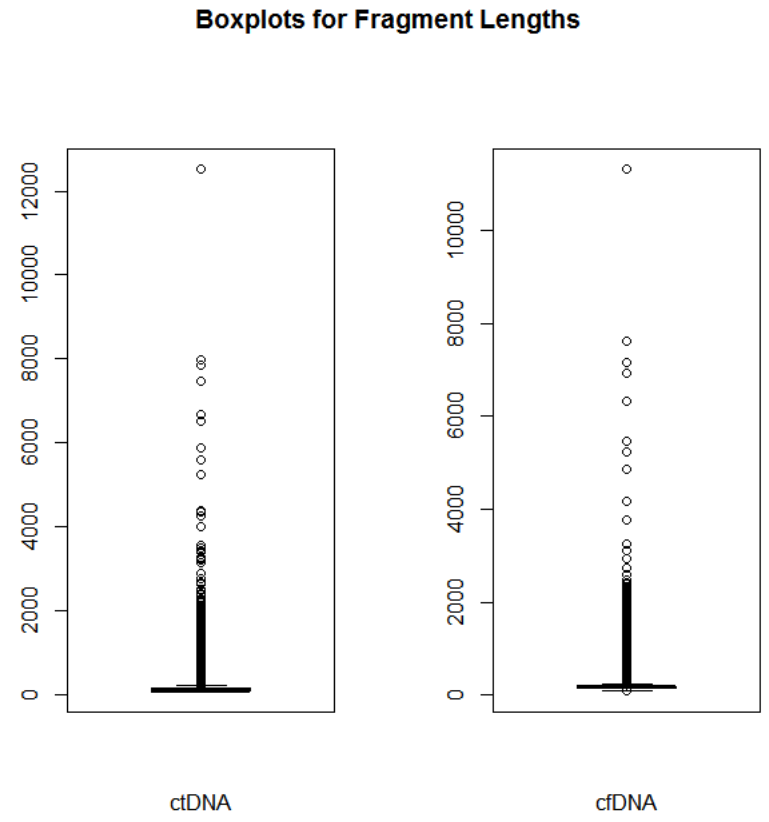

Figure 1: Boxplots for Fragment Lengths of ctDNA and cfDNA

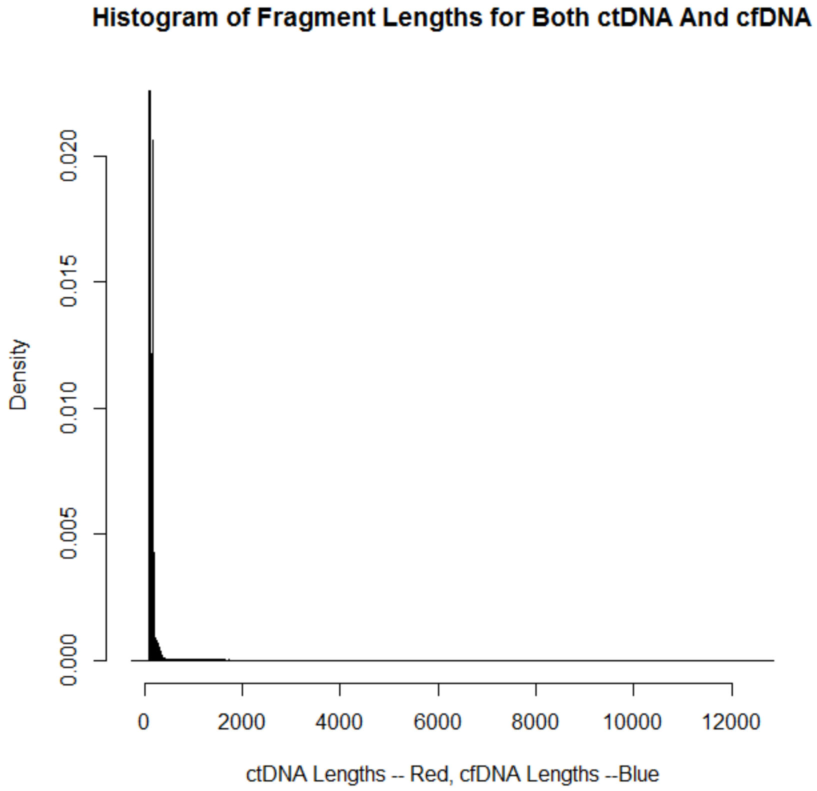

Figure 2: Histograms for Fragment Lengths of ctDNA and cfDNA

To overcome these difficulties, I made the logarithmic transformation of the fragment lengths, which significantly improved the separation of the fragment length distributions between ctDNA and cfDNA as seen in Figure 3. Let $X_1 = \log(\text{Fragment Length})$ and $X_2 =$ GC-content. Then the logistic regression is specified by the logit model:

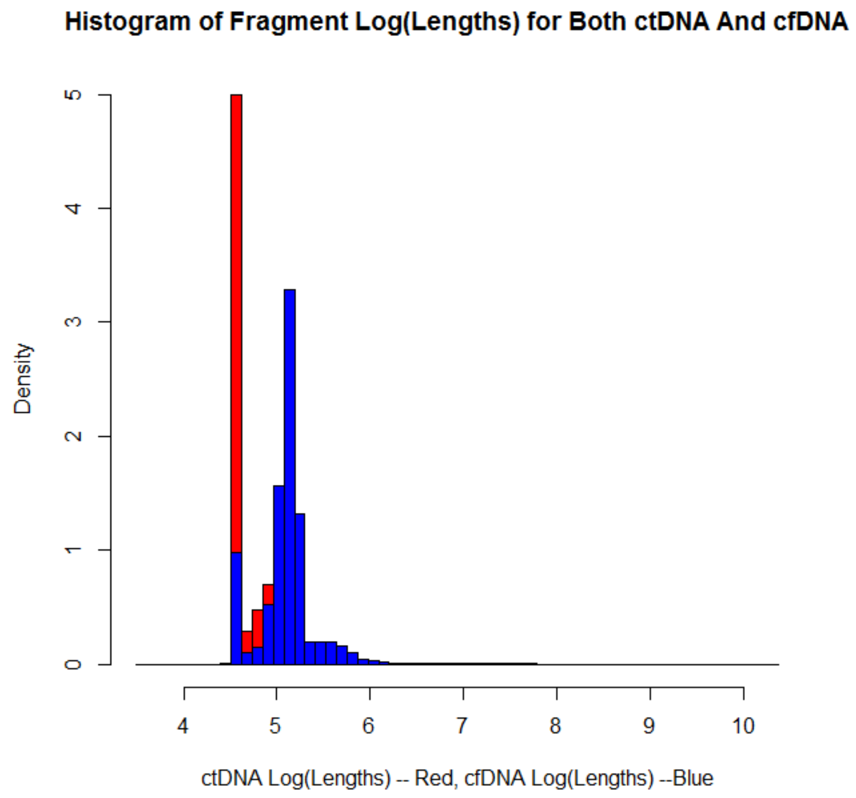

Figure 3: Histograms for Log Fragment Lengths of ctDNA and cfDNA

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{1}$$

where

$$p = P(Y = 1|X_1, X_2), \quad Y = \begin{cases} 1 & \text{if the sequence is ctDNA} \\ 0 & \text{if the sequence is cfDNA .} \end{cases}$$

To develop a probabilistic classifier based on logistic regression, I partitioned the DNA sequence data randomly into training data set (75% of the data) and testing data set (25% of the data). Using the training data, all unknown model parameters were estimated by the maximum likelihood (ML) method. Table 3 displays the results of my logistic regression.

As seen from Table 3, both feature variables have statistically significant Z scores with p-values less than 2×10^{-16}. Each of these Z scores corresponds formally to a Wald test of the null hypothesis that the coefficient is zero. A nonsignificant Z score would suggest that the corresponding coefficient can be dropped from the model. Since the Z scores for both log(Fragment Length) and GC-content have extremely small p-values, they provide strong evidence that these coefficients are not zero.

In addition, the signs of these estimated coefficients are also consistent with the log odds interpretation. For example, the negative sign of the estimated coefficient (-4.41) of log(Fragment Length) suggests that holding GC-content constant, for every increase

Table 3: Significance of log(Fragment Length) and GC-content

Coefficients	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	20.28	0.065	309.95	$< 2 \times 10^{-16***}$
log(Fragment Length)	-4.41	0.013	-336.12	$< 2 \times 10^{-16***}$
GC-content	0.35	0.033	10.68	$< 2 \times 10^{-16***}$

of 1 unit in log(Fragment Length), the odds of the sequence as ctDNA decreases by a factor of $\exp\{-4.41\}$. In other words, for any given DNA sequence, the larger the value of log(Fragment Length) is, the less likely it is ctDNA sequence, adjusting for the other feature variable.

Given the fitted logistic regression, the classification rule is based on the estimated logistic probability $\hat{p}$. My decision rule is that if $\hat{p} > 0.5$, classify the sequence to be ctDNA; otherwise, classify it to be cfDNA. The performance of this logit classifier was evaluated by the testing data. Figure 4 shows the performance of the classifier as measured by the receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) for the logit classifier is 80% and the classification accuracy is 82%.

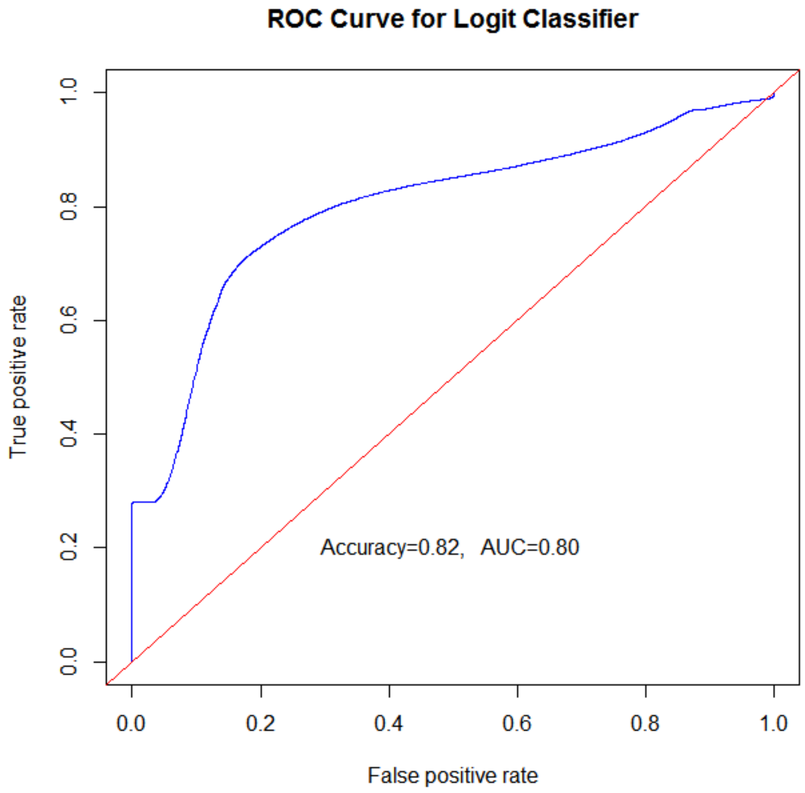

Figure 4: ROC Curve for Logistic Regression Classifier

2.3 Relative Importance of Predictors

Given both log(Fragment Length) and GC-content are highly significant, it would be worth investigating the relative importance of these two predictors. To assess the sole effect of GC-content on the response variable, I used the training data and ran the logistic regression of my indicator variable on the GC-content predictor alone and the resulting output is provided in Table 4.

Table 4: Significance of GC Content

Coefficients	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	-1.26	0.015	-86.18	$< 2 \times 10^{-16***}$
GC-content	-0.15	0.030	-5.08	$3.71 \times 10^{-7***}$

As seen from Table 4, GC-content is highly statistically significant and plays a significant role in explaining the categorical response variable, which confirmed previous findings in Table 3. To assess its predictive power, I used the testing data to evaluate its performance in terms of ROC curve.

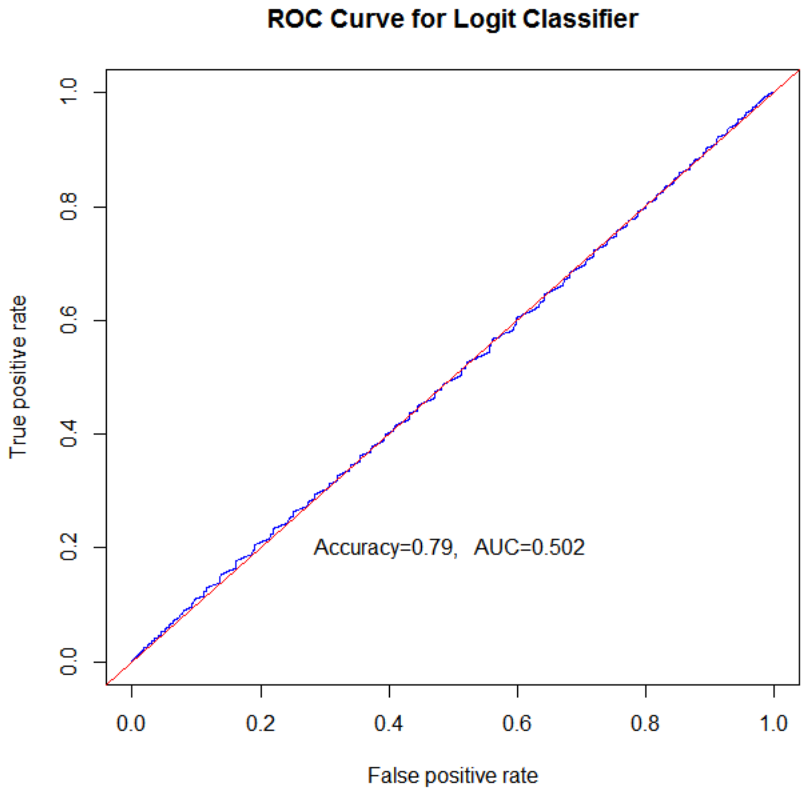

Figure 5: ROC Curve for Logistic Regression Classifier Based on GC Content

Figure 5 shows that the ROC curve for the classifier based on GC content almost coincides with the chance diagonal indicating that its accuracy as measured by the AUC of this classifier is slightly better than random guessing. Consequently the statistical significance of GC-content did not translate into predictive power in classifying ctDNA. This finding is also consistent with the largely overlapping histograms given in Figure 6. Similar analysis of log fragment lengths showed that the log fragment lengths was the dominant predictor in classifying ctDNA.

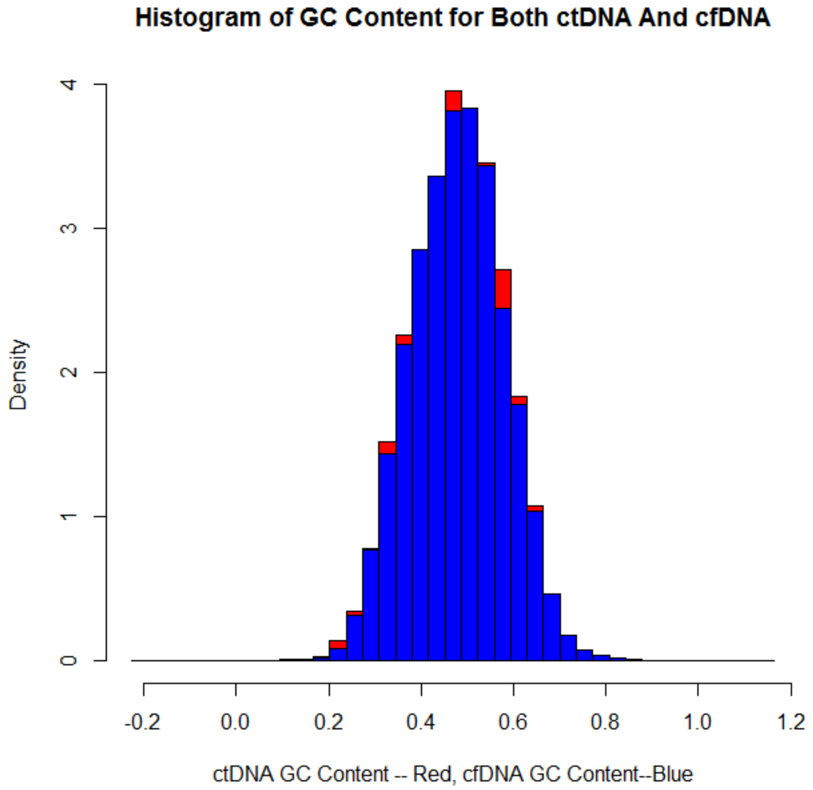

Figure 6: Histogram of GC Content for both ctDNA and cfDNA

3. Likelihood Ratio Classifier

3.1 Negative Binomial Distributions for Modeling DNA Fragment Lengths

As we have seen, the length of DNA sequences plays a major role in tumor DNA classification, it would be interesting to model fragment lengths directly using some parametric distributions. As seen in Figure 1 and Figure 2, the distributions of fragment lengths are highly skewed by some extremely large observations leading to huge variations. In addition, fragment lengths as measured in base pairs are positive integers and the probability that fragment lengths take the zero value is zero. These considerations motivated me to use the negative ninomial distribution for modeling DNA fragment lengths. Let X denote the fragment length. I assme that X has the negative binomial distribution with the probability mass function given by

$$f(x|r,p) = P(X = x|r,p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \cdots, \tag{2}$$

Specifically, I assume $X \sim f(x|r_{ct}, p_{ct})$ for ctDNA and $X \sim f(x|r_{cf}, p_{cf})$ for cfDNA.

My classification rule is based on the likelihood ratio of these probabilities

$$LR(x) = \frac{P(X = x|\text{ctDNA})}{P(X = x|\text{cfDNA})} = \frac{f(x|r_{ct}, p_{ct})}{f(x|r_{cf}, p_{cf})} \tag{3}$$

For any given DNA sequence with length x, if $LR(x) > 1$, the sequence is classified as ctDNA, otherwise, the sequence is classified as cfDNA. In practice, the negative binomial model parameters are unknown and need to be estimated based on the training data. They can be estimated either by the method of moments (MM) or maximum likelihood (ML) method.

3.2 MM and ML Parameter Estimation

The MM estimates follow easily from the mean and variance of X, which are given by $E(X) = r/p$ and $Var(X) = r(1-p)/p^2$, respectively. Consequently, given the training data $X_1, \cdots, X_n$, the MM estimates for r and p are computed as

$$\tilde{r} = \bar{X}\tilde{p}, \quad \tilde{p} = \frac{\bar{X}}{\bar{X} + \sum_{i=1}^{n} X_i^2/n - \bar{X}^2} \tag{4}$$

The ML estimates $\hat{r}$ and $\hat{p}$ for r and p are given by $\hat{p} = \hat{r}/\bar{X}$, where $\hat{r}$ is obtained by solving the following log likelihood equation:

$$\log(\frac{\hat{r}}{\bar{X} - \hat{r}}) - \psi(\hat{r}) + \frac{1}{n}\sum_{i=1}^{n} \psi(X_i - \hat{r} + 1) = 0 \quad \text{for } 1 \le \hat{r} \le \min_{1\le i\le n}\{X_i\}, \tag{5}$$

where $\psi(x)$ denotes the digamma function. Since the equation cannot be solved for $\hat{r}$ in closed form, it is solved numerically by root-finding algorithms such as Brent's method. I applied both MM and ML methods to estimating the negative binomial model parameters based on the training data (75% of my data) and used the remaining 25% of the data for testing the likelihood ratio classifiers. Their performance and accuracy are measured by the ROC curves and the AUCs shown in Figure 7 and Figure 8. These results demonstrated that the likelihood ratio classifier based on ML estimates of the model parameters outperformed the likelihood ratio classifier based on MM estimates.

4. Conclusions

In this paper, I have developed two types of probabilistic classifiers for distinguishing ctDNA from cfDNA. The first type of probabilistic classifiers is the logit classifier based on the logistic regression. The logit classifier incorporates the log fragment lengths and GC-content of DNA sequences as distinguishing feature variables in the logistic regression model with the potential to include additional predictors as more distinguishing properties of ctDNA become available. The second type of probabilistic classifiers is the likelihood ratio classifier based on the negative binomial distributions for modeling the fragment lengths. The performance of both types of classifiers is evaluated and measured by the receiver operating characteristic (ROC) curve. Its accuracy is demonstrated by the area under the ROC curve. Both types of classifiers are easy to compute and fairly accurate. Consequently they have the potential to become a relatively cheap and applicable tool for the early detection of cancers.

References

[1] Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, Pacey S, Baird R, and Rosenfeld N *Liquid biopsies come of age: towards implementation of circulating tumour DNA*, Nat Rev Cancer **14**, 223-238, 2017

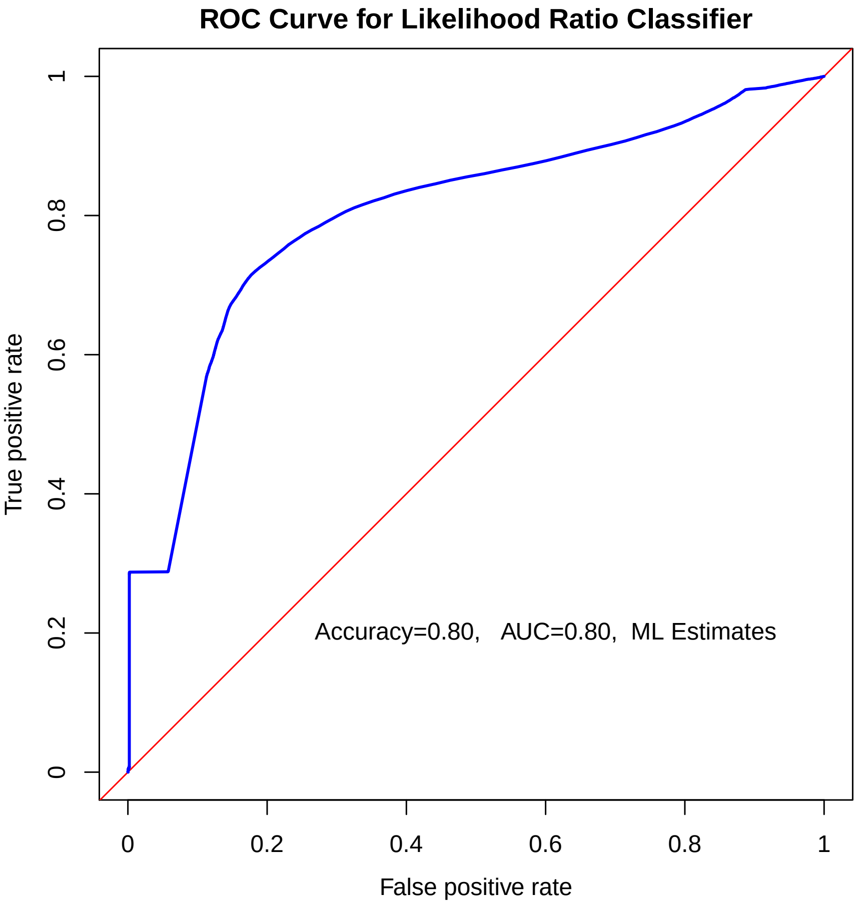

Figure 7: Negative Binomial Models Estimated by ML Method

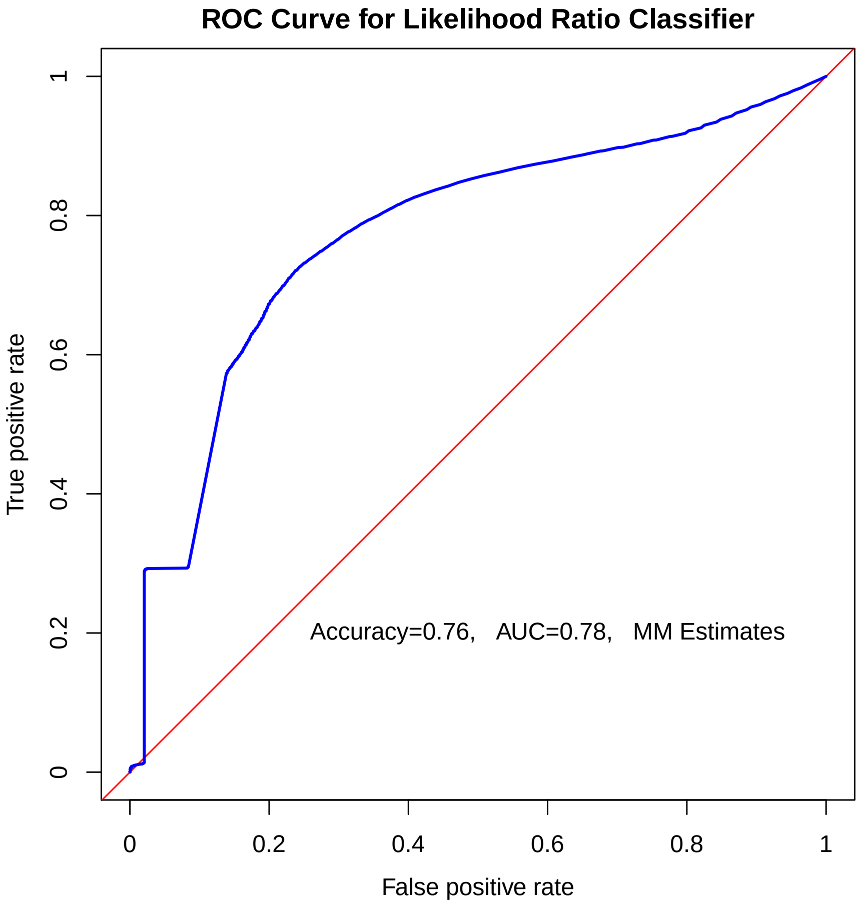

Figure 8: Negative Binomial Models Estimated by MM Method

[2] Mouliere F, Robert B, Peyrotte E, Del Rio M, Ychou M, Molina F, Gongora C and Thierry, A R *High fragmentation characterizes tumour-derived circulating DNA*, PLoS ONE **e23418**, 2011

[3] Mouliere, F, El Messaoudi, S, Pang, D,Dritschilo, A and Thierry, A R *Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer*, Mol. Oncol. **8**, 927941 , 2014

[4] Leinonen R, Sugawara H, and Shumway M, *The Sequence Read Archive*, Nucleic Acids Res. **39**, D19D21, 2011

[5] Li H and Durbin R *Li H. and Durbin R.*, Bioinformatics **26**, 589-95, 2010

[6] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D *The human genome browser at UCSC*, Genome Res. **12**, 996-1006, 2002

[7] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup *The Sequence alignment/map (SAM) format and SAMtools*, Bioinformatics **25**, 2078-9, 2009

[8] Quinlan AR and Hall IM *BEDTools: a flexible suite of utilities for comparing genomic features*, Bioinformatics **26**, 841842, 2010

[9] Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, Wong VW, Wong GL, Chan SL, Mok TS, Chan HL, Lai PB, Chiu RW, Lo YM.*Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients*, Proc Natl Acad Sci U S A. **112**, E1317-25, 2015