Predicting DNA methylation using fragmentation patterns in cell-free DNA

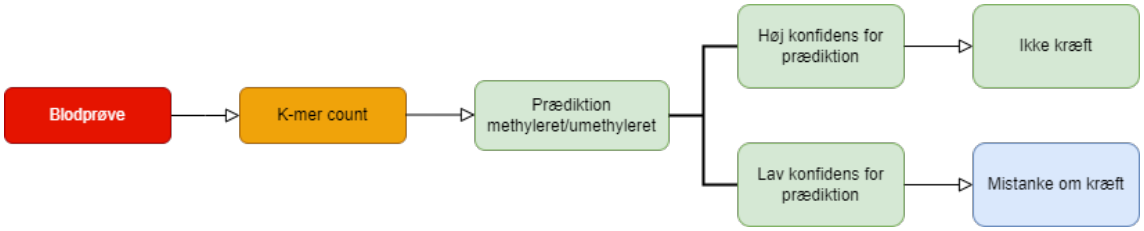
Daniel Jensen, Thomas Brejner

maj 2024

Hvad går projektet overordnet ud på?

I dette projekt vil vi lave en model for at bestemme om en person muligvis har kræft ved at analysere data fra en blodprøve (Se "figure1"). Dette kan bestemmes ved hjælp af *cfDNA* (cell-free DNA) og *ctDNA* (circulating tumor DNA), hvilket kan analyseres ved hjælp af en blodprøve. Dette er relevant idét det både er mindre invasivt og billigere end konventionelle metoder brugt til at diagnosticere kræft, som for eksempel CT skanninger, røntgen billeder, og andre biopsier. Endvidere at det kun kræver analyse fra en almindelig blodprøve så der ikke behøver at være tidligere mistanke om kræft.

Figure 1: *Proces flowchart*



Rød: Bliver taget i et laboratorie, Orange: Blodprøven er blevet sekvenseret og vi får de rå k-mer counts, som vi så normaliserer, Grøn: Vi kører dataen gennem vores kode, Blå: Tilbage til laboratoriet.

Hvad er den faglige problemstilling?

Tidligere forskning har vist at DNA har forskellige fortrukne *fragmenterings områder*[1] (Se figure "figure 2") hvis det kommer fra et *methyleret*[2] eller ikke-methyleret område. Vi kan benytte os af den forskel for at identificere om et område er methyleret eller ikke-methyleret, baseret på hvor høj counten på de forskellige *k-mer*[3] (Se "figure 3" for et eksempel på hvordan nogle 4-mers kunne se ud) er. Hvis man har en kræftknude bliver dens ctDNA samlet op sammen med det almindelige cfDNA, så der kommer DNA som ikke burde være der. Dette betyder at for en klassifiseringsmodel som kan bestemme om området er methyleret eller ikke-methyleret vil vi have en størrer usikkerhed i vores prædiktion om

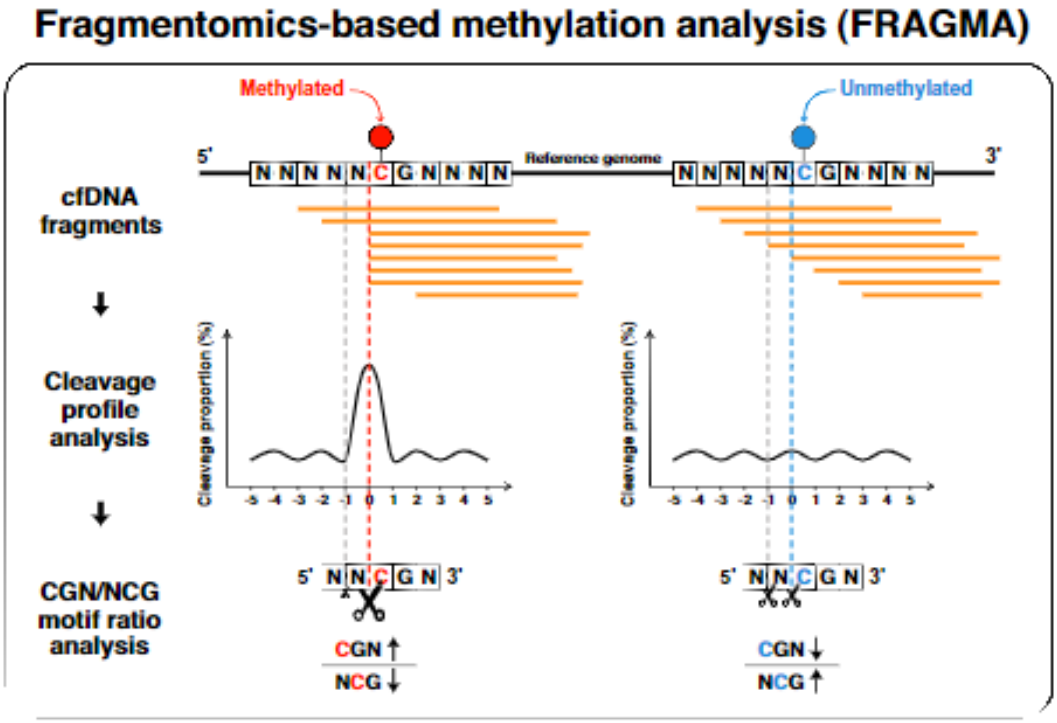
[1] Et fragmenterings område er den base hvor DNA sekvensen "knækker" ved

[2] Methylering er en biologiske process hvor der bliver tilføjet en methyl gruppe. For dette projekt er det ikke vigtigt at vide noget om processen, men kun, at der er forskel på methyleret og ikke-methyleret områder

[3] En k-mer er en lille del af en DNA sekvens på størrelse k

methyleringen i området. På en mere faglig formulering, så har vi en methyleret fordeling, en umethyleret fordeling, og diverse kræft fordelinger. Hvis man ikke har kræft kommer samplen fra enten den methyleret eller umethyleret fordeling. Hvis man har kræft så har man en mixturefordeling med enten den methylerede eller umethylerede og én kræft fordelingen. Mixturenfordeling vil have et højt bidrag fra den methyleret/umethyleret fordeling, og et lavt bidrag fra kræftfordeligen. Først vil vi lave en model som kan bestemme om en given sample er fra enten den methyleret eller umethyleret fordeling. Derefter ville vi finde et idealt konfidens cutoff for vores prædiktioner for methylering, for at klassificer kræft og ikke kræft. Problemstillingen ligger i hvordan vi bestemmer om et område er methyleret eller ikke-methlyeret, og dernæst hvor lavt vores konfidens cutoff for methyleret eller ikke-methyleret skal være.

Figure 2: *Viser hvordan methylerede og ikke-methylerede områder har forskellige præferede endepunkter*



Hvilke data kigger I på, og hvad skal man være særlig opmærksom på?

Dataet er en lang liste af samples for 475 personer. Hvor 244 er raske og 231 har diverse former for kræft. Inde for hver sample har vi igen en liste som viser alle de mulige k-mer for de samlede DNA fragmenter samlet i samplen. Hvis samplen kommer fra nogen med kræft vil der være nogle "støj" DNA fragmenter, som blev taget med i analysen (se "figure 4). Vi har 3 forskellige størrelser af fragmentering vi kan kigge på: 2, 4 og 6 -mer. Man skal være opmærksom på at vi også har en "background" fil for hver k-mer, der detaljerer alle de mulige fragmenteringer for hele DNA sekvensen.

Figure 3: *Eksempel der viser alle 4-mer for en lille DNA sekvens*

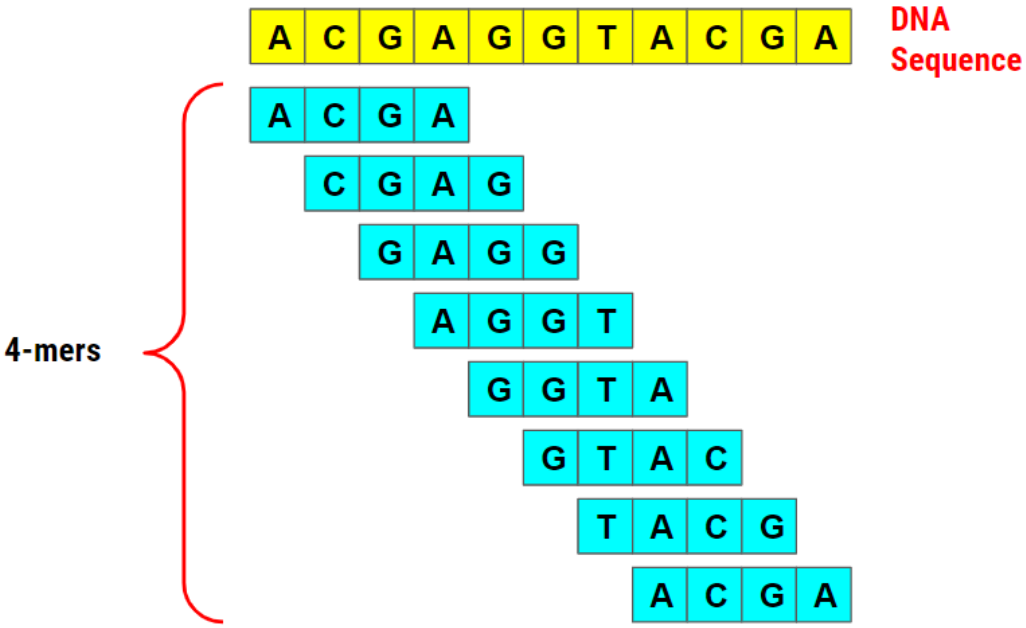
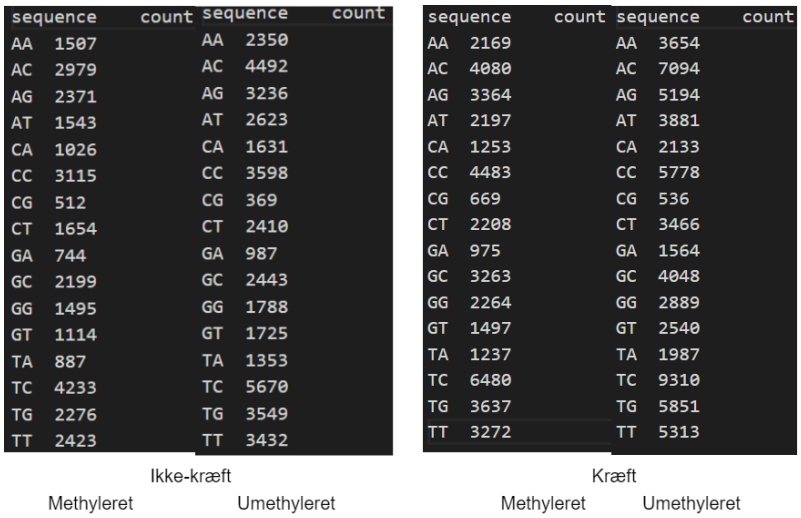


Figure 4: *Et eksempel på hvordan methylerede og umethylerede counts ser ud for en rask sample og en kræft sample.*



sequence	count	sequence	count
AA	1507	AA	2350
AC	2979	AC	4492
AG	2371	AG	3236
AT	1543	AT	2623
CA	1026	CA	1631
CC	3115	CC	3598
CG	512	CG	369
CT	1654	CT	2410
GA	744	GA	987
GC	2199	GC	2443
GG	1495	GG	1788
GT	1114	GT	1725
TA	887	TA	1353
TC	4233	TC	5670
TG	2276	TG	3549
TT	2423	TT	3432

Ikke-kræft

Methyleret Umethyleret

sequence	count	sequence	count
AA	2169	AA	3654
AC	4080	AC	7094
AG	3364	AG	5194
AT	2197	AT	3881
CA	1253	CA	2133
CC	4483	CC	5778
CG	669	CG	536
CT	2208	CT	3466
GA	975	GA	1564
GC	3263	GC	4048
GG	2264	GG	2889
GT	1497	GT	2540
TA	1237	TA	1987
TC	6480	TC	9310
TG	3637	TG	5851
TT	3272	TT	5313

Kræft

Methyleret Umethyleret

Hvad håber i at finde ud af?

Som tidligere nævnt håber vi på at finde ud af om vi kan bruge en blodprøve til at bestemme hvorvidt en person har kræft, på en langt mindre invasiv måde end man normalt ville gøre. Vi håber på at det vil være muligt at kunne se forskel på konfidensen i modellens prædiktioner hvis det kommer fra en person med kræft.

Er der nogle særlige udfordringer?

Vi antager at det er denne forskel mellem methyleret og ikke-methyleret, og at vi kan se den i vores blodprøve. Vi havde nogle problemer med at lave vores methylering model, som skal klassificere mellem methyleret og umethyleret. Vi forsøgte med nogle forskellige klassifikations metoder til vores methylerings model og endte med en *L1 Regularized Logistic Regression*, med *Cross Validation*. Dette mindsker overfitting og sørger for at vi får en bedre model. Så antager vi også at når vores første model giver et laverer konfidens i klassificeringen, at det betyder at der er muligvis er tale om kræft. Vi vidste også ikke præcis hvordan vi ville finde et godt cutoff for konfidens men endte med og bruge en "ROC curve" analyse.

Hvad skal projektet munde ud i?

Vi har lavet et Github-repository[4] som også ville inkludere en rapport som dokumentérer hvad vi har gjort og hvorfor, samt forklarer det kode vi har skrevet.

[4] https://github.com/dwarfy35/Dataproject