Navne: Julie Malmkjær, Mathias Pedersen og Naja Andrup Pedersen
Vejleder: Kristoffer Pors Jakobsen og Søren Lykke Andersen (MBG)

Identifikation af RNA-modifikationer i polyA-haler med Nanopore-sekventering (og machine learning)

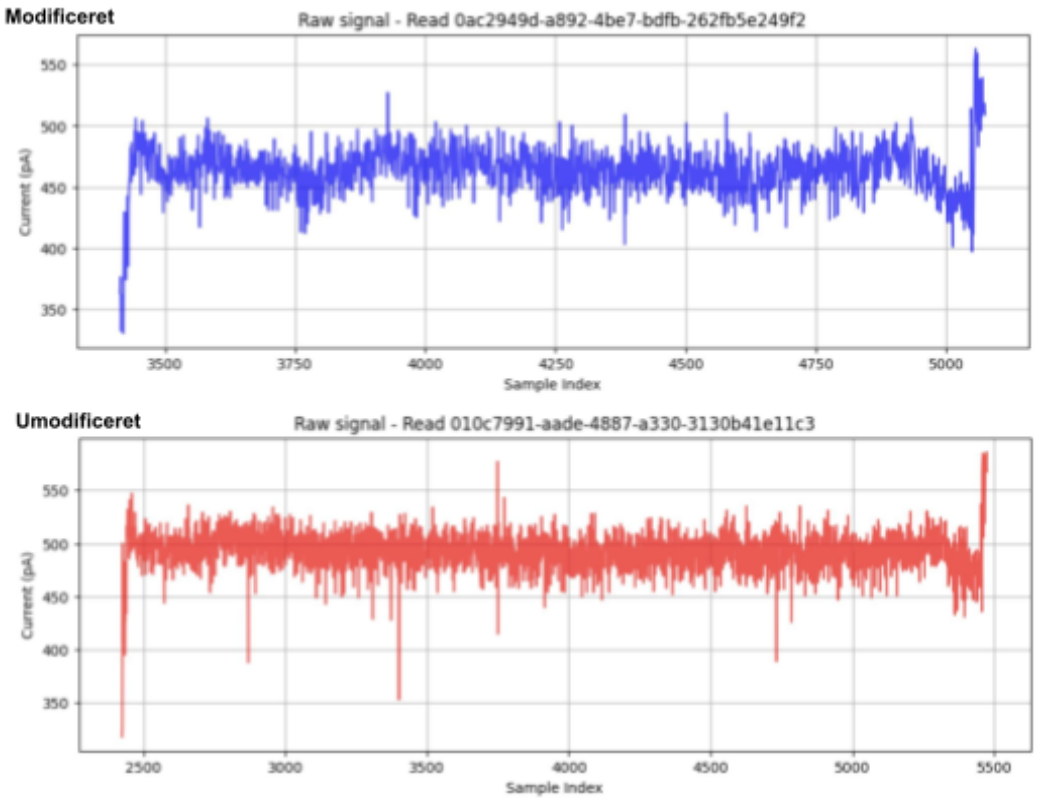
Formålet med dette projekt er at udvikle en model, der kan identificere og lokalisere (specifikke) modifikationer i polyA-halen af messenger RNA (mRNA) ved hjælp af signaldata fra Nanopore-sekventering.

PolyA-halen er en lang sekvens af adenin (A)-nukleotider, som sidder i slutningen af de fleste mRNA-strenge, hvor den spiller en central rolle i mRNA-stabilitet. Modifikationer sat på nukleotider i polyA-halens sekvens kan have biologisk betydning for f.eks. proteindannelse, og det kan have potentielle terapeutiske anvendelser.

For at kunne finde de specifikke modifikationer, vil vi analysere de elektriske signaler fra Nanopore-sekventering og forsøge at skelne mellem modificerede og umodificerede nukleotider.

Den faglige problemstilling

Nanopore-sekventering kan registrere nukleotidsekvenser, herunder identificere polyA-halen, men de rå signaldata er komplekse og støjfyldte. Dette gør det vanskeligt at skelne mellem modificerede og umodificerede nukleotider. Udfordringen ligger derfor i at udvikle en model, der kan netop det på trods af variation og støj i dataene.



Figur 1: Eksempel på rå signal data plottet med picoampere for hver måling (sample), hvor øverste aflæsning er fra en mRNA-streng med modifikation, og den nederste har ingen modifikation.

Navne: Julie Malmkjær, Mathias Pedersen og Naja Andrup Pedersen
Vejleder: Kristoffer Pors Jakobsen og Søren Lykke Andersen (MBG)

Data

Dataen kommer fra syntetiske mRNA-strenge fremstillet i et laboratorium, hvor længden og positionerne af eventuelle modifikationer på polyA-halen på forhånd er kendt.

Der er datasæt, hvor polyA-halen har en længde på 60 baser, enten uden modifikation eller med en enkelt modifikation på den midterste eller sidste base. Derudover er der også datasæt, hvor polyA-halen er 120 baser lang og enten har ingen, én, to eller 4 modifikationer tilfældigt placeret på 6 positioner.

Efter Nanopore-sekventering består datasættene af de målte elektriske signaler, fra da mRNA-streng blev trukket igennem sekventeringsenheden. Det er ændringer i de elektriske signaler som rådataene består af. Rådataene konverteres til et læsbart format, og signalværdierne gemmes som en lang sekvens, hvor start- og slutpunktet for polyA-halen er angivet. Når målingerne fra en mRNA-streng visualiseres, så ser det ud som i Figur 1, hvor førsteaksen repræsenterer signalernes rækkefølge, mens andenaksen angiver den målte elektriske signalstyrke.

Udfordringer

De største udfordringer ved dataene er:

- **Støj og variation**: Signaldata fra Nanopore-sekventering er støjfyldte, da de afhænger af eksperimentelle forhold i laboratoriet. En del af støjen kan skyldes, at mRNA er meget skrøbeligt og skal håndteres forsigtigt, så det ikke går i stykker.
- **Varierende signalrepræsentation**: Selvom polyA-haler kan have samme længde i nukleotider, så kan deres signaldata variere betydeligt. For eksempel er der et datasæt med polyA-haler af ens længde (60 nukleotider), som har signaldata der spænder fra under 300 målepunkter til over 23.000 målepunkter.

Disse udfordringer kan gøre det vanskeligt at udvikle en model, der fungerer på tværs af naturligt forskellige datasæt.

Projektets mål

Det endelige mål er at udvikle en model, der kan identificere og lokalisere (specifikke) modifikationer i polyA-halen med data fra en Nanopore-sekventering.

Grundet udfordringerne med meget varierende og støjende data, så er første delmål at klargøre et mere ensartet datasæt med vektorer, som kan anvendes til en model, der kan identificere modifikationerne.