Application and evaluation of different unsupervised machine-learning techniques on single cell data

Anna Sofie Christensen, Elisa Lægsgaard, Marcus Olesen & Sofie Schubert Elving

Baggrund

I 2019 muterede coronavirus ud i en ny gren, som fik navnet SARS-CoV-2. Versionen berørte kort efter store dele af verdens befolkning med sygdom og isolation. Mutationen gjorde virussen i stand til nemt at infiltrere immunforsvaret og bruge værten til at sprede sig. De inficerede celler viste store ændringer i fordelingen af, hvilke gener cellen udtrykte.

Det menneskelige genom har omkring 20.000 gener, og ved brug af en teknik kaldet RNA-sekventering kan man måle deres "genekspression", et mål for hvor meget hvert gen bliver udtrykt for hver individuelle celle (single-cell data).

I projektet ønskes der at sammenligne og vurdere hvordan forskellige unsupervised machine learnings-teknikker kan benyttes til at undersøge genomdata og dets strukturer, og potentielt visualisere de celleafhængige ændringer, der sker ved infektion, for bedre at kunne forstå COVID-19 og hvordan det påvirker immunforsvaret.

Herunder anvendes og undersøges standard dimensions-reducerende metoder til visualiseringen af dataet som f.eks. PCA og t-SNE, for at undersøge strukturen af immunlandskabet samt de ændringer, der sker under infektion. I denne forbindelse benyttes og evalueres forskellige clustering teknikkers evne til at identificere den specifikke immunforsvars celletype samt graden af infektion i cellen.

Yderligere ønskes der at undersøge hvordan nyudviklede metoder baseret på neurale netværk fanger strukturer og grupperinger i dataet til sammenligning med standardteknikkerne.

Data

Til at undersøge problemstillingen har vi fået adgang til et single-cell datasæt bestående af omkring 1,4 millioner raske celler, samt et datasæt med alle de inficerede celler bestående af ca. 3.000 celler. Genom-dataet består af mange høj-dimensionelle observationer med rig struktur. Hver observation indeholder information for en celleprøve, bestående af 20.000 gener samt anden information for prøven som infektionsstatus, celletype og karakteristika af patienter/kontroller. Hver celle indeholder alle menneskets gener, men forskellige celler benytter bestemte gener mere/mindre/slet ikke, alt efter celletype, arbejdsopgave og andre forhold, som for eksempel tilstedeværelse af vira i cellen, som kan ændre på genudtrykket.

Datasættets størrelse kræver adgang til f.eks. et computer cluster for at opbevare og arbejde med dataen. Det er derfor attraktivt at finde metoder til bedst at repræsentere høj-dimensionelle data i low-dimensional space, der stadig fanger de vigtige karakteristika og strukturer i dataet.

Udfordringer

Den største udfordring ligger i tværfagligheden af arbejdet. Det er ikke nok at forstå de forskellige dimensions-reducerende teknikker, man skal også have overblik over hvordan det

relaterer sig til “biologien” bag dataet. Derudover gør dataens høj-dimensionalitet det svært at bevare eller identificere interessante strukturer/karistika i dataet ved dimensionsreduktion. Herunder kan det være svært at få indblik i hvilke strukturer, som celletype, infektionsstatus, patientkarakteristika eller potentielt andre underliggende strukturer, som kommer til udtryk i en low-dimensional projektion.

Mål

Målet med projektet er at benytte de tidligere omtalte dimensionsreduktions-teknikker til bedst muligt at repræsentere data i færre dimensioner. Vi vil i den sammenhæng fremstille visualiseringer, der forhåbentligt kan afsløre underliggende karakteristika/strukturer som infektionsstatus, celletype mm. Vi kan, kvalitativt, evaluere i hvilken grad de forskellige teknikker opnår dette. Yderligere forventes det at udføre visualiseringer af forskellige clustering metoder, hvor disse evalueres og sammenlignes efter deres evner til at gruppere de dimensions-reducerede data efter underliggende grupperinger/opdelinger i dataet.

Helt konkret vil vi kigge på dimensionsreduktions-teknikkerne; PCA, t-SNE og UMAP. Til clustering vil vi kigge på K-means- og hierarchical clustering. Til neurale netværk vil vi kigge på single-cell Variational Inference (scVI). Og som præcisions-mål vil vi benytte adjusted rand index (ARI) til at sammenligne de resulterende grupper med de annoterede grupperinger givet i datasættet.

Der ønskes at arbejdet munder ud i en præsentation, der består af key-visualizations og tabeller, som illustrerer/opsamler resultaterne af de forskellige benyttede machine learnings-teknikker.

Dette forventes at blive delt via f.eks. et link til en GitHub-side.