Dansk resumé

High-throughput omics-teknologi har gjort det muligt at måle tusindvis af molekyler, såsom DNA, RNA eller proteiner, på samme tid og på den måde få et detaljeret indblik I cellens sammensætning. Molekylært profileringsdata er desværre komplekst og plaget af støj, og sofistikerede metoder er derfor nødvendige for effektivt at kunne opdage biologisk og klinisk relevante mønstre. Denne afhandling er en samling af separate manuskripter, der alle beskæftiger sig med udviklingen af beregningsmæssige metoder til at analysere komplekst biomedicinsk data.

Det første manuskript introducerer en itereret lokalsøgningsalgoritme til at finde den største fælles delgraf for to eller flere store netværker. Algoritmen optimerer et nyt kantbevaringsmål, der både kan finde helt og delvist bevarede kanter. Metoden gøres tilgængelig som en udvidelse til Cytoscape-platformen.

Det næste manuskript beskriver CoNVaQ, en metode til at foretage genomdækkende associationsstudier (GWAS) baseret på kopi-antal-varianter (CNV). Vores værktøj inkluderer en algoritme til at opdele CNV-kald i diskrete regioner, samt to modeller for at finde associationer: en statistisk test baseret på Fishers eksakte test og en forespørgselsbaseret model. Vi demonstrerer metoden ved at finde varianter forbundet med HPV-status i et peniscancer kohorte.

Det tredje manuskript er en systematisk evaluering af sammenhængen mellem regulerende interaktioner og målt genekspression i *E. coli*. Studiet undersøger antagelsen, at en op- eller nedregulering af en transkriptionsfaktor medfører en ændring i ekspressionen af de gener den regulerer. Vi observerer, at både fremmende og blokerende interaktioner er forbundet med en svag positiv korrelation, og tilfældige netværksmodeller stemmer lige så godt overens med ekspressionsdataet som det rigtige netværk. Vi diskuterer mulige årsager til denne konklusion.

Det fjerde manuskript introducerer en beslutningstræensemble-baseret metode til at integrere et molekylært interaktionsnetværk med genekspressionsdata for at finde et beriget delnetværk. Vi sammenligner metoden med andre moderne metoder, og observerer, at den finder interaktionstætte, mere biologisk relevante genmoduler. Disse resulter viste sig imidlertid at være drevet primært af bias i netværkstopologien og ikke relevante ekspressionsmønstre.

Det sidste manuskript beskriver et framework til at analysere genetiske varianter ved hjælp af en hierarkisk cellemodel. Vi konstruerer et genhierarki ud fra Gene Ontology og søger efter berigede cellebestanddele og -processer ved hjælp af en generaliseret regressionsmodel. Vi introducerer to betingede tests til at filtrere termer der er redundante eller berigede pga. et enkelt gen. Vi anvender metoden til at analysere et kronisk obstruktiv lungesygdom (KOL) kohorte og finder to mekanismer der ikke var fundet i standard GWAS. Vi diskuterer

yderligere begrænsningerne ved at bruge Gene Ontology til berigelsesanalyse pga. termstørrelser og bias.

For at opsummere, så udviklede vi metoder til at understøtte bioinformatisk analyse af biomedicinsk data. Vores resultater demonstrerer, at selvom beregningsmæssig analyse af molekylært profileringsdata, særligt kombineret med sekundær information fra netværks- eller pathway-databaser, kan være en stor hjælp til at opdage ny biologisk indsigt, så kan bias og tekniske artefakter i det underliggende data have stor indflydelse på resultaterne.