

DOKTORAND: Gro Nilsen
GRAD: Philosophiae doctor
FAKULTET: Det matematisk-naturvitenskapelige fakultet
INSTITUTT: Institutt for Informatikk
FAGOMRÅDE: Bioinformatikk
VEILEDERE: Ole Christian Lingjærde, Knut Liestøl, Ørnulf Borgan
DISPUTASDATO: 7. mai 2015

AVHANDLINGENS TITTEL: *Statistical learning in genomics: Uncovering patterns and groups in high dimensions*

Stipendiat Gro Nilsen har i sin doktoravhandling utviklet statistiske metoder for analyse av data som måler genetiske endringer i kreft. Metodene gjør det blant annet mulig å avdekke grupper av pasienter med ulike endringsmønstre.

Kreft er en sykdom som blant annet kjennetegnes av en rekke genetiske endringer i arvematerialet. Slike endringer medvirker til at kreftceller kan vokse og spre seg til andre organer. I mange studier er det av interesse å identifisere grupper av pasienter som deler mange av de samme forandringene, og å avdekke om disse gruppene for eksempel responderer forskjellig på behandling eller har forskjellig prognose. Data analysen i slike studier er imidlertid utfordrende fordi dataene er av svært høy dimensjon (man har flere titalls tusen til millioner av målinger for hver pasient) og målingene er unøyaktige. I sin avhandling har stipendiat Gro Nilsen utviklet statistiske metoder som gjør det mulig å redusere både støyen og dimensjonen på dataene, og som kan avdekke pasientgrupper med ulike endringsmønstre.

I første del av arbeidet foreslås en rekursiv metode som finner undergrupper blant eksisterende hovedgrupper. Slike undergrupper er blant annet kjent fra brystkreft, og i anvendelser av metoden på data fra ulike krefttyper så man at denne tilnærmingen kunne avdekke slike kjente undergrupperinger. Andre del av avhandlingen er spesifikt rettet mot analyse av såkalte kopitallsdata hvor man måler om deler av arvematerialet har blitt slettet eller duplisert i kreftcellene. Stipendiaten og medforfattere foreslår først en statistisk segmenteringsmetode som på en rask og nøyaktig måte identifiserer hvilke regioner som er endret og i hvilken grad. Dette bidrar til å redusere støyen i dataene samtidig som antall variable reduseres kraftig, noe som gjør det lettere å gjøre videre statistiske analyser. På bakgrunn av denne segmenteringen foreslås deretter et sett av matematiske mål som fanger opp ulike endringsmønstre i arvematerialet. Når denne metoden ble anvendt på data fra brystkreftpasienter fant man blant annet at disse målene kunne knyttes til de ulike undergruppene av brystkreft og brukes til å predikere pasientenes prognose.