Laskien pidemmälle molekyyligenetiikassakin

Olli Pietiläinen[1]

Olen opiskellut yliopistossa pääaineena orgaanista kemiaa ja sivuaineena fysiikkaa ja biokemiaa, mutta en koskaan juuri tuntenut vetoa matematiikkaa kohtaan. Ajauduin jatko-opinnoissani kemiasta perinnöllisyystieteen ja molekyyligenetiikan pariin, jossa matematiikan ja tietojenkäsittelytieteen merkitys tuli minulle yllätyksenä. Jälkeenpäin ajateltuna sehän oli itsestään selvää.

DNA:n neljän emäksen vaihtelusta rakentuva noin kuuden miljardin yksikön mittainen geneettinen koodi muodostaa yhdessä rekombinaation, uusien mutaatioiden ilmaantumisen ja ympäristön vuorovaikutuksen kanssa synteesin, joka määrittää suurta osaa ominaisuuksistamme. Synteesin tutkiminen on monelta osin jo käytännön tasolla matemaattinen ja tietojenkäsittelytieteellinen haaste.

Seuraavaksi esittelen esimerkkejä geneetikoiden lähes päivittäin kohtaamista matematiikan ja tietojenkäsittelytieteen sovelluksista. Ne osoittavat, miten tärkeää matematiikka alallani nykyään on. Niistä voi myös päätellä, miten monenlaiseen matematiikkaan olen työssäni joutunut tutustumaan.

Nykyisillä genotyypitys- ja sekvensointiteknologioilla voidaan määrittää miljoonia DNA-polymorfioita aina koko perimän emäsjärjestykseen asti tuhansista henkilöistä. Näin tuotettu genotyyppidata muodostaa valtavia giga- ja teratavujen kokoisia tiedostoja. Niiden käsittely ja yksinkertainenkin tilastollinen analyysi vaatii tekijältään jotain enemmän kuin perinteisten taulukkolaskentaohjelmien hallintaa.

Kromosomit jakaantuvat sattumanvaraisesti sukusoluihin. Kuitenkin samassa kromosomissa kaksi lähekkäin olevaa polymorfiaa kulkeutuu useammin yhdessä kuin olisi odotettavissa sattumalta. Tästä muodostuu niin sanottu kytkeytymisepätasapaino, jossa lähekkäin olevat geenimerkit korreloivat keskenään. Tätä sisäistä korrelaatiorakennetta voidaan käyttää laskennallisesti esimerkiksi ennustamaan kokeellisesti määriteltyjen varianttien genotyyppien avulla sellaisia variantteja, joita ei olla suoraan tutkimushenkilöissä mitattu.

Joukko yksittäisiä geenivariantteja, jotka assosioituvat tiettyyn ominaisuuteen, kuten alttiuteen sairastua sydän- ja verisuonitauteihin tai syöpään, muodostavat yhdessä monimutkaisen verkoston. Yksittäiset variantit vaikuttavat eri biologisten prosessien kautta ja toisinaan useisiin prosesseihin yhtä aikaa, jotka lopulta näyttäytyvät tilastollisesti sairastumisriskin nousuna tai laskuna. Todellinen kausaalisuus selvitetään matemaattisesti empiirisestä datasta. Ongelmat ovat vaikeita. Esimerkiksi vaikuttaako keuhkosyöpään assosioituva geenivariantti tupakoimiskäyttäytymiseen vai biologiseen prosessiin, joka tekee terveestä solusta hallitsemattomasti jakautuvan syöpäsolun? Teknologioiden kehityksen myötä voidaan myös ensimmäistä kertaa systemaattisesti tutkia harvinaisten varianttien ja tutkimushenkilöissä syntyneiden uusien mutaatioiden osuutta eri ominaisuuksiin. Yksittäisten varianttien harvinaisuuden vuoksi tilastollisen analyysin voima pienenee. Toisaalta uusia mutaatioita syntyy 10^{-8} suuruusluokan taajuudella per emäs per sukupolvi, ja suuri osa kahden henkilön välisistä geneettisistä eroista aiheutuu juuri harvinaisista varianteista. Siksi näiden mutaatioiden ja jonkin ominaisuuden välisen suhteen tutkimiseen tarvitaan yhä hienostuneempia tilastollisia malleja, jotka huomioivat esimerkiksi mutaatioille ennustetun haitallisuuden asteen, mutaatiofrekvens-

[1] Kirjoittaja on väitöskirjatutkija Suomen molekyylilääketieteen instituutissa ja tutkii vakaville mielenterveyden häiriöille ja hermoston kehityksen sairauksille altistavia geenimuotoja.

sin ja kuinka usein verrokkiaineistossa nähdään vastaavia muutoksia kyseisellä perimän alueella. Toisaalta varianttien toiminnallisuuden tai haitallisuuden ennustaminen perustuu sekin osaltaan matemaattiseen ennustamiseen. Esimerkiksi, kokeellisesti tuotettua tietoa DNA:n toiminnallisuuteen liittyvistä sekvenssin motiiveista voidaan laskennallisesti käyttää ennustamaan eri perimän alueiden toiminnallisuutta pelkästään niiden DNA-sekvenssin perusteella.