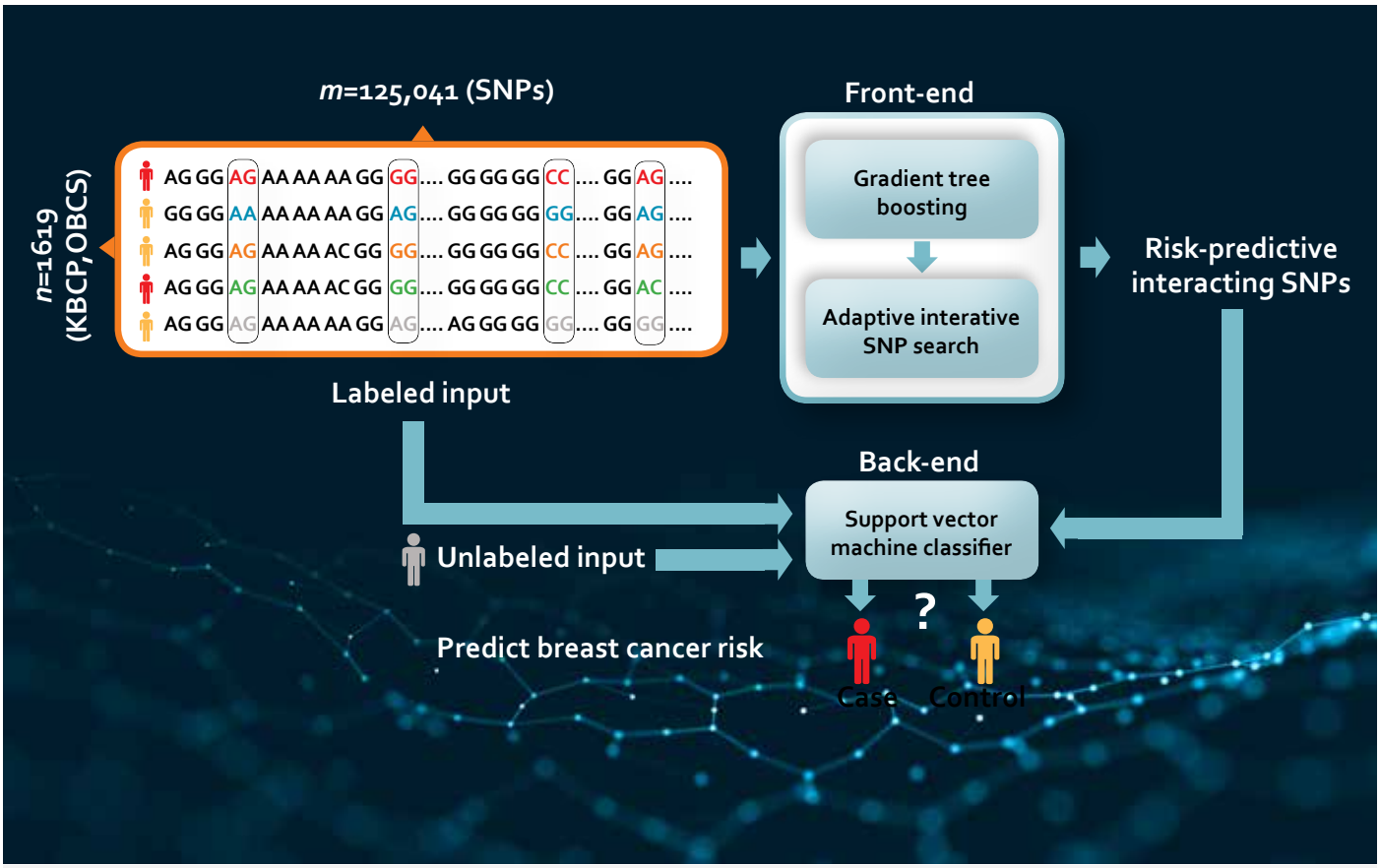
Koneoppimisella etsitään merkkejä rintasyövästä

Geenivarianttien lisäksi on genomisia variantteja yksittäisissä DNA:n emäsparijaksoissa. Nämä variaatiot aiheuttavat yksilöiden väliset erot mutta ne voivat myös auttaa paikallistamaan tautia aiheuttavia geenejä. Nämä yhden emäsparin vaihtelut eli snipit (single nucleotide popymorphism, SNP) voivat toimia markkereina, jotka viittaavat sairauteen. Itä-Suomen yliopistossa kehitetty tekoälymalli etsii rintasyöpään viittaavia snippejä.

Genomidatan valtava määrä on mahdollistanut sen, että tutkijat voivat laskea, mitä geenimuunnoksia on niissä ryhmissä, jotka ovat sairastuneet syöpään. Yhteen tautiin voi vaikuttaa satoja tai tuhansia geenimuunnoksia.

Tilastollisten menetelmien ansiosta tutkijat voivat arvioida, miten yhden ihmisen geenimuunnokset lisäävät riskiä sairastua tautiin eli näin saadaan monitekijäisten geenien riskiarvo. Mutta variaatioita on myös DNA:n emäspareissa eli nukleotideissä. Ne tunnetaan genomisina variantteina eli snippeinä. DNA:n sekvenssivariaatiot tapahtuvat kun yhdessä emäsparissa genomisekvenssi (adeniini-tymiini, sytosiini-guaniini) muuttuu. Jokainen SNP edustaa muutosta yhdessä emäsparissa. Esimerkiksi yksi SNP voi vaihtaa jossakin DNA-ketjun emäsparissa sytosiinin tymiiniksi. Se tarkoittaa, että sytosiini-guaniini -emäspari voi muuttua DNA-ketjussa esimerkiksi tymiini-adeniini -pariksi. Toisin kuin geenimuunnokset, snipit eivät välttämättä sijaitse geeneissä. Snippejä sijaitsee myös ei-koodaavissa geeneissä tai geenien välissä. Ihmisen genomissa on paljon snippejä. Niitä on keskimäärin melkein joka tuhannen emäsparin jälkeen,

125 000 snipin genotyypitys tehtiin iCOGS-sirulla yteistyössä BCAC:n (Breast Cancer Association Consortium) kanssa. iCOGS on genotyypittävä siru, joka on suunniteltu testaamaan kolmea hormoniperäistä syöpää: rinta,-munasarja,- ja eturauhassyöpiä. Sirulla on genotyypitetty yli 250 000 yksilöä ja snippiä yli 50 eri alueelta, joissa tiedetään lymyilevän joidenkin näiden tautien epäilyttäviä variantteja.



Tutkijat ovat löytäneet yli miljoona snippiä (single nucleotide polymorphisms) populaatioissa kaikkialla maailmassa. Kaikkein yleisimmin nämä variaatiot löytyvät DNA:sta geenien välistä. Nämä variaatiot voivat olla ainutlaatuisia tai esiintyä monella yksilöllä. https://upload.wikimedia.org/wikipedia/commons/2/2e/Dna-SNP.svg

mikä tarkoittaa, että ihmisen genomissa on arviolta 4–5 miljoonaa snippiä.

Snipit voivat olla hyödyllisiä, kun etsitään syövän geneettisiä riskitekijöitä. Biolääketieteellisessä tutkimuksessa snippejä käytetään tutkimusaineistossa vertailemalla genomialueita sairastuneiden ja terveiden välillä.

"Kun snipit ilmaantuvat geenissä tai regulatiivisella alueella lähellä geeniä, nillä voi olla suora rooli taudin syntymiseen, koska ne vaikuttavat geenin toimintaan. Meillä on uudenlainen koneoppimisen lähestymistapa, jolla voidaan tunnistaa joukko vuorovaikuttavia snippejä, jotka ovat eniten osallisina rintasyövän riskitekijöissä", sanoo tutkija **Hamid Behravan** Itä-Suomen yliopistosta. Hän työskentelee Kuopiossa Kliinisen lääketieteen yksikössä.

"Olemme julkaisseet useita tuloksia siitä, miten geneettinen osatekijä rintasyövän riskissä tunnistetaan, jolloin erotettaisiin luotettavasti sairastapaukset terveiden vertailuryhmästä. Rintasyöpään liittyvien snippien tunnistaminen on erityisen hyödyllistä, koska rintasyövän ennustettavuutta voidaan parantaa ja kehittää yksilöllisiä hoitosuunnitelmia, sanoo Behravan.

Standardeilla hypoteesien testausmenetelmillä on mitattu ainoastaan yhden snipin yhteyttä tautiin. Kuitenkin Itä-Suomen yliopiston tutkimukset ovat osoittaneet, että rintasyövän riskitekijät voidaan ennustaa paremmin kun snippejä tarkastellaan ryhminä, jotka itse asiassa vuorovaikuttavat toistensa kanssa.

Genominlaajuisten assosiaatiotutkimusten (GWAS) idea on tunnistaa snipit DNA:ssa. Se auttaa selvittämään geneettiset osatekijät tutkittavassa fenotyypissä joukossa genotyypitettyjä ihmisiä. Genotyypityksessä luetaan vain ne tiedossa olevat kohdat kromosomeissa, joissa esiintyy tutkittavaan tautiin liittyviä geenivariantteja.

"Genominlaajuiset assosiaatiotutkimukset mittaavat yksittäisen snipin yhteyttä sairauteen, mutta jättävät huomioimatta mahdollisen korrelaation snippien välillä", sanoo Behravan.

"Tähän päivään asti koko populaation kattavat GWAS-tutkimukset ovat usein käyttäneet ns. PRS- pisteytystä (polygenic risk scoring, PRS), joka kerää yhteen riskialleelien (geenien vaihtoehtoiset muodot) vaikutukset tautiin. Kuitenkin PRS olettaa, että tauteihin liittyvät snipit ovat riippumattomia toisistaan ja että riskivaikutukset ovat lineaarisia ja yhteenlaskettavissa. Olemme osoittaneet, että sen sijaan, että arvioisimme yksittäisiä osatekijöitä (snipit) yksi kerrallaan, olisi erityisen hyödyllistä parantaa rintasyöpäriskin ennustettavuutta tutkimalla vuorovaikuttavien snippien ryhmää käyttäen koneoppimista."

Snipit, joilla on todellista biologista merkitystä, löydettiin koneoppimisen avulla

Itä-Suomen yliopistossa kehitetty koneoppimisen menetelmä on osoittautunut tehokkaaksi.

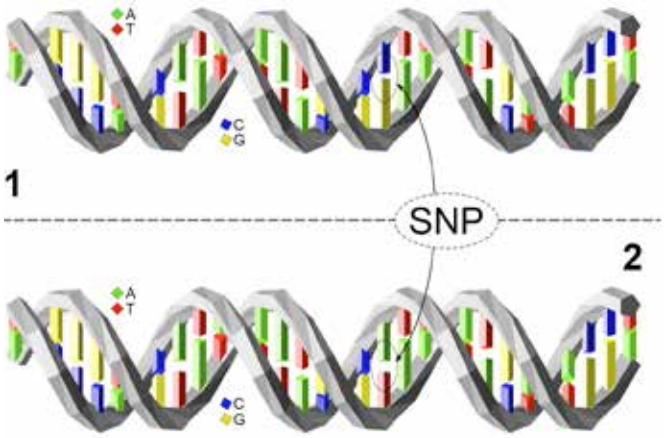
“Löysimme ryhmän vuorovaikuttavia snippejä, joilla on todellista biologista merkitystä. Tunnistettujen snippien biologinen analyysi paljasti geenejä, jotka liittyivät tärkeisiin rintasyöpään viittaaviin mekanismeihin, kuten estrogeeniaineenvaihduntaan ja ohjelmoituun solukuolemaan, apoptosikseen.

Kohonneet estrogeenitasot liittyvät vaihdevuosien jälkeeen kasvaneeseen rintasyövän riskiin. On myös vahva näyttö, että kasvaimen kasvu ei johdu pelkästään rajoittamattomasta leviämisestä vaan myös pienentyneestä solukuolemasta.

”Löysimme siis menetelmämme avulla geenit noiden tunnistettujen snippien taustalta. Laadimme näistä geeneistä interaktiivisia karttoja. Sitten tarkkailimme useita erilaisia rintasyöpään liittyviä geenien vuorovaikutusverkostoja, kuten estrogeeniaineenvaihduntaa ja ohjelmoidun solukuoleman verkostoja. Meidän systeemimme ei ainoastaan löytänyt mahdollisimman hyvin vuorovaikuttavia rintasyövän riskejä ennustavia snippejä, vaan se myös tunnisti ne snipit, jotka muodostivat merkittävän määrän tärkeitä biologisia rintasyövän osa-alueita. Näin ollen, vuorovaikuttavat snipit ilmaisevat myös ne snipit, jotka ovat mukana syöpään liittyvissä biologisissa verkostoissa.”

Kone oppii etsimään geneettisiä variaatioita

Kuopiossa kehitetty koneoppimisen lähestymistapa perustuu gradienttipuun tehostamismenetelmälle, jossa on iteratiivinen hakualgoritmi. Tehostaminen on ensimmäinen moduuli ja haku toinen.



Tutkijat ovat löytäneet yli miljoona snippiä (single nucleotide polymorphisms) populaatioissa kaikkialla maailmassa. Kaikkein yleisimmin nämä variaatiot löytyvät DNA:sta geenien välistä. Nämä variaatiot voivat olla ainutlaatuisia tai esiintyä monella yksilöllä. (SNP model by David Eccles).

Tehostaminen (boosting) on algoritmi ja metodi, jolla heikot oppijat muutetaan vahvoiksi. Heikolla luokittelijalla tarkoitetaan sellaista luokittelijaa, joka on vähintään puolessa tapauksista oikeassa. Algoritmi käynnistyy opettamalla päätöspuuta. Heikot luokittelijat lisätään peräkkäisesti korjaamaan olemassaolevien luokittelijoiden virheet, jotta rakennetaan vahvaa luokittelija.

”Ensimmäinen moduuli arvioi tunnusmerkkien tarkkuutta, tässä tapauksessa snippejä, rintasyövän ennustettavuudessa. Ensimmäinen moduuli antaa alustavan kandidaattilistan snipeistä, jotka voivat ennustaa rintasyöpäriskistä.”

Toinen moduuli sitten käyttää kandidaattisnippejä adaptiivisessä ja iteratiivisesä haussa, jotta se voisi kaapata nuo vuorovaikuttavat piirteet. Parhaimmat tunnistetut vuorovaikuttavat snipit käytetään ennustamaan tuntemattoman yksilön rintasyövän riskiä testivaiheessa käyttäen koneluokittelijaa. Luokittelija opetettiin erottamaan rintasyöpätapaukset (positiiviset näytteet) terveistä kontrolleista (negatiiviset näytteet).

Koska syöpä on monitekijäinen tauti, jonka aiheuttavat elintavat sekä geneettiset ja ympäristötekijät, geneettisiin variantteihin perustuva yksilöllinen analyysi ei ehkä ole riittävä, jotta saataisiin kokonaisvaltainen kuva tautiriskistä. Behravanin mukaan myös muita datalähteitä tarvitaan.

“Kehitämme integroivia koneoppimisen lähestymistapoja, jossa yhdistetään eri datalähteitä, kuten väestötieteellistä dataa.”

Ari Turunen

LISÄTIETOJA:

Lääketieteen laitos, Itä-Suomen yliopisto
https://www.uef.fi/fi/web/laake

CSC – Tieteen tietotekniikan keskus Oy
on valtion omistama, opetus- ja kulttuuriministeriön hallinnoima, voittoa tavoittelematon osakeyhtiö. CSC ylläpitää ja kehittää valtion omistamaa keskitettyä tietotekniikkainfrastruktuuria.
http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR
rakentaa infrastruktuurin bioalan tutkimuksen tueksi. Se yhdistää 21 Euroopan maan ja Euroopan molekyylibiologian laboratorion EMBL:n johtavat organisaatiot yhteiseksi biologisen informaation infrastruktuuriksi. Sen Suomen keskus on CSC – Tieteen tietotekniikan keskus Oy.
http://www.elixir-finland.org
http://www.elixir-europe.org