Algoritmi opetetaan tunnistamaan syöpä sekvenssidatasta

Syväoppiminen on mullistanut syöpäsairauksien tutkimisen. Syvillä neuroverkoilla voidaan automaattisesti löytää potilaan näytedatasta piirteitä, joiden perusteella voidaan tunnistaa syöpiä. Oppivat algoritmit voivat tunnistaa jatkossa verinäytteestä mahdollisia syövän esiasteita. Esa Pitkänen ja hänen tutkimusryhmänsä Suomen molekyylilääketieteen instituutista kehittävät uuden sukupolven syväoppimisen algoritmeja.

Algoritmeja on hyödynnetty kudosnäytteiden leikekuvien solujen tunnistamisessa. Esimerkiksi jos kudoksen solut näyttävät epätyypillisiltä, algoritmi tunnistaa sen ja päättelee onko kyseessä syöpä. Nyt kuvantamisdatan rinnalla käytetään syöpien tunnistamisessa kasvaimista saatua DNA-sekvenssidataa.

”Aikaisemmin on ollut vaikea sanoa DNA-sekvenssin perusteella, minkälaisesta kasvaimesta sekvenssi on tullut. Nyt on luotu uusia tekniikoita ja syväoppimisen algoritmeja”, sanoo tutkija Esa Pitkänen.

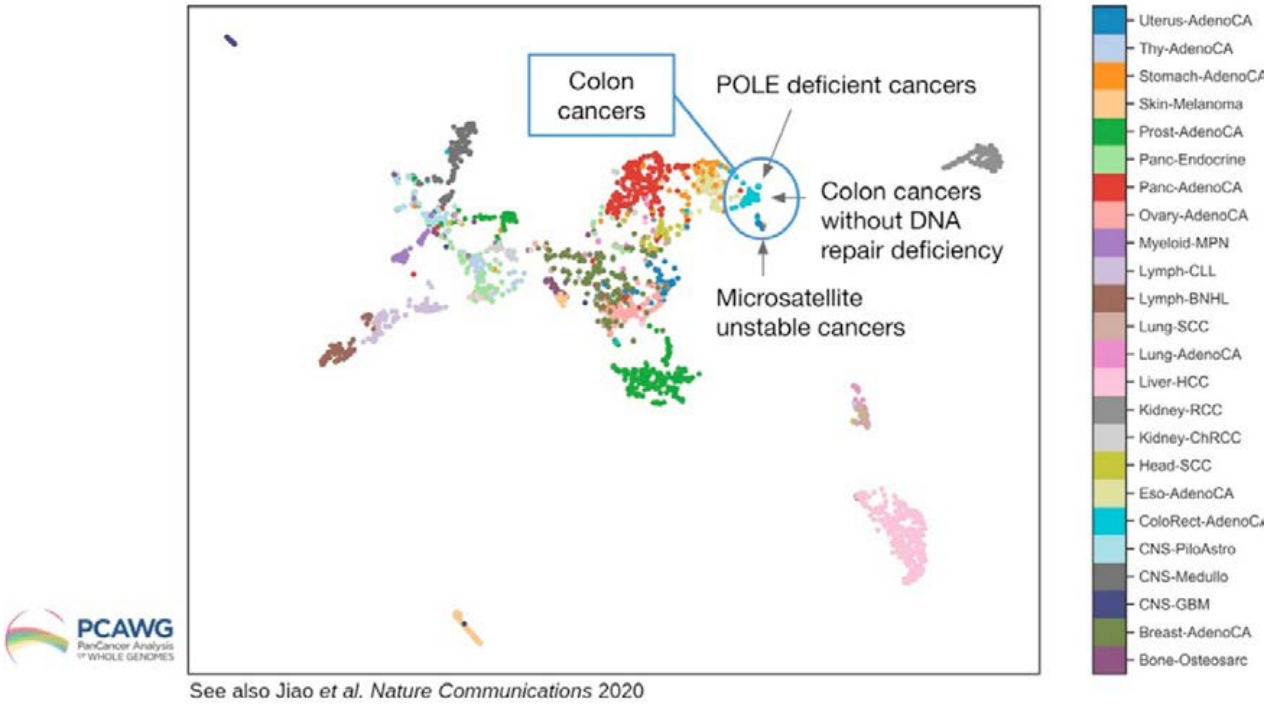
Pitkänen ryhmineen kehittää algoritmeja, jotka tunnistavat DNA-sekvensseistä lyhyitä, toisteisia pätkiä. Algoritmien avulla voidaan löytää pätkiä, jotka mutatoituvat tietyssä syöpätyypissä usein tai joihin tietyt geenien säätelyyn osallistuvat proteiinit sitoutuvat. Näitä pätkiä analysoimalla voidaan saada tietoa esimerkiksi syöpäsairauksien syiden kartoittamiseen ja lääkkeiden kehittämiseen.

”DNA:n kopioituminen solun jakautumisen yhteydessä ei ole täydellistä. Kun solu jakautuu niin on mahdollista, että mutaatioita syntyy. Kun solu jakautuu, kopioitavaa DNA:ta on kuuden miljardin merkin verran eli virheitä tapahtuu. Pienikin todennäköisyys riittää että mutaatioita tulee”, sanoo Pitkänen.

”Jos riittävästi mutaatioita tapahtuu esimerkiksi kasvaimen syntyä ehkäisevissä geeneissä, syöpä voi alkaa kehittyä.”

Esimerkiksi pistemutaatiossa yksi emäs vaihtuu toiseksi DNA-ketjussa. Virhe voi syntyä, kun solun jakautuessa DNA kopioidaan ja kopioinnista vastaavat entsyymit korjaavat esimerkiksi auringonvalon ultraviolettisäteilystä vaurioituneen kohdan väärin. Ihosyöpää aiheuttavan ultraviolettisäteilyn aikaansaama tyypillinen mutaatio on se, että ihmisen DNA:n emäspareissa kaksi peräkkäistä sytosiinia (C) muuttuvat kahdeksi tymiiniksi (T). Kun tällaisia, ihosyövälle tyypillisiä mutaatioita havaitaan riittävästi, oppivat algoritmit yhdistämään mutaatiot tiettyyn syöpätyyppiin.

”Yritämme ennustaa mutaatioiden perusteella mikä syöpätyyppi ja kasvain on kyseessä. Samalla saadaan tietoa, joka voi vaikuttaa vaikuttaa hoitoon.”



Esa Pitkänen ja hänen tutkimusryhmänsä hyödynsivät yhtä suurimmista syöpänäytteiden (PCAWG) datakokoelmaa, joka koostuu 47 miljoonasta mutaatiosta. Data on peräisen 2600 potilaan kasvainnäytteistä, jotka on sekvensoitu. Kokoelmassa oli 37 eri kasvaintyyppiä eri syövistä, kuten paksusuolensyövästä, keuhkosyövästä ja melanoomista. Prima Sanjaya teki koneoppimismallin syvillä neuroverkoilla, joka ottaa huomioon kunkin potilaan sekvenssidatan ja ikään kuin heijastaa tämän datan kaksiulotteiseen karttamuotoon. Tässä kuvassa jokainen piste on yksi erillinen potilaalta saatu kasvain . Värit ovat eri kasvaintyyppejä. Mielenkiintoisesti malli ryhmittelee paksusuolensyövät yhteen mutta myös näkee eron kolmen alatyypin välillä (merkitty kuvaan nuolilla).

Algoritmi tunnistaa verinäytteestä saadusta DNA:sta syövän

Pitkänen ryhmineen analysoi sekvenssijaksoja ja algoritmeja opetetaan tunnistamaan sekvenssijaksojen poikkeavuuksia. Näistä poikkeavuuksista algoritmi pystyy tunnistamaan, että kyseessä on kasvain ja luokittelemaan kasvaimet eri syöpätyyppeihin.

“Ennen siirtymistäni Suomen molekyylilääketieteen instituuttiin olin Euroopan molekyylibiologian laboratoriossa EMBL Heidelbergissä, jossa osallistuin PCAWG-syöpägenomiprojektiin. Projektissa analysoitiin yli 2600 syövän kokogenomia. PCAWG-data toimii aineistona useassa ryhmäni syöpägenomiikkaa käsittelevissä projekteissa.”

Esa Pitkäsen ryhmän kehittämälle algoritmille on opetettu näiden 2600 syöpäpotilaan kasvainnäytteistä löydetyt löytyneet mutaatiot, joita on yhteensä 47 miljoonaa.

“Algoritmi on koulutettu siten, että se yrittää näistä sekvenssien muutoksista päätellä syöpätyypin. Kun algoritmille on annettu eri kasvainten kaikki mutaatiot sekvensseineen, se pystyy jatkossa päättelemään minkälainen kasvain on kyseessä. Päättely perustuu siihen, että algoritmi oppii nämä yhteydet.”

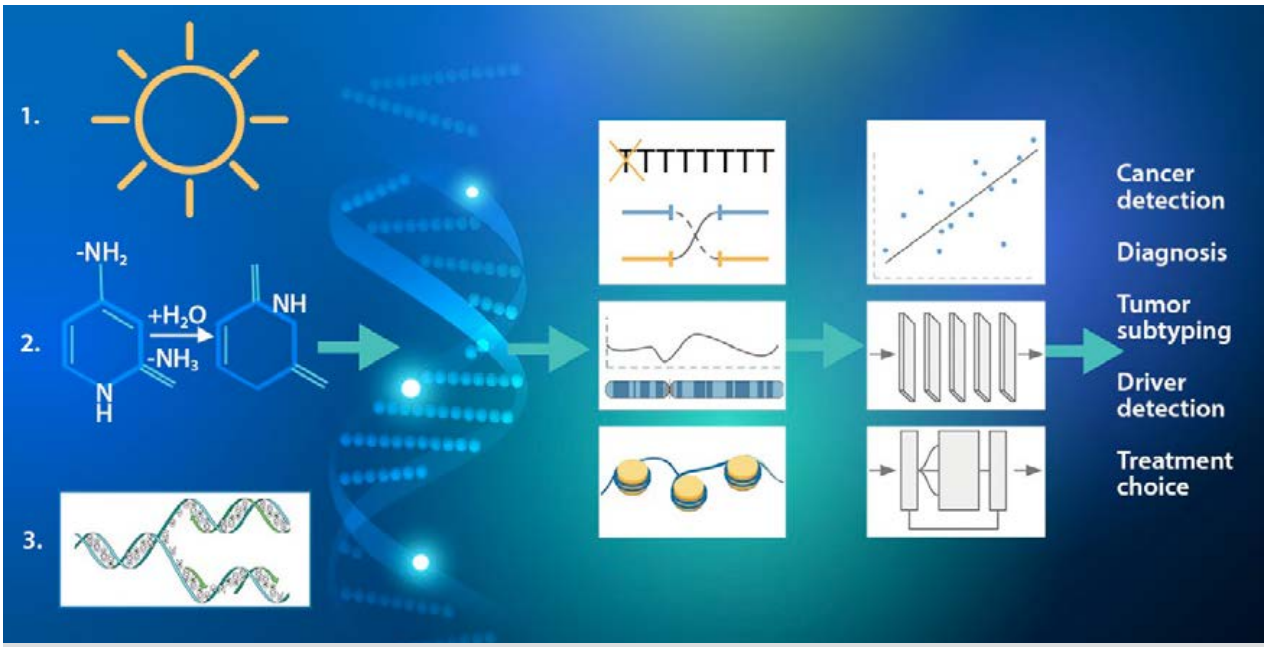
Algoritmi oppii kasvaimissa olevan sekvenssidatan poikkeamien kautta tunnistamaan, että kyseessä on tietylle syövälle olennainen mutaatio. Algoritmi pystyy ryhmittelemään kasvaimet pelkän sekvenssidatan perusteella.

”Ryhmässäni tutkija Prima Sanjaya on kehittänyt neuroverkkomalleja sekvenssidatan analysoimiseen. Silloin tällöin törmätään metastaattisiin eli levinneisiin syöpiin, josta ei tiedetä mistä se on levinnyt. Tulevaisuudessa voidaan hyödyntää myös ns. nestebiopsiaa. Tällöin pystytään toivottavasti verinäytteestä sanomaan, onko potilaalla syöpä ja jos on niin minkälainen.”

Nestebiopsia perustuu siihen, että elimistön solut vapauttavat verenkiertoon ja ruumiinnesteisiin DNA:ta, jota kutsutaan solunulkoiseksi tai soluvapaaksi DNA:ksi (cell free DNA, cfDNA). Myös syöpäsoluista vapautuu DNA:ta, joka mahdollistaa syöpämutaatioiden etsimisen veren plasmasta.

“Jos nestebiopsiassa näkyy jälkiä syövästä, emme tiedä suoraan mikä syöpä on kyseessä, koska se voi tulla verenkiertoon mistä vain kehosta. Jos meillä on keinoja katsoa tarkemmin, kuten syväoppimisen algoritmit, saamme arvokasta tietoa, mihin kohtaan potilaan kehossa tutkimus pitää suunnata. Algoritmi voi kehottaa katsomaan esimerkiksi paksusuoleen. Uskon, että tulevaisuudessa tällaisilla algoritmeilla on suuri merkitys. Nestebiopsian ja algoritmien ansiosta voidaan tehdä tutkimusta ilman potilasleikkauksia”

Mutaatioiden lähteet



Mutaatioiden lähteinä ovat **1.ulkoiset tekijät:** esimerkiksi auringon UV-säteily. **2.sisäiset tekijät:** spontaani deaminaatioreaktio eli emäksen amiiniryhmän muutos, jolloin alkuperäinen emäs muuttuu joksikin toiseksi, esimerkiksi adeniini urasiiliksi **3. DNA:n kopioinnissa aiheutuneet virheet.**

Mutaatio tarkoittaa muutosta DNA:n tai RNA:n nukleotidijärjestyksessä. Nukleotidiin kuuluu emäs, sokeri ja fosfaatti. DNA:n sokeri on D-deoksiriboosi ja RNA:n D-riboosi. DNA:n emäksiä ovat guaniini (G), adeniini (A), sytosiini (C) ja tymiini (T). RNA:n emäsosassa tymiinin tilalla on urasiili (U). Mutaatio voi olla vain yhden nukleotidin muutos eli pistemutaatio, tai se voi käsittää useita nukleotideja. Pistemutaatiossa yksi emäs vaihtuu toiseksi RNA- tai DNA-ketjussa. Iso mutaatioita, jotka voivat käsittää tuhansia nukleotideja, kutsutaan rakennemuutoksiksi.

Rakennemuutos voi vaikuttaa yhtä aikaa useaan geeniin. Syövät ovat yleensä useiden somaattisten mutaatioiden aiheuttamia; somaattiset mutaatiot eivät periydy, ja niitä voi syntyä milloin tahansa alkionkehityksen aikana ja sen jälkeen. Mutaatioiden seurauksena normaalin solun toiminta voi muuttua siten, että solu alkaa jakautua hallitsemattomasti. rilaisia mutaatiotyyppejä mutaatioiden jakautuminen kromosomeihin epigeneettinen tieto. Epigeneettiseen periytymiseen vaikuttavat monet ulkoiset tekijät, kuten esimerkiksi ravinto. Esimerkiksi identtiset kaksoset, voivat kehittyä ulkoisilta olemuksiltaan erilaisiksi. Mutaatioiden mallintaminen lineaariset mallit syvät neuroverkot transformer-mallit. Transformerit ovat syväoppimismalliperhe, jotka toimivat erityisen hyvin esim. tekstimuotoiseen dataan, sovelluksena vaikkapa konekääntäminen. Syöpätutkimuksessa transformer-mallit voivat kiinnittää huomiota mutaatiotyyppeihin, jotka ovat tärkeitä tietyn syöpätyypin tunnistamiseksi. Esimerkiksi ihosyövissä, joissa on paljon auringonvalon aiheuttamia mutaatioita (C>T, CC>TT), huomio kohdistuu juuri näihin mutaatioihin.

Kuvassa keskellä erilaisia mutaatiotyyppejä ja miten mutaatiot jakautuvat kromosomeihin. Mutaatioihin liittyy epigeneettinen tieto. Epigeneettiseen periytymiseen vaikuttavat monet ulkoiset tekijät, kuten esimerkiksi ravinto. Esimerkiksi identtiset kaksoset, voivat kehittyä ulkoisilta olemuksiltaan erilaisiksi.

Mutaatioiden mallintaminen:

lineaariset mallit

syvät neuroverkot

transformer-mallit. Transformerit ovat syväoppimismalliperhe, jotka toimivat erityisen hyvin esim. tekstimuotoiseen dataan, sovelluksena vaikkapa konekääntäminen. Syöpätutkimuksessa transformer-mallit voivat kiinnittää huomiota mutaatiotyyppeihin, jotka ovat tärkeitä tietyn syöpätyypin tunnistamiseksi. Esimerkiksi ihosyövissä, joissa on paljon auringonvalon aiheuttamia mutaatioita (C>T, CC>TT), huomio kohdistuu juuri näihin mutaatioihin.

Algoritmi suolistosyövien tunnistamisessa

Syövän syntyyn vaikuttavat perintötekijöiden lisäksi elintavat. Helsingin yliopistossa on tutkittu paljon esimerkiksi suolistosyöpiä.

”Se tiedetään, että punaisen lihan syömisellä on yhteys paksunsuolen syövän syntyyn. Syntymekanismit vaativat vielä lisätutkimuksia mutta esimerkiksi punaisen lihan aiheuttamien DNA:n alkylaatioreaktioiden merkitystä on selvitetty viime vuosina paljon.”

Paksunsuolen syöpä (CRC) on yksi vaarallisimpia syöpiä länsimaissa ja johtaa 30% tapauksissa esimerkiksi Suomessa kuolemaan. Noin 15% paksunsuolen syövistä kuuluvat joukkoon, jossa esintyy ns. mikrosatelliiti-instabiliteettia (MSI). Mikrosatelliitit ovat DNA:n toistojaksoja, joiden pituus vaihtelee yksilöstä toiseen ja ovat siten yksilöllisiä “sormenjälkiä”. Mikrosatelliiti-instabiliteetissa solun DNA:n replikaation jälkeinen korjausmekanismi ei toimi, jolloin mutaatioita alkaa kertyä erityisesti mikrosatelliitteihin.

”MSI-kasvaimessa mikrosatelliitteihin tulee helposti yhden emäksen lisäyksiä tai poistoja. Esimerkiksi kahdeksan peräkkäisen adeniinin mikrosatelliitista häviää yksi adeniini. Osuessaan geeniin tällainen muutos aiheuttaa geenin koodaaman proteiinin aminohappoketjun sisällön muuttumisen täysin. Jos riittävästi muutoksia tapahtuu hallitsematonta solujakautumista estävissä geeneissä, saattaa syövän kehittyminen alkaa.”

MSI liittyy usein paksunsuolensyövän lisäksi muihin syöpiin, kuten vatsasyöpiin, kohdunrungon ja munasarjan syöpään tai aivosyöpään. Syövän ennusteen arvioinnissa voidaan käyttää apuna MSI-analyysiä. Analyysin perusteella on joskus mahdollista määrittää sopiva hoito.

”Mielenkiintoista on, että syvä neuroverkko oppii myös luokittelemaan eri syöpien alalajeja. Se tunnisti esimerkiksi suolisyöpien MSI-alatyypin”, Pitkänen sanoo.

Suomen ELIXIR-keskus CSC on yksi päätapartnereita PerMedCoE-hankkeessa. Kolmevuotisen HPC/Exascale Centre of Excellence in Personalised Medicine -hankeen (PerMedCoE) avulla esimerkiksi syöpään liittyvä data saadaan tehokkaasti terveydenhoidon käyttöön ja diagnoosit nopeutuvat.

”Tulevaisuuden yksilöidyt hoidot kuten syöpähoidot rakentuvat täsmälliseen käsitykseen potilaasta ja hänen sairaudestaan. Tämä käsitys muodostetaan keräämällä suuri määrä erilaista tietoa, kuten syöpää hoidettaessa kasvaimen genomi- ja kuvantamistietoa. Monet tiedonkeruumenetelmät tuottavat valtavan määrän tietoa, joiden analysoimiseksi kehitetyt uudet laskennalliset menetelmät puolestaan vaativat suuria laskentaresursseja”, Pitkänen toteaa.

”Uuden laskennallisen menetelmän kehittäminen ideasta toimivaksi, terveydenhoidossa käytettäväksi työkaluksi on tällaisessa toimintaympäristössä valtava haaste. Erityisesti syöpähoidoissa on tärkeää, että potilaan hoitoon vaikuttava tieto saadaan lääkärin käyttöön mahdollisimman nopeasti. Uskon, että PerMedCoE:n tuloksilla luodaan pohjaa sille, että valtavasta terveystietomäärästä voidaan lääkärin avuksi jalostaa merkityksellistä tietoa ja näin parantaa hoitotulosta merkittävästi.”

Ari Turunen

LISÄTIETOJA:

HPC/Exascale Centre of Excellence in Personalised Medicine (PerMedCoE)
https://permedcoe.eu

Suomen molekyylilääketieteen instituutti FIMM
https://www.fimm.fi

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