Avoimesti tieteestä!: Jarkko Koivunen – Avoin data tutkimusta helpottamassa
Äänitteen kesto: 16 min

Litterointimerkinnät

sa-	sana jää kesken
(sana)	epävarmasti kuultu jakso puheessa tai epävarmasti tunnistettu puhuja
(-)	sana, josta ei ole saatu selvää
(--)	useampia sanoja, joista ei ole saatu selvää
, . ? :	kieliopin mukainen välimerkki tai alle 10 sekunnin tauko puheessa

---

[musiikkia]

Avoimesti tieteestä, keskustelua tieteen, tutkimuksen ja oppimisen avoimuudesta ja vastuullisuudesta.

Henriikka Mustajoki: Tervetuloa Avoimesti tieteestä -podcastin pariin. Minä olen Henriikka Mustajoki, ja tänään keskustelen sidekudosbiologian dosentin Jarkko Koivusen kanssa. Tervetuloa Jarkko.

Jarkko Koivunen: Kiitos.

Henriikka Mustajoki: Kerropa Jarkko ensin, että minkälaista dataa tuottaa sidekudosbiologi?

Jarkko Koivunen: Hyvin vaihteevaa, kaikkea mahollista mitä biolääketieteen tai solubiologian alalla voi tutkia, niin käytännössä kaikennäköstä dataa sieltä tulee. Solukokeita tehään paljon, proteiini-molekyylitasolla tehdään kokeita, ja tietysti hyvin paljon noita esimerkiks hiirikokeita. Niistä kokeista tulee hyvin paljon kuvia, mut kyllä sielt tulee sitte iha numeroita. Aika paljon se data, mitä iteki tekee, niin se on nimenomaan sitä kuvantamista. Eli se on aika pitkälle siinä kuvien ja numeroiden välissä hyppimistä, mimmosta se data on.

Henriikka Mustajoki: Mimmonen on sun työssä sen datan elinkaari? Sä oot katsonu niitä kuvia ja lukenut niitä numeroita. Niin mitä sille datalle sen jälkeen tapahtuu?

Jarkko Koivunen: Analysoidaan, kirjoitetaan tulokset ylös. Käytännössä uudet kokeet, uusi analyysi ja uudet kokeet. Sitä aikansa pyöritellään näitä asioita ja sit kirjoitetaan artikkeli, jossa sit julkaistaan ne numerot ja tulokset. Ja se, että missä muodossa ne julkaistaan, niin se vähä vaihtelee tietysti julkaisufoorumista, että mitä lehti vaatii tai mitä se sallii. Nykyisin on se suplementtiavaruus, johon työnnetään hyvin paljon sitä dataa, tehään enemmänki niitä numeroita, ja kaikkia ei laitetaan aina graafeissa esimerkiks niitä mitä saadaan. Voiaan tehdä graafi ja sit laittaa ne numerot sen lisäks sinne supplementtiavaruuteen, mut kaikki foorumit ei

tätä esimerkiks hyväksy. Että pitää tehdä sillä tavalla ku se lehti, missä pyritään julkaisemaan, sen tekee. Se on tavallaan yhen projektin päätös aina kun saahaan se julkastua, mutta sit harvoin se koko tutkimus siihe loppuu. Et sit otetaan uus pomppu ja jatketaan eteenpäin heti siihen perään. Tutkimushan luo aina uusia avauksia, et ei se siihen pääty ku yhen kerran julkaistaan joku. Toivottavasti mikään ei pääty siihen, vaan siitä lähdetään sit vähä eri suuntaan tai jatketaan sitä samaa suuntaa.

Henriikka Mustajoki: Jos ajattelet, että sen jutun lukee joku toinen joka solukudosbiologiaa ymmärtää, niin lehtien kautta hän pääsee kuitenki näkemään sen datan, joka on ollu teidän tutkimusartikkelin pohjana. Toimiiko teidän alalla lehdet niin?

Jarkko Koivunen: Kuvat siellä aina lukee lehessä. Se vähä vaihtelee sitte ne numeeriset arvot, että halutaanko se graafeina vai miten. Esimerkiks tilastotiede on semmonen, josta väännetään meiän alalla hirveen paljon. Se on itse asias semmonen mikä vaikuttaa kustannuksiin ja kaikkeen muuhun. Ku tehään eläinmalleilla töitä, niin se että vaaditaanko tilastotieteellisesti että suurennetaan otoskokoa määrättömästi tämmösiin mitä ihmisist tulee. Että miljoonia COVID-rokotteita nyt arvioidaan, niin siel lähetää kymmenistä tuhansista, sadoista tuhansista ihmistä mistä arvioaan. Ja jos ajatellaa, että pitäs ottaa eläinkokeita, niin se ei oo hirveen vastuullistakanaa toimia sillä tavalla, että yhtä koetta kohti ois kymmeniä tuhansia eläimiä, se ei oo eettisesti oikein. Tää on semmoinen filosofinen kysymys, mistä on vähän tieteilijät jakaantunu senki suhteen, et mihin se tilastotiede vie. Se on osa sitä koko prosessia, että kuinka paljon avata. Se on hyvin haastavaa, et mun mielest on helpompi näyttää kuvat, näytetään numerot, ja jokainen tekee niistä omat johtopäätökset. Meiän johtopäätökset ei välttämät oo aina samat kun jonkun toisen joka sitä lukee.

Henriikka Mustajoki: Ja jos sä tiedät et sä jaat ne numerot, ni onko ne sellasia numeroita, että niitä voi joku muu hyödyntää vähän toisenlaisessa saman alan tutkimuksessa? Ku sä mietit tätä skaalautumista, joka kuulostaa kauheen loogiselta, et sielt tulee aika nopeesti rajat vastaan jos koe-eläinten kanssa toimitaan. Niin löytyykö sun alalta semmost synergiaetua? Et voi käyttää muiden tuottamaa dataa ilman et täytyy kaikkia kerätä itse.

Jarkko Koivunen: Kyllä, hirveän paljon. Suomessa rahoitusmallit on vähän enemmän sinne palkkaan kallellaan, ja esimerkiks sekvensointidatan ja tämmösten tuottaminen, niin se maksaa, se ei oo ilmaista sen datan tuottaminen. Mullaki on ollu semmonen visio, et ois kauheen kiva ku tulis semmonen rahoitus, jossa vaan annettais rahaa ja sanotaa et tuottakaa dataa, eli tuottakaa sekvenssejä, tuottakaa proteomiikkadataa. Eikä sitä, että tehään kokonaiskustannusmalli, josta aina pitää olla mahdollisimman iso osa palkka. Siinä Suomi on vähän ehkä perässä. Mutta sen sijaan maailmalla tuotetaan hirveen paljon ja on näitä servereitä, webbisivuja avataan, jossa monet eri tutkijaryhmät laittaa heidän oman datansa sinne. Siihen löytyy raakadata, siihen löytyy yleensä avoin lähdekoodi, et miten se on analysoitu. Ja sit sieltä löytyy vielä se, että siitä on tavallaan tehty helpoks se, että

ei tarvi olla alan asiantuntija, ni pystyy kaivamaan sieltä semmosia. Yleensä se on sillai et jokaisel on oma suosikkiproteiini tai -geeni, jonka he haluaa löytää mahollisimman helposti ilmaa et he tietää mitään tästä datankaivuusta. Ja he kirjoittaa, ne on tehty semmoseks ne serverit, että kirjoitat sinne sen oman sanan mikä sua mietityttää, ja se antaa sieltä dataa ulos. Eli koko ajan enemmän ja enemmän mennään, data tulee enemmän ja sit sitä pyritään myöski avaamaan koko ajan tällä tavalla.

Henriikka Mustajoki: Mikä motivoi tutkijoita laittamaan sitä dataa sinne serverille?

Jarkko Koivunen: Viittauksiahan siitä tulee, se on varmasti se ensisijainen juttu. Sit tunnettuvuus lisääntyy ja sitä mukaa viittaukset myöski sinne. Et se on varmaan, maine ja kunnia, siitähän se lähtee. Ja toki se vie tietysti tutkimusta valtavan paljon eteenpäin. (- keuhkofibroosi kuuntelin) [05:51], seki oli hieno, vähän liittyy avoimeen dataa, YouTubessa on heidän konferenssit pääsääntöisesti. Sielt on jotain leikattu pois, mut pääsääntöisesti konferenssit löytyy YouTubesta, ja ei tarvi matkustaa Kaliforniaan eikä maksaa hyvin paljon siitä, aina pystyy kattoo jälkikäteen. Nii siellä esimerkiks mones puheessa sanottiin, että pitäs tuottaa mahdollisimman paljon dataa. Et jos halutaan keuhkofibroosia ja tauteja parantaa, että meillä ois valoisampi tulevaisuus, niin sitä dataa pitää tuottaa ja se pitää avata. Eli kyl siinä on tämmöstä, ja pyritään koko ajan enemmän ja enemmän täntyyppiseen asiaan. Syöpätutkijat on ollu jo vuosikausia edellä ehkä muita, mutta nyt alkaa muitakin tauteja ja sairauksia tutkivat avaamaan enemmän ja enemmän sitä dataa. Se vaatii tietysti aina enemmän ku yhen porukan, joka sitä tekee. Mitä useampi sen tekee, niin sen parempi siitä tulee.

Henriikka Mustajoki: Jos sä mietit nyt sun tutkijanuraa, ja datan avaaminen on selvästi lisääntynyt sen aikana. Niin millä tavalla se on muuttanu sun tapaa tehä tutkimustyötä?

Jarkko Koivunen: Asiat on helpottunu hirveen paljon. Ei pelkästään että se sekvensointidata on julkista, mutta sit tarvii olla alan asiantuntija tai ekspertti, että sitä voi analysoida. Että se laitetaan semmosessa muodossa, että kaikilta ei löydy softia siihen avaamiseen, vaan pitää olla tosissaan saman alan spesifinen tutkija, joka sitä pystyy hyödyntämään. Nii nykyisin se on ehkä enemmän sitä, että päinvastoin pyritään tekemään siitä mahdollisimman laadukasta siitä datasta ja yksinkertasta. Meillä tieteenalan sisälläki on semmosia, jotka enemmän on geneetikkoja ja sitten on semmoisia jotka tutkii soluja, solubiologeja, ja hiirimalleihin keskittyviä. Niin se että jokainen pystyy helposti löytämään sieltä tiedonmurusia, jotka voi viiä sitä tutkimusta eteenpäin, nii se on ehkä se suurin muutos mikä on koko ajan. Nää serverit helpottuu ja helpottuu koko ajan. Sieltä pystyy jo, jos tietää mikä geeni on, ni pystyy kattomaan että mikä geeni on ylhäällä missäkin syövässä ja niin edes päin. Se on menossa siinä mielessä parempaan suuntaan. Että tutkijapohja ja määrä jotka pystyy hyödyntämään sitä dataa niin kasvaa ehkä eniten. Se tulee sitä kautta, et se on helpommin hyödynnetty.

Henriikka Mustajoki: Sä mainitsit keuhkofibroosin omana tutkimusaiheena ja että siinä on nyt tapahtunu ilmeisesti ihan viime aikoina muutos, siinä datan avaamisessa. Siinä ilmeisesti tulee joku semmonen eksponentiaalisen kasvun kokemus sitten, et sitä on entistä enemmän nyt ja aina vaan enemmän. Onks tämä se kokemus?

Jarkko Koivunen: No joo, se osin tulee siitä, että nää yksisolusekvensointiporukat löi hynttyyt yhteen. Niitä julkaistiin muutama tossa pari kolme vuotta sitte, ja he tavallaan julkas ne asiat samoihin aikoihin, useempi eri tutkimusryhmä. Ja sit niistä laitettiin semmoinen serveri, nettisivu pystyy. Eli se on osin myös siinä, että siinä vaihees ku he sai tutkimuksen valmiiks, nii heillä oli jo semmonen, et sit laitettii jo melkein saman tien tämmöselle serverille missä niitä pysty kaikki analysoimaan. Hyvä ettei ollut serveri samoihin aikoihin ku oli näissä preprintti bioRxivissa julkastu osa niist dataseteistä. Ei ollu vielä varsinaisest lehdessä vertaisarvioitu, vaan oli vasta preprinttä julkastu, ku ne oli jo siellä serverillä. Se osin oli myös siitä, että data puuttu vielä ihan viime vuosiin asti.

Henriikka Mustajoki: Jos ihmiset ei jakais niille servereille sitä dataa niin ku ne nyt jakaa, niin mullistuisko sun tutkimusmahdollisuudet jos sitä ei olisi saatavilla?

Jarkko Koivunen: Totta kai, siinä tulis tehtyy valtavasti turhaa työtä. Eli sehän se on tän datan avaamisen, ja mitä paremmin kirjallisuutta lukee ja tutkii, nii tiede on kuitenki aika paljon valistuneita arvauksia ja hypoteeseja. Ni se data, mitä muut julkasee, nii sen pohjalta voi hirveen paljon ohjata sitä mitä ite kannattaa tehä. Pystyy tekemään ne oikeet kokeet, eikä niinkään semmosia asioita, jotka ei välttämättä oo niin järkeviä. Eli se muitten data ja kirjallisuus, nii se ohjaa sitä omaa tutkimusta. Ja ehkä se, että ku kirjallisuudessa, mitä avoimemmin data on näkyvillä, nii sen enemmän siitä datasta voi tehä johtopäätöksiä sillai omasta näkökulmastaan. Ja pystyy luottamaan siihen, että mitä artikkelissa sanotaan tai mitä johtopäätöksiä on tehty, että se data tukee sitä artikkelin johtopäätöksiä ja löydöksiä. Eli se parantaa toistettavuutta ja toisaalta myös ohjaa sitä omaa tutkimusta, toivon mukaan parempaan suuntaan, ettei tuu tehtyy ihan jokaista virhettä. Se säästä rahaa ja aikaa huomattavan paljon.

Henriikka Mustajoki: Sä oot tällä hetkellä Oulun yliopistossa. Ja datan avaamiseen liittyy kaikenlaisia työvaiheita ja osaamista. Mitä sä tutkijana koet että on oleellista siinä tuessa, jota organisaatio tarjoaa tutkijoille?

Jarkko Koivunen: Geneetikot ja rakennebiologit esimerkiks on jo vuosikymmeniä edellä kaikkia muita. Rakennebiologit, ne on julkaissu kaiken datan avoimena, metadatat mukaan lukien jo, millon se PDB on, 70- vai 80-luvulla perustettu. Eli he on siellä mihin kaikki muut on vasta pyrkimässä. Ja geneetikoil on ihan sama juttu, että se data on ollu aina avoimempaa. Et yks asia on, et jos ei oo ihan selvää että mitä kaikkee voi avata, niin tarvii tukea siihen että mitä kannattaa, mihin se laitetaan se data, IDA vai mitä on kansallisia palveluita, tai Zenodo, kansainvälisiä palveluita ja muuta. Niitä pitää tehä tunnetuksi. Siellä se on, että mitä avataan ja

ja missä avataan. Kyl se miksi, se on jokaisen tutkijan omassa päätäntävallassa. Se on sit kuitenki sillä taval subjektiivinen aina, et kannattaako avata vai ei.

Henriikka Mustajoki: On jossain määrin puhetta semmosest pelosta, et mitä mulle jää jos mä jaan kaiken mun datan, esimerkiks ensikäyttöoikeudesta ja muusta. Sun kuvauksen perusteella sidekudosbiologian alalla ne meriitit on siitä datan julkaisemista sellasia, että siitä ei tarvitse huolehtia. Vai vieläkö se huolettaa?

Jarkko Koivunen: Ei se datan julkaiseminen, siinä vaiheessa kun se on ulkona. Tietysti sidekudosbiologia on aika spesifinen ala, mutta tiedän paljon rakennebiologiaa ja solubiologiaa ja muuta. Niin siinä vaiheessa jos se oma data on jo ulkona, niin periaatteessa onhan se ennenki ollu aina, jo vuosia, että pyydettäessä se pitäs luovuttaa. Mutta nyt siitä vaan tehään helpompaa ja helpompaa. En koe, et nykymaailmas se on semmonen että siinä joku karkaa. Koe-eläimistä hiirimallitki laitetaan nykysin julkisesti saataville tonne ja ne kaikki pitää olla saatavissa.

Toki voi siinä jäähä meriittejä saamattaki. Esimerkiks mulla on 10 vuotta sitte julkastu artikkeli, jossa me julkaistiin tiettyjä molekyylejä ja mitä ne tekee. Yks firma myy niitä tänäki päivänä, on myyny monta vuotta jo. Ja siellä on kyllä viittaus meiän alkuperäseen artikkeliin, mutta kukaan joka ostaa niitä molekyylejä ei ikinä viittaa meiän juttuun. Eli tavallaan esimerkiks kaupallisesti hyödynnetään tiedettä, mutta sit se ois ihan kauheen kiva, että ku kerran siellä on viite siinä minkä ostat, molekyylin tai vasta-aineen tai minkä tahansa. Ni ois se kauheen kiva että tutkijat myöski viittais siihe. Voi siinä hävitäki vähäsen. Mutta nykypäivänä mä uskon et se on enemmän marginaalista, eikä sille oikein mitään voi. Emmää koe sitä mitenkää isona uhkana. Se on sit eri asia, että jos julkasee preprintin ennen ku on lehdessä, ni sit nykypäivänä tulee hyvinki paljon hankaluuksia. Kyllä niitä skuupattuja juttuja meiänki ihan läheltä löytyy useempiki tarina iha viime vuosina. Että ollaan laitettu preprinttiä ja joku muu ehtii tehä sen jutun, toistaa kaikki asiat ja julkasta ensin. Näitäkin sattuu, valitettavasti

Henriikka Mustajoki: Jos saisit toivoo yhtä asiaa joka liittys dataan ja datan avaamiseen, ni mitä sä toivoisit että seuraavaks tapahtuu?

Jarkko Koivunen: Kyl mä edellee pyytäisin sitä [naurahtaa] rahotusta sen datan tuottamiseen mieluummin ku niitä paikkoja mihin sitä dataa voi laittaa ja avata. Niitä on kuitenki suhteellisen hyvin tulossa. Toki voi olla, että kohta loppuu CSC:ltäki tila, et taivaanrannassa näkyy datapakettien kasvu. En pysty sanoo yhtä asiaa. Sitä että ois paikka missä säilyttää dataa ja sitä pystys tuottamaan, se on ehkä se. Että kaikki rahat ei menis pelkästään ihmisiin ja palkkoihin, vaa että vois tehä ja tuottaa dataa. Ihan koska sitä ois kauheen kivaa tehä täällä Suomessaki, eikä vaan tutkia jenkkien tekemää dataa.

Henriikka Mustajoki: Kiitoksia Jarkko Koivunen tästä tutkijan näkökulmasta solubiologian tutkimusalan edustajana, siitä miltä avoimen datan käyttö näyttää,

sen merkityksestä sun tutkimukselle, ja siitä mitä mahtaa olla tulossa seuraavaksi. Kiitos Jarkko.

Jarkko Koivunen: Kiitos.