2. FÉLÉVI BESZÁMOLÓ

Udvarnoki Zoltán András (udvzoli@gmail.com)

Statisztikus Fizika, Biológiai Fizika és Kvantumrendszerek Fizikája PhD program

Témavezető: Csabai István

A dolgozat címe: Gépi tanulás a tudományban

2020.06.18.

Munkámat ebben a félévben is a korábbi két területen folytattam. Az egyik a következő generációs szekvenálásból származó genomikai adatok feldolgozása, elemzése, a másik MHC-peptid kötődési affinitás becslése.

Az új koronavírus járvány megjelenése új kérdéseket vetett fel mindkét területen, mely aktualitása és fontosság okán a féléves kutatási tevékenységemet is befolyásolta.

Mint tudjuk, a 2019-es év végén egy új típusú koronavírus jelent meg feltehetően a kínai Wuhan tartományban, majd néhány hónap alatt az egész világon elterjedt, és emberéletek százezreit követelte. A vírussal kapcsolatban eddig nem látott gyorsasággal jelentek meg adatok, mind járványterjedést, mind a vírusgenetikát tekintve. Néhány hónap alatt több 10 ezer koronavírus genomot szekvenáltak a világ különböző pontjain a tudósok. Mindez lehetővé tette számunkra is, hogy kutatócsoportunkban a téma felé forduljunk, és elemzéseket végezzünk a különböző területeken.

Kutatási tevékenység:

A félévben a Semmelweis Egyetemmel közös projektben magyar emlődaganatos tumor-normál mintapárok teljes genom szekvenciáinak elemzését folytattam, majd új melanoma minták elemzését kezdtük el.

A korábban elkezdett emlődaganatos minták öröklött mutációinak elemzését befejeztük. Sajnos nem sikerült statisztikailag szignifikáns felfedezéseket tennünk, ami a populációs átlagoktól való eltérést illeti, és esetlegesen az emlődaganatra való hajlamosságot mutatná. Az adatokat összevetettem magyar kolorektális daganatminták adataival, hátha a két kohort nagyobb bizonyosságot ad, de nem voltak reményre okod adó eredmények.

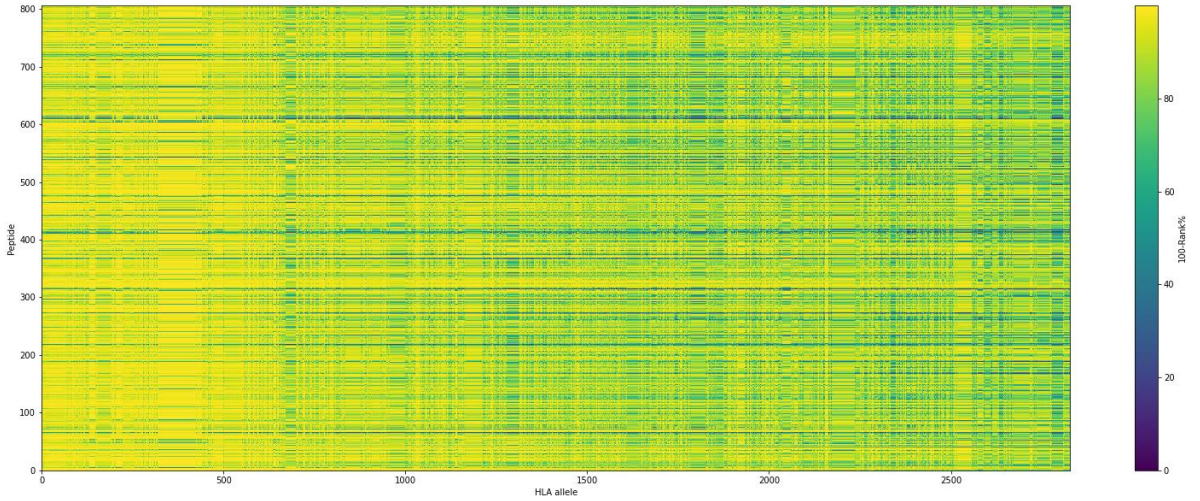
A minták elemzése mellett fontos célunk megosztható adatok közzététele, hogy a nagyobb tudományos közösség is hozzáférhessen. Erre két platform fejlesztését végeztük el, egyik a szomatikus, másik az öröklött mutációk adatbázisát tartalmazza, mindkettő fejlesztésében részt veszek, finomítások még folyamatban vannak, de már a kollaborációban részvevő partnerek számára elérhetővé tettük a működő portalt. Az adatbázis nyilvános közzétételével együtt tervezzük, hogy egy publikációban részletesen leírjuk az adatbázist.

A koronavírus peptidjei és MHC allélok közötti kötődés vizsgálata prognosztikus szempontból rögtön felmerült, hiszen a specifikus immunválasz egyik legszelektívebb lépése ennek a kötődésnek a kialakulása (Prachar et al., 2020; Nguyen et al., 2020; Kivotani et al., 2020).

Két szempontot érdemes figyelembe venni, az egyik az allélok globális eloszlása, a másik pedig a koronavírus peptidome (peptidkészlet) globális eloszlása. Két szoftvert használtam a peptid-MHC kötődés becslésére, az egyik a netMHC a másik a netMHCpan. A különbség a kettő között, hogy míg az első csak elegendő mért adattal rendelkező allélokra

végez becslést, utóbbi bármely allélra, amelynek aminosavszekvenciája elérhető, cserébe kevésbé pontos.

Először olyan allélok keresésével próbálkoztam, amelyek különösen nagy kockázatot jelenthetnek azáltal, hogy nem kötik meg a koronavírus peptidjeit. Az összes allélra megnézni az összes 8-12 hosszúságú peptidet, akárcsak a referenciagenomból, időben túl költséges. Ezért először netMHC-val megszűrtem a kötődő peptideket a netMHCpan-nel pedig ezekre végeztem predikciót. Összefoglaló eredmény látható a 1. ábraán. Ami látszik az ábrán, hogy a név szerint sorbarendezett allélok és lókuszok (HLA-A, HLA-B, HLA-C) egyre rosszabbul teljesítenek, de olyan allél nincs, amely ne kötődne jónéhány peptidhez.



1. ábra Megszűrt peptidek kötődési affinitása az egyes allélokhoz. Nagyobb érték erősebb kötődést jelent.

Megpróbálkoztam az allélfrekvencia és a halálozás közötti kapcsolat felderítésére is.

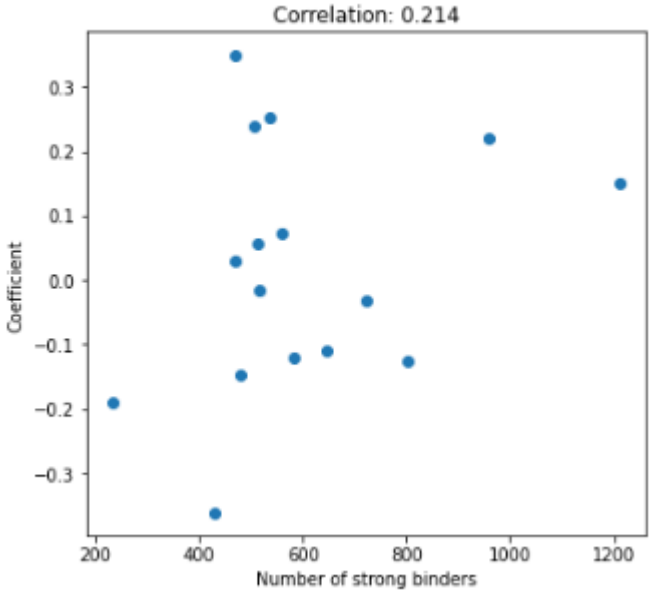
A koronavírusos esetek számai nyilvánosan elérhetők (innen). A mortalitást két módon mérhetjük:

$$mort1 = \frac{haláleset}{lakosság} \quad vagy \quad mort2 = \frac{haláleset}{betegek}.$$

Mindkét módszernél nagy a bizonytalanság. Az első esetben az országonként eltérő járványügyi intézkedések, második esetben a tesztelési irányelvek terhelik zajjal az értékeket. Az egészségügy, egészségügyi morál, és általános egészségügyi állapot különbségeit nem is említem. Mindenesetre reménykedtem, hogy statisztikai módszerekkel felfedezhetők összefüggések.

Az allélfrekvenciákra is rendelkezésre áll egy adatbázis, viszont ez nehezen használható, amelynek legfőbb oka, hogy sok-sok különböző publikáció adatait összesíti. Az egyes országokban különböző népcsoportok különböző arányban szerepelnek, gyakori, hogy csak allélcsoportok szintjén, ez további bizonytalanságot szül.

A halálozás, és az országhoz tartozó allélfrekvencia adatokra regularizált regressziós modellt (EalsticNet) illesztettem. Az allélcsoportokhoz tartozó koefficienseket pedig összevetettem az allélcsoportra becsült kötő peptidek számával. Az eredmény meglepő, ahogy az 2. ábraán is látszik, az allélcsoportok szerepe a halálozásban pozitívan korrelál a kötődő peptidek számával. Ez az eredmény ellentétes a várakozással, miszerint a több kötődő peptid erősebb immunválaszt indukál, az pedig javítja a túlélési esélyeket. Magyarázat lehet rá a citokin vihar, amely túlzott immunválasz miatti halálozást vetíthet előre. Sajnos az adatokban tapasztalt bizonytalanság erős következtetések levonását nem teszi lehetővé.



2. ábra A halálozási regresszióhoz tartozó együttható és a kötődő peptidek száma közötti kapcsolat. Minden pont egy allélcsoportot takar.

Tanulmányi tevékenység:

A félév során az alábbi kurzusokat végeztem el:

- Elméleti evolúcióbiológia (még nem értékelt)
- Sejtszignalizációs hálózatok kvantitatív analízise (még nem értékelt)

Konferencia részvétel:

2. Szint+ Tématerületi Kiválósági Program konferenciája: a konferencia MSTeams-ben került megrendezésre, és többek között aktuális, a koronavírussal kapcsolatos kutatásokról is beszámolt.

Jelentkeztem a London Mathemtical Laboratory nyári iskolájára a Reconstructing Dynamical Systems with Machine Learning témára, viszont ez sajnos a koronavírus járvány miatt elmarad.

Publikációk:

Az emlődaganatos minták elemzéséből egy publikáció megírása a közeljövőben célként van kitűzve.

Mivel az eredményeink egyelőre nem mutattak publikálható újdonságokat, ezért egyelőre az emlő és kolorektális daganatok magyar mintáinak adatait szeretnénk publikálni, amely cikk megírása folyamatban van.

Oktatási tevékenység:

A félévben egy gyakorlati tárgy oktatója voltam.

A Fizika numerikus módszerei 2., Fizika BSc 4. féléves tárgy (fiznum2f18va, mf1c2m04, fiznum2f17va) 3 kurzusánál voltam gyakorlati oktató, összesen heti 6 órában. A távolléti oktatásra való átállás után minden óra kötelező jelleggel megtartásra került, személyes MSTeams konzultáció formájában, így a tárgy oktatásának heti ideje megközelítőleg 10 órára nőtt.

Hivatkozások

Prachar, M., Justesen, S., Steen-Jensen, D. B., Thorgrimsen, S. P., Jurgons, E., Winther, O., & Bagger, F. O. (2020). Covid-19 vaccine candidates: Prediction and validation of 174 sars-cov-2 epitopes. *bioRxiv*.

Nguyen, A., David, J. K., Maden, S. K., Wood, M. A., Weeder, B. R., Nellore, A., & Thompson, R. F. (2020). Human leukocyte antigen susceptibility map for SARS-CoV-2. *Journal of virology*.

Kiyotani, K., Toyoshima, Y., Nemoto, K., & Nakamura, Y. (2020). Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. *Journal of Human Genetics*, *65*(7), 569-575.