1. 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

2019.01.23.

Munkám a félév során két nagyobb területet érintett. Az egyik a következő generációs szekvenálásból származó genomikai adatok feldolgozása, elemzése, a másik MHC-peptid kötődési affinitás becslése.

A következő generációs szekvenálás megjelenésével és elterjedésével az elmúlt évtized során nagy mennyiségű és értékes adathalmaz került a kutatók kezébe. Azonban az eljárásbeli hibák, a módszer korlátai, illetve sok megválaszolatlan kérdés miatt ezen adatok feldolgozása, a benne rejlő információk megértése, és következtetések levonása máig kihívást jelentő feladat. Az elemzések során célunk főként a szomatikus, illetve öröklött mutációk feltárása, a genom strukturális variációinak megtalálása, illetve egyéb, a referenciagenomtól való eltérések megfigyelése, amelyek betegségek szempontjából prognosztikus vagy terápiás jelentőséggel bírhatnak. Ezeknek az aspektusoknak a vizsgálatát megelőzően azonban a mintáról kell általános tulajdonságokat megállapítani, mit például tumor minta esetén a tumor aránya a mintában, illetve a ploiditás, vagyis a sejt homológ kromoszómáinak száma. Hasonlóan, a kisebb léptékű elváltozások előtt a nagyobb léptékűek meghatározása szükséges, mint például az adott genomrégió kópiaszáma. Mindezek viszonylag egyszerű naiv módszerekkel becsülhetők, illetve kész szoftverek is rendelkezésre állnak.

Az MHC-peptid kötés az immunrendszer egyik fontos és legszelektívebb lépése, amely eldöntheti, hogy egy adott sejtet az egyén immunrendszere megtámad-e. Így mind a szervátültetés, mind a daganatkutatás során fontos szerepe van. A kölcsönhatás lényege, hogy a különböző MHC-I fehérjékhez – melyeknek rengeteg típusa (allélja) van, de egy emberben maximum 6 típusa található meg – megpróbáljuk megmondani, hogy egy adott 8-12 aminosav hosszú peptid milyen affinitással kötődik. Ez már egy régóta, de máig folyamatosan kutatott téma, melyben mindig újabb és jobb módszerek jelennek meg.

Kutatási tevékenység:

A félévben a Semmelweis Egyetemmel közös projektben magyar emlődaganatos tumor-normál mintapárok teljes genom szekvenciáinak elemzésében vettem részt. E részeként több módszerrel tettem kísérletet a tumor arány becslésére. Például a heterozigozitás elvesztésének régiójában (LOH régió) az allélfrekvenciából (AF) a következő módon becsülhető a tumor arány (c):

$$c = 2 \cdot |AF - 0.5|$$

Ebben a számolásban felmerül egy probléma, ami az allélfrekvenciák zajosságából következik, és így szükségessé teszi az allélfrekvenciák futóátlaggal történő normálását az LOH régió meghatározásához. A futóátlag hagyományos módon a teljes genomra nem használható, hiszen több mint 3 milliárd bázispárról van szó, így a mutáns pozíciókra szorítkozunk, ezek viszont nem egyenletes távolságra helyezkednek el. Így a futóátlagolást

egy nem egyenletesen elhelyezkedő index alapján kell elvégezni, amire nem találtam kész elérhető algoritmust, így ezt implementálni kellett.

Másik módszer, amely felmerült, egy CNVkit nevű program kimenete alapján számolva a tumor arány. A program egy úgynevezett copy ratio-t (CR) számol, amelynek értéke a tumor minta kópiaszámának (T) és a tumor arány (c) ismeretében:

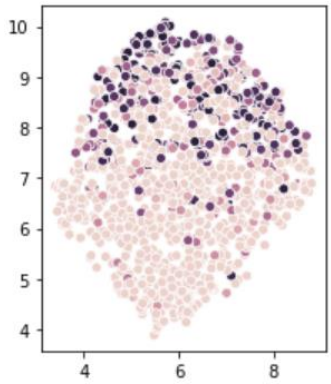
$$CR = \frac{xT + (1 - x) \cdot 2}{2}$$

Ebből T változik a genomi pozíciókra, viszont elméletileg csak egész szám lehet, míg x konstans a teljes genomra. Így x-ben végigpásztázva a 0 és 1 közötti tartományt, amely x mellett a T értékek egész számoktól vett távolsága a legkisebb, az jó becslést adhat a tumor arányra.

A fent említett két naiv módszert összehasonlítva a Sequenza nevű szoftver eredményeivel, illetve a minta preparálásakor végzett vizuális becsléssel néhány esetben jó egyezést kaptam a tumor arányra, azonban mindhárom módszer figyelmen kívül hagyja a tumoron belüli heterogenitást. Így ennek figyelembevétele egy fontos további teendőként merül fel a Sequenza nevű szoftver továbbfejlesztése szempontjából is, amit célként tűztünk ki. A Sequenza programon kisebb javításokat végeztem, így a félév során mélyebben is megismerkedtem a működésével.

A minták további elemzésében is részt vettem, itt a szomatikus mutációk bevett analízisét végeztem el, amely során mutációs spektrumot számoltattam, és ez alapján nem felügyelt gépi tanulási módszerekkel (t-SNE, PCA) a minták klaszterezését végeztem el, és hasonlítottam a daganat fenotípusához, típusához. Az öröklött mutációk elemzését is elvégeztem, ahol ismert emlődaganattal összefüggésbe hozható gének mutációit, illetve a világpopulációtól való eltéréseket kerestünk, esetlegesen a magyar onkogenom meghatározásához. A munka ezen része még folyamatban van.

A MHC-peptid kötések affinitásának becslése témakörben korábban már dolgoztam, és így merült fel, hogy valamilyen módon a 8-12 aminosav hosszúságú peptideket az őket jellemző $8^{20} - 12^{20}$ dimenziós térből egy alacsonyabb dimenziós térbe vetítsük, amelyben esetlegesen a kötődési affinitás szerinti klaszterezettség vizuálisan láthatóvá válik. A UMAP nevű dimenzióredukciós algoritmust használva meglepően jó eredményeket kaptam felügyelet nélküli klaszterezés esetén is néhány MHC allélra. (1. Ábra)



1 Ábra 9 aminosav hosszú peptidek beágyazása 2 dimenzióban. A színek egy MHC allélhoz való kötődési affinitást jelölik.

A módszer használható lenne erős affinitású peptidek jóslására, ami fontos lehet bizonyos immunterápiák szempontjából.

Tanulmányi tevékenység:

A félév során az alábbi kurzusokat végeztem el:

- A káoszelmélet alkalmazása (jeles)
- Evolúciós játékelmélet (jeles)
- Preklinikai modellek a daganatkutatásban (még nem értékelt)

Konferencia részvétel:

Októberben elején részt vettem egy NVIDIA Deep Learning Institute által szervezet Fundamentals of Deep Learning for Natural Language Processing című workshopon és kurzuson, melyet sikeresen teljesítettem.

Novemberben végén pedig HEPTech AIME (Academia-Industry Matching Event) nevű nemzetközi konferencián vettem részt, ahol a gépi tanulás haza és nemzetközi alkalmazási irányairól tájékozódhattam.

Publikációk:

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

Oktatási tevékenység:

A félévben két kurzussal kapcsolatban végeztem oktatási tevékenységet.

A Számítógépes szimulációk Fizikus MSc kurzus (compsimf17em) 2 projektmunkájának javítása az évfolyam hallgatóinak 1/3-a (10 fő) számára, illetve előzetesen tanácsadás a munka elvégzéséhez. Az Adatbányászat és gépi tanulás Fizikus MSc kurzus (dsminingf17vm) házi feladatainak javításában segítettem kb. 2 hetente.