HÁLÓZATI MODELLEK ALKALMAZÁSA A BIOINFORMATIKÁBAN

A DOKTORI ÉRTEKEZÉS TÉZISEI

Ligeti Balázs

Roska Tamás Műszaki és Természettudományi Doktori Iskola

Pázmány Péter Katolikus Egyetem

Információs Technológiai és Bionikai Kar

Témavezető:

Prof. Pongor Sándor

Budapest, 2016

1. Bevezetés

A biológiai adatok hálózatalapú szemlélete alapvetően befolyásolja azt, ahogy az élettudományokban ma a diagnosztikai és terápiás problémákra tekintünk. A hagyományos paradigmák számára az adatok nem mások, mint rendezett adatbázisokban tárolt, alapvetően elszigetelt egységek. Manapság azonban egyre inkább úgy tekintünk ezekre, mint egy összekapcsolt hálózatra.

Számtalan típusú kapcsolat létezik – lehet szó gyógyszerek és betegségek kapcsolatáról, gyógyszerek és fehérjecélpontok, vagy az ezeket termelő gének kapcsolatáról, de vizsgálható a kapcsolat gyógyszerek és más, ezeket helyettesíthető vagy semlegesítő gyógyszerek között is. Ezeken kívül vizsgálhatjuk fehérjék és más, velük fizikailag kölcsönhatásban lévő proteinek, vagy az általuk szabályozott gének kapcsolatát, vagy fehérjék kapcsolatát olyan betegségekkel, amelyekben szerepük van stb.

A kép tehát igen összetett, számtalan típusú egységről és kapcsolatról beszélhetünk, amelyek különböző ontológiákban reprezentálhatóak, mely ontológiák szintén tekinthetők hálózatoknak: koncepciók (fogalmak) hálózatának.

Nyilvánvaló, hogy ilyen hatalmas mennyiségű adat tárolása és kezelése a jelenlegi számítógépek számára igen nagy kihívást jelent. Ráadásul ezek az adathálózatok nem csak hiányosak, de zajosak is. Egészen pontosan rendelkezésünkre áll látszólag hatalmas mennyiségű fehérje, de az ezekről tárolt tudásunkat ritkán igazolják kísérletek, és az annotációk jelentős része csak átvétel különböző organizmusok homológ fehérjéiből. Abban sem lehetünk biztosak, hogy vajon az

adott két fehérje minden egyes szövetben és / vagy a sejtciklus minden egyes fázisában kölcsönhatásban áll egymással.

Ezekre a problémákra az jelentheti a megoldást, ha a feltételezett adathálózatot kisebb, specifikus – betegségspecifikus, szövetspecifikus, patogénspecificus stb. –, manuálisan gyűjtött részekre bontjuk, melyek egy adott problémával kapcsolatosan tartalmaznak megbízható információt. Mindez azonban egy fárasztó és munkaigényes megoldás, amelynek csak a legfontosabb területeken van létjogosultsága. A rákspecifikus adathálózatok kiváló példái ennek a megközelítésnek.

Ezen kívül két fő információforrás segíthet még az adatok hiányosságából fakadó problémákon. Egyrészt a számtalan nagy áteresztőképességű kísérleti módszer (két hybrid rendszerek, DNS-szekvenálás, chip-seq stb.) új típusú molekuláris interakciós adatokat kínál, melyek könnyedén hozzáadhatóak a meglévő adatbázisokhoz. Meg kell azonban jegyezni, hogy a nagy áteresztőképességű adatok általában igen zajosak, amit kezelni szükséges. Ilyen esetekben a hierarchikus adathálózatok (pl. ontológiák) kiváló keretet kínálhatnak a zajcsökkentés és érzékenység közti egyensúlyozáshoz, hogy a kísérletekben lehetőség nyíljon új adatkapcsolatok felfedezésére.

Másrészt a tudományos cikkek absztraktjait vagy teljes szövegét tartalmazó irodalmi adatbázisok hatalmas méretű új tudást nyújtanak, amelyek elvben összeköthetők molekuláris adatokkal. Ugyanakkor ez a folyamat sem egyértelmű: a tudományos szövegek természetes nyelven íródnak és a fogalmak bennük nem mindig analógak azokkal, amelyeket a molekuláris adatbázisok szövegei használnak.

Kutatók nagy csoportja igyekszik kezelni a fenti problémákat betegségspecifikus adatbázisok és eszközök segítségével. A rákadatbázisok és -eszközök egy jellegzetes példái ennek, mivel a rák egyike a legfontosabb komplex betegségeknek, amely az összes halálozás 15%-áért felelős, és amelynek 100-nál is több többé-kevésbé jól körülírható típusa van, és 500-nál több gént kapcsoltak már hozzá [1].

Habár az onkológusok különböző hagyományos adatbázisokat használnak, számos kísérletet szenteltek már arra, hogy különböző ráktípusokról gyűjtsenek össze adatokat. Mindez egy biztos tudásbázist garantál integrált adathálózatok tervezéséhez, amelynek segítségével új, a rákterápiával kapcsolatos kérdésekre lehet majd választ adni.

Jelen dolgozatban három kérdéscsoportra fókuszálok, melyek integrált adathálózatok segítségével vizsgálhatóak: i) rákterápiában hasznos gyógyszer-kombinációk keresése. Ennek a problémának a megoldása egy egyszerű hálózati átfedés-mérték adathálózatokon történő alkalmazásával történik. Valamint ii) új gén-betegség asszociációk felfedése a petefészekrákban potenciális biomarkerek listájának generálása céljából. Ennek a problémának a vizsgálatához egy szövegbányászati megközelítést alkalmaztam a MEDLINE absztraktokon [2] és a STRING adatbázison [3]. iii) Végül egy gyakorlati alkalmazást mutatok be, ami a taxonómiai azonosítást segíti adathálózatok felhasználásával. Itt a taxonómiai és a funkcionális részhálózatoknak azt a tulajdonságát használom ki, hogy ezek hierarchikus gráfok, amely lehetővé tesz egy jelentős sebességnövekedést a jelenlegi algoritmusokéhoz képest.

2. Módszerek

Logikai szempontból minden interakciós és adathálózat gráfnak tekinthető, amelyben a csomópontok olyan entitások lehetnek, mint a molekulák, betegségek, azaz biológiai, fizikai vagy absztrakt objektumok, míg a köztük lévő élek különféle kapcsolatokat jelenthetnek, mint például molekuláris kölcsönhatást, gyógyszer-betegség kapcsolatot stb.

A dolgozat a hálózati szomszédság fogalmát tárgyalja, amely egy csomópont körüli részhálózatként vagy részgráfként definiálható. A részhálózat meghatározása egy adathálózatban történhet statikus vagy dinamikus módszerek segítségével, valószínűségi alapokon.

Abból a feltevésből indulunk ki, hogy egy zavaró hatás egy központi csomópontból, például egy gyógyszer-célpontból tovább terjed. Ez egy dinamikus megközelítés, mivel a hálózat csomópontjai a propagáció során iteratív módon lesznek súlyozva, és végül kiválaszthatóak azok a csomópontok, amelyeknek a súlya valamilyen küszöbérték feletti. Kétfajta, az informatika különböző területein, széles körben alkalmazott propagációs algoritmust használtam: *PageRanket* [4-6] és gráfokon definiált diffúziót [7-10].

A *PageRank* algoritmus a hálózatokon történő véletlen bolyongás egy speciális esete: a bolyongó egy adott csomópontról indul, majd véletlenszerűen választja ki a következő pontot a közvetlen szomszédjai közül, és tovább mozog oda, és így tovább. A *PageRank* esetében a bolyongó nem csak a közvetlen szomszédokhoz mehet, hanem bizonyos valószínűséggel bármelyik más csomóponthoz („visszatérési valószínűség”). Ha a bolyongó számára csak a

csomópontok egy meghatározott halmazára engedélyezett a visszatérés, akkor a *PageRank with prior* algoritmusról beszélünk [5, 6, 11]. Ha rendelkezésünkre áll előzetes tudás arról, hogy mely csomópontok a fontosabbak, akkor ez az információ felhasználható az eredeti *PageRank* értékek újrasúlyozásához. Egyéb, bolyongáson alapuló ismert algoritmus például a *k-step Markov* [11], a *HITS* [12], vagy a *HITS with Prior* [11].

A diffúzió egy fizikai metafora a hálózatokon történő transzport-jelenségek modellezésére. Esetünkben egy képzetes mennyiséget, mint például „energia”, „gyógyszerhatás”, rendelünk a csomópontokhoz – például a gyógyszer által célzott génhez – és ezt követően egy iteratív eljárással számoljuk ki, hogy ez a mennyiség hogyan diffundál a hálózaton. A *PageRank with priorhoz* hasonlóan lehetséges előzetes tudást beépíteni a hálózatról, például egy betegséghez kapcsolódóan a releváns gyógyszerekről, a Laplace mátrix regularizációjával [7]. A regularizáció értelmezhető úgy, mint a diffúziós folyamat módosítása i) egy csomópont energiaveszteségének irányításával (növelésével vagy csökkentésével); ii) bizonyos élek bejövő energiaáramlásának módosításával (növelésével vagy csökkentésével); iii) mindkettővel. A fenti módosítások mindegyike leírható különböző regularizációs paraméterek segítségével.

Közönséges differenciálegyenlet-rendszerek kiértékelése komoly kihívást jelenthet, például a diffúzió esetében is. Ugyanakkor ritka lineáris algebra alkalmazásával és az adathálózatokra jellemző ritkaság tulajdonságának kiaknázásával a számítási idő csökkenthető. A mátrix exponenciális kiértékelése helyett fókuszálhatunk közvetlenül a

mátrix-vektor szorzat közelítésére, ezáltal jelentős gyorsulás érhető el. A $e^{-Lt}x(0)$ kifejezés iteratív módszerek, például az Arnoldi algoritmus használatával közelíthető [13-15].

A hálózati szomszédság meghatározásához mind a *PageRank*, mind a diffúziós módszerek esetében szükséges egy küszöbérték meghatározása, amely alapján a csomópontok (és az élek) kizárhatók vagy bevehetők a hálózati szomszédságba. A küszöbérték meghatározása Monte-Carlo-szimulációk segítségével történhet, amelynél nagyszámú (például 10 000) iterációt indítunk véletlenszerűen, és az így kapott szignifikanciák segítenek a hálózati szomszédság kijelölésében.

3. Új tudományos eredmények

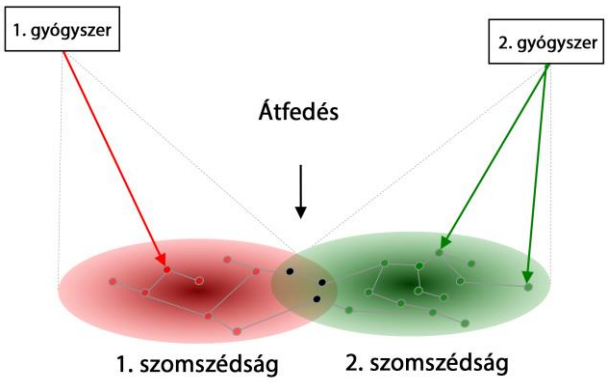
I. Hatékony gyógyszer-kombinációk előrejelzése

A szerző kapcsolódó publikációi: [J1][J3][C7]

A gyógyszer-kombinációk hatékonynak bizonyultak komplex hátterű betegségek szisztémás kezelésében, mint amilyenek a daganatos megbetegedések, a cukorbetegség, az *atrhritis* vagy a magas vérnyomás. A legtöbb jelenleg használt kombinációt tapasztalati úton fejlesztették, ami jelentősen korlátozza új és hatékony kombinációk felfedezésének a gyorsaságát.

I.1. TÉZIS. Bemutattam egy új módszert, ami azon a feltevésen alapszik, hogy a több gyógyszer által generált perturbáció propagál egy interakciós hálózaton, és nem várt amplifikációt okozhat olyan célpontokon, amelyek az eredeti gyógyszerek által közvetlenül nem érintettek. A jelenség megragadásához bevezettem egy új, ú.n. *Célpont Átfedés Értéket* (*Target Overlap Score* - TOS), ami két gyógyszerágensre úgy definiálható, mint a közösen perturbált célpontok száma osztva minden, a két ágens által lehetségesen érintett célpont számával. Tehát két részhálózatra, net_1 és net_2:

$$TOS(net_1, net_2) = \frac{\left| V_{net_1} \cap V_{net_2} \right|}{\left| V_{net_1} \cup V_{net_2} \right|}$$



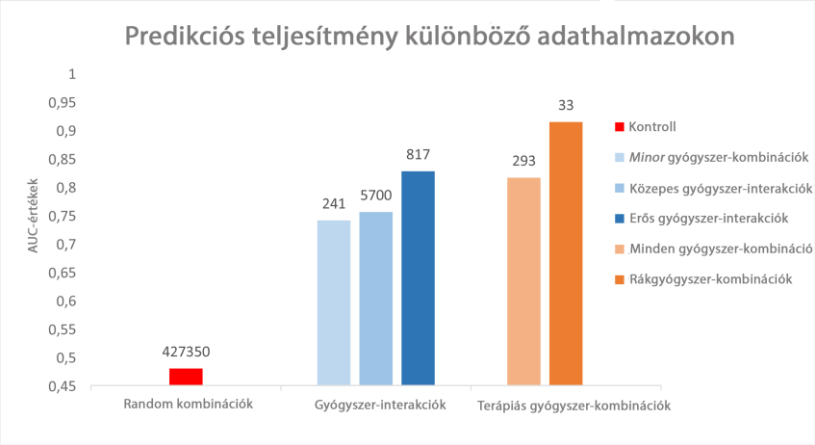
1. ábra. *A részhálózatok átfedését bemutató hipotézis. A két gyógyszer (1. gyógyszer, 2. gyógyszer) először a közvetlen célpontokra van hatással, majd ez tovább terjed a célpontok szomszédjára is (piros, zöld színnel jelezve). Az átfedésben lévő célpontokra (feketével jelezve) mindkét gyógyszer hatással van. Feltesszük, hogy azok a gyógyszerek, amiknek sok közös célpontjuk van, egymás hatását felerősítik.*

1.2. TÉZIS. Megmutattam, hogy mind a gyógyszer-gyógyszer interakciókat, mind pedig a gyógyszer-kombinációkat lehetséges megkülönböztetni a véletlenszerű kombinációktól a TOS-érték segítségével. Azt is bemutattam, hogy ez a mérték jól korrelál a DCDB, TTD és a Drugs.com adatbázisokból vett jótékony és káros gyógyszer-kombinációk ismert hatásaival (2. ábra).

I.3. TÉZIS. Megmutattam, hogy két további, az irodalomban gyakran és sikeresen alkalmazott adattípus integrálásával (úgy mint a közvetlen célpontok közötti funkciók hasonlóságának mértéke <GO> és az ATC-kódok hasonlósága) a TOS-érték teljesítménye nem javult szignifikánsan.

A gyógyszerekről szerzett minél több információ beépítésével a gyógyszer-gyógyszer interakciók előrejelzésének javulását várjuk. Az egyik gyakran sikeresen használt gyógyszerleírás a funkcionális

annotáció (azaz ontológiai kifejezések hozzárendelése a gyógyszerekhez a célpontjaikon keresztül), illetve az ATC-kódok alapján számolt terápiás hasonlóság. A betanított osztályozó (egy logisztikus regressziós modell) különböző mértékekkel (TOS, GO, ATC) nem mutat fejlődést a rangsorolási teljesítményében ahhoz képest, amit a TOS önmagában való alkalmazása nyújtott.



2. ábra. *A predikciós teljesítmény különböző tanítóhalmazokon, rákos terápiákhoz köthető gyógyszer-gyógyszer interakciókon és gyógyszer-kombinációkon volt mérve. A predikciós eljárás egy egyszerű mértéken, a Célpont Átfedés Értéken (*Target Overlap Score – TOS*) alapszik. A predikciós eljárást száz alkalommal ismételtem különböző negatív halmazokkal, majd ezek után alakultak ki az átlag értékek. Az AUC értékek szórása a különböző adathalmazok esetében 0.0001 és 0.006 közé esik.*

I.4. TÉZIS. Megmutattam, hogy a TOS-érték jó korrelációt mutat olyan klinikai vizsgálatok eredményeivel, ahol a trastuzumabot alkalmazták más célzott és kemoterápiás szerekkel HER2-receptor pozitív emlőrákos betegek terápiájában.

Összehasonlítottam a jelenleg is klinikai kutatások alatt álló, mellrák kezelésére javasolt kombinációkat a TOS-értékekkel. A keresést leszűkítettem azokra a kísérletekre, amelyek alkalmazták a RECIST rendszert (Válaszkiértékelési kritériumok *solid* tumorok esetén – *Response Evaluation Criteria In Solid Tumors*). A TOS jól-korrelált számos vizsgált függőváltozóval, mint az *overall response* (r=0.64; p=0.0028), *overall survival rate* (r=0.87; p=0.017) vagy a *confirmed clinical benefit* (r=0.84; p=0.0021).

II. Tumormarkerek előrejelzése szöveg- és adathálózatok integrálásával

A szerző kapcsolódó publikációi: [J1][J6]

A szövegbányászati módszerek elősegíthetik orvosbiológiai hipotézisek generálását azáltal, hogy új asszociációkat fednek fel gyógyszerek és gének között. Korábban kifejlesztettünk egy RaJoLink nevű „ritka-term" modellt ([16]), amelyben olyan kifejezések alapján fogalmazunk meg új hipotéziseket, amelyek ritkán jelennek meg a cél-domain irodalmában.

II.1. TÉZIS. Tovább javítottam a RaJoLink „ritka-term" szövegbányászati modellnek a szenzitivitását a hálózatok elemzésében bevett módszerekkel (*personalized diffusion ranking, PageRank with Prior*) és a STRING fehérje-fehérje asszociációs hálózat használatával.

Mivel számos jelenlegi orvosi hipotézist molekuláris egységek és molekuláris mechanizmusok szempontjából fogalmaznak meg, kiterjesztettem a módszert fehérjékre és génekre egy sztenderdizált szótár és egy gén/fehérje hálózati modell alkalmazásával. A javasolt, tovább fejlesztett RaJoLink ritka-term modell szövegbányászati és génprioritizációs módszereket ötvöz.

Hasznát már ismert, valamint potenciális, petefészekrákkal kapcsolatos gén-betegség asszociációk megtalálásával illusztráltam

MEDLINE absztraktok és a STRING adatbázis használatának segítségével.

II.2. TÉZIS. Az új rangsorolások alapján kiválasztottam 10 olyan gént (RUNX2, SOCS3, BCL6, PAX6, DAPK1, SMARCB1, RAF1, E2F6, P18INK4C <CDKN2C>, és PAX5), amelyek potenciális biomarkerei lehetnek a petefészekráknak, viszont korábban nem szerepeltek a petefészekrák irodalmában. Ezek közül kettőt (RUNX2, BCL6) megerősítettek 2012 után.

A RUNX2 transzkripciós faktor egy feltételezett tumor-szuppresszor gén. Összefüggésbe hozták számos daganatos megbetegedéssel, mint a prosztata-, tüdő-, vagy mellrák, csontvelőrák vagy pajzsmirigyrák. A gén lehetséges szerepét ebben az összefüggésben alátámasztja a hormonreceptorok előrejelző képessége a petefészekrákban [17]. 2012-ben megerősítették, hogy a RUNX2 kapcsolatban áll az előrehaladott tumorprogresszióval epitheliális petefészekrák esetén [18]. Ezen túl a RUNX2 gátlása szignifikáns csökkenést eredményezett a sejtproliferációban.

A BCL6 (B-cell CLL/lymphoma 6) egy másik transzkripciós faktor, amely gyakran mutálódott diffúz nagysejtes limfóma esetén. A gént nem csak limfómához és leukémiához, de mell-, gyomor- és tüdőrák súlyosbodásához is kötötték. Érdekes, hogy mind a BCL6-ra, mind a RUNX2-re hat a prolaktin-elválasztás. Wang és munkatársai kimutatták, hogy a BCL6 egy negatív előrejelző faktor a

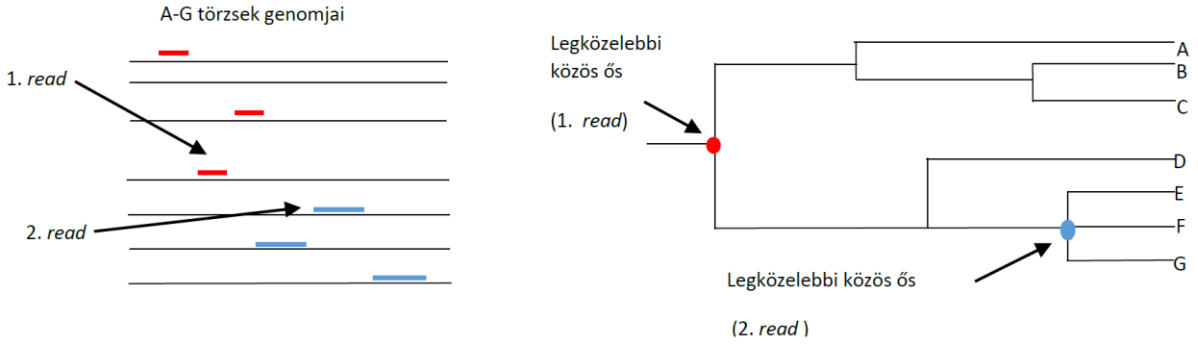
petefészekrákban [19], és a BCL6 NACCI-val együtt történő gátlása csökkentette a rákos sejtek inváziós képességét [20].

III. Mikrobiális vizsgálatok gyors és érzékeny karakterizációja

A szerző kapcsolódó publikációi: [J5]

A metagenomikai minták újgenerációs szekvenálása (*next generation sequencing* – NGS) egyre inkább egy sztenderd eljárássá válik mikroorganizmusok egyedi fajainak vagy patogén törzseinek detektálására. Az NGS-ben használt számítógépes programoknak sebesség és szenzitivitás között kell egyensúlyozniuk, emiatt a faj- vagy törzsszintű azonosítás gyakran pontatlan, és a kis mennyiségű patogének könnyen észrevétlenek maradhatnak.

III.1. TÉZIS. Megmutattam, hogy a taxonómiai fa segítségével a nagy mennyiségű NGS adatok kiértékelése lehetséges, akár egy egyszerű asztali PC-n is, "normális" időben. A módszer alapja, hogy az egyes *readeket* gyors illesztő (bowtie2) segítségével hozzárendeli a taxonómiai fa azon csúcsához, amely a legalacsonyabb közös őse azon mikrobáknak, amelyek genomjára az adott *read* illeszkedik (3. ábra).



3. ábra. A Taxoner algoritmus
Első lépésben a rövid readeket *hozzárendeljük a mikrobiális genomokhoz. Ezután előfeldolgozzuk az illesztéseket és csak azokat a találatokat rendeljük hozzá a taxonómiai fához, melyek egy bizonyos küszöbérték felettiek. Az osztályozási vagy* binning *lépés során a* readet *azoknak a taxonoknak a közös őséhez rendeljük hozzá, amelyek genomjára az adott* read *illeszkedett.*

III.2. TÉZIS. Különböző platformokon (Roche 454, Ion Torrent, Illumina) végzett, ismert vagy ismeretlen patogének (*Staphylococcus aureus*, *Bacillus anthracis*) szekvenálási eredményeit kiértékelve kimutattam, hogy a Taxoner olyan jól, vagy jobban teljesít, mint a BLAST-alapú klasszifikációs módszerek, ugyanakkor két nagyságrenddel gyorsabb azoknál.

Az elmúlt években számtalan szekvenáló platformot fejlesztettek ki. Ezért a Taxoner osztályozási teljesítményét többféle platformról (Roche454, Ion Torrent, Illumina) származó szekvenálási adathalmazon (*Staphylococcus aureus)* többféle programmal (MetaPhlAn [21], Megan-nal [22, 23] kombinált BLAST [24, 25]) vetettem össze. Az alacsony hamis negatív ráta azt jelzi, hogy a Taxoner csaknem olyan megbízható, mint a BLAST+Megan, ugyanakkor jóval kevesebb számítási kapacitást és időt igényel.

III.3. TÉZIS. Megmutattam, hogy a Taxonerrel lehetséges metagenomikai adatsorok karakterizálása a legalacsonyabb taxonómiai szinteken is.

Elemeztem a Humán Mikrobiom Projekt által közreadott MOCK adathalmazt, amelyben 22 mikrobiális törzs és faj található egyenlő arányban. Az adathalmaz 6.5 millió Illumina rövid-*readet* tartalmaz. A Taxoner megbízhatóan azonosította a taxonok nagy részét (14/22), még törzsszinten is.

III.4. TÉZIS. Kimutattam, hogy bár a Taxoner lassabb, mint a MetaPhlAn, két nagyságrenddel kevesebb számú *readre* van szüksége fajszinten történő pozitív detekcióra, ezért a módszer alkalmas kis mennyiségben jelenlévő patogének azonosítására, valamint a rendszer olyan esetekben is alkalmazható, ahol nem ismerjük a mikroba genomját.

Patogének azonosítása esetén az analízis szenzitivitása kulcskérdéssé válik. A szenzitivitás egy adott faj detektálásához szükséges *readek* száma. Egy kísérleti anthrax adathalmaz véletlenszerű mintavételezése után az elemzés kimutatta, hogy a Taxoner 10 *readből* biztosan tudta azonosítani az anthraxot, míg a MetaPhlAnnak ugyanehhez 200-350 *readre* volt szüksége.

Az ismeretlen szekvenciájú mikroorganizmusok szenzitív detektálása szintén fontos kérdés, mivel nagy részük még mindig ismeretlen. A Taxoner ismeretlen fajokon – azaz olyanokon, amelyeknek a genomszekvenciája hiányzik az adatbázisból – való osztályozási teljesítményének értékeléséhez elemeztem egy kísérleti anthrax adathalmazt (*B. anthracis* törzs BA104; NCBI taxon azonosító: nem elérhető). A Taxoner a *readek* többségét (96.5%) *Bacillus anthracis*-ként osztályozta, s csupán egy kis részét (1.2%) a *Bacillus genus*-on belüli más fajként.

4. Az eredmények alkalmazási területei

Munkám során azt vizsgáltam, hogy különböző gráfmodellek hogyan segíthetnek bioinformatikai problémák megközelítésében. Kutatásom a következő területekre fókuszált: i) új gyógyszer-kombinációk keresése, ii) új, váratlan biomarkerek felfedezése az irodalomban, iii) metagenomikai *readek* osztályozási teljesítményének növelése.

A hálózatelemzési módszerek nem csak a betegségek és a gének közötti új kapcsolatok felfedezésében nyújthatnak segítséget, hanem új, jótékony gyógyszer-kölcsönhatások előrejelzésére is szolgálhatnak, ezáltal lehetővé téve a daganatos megbetegedések terápiájának a javítását.

Bemutattam, hogy a szövegbányászat hálózatelemzéssel való kombinálása segíthet új petefészekrák-biomarkerek azonosításában. 2012 óta az algoritmusom által választott tíz, teljesen újnak számító asszociációból kettőnek a szerepe megerősítésre került más tanulmányokban [18-20, 26].

A gráfmodellek alkalmazása nem korlátozódik a betegség-gén-gyógyszer kapcsolatok felfedésére. Ezeken túl a nagy áteresztőképességű adatok, például metagenomikai adatsorok elemzését is segíthetik.

A dolgozat világossá teszi, hogy a hálózatelmélet alapelveinek alkalmazása segíthet gyógyszerek, betegségek és mikrobák közötti váratlan, nem-triviális kapcsolatok felfedezésében.

5. Publikációk

[J1] **Ligeti, B.,** Menyhárt, O., Petrič I., Győrffy, B.; Pongor, S. (2016). Propagation on Molecular Interaction Networks: Prediction of Effective Drug Combinations and Biomarkers in Cancer Treatment. *Current Pharmaceutical Design*, in press.

[J2] **Ligeti, B**.; Vera, R.; Juhász, J.; Pongor, S. (2016). CX, DPX and PCW: Web servers for the visualization of interior and protruding regions of protein structures in 3D and 1D. *Springer Protocols: Methods in Molecular Biology*, in press.

[J3] **Ligeti, B.**; Pénzváltó, Z.; Vera, R.; Győrffy, B.; Pongor, S. (2015). A Network-Based Target Overlap Score for Characterizing Drug Combinations: High Correlation with Cancer Clinical Trial Results. *PLoS One*. **10** (9), e0129267.

[J4] Hudaiberdiev, S.; Choudhary, K.; Vera, R.; Gelencsér, Zs.; **Ligeti**, B.; Lamba, D.; Pongor, S. (2015); Census of solo LuxR genes in prokaryotic genomes. *Front. Cell. Infect. Microbiol*. 5:20. doi:10.3389/fcimb.2015.00020

[J5] Pongor, L. S.; Vera, R.; **Ligeti** B. (2014). Fast and Sensitive Alignment of Microbial Whole Genome Sequencing Reads to Large Sequence Datasets on a Desktop PC: Application to Metagenomic Datasets and Pathogen Identification. *PLoS One*, published 31 Jul 2014, 10.1371/journal.pone.0103441

[J6] Petrič, I.; **Ligeti**, B.; Gyorffy, B.; Pongor, S. (2014). Biomedical Hypothesis Generation by Text Mining and Gene Prioritization. *Protein Pept Lett*. 20, 1-1.

[C7] **Ligeti**, B.; Vera, R.; Lukacs, G.; Gyorffy, B.; Pongor, S. (2013). Predicting effective drug combinations via network propagation. *Biomedical Circuits and Systems Conference*, 378-381.

[J8] Vera, R.; Perez-Riverol, Y.; Perez, S.; **Ligeti**, B.; Kertész-Farkas, A.; Pongor, S. (2013). JBioWH: an open-source Java framework for bioinformatics data integration. *Database*. 2013, bat051.

6. Hivatkozások

[1] Pavlopoulou A, Spandidos DA, Michalopoulos I. Human cancer databases (Review). Oncology reports, 2015; 33: 3-18.

[2] Petric I, Ligeti B, Gyorffy B, Pongor S. Biomedical hypothesis generation by text mining and gene prioritization. Protein and peptide letters, 2014; 21: 847-857.

[3] Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res, 2013; 41: D808-15.

[4] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. 1999.

[5] Haveliwala TH. Topic-sensitive pagerank. In: ed.^eds., Proceedings of the 11th international conference on World Wide Web. ACM, 2002; pp. 517-526.

[6] Jeh G, Widom J. Scaling personalized web search. In: ed.^eds., Proceedings of the 12th international conference on World Wide Web. ACM, 2003; pp. 271-279.

[7] Ito T, Shimbo M, Kudo T, Matsumoto Y. Application of kernels to link analysis. Proceedings of the eleventh …, 2005: 586-592.

[8] Kandola J, Shawe-Taylor J, Cristianini N. On the application of diffusion kernel to text data. In: ed.^eds. Technical report, Neurocolt, 2002. NeuroCOLT Technical Report NC-TR-02-122, 2002.

[9] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge university press 2004.

[10] Kondor RI, Lafferty J. Diffusion kernels on graphs and other discrete input spaces. In: ed.^eds., ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning. Morgan Kaufmann: San Francisco, CA, USA, 2002; pp. 315-322.

[11] White S, Smyth P. Algorithms for estimating relative importance in networks. In: ed.^eds., Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003; pp. 266-275.

[12] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 1999; 46: 604-632.

[13] Eiermann M, Ernst OG. A restarted Krylov subspace method for the evaluation of matrix functions. SIAM Journal on Numerical Analysis, 2006; 44: 2481-2504.

[14] Moler C, Van Loan C. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM review, 2003; 45: 3-49.

[15] Saad Y. Analysis of some Krylov subspace approximations to the matrix exponential operator. SIAM Journal on Numerical Analysis, 1992; 29: 209-228.

[16] Petrič I, Urbančič T, Cestnik B, Macedoni-Lukšič M. Literature mining method RaJoLink for uncovering relations between biomedical concepts. J Biomed Inform, 2009; 42: 219-227.
[17] Fekete T, Rásó E, Pete I, Tegze B, Liko I, Munkácsy G, Sipos N, Rigó JJ, Györffy B. Meta-analysis of gene expression profiles associated with histological classification and survival in 829 ovarian cancer samples. International Journal of Cancer, 2012; 131: 95-105.
[18] Li W, Liu Z, Chen L, Zhou L, Yao Y. MicroRNA-23b is an independent prognostic marker and suppresses ovarian cancer progression by targeting runt-related transcription factor-2. FEBS letters, 2014; 588: 1608-1615.
[19] Wang Y-Q, Xu M-D, Weng W-W, Wei P, Yang Y-S, Du X. BCL6 is a negative prognostic factor and exhibits pro-oncogenic activity in ovarian cancer. American journal of cancer research, 2015; 5: 255.
[20] Shan W, Li J, Bai Y, Lu X. miR-339-5p inhibits migration and invasion in ovarian cancer cell lines by targeting NACC1 and BCL6. Tumor Biology, 2016; 37: 5203-5211.
[21] Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods, 2012; 9: 811-4.
[22] Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res, 2007; 17: 377-86.
[23] Huson DH, Richter DC, Mitra S, Auch AF, Schuster SC. Methods for comparative metagenomics. BMC Bioinformatics, 2009; 10 Suppl 1: S12.
[24] Shiryev SA, Papadopoulos JS, Schaffer AA, Agarwala R. Improved BLAST searches using longer words for protein seeding. Bioinformatics, 2007; 23: 2949-51.
[25] Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol, 2000; 7: 203-14.
[26] Wang Z-Q, Keita M, Bachvarova M, Gobeil S, Morin C, Plante M, Gregoire J, Renaud M-C, Sebastianelli A, Trinh XB. Inhibition of RUNX2 transcriptional activity blocks the proliferation, migration and invasion of epithelial ovarian carcinoma cells. PloS one, 2013; 8: e74384.