

Bakteriális patogén és ember közötti molekuláris hálózatok vizsgálata

Horváth Balázs

2015

1. Tartalomjegyzék

2. Rövidítésjegyzék

PPI - protein protein interaction NCBI GABARAP -Gamma-aminobutyric acid receptor-associated protein MAP-ok MAP1LC3B - Microtubule-associated proteins 1A/1B light chain 3B

3. Bevezetés

3.1. A bél mikrobióta fontosságának ismertetése

Miért van szükség a bél mikrobióta vizsgálatára?

A humán bél mikrobióta egy komplex ökoszisztéma. A mikrobiomot alkotó sejtek száma nagyjából a humán szomatikus és csírasejtek összegének tízszerese. A bél mikrobiom mind metabolikusan, mind immunológiaiilag komplex kapcsolatban áll az emberrel. [?] Eddig több mint három millió nem redundáns mikrobiális gént sikerült kimutatni az emberben [?]. Ez a nagy genetikai állomány lehetővé teszi, hogy olyan metabolikus folyamatok játszódjanak le a humán bélben, melyeket az emberi sejtek nem képesek végrehajtani. [?] A bél mikrobióta felelős bizonyos glikánok, aminosavak és xenobiotikumok metabolizmusáért valamint rövid láncú zsírsavak (*short chained fatty acids* - SCFA-k), vitaminok és kofaktorok termeléséért. A gazda által meg nem emésztett poliszacharidok bontását a bél mikrobióta végzi, mely folyamat eredményeképpen olyan rövid láncú zsírsavak keletkeznek mint az acetát, proprionát és vajsav. [?]

A bélflóra kulcsszerepet játszik az immun-homeosztázis fenntartásában. Az immunrendszerrel bakteriális mintázatokat észlelő receptorokon keresztül és GPCR-ek által van kapcsolatban. A mikroorganizmusok által termelt SCFA-k képesek GPCR-eken keresztül sejtszignalizáció indítására. A veleszületett immunrendszer nagy részét alkotó monociták és neutrofil granulociták rendelkeznek GPR43 receptorral, mely szintén SCFA érzékeny, tehát a bélflóra metabolitokon keresztül is kapcsolatban áll az immunrendszerrel. [?]

A bélflóra hatással van még a gazda metabolizmusára is. Az *Eubacterium spp.* által oligoszacharidokból képzett vajsav részt vesz az emberi szervezet energia egyensúlyának szabályzásában. [?] Az enteroendokrin sejtek és az adipociták is rendelkeznek a GPR41 receptorral mely vajsavra és proprionátra is érzékeny. Adipocitáknál ez a GPR41 szignalizáció *leptin* elválasztást eredményez. [?] A vajsav segít a karcinogenezis kivédésében mivel apoptózis indukáló és proliferáció gátló hatása van. Éppen ezen okokból a bél mikrobióta tekinthető egy új metabolikus szervnek is. [?] Kapcsolatok mutathatók ki a bél mikrobiom megváltozása és olyan betegségek között mint az IBD (*inflammatory bowel disease*), elhízás vagy a különböző rák típusok. [?]

A bél mikrobióta vizsgálatának módszerei

A mikrobióta vizsgálatát elsősorban a különböző meta omikák eszköztárával közelítik meg. Ezek közül is a legfontosabb eszköztár a metagenomika, de alkalmaznak már metabolomikai, metatranszkriptomikai és metaproteomikai megközelítést is. A metagenomikai vizsgálatok során a környezetből származó mintát megfelelő előkészítés után közvetlenül *shotgun* szekvenálásnak vetik alá. [?]

Quin és társai 2010-re meghatározták a minimális bél metagenomot. A vizsgálat során Illumina GA short-read alapú technológiával 124 egy kohortba tartozó nordikus és mediterrán személy székletmintáját elemezték. Az ebből kinyert 576,7 gigabázisnyi DNS-ből 3,3 millió nem redundáns mikrobiális gént mutattak ki. Az így kimutatott gének az emberi genom százötvenszeresét teszik ki. A minták egészére jellemző, hogy a bennük található gének két fő részre osztható: A legnagyobb csoportba (86%) a sűrűn előforduló mikrobiális gének, míg a másik fő csoportba pedig a kifejezetten a humán bélflórára jellemző mikrobiális gének tartoznak. Az összes személyből származó vizsgált génhalmaz 99,1%-a *Eubacteria*, 0,8%-a *Archea* és a fennmaradó 0,1%-a pedig vegyesen *Eucaryota* és virális eredetű. A bakteriális eredetű gének összesen 1000-1150 uralkodó baktériumfajhoz tartozhatnak, ami személyenként kb. 160 domináns fajt jelent. A személyekre jellemző nagyjából 160 uralkodó baktériumfaj listái között a személyeket összevetve nagyfokú hasonlóság figyelhető meg. Egy adott személy bél metagenomjának minimálisan 40%-a megtalálható a minták legalább felében. A közelítőleg ezer fajból 75 faj található meg a minták több mint felében és 57 faj van ami a minták nagyobb mint 90%-ban kimutatható. [?]

3.2. A Humán-Salmonella kapcsolat ismertetése és hatása az autofágiára

Salmonella spp.

A *Salmonellák* olyan Gram-negatív patogének melyek az állatok széles skáláját képesek fertőzni. A tudomány jelenleg több ezer szerotípust ismer, melyek két fő típusra oszthatók. Az egyik fő típus a *Typhoid*, ebbe a csoportba tartozik a *Typhi* és *Paratyphi* melyek kifejezetten embert fertőznek. A másik fő csoport a *Non-typhoid* amelybe tartozó baktériumok már széleskörű gazdaspecifitással rendelkeznek.

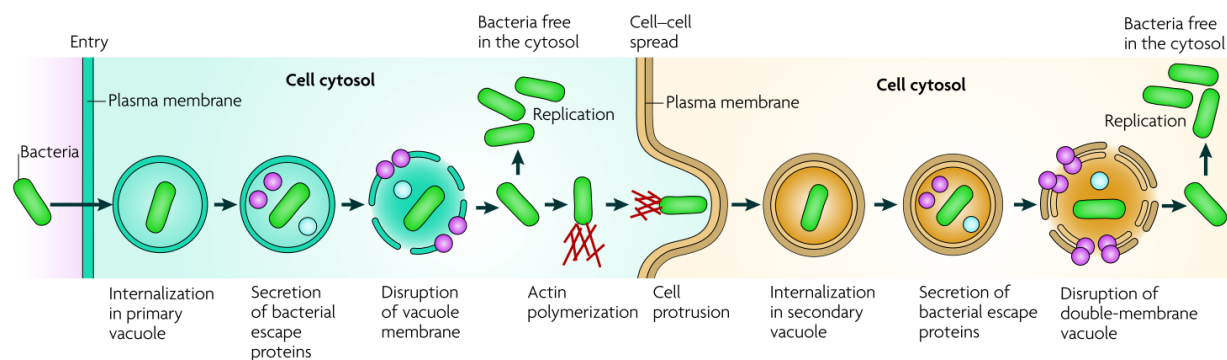
A fertőzés kontaminált étel vagy folyadék fogyasztásával történik. A *Salmonellák* az alacsony pH és oxidatív stressz ellen adaptív toleranciával rendelkeznek, így képesek

eltűnni a gyomor savasságát és a veleszületett immunrendszer egyéb hatásait. A vékonybélbe jutva az epithélium sejtjeit fertőzik. Fő célpontjaik a *microfold* (*M cells*) sejtek, melyek fő feladata, hogy pinocitózissal mintákat vegyenek a középbél atnigénjeiből és ezt antigén prezentáló sejteknek adják. Azonban a *Salmonellák* úgynevezett baktérium-közvetített endocitózissal képesek még a nem fagocita típusú enterocitákba is bejutni [?]

A *Salmonella* életciklusa

Az intracelluláris baktériumok életciklusa általánosan három stádiumra osztható: A bejutáshoz használt vakólum elhagyása, replikáció a citoszólban és a citoszólikus veleszületett immunitás elemeinek manipulációja. A *Salmonella* az úgynevezett *trigger* mechanizmussal jut be a sejtbe. A mechanizmus során a baktérium olyan fehérjéket juttat be az eukarióta sejtbe, melyek képesek a sejtvázzal kölcsönhatni. Ezek a bakteriális effektorfehérjék nagyfokú sejtvál-átrendeződést váltanak ki az eukarióta gazdában. A folyamat végén a baktérium egy vakólummal határolva a sejt belsejébe kerül. [?] Ezt a képletet a szakirodalomban SCV-nek nevezik (*Salmonella containing vacuole*). [?]

A fagocitózis végeztével a *Salmonellák* átesnek egy úgynevezett bakteriális felszín átformázáson (*bacterial surface remodeling*). A folyamat során represszálódnak az olyan bakteriális gének expressziója amit a gazda könnyen fertőzési jelnek tekinthet. Ilyen gének például a SPI1, a T3SS és a flagellin. Mindezek mellett megváltozik a baktériumok felszíni lipopoliszacharid mintázata is. [?]



1. ábra. Az intracelluláris baktériumok életciklusa

Bejutáskor a baktériumok egy elsődleges vakólumba érkeznek. A sejt a belsejében a mikrobák olyan fehérjéket szekretálnak, melyek felbontják az őket határoló elsődleges vakólum membránját. A legtöbb intracelluláris baktériumra jellemző, hogy befolyásolni tudja az aktin polimerizációt és ezáltal képes az intra- és intercelluláris mozgásra. A szomszédos sejtbe átjutott baktériumok egy másodlagos membránburokba kerülnek, melyet ugyancsak felbontanak.

!TODO kép magyarázás és formázás, jobban látható feliratok

Normális körülmények között a vakólum pH-ja mindaddig fokozatosan csökken amíg érett degradatív fagolizoszómává nem válik. A baktériumok kétféleképpen képesek életben maradni ebben a környezetben: A vakólum-lizoszóma fúzió gátlásával, vagy a fagolizoszóma összetételének aktív módosításával. [?] A szakirodalomban még nincs kialakult álláspont arról, hogy a *Salmonellák* melyik mechanizmust használják. Bizonyítottan képesek életben maradni, olyan SCV-ben mely már fuzionált a lizoszómával, viszont a fő útvonal valószínűleg a vakólum savanyítási folyamatának késleltetése lehet. [?]

Az SCV-n belül a *S. typhimurium* képes a replikációra. A hármastípusú szekréciós rendszer segítségével a baktériumsejtek olyan anyagokat tudnak kibocsátani, melyek lehetővé teszik az SCV-ből kijutást és citoplazma invázióját. [?]

A hármastípusú szekréciós rendszer (T3SS vagy TTSS)

A T3SS evolúciósan a flagelláris export rendszerrel mutat rokonságot. Jelenléte esszenciális ahhoz, hogy a *Salmonella* képes legyen a fertőzésre és gazda sejtjeinek kolonizálására. A T3SS felelős a baktérium virulencia vagy effektor fehérjéinek átviteléért. Az effektorok az eukarióta sejtbe jutva megváltoztatják annak sejtfunkcióit. Az virulenciafehérjék átalakítják a gazda citoszkeleton architektúráját, membrán anyagáramlását, szignál transzdukcióját és citokin expresszióját, ezzel segítve a baktériumok túlélését és további kolonizációját. [?]

A *Salmonella* és az autofágia kapcsolata

Az autofágia egy intracelluláris katabolikus folyamat melynek szerepe van a fehérjeaggregátumok és károsodott sejtorganellek eltávolításában és a veseszűletett immunrendszer működésében.

A *xenofágia* az autofágiának azon formája mely során az intracelluláris baktériumok és vírusok szelektív felismerése és lebontása történik. A szelektív felismerésért az autofágia adaptor fehérjéi felelősek. Ilyen receptor fehérje például a p62 (SQSTM1), a NDP52, optineurin (OPTN) és az NBR1. Az előbb felsorolt receptorok a szubsztátjuk megkötése után kargo adaptorként viselkednek az LC3 (ATG8) számára. *Salmonella* fertőzéskor a sérült SCV-ből kilépett baktériumok sejtfelszíni fehérjéi poliubiquitin borítást kapnak amit a kargo adaptor fehérjék érzékelnek. *S. typhimurium* fertőzéskor a poliubiquitinált baktériumokat NDP52 és a p62 is felismeri. Az így megkötött baktériumok xenofágia útján eltávolítódnak. [?]

3.3. Ökológiai hálózatok elemzésére használt topológiai mérőszámok

Miért van szükség topológiai mérőszámokra?

A konzervációs biológia az élettudományok azon ága mely a Föld biodiverzitásának megőrzésével foglalkozik. Mivel az összes faj védelme nem megoldható, ezért szükségessé vált olyan fajok kiválogatása melyek kiemelt figyelmet igényelnek konzervációs biológiai szempontból. [?] Az 1990-es évek előtt a védelemre való kiválasztás fő szempontja a faj ritkasága volt. A fajok ilyen alapú szelekciója nem veszi figyelembe hogy például az adott taxon kulcsszerepet játszik-e az ökoszisztéma funkciók ellátásában. [?]

Kulcsfajok

1966-ban Robert Paine megalkotta a kulcsfaj koncepciót(*keystone species*). Megfigyelte hogy ha kiesik a Kaliforniai sziklás tengerparti közösségből a *Piaster ochraceus* csúcsragadozó tengeri csillag akkor az egész közösség fajösszetétele összeomlik. A mai legelfogadottabb kulcsfaj definíció szerint ezek olyan fajok, melyek ökológiai hatása aránytalanul nagy az abundanciájukhoz képest. A fogalommal kapcsolatban azonban további kérdések merülnek fel: Milyen hatás számít nagyknak? Pontosan mekkora biomassa hányad után mondható az adott faj ereje aránytalanannak? [?] Ez utóbbi kérdések megválaszolásához szükség van olyan mérőszámokra, melyek segítségével kvantitatívvá tehető egy adott faj ökológiai fontossága. Másrészt így lehetővé válik a fajkiválasztás során a szubjektivitás

csökkentése is. Az ilyen mérőszámok használatával objektív fontossági sorrendet lehet felállítani az adott élőhelyen előforduló taxonok között. [?]

Rangsorolásra használt topológiai mérőszámok az ökológiában

Ma már a kulcsfajok kiválasztása részben ökológiai interakciós hálózatok elemzése alapján történik. A használt hálók kizárólag biotikus-biotikus (faj-faj) kapcsolatokat tartalmaznak. Erre azért van szükség, mert például minden élőlény összekötésben áll a detritusszal és ez eltorzítaná az analízis eredményét. Sőt ilyen esetben a detritusz maga is struktúrális kulcsfajnak számítana. Egy adott fajnak az ökológiai interakciós hálóban betöltött szerepét pozicionális fontossági mérőszámokkal, vagy más néven centralitási indexekkel lehet jellemezni. A konzervációs biológiában sokfajta ilyen mérőszámot használnak, melyeknek közös tulajdonsága, hogy mindegyik valamilyen egyedi tulajdonságra fekteti a hangsúlyt és az alapján rangsorolja a hálózatban szereplő fajokat. Ilyen eltérés lehet két index között például, az hogy az egyik egy adott pont lokális kapcsolati mintázatára, míg a másik az egész hálózatra vonatkozó hatását számszerűsíti. Adott hálóra különböző mérőszámok eltérő fajsorrendeket adnak, de a hasonló tulajdonságok figyelembevételén alapuló mérőszámok között felállíthatók konszenzus fák. [?]

Főbb topológiai mérőszámok

Normalised degree - D

Az adott ponttal kapcsolódó pontok száma elosztva a hálózat összes pontjának számával. [?]

Closeness centrality - CC vagy C

A pontok száma elosztva az adott pontból eredő azt minden más ponttal összekötő legrövidebb topológiai távolságok összegével. [?] Ez a mérőszám megmutatja, hogy egy adott pontnak mekkora az átlagos távolsága a hálózat összes többi pontjától. Az index kici szám olyan pontokra melyek rövid legrövidebb útvonalakon vannak a többi ponttal összekötve. Az ilyen pontok valószínűleg könnyebben elérnek más pontokat vagy nagyobb hatást tudnak gyakorolni más pontokra. Adott i pont átlagos legrövidebb távolságát a

többi ponttól a következőképpen lehet kiszámolni: [?]

$$\ell_i = \frac{1}{n-1} \sum_j d_{ij} \quad \text{vagy,} \quad \ell_i = \frac{1}{n} \sum_{j(\neq i)} d_{ij} \quad (1)$$

Ahol:

ℓ_i : Az i pont átlagos legrövidebb távolsága a hálózat többi pontjától.

d_{ij} : Az az i pontot a j ponttal összekötő legrövidebb útvonal (geodézikus útvonal) pontjainak száma.

n : A hálózat pontjainak száma.

A két számítás között stratégiai különbség van. A baloldali egyenlet azt feltételezi, hogy adott pontnak önmagára mért hatása nem releváns a hálózat működésének szempontjából. Azonban még erre az esetre is jellemző, hogy mivel definíció szerint a d_{ii} távolság 0, ezért az összeget ez az érték nem növeli csupán az osztót. [?]

Az ℓ_i érték önmagában még nem centralitási index, mert kis számokat ad a magas központiságú pontokra. Ahhoz, hogy megkapjuk a *Closeness Centrality*-t az ℓ_i inverzét kell vennünk: [?]

$$C_i = \frac{1}{\ell_i} \quad (2)$$

Betweenness centrality - BC

A vizsgálni kívánt ponton áthaladó a hálózat többi pontpárját összekötő legrövidebb utak összege elosztva a hálózat többi pontpárját összekötő összes legrövidebb út összegével. [?] Ez a mérőszám azt mutatja meg, hogy egy adott pont milyen arányban szerepel a többi pont között futó útvonalakban. A *betweenness centrality* vagy röviden *betweenness* olyan hálózatok jó jellemzője, melyekben valamilyen természetű „áramlás” folyik a pontok között. Ha feltételezzük, hogy egy ilyen hálózat minden kapcsolata között az áramlás során ugyanannyi kicserélődés történik egy egységnyi idő alatt és a kicserélődés a legrövidebb útvonalakon folyik, akkor az összes geodézikus útvonalon is azonos rátával történik az áramlás. Ez azt jelenti, hogy egy adott ponton átmenő áramlás mennyisége arányos azzal, hogy a hálózat legrövidebb útvonalainak milyen arányában szerepel. [?]

Topological importance - TIⁿ

Ez egy teljesen topológiai alapú mérőszám mely összegzi az egy adott pontból kiinduló összes lehetséges n lépéshosszúságú útvonal hatását. A hálózat összes direkt kapcsolatára kiszámítható azok topológiai erőssége:

$$d_{X,Y} = \frac{1}{x} \quad (3)$$

Ahol:

$d_{X,Y}$: Az Y pont hatása X pontra.

x : Az X pont első szomszédainak száma.

Az így kiszámolt közvetlen kapcsolatok hatását egy mátrixban lehet ábrázolni, melynek indexelése a populációdinamika konvencióit követi: d_{ij} jelenti a j pontnak az i pontra gyakorolt hatását. Adott direkt kapcsolat hatásának nagysága a kapcsolat irányától is függ, tehát d_{ij} nem feltétlenül ugyanakkora mint d_{ji} . Egy n lépés hosszú útvonal erejét az ezt alkotó direkt kapcsolatok hatásának szorzataként értelmezzük:

$$d_{p_{XY}}^n = \prod_{i=1}^{n-1} d_{i,i+1}^1 \quad (4)$$

Ahol:

p_{XY} : Útvonal amire igaz hogy $p \in \{X \text{ és } Y \text{ közötti } n \text{ lépés hosszúságú útvonalak}\}$

$d_{p_{XY}}^n$: Az X és Y pontok közötti n lépés hosszú p útvonal ereje.

$d_{n,n+1}^1$: Az útvonal i és $i + 1$ -ik pontja közötti direkt kapcsolat erőssége

Ez alapján egy Y pont hatása X -ra n lépés távolságban:

$$d_{XY}^n = \sum d_{p_{XY}}^n \quad (5)$$

Ahol:

p_{XY} : Útvonal amire igaz hogy $p \in \{X \text{ és } Y \text{ közötti } n \text{ lépés hosszúságú útvonalak}\}$

d_{XY}^n : Az összes Y pontból eredő és X -ben végződő n hosszúságú útvonalak erejének összege.

Mivel a direkt kapcsolatok ereje függ a kapcsolat irányától, így a TI tükrözi a kapcsolat asszimmetrikusságát is. Egy adott pontra TI^n a következő képen számítható ki:

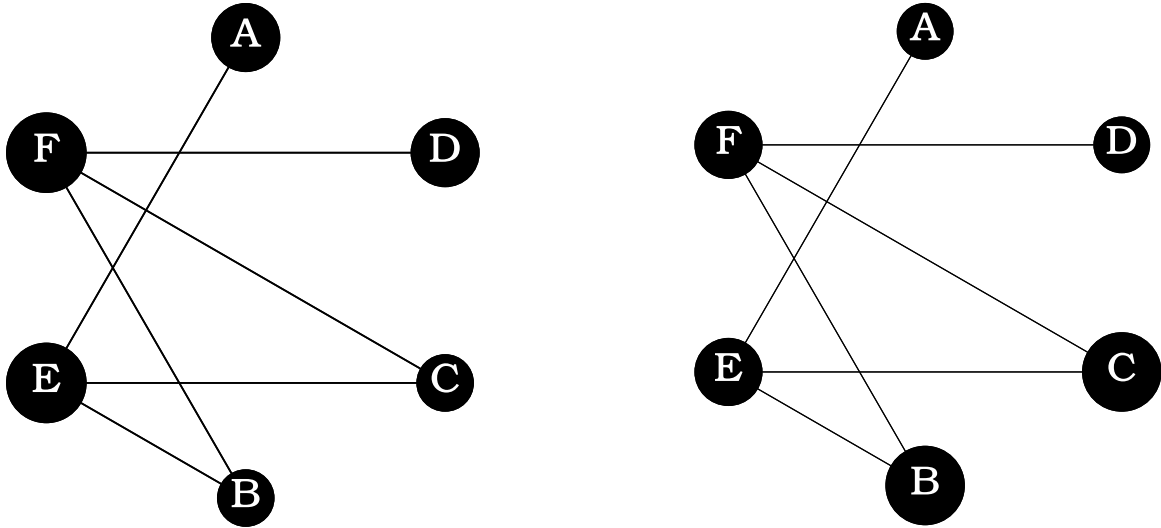
$$TI_A^n = \sum d_{j,A}^n \quad (6)$$

Ahol:

TI_A^n : A pont n lépésre számított topológiai fontossága.

$d_{j,A}^n$: A és j pont közötti n hosszúságú útvonalak ereje.

A TI^n -t a hálózat összes pontjára ki lehet számítani és ez alapján sorrendet lehet felállítani a nódusok között.



2. ábra. d (bal) és TI^2 (jobb) szemléltetése ugyanazon a példagráfon

A pontok átmérője arányos az adott nódusra kiszámolt d (direkt vagy közvetlen topológiai kölcsönhatás) és TI^2 (topológiai fontosság két lépésre) értékekkel. [?] alapján módosítva.

Az 2. ábrán látható példagráfra rendre felírhatóak a közvetlen kölcsönhatások (d) és a két lépésnyire közvetített indirekt kölcsönhatások (d^2) értékeit tartalmazó mátrixok:

	A	B	C	D	E	F
A	0	0	0	0	$\frac{1}{3}$	0
B	0	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$
C	0	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$
D	0	0	0	0	0	$\frac{1}{3}$
E	1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
F	0	$\frac{1}{2}$	$\frac{1}{2}$	1	0	0

d értékek

	A	B	C	D	E	F
A	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0
B	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
C	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
D	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	0	0
E	0	0	0	0	$\frac{2}{3}$	$\frac{1}{3}$
F	0	0	0	0	$\frac{1}{3}$	$\frac{2}{3}$

d^2 értékek

d^2 értékek

Ugyancsak mindkét mátrixra jellemző, hogy a sorok összege azt mutatja meg, hogy egy adott pont mennyire erős kölcsönható, tehát mekkora TI^n értéke. Például B pont két lépés távolságban összesen $\frac{4}{3}$ erővel hat, ez alapján erősebb kölcsönhatónak mondható mint az A pont a maga $\frac{2}{3}$ értékű összesített kimenő két lépés hosszú hatásaival. [?]

Az 2. ábrán az is jól megfigyelhető, hogy C pont a gyengébb közvetlen kölcsönhatók közé tartozik. Ugyanakkor mivel a C -ből eredő két lépéses útvonalak erős elsődleges kölcsönhatókon keresztül érik el végpontjaikat, ezáltal két lépés távolság viszonylatában már C is az erős kölcsönhatók közé tartozik.

Az $n > 1$ lépésszámú d^n értékeket tartalmazó mátrixokban már egy adott pont indirekt hatása önmagára is kiterjedhet. Páros számú lépések esetén viszont mindenképpen felírhatók olyan útvonalak melyeken a pont eléri önmagát, [?] vagyis $d_{X,X}^n \neq 0$ ha $n \in \{ 2k : k \in \mathbb{Z} \}$. Az 2. ábrán látszik, hogy például az F pont két lépés távolságban a következő útvonalakon hat önmagára: $F \rightarrow B \rightarrow F$, $F \rightarrow C \rightarrow F$ és $F \rightarrow D \rightarrow F$.

4. Célkitűzések

A diplomamunka célja

A diplomamunkám célja egy több adatbázisból integrált fehérje-fehérje kapcsolatokat tartalmazó humán-*Salmonella* gazda-patogén hálózat létrehozása különböző adatbázisok alapján és az elkészült háló topológiai elemzése.

Az elkészítendő hálózatnak a következőket kell tartalmaznia:

1. Kurált *H. sapiens* fehérje-fehérje kapcsolatok
2. Kurált *Salmonella* fehérje-fehérje kapcsolatok
3. *H. sapiens* és *Salmonella* közti prediktált fehérje-fehérje kapcsolatok

A topológiai elemzés során kapott adatok alapján véleményt szeretnék alkotni arról, hogy felhasználhatók-e az ökológiában fajok közti kapcsolatok vizsgálatára használt tisztán topológiai adatokon alapuló mérőszámok a molekuláris kapcsolati hálók elemzésére. Valamint, hogy az így előállított rangsorok mennyire korrelálnak a jelenleg használt *Salmonella* és humán belsejteket vizsgáló módszerek eredményeivel.

A célok eléréséhez tervezett feladatok

1. Program írása mely képes az ARN (Autophagy Regulatory Network) adatbázis kurált humán autofágia specifikus fehérje-fehérje kapcsolati rétegének („ARN core”) MiTab SQL formátumra átalakítására.
2. Program írása mely képes a Salmonet adatbázis kurált *Salmonella* fehérje-fehérje kapcsolati hálózatának MiTab SQL formátumra átalakítására.
3. Programok írása melyek képesek a Krishnadev és Skhirshagar féle humán-*Salmonella* fehérje-fehérje kapcsolati predikciók MiTab SQL formátumra alakítására.
4. Program írása mely képes a létrehozott MiTab SQL fájlokban a fehérjék azonosítójának *Uniprot* azonosítóra fordítására.
5. Program írása mely képes a már csak *Uniprot* azonosítókat tartalmazó adatbázisok összeállítására.

6. Program írása mely képes hálózatokban a megadott útvonalhosszra kiszámolni a topológiai fontosságot.
7. A kapott adatsorok értékelése, hálózatok ábrázolása, biológiai relevancia keresése.

5. Források és módszertan

5.1. Informatikai módszerek

A problémák megoldására használt programnyelvek

A teljes adatbázisok feldolgozására valamint az adatbázisokból származó adatok rendszerezésére és megfelelő formátumúra alakítására *Python* programnyelven írtam szkripteket. A diplomamunkám során a *Python* 2.7-en és 3.4-en futtatható szkripteket alkalmaztam. A fehérjék azonosítójának fordítását végző szkriptek egyike témavezetőm Kadlecsek Tamás *Javascript*-ben írt fordítószkriptjének kismértékű módosítása.

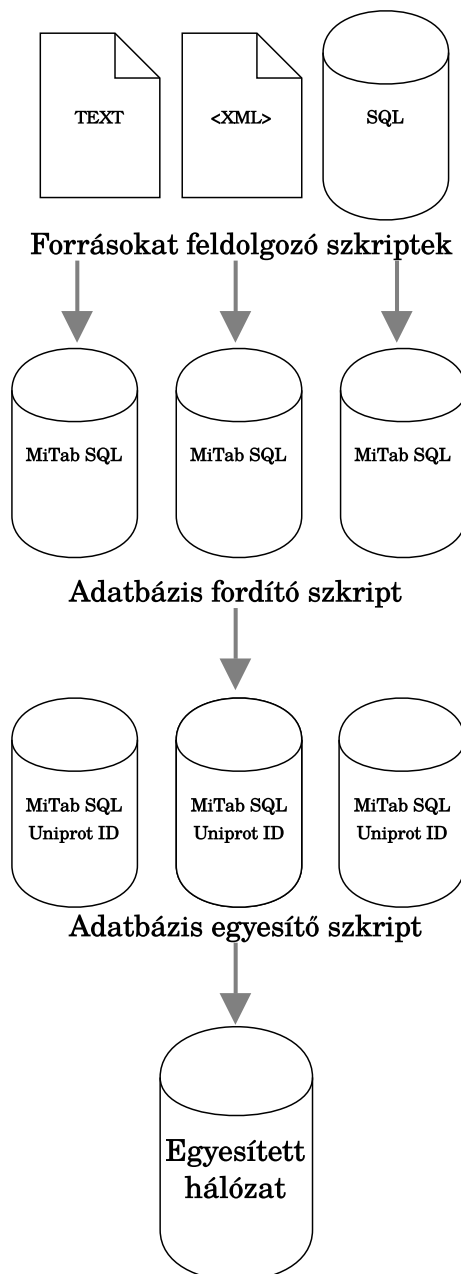
A diplomamunkám során alkalmazott szkriptek egy részét a *Signalink* 3 szignalizációs adatbázis kézzel gyűjtött és külső adatokat tartalmazó rétegeinek (L0 és L3) létrehozásakor készítettem. Mivel a *Signalink* rétegei is több adatbázisból integrál fehérje-fehérje interakciókat, így az ott alkalmazott munkafolyamat felhasználható volt a diplomamunkám gazda-patogén hálózatának létrehozásakor is. (3. ábra) A humán-*Salmonella* hálózat szerkezete azonban különbözik a *Signalink* 3-étól. A *Signalink* adatbázisának készítésekor kifejezetten szűrtük például az interspecifikus kapcsolatokat. A *Signalink*-kel szemben a humán-*Salmonella* háló prediktált éleket is tartalmaz. A két hálózat különbségei miatt, a diplomamunkámban az adatokat kezelő algoritmusok bár hasonlítanak a *Signalink*-et létrehozókra, de azokkal nem azonosak.

Az adatok tárolása

Az adatok ideiglenes tárolására, már a *Signalink* 3 készítése óta témavezetőm Kadlecsek Tamás javaslatára *SQLite* 3 adatbázis fájlokat alkalmazunk.

Az *SQLite* 3 egy nyílt forráskódú, C nyelven írt API-val rendelkező, beágyazott relációs adatbázis motor. Az SQL sztenderd szintaxisának nagy részét tartalmazza. Sok népszerű programnyelv rendelkezik már beépített *SQLite* támogatással, ilyen például a *Python* is. [?]

Az adatok ilyen módú tárolása lehetővé teszi azok gyors szűrését, kategorizálását és átalakítását SQL parancsok segítségével. Ilyen módon még azelőtt gyorsan információkat nyerhetünk nagy méretű hálózatokról, mielőtt azokat olyan jóval lassabb működésű hálózatkezelő programokkal elemezni kezdenénk mint például a *Cytoscape*.



3. ábra. A hálózat létrehozásának folyamata

A különböző forrásokból származó adatok esetén először a forrás formátumokat feldolgozni képes szkriptek átalakítják azokat a belső szabványként használt MiTab SQL formátumra. Általában a különböző adatforrások különféle azonosítókkal illetik a komponenseiket. Ahhoz, hogy több hálózatot egyesíteni tudjunk, szükség van arra, hogy egy adott biológiai entitás csak egyfajta azonosítóval szerepeljen. A fordító szkript MiTab SQL fájlból olyan MiTab SQL fájlt készít, amiben az elsődleges azonosító már a kívánt, esetemben *Uniprot* azonosító. Legvégül az adatbázis egyesítő szkript úgy „összefűzi” a különböző hálózatok pontjait és éleit, hogy ne legyen benne redundáns információ.

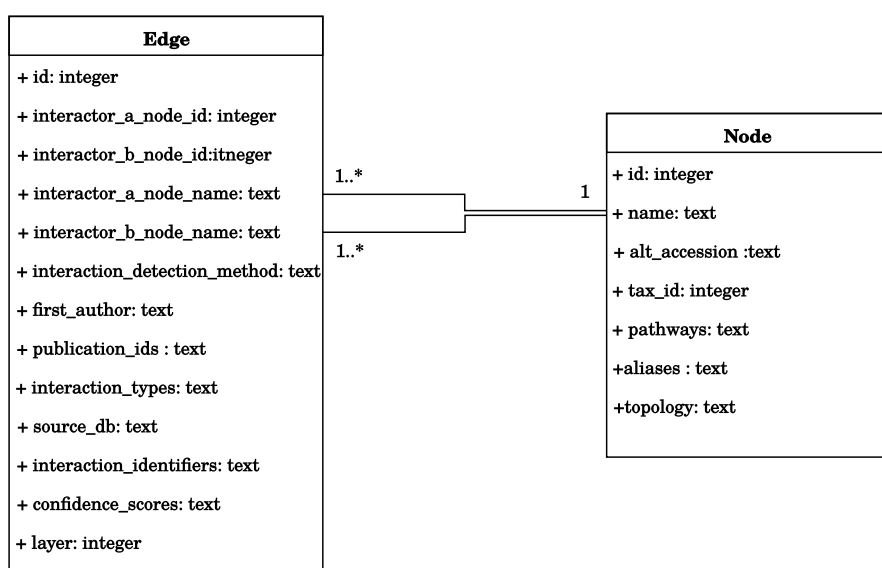
Az *SQLite* adatbázisfájlok másik előnye, hogy rendelkezésünkre áll az SQL nyelv. Mivel SQL parancsok segítségével gyorsan kezelhetők a feldolgozott adatok, így csak

ritkán van szükség adatmanipulálási célból egy újabb szkript írására. Amennyiben mégis szükséges újabb szkript írása, a legtöbb szkriptnyelv rendelkezik valamilyen *SQLite* adatbázis kezelési opcióval. Nagy méretű és mennyiségű biológiai adatot tároló *SQLite* fájlban a keresés is igen gyorsan megoldható az adatbázis beindexelésével, sőt még gyorsabb keresés is megvalósítható az indexelt táblák memóriába csatolásával. Az *SQLite* segítségével könnyen lehet importálni és exportálni a legtöbb népszerű adattárolási formátumba.

Ezen indokok miatt használjuk a csoportban az *SQLite* adatbázisokat az adatok köztes tárolására, a szöveges fájlok helyett.

Hálózatok tárolására a csoport által létrehozott MiTab SQL formátumot használtam. A MiTab SQL egy *SQLite* 3-ban tárolt a PSI-MI Tab formátummal közel megegyező adatstruktúra. A PSI-MI Tab egy *HUPO Proteomics Standards Initiative* (PSI) szervezet által meghatározott proteomikai adatok tárolására használt formátum. A PSI-MI Tab formátum specifikációja a szervezet honlapján elérhető. **TODO: hivatkozást elkérni**

A pontok és az élek külön táblában vannak letárolva az adatbázisban, így a PSI-MI Tab specifikáció pontra és az élre vonatkozó tulajdonságai a megfelelő táblába kerülnek. A MiTab SQL táblák oszlopai azonban nem teljesen egyeznek a PSI-MI Tab kategóriákkal. Ilyen különbség például, hogy a MiTab SQL nem használ néhány opcionális PSI-MI kategóriát viszont tartalmaz a PSI-MI-re nem jellemző tulajdonságokat is mint a topológia. Az éleket tartalmazó táblában a forrás (*interactor_a_node_name*) és a cél pont név oszlopa a *node* tábla azonosító oszlopának idegen kulcsai. (4. ábra)



4. ábra. A MiTab SQL sémája

Verziókövetés

Diplomamunkám készítése során a *Git* verziókövető rendszert használtam, melynek tartalmát a web-alapú *GitHub* tárhely szolgáltatásra töltöttem fel. A diplomamunkám *Git*-*Hub* tárhelyén (10. hivatkozás) a következő általam írt kódok érhetők el:

- A 3. ábrán ábrázolt munkafolyamatot lebonyolító szkriptek
- A fordításhoz használt adatbázist megépítő szkript (Kadlecsik Tamás szkriptje alapján)
- A fordítást végző szkriptek
- Az adatbázisokat összeajtó szkriptek
- A MiTab SQL formátumot kezelő osztály
- A topológiai elemzést végző szkript
- Az adatszűrésre használt SQL szkriptek

Tesztelés

A bonyolultabb algoritmusok esetén *egységteszteket* (*unit test*) alkalmaztam. A topológiai fontosságot kiszámító *Python* osztály összes metódusának működését ilyen módon ellenőriztem. A teszteléshez a *python* saját *unittest* nevű csomagját használtam.

5.2. Az források feldolgozásának eszközei

MiTab SQLite adatbázis API

A *PsimiSQL* egy *Python* 2.7-es szintaxisban írt osztály, melyet még a *Signalink 3* összeállításához készítettem, de azóta más projektekben is használtam és továbbfejlesztettem. A *PsimiSQL* segítségével a molekuláris biológiai hálózatok könnyen átalakíthatók MiTab SQLite adatbázisokká. Az osztály számos függvényével megkönnyíti a MiTab SQLite adatbázisok kezelését *Python* alól. Ilyen függvény például a redundáns adatok képzését gátló *insert_unique_node()* mely ellenőrzi, hogy az adott hálózatban szerepel-e már az importálni kívánt pont. Az osztály példányosításakor a memóriában létrejön egy példányhoz kötött MiTab SQL sémával rendelkező SQLite 3 adatbázis. Az adatbázis benépesítése és az adatok keresése tehát nagy sebességgel történik. Az osztálynak vannak függvényei melyekkel könnyen importálni és exportálni lehet MiTab SQL adatbázis fájlokat.

A szótárak építése és a fordítás

Ahhoz hogy a feldolgozott forrásokat össze lehessen fűzni egy nagy adatt szettbe, szükség van arra, hogy a hálózatokban ne szerepeljen ugyanaz a biológiai entitás más azonosítóval. Ennek érdekében a mindegyik hálózat fehérjéit a legfrisebb *Uniprot* adatbázis azonosítókra fordítottam. Ehhez két szkriptet kellett írnom.

A Salmonet és az ARN már eleve *Uniprot* azonosítókat használ. Azonban az *Uniprot* adatbázis állandó frissítései miatt, fenn áll a lehetőség, hogy nem egy időben készült fájlok ugyanarra a fehérjére más *Uniprot* azonosítót használnak. Egy másik hibaforrás az lehet, hogy a *Uniprot* adatbázis egy fehérjét több azonosítóval is tárol. Előfordulhat, hogy egy fehérje többször is szerepel csak más *Uniprot* azonosítókkal. Amikor egy fehérjét beletesznek a *Uniprot* adatbázisba, akkor kap egy elsődleges azonosítót. Primer azonosítót kapnak még olyan fehérjék is, melyek már benne voltak az adatbázisban de később külön izoformákra lettek szétválasztva. Új elsődleges azonosítót kapnak olyan fehérjék is melyeket több vélt fehérjéből egyesítettek. Minden ilyen művelet után, a legfrissebb elsődleges azonosító marad az új elsődleges, az összes többi pedig másodlagos azonosítók lesznek. A *Uniprot* azonosítókat még csoportosítani lehet az alapján is, hogy a fehérje manuálisan vagy automatikusan lett annotálva. Az első típusba az úgynevezett *Swissprot* az utóbbiba pedig a *trEMBL* azonosítók tartoznak. A szkriptem az összes pont azonosítójára, ha az nem *Swissprot*, kikeresi a *swissprot* azonosítót ha létezik, vagy az elsődleges *trEMBL* azonosítót. Az ARN és a Salmonet fordításához szükség volt egy *Salmonella-Salmonella* és egy humán-humán szótárra, amiket a Kadlecsek Tamás féle szótárépítő scripttel állítottam elő. Az azonosítókat a szótárak alapján saját készítésű szkripttel fordítottam.

A predikciókhoz egy olyan szótárat kellett létrehozni, mely *Salmonella* génazonosítókhoz rendel *Salmonella Uniprot* azonosítókat. Több okból is lehetséges ebben az esetben a génről fehérjére fordítás. A *Salmonella* adatbázisok általában csak génazonosítókat tartalmaznak, sőt vannak publikált fehérje-fehérje kapcsolatokat leíró szövegfájlok is melyek csak génneveket tartalmaznak. Másik fő ok, az hogy *Salmonella*-ra még nincs megfelelő mennyiségű és felbontású adat. Valamint prokariótákban nincs alternatív splicing, így egy génről készülő fehérjék is jobban megfeleltethetőek. Ezt szintén Kadlecsek Tamás szkriptjével állítottam elő. Egy általam írt másik fordítószkript segítségével pedig az előzőhöz hasonló módon fordítottam a predikciókat.

5.3. A források

5.3.1. Autophagy Regulatory Network (ARN)

Az ARN egy széles terjedelmű autofágia adatbázis. Az adatbázis az irodalomból kézi gyűjtéssel kapott élek mellett tartalmaz még 19 más adatbázisból importált valamint 4 féle predikcióval készült feltételezett kapcsolatokat is. Az ARN-ben található 1485 darab fehérje között 4013 kapcsolat van. Az adatbázis komponensei között vannak az autofágia mechanizmusában szerepet játszó fehérjék és ezek regulátorai valamint transzkripciós faktorai. Az adatbázisban 413 transzkripciós faktor valamint 386 olyan miRNS melyek képesek lehetnek autofágia komponensek szabályzására. [?]

Az ARN hat rétegből épül fel:

1. Autofágia fehérjék.
2. Az első réteg fehérjeinek autofágia specifikus forrásokból származó regulátorai.
3. Olyan poszt-transzlációs regulátorok melyek közvetlenül hatnak az első két réteg fehérjéire.
4. Az első három réteg transzkripciós szabályzói.
5. Az első négy réteg poszt-transzlációs regulátorai.
6. Olyan jelátviteli útvonalak és fehérje-fehérje interakciók melyek különböző útvonalakat az autofágia szabályzókhöz kötnek.

[?]

5.3.2. Az ARN feldolgozása

A gazda patogén hálózat összeállításához az ARN adatbázisnak csak az első, autofágia fehérjéket tartalmazó rétegét használtam fel. A hálózat letöltését követően azt egy általam írt *Python* scripttel MiTab SQL formátumba alakítottam. A fordító szkripttel az akkor legfrissebb *Uniprot* adatbázis azonosítókra fordítottam. Mindezekre azért volt szükség, mert az *ARN* létrehozásakor még nem használtuk a MiTab SQL formátumot.

5.3.3. Salmonet

A *Salmonet* a csoportunk által készített de még nem publikált molekuláris hálózat. A hálózat adatai kézi adatgyűjtésből, nagy áteresztőképességű módszerekből valamint predikciókból származnak. Az adatokból integrált hálózat transzkripcionális szabályzási, metabolikus és fehérje-fehérje kölcsönhatási szinteket tartalmaz. A *Salmonet* összesen öt-öt gasztrointesztinális és extraintesztinális *Salmonella* törzs hálózatából lett egyesítve. Egy törzs hálózata a metabolikus, szabályzási és fehérje-fehérje alhálózatok egyesítéséből készült. A *Salmonet* ezen tíz törzs hálózatainak összeállításával jött létre. [?]

A metabolikus hálózatok

Ha egy metabolit egy bizonyos reakció képződménye és egy másik szubsztrátja, akkor a két reakciót katalizáló fehérje kapcsoltnak tekinthető. A reakciók adatai két forrásból származnak. Az adatok egy része a *Salmonella Typhimurium LT2* kézi gyűjtésű metabolikus modelljéből származik, a másik része pedig a *BioModels* adatbázisból származó predikciókból. [?]

A szabályzási hálózatok

Szabályzási kapcsolatnak tekintjük a transzkripció faktor promóterhez való kötődését. Az ilyen típusú kapcsolatok is a szakirodalomból kézzel gyűjtött kísérletesen bizonyított valamint nagy áteresztőképességű kísérletek adatszettjei alapján lettek összeállítva. [?]

A fehérje-fehérje kapcsolati hálók

A fehérje-fehérje kapcsolatok is több forrásból kerültek a *Salmonet*be. Kézi gyűjtéshez a csoportunk által a *Signalink* adatbázis létrehozásához használt módszert alkalmaztuk. A kézi gyűjtés mellett kísérletesen igazolt kapcsolatokat az *IntAct* adatbázisból is integráltuk. Az előző két módszer mellett ez a réteg is tartalmaz predikciókból származó kapcsolatokat. [?]

5.3.4. A Salmonet feldolgozása

Az *Salmonet* átalakítása az ARN-hez hasonló módon történt, azzal a különbséggel, hogy fordításkor *Salmonella-Salmonella* szótárat használtam.

5.3.5. A predikciók

Az eddig ismertetett források csak fajon belüli kapcsolatokról felépülő hálózatokat tartalmaztak. Ahhoz, hogy szakdolgozatomban tudjam tanulmányozni a humán-*Salmonella* kapcsolatot szükségem van még interspecifikus élekre is. A predikciós forrásokból származó interspecifikus kapcsolatok fogják összekapcsolni a gazda hálózatát a patogénével. Szakdolgozatomban [?] és [?] humán-*Salmonella* predikcióit használtam.

5.3.6. A predikciók feldolgozása

A két predikció feldolgozására külön *Python* szkripteket írtam. Csakúgy mint az előző forrásokat, az így elkészült adatbázisokat a legújabb *Uniprot* azonosítóra fordítottam.

5.3.7. Az adatbázisok egyesítése

Az adatbázisok egyesítésekor a fő szempont az, hogy az végleges hálózatban ne legyenek redundáns pontok vagy élek. Az adatbázis egyesítő szkript beolvassa az összes adatbázisfájlt és egy *has-map*-ben eltárolja a pontokat és éleket és ezek tulajdonságait. A *hasm-map*-ből sokkal gyorsabban ki kereshető, hogy egy adott él vagy pont benne van-e már, mintha SQL lekérdezésekkel oldanám meg. Amíg a szkript végigmegy az összes adatbázisfájlon a *has-map*-ek tartalmából létrehoz egy MiTab SQL fájlt mely nem redundánsan tartalmazza az összes forrás adatbázis tartalmát.

Az egyesített adatbázis a két kézi gyűjtésű hálózat pontjait és kapcsolatait valamint a predikciókból csak az éleket tartalmazza. A predikciós élek közül is csak azokat tartottam meg, melynek mindkét nódusa megtalálható valamelyik kézi gyűjtésű forrásban. Tehát a predikciókból új pont nem került bele a hálózatba.

Ezt az adatbázist már újabb *Python* szkript írása nélkül csupán SQL lekérdezésekkel létre lehet hozni.

5.4. A hálózatok topológiai elemzése

5.4.1. A TopologyAnalyser osztály

A *TopologyAnalyser* osztály *Python 3* szintaxist használ. Az osztály segítségével kiszámítható a Jordán Ferenc féle topológiai fontosság (TI^n). Az *TopologyAnalyser* egyetlen külső

függősége a NetworkX csomag. A NetworkX egy *Python*-ban írt, komplex hálózatok létrehozására, manipulálására és elemzésére használható ingyenes csomag.

A NetworkX egy teljesen *Python*-ban írt csomag. Vannak még más ingyenes hálózatkezelő *Python* modulok, mint például a *graph-tool* és az *Igraph*. Az utóbbi kettő csomag már részben *C/C++*-ban íródott, így kisebb futásidőre képesek. Választásom azért esett a kisebb teljesítményű *NetworkX*-re mert így felhasználhattam régebben megírt szkriptjeimet. A *Signalink* adatbázisának építésekor ugyanis szükségem volt például XGMML fájlok gyors feldolgozására, és erre a célra a *NetworkX* egy alosztályt használtam.

Az osztály konstruktorának egyetlen paramétere egy éllista. Példányosítás után a *TopologyAnalyser* típusú objektum, egy éllistát, egy *Graph* típusú objektumot és egy egység-élerősségi mátrixot tartalmaz.

A *Graph* osztály a NetworkX csomag része és irányítatlan gráfok tárolására alkalmas. A hálózat példányhoz kötött *Graph* típusú objektum (*self.graph*) tárolása előnyös, mert így elérhetők a NetworkX csomag különböző gráf elemzésre használható metódusai. Ilyen függvény például a *Graph.neighbors_iter(node)* mely egy adott pont szomszédjainak iterálható listáját adja vissza.

Az élerősség mátrix (*self.edgeStrength*) egy példányhoz kötött *Python Dictionary*. A *Dictionary* osztály *Python*-ban *hashmap* adatstruktúrával van implementálva, ezáltal a benne tárolt adatok gyorsan elérhetők. A szótár kulcsai a mátrix indexei, az értékei pedig az élerősségek *Fraction* típusú objektumokként letárolva. Az egységerő mátrix celláinak indexei maguk az élt alkotó pontok *Uniprot* azonosítói. Az élerősség mátrix 0 értékkel rendelkező cellái nincsenek letárolva a *hasmap*-ben.

Az osztály fontosabb metódusai és működésük

- *getNthNeighbors()*

A függvény rekurzívan kikeresi az adott pont *n*. szomszédját.

- *pathFinder()*

Ez a függvény megkeresi az összes n hosszúságú útvonalat egy forrás és egy cél nódus között.

```
def pathfinder(mélység, forrás_nódus, cél_nódus, útvonal, összegyűjtött útvonalak):
    if útvonalhossz == 0 then
        forrás_nódus hozzáadása az útvonalhoz ;
        mélység -= 1
    else if mélység == 0 és forrás_nódus == cél_nódus then
        return útvonal
    else if mélység == 0 és forrás_nódus ≠ cél_nódus then
        return
    for forrás_nódus első szomszédai do
        szomszéd_nódus hozzáadása az útvonalhoz;
        pathfinder( mélység - 1, szomszéd_nódus, cél_nódus, útvonal, összegyűjtött útvonalak) ;
        az utolsó pont eltávolítása az útvonalból;
```

- *buildEdgeUnitStrengthDict()*

Ez a függvény egyszer hívódik meg a konstruktor lefutása során. A metódus imperatív módon egy **for** ciklussal végig iterál a kapott éllistán. A függvény visszatérési értéke egy egységerő mátrix.

- *getNthPathwaysForNodes()*

A függvény először minden pontnak megkeresi az n . szomszédját. Mint az irodalmi bevezetőben ismertettem, a topológiai fontosság számolásánál páros lépésszámnál hurkon keresztül egy adott pont mindig önmagának a szomszédja is. A metódus két egymásba ágyazott **for** ciklust használ. A külső ciklus minden ponra előállítja a pont n . szomszédjainak halmazát. A belső ciklus pedig minden cél és forrás pont párra meghívja a *pathFinder()* függvényt, majd a kapott útvonalakat imperatív módon hozzáadja egy listához, amit még a ciklusokon kívül definiáltam. A metódus visszatérési értéke egy lista amelyben minden lehetséges forrás és célpont közötti n hosszúságú útvonal benne van.

- *getEdgesFromPathway()*

A metódus egy útvonalat fogad és az azt alkotó élek listájával tér vissza.

- *countPathwayStrength()*

A metódus egy útvonalat kap, amire meghívja a *getEdgesFromPathway* függvényt. A kapott éllistán végigmegy és minden élre kikeresi az egységerő mátrixból a megfelelő értéket. Az összes él erejének összeszorozása után az út erősségével tér vissza.

- *getPathWaysToNthNeighbours()*

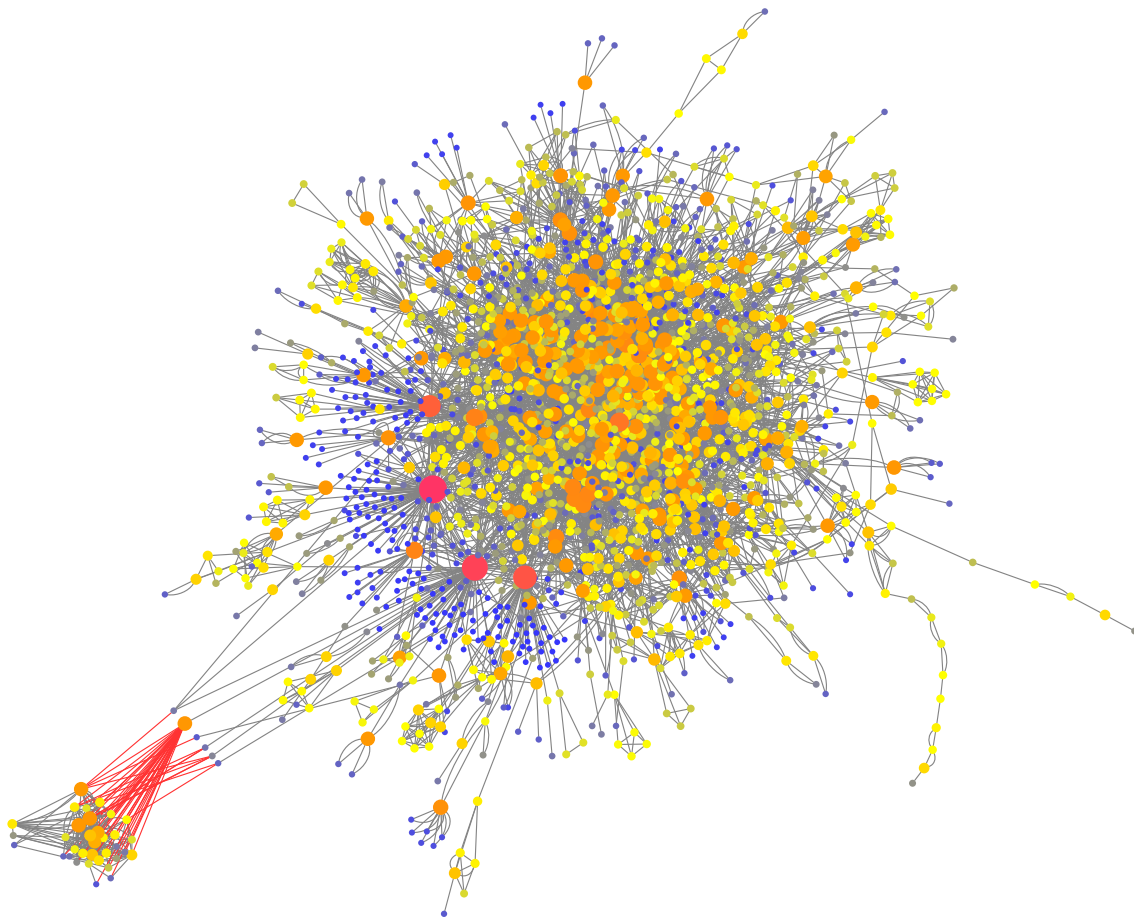
Ez a függvény kikeresi egy adott pont összes *n.* szomszédját és egy listában letárolja. Amikor elkészült a szomszédok listájával, akkor ezen egy **for** ciklus segítségével végigmegy. Ekkor minden periódusban meghívja a forrás és az aktuális cél nódusra a *pathfinder* függvényt, majd a kapott útvonalakat imperatívén hozzáfűzi egy a cikluson kívül definiált listához. A ciklus után a függvény visszatér az elkészül listával.

- *countTI()*

A függvény egy adott pontra meghatározza annak topológiai fontosságát. Első lépésként létrehoz egy útvonal listát a *getPathWaysToNthNeighbours* metódus segítségével, majd végigmegy az útvonalak listáján egy **for** ciklusban. Minden periódusban meghívja a *countPathwayStrength(pathway)* függvényt az aktuális útvonalra, és annak eredményét egy a cikluson kívül definiált változóba akkumulálja.

6. Eredmények

6.1. Topológiai indexek az integrált hálózatban



5. ábra. Az egyesített hálózat

A hálózaton piros élek jelölik az *ARN*-t (*baloldalt alul*) és a *Salmonet*-et elválasztó prediktált éleket. Az ábrán a színátmenetek és méretek arányosak a pontokra kiszámolt TI^3 -al. A topológiaiailag fontos pontok kiemelése érdekében három csoportot állítottam fel: Az első csoportba azok a pontok tartoznak, melyekből távozó hatás kisebb mint a beérkező ($TI^3 \leq 1$). A hálózatban 1674 ilyen nódus található, ami a háló 68,1%-át teszi ki. Az utóbbi pontok színe a sötétkéktől a citromsárgáig terjed. A középső citromsárgától a narancsig ábrázolt pontok azok melyeknek már van nettó kimenő hatásuk ($1 < TI^3 < 2$). Ebbe a halmazba 624 pont azaz a hálózat 25,4%-a tartozik. A harmadik narancstól vörösig terjedő csoportot azok a nódusok alkotják

melyeknek a legnagyobb a topológiai fontosságuk. Ez utóbbi 161 pont a hálózat 6,5%-át teszi ki.

ide mindenképpen kell egy legend, a pont mérethez és a színhez is, a színhez lehetne egy táblázat ami ezeket a

1. táblázat. Az egyesített hálózat legnagyobb topológiai fontosságú pontjai különböző lépésszámoknál

Azonosító	TI ¹	Azonosító	TI ²	Azonosító	TI ³
phoP	64,38	trxA	16,61	phoP	53,53
ssrB	59,17	phoP	10,31	ssrB	46,15
rpoN	46,21	yajL	9,95	rpoN	37,36
hilD	39,79	hilD	8,71	hilD	31,07
trxA	23,77	ssrB	7,93	trxA	20,10

~~Elemzéseimet rossz a sorrend, először legyen egy kis bevezetés ami kb az első bekezdés és utána először~~

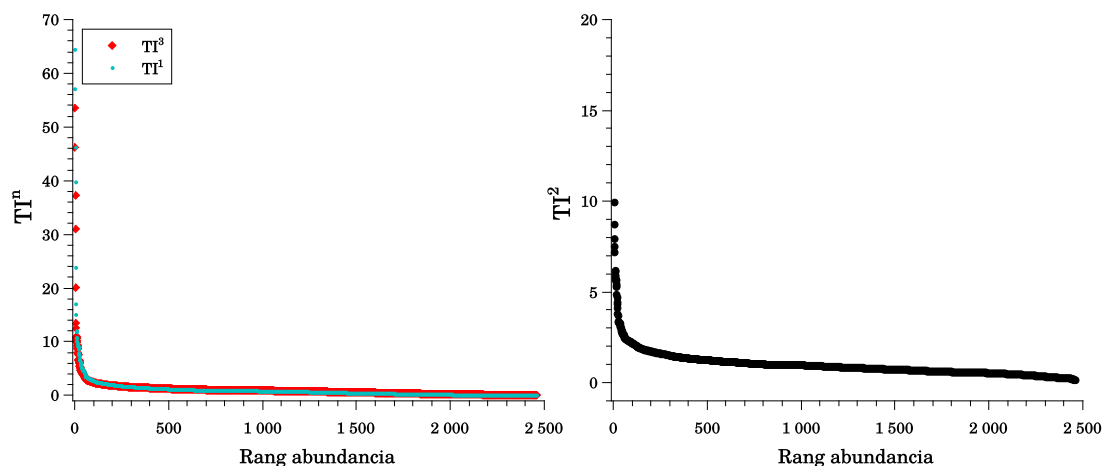
~~Az elemzéseimet~~ a *Salmonet*, az *ARN* és az egyesített hálózatra ~~végeztem el is el végeztem.~~

A topológiai fontosságot (TI) mind a három hálóra kiszámoltam 1, 2 és 3 lépés távolságra.

A többfajta lépéshosszra kiszámolt indexek információt szolgáltatnak arról, hogy egy adott pont topológiai fontossága a topológiai fontosság ugyanaz mint a TI? hogyan változik egy hálózaton belül. Fontos információt nyújt az is, hogy egy pont topológiai fontossága hogyan változik azzal, hogy az eredeti vagy az egyesített hálózatban vizsgáljuk.

A feldolgozott *ARN* hálózat 34 pontot tartalmaz amik között 186 él van. A *Salmonet*-ben pedig 2425 pont van amik között 7973 él húzódik. A predikciók az *ARN*-ből és *Salmonet*-ből származó 2459 nódus közé még 40 kapcsolatot adnak. A végleges hálózat így 2459 pontot és 8199 élt tartalmaz. (5. ábra) erről inkább egy táblázat kellene

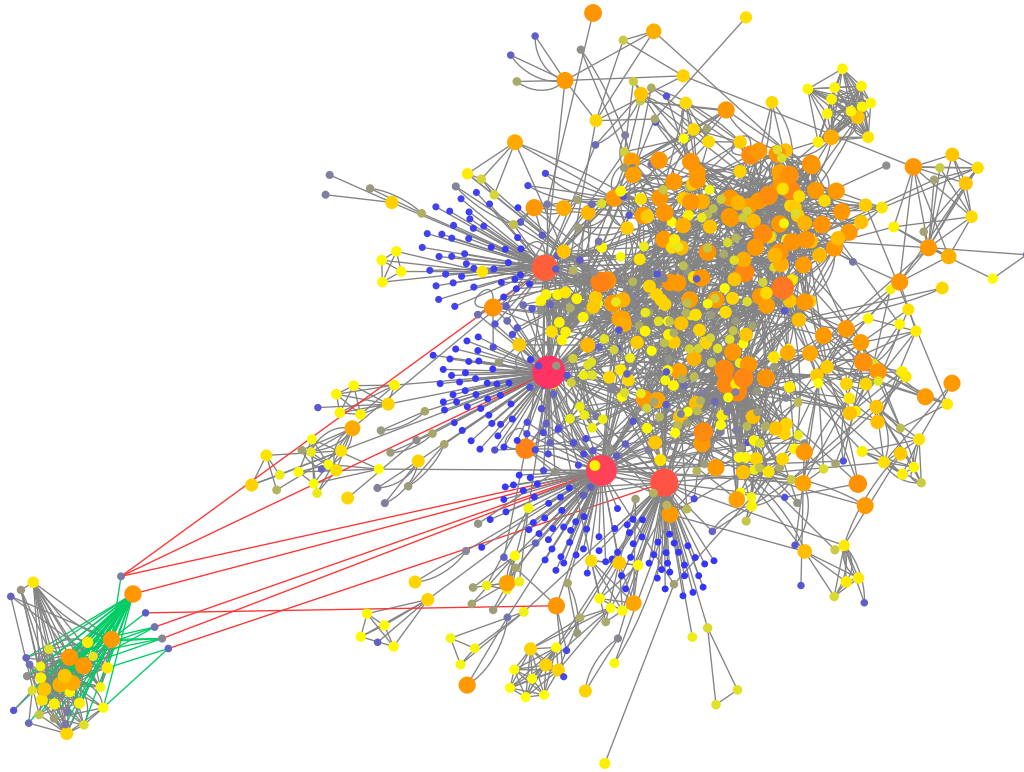
Az integrált hálózatban a *Salmonet* tulajdonságai érvényesülnek. Az első öt nódus mind a három topológiai fontosság kategóriájában *Salmonet* eredetű. (1. táblázat) A hálózatban a legerősebb hatással rendelkező humán fehérje a GABARAP, ami TI¹ (2,54) és TI³ (2,12) esetén a 88., TI²-nél (2,61) pedig a 112. legerősebb hatású pont.



6. ábra. Az egyesített hálózat rang abundancia görbéi

A hálózat fontossági rangsorára felrajzolhatók az ökológiában alkalmazott rang abundancia diagramok. (6. ábra) A rang abundancia diagramok értelmezési tartománya a pontok rangsorszámainak halmaza, az értékkészlete pedig valamilyen pontra vonatkoztatott mérőszámok. Ebben az esetben az y értékek különböző lépéshosszhoz számított topológiai fontosságok.

A 6. ábrán megfigyelhető, hogy a hálózatban csupán néhány kiugróan magas topológiai fontosságú pont van. Az is jól látszik, hogy a visszahatást figyelmen kívül hagyó TI^1 és a csak közvetett útvonalakon végbemenő visszahatást figyelő TI^3 közel azonos lefutású. TI^2 -nél minden pont eléri önmagát az összes első szomszédján keresztül. Egy pont hatása a cél pontja fokszámával fordítottan arányos. A megnövekedett bemenetszámok miatt a TI^2 értékek is kisebbek, viszont jól szemléltetik azt, hogy az adott pontok mennyire hatnak vissza önmagukra. [itt nem lenne jobb egy log skála? eléggé indokolt hogy a TI2 külön diagram](#)

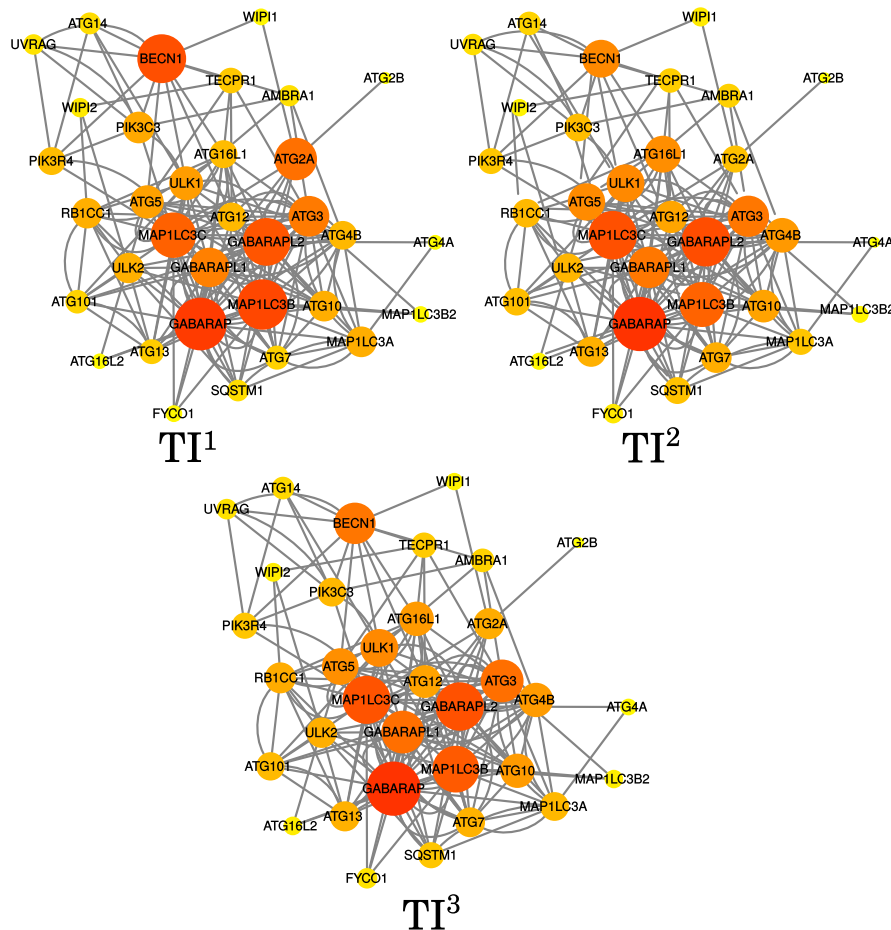


7. ábra. A három lépésre elérhető pontok

Az ábra az egyesített hálózat azon részét tartalmazza mely elérhető három lépésre a *Salmonella* hat darab humán fehérjével kapcsolatot létesítő fehérjétől ezeket hívhatjuk interfész fehérjéknek. A zöld színű élek a predikciókból származnak. A piros színű élek ez esetben a *Salmonella* interfész fehérjéi és ezek *Salmonellán* belüli első szomszédai között húzódnak. Látványos, hogy ez utóbbi élekből több is a nagy topológiai fontosságú pontokhoz tart. A pontok színezése megegyezik az 5. ábránál ismertetett sémával. ide is kell legend, bocsi de ebben tuti belekötnek, akármilyen szép az ábra

6.2. Topológiai indexek az alhálózatokban

6.2.1. ARN



8. ábra. Az *ARN* 1, 2 és 3 lépésre számított topológiai fontosságai

A pontok színmélysége citromsárgától narancsig, valamint átmérője egyenesen arányos az nódus TI értékével. Az *ARN*-en belül a innen mintha lemaradt volna valami legend, legend mindenhova, a legendnek is legyen legendje :D

2. táblázat. Az *ARN* legnagyobb topológiai fontosságú pontjai különböző lépésszámoknál

Azonosító	TI ¹	Azonosító	TI ²	Azonosító	TI ³
GABARAP	2,41	GABARAP	2,21	GABARAP	2,33
MAP1LC3B	2,27	GABARAPL2	1,96	GABARAPL2	2,00
BECN1	2,17	MAP1LC3C	1,93	MAP1LC3C	1,99
MAP1LC3C	2,11	MAP1LC3B	1,75	MAP1LC3B	1,68
ATG2A	1,99	ATG3	1,56	GABARAPL1	1,68

ezekkel a táblákkal kell valamit csinálni, mert itt egy sor nem egy sor, az előzővel is ugyan ez a gond

Az *ARN* az egyesített hálózatnál jóval kompaktabb ide esetleg lehetne egy denzitást írni (össze kapc a potok gazdagon kapcsolnak egymással. Az erősebb belső kapcsoltságot az is jól jellemzi, hogy ebben a hálózatban nagyobb az olyan nódusok aránya melyeknek nagy a topológiai fontossága. Az egyesített hálózatban csupán a pontok 31,9%-a rendelkezik egynél nagyobb TI^3 értékkel, ugyanakkor ez arány az *ARN* esetén már 44,1%.

Az 2. táblázaton ábrázoltam a TI^1 , TI^2 és TI^3 kategória első öt legfontosabb fehérjét. A táblázatból látható, hogy egy adott fehérje különböző lépéshosszoknál számolt topológiai fontossága nagyjából hasonló értékeket vesz fel. A TI^2 adatok mindenképp magukba foglalják azt, hogy a pontok hogyan hatnak vissza önmagukra. A TI^3 -as értékek is tartalmazhatnak olyan útvonalakat melyeken egy pont önmagára visszahat, de csak akkor ha a két nódusnak van közös első szomszédja. A TI^3 tehát magába foglalja az adott pont hatását önmagára de csak közvetett módon. TI^3 -nál a pontból származó egységnyi erő nem oszlik több részre, mert a pont nem éri el magát még páros lépésszámú útvonalakon is.

Az *ARN* legnagyobb TI értékű fehérjéi

A következő felsorolásba összegyűjtöttem azokat a fehérjéket amelyek mind a három topológiai fontosság kategóriában megtalálhatók a tíz legfontosabb fehérje halmazában. A következő nyolc fehérje tehát a közvetlen topológiai hatást figyelembe vevő TI^1 , az önhatást hangsúlyozó TI^2 és távolabbi szomszédságot is elérő közvetett hatást hangsúlyozó TI^3 listáján is magas pontszámmal végzett értéke is a legmagasabbak:

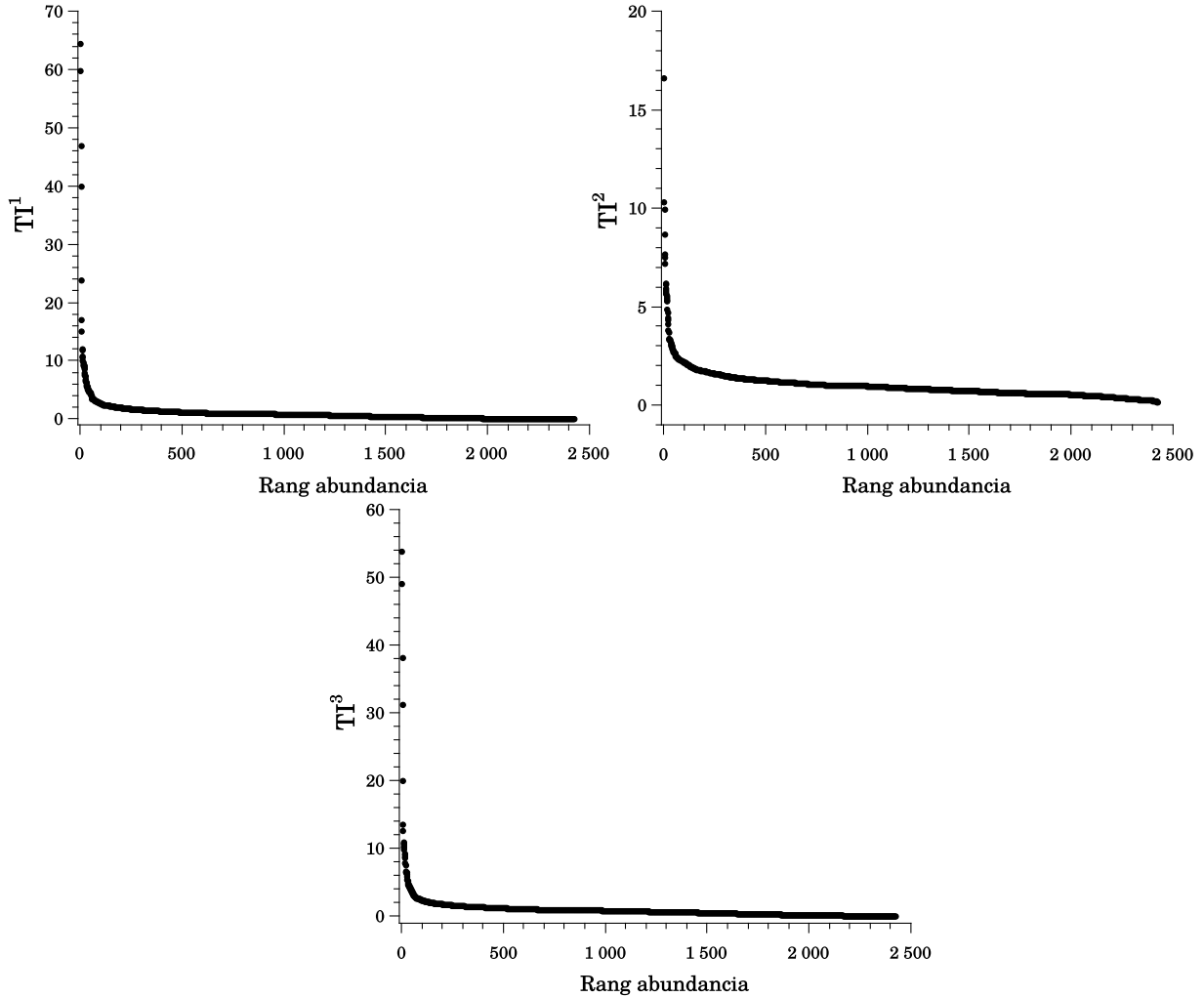
- **ATG3** – Más néven *ubiquitine-like-conjugating enzyme ATG3*, egy konjugáló enzim, mely az autofágiában és a mitokondrium homeosztázisban tölt be szerepet. Autofágia során ez az enzim katalizálja a foszfatidiletanolamin konjugációját az ATG8 szerű fehérjék C-terminális glicinjére. Ilyen ATG8 szerű fehérje a listán szereplő GABRAP, GABARAPL1, GABARAPL2 valamint a listán nem szereplő MAP1LC3A is. [?] milyen lista?
- **BECN1** – Más néven *Beclin-1*. Az autofágia egyik központi-kulcs fehérjéje. Bonyolult jelátviteli útvonalak hatására a VPS34-el együtt a fagofór kialakulását szabályozza.

A *Beclin-1* egy PI3K komplex alegysége, VPS34-el való kapcsolódása fokozza annak katalitikus hatását. [?]

- **GABARAP, GABARAPL1 és a GABARAPL2** – A GABARAP a nevét onnan kapta, hogy neuronokban a GABA_A receptorokhoz kapcsolódik. A GABARAP egy 14kDa-os citoplazmatikus molekuláris ~~chaperone~~chaperone (erre mi a jó magyar szó? a chaperone jó, csak legyen dőlt). Ma már tudjuk, hogy a fehérje ennél sokkal szélesebb körű kapcsolatrendszerrel rendelkezik. Autofágiában a fagofór formálódásánál csakúgy mint az LC3-ak, a GABARAP-ok is liopidálódnak. Amikor ~~a sejt erős autofágiás állapotban van~~ egy setben az autofágia felpörög akkor a GABARAP-ek foszfatidiletanolaminra konjugálódnak, ezzel ellentétes helyzetben viszont konstitutív proteolitikus bomlásra vannak ítélve. [?]
- **MAP1LC3B** – Ez a fehérje a MAP-ok családjába tartozik és ~~humánban~~emberben olyat nem írunk soha hogy humánban két majdnem identikus kódoló régióval rendelkező génről íródik át ez azt jelenti hogy két izoformája van? akkor azt írjuk. Első izoláláskor a MAP-1 alegységének vélték és ezért kapta LC3 (*light chain 3*) nevet. Az LC3 és a MAP1LC3B név tehát ugyanarra a fehérjére vonatkozik. Ma már tudjuk, hogy a mikrotubulusok mellet, az autofágiában is kulcsszerepet játszik. [?] Az autofágiában két ubikvitin rendszer szerű enzimkomplex dolgozik. Az LC3 autofágia indukciójakor proteolitikus hasítást szenved a citoplazmában és így jön létre az LC3B-I-es izoforma. Az LC3B-I-et az Atg7 egy E1-szerű enzim aktiválja, majd az E2 szerű ATG3-ra kerül. Ezután LC3B-I C-terminálisán helyezkedő glicinje foszfatidiletanolaminnal konjugálódik. Ez utóbbi formát nevezzük LC3B-II-nek. Az LC3B-II a növekvő fagofór külső és belső felszínén is megtalálható. Jelenlegi ismereteink alapján az LC3B-II két fő szerepei a membránok hemifúziója és a lebontásra irányított fehérjék szelekciója az autofagoszómához. [?]
- **MAP1LC3C** – Ez a fehérje egy LCH3 paralóg, az LC3C génről íródik át melynek szabályozása független az LC3-tól. Az LC3-hoz hasonlóan ez a fehérje is a makroautofágia szelektivitását növeli. [?]
- **ULK1** – Az ~~Ulk-1~~ULK-1 egy szerin-threonin-protein-kináz, az ATG1 emlős homológja. Az Ulk-1 a fagofór kéződés egyik kulcs faktora. Aktivitását az az mTORC1

szabályozza. Az mTORC1 egyik legfontosabb feladata emlős sejtekben az energiaszint érzékelés. Az mTORC1 magas tápanyagtartalom vagy aktiváló jelátvitel hatására foszforillálja az Ulk-1 szubsztrátját, így az Ulk-1 nem tudja beteljesíteni a fagofór kéződésében játszott szerepét és az autofágia gátlódik. [?]

6.2.2. Salmonet



9. ábra. A *Salmonet* pontjainak rang abundancia diagramjai

A képen a *Salmonet* pontjai a különböző topológiai fontossági értékek szerint sorba vannak rendezve.

A *Salmonet* 2425 nódust és 7973 élt tartalmaz. A *Salmonet* jobban hasonlít az egyesített hálózatra. Az *ARN*-el ellentétben de az egyesített hálózathoz hasonlóan, jóval magasabbak a legnagyobb topológiai fontosságú pontok értékei. Sőt a *Salmonet* különböző

lépésszámokhoz rendelt öt legfontosabb fehérjének listája megegyezik az egyesített hálózatával.

(3. és 1. táblázat)

3. táblázat. A *Salmonet* legnagyobb topológiai fontosságú pontjai különböző lépésszámoknál

Azonosító	TI ¹	Azonosító	TI ²	Azonosító	TI ³
phoP	64,46	trxA	16,61	phoP	53,84
ssrB	59,77	phoP	10,30	ssrB	49,06
rpoN	46,89	yajL	9,95	rpoN	38,11
hilD	39,87	hilD	8,70	hilD	31,20
trxA	23,77	ssrB	7,70	trxA	20,11

A *Salmonet*-ben négy olyan fehérje van mely mind a három topológiai index szempontjából a legerősebb tíz helyezettben van:

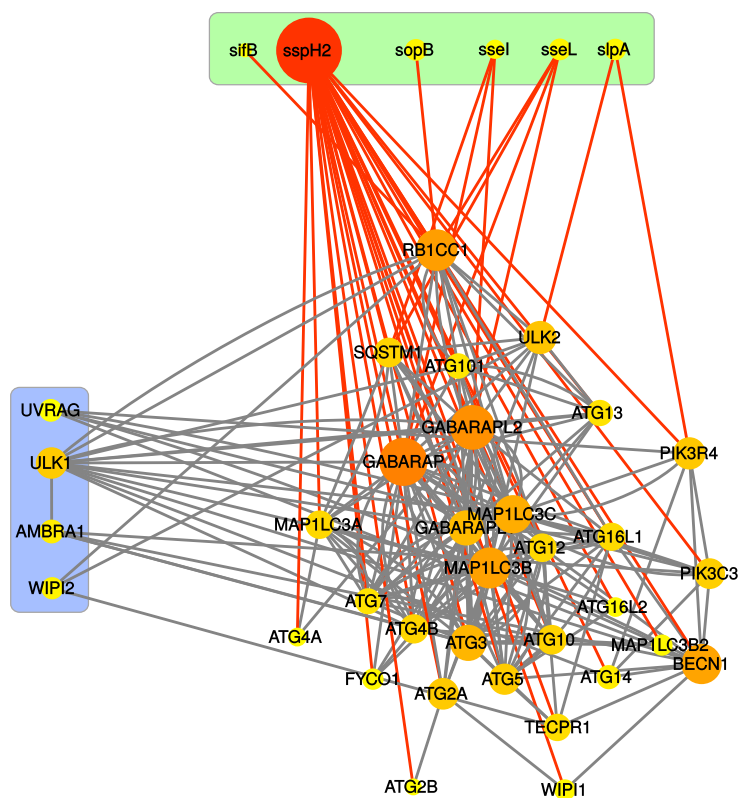
- **hilD** – *Salmonella typhimurium*ban a patogenitási sziget 1 (SPI-1) kódolja a T3SS fehérjéit amik az invázióhoz szükségesek. Az SPI-1 génjei többnyire a szigeten kívüli transzkripció faktorok által szabályoznak. A sziget mester ~~kapcsolója a hilA~~ transzkripciós faktor transzkripciós faktora a hilA, melyet két másik fehérje mellett a hilD szabályoz. A hilD-nek eddig 17 kötési helyét írták le a *Salmonella* genomban. [?]
- **phoP** – A T3SS komponenseit kódoló patogenitási sziget mellett egy másik fontos virulencia rendszer is szerepet játszik a *Salmonella* fertőzés sikerességében. A phoP-phoQ egy kétkomponensű rendszer mely intracelluláris szignálokra aktiválódik, és nélkülözhetetlen a baktériumsejtek túléléséhez és replikációjához. A phoQ egy kináz mely a megfelelő jelátvitel hatására foszforilálja a phoP-t ami ekkor ki tudja fejezni transzkripciós aktivátor hatását. [?]
- **ssrB** – Az ssrA-ssrB is egy kétkomponensű rendszer mely egy szenzor kinázból és egy regulátor transzkripciós faktorból áll. Az ssrA-ssrB rendszer szükséges a baktériumsejtek makrofágokon belüli túléléséhez és replikációjához. A patogenitási sziget 2 több ssrB promótert is tartalmaz. Az ssrB több gén mellett az SPI-2-ben kódolt T3SS fehérjének átírásáért is felelős. [?]
- **trxA** – A trxA is egy több komponensű rendszer része. A trxA egyike a *Salmonella* genom két ismert thioredoxinjának. A trxA által kódolt fehérje a trx1 egy kicsi

oldható diszulfid reduktáz enzim. A *trx1* sok specifikus célfehérjéinn thiol-diszulfid redox folyamatot hajt végre, így fontos szerepet játszik az oxidatív stressz elleni védelemben. Az oxidált *trx1* a redukált *trxB*-ből regenerálódik, mely NADPH-ből szerzi a redukálóerőt. [?]

6.3. A hármás típusú szekréciós rendszer kapcsolatai

6.3.1. A *Salmonella* kapcsolatai az autofágia fehérjékkel

Az *ARN* pontjaiból 30 létesít összesen 40 kapcsolatot valamilyen *Salmonella* fehérjével. A predikciók alapján tehát hálózatban szereplő humán autofágia fehérjék 88%-a létesíthet kapcsolatot *Salmonella* fehérjével (10. ábra).



10. ábra. Az *ARN* és kapcsolatai az egyesített hálózattal

A képen az ábrán az *ARN* és valamilyen *ARN*-t és a *Salmonet*-et összekötő hat pont-fehérje látható az egyesített hálózatból kiragadvakiemelve. Az bal oldali kék háttérű pontok jelképezik az *ARN*-nek azon részét, mely csak belső kapcsolatokkal rendelkezik. Az ábra felső részén helyezkedő elhelyezkedő pontok a *Salmonet* azont kiragadott pontjai, melyek az *ARN*-nel kapcsolatot létesítenek. A szűkre színű élek a humán-humán az emberi fehérjék közti kapcsolatok, a narancs színűek pedig a predikciókból származó interspecifikus kapcsolatok. A pontok nagysága és színmélysége egyenesen arányos az egy lépésre számított topológiai fontosságukkal. Mivel a hálózat az egyesített hálózatból lett kiragadvakiemelve, így a pontok

TI értékei is a nagy hálóra vonatkoznak, de szemléletesség kedvéért a színezés és a méret kis hálózat minimum és maximum értékeihez van viszonyítva.

ahol az ábráról írsz pár mondatban le kellene írni hogy a nagyon jól látszik hogy azok a fehérjék amiknek nincsenk és megint csak, legend

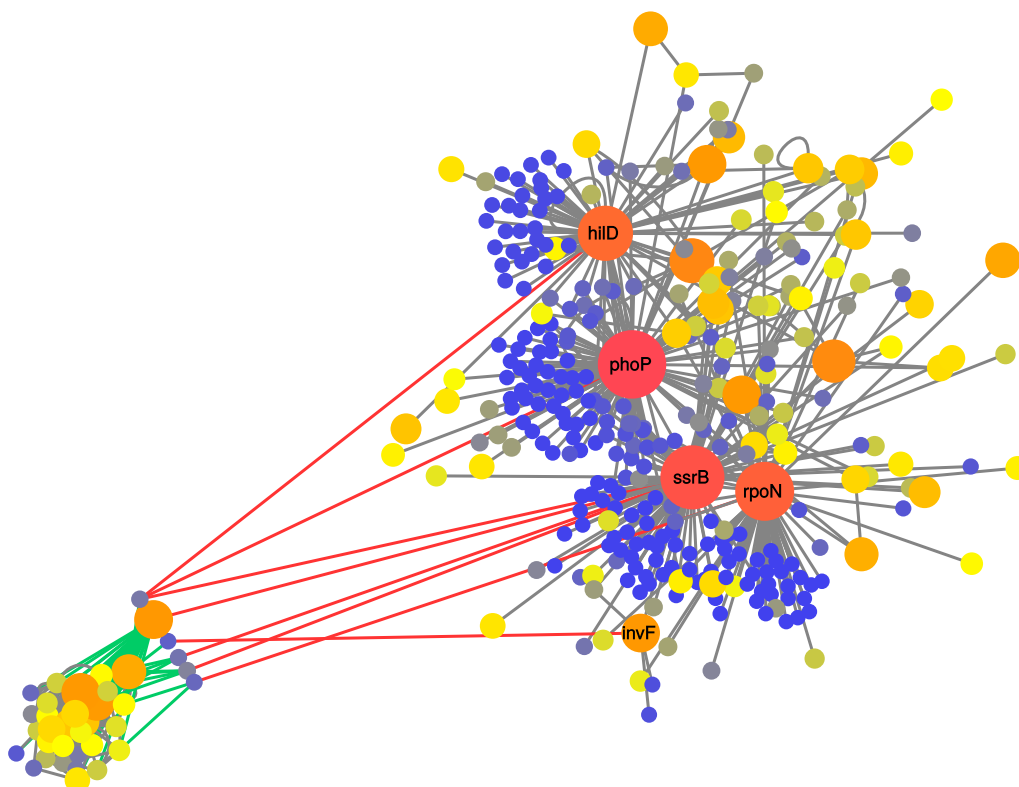
A 40 darab interspecifikus kapcsolatért *Salmonella* oldalról csupán a következő hat *Salmonella* fehérje a felelős: *sifB*, *slpA*, *sopB*, *sseI*, *sseL*, *sspH2* (10. ábra). A predikciók alapján tehát a *Salmonella* fehérjék mindössze 0,2%-a létesít csak kapcsolatot a gazda autofágia fehérjeivel.

A hat fehérje közül a következő öt bizonyítottan a T3SS-el jut ki a baktériumsejtből, míg az utolsóról kevés információnk van:

- **sifB** – A virulencia vagy más néven effektor fehérjék közé tartozik. Fertőzéskor *Salmonella* által kiváltott filamentumok (*Sif*) mentén eltávolodnak az SCV-től. [?]
- **sspH2** – Szintén effektor fehérje. Egy olyan E3 ubikvitin ligáz mely a gazda ubikvitinációs szignalizációját befolyásolja. A szubsztrát hiányában öngátló állapotban van. Szubsztrát kötésre bekapcsolódik. [?] Az *sspH2* gátló hatással van az aktin polimerizációra. [?]
- **sopB** – Egy olyan esszenciális virulencia fehérje mely inozitol-foszfát-foszfátáz mechanizmusával közvetetten aktiválja a gazda Rho-GTP-ázait. Ilyen kis-G fehérje a Cdc42 és a Rho-G. Az aktiválások hatásának eredője az aktin sejtváza átrendeződése, ennek következtében a *Salmonella* bejutásának segítése a hoszt sejtbe átfogalmazni. [?]
- **sseI** – Egy olyan effektor fehérje melynek aminoterminalisa nagy hasonlóságot mutat a *SspH2*-vel. Filaminon keresztül képes a polimerizált aktin kötésére. [?]
- **sseL** – Egy deubikvitináz enzim, mely mono és polyubikvitin láncok hidrolizálását végzi. Az *sseL*-nek *in vitro* a lizin-63 kapcsolt ubikvitin láncok sokkal jobb szubsztrátjai, melyből arra lehet következtetni, hogy a proteasómális degradáció befolyásolása helyett inkább az ubikvitin szignalizáció módosításáért felelős. [?]
- **slpA** – Ez a fehérje egy peptidil-prolil cisz-transz izomeráz. A szakirodalomban a *Salmonella* ezen fehérjéjéhez nem találtam olyan cikket, mely részletesen leírta volna a funkcióját. Az *Uniprot* adatbázis segítségével azonban találtam a legalább 90%-os hasonlóságot mutató fehérjék listájában egy *Swissprot* bejegyzést. Ez a hasonló

de annotált fehérje az *E. coli* FKBP típusú 16 kDa-os peptidil-prolil izomeráza, mely fehérjék hajtogatási sebességét növeli. [TODO link](#) A két fehérje szekvenciáját az *NCBI BLAST* algoritmusán alapuló *blastp suite-2sequences* és az *ExPASy* fehérjeszekvencia összehasonlító programjával is összehasonlítottam. Az első 92%-os egyezést adott ($E=5 \cdot 10^{-104}$) a második pedig 91,9%-t.

6.3.2. A T3SS kapcsolatai és a *Salmonella* belső kapcsolatai



11. ábra. A T3SS fehérjék első és második szomszédai

A kép a 7. ábrán látható hálózat egy részét tartalmazza, azonos színezéssel. A háló az interfész fehérjék első és második szomszédjait tartalmazza. Az öt *Salmonella* oldali első szomszéd nevei fel vannak tüntetve.

A 11. ábrán az öt feltüntetett fehérje közül csupán egy az *invF* olyan amelyik nincs benne se a 1. se a 3. táblázatban. Így elmondható, hogy a T3SS fehérjék *Salmonella* oldali első szomszédjai a mind a három topológiai kiszámolt topológiai fontosság szempontból a legmagasabb rangú pontok közé tartoznak.

7. Diszkusszió

8. Összefoglalás

9. Summary

10. Internetes hivatkozások jegyzéke

1. link [TODO github link](#)
2. link [TODO](#) homológ fehérje <http://www.uniprot.org/uniprot/P0AEM2>
3. link [ATG3](#) <http://www.uniprot.org/uniprot/Q9NT62>
4. link [Beclin-1](#) <http://www.uniprot.org/uniprot/>

11. Köszönetnyilvánítás

12. Nyilatkozat