Bakteriális patogén és ember közötti molekuláris hálózatok vizsgálata

Horváth Balázs

2015

1. Tartalomjegyzék

2. Rövidítésjegyzék

PPI - protein protein interaction

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 immunológiailag komplex kapcsolatban áll az emberrel. [Karlsson et al., 2011] Eddig több mint három millió nem redundáns mikrobiális gént sikerült kimutatni az emberben [Qin et al., 2010]. Ez a nagy genetikai állomány lehetővé teszi, hogy olyan metabolikus folyamatok játszódjanak le a humán bélben, melyeket az emberi sejtek nem képesek végrehajtani. [Karlsson et al., 2011] 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. [Backhed et al., 2005a]

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. [Brown et al., 2003]

A bélflóra hatással van még a gazda metabolizmusára is. Az *Eubacterium spp.* által oligoszacharidokból képzett vajsav részt vesz az emberi szervezet energia egyensúlyának szabályzásában. [Karlsson et al., 2011] 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. [Brown et al., 2003] 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. [Backhed et al., 2005b] Kapcsolatok mutathatók ki a bél mikrobiom megváltozása és olyan

betegségek között mint az IBD (*inflammatory bowl disease*), elhízás vagy a rák. [Karlsson et al., 2011]

A bél mikrobióta vizsgálatának módszerei

A mikrobióta vizsgálatát elsősorban a különböző meta omikák eszköztárával közelítik meg. Ezek közül is a legfőbb 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. [Karlsson et al., 2011]

Quin és társai2010-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 gigabázásnyi 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. [Qin et al., 2010]

3.2. A szakirodalomban publikált gazda patogén hálózatok

!TODO

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

Salmonella spp.

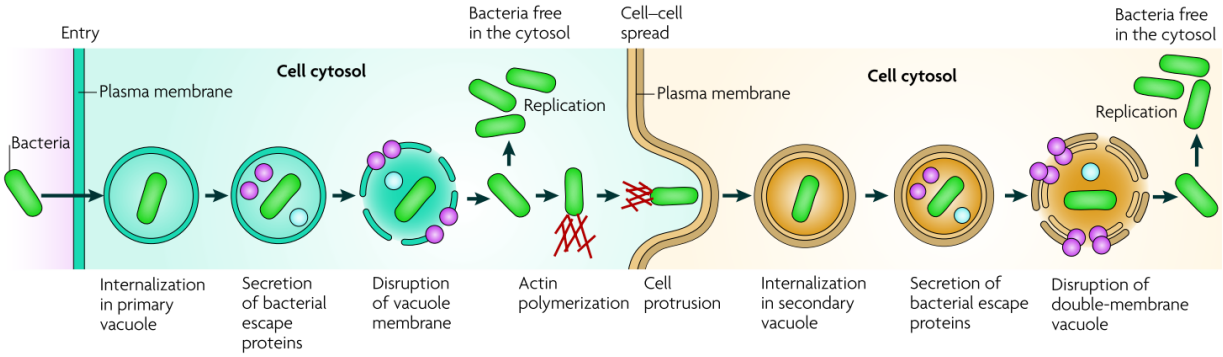
A *Salmonell*ák olyan Gram-negatív patogének melyek az állatok széles skáláját képesek fertőzni. A tudomány jelenleg több ezer szerotípust ismer, melyek két fő típusra oszthatók. Az egyik fő típus a *Typhoid*, ebbe a csoportba tartozik a *Typhi* és *Paratyphi* melyek kifejezetten embert fertőznek. A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

A fertőzés kontaminált étel vagy folyadék fogyaztásával történik. A *Salmonell*ák az alacsony pH és oxidatív stressz ellen adaptív toleranciával rendelkeznek, így képesek eltűrni 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 *Salmonell*ák úgynevezett baktérium-közvetített endocítózissal képesek még a nem fagocita típusú enterocitákba is bejutni [Haraga et al., 2008]

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 sejtváz-átrendeződést váltanak ki az eukarióta gazdában. A folyamat végén a baktérium egy vakólummal határolva a sejt belsejébe kerül. [Ray et al., 2009] Ezt a képletet a szakirodalomban SCV-nek nevezik (*Salmonella containing vacuole*). [Haraga et al., 2008]

A fagocitózis végeztével a *Salmonell*ák 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. [Haraga et al., 2008]



1. ábra. **Az intracelluláris baktériumok életciklusa**

Bejutáskor a baktériumok egy elsődleges vakólumba érkeznek. A sejt a belsejében a mikróbák 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 másdlagos memránburokba kerülnek, melyet ugyancsak felbontanak. !TODO kép magyarítás 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. [Ray et al., 2009] A szakirodalomban még nincs kialakult álláspont arról, hogy a *Salmonell*ák 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. [Haraga et al., 2008]

Az SCV-n belül a *S. typhimurium* képes a replikációra. A hármas típusú szekréciós rendszer segítségével a baktériumsejtek olyan anyagokat tudnak kibocsájtani, melyek lehetővé teszik az SCV-ből kijutást és citoplazma inváziójáát. [Jo et al., 2013]

A hármas típusú 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. [Haraga et al., 2008]

A *Salmonella* és az autofágia kapcsolata

Az autofágia egy intracelluláris katabolikus folyamat melynek szerepe van a fehérjeaggregátumok és károsodott sejtorganellumok eltávolításában és a veleszületett 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 szubsztrátjuk 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. [Jo et al., 2013]

3.4. Ökológiai hálózatok elemzésére használt topológiai mérőszámok

Miért van szükség topológiai mérőszámokra?

A konzervrációs 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. [Payton et al., 2002] 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. [Jordán et al., 2007]

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 nagynak? Pontosan mekkora biomassza hányad után mondható az adott faj ereje aránytalannak? [Payton et al., 2002] Ez utóbbi kérdések megválaszolásához szükség van olyan mérőszámokra, melyek segítségével kvantitatívvé 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. [Jordán et al., 2007]

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. [Jordán et al., 2007]

Főbb topológiai mérőszámok

Normalised degree - D

Az adott ponttal kapcsolódó pontok száma elosztva a hálózat összes pontjának számával. [Baranyi et al., 2011]

Closeness centrality - CC vagy C

A pontok száma elosztva az adott pontból eredő azt minden más ponttal összekötő legrövidebb topológiai távolságok összegével. [Baranyi et al., 2011] Ez a mérőszám megmutatja,

hogy egy adott pontnak mekkora az átlagos távolsága a hálózat összes többi pontjától. Az index kicsi 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: [Newman, 2010]

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

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. [Newman, 2010]
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: [Newman, 2010]

$$C_i = \frac{1}{\ell_i} \tag{2}$$

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. [Baranyi et al., 2011] 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. [Newman, 2010]

Topological importance - TI^n

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:

$$d_{X,Y} = \frac{1}{x} \tag{3}$$

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:

$$d^n_{p_{XY}} = \prod_{i=1}^{n-1} d^1_{i,i+1} \tag{4}$$

Ahol:

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

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

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

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

$$d^n_{XY} = \sum d^n_{pXY} \tag{5}$$

Ahol:

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

d^n_{XY}: 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:

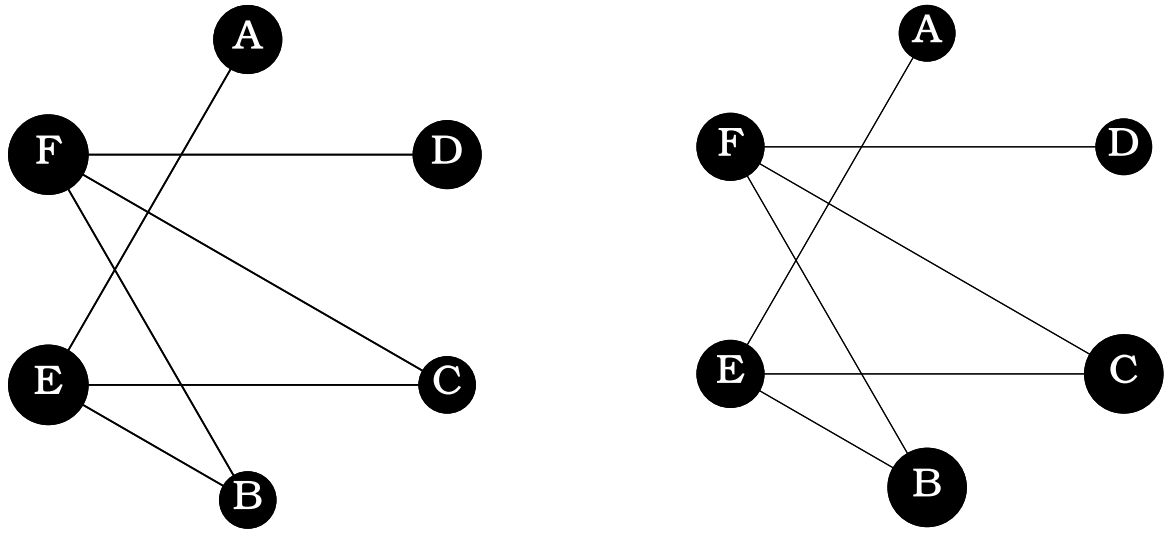
$$\mathrm{TI}^n_A = \sum d^n_{j,A} \tag{6}$$

Ahol:

TI^n_A: A pont n lépésre számított topológiai fontossága.

$d_{j,A}$: 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** $\mathbf{TI^2}$ ***(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. [Jordán et al., 2003] 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:

$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}
\begin{array}{c}
\begin{array}{cccccc} A & B & C & D & E & F \end{array} \\
\begin{bmatrix}
0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\
0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\
0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\
0 & 0 & 0 & 0 & 0 & \frac{1}{3} \\
1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\
0 & \frac{1}{2} & \frac{1}{2} & 1 & 0 & 0
\end{bmatrix}
\end{array}
$$

d értékek

$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}
\begin{array}{c}
\begin{array}{cccccc} A & B & C & D & E & F \end{array} \\
\begin{bmatrix}
\frac{1}{3} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\
0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{2}{3} & \frac{1}{3} \\
0 & 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3}
\end{bmatrix}
\end{array}
$$

d^2 értékek

Az ábrázolt mátrixok elrendezése követi a populációdinamikai konvenciókat, tehát például $d_{BF} = \frac{1}{2}$ azt jelenti, hogy F pont a B-re $\frac{1}{2}$ erővel hat. Mindkét mátrixra érvényes az, hogy az adott oszlop értékeinek összege egy. Ez a tulajdonság a d érték definíciójából fakad. d azt mutatja meg, hogy adott pont a cél pont kapcsolatainak hányad részét adja. Ezáltal minden pont egy egységnyi hatást kap ami eloszlik a vele kapcsolatban álló pontok között. [Jordán et al., 2003] Ezt a hatást jól szemlélteti az 2. ábra jobb oldali része amin látható, hogy B, C, D és A pontok kimenő hatása kisebb, mivel célpontjaik sok hatást fogadnak.

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. [Jordán et al., 2003]

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, [Jordán et al., 2003] vagyis $d^n_{X,X} \neq 0$ 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: $F \to B \to F$, $F \to C \to F$ és $F \to D \to F$.

Weighted Topological Importance - WI^n !TODO

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 következőeket 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 mérőszámok a molekuláris kapcsolati hálók elemzésére. Valamint, hogy az így előállított rangsorok mennyire korrelálnak a jelenleg használt Salmonella és humán bélsejteket 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-fehéjre 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 összeejtésére.

6. Program írása mely képes hálózatokban a megadott útvonalhosszra kiszámolni a topológiai fontosságot.

7. A kapott adatsorok értékelése, hálózatok ábrázolása, biológiai relevancia keresése.

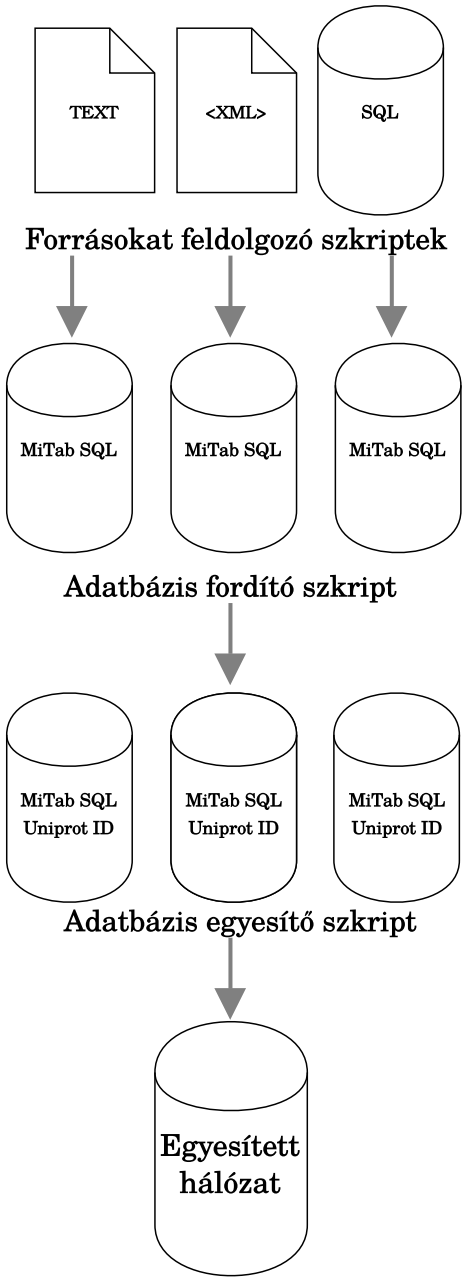
5. Források és módszertan

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-on futtatható szkripteket is alkalmaztam. A fehérjék azonosítójának fordítását végző szkriptek egyike Kadlecsik Tamás *Javascript*ben írt fordítószkriptjének kismértékű módosítása.

A diplomamunkám során alkalmazott szkriptek egy részét a *Signalink* 3 szignalizációs adatbázis nulladik és harmadik rétegének létrehozásakor készítettem. Mivel a *Signalink* nulladik rétege 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 nullás rétegétől, mert például 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.



3. ábra. **A hálózat létrehozásának folyamata**

A különböző forrásokból származó adatok esetén először a forrás formátumokat feldolgozni képes szkriptek átalakítják azokat 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 gyárt, 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 adatok tárolása

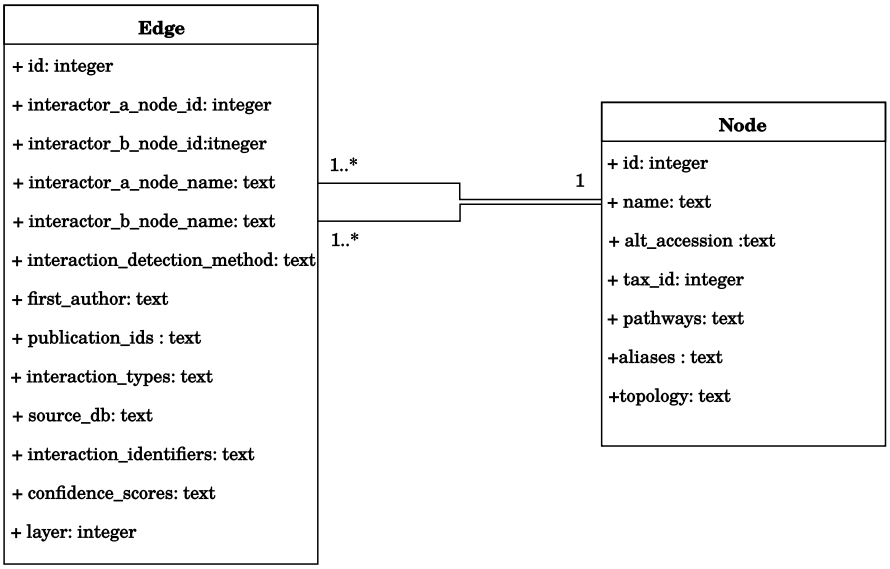
Az adatok ideiglenes tárolására, már a *Signalink 3* készítése óta Kadlecsik Tamás javaslatára *SQLite 3* adatbázis fájlokat alkalmazunk.

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. [Owens, 2006]

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*. Az *SQLite* adatbázisfájlok másik előnye, hogy rendelkezésünkre áll az SQL nyelv. Mivel SQL parancsok segítségével gyorsan kezelhetőek 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.

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.

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 GitHub 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 összeejtő szkriptek
- A MiTab SQL formátumot kezelő osztály
- A topológiai elemzést véző szkript
- Az adatszűrésre használt SQL szkriptek

5.2. A források feldolgozása

5.2.1. Az források feldolgozásának eszközei

MiTab SQLite adatbázis API

A *PsimiSQL* egy *Python* 2.7-es szintaxisban írt osztály, melyet még a *Signalink 3* összeállításához

kezdtem el készíteni, 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 gráfba 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. Elsődleges azonosítót kapnak még olyan fehérjék is, melyek már benne voltak az adatbázisban de csak 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. Minden swissprot azonosító egyben elsődleges azonosító is. 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 Kadlecsik Tamás féle szótárépítő scripttel állítottam elő. Az azonosítókat a szótárak alapján saját készítésű szkripttel fordítottam.

A predikciókhoz egy olyan szótárat kellet létrehozni, mely *Salmonella* génazonosítókhoz rendel *Salmonella* uniprot azonosítókat. Ezt szintén Kadlecsik Tamás szkriptjével állítottam elő. Egy általam írt másik fordítószkript segítségével pedig az előzőhöz hasonló módon fordítottam a predikciókat.

5.2.2. A források

***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. [Turei et al., 2015]
Az ARN hat rétegből épül fel:

1. Autofágia fehérjék.
2. Az első réteg fehérjéinek 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 szabályzókhoz kötnek.

[Turei et al., 2015]

Az ARN feldolgozása

Az ARN adatbázisnak csak az első, autofágia fehérjéket tartalmazó rétegét használtam fel a gazda patogén hálózat összeállításához. A hálózat letöltését követően azt egy általam írt *Python* scripttel MiTab SQL formátumba alakítottam. A fordító szkripttel az akkor legfrissebb *Uniprot* adatbázis azonosítókra fordítottam.

5.2.3. Salmonet

5.2.4. A predikciók

A predikciók forrása

Az eddig ismertetett források csak fajon belüli kapcsolatokból 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 [Krishnadev and Srinivasan, 2011] és [Kshirsagar et al., 2012] humán-*Salmonella* predikcióit használtam.

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.

5.2.5. 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 eg *has-map*ben eltárolja a pontokat és éleket és ezek tulajdonságait. A *hasm-map*-ből gyorsan ki kereshető, hogy egy adott él vagy pont benne van-e már. Ami 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.

5.3. A hálózatok topológiai elemzése

5.3.1. 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 létrehozott ingyenes csomag. A *TopologyAnalyser* topológiai fontosság meghatározásához használt függvényeit én írtam.

Az osztály konstruktorának egyetlen paramétere egy éllista. Példányosítás után a *TopologyAnalyser* típusú objektum, egy éllistát, egy *Networkx.Graph* objektumot és egy

egység-élerősségi mátrixot tartalmaz *hashmap*-ként letárolva. Az élerősség mátrix (*self.edgeStrength*) *hashmap* kulcsai a mátrix indexei, az értékei pedig az élerősségek *Fraction* típusú objektumok. A mátrix indexei az egység-élerő mátrix esetén maguk az élt alkotó pontok *Uniprot* azonosítói. Az élerősség mátrix 0 értékkel rendelkező cellái nincsenek letárolva a *hasmap*-ben. Az egység

5.3.2. Algoritmusok, alkalmazott technológiák

6. Eredmények (A hálózat elemzése)

6.1. Főbb statisztikák

6.2. Kapott topológiai adatok és jelentésük

7. Diszkusszió

8. Összefoglalás

9. Summary

10. Hivatkozások jegyzéke

1. link TODO github link

11. Köszönetnyilvánítás

12. Nyilatkozat

Hivatkozások

[Backhed et al., 2005a] Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005a). Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–1920.

[Backhed et al., 2005b] Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005b). Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–1920.

[Baranyi et al., 2011] Baranyi, G., Saura, S., Podani, J., and Jordán, F. (2011). Contribution of habitat patches to network connectivity: Redundancy and uniqueuness of topological indices. *Ecological Indicators*.

[Brown et al., 2003] Brown, A. J., Goldsworthy, S. M., Barnes, A. A., Eilert, M. M., Tcheang, L., Daniels, D., Muir, A. I., Wigglesworth, M. J., Kinghorn, I., Fraser, N. J., Pike, N. B., Strum, J. C., Steplewski, K. M., Murdock, P. R., Holder, J. C., Marshall, F. H., Szekeres, P. G., Wilson, S., Ignar, D. M., Foord, S. M., Wise, A., and Dowell, S. J. (2003). The Orphan G protein-coupled receptors GPR41 and GPR43 are activated by propionate and other short chain carboxylic acids. *J. Biol. Chem.*, 278(13):11312–11319.

[Haraga et al., 2008] Haraga, A., Ohlson, M. B., and Miller, S. I. (2008). Salmonellae interplay with host cells. *Nat. Rev. Microbiol.*, 6(1):53–66.

[Jo et al., 2013] Jo, E. K., Yuk, J. M., Shin, D. M., and Sasakawa, C. (2013). Roles of autophagy in elimination of intracellular bacterial pathogens. *Front Immunol*, 4:97.

[Jordán et al., 2007] Jordán, F., Benedek, Z., and Podani, J. (2007). Quantifying positional importance in food webs: A comparison of centrality indices. *Ecological modelling*.

[Jordán et al., 2003] Jordán, F., Liu, W.-C., and van Veen, F. (2003). Quantifying the importance of species and their interactions in a host-parasitoid community. *Community ecology*.

[Karlsson et al., 2011] Karlsson, F. H., Nookaew, I., Petranovic, D., and Nielsen, J. (2011). Prospects for systems biology and modeling of the gut microbiome. *Trends Biotechnol.*, 29(6):251–258.

[Krishnadev and Srinivasan, 2011] Krishnadev, O. and Srinivasan, N. (2011). Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int. J. Biol. Macromol.*, 48(4):613–619.

[Kshirsagar et al., 2012] Kshirsagar, M., Carbonell, J., and Klein-Seetharaman, J. (2012). Techniques to cope with missing data in host-pathogen protein interaction prediction. *Bioinformatics*, 28(18):i466–i472.

[Newman, 2010] Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

[Owens, 2006] Owens, M. (2006). *The Definitive Guide to SQLite*. Apress.

[Payton et al., 2002] Payton, I. J., Fenner, M., and Lee, W. G. (2002). Keystone species: the concept and its relevance for conservation management in New Zealand. *Science for canservation*.

[Qin et al., 2010] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J. M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., Wang, J., Antolin, M., Artiguenave, F., Blottiere, H., Borruel, N., Bruls, T., Casellas, F., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Forte, M., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Jamet, A., Juste, C., Kaci, G., Kleerebezem, M., Knol, J., Kristensen, M., Layec, S., Le Roux, K., Leclerc, M., Maguin, E., Minardi, R. M., Oozeer, R., Rescigno, M., Sanchez, N., Tims, S., Torrejon, T., Varela, E., de Vos, W., Winogradsky, Y., and Zoetendal, E. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

[Ray et al., 2009] Ray, K., Marteyn, B., Sansonetti, P. J., and Tang, C. M. (2009). Life on the inside: the intracellular lifestyle of cytosolic bacteria. *Nat. Rev. Microbiol.*, 7(5):333–340.

[Turei et al., 2015] Turei, D., Foldvari-Nagy, L., Fazekas, D., Modos, D., Kubisch, J., Kadlecsik, T., Demeter, A., Lenti, K., Csermely, P., Vellai, T., and Korcsmaros, T. (2015). Autophagy Regulatory Network - a systems-level bioinformatics resource for studying the mechanism and regulation of autophagy. *Autophagy*, 11(1):155–165.