dc_901_14

Átfedő modulok molekuláris biológiai kölcsönhatási hálózatokban

MTA doktori értekezés

Farkas Illés

Magyar Tudományos Akadémia
Támogatott Kutatóhelyek Irodája
MTA-ELTE Statisztikus és Biológiai Fizika Kutatócsoport

Budapest, 2015 április

dc_901_14

Tartalomjegyzék

Bevezetés

A dolgozat a biológia és a statisztikus fizika határterületén található kutatási témákat és eredményeket ismertet. A bevezetésben elsőként szeretném bemutatni a dolgozat tudományterületét a terület keletkezésének leírásával. Ezután megpróbálom a dolgozat három fő fejezetének eredményeit elhelyezni a tudomány „palettáján”. A bevezetés végén felsorolt további információk a dolgozat olvasását és használatát segítik. Célom, hogy a dolgozat színvonalas tudományos eredményeket jól átlátható, könnyen olvasható és érdekes módon mutasson be.

A dolgozatban a szakirodalom ismertetése során mindenütt nagyobb súlyt kapnak azok az eredmények, amelyek a bemutatott saját munkákkal kapcsolatosak. A legtöbb esetben ezeket az irodalmi eredményeket és módszereket a használatuk során szerzett saját tapasztalataim alapján foglalom össze. Általános elvként próbálom követni azt az elrendezést, hogy a társszerzőségemmel készült cikkek előtt (időben korábban) megjelent szakirodalmi eredmények az egyes fejezeteken belül az „Előzmények” című alfejezetbe kerüljenek.

A bemutatott kutatás célja

A dolgozatban bemutatott kutatás célja, hogy az élő rendszerek molekuláris biológiai szintű jelenségeiben a modern hálózatkutatás és a hálózati modul keresés segítségével általános érvényű összefüggéseket azonosítson, valamint ilyen összefüggések felismerésére alkalmas eszközöket biztosítson. A kutatási eredményeket a dolgozat három fejezetben mutatja be. Az első és második fejezet döntően alapkutatási eredményeket ismertet, míg a harmadik fejezetben az alkalmazásokon van a hangsúly.

A dolgozat tudományterülete

A biológia hosszú ideig döntően leíró tudományág volt, az élettel kapcsolatos jelenségek felismerését, rendszerezését és részletes leírását végezte. Ennek a feladatnak a nagyságát érzékelteti, hogy néhány évvel ezelőtti számítások szerint a jelenleg élő szárazföldi fajok 86 százalékát és a vízben élő fajok 91 százalékát még nem ismerjük [1]. A fajok azonosításán túl további feladatot jelent a már azonosított fajok változatos és igen összetett egyedfejlődésének, az egyedeik testfelépítésének, az egyedeik közötti kölcsönhatásoknak és további térbeli/időbeli jelenségeiknek a leírása. Az elmúlt évszázadokban a fajok leírása és az összehasonlítások során nyert felismerések fokozatosan elvezettek olyan kérdésekig, amelyek a konkrét megfigyelt eredményeken túl már a megfigyelt eredmények mérhető okait keresik. Például felmerült, hogy miért hasonlítanak egymásra egyes fajok, egy élőhelyen mi határozza meg a fajok számát, hogyan hatnak egymásra a fajok és hogyan hatnak a fajokra az élőhelyek [2].

Az okok vizsgálata újabb nagy lendületet kapott a 20. században, amikor – a század elejétől kezdődően – a fizikai és kémiai kutatások eljutottak az anyag korábban ismeretlen alkotóelemeinek leírásáig. Az anyag szerkezetének részletes megismerése nyomán a biológia eszköztára hatalmas változáson ment át és megszületett (többek között) a molekuláris biológia és a hozzá kapcsolódó számos új kutatási irány. Így a biológia és a kapcsolódó élettudományi és gyógyászati területek által kísérletekkel és mérésekkel vizsgálható kérdések köre korábban nem látott módon bővülhetett és jelenleg is folyamatosan bővül. A bővülő mérési lehetőségek miatt az előző bekezdésben említett – és önmagában is hatalmas – „klasszikus” leíró típusú biológiai ismeretanyaghoz napjainkban összetett ismeretek társulnak a biológiai rendszerek működéséről, azaz dinamikájáról. A legintenzívebben kutatott biológiai folyamatok közé tartoznak a sejtosztódás, a szervfejlődés, a sejtek kommunikációja és vezérelt pusztulása. Mindegyik esetben a kutató számára az egyik legnagyobb kihívás az, hogy a rendelkezésre álló sok és nagyon különböző biológiai jelentésű mérési információt összefüggő egészként „megértse”. A megértés érdekében az egyik legfontosabb lépés a megfelelő matematikai leíró/elemző eszközök kiválasztása, amelyek kiemelik az adott biológiai kérdés szempontjából fontos mérési információkat, és csökkentik a kutató által a vizsgált biológiai kérdés szempontjából kevésbé fontosnak ítélt információk súlyát.

A biológiai mérési eredmények értelmezése során a megfelelő matematikai eszközök használata azért is különösen fontos, mert a biológiai mérések napjainkban egyre

több számszerű eredményt adnak. A számok (mért értékek) közötti összefüggések feltárására a kvantitatív természettudományos (például a fizikai és kémiai) és a mérnöki gondolkodásmód kíválóan alkalmas. Néhány évtizeddel ezelőtt a biológia és kvantitatív természettudományok szemlélete között még alapvető eltérés volt. Míg a biológia egy rendszer minél részletesebb leírására törekedett (tehát az egyedi tulajdonságok kiemelésére), addig a fizikai, a kémiai és a mérnöki munka célja az általánosítás volt (tehát az egyedi tulajdonságok kiküszöbölésére). Az utóbbi évtizedek egyik fontos fejleménye, hogy ez az erős eltérés csökken. A kvantitatív természettudományok és a mérnöki tudományok egyre gyakrabban vizsgálják sok alkotóelemből álló és erősen összetett rendszerek („komplex” rendszerek) minőségileg eltérő jelenségeit. Ezzel egyidejűleg a biológia számos, erősen kvantitatív területtel bővült, amelyeknek a célja gyakran az általános érvényű elvek felismerése az egyedi részletek figyelmen kívül hagyásával.

A biológiai és a fizikai/matematikai gondolkodásmód közeledésének egy kíváló példája a hálózatos leírási módszer. A fizikában a ’90-es évek végén kapott nagy lendületet a kölcsönható sokrészecske-rendszerek hálózatokkal történő elemzése. Ennek az alapja az, hogy a megfigyelt jelenséget – mint egészt – alkotóelemeire bontjuk, és megpróbáljuk az alkotóelemek közötti összes bináris (más néven páros, azaz két résztvevő között fellépő) kölcsönhatást azonosítani. Ha (i) sikerül a pár kölcsönhatások nagy részét azonosítani, és ha (ii) a kölcsönhatások nagy része valóban két elem között lép fel, akkor a hálózatos leírási módszer igen intuitív és hatékony lehet. A megfigyelt valós jelenség modelljeként használt hálózat könnyen lerajzolható és értelmezhető, továbbá számos matematikai (például lineáris algebrai és kombinatorikai) eszközhöz kíválóan kapcsolható. Ha egy rendszerben nem sikerül a kölcsönhatások nagy részét megismerni, vagy a kölcsönhatások között jelentősek a kettőnél több résztvevőt tartalmazó kölcsönhatások, akkor a hálózatos leírás kiegészítésre szorul.

A technológiai és társadalmi rendszerekhez képest a biológiai (élő) rendszerekben jóval gyakoribb, hogy két résztvevő (például fehérje) kölcsönhatását sok másik (fehérje) állapota módosíthatja. Például ha két fehérje térbeli szerkezete (felszíne) olyan, hogy ez képessé teszi kettőjüket a kapcsolódásra, akkor is még nagyon sok körülmény (fizikai, kémiai paraméter és más fehérjék állapota) befolyásolhatja, hogy a kölcsönhatásuk valóban bekövetkezik-e. A biológiai feladatokat ugyanis gyakran kettőnél több, egymással kapcsolatban álló résztvevő (például fehérje) végzi [3]. Ha a fehérjék egy

csoportja képes egy biológiai feladat elvégzésére, akkor a csoportot szokás modulnak nevezni[1]. Egy fehérje modul tagjai gyakran nincsenek azonos időpontban mindannyian szorosan összekapcsolódva, mégis az adott feladat elvégzéséhez gyakori kölcsönhatásaikra szükség van. A „fehérje modul”-tól kissé eltér a „fehérje komplex”, mert ez utóbbi általában azt jelöli, hogy a résztvevő fehérjék mindannyian folyamatosan fizikailag kapcsolatban vannak, tehát azonos időpontban jelen vannak ugyanazon a helyen.

A modulok azonosításához használhatjuk a már feltérképezett pár kölcsönhatási hálózatokat. Így megtarthatjuk a hálózatok nagyon intuitív leírásmódját és együtt használhatjuk a modulok (fehérje csoportok) biológiai szempontból jóval pontosabb eszköztárával. A hálózati modulok azonosítása érdekében először soroljuk fel az összes megfigyelhető pár kölcsönhatást. Ezután keressünk a pár kölcsönhatásokból kapott hálózatban olyan csoportokat, amelyeken belül sokkal nagyobb a kapcsolatok sűrűsége, mint a hálózatban máshol átlagosan. A kapcsolatok sűrűségének egy gyakori mérőszáma, hogy egy csoport két véletlenszerűen kiválasztott eleme milyen valószínűséggel van összekapcsolva (egy pár kölcsönhatás által).

Ha egy biológiai rendszerben a mérések alapján megismert pár kölcsönhatási listát „kicseréljük” fehérjék (vagy más biológiai alkotóelemek) nagyobb csoportjaira, akkor gyakran ezek a csoportok már jóval informatívabbak, hiszen a valóságban a biológiai feladatokat nem egy (vagy két egymással kapcsolatban álló) fehérje végzi, hanem fehérjék kisebb-nagyobb csoportjai. A csoportok (más néven: modulok, klaszterek) azonosításához használható legáltalánosabb kiindulási kölcsönhatási lista a fehérje-fehérje kölcsönhatások listája, amely legtöbbször PPI (protein-protein interaction) vagy PIN (protein interaction network) néven található meg a szakirodalomban. A „PPI” betűszóban található „interaction” szó kezdetben (a ’90-es évek végén és a 2000-es évek legelején) azt jelentette, hogy a PPI hálózatként felsorolt fehérje párok valóban fizikai kölcsönhatásokat jelölnek. Ennek a hálózatnak a (csúcs)pontjai fehérjék, és a hálózat egy éle (egy pont-pont kapcsolat) a két pont által jelölt egy-egy fehérje közötti kölcsönhatást mutatja. A PPI hálózat élek (azaz fehérje párok) listája. A 2000-es évek elejétől kezdődően a „PPI”-nek nevezett hálózatok egyre több, a fizikai kölcsönhatásokon túli adatot is integráltak, így napjainkban már helyesebb a „PPA” (protein-protein association) hálózat megnevezést használni. Az ezekben összesített adat típusokról a dolgozat első fejezetében részletes leírás található.

[1] A „modul” szó és a „fehérjék által alkotott funkcionális modul” kifejezés a biológiai hálózatos irodalomban igen elterjedt és elfogadott. A „modul” szónak a molekuláris biológiában van más – jóval régebben használatos – jelentése is. A fehérjék szerkezetét és funkcióit kutató szakirodalomban az aminosav szekvencia, a szerkezet és a funkció alapján együttesen megállapítható jellegzetes fehérje szakaszokat gyakran doménnek hívják, és a több élőlényben előforduló doméneket szokás modulnak nevezni.

A dolgozatban tárgyalt kérdések

A dolgozat első fejezete fehérje-fehérje kölcsönhatási (fehérje-fehérje asszociációs) hálózatokban történő modul keresés eredményeit mutatja be. Mivel az élő sejtekben a biológiai funkciók száma meghaladja a funkciók elvégzéséhez rendelkezésre álló fehérjék számát, ezért természetes, hogy a modul (fehérje csoport) keresési módszer átfedő modulokat azonosít. Az általunk (és a módszerünk segítségével mások által) azonosított átfedő fehérje-fehérje kölcsönhatási modulok [T1, T2, T3] azért jelentősek, mert (i) hozzájárulnak ismeretlen funkciójú fehérjék biológiai funkcióinak előrejelzéséhez és (ii) segítik már ismert funkciójú nagyobb csoportokon belüli kisebb csoportok azonosítását. Az (i) pontban említett egyedi fehérje funkciók előrejelzésével kapcsolatban egy irodalmi összefoglaló [4] beszámol munkánkról, és nagy számú további független hivatkozás is érkezett cikkeinkre. A (ii) pontban említett fehérje-fehérje kölcsönhatási részcsoportoknak a gyógyászati szerepe lehet jelentős, mert egy modul fehérjéinek részleges gátlásával elérhető, hogy az adott modul biológiai funkciója jóval erősebben legyen gátolva, mint a vele átfedő más modulok funkciói. Ez azért jelentős lehetőség, mert egyedi fehérjék gátlása esetén az adott fehérje részvételével végzett feladatok közül nehezebb kiválasztani, hogy melyiket kívánjuk gátolni, és melyeket kívánunk sértetlenül hagyni a mellékhatások csökkentése érdekében.

A dolgozat első fejezetében bemutatott CFinder algoritmus nem csak molekuláris biológiai rendszerekben alkalmazható, hanem számos más típusú rendszerben is, például társadalmi, technológiai, gazdasági és kognitív hálózatokban. Ezekben a hálózatokban a folyamatok egyes lépései (például az emberi kapcsolatok eseményei, a számítógépek kommunikációjának lépései vagy akár szerződések) sokszor igen pontosan leírhatóak a résztvevő párok közötti kölcsönhatásokkal (kapcsolatokkal). A molekuláris biológiai rendszerekben a pár kölcsönhatás gyakran pontatlan közelítés, mert a molekuláris biológiai események jelentős részét kettőnél több résztvevő határozza meg. Ezért az átfedő modulok keresésének a biológiában kiemelt jelentősége van.

Az említett általános fehérje-fehérje kölcsönhatások (PPI) csupán két fehérjét és a köztük lévő kapcsolat létét vagy hiányát mutatják. Természetesen ehhez az egyszerű irányítatlan kapcsolathoz képest a biológiai mérésekből gyakran több információ is ismert. A részletesebb adatok már nem állnak rendelkezésre minden fehérje-fehérje kölcsönhatás esetén, hanem csak speciális esetekben. Három ilyen speciális eset a

transzkripció és a transzláció szabályozása, valamint a jelátvitel. A transzkripció során a DNS egy szakaszáról mRNS (messenger RNS, hírvivő RNS) másolat készül és a transzláció során az mRNS másolatból fehérje készül. A jelátvitel a sejtet érő (általában) külső jeleknek a sejtmembrántól a sejtmagig történő továbbítását és feldolgozását végzi. A jelátvitel egyik jellemző kölcsönhatási formája a fehérje térszerkezet módosítás kis funkciós csoportok (például foszfát csoport) elhelyezésével vagy eltávolításával. A transzkripció és a transzláció szabályozása valamint a jelátvitel esetén egyaránt szokás egy „A” gént (pontosabban: a DNS-en található leolvasási szakaszt), az „A” gén alapján készült mRNS másolatot és az mRNS másolat alapján készült fehérje minden módosulatát egyetlen hálózati csúcspontnak tekinteni. A transzkripció szabályozási hálózat irányított kapcsolatokat (éleket) tartalmaz, és az A géntől akkor mutat él a B génhez, ha ismert, hogy az A gén alapján készült fehérje a B génnek a DNS-en található leolvasási szakasza elé képes kapcsolódni („bekötni”) és ezáltal szabályozni, hogy a B génről időegység alatt hány mRNS másolat készül.

A dolgozat második fejezete transzkripció és transzláció szabályozási hálózatokat tárgyal. Ezekben a hálózatokban a konkrét (irányított) szabályozási kapcsolaton túl legtöbbször az is ismert, hogy az egymással kapcsolatban lévő A és B fehérje közül az A fehérje segíti (serkenti) vagy gátolja a B gén transzkripcióját. Ennek alapján szokás a transzkripció szabályozási hálózat irányított éleit súllyal is ellátni és a súlyokat zöld (serkentés, pozitív él súly) és piros (gátlás, negatív él súly) színnel jelölni. A transzláció szabályozásának többféle módja ismert. A dolgozatban tárgyalt esetben az mRNS-ből történő aminosav lánc (fehérje) készítést – azaz a transzlációt – az mRNS-ektől különböző rövid RNS láncok képesek gátolni. Ezeknek a rövid, transzlációt gátló – más szóval: „csendesítő” – RNS-eknek a szakirodalomban az átfogó neve „short silencing RNA” [2] és egy nagy csoportjukat miRNS-nek (mikroRNS-nek) nevezik. A dolgozatban vizsgált transzláció szabályozási hálózatban minden irányított él egy transzláció gátlási kölcsönhatást jelöl, az él egy rövid csendesítő RNS-ből indul ki és egy mRNS-be fut be. A második fejezetben bemutatott eredményeink közül kiemelem, hogy számítási módszert dolgoztunk ki a rövid csendesítő mikroRNS-ek relatív fontosságának mérésére [T7]. A módszer célja annak az előrejelzése, hogy az

[2] Friss kutatási eredmények szerint léteznek nem csak rövid, hanem hosszú „csendesítő” RNS-ek is. A mikroRNS-ek nem kódolnak fehérjét, ezért a nem kódoló RNS („non-coding RNA”) molekulák nagyobb csoportjába tartoznak.

emberi sejtekben melyik mikroRNS eltávolítása esetén várható a legnagyobb eséllyel fenotípus változás.

A dolgozat harmadik fejezete sejteken belüli jelátviteli hálózatokat mutat be három fajban. Érdemes megjegyezni, hogy amint az általános fehérje-fehérje kölcsönhatási hálózat felől haladunk egyre speciálisabb biológiai hálózatok felé, úgy lesz maga a hálózat (a vizsgálataink kiindulópontja) is egyre kevésbé elérhető az irodalomban. Emiatt magát a hálózatot is fel kell építeni a rendelkezésre álló sok és sokféle kísérleti adatból megbízható kritériumok alapján. A sejteken belüli jelátvitel esetében az ismert jelátviteli útvonalak irodalmának manuális – tehát nem automatizált – feldolgozásával kaptuk meg a kiindulásként használt irányított kölcsönhatás listát. A 3. fejezetben leírt eredmények közül kiemelem, hogy az így felépített jelátviteli útvonalak segítségével a *C. elegans* fajban (fonálféreg) előrejeleztük 6 fehérje részvételét a Notch jelátviteli útvonalban, és ezt az előrejelzést sikerült a számítógépes előrejelzéssel azonos cikkünkben kísérletes módon igazolni [T10].

További információk

A dolgozat mindhárom fejezetében az utolsó két alfejezet (i) az adott fejezetben a társszerzőségemmel készült eredmények összefoglalása és (ii) a saját hozzájárulásom az eredményekhez. A dolgozatban a többes szám első személy arra vonatkozik, hogy az eredményeket az adott publikációkban szereplő társszerzőkkel együtt közösen értem el. A dolgozat ábrái közül 9 ábrát társszerzőségemmel készült publikációkból vettem át. Ezeknek az ábráknak az aláírásában leírtam, hogy ki milyen munkát végzett az adott ábra elkészítése során. További 5 ábrát a dolgozathoz készítettem. Az ábrák – néhány kivétellel – a dolgozat PDF fájljában színesek, A dolgozat PDF fájljának szövege tartalmaz klikkelhető hiperlinkeket. Ezeknek nagy része a dolgozat ábráira és a hivatkozott publikációkra mutat. Ha az olvasó klikkel a PDF fájlban egy ilyen hiperlinkre (ami a dolgozaton belülre mutat), akkor az adott ábrától vagy az irodalomjegyzékből vissza tud ugrani a PDF fájlnak arra a helyére, ahol a belső hivatkozásra klikkelt. Ehhez Acroread olvasóval Windows-on az „ALT - balra nyíl” billentyűket kell leütni, Macintosh-on a „Cmd - balra nyíl” billentyűket (Linux-on még nem találtam megoldást erre a kérdésre).

A doktori dolgozat (a pályázati kiírásban: „doktori mű”) és a tézisfüzet áttekinthetősége érdekében a dolgozat mindhárom fejezetének „Eredmények” részében a számozott alfejezetek számozása azonos a tézisfüzetben található tézispontok számozásával: **1.1**, **1.2**, **1.3**, **1.4**, **1.5**, **2.1**, **2.2**, **3.1** és **3.2**. A két dokumentumban a tézispontokhoz kapcsolódó saját publikációk számozása szintén azonos: [T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12]. A dolgozat minden alfejezetének címe mellett megtalálható azoknak a saját publikációknak a sorszáma, amelyeknek az eredményeit az adott alfejezet használja. A Ph.D. fokozatom megszerzése (2004 március) óta részt vettem több olyan kutatásban, amelyeknek a témája eltér a D.Sc. dolgozat témájától. Ezeknek a kutatásoknak az eredményeként nemzetközi referált folyóiratokban megjelent 7 cikk, amelyeknek a számozása a doktori dologzatban és a tézisfüzetben szintén azonos: [M1, M2, M3, M4, M5, M6, M7].

A dolgozat PDF fájljának végén az irodalmi és saját hivatkozások mindegyikénél található hiperlink a cikk vagy könyv online elérhető változatára. A dolgozat téziseihez felhasznált saját publikációk esetében a link szövege általában „Teljes cikk PDF”, és a klikkelés egy ingyenes PDF fájlhoz vezet, ami az adott folyóirat oldalán vagy a honlapomon található. A dolgozat végén szerepelnek további saját (a társszerzőségemmel készült) és nem saját közlemények. Ezeknél a hiperlink szövege általában a publikáció DOI (Digital Object Identifier) száma, ez a DOI szám klikkelhető, és a klikkelés a publikációnak a megjelenési folyóiratnál látható weboldalára vezet. Ha a publikáció eredeti honlapján az Olvasó számára előfizetés hiányában egy PDF fájl nem hozzáférhető, akkor az adott cikk teljes (pontos) címét érdemes bemásolni két idézőjel között a Google Scholar keresőmezőjébe. Ezután a Google Scholar által talált cikk alatt az „Összes változat” linkre történő klikkelés után a kapott összes változat listájában általában megjelenik legalább egy ingyenes PDF verzió is, például valamelyik szerző honlapján.

1. fejezet

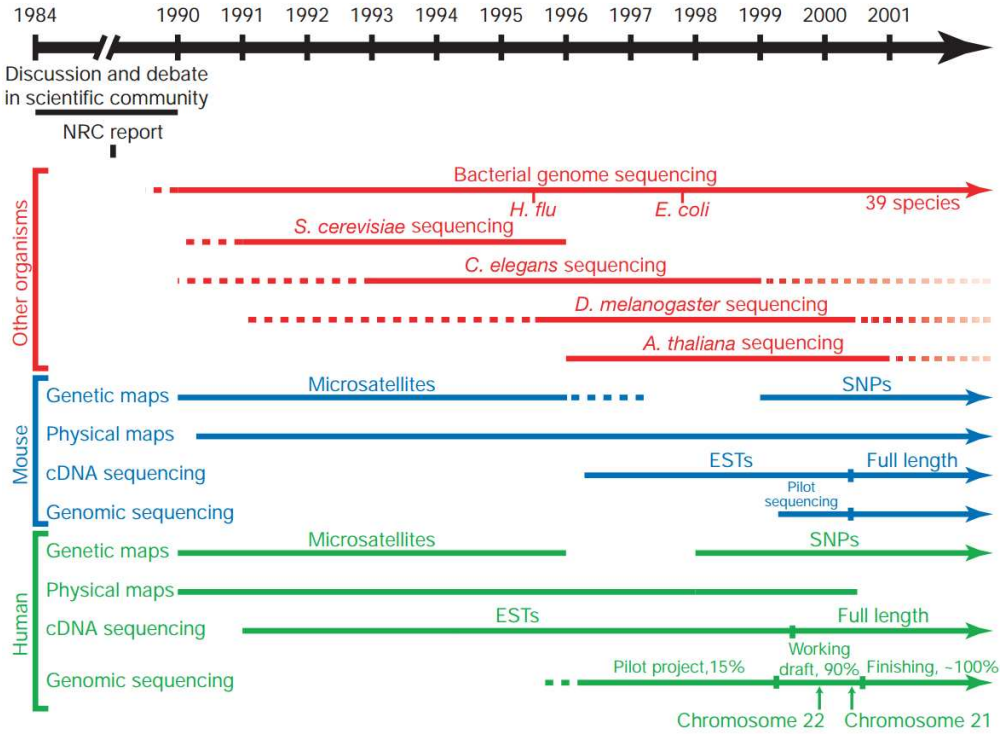
Fehérje-fehérje kölcsönhatási hálózatok moduljai

Előzmények

Az élő sejtek legváltozatosabb funkciójú molekulái a fehérjék. Meghatározó résztvevői az összes biológiai molekula típus építésének, módosításának és lebontásának. Az 1990-es évek második felére kialakult az a felismerés, hogy a fehérjék ezeket a feladatokat általában nem egyenként, hanem csoportosan (modulokban), együttműködve hajtják végre [3]. A molekulák együttműködő csoportjai kapcsán az is ismert, hogy bennük nem csak fehérjék, hanem a fehérjéktől eltérő típusú molekulák (például RNS-ek) is jelen vannak. A doktori dolgozat első fejezete fehérje-fehérje kölcsönhatásokkal és a fehérjék által alkotott funkcionális csoportokkal (modulokkal) foglalkozik. Ez a megkötés megfogalmazható úgy is, hogy az első fejezet a teljes (minden molekula típust tartalmazó) modulok helyett csupán a moduloknak a fehérjék által alkotott részhálózatát („vázát”) elemzi.

A természettudományok más területeihez hasonlóan a fehérje-fehérje kölcsönhatási hálózatok eredményeinek megismeréséhez is az első lépés a mérési (kísérleti) módszerek megismerése. A számítógépes és elméleti elemzések ezeknek a méréseknek az eredményeit használják. A fehérje-fehérje kölcsönhatási mérések szempontjából központi szerepű az élesztőgomba[1] nevű egysejtű faj. Az élesztőgomba a kutatás szempontjából egy „modell” faj (modell szervezet): (i) a fajról rendelkezésre álló részletes biokémiai és genetikai ismeretek következtében ebben a fajban dolgoztak ki kutatók számos olyan módszert, amelyek később más szervezetekben is alkalmazhatóak

[1] Latin neve *Saccharomyces cerevisiae*, angol neve „baker’s yeast”. A „sarjadzó” módon osztódó élesztő fajok közé tartozik. Szintén gyakran használt neve a „sörélesztő”.



1.1. ábra. Modell szervezetek DNS-ének és az emberi DNS szekvenálásának főbb évszámai. A felsorolt élőlények pontos latin nevei, zárójelben a magyar és angol teljes vagy egyszerűsített nevek: *Saccharomyces cerevisiae* (élesztőgomba, yeast), *Caenorhabditis elegans* (fonálféreg, nematode), *Drosophila melanogaster* (gyümölcslégy, fruit fly) és *Arabidopsis thaliana* (lúdfű, thale cress). Az ábrán szereplő rövidítések: SNP (Single Nucleotide Polymorphism), EST (Expressed sequence tag), cDNA (complementary DNA). Az ábra átvétel az [5] publikációból.

lettek, és (ii) az élesztőgombában talált eredményekből a kutatók gyakran következtetnek összetettebb élőlények működésére. Ennek megfelelően az eukarióta (sejtmaggal rendelkező) élőlények közül elsőként ennek a fajnak vált elérhetővé a teljes DNS szekvenciája. A széles körben használt „modell” élőlények és az ember DNS-ével kapcsolatos fontosabb mérföldköveket és évszámokat mutatja be (az előző három évtizedből) az 1.1. ábra.

Fehérjék kapcsolódását mérő kísérleti módszerek

A fehérje-fehérje kölcsönhatásokat mérő kísérleti módszerek között vannak olyanok, amelyek közvetlenül fehérjék fizikai összekapcsolódását („összetapadását”) tesztelik és vannak olyanok, amelyekből indirekt módon lehet következtetni a fizikai összekapcsolódásra. Mindkét esetben fontos különválasztani (i) a kölcsönhatásokat

egyesével („small-scale”) elemző kísérleteket és (ii) a sok kölcsönhatást egyetlen mérésben tesztelő kísérlet-sorozatokat. Utóbbiakat az egyszerre tesztelt nagy számú kölcsönhatás miatt szokás „large-scale” vagy high-throughput [2] (HTP) méréseknek is nevezni. A doktori dolgozat első fejezetében bemutatott saját eredmények 2004 és 2008 között készültek, ezért a következő leírásban az ezt megelőző években kidolgozott mérési módszerek nagyobb súllyal szerepelnek. Fontos, hogy az itt felsorolt általános mérési módszerek egyike sem rögzíti a fehérjéknek a kölcsönhatási képességén túl meglévő számos további tulajdonságát (amelyek befolyásolhatják a kölcsönhatásokat), például a foszforilációs vagy metilációs állapotot és a konformációt.

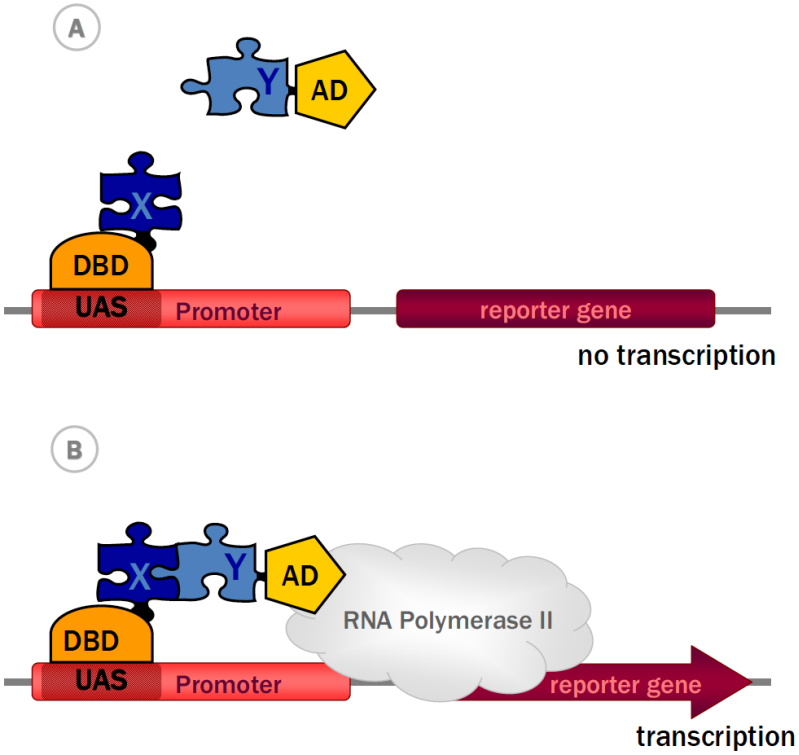
A fehérje-fehérje kölcsönhatások közvetlen kísérletes tesztelésének egyik leggyakrabban használt HTP (high-throughput) módszere a két-hibrid mérés („yeast-2-hybrid”, Y2H) [6, 7, 8, 9, 10, 11, 12], amelyben két rögzített fehérje összekapcsolódása esetén elindul egy harmadik (fluoreszcens) fehérje előállítása. A Y2H módszert mutatja be az 1.2. ábra. Ennek a módszernek előnye, hogy a mérésben ténylegesen össze kell tapadnia a két tesztelt fehérjének. Főbb hátrányai, hogy (i) a két tesztelt fehérje rögzítése (preparálása) miatt a mérés során a kapcsolódásukhoz nem áll rendelkezésre az in vivo esetben elérhető összes felszínük és (ii) az élesztőgombában történő mérés más élőlények fehérjéi számára a saját eredeti sejtjeikben meglévő in vivo molekuláris környezettől (például a pH, membránhoz való kötöttség és a kis molekula koncentrációk alapján) eltérő feltételeket jelenthet.

A fehérje-fehérje kölcsönhatások közvetett tesztelésének két gyakran használt „high-throughput” módja az ismételt affinitásos tisztítás [3] (TAP) [14, 15] és a fehérje komplex-ek[4] immunválasszal történő kiválasztása (Co-IP, protein complex immunoprecipitation). A yeast-two-hybrid módszerrel ellentétben a TAP és a Co-IP mérések azt tesztelik, hogy egy ismert fehérje melyik másik fehérjékkel van jelen azonos komplexben. Emiatt a TAP és a Co-IP eredményeként csak egy-egy fehérje komplex (fizikailag összekapcsolódott fehérjékből álló egység) fehérjéinek listája áll rendelkezésre a fehérjéknek a komplex-en belüli pontos kölcsönhatásai nélkül. Másképpen megfogalmazva: ha a TAP és a Co-IP eredménye az, hogy két fehérje összekapcsolódik egy

[2] A nagy áteresztőképesség (high throughput) itt a mérési folyamatból keletkező sok adat kapcsán az adatokat „előállító folyamat” nagy sebességére utal.

[3] A Tandem Affinity Purification (TAP) névben szereplő „tandem” szó arra utal, hogy a mérést két jelölő molekularészlet egymás utáni felismertetése teszi pontosabbá.

[4] A fehérje-fehérje kölcsönhatási hálózatok irodalmában gyakran „komplex”-nek nevezik a fizikailag összekapcsolódó (összetapadó) fehérjék egy csoportját, és „modul”-nak vagy „funkcionális modul”-nak nevezik a szorosan együttműködő fehérjék egy csoportját.



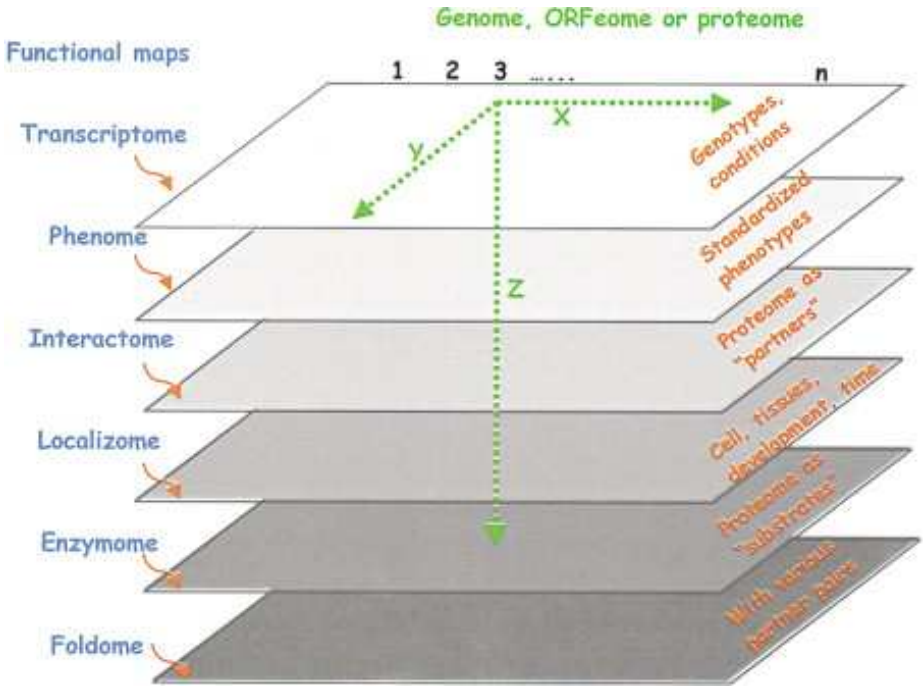
1.2. ábra. A klasszikus yeast-2-hybrid (élesztő-két-hibrid) kísérleti rendszer alkalmas nagy számú fehérje-fehérje kölcsönhatás automatizált (high-throughput, HTP) mérésére. (A) A vizsgált X fehérje és a vele összekapcsolt DNS-kötő fehérje domén (DBD) neve együtt „bait” (csali). A mérés azt teszteli, hogy az Y fehérje képes-e hozzákapcsolódni az X fehérjéhez. Az Y fehérje és a vele összekapcsolt aktivációs domén (AD) neve együtt „target” (cél fehérje). Tehát a „csali” elnevezés arra utal, hogy a kísérletező a DBD-X egység segítségével próbál „kihalászni” olyan fehérjéket, amik az X fehérjéhez képesek kapcsolódni. (B) A DBD-X egység (fúziós fehérje) hozzákapcsolódik a jelző gén (reporter gene) leolvasását serkentő DNS szakaszhoz (UAS, upstream activator sequence of promoter). Ha az Y fehérje hozzátapad az X fehérjéhez, akkor az AD aktiváló domén RNS polimeráz kötési képessége nyomán elindul a jelző gén átírása. Az ábra átvétel a [13] publikációból.

komplex-en belül, akkor ebből a kapcsolódásból nem következik az, hogy a két fehérje a komplex-en kívül (önálló párként) is képes összekapcsolódni. Az irodalomban a kísérletek alapján ismert komplex-ekből pár kölcsönhatásokra kétféle módon szokás következtetni. Az első lehetőség, hogy feltesszük azt, hogy a komplex-en belül az összes lehetséges fehérje-fehérje pár kölcsönhatásban áll. A második lehetőség, hogy azt feltételezzük, hogy a komplex azonosítására használt „csali” (bait) fehérje kölcsönhatásban van az összes többi fehérjével, de más kölcsönhatás nincsen a komplex-en belül. Ezt a két módszert a teljes összekapcsoltság alapján "matrix" illetve a csillagszerű

(küllő-szerű) kapcsolatok alapján "spoke" módszernek szokták nevezni [16].

Összefoglalva: az eddig említett három kísérleti módszer közül csupán egy (a yeast-2-hybrid) teszteli közvetlen módon két fehérje összetapadását, a másik két módszer (a TAP és a Co-IP) „csupán” azt teszteli, hogy a két fehérje előfordul-e egy komplex-en (fizikailag összetapadt fehérje csoporton) belül. Természetesen a felsorolt módszerek kidolgozása óta a mérések sokat fejlődtek, és a fehérje komplex-ek felsorolása az élesztőgombában folyamatosan növekvő pontosságú méréseket tesz lehetővé [17, 18]. A felsoroltakon túl további nagyskálájú közvetlen mérési lehetőségeket biztosítanak például a fehérje csipek [19] és a tömegspektrometria közvetlen alkalmazása a fehérje komplex-eket alkotó fehérjék azonosítására (HMS-PCI) [20]. Szintén jelentős fejlemény, hogy napjainkra – technikai okok miatt – a két-hibrid rendszerű mérésekben az élesztő mellett (helyett) elterjedt az *Escherichia coli* baktérium használata [21]. A kísérleti technológiák fejlődése miatt az elmúlt években már összetett soksejtű szervezetek fehérje komplex-eiről is készültek részletes mérések [22].

Az összes felsorolt – fehérje kölcsönhatásokat mérő – módszer esetében a mérések után kapott adatokat érdemes szűrni azért, hogy az aspecifikus kölcsönhatások az elemzésekben ne jelenjenek meg. Az aspecifikus kölcsönhatások kiszűrésnek a biológiai jelentése az, hogy például a fehérjéket előállító riboszómával, a fehérjéket „javító” dajkafehérjékkel („chaperon” fehérjékkel) és a fehérjék lebontását végző fehérjékkel való kölcsönhatásokat az adatokból eltávolítjuk. Ezekkel majdnem minden fehérje kölcsönhatásba tud lépni, ezért biológiai szempontból a velük való kölcsönhatás ténye nem informatív. A legtöbb kölcsönhatással rendelkező fehérjék kiszűrése egyszerűnek tűnhet, ám mégis jelentős szisztematikus hibát vihet bele az adatokba. Sok esetben ugyanis nem lehet egyértelműen azonosítani a fehérjék közül azokat, amelyek a többinél lényegesen több, de gyenge kölcsönhatással rendelkeznek. Nem lehet megmondani, hogy hol van a „legnagyobbak” alsó határa. Ennek oka az, hogy – más típusú hálózatokhoz hasonlóan – gyakran a molekuláris biológiai kölcsönhatási hálózatokra is jellemző a skálafüggetlen fokszám eloszlás [23]. Azaz, ha a fehérjéket sorba rendezzük a velük kölcsönható partnerek száma (a fehérjéket jelölő hálózati csúcspontok fokszáma) alapján, akkor egy folytonos, hatványfüggvény-szerű eloszlást kapunk, amiben nincsen a „nagyok” és a „kicsik” közötti éles határ. Másként megfogalmazva: a fehérje-fehérje kölcsönhatási hálózatokban a fehérjék kapcsolat számának eloszlása általában nem bimodális (egy fehérjének vagy kevés vagy sok kapcsolata van), hanem folytonos, ezért nem lehet az eloszlást „szétválasztani”.



1.3. ábra. A 2000-es évek elejétől kezdődően a DNS-re utaló „genome” szóhoz hasonlóan megjelentek a szakirodalomban az RNS-re utaló „transcriptome”, a fehérje kölcsönhatásokra utaló „interactome” szavak és további „ome” végű kifejezések. Az ábra átvétel a 2001-ben megjelent [24] publikációból.

Fehérjék kapcsolódását közvetve mérő módszerek

Az előző alfejezet bemutatta azokat a főbb kísérleti módszereket, amelyekkel két fehérje összekapcsolódása vagy két fehérje egy komplex-ben történő részvétele mérhető. Ezeken a közvetlen módszereken túl az irodalomban számos más módszer is ismert, amellyel közvetett módon fehérjék közötti kapcsolatokra lehet következtetni. Fontos, hogy ezeknél a módszereknél már nem csupán két fehérje közötti fizikai kölcsönhatásról van szó, hanem funkcionális kapcsolatokról is. A funkcionális kapcsolatok felhasználása miatt az ilyen eredmények összesítése után kapott kölcsönhatás listát PPI (protein-protein interaction) hálózat helyett szokás PPA (protein-protein association) hálózatnak nevezni. Szintén gyakori elnevezés – a „genome” szó mintájára – az „interactome” is [24] (ld. 1.3. ábra). A 10-15 éve elterjedt „interactome” szónak többféle definíciója létezik, ezek közül a leggyakrabban használt definíció a következő: a gének által kódolt fehérjék közötti összes lehetséges kölcsönhatás hálózata [25]. Bármilyen konkrét helyzetben (például sejt állapotban) ezeknek a kölcsönhatásoknak csak egy részhalmaza valósul meg.

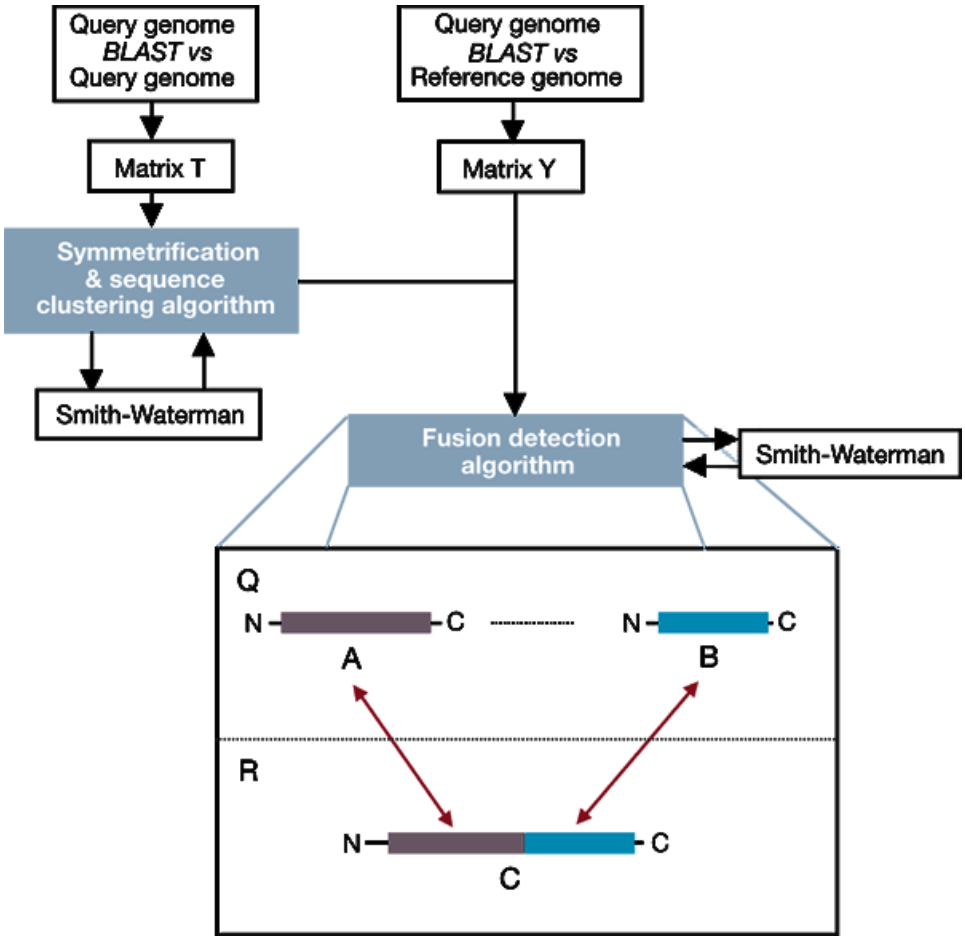
A közvetett kölcsönhatást (funkcionális kapcsolatot) mérő módszerek közül a leghosszabb ideje a DNS szekvenciát felhasználó módszereket használják. Ezek a módszerek egynél több faj (élőlény) génjeit hasonlítják össze, általában faj páronként. Az összehasonlításhoz szükség van szekvencia illesztési módszerekre: az egzakt módszerek közül az egyik gyakran használt a Smith-Waterman [26], a heurisztikus módszerek közül a legnépszerűbbek a BLAST [27] és változatai. A BLAST szekvencia illesztés segítségével egy F_1 faj g_1 génjét és egy F_2 faj g_2 génjét szokás egymásnak megfeleltetni, ha ez a két gén egy „Bidirectional Best Hit”-et (BBH) alkot. A BBH definíciója a következő: a g_1 génhez az F_2 élőlény génjei közül leginkább hasonló a g_2 és a g_2 génhez az F_1 élőlény génjei közül leginkább hasonló a g_1. Így például a biológiában közismert $p53$ gén (a sejtciklus, a DNS javítás és a programozott sejthalál egyik irányítója) megfelelői számos fajban megtalálhatóak.

A „Rosetta kő” – más néven fúziós fehérje – módszer azt vizsgálja, hogy az F_1 fajban található a_1 és b_1 gének esetében előfordul-e az, hogy valamelyik másik F_2 fajban az a_1 és b_1 génnek megfelelő a_2 és b_2 gének egymással összeolvadva egyetlen gént alkotnak. Ha igen, akkor ez alapján a módszer [5] arra következtet, hogy az F_1 fajban az a_1 és b_1 gén egymással funkcionális kapcsolatban van [28]. Minél több ilyen F_2 fajt találunk (ahol az a és b gén összeolvadt egymással), annál nagyobb az esély arra, hogy az F_1 fajban is funkcionális kapcsolat van köztük (ld. 1.4. ábra).

A fúziós fehérje módszerhez igen hasonló, de annál valamivel általánosabb kritérium, hogy két gén számos élőlényben a DNS-ben egymáshoz közel legyen [30]. Két gén funkcionális kapcsolatának megállapítására a fúziós fehérje módszerhez szintén hasonló, de még általánosabb (enyhébb) kritérium a filogenetikus profilok összehasonlítása. A filogenetikus profil módszer szerint ha minden génhez bináris vektorként felírjuk, hogy egy fajban a gén jelen van (1-es érték) vagy nincsen jelen (0 érték), akkor az egymáshoz hasonló filogenetikus vektorok hasonló funkciójú géneket jelölnek.

A fehérjék közötti funkcionális kapcsolatok feltárására az irodalomban ismertek olyan módszerek is, amelyek a fehérjékhez tartozó mRNS-ek mért koncentrációit használják. Ebben az esetben ismét minden egyes génhez egyetlen mRNS-t és egyetlen fehérjét rendelünk hozzá, tehát elhanyagoljuk például az egyetlen génből többféle mRNS-t előállítani képes „alternative splicing” jelenséget és a transzláció utáni fehérje

[5] A módszer neve a rosettai kő nevű ókori táblára utal, amelyen azonos szöveg három különböző írás rendszerrel leírva szerepel, és a három írás rendszer közül az egyiket a táblán lévő másik kettő segítségével sikerült részben „megfejteni”.



1.4. ábra. A fúziós fehérje módszer bemutatása a [29] publikációból. A módszer a vizsgált Q fajban (Query genome) található géneket és az R fajban (Reference genome) található géneket hasonlítja össze az egzakt Smith-Waterman [26] és a heurisztikus BLAST [27] szekvencia-illesztési algoritmus felhasználásával. A T mátrix a Q fajban található géneket hasonlítja össze egymással, míg az Y mátrix a Q faj génjeit az R faj génjeivel hasonlítja össze. A bekeretezett panel felső része mutatja a Q fajban található A és B gént (a gének végpontjainak neve mindenütt N és C). Ugyanennek a panelnek az alsó része mutatja az R fajban található C gént, amely a szekvenciák hasonlósága alapján az A és B gén fúziójaként értelmezhető. Az így megtalált fúzió és a szekvenciák hasonlósága miatt valószínűsíthető, hogy az A és B gén között nem csak az R fajban van funkcionális kapcsolat, hanem a Q fajban is.

módosításokat[6]. A génhez tartozó mRNS koncentrációját – megfelelő normálással – szokás (mRNS) expressziós szintnek is nevezni. Sok, egymás utáni kísérlet elemzése esetén expressziós mátrix-nak szokás nevezni azt a mátrix-ot, aminek az (i, j) eleme

[6]A fehérjék transzláció utáni módosításának (Post-Translational Modification, PTM) két gyakori formája a foszforiláció és a konformáció módosítás.

az $i.$ génnek a $j.$ kísérletben mért expresszióját (a gén mRNS-ének koncentrációját) tartalmazza. Mindezt felhasználva akkor következtethetünk két fehérje funkcionális kapcsolatára, ha a két fehérje expressziós vektora (a mátrix-ban hozzájuk tartozó sorok) erős korrelációt mutatnak. Két fehérje erős kapcsolata nem csak úgy jelenhet meg egy sejtben, hogy a két fehérjéhez tartozó mRNS koncentrációk azonos módon változnak, hanem úgy is, hogy pontosan ellentétesen változnak. A korreláció csak az azonos változások esetén magas, míg a korreláció abszolút értéke az azonos és ellentétes változások esetén egyaránt magas. Emiatt két fehérje erős funkcionális kapcsolatát az mRNS-eik expressziójának korrelációja helyett megfelelőbb mérni a korreláció abszolút értékével.

Két gén eltávolítása utáni életképesség mérésén alapszik a „synthetic lethality" nevű módszer, aminek szintén elterjedt nevei az SGA (synthetic genetic array) [31], a „genetic interaction" és a „double knock-out". A módszert főként élesztőgombában használják. Első lépésként a kísérletező az élesztőgomba minden egyes génjéhez (a gyakorlatban 6000 és 7200 közötti gén) előállít egy élesztőgomba törzset (mutánst), amelyből pontosan az az egy gén hiányzik az eredeti élesztőgombához (vad típus, wild type, WT) képest [32]. Ezt a gén eltávolítást az irodalomban gyakran nevezik gén „kiütésnek" (gene knock-out). Egy gén kiütése után általában életképes marad az élesztőgomba, például a $zds1$ gén törlése után életképes marad, ezt jelzi a http://www.yeastgenome.org/cgi-bin/locus.fpl?dbid=S000004886 weblapon a „Mutant phenotype" kategóriában szereplő „viable" bejegyzés. A synthetic lethal kölcsönhatás definíciója két gén együttes kiütése utáni állapoton alapszik. Ha külön-külön a g_1 és a g_2 gén kiütése után az élőlény életképes marad, de a g_1 és g_2 együttes kiütése után nem, akkor valószínűsíthető, hogy a két gén egymás kiesésének pótlásában szerepet játszik. Ilyen esetben a (g_1, g_2) gén pár között synthetic lethal kölcsönhatás van. Ez a kölcsönhatás gyakran párhuzamos (egymást pótolni képes) útvonalakban való részvételt jelent vagy egy komplex-ben azonos helyre való kötési képességet. A módszer pontosabbá tehető úgy, hogy a kísérletező az eredeti törzsben (más néven: vad típus, wild type, WT) és minden egy- és két gén kiütéses törzsben megméri az élesztőgomba populáció osztódási sebességét. Így részletes kép kapható arról, hogy az élesztőgombában (majdnem) tetszőlegesen kiválasztott két gén együttes kiütése a két gén különkülön történő kiütéséhez képest pontosan mennyivel erősebb vagy gyengébb hatással jár [29].

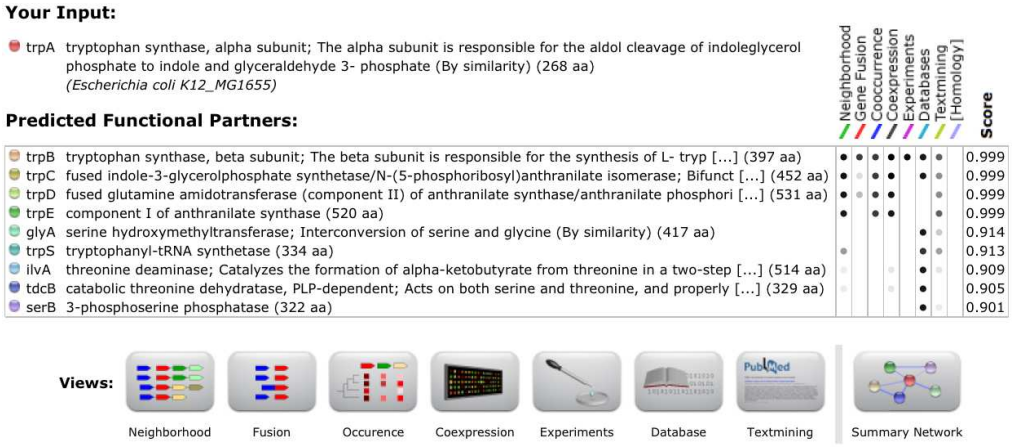
A fehérjék közötti funkcionális kapcsolatok mérésére szintén elterjedt módszer az élettudományi szakirodalom automatizált feldolgozása (gyakori neve: „literature mining”). Ennek a módszernek egyik változata alapján két fehérje egymással funkcionális kapcsolatban van, ha a két fehérje neve együtt szerepel megfelelő számú cikk kivonatában. Ez a kölcsönhatás előrejelzési mód alacsony találati arányt ad, de a keresés sok cikkre alkalmazható automatizáltan [33].

Fehérje-fehérje kölcsönhatási (PPI) hálózatok összeállítása

A fehérjék kapcsolati hálózatának összeállítása több lépésből áll [30, 34, 35, 36]. Az eltérő forrásokból rendelkezésre álló és eltérő biológiai jelentésű adatok integrálását szemlélteti az 1.5. ábra. A végső adatok információ tartalmának és hiba forrásainak ismerete érdekében hasznos a PPI[7]. hálózatok összeállításának egymás utáni lépéseit áttekinteni.

A sokféle típusú és jelentésű mérési eredmény összesítéséhez az első lépés a mérésekben szereplő fehérjék (gének) neveinek átalakítása azonos név típusra. Minden gén (és a hozzá tartozó fehérje) többféle névvel rendelkezik. Ez részben amiatt van, mert az eredetileg különböző jelenségeknél látott fehérjékről idővel kiderülhet, hogy azonos fehérjéről van szó, részben amiatt, mert többféle adatbázis katalogizálja a géneket/fehérjéket. Példaként érdemes ránézni az élesztőgomba ZDS1 nevű fehérjéjének adatlapjára a közismert UniProt adatbázisban található adatlapon a következő címen (a doktori dolgozat PDF fájljában ez a hiperlink klikkelhető): http://www.uniprot.org/uniprot/P50111. A fehérje ajánlott neve (recommended name, ZDS1) mellett használatban van két „alternative name” (NRC1 és RT2GS1), a gén névhez szerepel 6 szinonima, egy hely (locus) név a DNS-ben (ordered locus name) és egy leolvasási szakasz név (ORF, Open Reading Frame). Ezeken túl még gyakran előfordul az élesztőgomba adatbázis (Saccharomyces Genome Database, SGD) által használt S000004886 elsődleges azonosító (Primary SGDID), az NCBI Protein adatbázisban használt 1709345 GI (Gene ID) és a ZDS1_YEAST azonosító. Ebből a konkrét példából látható, hogy egy-egy kiértékelés során az összes fehérje/gén név azonos név típusra történő átalakításához szükség van speciális, erre a célra írt programokra (a programokkal készített, folyamatosan frissített konverziós táblákra). Az itt bemutatott saját munka során a fehérje név átalakításokat magam végeztem erre a célra írt Perl

[7] A szakirodalomban gyakori hogy a „PPA” (Protein-Protein Association) hálózatokat is „PPI” (Protein-Protein Interaction) hálózatnak nevezik.



1.5. ábra. Egy fehérje-fehérje kapcsolatokat kereső és elemző online szolgáltatás (STRING) eredményei táblázatos formában. A kereső a http://string.embl.de helyen található és az ábra a weboldalon 1. példaként felkínált fehérjére (az *E. coli* egysejtű trpA fehérjéje) történő keresés eredményét mutatja. A kereső az ábra alján felsorolt 7 adat forrás mindegyikéből megkeresi a trpA fehérjéhez kapcsolódó más fehérjéket és az eredményeket összesíti hálózatos formában. Az adat források (zárójelben a táblázat jobb felső részében található oszlop címkék): gének közelsége (Neighborhood), összeolvadása (Gene fusion) és közös előfordulása fajokban (Cooccurrence, ez a korábban leírt phylogenetic profile módszer), több kísérleti körülmény (mint változó) esetére ismert mRNS szintek korrelációja (Coexpression), részletes biokémiai kísérletek (Experiments), más adatbázisok (Databases), élettudományi publikációk szövegének automatizált elemzése (Textmining). A táblázat alatt látható ikonok (a „Views” címkétől jobbra) közül az első három ikonon az egymás alatt található azonos szimbólumok egymásnak megfelelő géneket jelölnek eltérő szervezetekben.

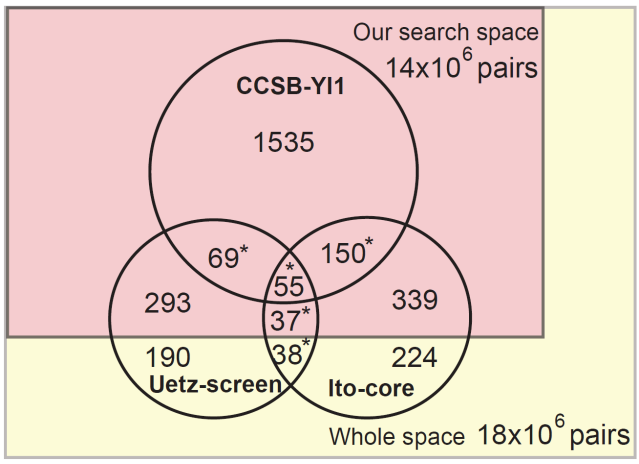
programjaim segítségével. A későbbiekben is naponta használt Perl programozási ismereteket 2004 óta folyamatosan oktatom, a kurzus honlapja http://hal.elte.hu/fij/perl. Ezen a címen elérhetőek a folyamatos órai házi feladatok, a házi feladatok hallgatói megoldásai és az általam összesített javasolt megoldások.

Amint a kutató a felhasználni kívánt összes adatsor fehérje neveit azonos név típusra konvertálta, a következő kérdés az, hogy hogyan lehet összehasonlítani a különböző mérés típusok (például a yeast-2-hybrid és a fúziós fehérje módszer) adatainak minőségét. Azaz, melyik mérés típus milyen arányban képes előrejelezni (jósolni) az adott szervezetben a fehérje-fehérje kölcsönhatásokat[8]. Ehhez az összehasonlításhoz először rögzíteni kell, hogy pontosan melyik fehérje-fehérje kölcsönhatások valódiak (helyesek). A pozitív és a negatív minta halmaz természetesen diszjunkt és az uniójuk

[8]Természetesen a találati arányon túl az is fontos, hogy egy mérés milyen típusú kölcsönhatásokat képes magas találati aránnyal előrejelezni és milyen kölcsönhatásoknál alacsony a találati aránya.

neve „gold standard” (rövidítve GS) halmaz. Az irodalomban az egyszerűség kedvéért gyakran csak a pozitív minta halmazt rögzítik és a pozitív minta halmazban található összes fehérje között elvileg lehetséges további kölcsönhatásokat tekintik negatív minta halmaznak. A biztosan létező (pozitív minta) és a biztosan nem létező (negatív minta) kölcsönhatások listáinak rögzítése után összehasonlítható bármely kettő (vagy több) módszer az alapján, hogy (i) hány olyan kölcsönhatást jósolt, ami a pozitív minta halmazban van és (ii) hány olyan fehérje-fehérje kölcsönhatást jósolt, amelyik a negatív minta halmazban van. Előbbiek alkotják „true positive” (TP) halmazt, utóbbiak a „false positive” (FP) halmazt. Néhány további, gyakran használatos fogalom a „precision” (a TP és a FP halmazok uniójában található összes jóslatból mekkora hányad TP, ez más néven a Positive Predictive Value, PPR) a „recall” (a TP halmaz mérete osztva a pozitív halmaz méretével, ez más néven a True Positive Rate, TPR) és a „False Positive Rate” (FPR, az FP halmaz mérete osztva a negatív minta halmaz méretével). A PPI kölcsönhatásokat előrejelző (jósoló) módszerekben a jósolt összes kölcsönhatás száma általában szabályozható egyetlen egyszerű paraméterrel. Ennek a paraméternek a neve gyakran „threshold” (küszöb érték), vagy „stringency” (szigorúság). és a változtatása során többek közt a TPR és az FPR mennyiségek szintén változnak. A TPR-t az FPR függvényében ábrázolva egy gyakran használt függvényt kapunk, a Receiver Operating Characteristic (ROC) függvényt, amely – több más felhasználása mellett – alkalmas a PPI kölcsönhatásokat jósoló módszerek egymással való összehasonlítására.

A mérésekből kapott kölcsönhatás listák általában még azonos kísérleti módszer esetén is jelentősen eltérhetnek (ld. 1.6. ábra). Ezért a gyakorlatban a fehérje-fehérje kölcsönhatások Gold Standard (referencia) listája általában egy megbízhatónak vélt – de mégis a mérés kiértékelői által önkényesen vagy szokás alapján kiválasztott – adatsorban található reakciókat jelenti. Így például gyakori referencia kölcsönhatás lista a MIPS (Munich Information Center for Protein Sequences) [38] vagy a KEGG (Kyoto Encyclopedia of Genes and Genomes) [39] adatbázis által a kiválasztott élőlényben publikált kölcsönhatás halmaz, vagy egy konkrét projektben „kézzel” (nem automatizálva, keresőprogramokkal) a szakirodalomból összegyűjtött kölcsönhatások listája. Az itt említett három Gold Standard példa (két adatbázis és a kézi gyűjtés) kapcsán természetesen felmerül az az általános kérdés, hogy melyik típusú mérések „jósolják” pontosabban (magasabb precision és recall értékekkel) az ismert fehérje-fehérje



1.6. ábra. Azonosságok és eltérések három különböző yeast-2-hybrid mérés sorozat eredményei között, amelyek az élesztőgomba fehérje-fehérje kölcsönhatás listáját („interactome”-ját) mérték. Az ábra átvétel a [10] publikációból. Az „Our search space” nevű halmaz a [10] publikáció szerzői által megvizsgált 14×10^6 fehérje párt jelöli. Ebből a „CCSB-Y1” (Center for Cancer Systems Biology - Yeast Interactome 1) halmaz által jelölt számú fehérje pár között találtak kölcsönhatást. A „Whole space” halmaz az összes lehetséges fehérje-fehérje párt jelöli, amelyek között elvileg előfordulhat kölcsönhatás. Az „Uetz-screen” halmaz a [37] publikáció eredményeinek egy részhalmaza és az „Ito-core” a [7] publikáció eredményeinek egy részhalmaza. Érdemes megfigyelni, hogy az azonos technológia (yeast-2-hybrid) ellenére a három mérés sorozat eredményei jelentősen eltérnek. Ennek egyik oka az, hogy a mérési módszer igen sok paramétertől függ és sok lépésből áll, amelyeket nem lehet teljesen pontosan rögzíteni.

kölcsönhatás listákat: (i) a részletes („small scale”) kísérletek, amelyek egy-két kölcsönhatásra fókuszálnak, vagy (ii) a nagyskálájú (high-throughput, HTP) kísérletek, amelyek azonos kísérleti eszközzel párhuzamosan tesztelik több ezer vagy akár több tízezer kölcsönhatás létezését. A kétféle megközelítés között alapvető különbségek vannak, amelyek meghatározzák előnyeiket és hátrányaikat. A részletes (small scale) kísérleteket egy-egy konkrét, tudományosan megalapozott hipotézis vezérli, és a sokféle tesztelés amit tartalmaznak, csökkenti a szisztematikus hibákat. A nagyskálájú (large scale) adatgyűjtésre optimalizált mérési módszerek legfőbb hátránya az, hogy minden kölcsönhatást azonos (egységesített, uniformizált) módon próbálnak azonosítani és emiatt az eredményeikben magas a hibák aránya. A nagyskálájú kísérleteknek előnye, hogy a céljuk a mérőeszközök által elérhető összes kölcsönhatást rögzíteni hipotézisek és előzetes várakozások nélkül elfogulatlanul. Általánosságban megállapítható, hogy a high-throughput kísérletek számának és minőségének fejlődésével a

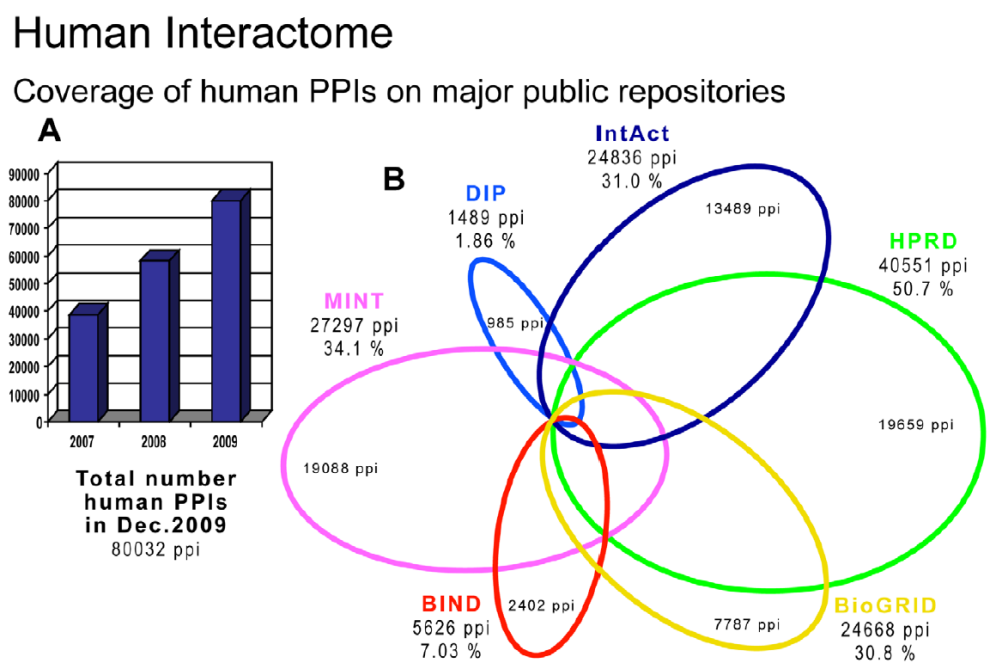
kapott HTP adatok minősége a small-scale kísérleti adatok minőségével összemérhetővé válhat [40, 41].

A Gold Standard kiválasztása és az egyes adat típusok minőségének megállapítása (kalibrálás) után a következő lépés az adatok összefésülése (integrálása) egyetlen fehérje-fehérje kapcsolat listába. Az eredmény lista minden egyes eleme egy (irányítatlan) fehérje-fehérje pár lesz, amihez egyetlen számot rendelünk hozzá annak jelzésére, hogy az adott fehérje-fehérje kölcsönhatásnak a felhasznált adat típusok és Gold Standard alapján mekkora az erőssége (valószínűsége, jelentősége). Ha az A–B fehérje pár esetén az $1, 2, 3, \ldots$ indexű adat forrásokból ismert kölcsönhatás erősségek $w_1, w_2, w_3, \ldots (0 < w_i < 1)$, akkor ezeknek a kölcsönhatás erősségeknek (valószínűségeknek) az összesítésére egy gyakran használt módszer a w_i értékekből súlyozott összeg képzése. Ennek egy konkrét megvalósítása szerepel részletesen leírva a [36] publikáció kiegészítő anyagának 7-11. oldalán. Első lépésként a szerzők minden A–B párhoz kiszámítják a w_i erősségek súlyozott összegét egy olyan paraméterrel, aminek a változtatásával lehet a nagy w_i értékeket még jobban kiemelni vagy a többi súlyhoz közelítve csökkenteni. Ezzel a paraméterrel állíthatja a kiértékelést végző kutató, hogy a magas True Positive Rate (vagy esetleg a jó ROC) alapján nagy megbízhatóságúnak talált adatok típusokban a súlyuknál még erősebben bízik vagy pont fordítva, a korábban kiszámított súlyuknál kevésbé bízik meg bennük. Sajnos előfordul olyan eset is, amikor egy széles körben használt fehérje-fehérje kölcsönhatási adatbázis kölcsönhatás konfidencia értékeinek kiszámítási módja gyengén dokumentált, és többszöri kérdezés után is csak igen nehezen elérhető [42].

Az előző bekezdésben leírt összesítési módszerekkel kapott fehérje-fehérje kölcsönhatások listája elérhető több adatbázisban. Ezek az adatbázisok általában vagy egy/néhány faj kölcsönhatásait sorolják fel[9], vagy sok faj esetén minden egyes fehérjéhez külön felsorolják, hogy az adott fehérje melyik másik fehérjékkel kapcsolódik[10]. Természetesen a felsorolt adatbázisok közül a legtöbb a fehérje-fehérje kölcsönhatásokon túl számos további információt is tartalmaz a vizsgált sejt típusokról és/vagy élőlény(ek)ről. Szintén fontos megemlíteni, hogy a felsorolt molekuláris biológiai adatbázisok közül a legnagyobbak (például az NCBI adatbázisai és a UniProt) fehérje-fehérje

[9] Egy vagy néhány élőlény fehérje-fehérje kölcsönhatásait sorolja fel például a HPRD / Human Proteinpedia [44], egy emberi fehérje komplex lista [22], a Human Protein Atlas [45], a WormBase [46], a FlyBase [47] és a „CCSB Interactome" nevű adatsorok több élőlényből [9, 10, 12, 48].

[10] Ilyenek például a STRING [30], UniProt [49], NCBI Protein [50], DIP [51], IntAct [52], BioGrid [53], BIND [54] és a MINT adatbázis [55]



1.7. ábra. Emberi fehérje-fehérje kölcsönhatás listák összehasonlítása. Az ábra fejlécében szereplő „coverage” a True Positive-ok számát jelenti. Az ábra átvétel a [43] publikációból, az adatok 2009 decemberi és korábbi állapotot mutatnak. (A) A teljes emberi PPI (fehérje-fehérje kölcsönhatási) hálózatban található kapcsolatok száma 2007 és 2009 között. (B) Az egyes adatbázisokban tárolt kölcsönhatok száma és az adatbázisok összesítése után kapott 80032 kölcsönhatáshoz képest számított százalékos értékek.

kölcsönhatási listáit jelentős részben a kisebb adatbázisok adatainak összesítése adja (a forrás pontos feltüntetésével). Néhány főbb adatbázis összehasonlítását mutatja az 1.7. ábra. Az adatbázisok összehasonlításában fontos az is, hogy a legtöbb és legrészletesebb ingyenes adatot – eddigi tapasztalatom szerint – az NCBI (National Center for Biotechnology Information) és a UniProt (Universal Protein Resource) ftp oldalai biztosítják[11]. Az adatok feldolgozásához két jelentős fejlesztői szoftver forrás a CPAN (Comprehensive Perl Archive Network) és a BioPython projekt[12].

[11] Az NCBI www és ftp site-ja sok adatbázis összesítését tartalmazza. A UniProt kisebb, de szintén sokféle információt tesz ingyenesen elérhetővé.

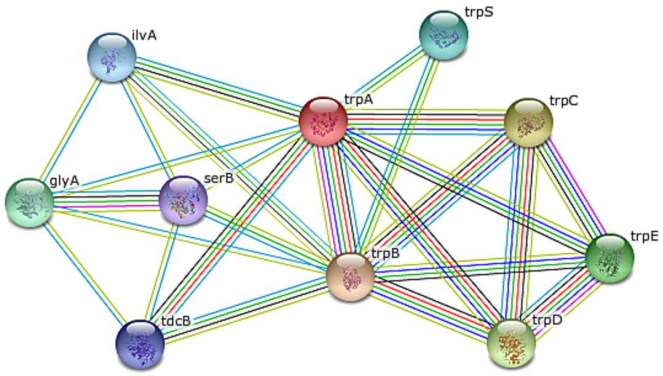
[12] A Perl a Pythonnál korábbi nyelv, és a környezet-függőség (context dependence) használata miatt az emberi nyelvekre jobban hasonlít. Többek között a szekvencia elemzésben és adatfeldolgozásban erős. A Python fiatalabb és kötöttebb struktúrájú nyelv. Többek között a szekvencia elemzésben és filogenetikában erős. Mindkét nyelv a „scripting” típusú nyelvek közé tartozik és gyakoriak a számítógépes rendszerek karbantartásában is. A TIOBE index alapján a Perl és a Python egyaránt túl van az összes programozási nyelvhez képest számított relatív elterjedtségi csúcsán. Mindkettő érett nyelv és széles körben használják.

PPI hálózatok szerkezete („topológiája”)

A leíráshoz használt matematikai eszközök

A fehérje-fehérje kölcsönhatási hálózatok szerkezetének leírásához szükséges matematikai alapfogalmak közül a két legegyszerűbb fogalom a hálózatban található pontok és kapcsolatok száma. A hálózat pontjait szokás csúcspontnak vagy csúcsnak is nevezni (angolul „node” vagy „vertex”). Ennek megfelelően a doktori dolgozatban a csúcsok számát N jelöli, és az irodalomban előfordul az N_V jelölés. A kapcsolatok gyakori megnevezése él („link” vagy „edge”). A dolgozatban az élek számát E jelöli, az irodalomban előfordul az N_E és néha az L jelölés. Egy csúcs kapcsolatainak száma az adott csúcs fokszáma, a dolgozatban az i. csúcs fokszámát k_i jelöli. Az irodalomban gyakori a d_i jelölés és előfordul a z_i jelölés is. Egy hálózatban minden él 2-vel növeli a csúcsok fokszámainak összegét, ezért a csúcsok fokszámainak átlaga $\langle k \rangle = 2E/N$ Az i. csúcspont tetszőleges két szomszédja közötti kapcsolat valószínűsége az i. csúcs (lokális) klaszterezettségi (más néven „csomósodási”) együtthatója, amelyet C_i-vel szokás jelölni. A C_i mérése a gyakorlatban (egy számítógépes programmal) úgy történik, hogy az i. csúcs szomszédjai között ténylegesen meglévő kapcsolatok számát (n_i) elosztjuk az i. csúcs szomszédai között maximálisan lehetséges kölcsönhatás számmal: $C_i = 2n_i/[k_i(k_i - 1)]$. Egy hálózat klaszterezettségét a szakirodalomban kétféle módon szokás definiálni. Az első (gyakrabban használt) definíció szerint egy hálózat klaszterezettsége a csúcspontok lokális klaszterezettségi együtthatóinak átlaga: $C = \langle C_i \rangle_{i=1\ldots N}$. A második (ritkábban használt) definíció szerint egy hálózat klaszterezettségét a háromszögek száma (N_h) és a legalább egy éllel összekötött csúcs-hármasok (tripletek) száma (N_t) a következő módon határozza meg: $C = 3N_\mathrm{h}/N_\mathrm{t}$.

A fokszámnál és a klaszterezettségnél nagyobb léptékű szerkezeti információt ad két csúcspont átlagos távolsága és a hálózat átmérője. Egy hálózat két csúcspontjának távolsága a kettőjüket összekötő legrövidebb útvonal hossza (az útban lévő élek száma). A dolgozatban az i. és a j. csúcspont távolságát $d_{i,j}$ jelöli, és a hálózat átmérőjét L jelöli. Az átmérőt a matematikai irodalomban $L = \max_{i,j}(d_{i,j})$ módon szokás definiálni, egy ettől kissé eltérő definíció a maximum helyett az átlagot használja: $L = \langle d_{i,j} \rangle$. Az átmérő mindkét definíciója feltételezi, hogy a gráf összefüggő, azaz egyetlen komponensből áll (másképpen fogalmazva: tetszőleges két csúcspontja



1.8. ábra. A STRING adatbázisban az 1.5. ábrán bemutatott példa keresés eredményeként kapott fehérje-fehérje kölcsönhatási részhálózat. Ebben a részhálózatban a trpA fehérje (az ábra közepén lévő csúcspont) betweenness-e magas, mert sok csúcspár közötti legrövidebb út tartalmazza ezt a csúcspontot. Természetesen két csúcspont között lehet egynél több legrövidebb út, például az $(A-B, A-C, B-D, B-C)$ élekből álló gráfban az A és B csúcsok között két legrövidebb (2 lépés hosszúságú) út van. További megjegyzés: az ábrán két csúcspont között általában több (eltérő színű) él látható. Minden él szín az 1.5. ábrán felsorolt 7-féle adat forrás valamelyikéből származó kapcsolat a két összekötött fehérje között.

között létezik út). Szintén a kölcsönhatási hálózat egészének szerkezetét méri a köztiség („betweenness centrality”, vagy egyszerűen „betweenness”). Az irányítatlan csúcs köztiség index szemléletes jelentésére az 1.8. ábra mutat példát.

Egy hálózat i. csúcsának betweenness-e (B_i) azon pont párok száma, amelyek között a legrövidebb út átmegy az i. ponton. Ha egy pontpár között több legrövidebb út van, akkor ez a pont pár a legrövidebb út elágazó szakaszain belül 1-nél kisebb járulékot ad az ott található pontok B_i-jéhez. Ezt a járulékot úgy kell kiszámolni, hogy minden elágazásnál az ott szétváló legrövidebb utak számával kell osztani az aktuális járulékot. A betweenness számítására gyakran használt módszer első lépésében szükség van a hálózat tetszőleges (k index-ű) pontjából kifelé haladó „buborékra”, amelyik megméri a k. ponttól az összes többi j. pont távolságát. Ezután egy visszafelé haladó buborék megadja, hogy ha j befutja a k-tól eltérő csúcsokat, akkor j-k pont párok legrövidebb útja(i) milyen járulékot ad(nak) a hálózatban található összes i. pont B_i betweenness-éhez. A többszörös legrövidebb utak miatt adódó 1-nél kisebb súlyokat a visszafelé haladó buboréknál kell figyelembe venni. Ebből a leírásból látható, hogy a betweenness számítása az általános módszerrel $\mathcal{O}(N^3)$ idejű. Lényegesen gyorsabb általános módszer nem ismert, ezért a gyakorlatban a betweenness kiszámítása egy

kis hálózat összes csúcspontjára praktikus, vagy egy nagy hálózat néhány csúcspontjára. A betweenness-t a (biológiai) hálózatos szakirodalom gyakran sorolja a „centrality” (központiság) típusú mérőszámok közé. A biológiai hálózatos kutatási területen néhány korai jelentős publikáció a csúcs fokszámát nevezte a csúcs „centrality”-jének (például [56]), így a centrality szónak ez a használata is előfordul. A csúcsok betweenness-én túl további fontos mennyiség az élek betweenness-e. Ez a mennyiség a csúcsok betweenness-éhez hasonlóan számolandó. A csúcs betweenness-hez hasonlóan minden elágazásmentes legrövidebb pont-pont út 1-gyel növeli az út mentén található összes él betweenness-ét és több legrövidebb út (elágazásos legrövidebb utak) esetén szintén az elágazó utak számával kell osztani az aktuális súlyt.

A betweenness-hez hasonlóan a teljes fehérje-fehérje kölcsönhatási hálózatról tájékoztat a csúcsok fokszámainak eloszlása (a fokszámok valószínűség sűrűsége, probability density function, p.d.f.), amelynek a jele $p(k)$. Szintén gyakran használt mennyiség a szomszédos csúcsok fokszámának korrelációja. Ha a szomszédos csúcsok fokszáma hasonló – tehát a magas fokszámú csúcsoknak főleg magas fokszámú szomszédai vannak és az alacsony fokszámú csúcsoknak jellemzően alacsony fokszámúak a szomszédai – akkor a hálózatot szokás fokszám szerint asszortatívnak vagy egyszerűen asszortatívnak (assortative) nevezni. Ellenkező esetben a hálózat diszasszortatív (disassortative). Utolsó hálózati tulajdonságként vizsgáljuk meg, hogy egy hálózatban melyik csúcsokat lehet a hálózat élein haladva egymásból elérni. Ez alapján egy irányítatlan gráf egy komponense azokat a csúcsokat tartalmazza, amelyek közül tetszőleges kettő elérhető egymásból. A legnagyobb méretű (legtöbb csúcsot tartalmazó) gráf komponens gyakori neve óriás komponens. A molekuláris biológiai hálózatokban az óriás komponens már igen kis hálózat méret esetén megjelenik.

Fehérje-fehérje kölcsönhatási hálózatok modellezése

A fehérje-fehérje kölcsönhatási hálózatok legfontosabb szerkezeti (topológiai) tulajdonságai az alacsony átmérő [57], a magas klaszterezettség [58] és a (gyakori) széles fokszám eloszlás [59]. Ez a három tulajdonság sok más hálózat típusnál is általános, viszont – számos más hálózattól eltérően – a legtöbb fehérje-fehérje kölcsönhatási hálózat diszasszortatív [60]. Az alacsony hálózat átmérő (alacsony átlagos távolság) részletesebben a következőt jelenti. Ha N csúcsot láncként összekapcsolunk, akkor a kapott gráf (hálózat) átmérője az N csúcs szám növelésével lineárisan nő. Ehhez

képest sok valódi (nem modell) hálózat átmérője $\mathcal{O}(N)$-nél lényegesen lassabban növekszik, és esetükben jóval pontosabb az

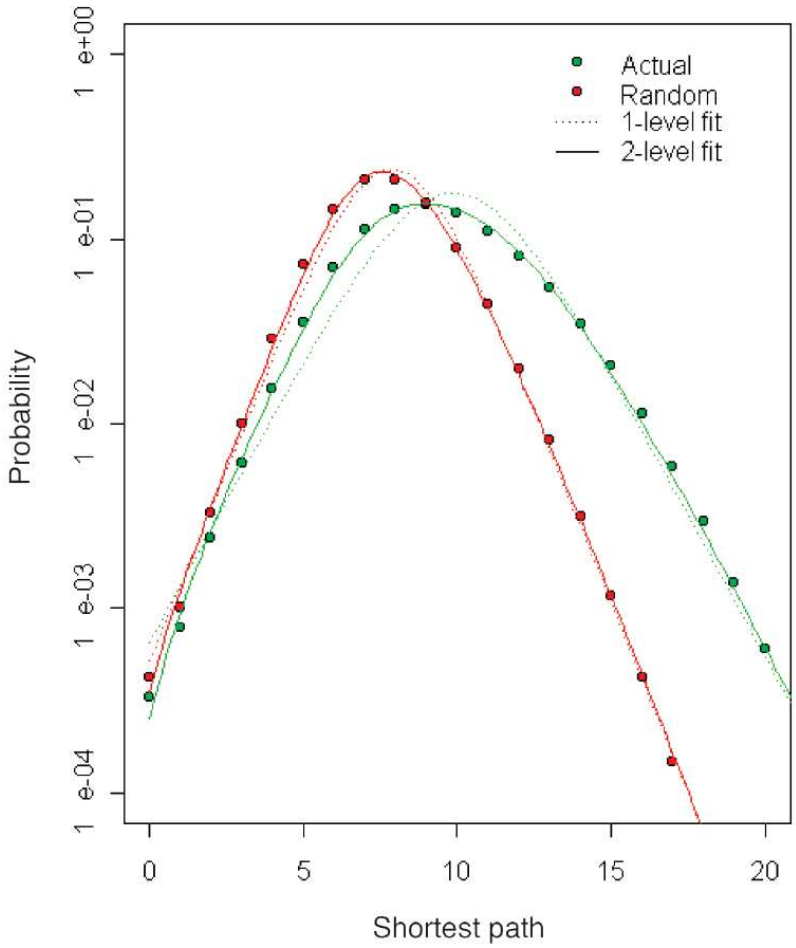
$$L(N) \approx \mathcal{O}(\log N) \tag{1.1}$$

közelítés. Ennek az összefüggésnek a neve kis világ (small world) tulajdonság[13].

A kis világ tulajdonság megfigyeléséhez nem szükséges több tízezer csúcsból álló hálózat, ez a tulajdonság már néhány ezer vagy akár néhány száz csúcspont esetén is jelentkezik. Például a gyümölcslégy (*Drosophila melanogaster*) PPI hálózatának $N = 3039$ csúcsát $E = 3659$ él köti össze, mégis két csúcs (fehérje) távolsága jellemzően 9 és 11 között van [11] (ld. 1.9. ábra), ami sokkal alacsonyabb a csúcsok számánál. Szintén biológiai példa, hogy a *Ceanorhabditis elegans* nevű fonálféreg faj $N = 282$ idegsejtjének kapcsolati hálózatában az idegsejtek közötti átlagos távolság 2.65 lépés [58]. Ez az érték szintén jóval alacsonyabb a csúcsok számánál.

A kis világ tulajdonsággal rendelkező hálózati modellek közül a legegyszerűbb a klasszikus véletlen hálózat (gráf) modell vagy röviden Erdős-Rényi gráf [61, 62]. Az Erdős-Rényi (ER) leírás szerint a hálózat (gráf) N csúcspontja között lehetséges $N(N-1)/2$ él hely közül véletlenszerűen kiválasztott E darab helyre kell elhelyezni egy-egy élt és a többi él hely betöltetlen marad. Fontos, hogy az elhelyezett E darab él helyei korrelálatlanok. Másképpen fogalmazva: tetszőleges két él hely betöltöttségét mutató bináris valószínűségi változók között nincsen korreláció. Emiatt az ER modellt szokás korrelálatlan véletlen gráf modellnek is nevezni. Erdős és Rényi eredményei előtt egy hálózat (gráf) leírása gyakran egy konkrét él lista vizsgálatát jelentette. Ehhez képest Erdős és Rényi egyik jelentős újítása az volt, hogy a leírás során nem egyetlen konkrét hálózatot próbáltak elemezni, hanem az összes olyan hálózatot, amelyik előállhat egy véletlenszerű szabály segítségével N darab csúcs és E darab él felhasználásával. Az ER modell és a későbbi véletlen gráf modellek egy konkrét hálózat helyett mindig azoknak a gráfoknak (hálózatoknak) az összességét elemzik, amelyek az adott modell által definiált szabály használatával létrejöhetnek. A gráf sokaság használatának egyik következménye az, hogy a korábban (egy konkrét gráf esetén) határozott értékkel rendelkező mennyiségek közül többnek a véletlen gráf modellekben már nincsen határozott értéke, csupán várható értéke és például szórása. Így például az ER

[13] A kis világ *tulajdonság* és a kis világ *modell* [58] egymással kapcsolatosak, de nem azonosak. Az általánosított kis világ modell definíciója az, hogy a hálózat rendelkezik a kis világ tulajdonsággal (alacsony átmérő) és magas a klaszterezettsége.



1.9. ábra. A gyümölcslégy (*Drosophila melanogaster*) fehérje-fehérje kölcsönhatási hálózatában a pont-pont távolságok eloszlásának a maximuma a $9 \dots 11$ tartományban van. Ez jóval alacsonyabb a hálózat csúcsainak és éleinek számánál ($N = 3039$ és $E = 3659$). Az Erdős-Rényi modell közelítőleg jól írja le ezt az alacsony várható értékű – $\mathcal{O}(N)$-nél jóval kisebb – mért pont-pont távolságot. Az ábra a távolságot „Shortest path”-nak nevezi, a mért távolságok eloszlását zöld pontok mutatják „Actual” felirattal. Az ER modell alapján kiszámítható távolság eloszlást piros pontok mutatják „Random” felirattal. Az ábra átvétel a [11] publikációból.

modell szerint ha a hálózat átmérőjének (L) várható értékéhez hozzáadjuk ugyanennek az átmérőnek a szórását (vagy akár a szórás háromszorosát), akkor jellemzően jóval alacsonyabb értéket kapunk, mint a hálózat csúcsainak száma (N). Ezt a tulajdonságot tárgyalja a következő bekezdés.

Az ER modellhez és a kis világ tulajdonsághoz is erősen kapcsolódik a gráfelmélet egyik fontos eredménye [61, 62, 63]. Ha az ER gráfban az átlagos fokszámra $\langle k \rangle = 2E/N > 1$ teljesül, akkor az $N \to \infty$ határesetben a gráf összes csúcspontja egyetlen

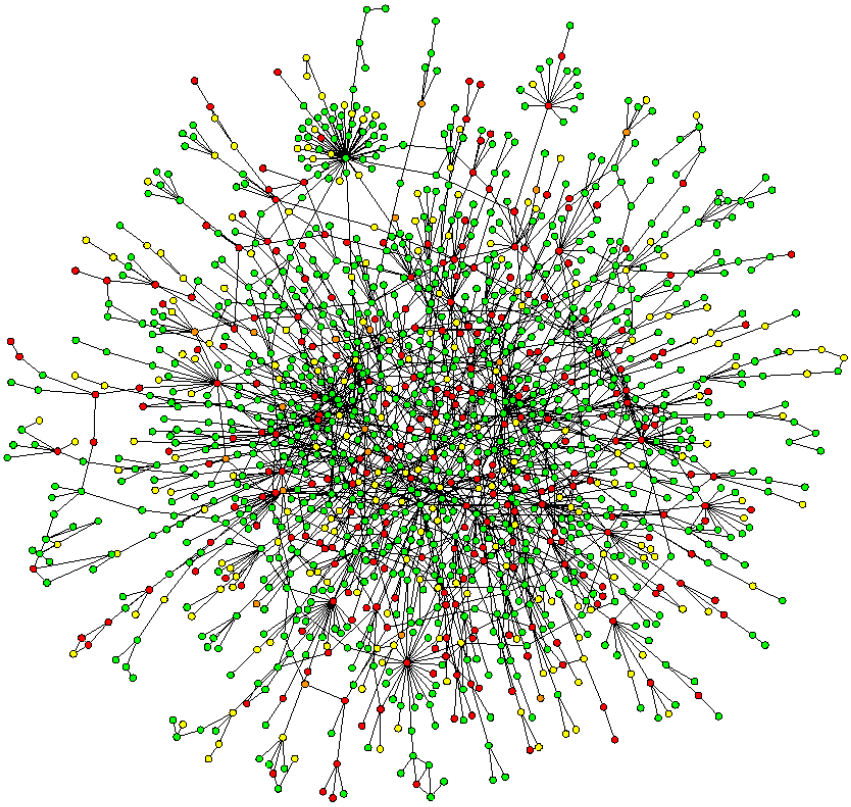
(óriás) komponensbe fog tartozni. Mivel ebben a határesetben minden csúcs azonos komponenshez tartozik, ezért tetszőleges két pont között van kapcsolat a gráfon, és – egy rövid érvelés[14] alapján – az ilyen pont-pont utak jellemző hossza $\mathcal{O}(\log N)$ szerint növekszik. A PPI hálózatokban az átlagos fokszám 1 fölött van, például az imént említett gyümölcslégy (*D. melanogaster*) esetében $\langle k \rangle = 2E/N \approx 2.4$, de ennél magasabb értékek is gyakoriak. Emiatt a PPI hálózattal azonos számú csúcsból és élből álló Erdős-Rényi (ER) gráfban a legtöbb csúcs egy komponensben található, a csúcs párok között van kapcsolat és a pont-pont távolságok jellemző értéke $\log N$-nel arányos. Tehát a PPI hálózatok $\mathcal{O}(N)$-nél jóval alacsonyabb átmérőjét az ER modell képes megmagyarázni egyedül a csúcsok és élek számának felhasználásával.

Az alacsony átmérővel ellentétben a magas klaszterezettséget az ER modell nem magyarázza. A konkrét példák előtt vizsgáljuk meg, hogy az ER modell alapján mekkora egy hálózat klaszterezettségi együtthatója. Egy $p = 2E/(N(N-1))$ sűrűségű ER gráfban a gráf bármelyik két csúcspontja között a kapcsolat valószínűsége p. Tehát az i. csúcspont tetszőleges két szomszédja között is p a kapcsolat valószínűsége, és ez az érték független a kiválasztott csúcs i indexétől. Emiatt az ER modellben tetszőleges csúcs klaszterezettségi együtthatója (átlagosan) p és a teljes gráf klaszterezettségi együtthatójának várható értéke szintén p. A valós gráfokban az élek száma általában nagyságrendileg a csúcsok számával lineárisan növekszik, ebből az ER modell alapján az következik, hogy a gráf klaszterezettsége $C = p = 2E/(N(N-1)) \approx \mathcal{O}(N^{-1})$ módon csökken. A PPI (és számos más) hálózat esetében ennél jóval nagyobb C érték mérhető. Például az élesztőgomba PPI hálózatának klaszterezettségi együtthatója 0.022 [64], míg az azonos csúcs- és él számú ER hálózaté[15] $(5.9 \pm 0.9) \times 10^{-4}$. Hasonló példa, hogy az előző bekezdésben szereplő gyümölcslégy (*D. melanogaster*) idegsejt hálózatában a klaszterezettségi együttható $C = 0.28$, míg az azonos számú csúcsot és élt tartalmazó ER hálózatban C várható értéke 0.05 [58]. A klaszterezettségi együtthatóval kapcsolatos további példát mutat az 1.10. ábra.

A PPI hálózatoknak az alfejezet elején említett négy fő tulajdonsága – alacsony átmérő, magas klaszterezettség, széles fokszám eloszlás és diszasszortativitás – közül

[14] Induljunk ki a vizsgált egyetlen gráf komponens egy tetszőleges csúcsából. Teljes indukcióval látható, hogy ettől a csúcstól $\ell = 1, 2, \ldots$ lépés távolságra összesen nagyjából (vezető rendben) $\langle k \rangle^{\ell}$ csúcs található. A gráf L átmérőjének nagyságrendjébe eső $\ell \approx L$ lépés elérésekor nagyjából az összes csúcsot érjük el, tehát $N \approx \langle k \rangle^{L}$, és így az átmérő $L(N) \approx = O(\log N)$ szerint növekszik.

[15] A véletlen hálózat modellekben (így az Erdős-Rényi modellben is) számos mennyiségnek nem határozott értéke van, hanem csupán várható értéke és szórása.



1.10. ábra. A klaszterezettségi együttható jól vizsgálható a szakirodalom egyik legismertebb fehérje-fehérje kölcsönhatási hálózata segítségével. Az ábra az élesztőgomba (*S. cerevisiae*) PPI hálózatát mutatja az [56] publikációból. (A csúcspontok színei az élesztőgombának az adott fehérje/gén eltávolítása utáni életképességét jelölik.) A PPI hálózat csúcsainak száma $N = 1870$, éleinek száma $E = 2277$, és a hálózat klaszterezettsége (az egyes csúcsok klaszterezettségi együtthatóinak átlaga) $C = 0.0960$. Az ER modell szerint generált – N csúccsal és E éllel rendelkező – 100 darab hálózatban klaszterezettség csupán $\langle C_{ER} \rangle \pm \sigma(C_{ER}) = 0.0010 \pm 0.0008$. Az ER modell alapján kapott átlag relatív távolsága a valódi értéktől jelentős: $Z(C_{ER}) = (C - \langle C_{ER} \rangle)/\sigma(C_{ER}) = 115.2$. Normál eloszlás esetén a 3-as – esetleg a 2-es – $|Z|$ értéket szokás szignifikánsnak tekinteni. A kapott $|Z(C_{ER})|$ ennél jóval magasabb. A számítások a doktori dolgozathoz készültek.

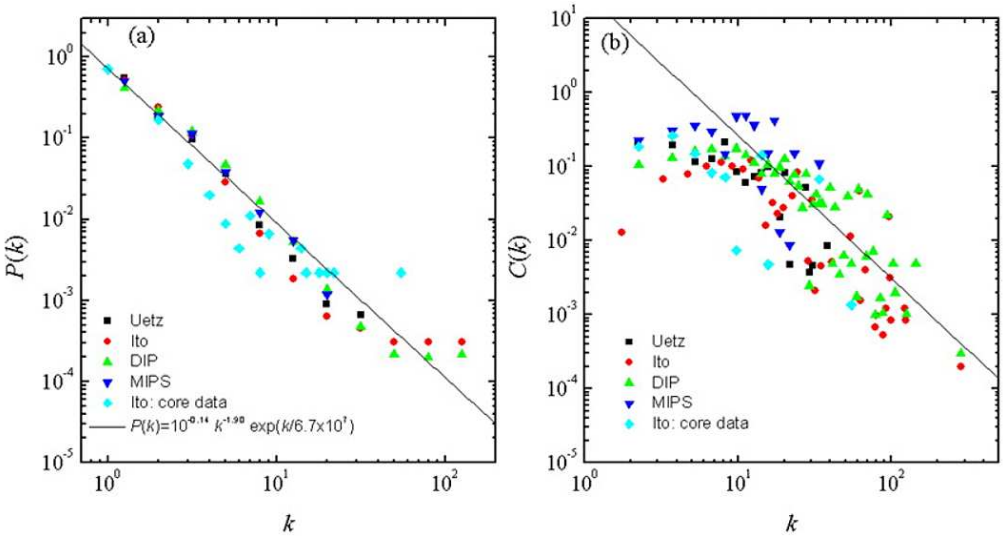
az ER hálózat az alacsony átmérőt jól írja le. Erre mutat egy példát a korábban látott 1.9. ábra. Ehhez képest a kis világ modell (a „Small World model”) már két tulajdonságot ír le helyesen, az alacsony átmérőn túl a magas klaszterezettséget is. A kis világ modellt Watts és Strogatz vezette be [58]. Watts és Strogatz konkrét definíciója előtt tekintsük át a modell általános megfogalmazását. Az általánosított kis világ modell elve az, hogy két hálózat egymásra illesztésekor (szuperpozíciójakor) az egyik háló-

zat kis átmérőt biztosít, míg a másik magas klaszterezettséget. Az alacsony átmérőt (azaz a kis világ tulajdonságot) biztosító első gráf lehet egy ER gráf, amelyben sok a térben távoli kapcsolat. A magas klaszterezettséget biztosító második gráf lehet egy térben lokalizált hálózat, például összeköthetünk minden fehérjét a vele leggyakrabban kapcsolatban lévő 10 másikkal.

A kis világ modell általános leírása után tekintsük Watts és Strogatz konkrét definícióját [58]. Egy kör mentén helyezzünk el N pontot és kössük össze mindegyik pontot az első és második szomszédjával. Az így kapott hálózatban tetszőleges i. pontnak $k_i = 4$ szomszédja van, és azok között összesen $n_i = 3$ él található. Tehát tetszőleges pont klaszterezettsége $C_i = 2n_i/[4(4-1)] = 1/2$, és így a teljes hálózat klaszterezettsége is $C = C_i = 1/2$. Ez az érték független a hálózatban található csúcsok számától[16]. Ezután a hálózatban kis számú, véletlenszerűen kiválasztott él egyik végpontját mozdítsuk el („huzalozzuk át") véletlenszerűen kiválasztott csúcsokhoz. Kis számú áthuzalozott él esetén ez a lépés a klaszterezettséget csak kicsit csökkenti, viszont a csúcsok között létrehoz egy ritka ER hálózatot. Megfelelő számú véletlenszerűen átkötött él esetén ez az ER részgráf fogja biztosítani a hálózat alacsony átmérőjét. Azonban ha már nagyon sok élt elmozdítunk, akkor az említett magas $C = 1/2$ klaszterezettségi együttható csökkenhet. Watts és Strogatz eredménye szerint az élek véletlenszerű elmozdítása során először lesz alacsony a hálózat átmérője, és csak jóval később lesz alacsony a klaszterezettségi együttható. Watts és Strogatz tényleges modellje az általuk leírt hálózatnak ez a köztes állapota, amelyben a klaszterezettség magas és az átmérő alacsony. A Watts és Strogatz által közölt konkrét eljárás részletei szerepelnek a [58] publikációban. Napjainkban a kis világ modell eredeti összeállítási szabálya helyett a modell elve és általánosabb (egyszerűsített) megfogalmazásai már jóval elterjedtebbek.

Az Erdős-Rényi (ER) modell és a kis világ (Small World, SW) modell szerint egy hálózatban a csúcsok fokszám eloszlása exponenciális függvénynél gyorsabban cseng le. Az ER modellben a fokszám eloszlás az $N \to \infty$ határesetben Poisson eloszláshoz tart, aminek a lecsengése a nagy fokszámok tartományában az exponenciális csökkenésnél gyorsabb. Az SW modell fokszám eloszlásának lecsengése – az előző bekezdésben leírt általános SW definíció alapján – az ER modellével azonos.

[16] Összehasonlításként: az Erdős-Rényi modell szerint a klaszterezettségi együttható jóval alacsonyabb, $C = \mathit{o}(N^{-1})$ szerint változik a hálózat méretével.



1.11. ábra. A fehérje-fehérje kölcsönhatási hálózatok (a) fokszám- és (b) klaszterezettség-eloszlása közelítőleg hatványfüggvény alakú. Mindkét rész ábrán a vízszintes tengely egy csúcs fokszámát (szomszédainak számát) mutatja. Az (a) részábrán a folytonos vonal $\gamma = -2.5$ meredekségű hatványfüggvény, a (b) részábrán a folytonos vonal $\gamma = -2$ meredekségű hatványfüggvény. Az ábrán jelölt adatok forrásai a következő publikációk: „Uetz” [37], „Ito” és „Ito core” [7], „DIP” [51], „MIPS” [38]. Az ábra átvétel a [65] publikációból.

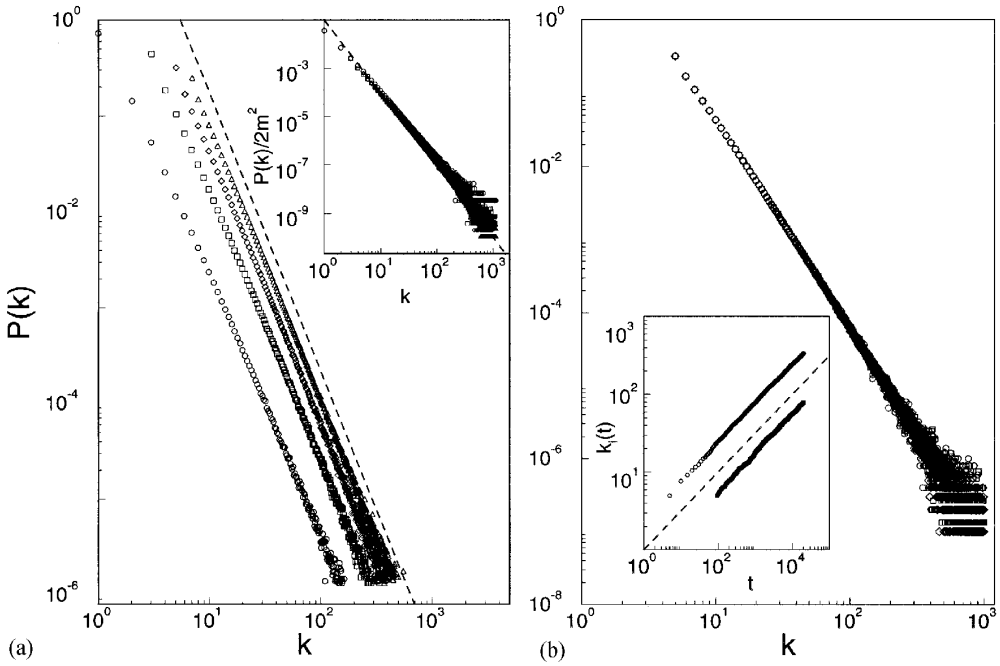
Összefoglalva: mindkét modell szerint a kiugróan nagy fokszámok nagyon ritkán fordulnak elő. Az ER és SW modellek fokszám eloszlásának lecsengése azért jelentős kérdés, mert a 2000-es évek elejétől kezdődően számos PPI (valamint további biológiai és nem biológiai) hálózatról ismertté vált, hogy tartalmaznak az exponenciális eloszláshoz képest kiugróan magas fokszámú csúcsokat, és a fokszámok eloszlása hatványfüggvény alakú [23, 59, 66, 67]. Tehát a PPI hálózatok négy említett tulajdonsága (alacsony átmérő, magas klaszterezettség, kiugróan nagy fokszámok, fokszámok diszasszortativitása) közül az SW modell megmagyarázza az első kettőt, de a harmadikat és a negyediket nem.

A valós hálózatokban megfigyelhető magas fokszámok magyarázatára a legelterjedtebb általános (nem speciálisan biológiai) modell típus a skálafüggetlen gráf. Ebből a modell típusból (modell családból) a leggyakrabban használt konkrét modell a fokszámok szerint lineárisan preferenciális kapcsolódással időben egyenletesen növekedő gráf [66, 68, 69, 70], röviden Barabási-Albert (BA) modell[17]. A speciálisan biológiai modell típusok jóval kevésbé elterjedtek. Esetükben a legfontosabb, hogy a modell a

[17] Ezt a modellt az irodalomban szokták „scale-free” modellnek is nevezni.

lehető legtöbb, kísérleti eredmény alapján ismert mechanizmust tartalmazza és a lehető legkevesebb feltételezéssel sok eredményt pontosan írjon le. Ezek alapján az egyik legmegfelelőbb PPI hálózat modell a duplikáció-mutáció modell. Egy további, általános modell – amely több tulajdonságot egyszerű módon képes együtt leírni – Holme és Kim "hármas csoport" (triad formation) modellje.

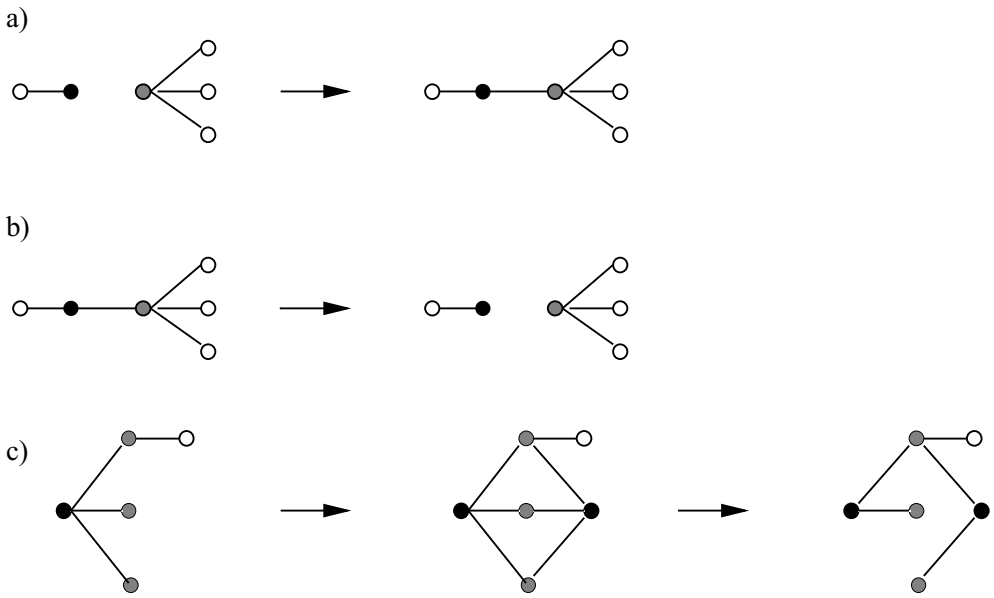
A BA modell szerint a hálózatok egyenletes sebességű növekedéssel érik el végtelen számú lépés után határesetként a pontosan hatványfüggvény szerinti fokszám eloszlást. A modell kiinduló lépésében a hálózat m_0 csúcsból és nulla élből áll. A t. és $(t+1)$. időlépés között egy új csúcs és m új él csatlakozik a hálózathoz. A gyakorlatban gyakran az $m = m_0 = 2$ paraméter értékeket használják. Minden új él egyik végpontja az új csúcs, és a másik végpontja a már meglévő („régi") $m_0 + t$ darab csúcs valamelyike. Mindegyik új él tetszőleges régi csúcspontot annak a csúcsnak a már meglévő fokszámával arányos valószínűséggel választ ki (a valószínűségek konkrét numerikus számításakor a nulla fokszámok helyett 1-es fokszámokat szokás használni). Fontos, hogy az m darab régi csúcspontot nem egymás után, hanem egyszerre kell kiválasztani. Ennek a kiválasztásnak a gyakorlati megvalósítása a következő. Egy egyenesre az origóból kiindulva mérjük fel egymás után a meglévő („régi") $m_0 + t$ csúcs fokszámának megfelelő hosszúságú intervallumot. Mivel ezeknek a csúcsoknak az összesített fokszáma kétszerese az élek számának, ezért az egymás után felmért $m_0 + t$ intervallum összesen 0-tól $2mt$-ig ér. Ezután válasszunk ki a $[0, 2mt)$ tartományból egyenletes eloszlással m darab véletlen pontot. Ha az $m_0 + t$ darab intervallum valamelyikébe egynél több pont esik, akkor ismételjük meg a kiválasztást egészen addig, amíg mindegyik intervallumba csak egy pont esik. Az így kiválasztott m darab intervallumhoz tartozó „régi" csúcspont lesz a hálózathoz a $(t+1)$. időlépésben csatlakozó m darab él végpontja, amit a $(t+1)$. időlépésben csatlakozó „új" csúccsal kötnek össze. A BA modellre néhány lépésben [68] levezethető a $t \to \infty$ esetben érvényes fokszám eloszlás: egy $\gamma = -3$ meredekségű hatványfüggvény. A levezetés első lépése a fokszám szerinti lineáris preferenciális kapcsolódást és a fokszámok időben egyenletesen növekvő összegét felhasználva folytonos idejűvé tett leírással megmutatja, hogy tetszőleges csúcs fokszáma minden időpontban (átlagosan) a hálózat $t = 0$ időpontjától eltelt idő négyzetgyökével arányosan növekszik. Erre építve a második lépés az időben egyenletes növekedés alapján a fokszámok kumulált sűrűségfüggvényéből (c.d.f.) kiindulva megmutatja, hogy a fokszámok sűrűségfüggvénye (p.d.f.) a



1.12. ábra. A skálafüggetlen hálózat modellek csoportjába tartozó BA (Barabási-Albert) modellben a csúcsok fokszám eloszlása egy $\gamma = -3$ lecsengésű hatványfüggvényhez tart és minden csúcs fokszáma a hálózat "kezdete" óta eltelt idő négyzetgyökével arányosan növekszik. A modell részletes leírása megtalálható a dolgozat szövegében. Az ábra átvétel a [68] publikációból. **(a)** A fő panel a fokszám eloszlást (a fokszámok sűrűségfüggvényét) mutatja azonos $N = m_0 + t = 300\,000$ számú csúcs és eltérő él sűrűségek esetén: $m_0 = m = 1$ (kör), $m_0 = m = 3$ (négyzet), 5 (rombusz) és 7 (háromszög). A szaggatott vonal meredeksége $\gamma = -2.9$. A bal felső sarokban lévő kis panel ezeket a fokszám eloszlásokat egybeskálázva mutatja, mindegyik fokszám eloszlás $2m^2$-tel való osztása után. A szaggatott vonal meredeksége $\gamma = -3$. **(b)** A fő panel a fokszám eloszlás konvergálását demonstrálja a hálózat méretének növekedésével. Három eltérő csúcs számú hálózat szerepel ezen a panelen ($N = 100\,000$ körrel, $N = 150\,000$ négyzettel és $N = 200\,000$ rombusszal) azonos $m_0 = m = 5$ él sűrűség paraméterrel. A kis panel két olyan csúcs fokszámának időbeli növekedését mutatja, amelyek a $t_1 = 5$ és $t_2 = 95$ időpontban csatlakoztak a hálózathoz. A szaggatott vonal meredeksége a modell által előrejelzett $1/2$.

következő: $p(k) = k^{-3}\, 2m^2/(m_0+t)$. A Barabási-Albert modell eredményeit mutatja az 1.12. ábra.

A PPI hálózatok szempontjából speciális fokszám-fokszám korreláció (a fokszámok diszasszortativitása) megtalálható a BA modellben: ha két csúcs fokszámának (1-nél nagyobb) arányát növeljük, akkor a két csúcs között a kapcsolat valószínűsége növekszik. Ennek a levezetéséhez induljunk ki az előző bekezdésben említett folytonos idő közelítés [68] egyik köztes eredményéből, ami szerint a BA modellben bármelyik



1.13. ábra. A [71] publikációban ismertetett duplikáció-mutáció modell elemi lépései a mutáció által kölcsönhatások (a) keletkezése és (b) eltűnése (a fekete és szürke színű csúcs között) valamint a (c) duplikáció. A duplikáció a (c) részábra első lépésében mutatott módon történik: a fekete színű csúcspont kettőződik. A (c) részábra második lépésénél látható, hogy a duplikáció után azonos kölcsönhatásokkal rendelkező két fehérje a mutáció révén már eltérő kölcsönhatásokkal rendelkezik és így megkülönböztethetővé válik.

csúcs fokszáma mindig a hálózat $t = 0$ időpontja óta eltelt idő négyzetgyökével arányosan nő. Ha az i. csúcs a t_i időpillanatban "születik" (csatlakozik a hálózathoz), akkor egy későbbi t időpillanatban a fokszáma $k_i(t) = m\sqrt{t/t_i}$. A BA modell növekedési szabálya miatt két csúcs között a kapcsolat csak akkor jöhet létre, amikor a "fiatalabb" csúcs csatlakozik a hálózathoz. Emiatt egy idősebb (i.) és egy fiatalabb (j.) csúcs közötti kapcsolat valószínűsége, $p(i, j)$, arányos az idősebb csúcsnak a fiatalabb születése pillanatában mérhető fokszámával. Továbbá fordítottan arányos az ugyanebben a pillanatban a hálózatban lévő fokszámok összegével. Tehát az i. és j. csúcs "születési" időpontját rendre t_i-vel és t_j-vel jelölve $t_i < t_j$ esetén az $m << t_i$ közelítésben a két csúcs közötti kapcsolat valószínűsége $p(i, j) = m^2\sqrt{t_j/t_i}/2m = m\sqrt{t_j/t_i}/2$. Tehát minél korábban "született" az idősebb csúcs (kis t_i) és minél később született a fiatalabb csúcs (nagy t_j), annál nagyobb köztük a kapcsolat valószínűsége. Ez a fokszámok diszasszortativitását jelenti a BA modellben.

A duplikáció-mutáció modell [71] a (i) gén duplikációt és a (ii) kölcsönhatások mutáció miatti változását veszi figyelembe (1.13. ábra). A gén duplikáció során egyet-

len DNS leolvasási szakasz másolata megjelenik az adott élőlény genomjának egy másik helyén is. Ilyenkor a két DNS leolvasási szakasz alapján készülő fehérjék azonosak, tehát a (fehérje-fehérje) kölcsönhatásaik is azonosak. Ha a két gén közül valamelyik génben mutáció következik be, akkor az a belőle "készülő" fehérje meglévő kölcsönhatásait változtathatja. Így a két fehérje már egymástól megkülönböztethetővé válik, és a PPI hálózatban külön csúcspontként fognak szerepelni. Ebben a modellben a mutáció a hálózatot "véletlenszerűsíti", tehát a kis világ tulajdonság érvényes lesz. Ha a duplikálódó fehérje képes önmagával kölcsönhatni, akkor a duplikáció utáni mutációs lépésben számos háromszög keletkezik, és a modell magas klaszterezettséget eredményez. A modell időben egyenletesen növekszik és bármely időlépésében minden fehérje esélye a duplikációra azonos. Emiatt egy időlépésben tetszőleges (i index-ű) fehérje szomszédainak száma a meglévő szomszédainak k_i számával arányos valószínűséggel növekszik $(k_i + 1)$-re, és a növekedés során a hálózat fokszám eloszlása hatványfüggvényhez konvergál. Továbbá a modellben – a fokszámok gyakorisága alapján várható valószínűségekkel normálva – nagyobb valószínűséggel fordulnak elő kapcsolatok egy kis és egy nagy fokszámú csúcs között, mint egy nagy és egy másik nagy fokszámú csúcs között. Tehát a modell a diszasszortatív fokszám-fokszám kapcsolatot is képes leírni. Összefoglalva: a duplikáció-mutáció modell konkrét mért adatok alapján két biológiailag jelentős folyamatot használ fel és ezek segítségével képes leírni a kis világ tulajdonságot, a hatványfüggvény szerinti fokszám eloszlást, a szomszédos fokszámok közötti negatív korrelációt és esetenként a magas klaszterezettséget.

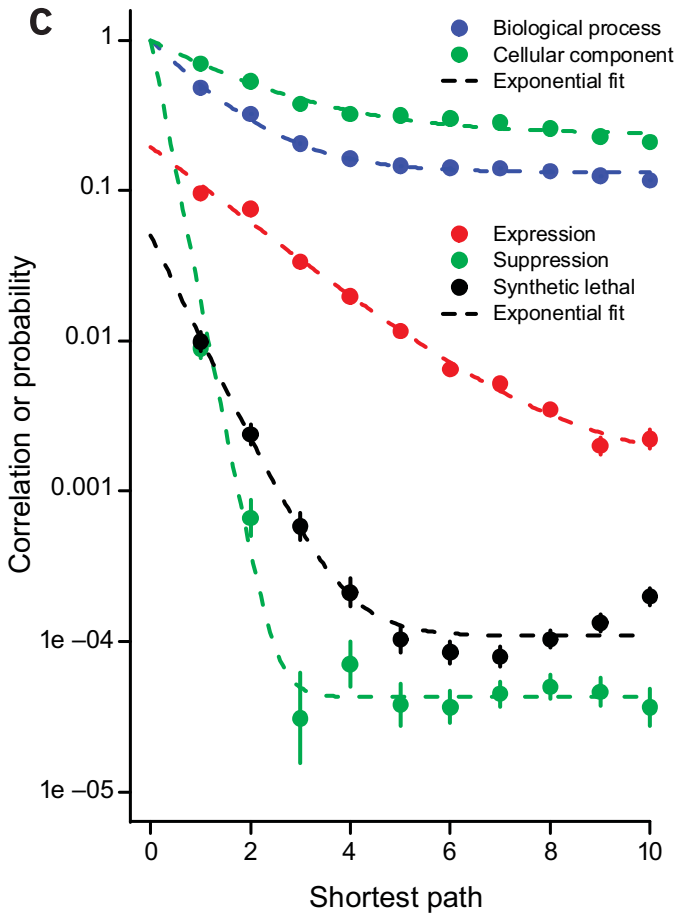
A hálózat modellek közül utolsóként vizsgáljuk meg Holme és Kim háromszögeket képző modelljét [72], ami a BA modell egy módosítása. Holme és Kim a BA modellt módosították úgy, hogy a növekedő hálózat minden egyes időlépésében az új csúcs nem korrelálatlanul kiválasztott m darab "régi" csúcshoz kapcsolódik, hanem egy "régi" csúcs kiválasztása után annak közelében (kapcsolatai között) keres további kapcsolatokat. Ez a feltételezés a fehérje-fehérje hálózatokban helytálló, mert (1) az "új" fehérje nagyobb eséllyel kerül kapcsolatba biológiailag egymáshoz hasonló "régi" fehérjékkel, mint teljesen eltérőekkel, és (2) vannak ilyen egymáshoz hasonló fehérjék, mert a biológiai tulajdonságok hasonlósága a PPI hálózatokban nagyjából 3 lépésnyi hálózati távolságnál szűnik meg [73] (1.14. ábra). A BA modell Holme és Kim szerinti módosítása háromszögeket képez, hiszen a hálózatba újonnan érkező csúcsot gyakran két, egymással már kapcsolatban lévő csúcshoz egyaránt hozzákapcsolja. A

leírási módszer további előnye, hogy a keletkező háromszögek száma egyetlen paraméter segítségével szabályozható. Összefoglalva: Holme és Kim módszere megtartja a BA hálózat alacsony átmérőjét és hatványfüggvény szerinti fokszám eloszlását, és kiegészíti azokat a magas klaszterezettséggel.

Az itt bemutatott hálózat modellek véletlen gráf modellek, amelyek konkrét hálózat definiálása helyett egy-egy gráf sokaság "elkészítését" (generálását) definiálják és erre a gráf sokaságra átlagolt tulajdonságokat (például átmérőt) vizsgálnak. Mindegyik leírt modellben teljesül a kis világ tulajdonság. Ezen túl megfigyelhető a magas klaszterezettség a Kis Világ modellben, a BA modell Holme-Kim-féle módosításában és gyakran a duplikáció-mutáció tpusú hálózatfejlődés eredményeként is. Utóbbi esetben a magas klaszterezettségnek egy elégséges feltétele az, hogy a duplikációs lépés után keletkező két fehérje kölcsönhatásban legyen egymással. Az itt leírt modellek közül a hatványfüggvény szerinti fokszám eloszlást leírja a skálafüggetlen modellek családjába tartozó BA modell, a duplikáció-mutáció modell és a Holme-Kim modell. A fehérje-fehérje hálózatokra speciálisan jellemző fokszám diszasszortativitást (az egymással kapcsolatban lévő csúcsok fokszámának negatív korrelációját) az itt vizsgált modellek közül a BA modell és a Holme-Kim modell írja le. A duplikáció-mutáció modellben a fokszám-fokszám korreláció előjele a duplikáció és a mutáció sebességének arányától függ, csak alacsony mutációs ráta esetén negatív. Molekuláris biológiai (alapkutatási és gyógyászati) szempontból a skálafüggetlen modellek egyik fontos következménye az, hogy a biológiai rendszerekben (és más rendszerekben is) gyakran a hálózati csúcspontok "fontossága" a legnagyobbaktól a kisebbek felé haladva folytonosan csökken, ezért sokszor nincsen lehetőség a fontos és nem fontos csúcsok (gének, fehérjék) egyszerű szétválasztására.

PPI hálózatok moduljai és a PPI modulokat kereső módszerek

Az előző alfejezetben leírt hálózat modellek a fehérje-fehérje kölcsönhatási hálózatok számos tulajdonságát helyesen írjak le, de a valódi hálózatokban megfigyelt moduláris szerkezetet nem vagy csak alig tartalmazzák. Tekintsünk ismét az 1.14. ábrára, amelyről leolvasható, hogy az élesztőgomba PPI hálózatában a biológiai tulajdonságok korrelációja 2-3 lépésnyi hálózati távolságnál tűnik el. Ugyanez az eredmény másképpen megfogalmazva azt jelenti, hogy a fehérje-fehérje kölcsönhatási hálózatban a fehérjék tulajdonságainak korrelációja 1 lépésnél szignifikánsan nagyobb



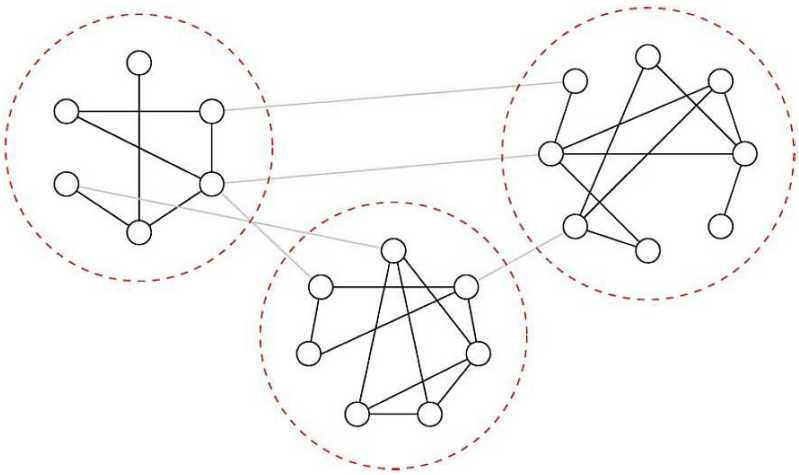
1.14. ábra. Az élesztőgomba (*S. cerevisiae*) yeast-2-hybrid és Co-IP kölcsönhatásainak összesített elemzése alapján a kölcsönhatási hálózatban 2-3 lépésnyi távolság után a biológiai tulajdonságok hasonlóságai eltűnnek. Az ábra átvétel a [73] publikációból. A Co-IP esetében a szerzők a mért komplex-ekből kiindulva a korábban említett "spoke" (küllő) módszerrel határozták meg közelítő módon a fehérje-fehérje pár kölcsönhatásokat. A PPI hálózatban a fehérjék nagy része rendelkezik – a Gene Ontology adatbázisban [74] található – egy vagy több "Biological process" (biológiai folyamat) és "Cellular component" (sejten belüli hely/alegység) besorolással. A függőleges tengely fehérjék besorolás (címke) listáinak a Jaccard korrelációit mutatja mindig a két összehasonlított fehérje vízszintes tengelyre felmért hálózati távolságának függvényében. Az ábrán a "Synthetic lethal" címkéjű görbe a fehérje pároknak a dolgozat szövegében leírt genetikai kölcsönhatását tesztelik.

távolságig megtalálható. Ezzel szemben az Erdős-Rényi modellben és a Kis Világ modellben két szomszédos csúcs minden tulajdonsága korrelálatlan, tehát a korrelációs hossz még 1-nél is rövidebb, nulla. Az előző alfejezetben leírt rövid levezetés alapján a Barabási-Albert modellben a szomszédos fokszámok a megfigyeléseknek megfe-

lelően antikorrelációt (diszasszortativitást) mutatnak. Továbbá a BA modellben még a másodszomszéd csúcsok között is van korreláció, akárcsak a BA modell Holme és Kim által javasolt változatában. Az általános modellek által mutatott fokszám-fokszám korrelációk azonban egyetlen esetben sem írják le azt a széles körben megfigyelhető jelenséget, hogy a – biológiai, társadalmi, technológiai és egyéb – hálózatok gyakran nagy számú, egymással átfedő, sűrű részhalmazt tartalmaznak. Ezek a csúcs halmazok (csoportok) sűrűek abban az értelemben, hogy a csúcsok lényegesen nagyobb valószínűséggel kapcsolódnak a csoporton belüli más csúcsokhoz, mint a csoporton kívüli csúcsokhoz.

Érdemes megjegyezni, hogy a molekuláris biológiai folyamatokban gyakran a hálózatok által adottnak feltételezett bináris fehérje-fehérje kölcsönhatásoknál pontosabb leírást adnak a kölcsönható fehérje csoportok, a modulok. Ennek legfőbb oka az, hogy az élő sejtekben gyakori a kooperatív kölcsönhatás, tehát nem kettő, hanem három vagy még több fehérje (és esetleg további molekula típus) kölcsönhatása befolyásol egy eseményt. Amikor ezeket a molekuláris biológiai modulokat a mérési módszerek és a mérési eredmények alapján a PPI hálózatokat definiáló eljárások hálózati élekké alakítják, akkor gyakori, hogy egy biológiailag ismert modulon belül kapunk magas páronkénti kapcsolati valószínűséget (és magas súlyokat), és azon kívül alacsony kapcsolati valószínűséget. Erre egy gyakori példa az, hogy a korábban említett Co-IP és TAP mérések által azonosított fehérje komplex-eken belül szokás az azonosított „mátrix” módszerrel a komplex-ben lévő minden fehérje-fehérje párt összekötni egy éllel. Ha a mérések egy fehérje komplex-nek többféle módosulatát is azonosítják, akkor a komplex központi részében lévő fehérje-fehérje párok sokszor fordulnak elő az eredményekben, és a köztük futó élek súlya nagy lesz. Így az eredményként kapott PPI hálózatban mindannyian nagyobb valószínűséggel szerepelnek, és a hálózatban modulokat kereső módszerek a komplex központi részét valóban, biológiailag helyesen azonosítják modulként.

Az előző bekezdésben leírtak alapján biológiai szempontból helyes eredményt adnak azok a matematikai módszerek (algoritmusok), amelyek a PPI hálózatban csúcsok lokálisan sűrűn összekötött halmazait keresik. Az általános hálózati modul kereső algoritmusok közül széles körben elterjedt Girvan és Newman eljárása [76], rövid nevén a Girvan-Newman (GN) módszer. A GN módszer abból a feltételezésből indul ki, hogy a vizsgált hálózatban egymástól jól elkülönülnek a „ritka” hálózati környezetnél



1.15. ábra. Moduláris szerkezet egy kis hálózatban. Azonos modulon belül lévő két csúcsot jóval nagyobb valószínűséggel köt össze él, mint két olyan csúcsot, amelyek különböző modulban vannak. Az ábra átvétel a [75] publikációból.

magasabb él sűrűséggel rendelkező tartományok (ld. 1.15. ábra). Ennek a szerkezetnek az egyik következménye az, hogy a sűrű tartományokat (csúcspont halmazokat) összekötő kis számú élen lényegesen több legrövidebb pont-pont kapcsolat halad át, mint a halmazok belsejében futó éleken. Tehát minél alacsonyabb egy él betweenness-e, annál inkább egy sűrű halmaz „belsejében” található, és minél magasabb egy él betweenness-e, annál több és nagyobb sűrű halmaz között biztosít kapcsolatot. Emiatt a a GN módszer a sűrű tartományokat (hálózati klasztereit) úgy azonosítja, hogy a hálózat legmagasabb betweenness-ű éleinek eltávolítja, és ezután felsorolja a megmaradó gráf komponenseit, ezek lesznek az eredeti gráf klaszterei.

Az előző bekezdésben leírtak után a GN eljárás pontos definíciójához egyetlen paramétert kell rögzíteni, az eltávolítandó élek számát. Ha a hálózatból csak az első néhány élt távolítjuk el (a legmagasabb betweenness-ű éleket), akkor megmaradnak a modulokat összekötő kapcsolatok, és a gráf nem esik szét komponensekre, tehát nem sikerül azonosítani a modulokat. Ha túl sok élt távolítunk el, akkor a megmaradó gráf komponensei az eredeti gráf moduljai helyett már csak azoknak kis részei lesznek. Az optimálisan eltávolítandó él szám meghatározása érdekében Newman és Girvan [75] a következőket javasolta: (1) minden lépésben töröljük a legnagyobb betweenness-ű élt, (2) az él kitörlése után számoljuk újra az élek betweenness-ét és így (3) annál a lépésnél találjuk a legjobb modul szerkezetet, amelynél a modulokon belüli és a modulok közötti élsűrűség a legjobban eltér. Ezt az eltérést a Newman-féle modularitás (Q)

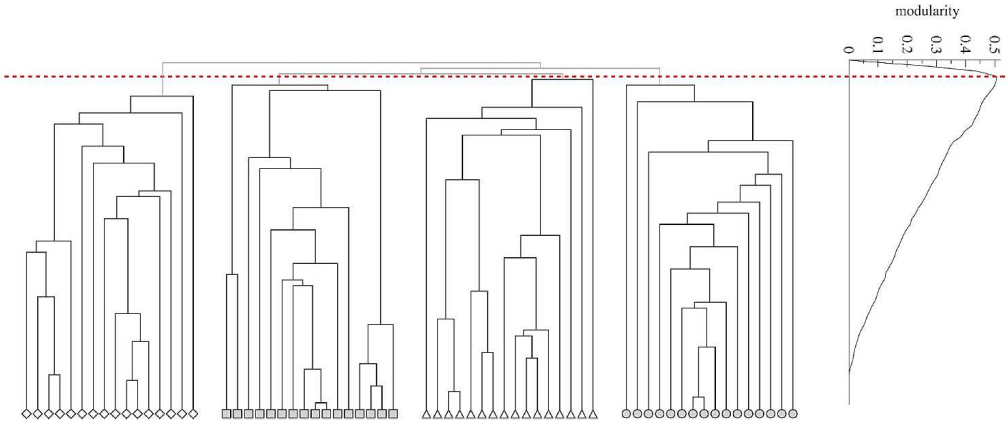
számszerűsíti. A Newman-féle modularitás definiálásához gondolatban csoportosítsuk egy hálózat csúcsait. Az i. csoporton belül futó élek számát jelölje e_{ii}, továbbá az i. csoporton belül legalább egy végponttal rendelkező élek számát jelölje a_i. A Newman-féle modularitás (Q) azt méri, hogy egy csoporton belül lényegesen több él fut-e, mint azok között:

$$Q = \sum_i \left(e_{ii} - a_i^2\right). \tag{1.2}$$

Egy összefüggő kiindulási gráf esetén az élek eltávolítása során a Newman-féle modularitás nulláról indul, növekszik, majd csökken és az összes él törlése után ismét nulla lesz. Végig nemnegatív értékeket vesz fel. A köztes lépések közül a maximális Q-hoz tartozó állapot lesz az, amelynek a gráf komponensei az eredeti hálózat csoportjait (moduljait) adják a GN módszer szerint. Erre mutat egy példát az 1.16. ábra.

A GN módszer előnye, hogy a (i) korábban már alaposan vizsgált agglomeratív („agglomerative") és felosztó („divisive") hierarchikus klaszterezésre épül, (ii) a hálózatban mérhető él betweenness jelentése egyszerű és az áthaladó legrövidebb útvonalakkal jól szemléltethető, továbbá (iii) kis hálózatokra hatékony algoritmusokkal megvalósítható. A GN módszer valódi hálózatokra történő alkalmazásának két főbb hátránya van. Az első hátrány az, hogy egy N csúcsból álló hálózatban az él betweenness számítása egzakt módon $\mathcal{O}(N^3)$ számítási időt igényel, és jelenleg (ismereteim szerint) nincsen olyan gyorsított változata, amelyik néhány tízezer csúcspont feletti hálózatokra alkalmazható és széles körben elterjedt (tehát már alaposan tesztelték a felhasználók). A GN módszer második hiányossága az, hogy az azonosított hálózati modulok között nem enged meg átfedéseket.

A PPI hálózatok moduljainak azonosítására a Girvan és Newman által javasolt algoritmuson túl számos további módszer ismert. Ezeknek a módszereknek a nagy része a statisztikus fizikában, az informatikában és a gráfelméletben korábban már kidolgozott klaszter (csoport) keresési módszereken alapszik. Ezek a – korábban más kutatási területeken már kidolgozott – csoport keresési módszerek általában a beágyazási térben definiált metrika (például euklideszi vagy hálózati távolság) segítségével azonosítják elemek (vektortérben lévő pontok, hálózati csúcspontok) olyan halmazait, amelyeknek elemei nagyobb valószínűséggel kapcsolódnak egymáshoz, mint a halmazon kívüli más elemekhez. A következő néhány bekezdés az a statisztikus fizikai, informatikai és a gráfelméleti eredetű főbb klaszterezési módszereket tekinti át. Az



1.16. ábra. Egy 64 csúcspontból álló teszt hálózat moduljainak keresése a Girvan-Newman módszerrel. Az ábra átvétel a [75] publikációból. A teszt hálózat 4 darab, egyenként 16 csúcsot tartalmazó modulján belül minden csúcsnak $z_{in} = 6$ kapcsolata van. Minden csúcsnak a modulján kívüli csúcsokkal $z_{out} = 2$ kapcsolata van. Az ábra alján található pontok mutatják a hálózat 64 darab csúcspontját. Az összes él törlése során a hálózat egyetlen komponensből 64 komponensre esik szét. Ezt a folyamatot követhetjük az ábrán látható hierarchikus fa tetejétől lefelé haladva. A hierarchikus fán látható elágazások a gráf komponensek egymás utáni szétesését mutatják. Az ábra jobb oldalán látható, hogy az élek törlése közben hogyan változik a Newman-féle modularitás (Q) értéke. Az ábra jobb oldalán lévő modularitás grafikon maximum értéktől szaggatott vonal vezet az ábra bal oldalán lévő fának a 4 modult tartalmazó állapotához (a szaggatott vonal az ábra bal oldalán lévő fát annál a pontnál „vágja el", ahol a fának 4 ága van). A konkrét példában négy gráf komponens esetén maximális a modularitás, tehát a GN módszer szerint ez a négy pont csoport az eredeti hálózat moduljai.

áttekintésből kimarad számos olyan módszer, amely a molekuláris biológiában is gyakori: k-means (k-medoids) [77], Self-Organized Maps (SOM) [78], neurális hálózatok[18] [79], Singular Value Decomposition (SVD) [80], és a Support Vector Machine (SVM) [81]. Ezek a módszerek azért maradnak ki a következő bekezdésekben szereplő felsorolásból, mert (i) nem hálózatokon vannak definiálva, hanem általában vektortereken és (ii) döntően gén expressziós (mRNS koncentrációs) adatok csoportosítására használatosak, nem fehérje-fehérje kölcsönhatások elemzésére. Egy érdekesség, hogy molekuláris biológiai hálózatokon szokás „távolságot"[19] definiálni úgy is, hogy bármely két (A és B) csúcshoz hozzárendeljük a köztük lévő ℓ hálózati távolság (lépés szám) inverz négyzetét: $d(A, B) = \ell^{-2}$ [82].

[18] A dolgozat hálózatokban történő modul (csoport) keresésről szól. A félreértések elkerülése érdekében érdemes megjegyezni, hogy a neurális hálózatokkal végzett csoportosítás nem a neurális hálózat csúcspontjait csoportosítja, hanem a neurális hálózat „input" adat vektorait.

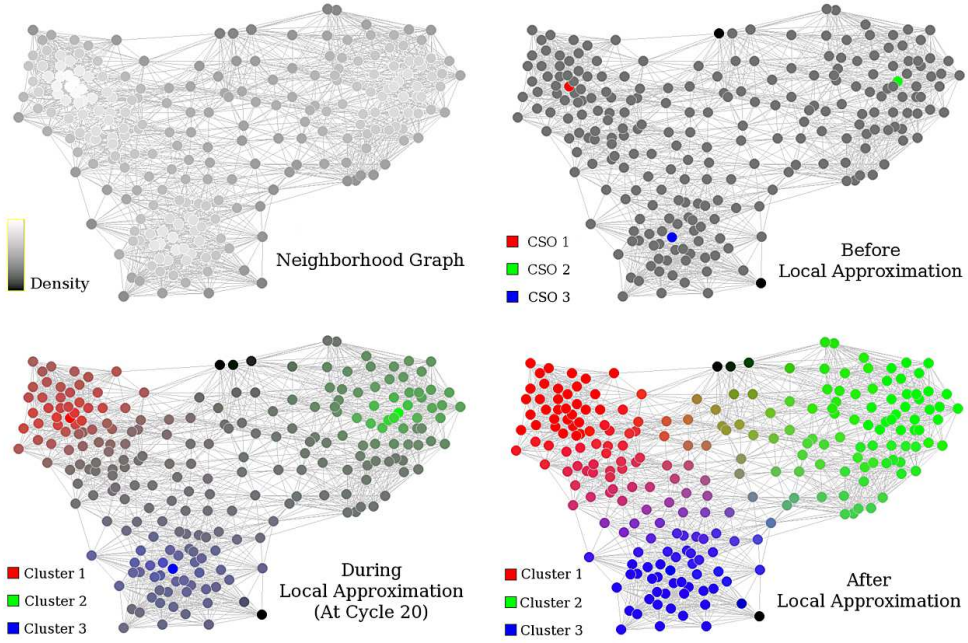
[19] Erre a „távolság" fogalomra nem érvényes a háromszög szabály.

A gyakran használt statisztikus fizikai módszerek közül kiemelkedőek a Monte-Carlo típusú módszerek, amelyek úgy keresnek csoportokat, hogy azokon maximalizálnak „belül sok kölcsönhatás, de kifelé kevés kölcsönhatás” típusú mennyiségeket [83]. Ezek közül az egyik legismertebb a szuperparamágneses klaszterezés (superparamagnetic clustering, SPC), amelynek alapja egy vektortérben ferromágnesesen kölcsönható részecskék rendeződése a pontok térbeli közelsége által meghatározott csoportok szerint [84]. Fontos, hogy az SPC módszer definíciója szerint csak diszjunkt csoportok (modulok) azonosítását végzi.

Az SPC-vel ellentétben átfedő klasztereket ad eredményként egy informatikai eredetű algoritmus család, a fuzzy clustering (elkent csoportosítás, ld. például [85]). Ezzel a csoportkereső módszerrel minden pont tetszőleges számú csoporthoz tartozhat, és mindegyik csoportjához egy 0 és 1 közötti súllyal tartozik. A fuzzy clustering típusú módszerekben egy-egy csoport azonosításához hagyományosan a csoport tagok közötti euklideszi távolságra volt szükség, ami hálózatokban nem létezik. Jelenleg már rendelkezésre állnak fuzzy módszerek hálózatok klaszterezésére is, ám az eredményként kapott „elkent” és nagyon részletes csoport szerkezet gyakran az eredeti adatokat kevésbé redukálja (egyszerűsíti), mint más, egyszerűbb módszerek. Az 1.17. ábra egy elkent modul keresési (fuzzy clustering) példát mutat egy molekuláris biológiai hálózatban a [86] publikáció alapján.

Két gyakran használt gráfelméleti eredetű hálózati csoportkereső módszer a klikk keresés és a k-core keresés. Ezekben a módszerekben az azonosított klikkek illetve k-core-ok a hálózat moduljai. A k-klikk egy k csúcsból álló teljes részgráf, azaz a hálózat csúcsainak egy olyan csoportja, amelyben bármely két csúcs kapcsolatban van egymással. Egy k-klikket természetesen tartalmazhat egy $(k+1)$-klikk. A k-klikktől eltérő fogalom a klikk. Ha egy k-klikket nem tartalmaz $(k+1)$-klikk, akkor egyben klikk is. Másként fogalmazva: a klikk egy maximális k-klikk. Egy hálózatban a k-core egy maximális méretű összefüggő részgráf, amin belül minden csúcs fokszáma legalább k. Fontos, hogy egy hálózat k-klikkjei átfedhetnek, viszont a k-core-ok között nem lehetséges átfedés.

A Monte-Carlo algoritmust használó hálózati modul kereső módszerek előnye, hogy – a determinisztikusság hiánya miatt – a keresési feltétel nem egzakt, tehát a biológiai rendszerekben és a mérésekben gyakori hibákra kevésbé érzékeny. Hátrányuk is ugyanebből fakad, mert a kapott eredmény nem jóldefiniált. A fuzzy clustering



1.17. ábra. Fuzzy clustering (elkent csoportok) azonosítási lépései egy hálózatban a [86] publikációban leírt módszer alapján. (1) A bal felső ábrán látható hálózatban a kereső módszer kijelöli a csoportok (klaszterek, modulok) középpontjait. A csoportok száma rögzített. (2) Ezek a klaszter középpontok láthatóak a jobb felső ábrán piros, zöld és kék színnel jelölve. (3) A bal alsó ábrán látható a modulok „növelése” a középpontjaikból kiindulva. (4) A jobb alsó ábra mutatja a végeredményt, amelyben minden pont több modulhoz is tartozhat és a színe az ábrán a modul tagsági súlyainak megfelelő. Az ábra forrása a Wikipédia.

előnye, hogy az azonosított hálózati modulok között definíció alapján megenged átfedéseket. Hátránya, hogy az eredmény nagyon részletes, komplex, és ezáltal kevésbé képes teljesíteni a hálózati modul keresés egyik fő célkitűzését: kevésbé képes redukálni a hálózatot annak fő összetevőire. A klikk és k-core keresés előnye, hogy pontosan definiált eredményt ad és matematikai értelemben valóban a lehetséges legsűrűbb halmazokat azonosítja. Viszont mindkettőnek hátránya, hogy a valódi modulok nem a lehetséges legsűrűbbek, hanem a teljes összekötöttséghez képest általában hiányzik belőlük számos csúcs és él. Tehát ez a két módszer igen érzékeny a biológiai és mérési hibákra.

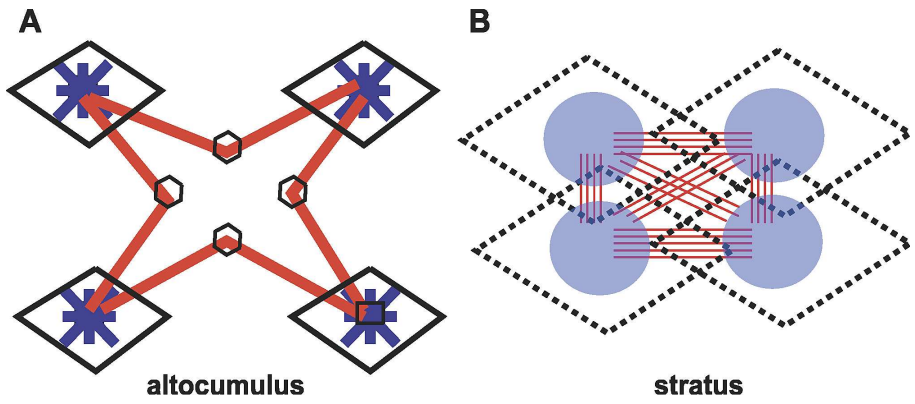
Az előző bekezdésekben bemutatott módszerek értékelése során sok szó esett a fehérje-fehérje kölcsönhatási hálózatokban található modulok átfedéséről. Ezek nagyságát egyszerűen érzékelhetjük, ha az élesztőgombában (*S. cerevisiae*) ismert fehérje komplex-eket (fizikailag összekapcsolódó fehérje csoportokat) megvizsgáljuk.

A CYGD (Comprehensive Yeast Genome Database) 2005-ös adatai alapján [87] az élesztőgombában ismert fehérje komplex-ekben összesen 2750 különböző fehérje vesz részt, de a komplex-ek fehérje számainak összege 8932. Ez azt jelenti, hogy átlagosan mindegyik fehérje 3-nál több komplex-nek tagja. Mivel a fizikailag kölcsönható fehérjék csoportjai (a komplex-ek) a fehérjék által alkotott funkcionális moduloknak csak egy részét jelentik, ezért a funkcionális modulok között még ennél is jelentősebbek az átfedések.

A fehérje kölcsönhatási hálózat moduljainak azonosítása a molekuláris biológia és a sejtbiológia számára lényeges útmutatást adhat. Az egyik kapcsolódó tudományos vita során a vitázó csoportok a modulok és a nagy fokszámú (sok kapcsolattal rendelkező) fehérjék kapcsolatát elemezték több cikk sorozatban. A hálózatokban sok kapcsolattal rendelkező csúcsokat (például fehérjéket) szokás középpontnak („hub”-nak) nevezni. Ezt a megnevezést használva egy 2004-es publikáció megállapította [88], hogy az élesztőgomba PPI hálózatában a nagy fokszámú fehérjék jelentős része két fő csoportba osztható: „party hub” és „date hub” (itt a „date” szó jelentése „randevú”). Előbbiek (a „partizó” központi fehérjék) jellemzően egyszerre lépnek kölcsönhatásba a kölcsönható partnereikkel (fehérjékkel), míg az utóbbiak (a „randizó” központi fehérjék) jellemzően különböző időpontban lépnek kapcsolatba a nagy számú kölcsönható partnerükkel. Ezt a szerkezetet később szemléletesen az „altocumulus” típusú felhőkhöz is hasonlították, Az altocumulus típusú felhőkben nagy számú, de kis méretű sűrű tartomány áll egymással laza kapcsolatban. Egy későbbi eredmény [89] szerint a lazán összekapcsolt sűrű tartományokat tartalmazó (altocumulus-szerű) szerkezet helyett a PPI hálózatok inkább egyetlen sűrűn összekötött felhőhöz (egy „stratus”-hoz) hasonlítanak. A PPI hálózatok szerkezetéről alkotott két eltérő képet az 1.18. ábra mutatja be.

A „party-hub, date-hub” (vagy „stratus-altocumulus”) néven ismertté vált vitának az elméleti hálózat dinamikai érdekességen túl számos konkrét kapcsolódási pontja van biológiai szempontból központi kérdésekhez. A vitázó felek között – a korábbi eredményeket [56] megerősítve – egyetértés alakult ki abban, hogy pozitív korreláció van aközött, hogy egy fehérjének sok kölcsönhatása van és hogy az adott fehérje esszenciális[20] [90]. Abban viszont már csak további kutatócsoportok munkái alapján jött létre egyetértés, hogy egy fehérje evolúciójának sebessége általában nem függ a

[20]Egy fehérje egy adott szervezetben esszenciális, azaz nem nélkülözhető, ha a fehérje génjének a szervezet DNS-éből történő eltávolítása után a szervezet nem marad életképes.



1.18. ábra. A fehérje-fehérje kölcsönhatási hálózatok szerkezetéről alkotott két eltérő leírás. (A) A „party hub, date hub” leírás [88] szerint a PPI hálózatban kis sűrű modulokat nagy fokszámú csúcsok kapcsolnak össze. Ennek a szerkezetnek a [89] publikáció szerint a neve „altocumulus” (gomolyfelhő), mert egymással kapcsolatban álló klaszterekből („gomolyagokból”) áll. (B) A „stratus” (réteges) felhő leírás [89] szerint a teljes PPI hálózat funkcionálisan eltérő részei mindannyian szoros kapcsolatban vannak a többi résszel. A szerkezeti képek közötti eltérés a fehérjék evolúciójának sebességével is kapcsolatos. Részletek a szövegben. Az ábra átvétel a [89] publikációból.

fehérje kölcsönhatásainak számától [91, 92]. A tudományterület gyors fejlődése ellenére még napjainkban is számos, molekuláris biológiai hálózatokat elemző publikációban kerül a középppontba a „hub”-ok és a modulok kapcsolata. Ennek fő oka, hogy a nagy fokszámú csúcsok gyakran több modulnak is nélkülözhetetlen tagjai. Ha egy-egy ilyen „hub”-ot eltávolítunk, gyakran több modul teljesen átalakul vagy akár megsemmisül. Ez a tulajdonság rávilágít a gyógyszeres kezelések során gyakran felmerülő fehérje funkció gátlás mellékhatásainak egyik fontos okára. Az itt leírt megfigyelésekből a modulokat azonosító módszerek számára a következtetés egyértelmű. A legtöbb fehérje-fehérje kölcsönhatási hálózatokban csak olyan módon szabad modulokat keresni, hogy a keresés közben a modulok közötti átfedéseket megengedjük.

Eredmények

1.1. Átfedő modulok azonosítása fehérje-fehérje kölcsönhatási (PPI) hálózatokban [T1, T2, T3]

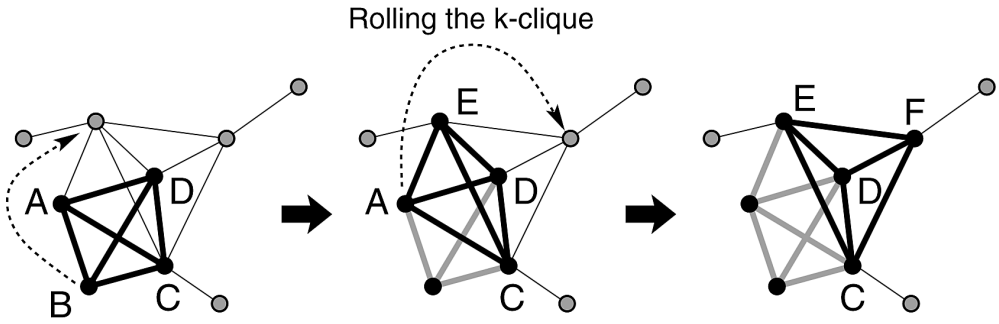
Az előző alfejezet szerint a fehérje-fehérje kölcsönhatási hálózatok moduljainak szerkezetéből akkor vonhatóak le biológiailag helyes következtetések, ha egy viszonylag egyszerű és hibatűrő algoritmus alkalmazásával egymással átfedő modulokat azonosítunk. A leírt módszerek közül a Monte-Carlo típusú optimalizáló módszerek, a

„fuzzy” (elkent) klaszterezés és a klikkek azonosítása biztosít átfedéseket a klaszterek között. A Monte-Carlo módszerekkel és a „fuzzy” módon kapott klaszterek nem biztosítanak egyértelmű eredményt és a klikkek azonosítása az adatok kis módosításaira érzékeny, azaz nem hibatűrő. A modul kereső módszer hibatűrése a következők miatt fontos követelmény. (1) A felhasznált kísérleti technikák mindegyike számos egyedi és szisztematikus mérési hiba lehetőséget tartalmaz. (2) A fehérje-fehérje kölcsönhatási hálózatok az adott sejtben lehetséges összes kölcsönhatás térképét mutatják. Ennek a térképnek a különböző részeit (a kölcsönhatási gráf részgráfjai) képesek mérni torzításokkal a dolgozatban korábban részletesen leírt mérési és adatfeldolgozási módszerek. Többek között a hibatűrés, a modulok közötti átfedések megengedése, és az egyszerűség követelményeinek próbáltunk megfelelni a klikk perkolációs módszer [93] (Clique Percolation Method, CPM) fehérje-fehérje kölcsönhatási hálózatokra történő alkalmazásával [T1, T2].

A klikk perkolációs módszer PPI hálózatokra történő alkalmazásakor abból a feltételezésből indultunk ki, hogy a biológiailag jelentős funkciókat fehérjék olyan csoportjai végzik, amelyek között a csoporton belül nagyobb valószínűséggel van kapcsolat, mint a csoporton kívül. A módszer a már említett k-klikkekből indul ki, és a pontos definíciójához első lépésként szükségünk lesz a k-klikkek közötti szomszédság fogalmára. Két k-klikk szomszédos egymással, ha pontosan $(k-1)$ olyan csúcspont van, amelyet mindkét k-klikk tartalmaz. Tehát két szomszédos k-klikk átfedése pontosan egy $(k-1)$-klikk. Ha két k-klikknek $(k-1)$-nél kevesebb csúcspontja közös, akkor a két k-klikk egymással nem szomszédos, hanem egymással átfed. A k-klikkek szomszédságának fogalma után induljunk ki egyetlen k-klikkből és járjuk be a vizsgált gráfnak (hálózatnak) azt a maximális részgráfját, amelyik k-klikk szomszédsági kapcsolatokon keresztül elérhető. Az így bejárt részgráf a hálózat egy k-klikk perkolációs klasztere[21].

A [T1] publikációnkban a k-klikk perkolációs klaszterekt tekintettük (definiáltuk) hálózati moduloknak egy, a konkrét hálózatban kiválasztott k értéket használva. A módszert az 1.19. ábra mutatja be egy kis példa hálózaton. A k érték kiválasztása azért fontos, mert a legtöbb hálózatban általában többféle k paraméter értékkel lehet k-klikk perkolációs klasztereket definiálni. A k paraméter érték kiválasztásakor az volt a célunk, hogy a kapott klaszterek a lehető leginformatívabbak legyenek hálózati

[21] A klikk perkoláció kifejezésben szereplő perkoláció szó arra utal, hogy a klasszikus perkoláció jelenségéhez hasonlóan szintén kapcsolatokon keresztül bejárható maximális méretű csoportokat azonosítunk.



1.19. ábra. A klikk-perkolációs (teljes névvel: k-klikk perkolációs) módszer szemléltetése $k = 4$-es k-klikk paraméterrel egy kis hálózaton. Az ábra átvétel a [T4] publikációnkból, Palla Gergely készítette. A szövegben leírt kiinduló k-klikket az ABCD csúcsok alkotják. Az ABCD k-klikknek szomszédja az ACDE k-klikk, és annak egy további szomszédja az ECDF k-klikk. Tehát a mutatott kis hálózatban az ABCD k-klikkből kiindulva a szomszédos k-klikkeken át bejárható maximális részgráf az ABCDEF. Figyeljük meg, hogy minden egyes szomszédsági lépésben az ábrán feketével rajzolt k-klikk templát egyetlen csúcsa mozdul el és a többi $(k-1)$ darab csúcsa helyben marad. Például az ABCD - ACDE szomszédság lépés során a fekete színű templát B csúcsa mozdul el az E csúcsba és az ADC csúcsok helyben maradnak. Ez alapján a szomszédos k-klikkre történő átlépés szemléletesen nevezhető a k-klikk templát „görgetésének” (k-clique rolling).

modulként. Magas k paraméter érték nagyobb él sűrűségű modulokat eredményez. A lehetséges legalacsonyabb k érték, amit választhatunk, a $k = 2$. Ebben az esetben minden k-klikk a hálózat egy éle, két szomszédos k-klikk $(k-1) = 1$ közös csúcsot tartalmaz, tehát a k-klikk szomszédság él szomszédságot jelent és minden k-klikk perkolációs klaszter egy gráf komponens.

A fehérje-fehérje kölcsönhatási hálózatok legnagyobb (gráf) komponense általában a hálózatnak majdnem az összes csúcspontját tartalmazza. Ez az eredmény biológiai modulként nem informatív. Emiatt a gráf komponenseket eredményező $k = 2$-es paraméter értéknél magasabbat kell választani. A legmagasabb választható k érték a hálózatban található legnagyobb klikk mérete. PPI hálózatokban gyakran 8-9 vagy több csúcsból álló teljes részgráfok (klikkek) is előfordulnak. Ilyenkor a k-klikk perkoláció eredményeként kapott modulok élsűrűsége magas, de a hálózat összes pontjához képest igen kevés pontot tartalmaznak. Összefoglalva: (i) ha a k (klikk méret) paraméter alacsony, akkor a kapott modulok között lesz egy nagy és sok kicsi; (ii) ha a k paraméter magas, akkor a kapott néhány modul mindegyike kicsi lesz (mindegyik modul a k klikk mérettel egyező vagy annál kicsivel több csúcsot tartalmaz). Közepes k-klikk méret esetén általában van kicsi, közepes és nagy méretű modul egyaránt, és

a modul méret eloszlás hatványfüggvényhez hasonló. A [T1] cikkünkben ezt a folytonos (gyakran hatványfüggvény-szerű) modul méret eloszlást tekintettük a biológiailag leginformatívabbnak, és a legjobb k-klikk paraméter azonosításához a következő kiválasztási módszert javasoltuk. A legnagyobb és a második legnagyobb modulban (k-klikk perkolációs klaszterben) található csúcspontok számának aránya legyen a lehető legközelebb a 2-höz.

A klikk perkolációs módszerhez (CPM) szükséges klikk keresés általános esetben NP teljes feladat [22] [94]. Azonban a valós PPI hálózatokban vagy befejezhető a keresés rövid időn belül, vagy a matematikailag egzakt eredmény biológiai szempontból nem szükséges és helyettesíthető egy heurisztikus (nem egzakt) módszer eredményével. Ezek alapján a CPM-et megvalósító (implementáló) CFinder szoftver a következő módszereket használja. A CFinder az alapértelmezett beállítások esetén a klikk perkolációs módszer egzakt eredményét számítja ki a beolvasott hálózat klikkjeinek megkeresésével. A klikkek azonosítása azért elegendő, mert a felsorolásukkal megkapható tetszőleges k mérethez az összes k-klikk és a k-klikk perkolációs klaszterek[23]. Ha egy hálózat nagyon sűrű (például az átlagos fokszám $\langle k \rangle = 2E/N > 5$), akkor előfordulhat, hogy több nagy méretű klikk páronként egymással erősen átfed. Ilyen esetben a klikkek egzakt azonosítása a felhasználó számára reális időnél sokkal tovább tart, ezért a CFinder-ben lehetőség van egy közelítő módszer használatára. A közelítő módszerben a felhasználó kijelölheti, hogy a CFinder csúcsonként maximálisan mennyi időt (hány másodpercet) töltsön el a kereséssel. Az 1.20. ábra egy példát mutat a CFinder használatára.

A PPI hálózatokban és más molekuláris biológiai is hálózatokban gyakori, hogy az éleknek súlya van. Ezek az él súlyok általában az élek által jelölt kölcsönhatások valószínűségét, jelentőségét vagy gyakoriságát mérik. Az él súlyok segítségével történő modul keresés érdekében a CFinder tartalmaz élsúly szerinti vágási lehetőséget: a felhasználó a modul keresés indítása előtt kiválaszthatja él súly alapján az összes él egy részét. Ha a felhasználó kijelöl egy minimális és egy maximális él súlyt (lehet

[22] A számítástudományban az algoritmizálható problémákat szokás csoportosítani aszerint, hogy a megoldáshoz a legrosszabb esetben szükséges számítási idő hossza a vizsgált rendszer méretével hogyan növekszik (skálázik). Egy probléma az NP csoportba tartozik, ha a megoldása nem determinisztikus módon, de polinomiális idő alatt elvégezhető. Egy probléma NP-nehéz, ha a megoldásához szükséges idő legalább olyan gyorsan növekszik a rendszer méretével, mint a legnehezebb NP probléma megoldásához szükséges idő. Egy problémát „NP teljes”, ha egyaránt NP és NP-nehéz.

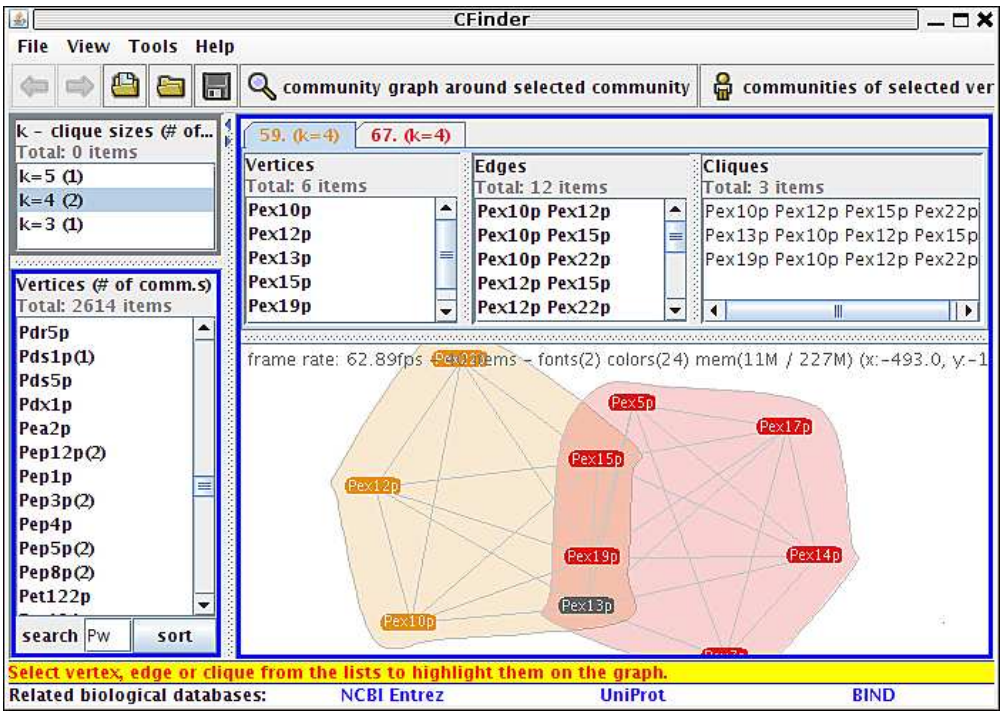
[23] A CFinder súlyozott és irányított élek esetén is használható. Az irányított klikk perkolációs módszert a dolgozat 2. fejezete definiálja és használja.

csak az egyiket), akkor a CFinder a hálózat összes éle közül csak az kijelölt tartományba eső súlyú éleket használja. Az élsúlyok szerinti „vágás” esetén a CFinder törli a kiválasztott intervallumon kívül lévő súllyal rendelkező éleket, és törli a kiválasztott (a modul kereséshez felhasznált) élek súlyát. Ha egy hálózatban az élek súlyának eloszlása folytonos és széles (nem egy-két csúcsban koncentrálódik), akkor jóval pontosabb megoldást ad az élsúlyok pontos felhasználásával történő modul keresés. Ennek érdekében a klikk perkolációs módszert kiterjesztettük súlyozott hálózatokra és az így kapott súlyozott klikk perkolációs módszert szintén elérhetővé tettük a CFinder-ben.

A súlyozott klikk perkolációs rövidített neve CPMw [24]. A módszer alapja a részgráf erősség („subgraph intensity”) nevű mennyiség segítségével definiált súlyozott k-klikk. Onnela és társszerzői egy súlyozott hálózatban (olyan hálózatban, amelyben minden élhez rendelkezésre áll egy súly) tetszőleges részgráf erősségét (intenzitását) a következőképpen definiálták: egy részgráf intenzitása a részgráf él súlyainak mértani közepe [95]. A [T3] publikációnkban erre a részgráf intenzitás definícióra építettük a súlyozott k-klikk perkolációs klaszterek definícióját. Ha egy k csúcsból álló teljes részgráf (teljes, azaz minden csúcs pár között van él) intenzitása nagyobb, mint egy előre rögzített I intenzitás küszöb érték, akkor ezt a részgráfot súlyozott k-klikknek neveztük el. A súlyozatlan klikk perkolációs módszerhez hasonlóan a súlyozott esetben is (i) pontosan akkor szomszédos két súlyozott k-klikk, ha pontosan $(k-1)$ csúcspontjuk közös, (ii) a súlyozott k-klikkek szomszédsági kapcsolatain keresztül bejárható maximális részgráfok a súlyozott (k-)klikk perkolációs klaszterek, és (iii) a CPMw módszer által definiált hálózati modulok a súlyozott klikk perkolációs klaszterek. A módszer két paraméterét (a k élsűrűség és az I intenzitás küszöb paramétert) a súlyozatlan esethez hasonlóan úgy javasoltuk megválasztani, hogy a két legnagyobb klaszter méretének (csúcs számainak) aránya a 2-höz legközelebb legyen.

A súlyozott klikk perkolációs módszert F. Zhang és munkatársai felhasználták daganatos emberi tüdőben található göbök (csomósodások) osztályozására (átfedő csoportokkal) [96]. Georgii és szerzőtársainak elemzése alapján súlyozott fehérje-fehérje kölcsönhatási hálózatok moduljainak azonosítása esetén a CPMw módszer ROC görbéje jobb (magasabban halad), mint a CPM módszeré, de mindkettőé alacsonyabban halad, mint az általuk javasolt élsűrűség-alapú DME (Dense Module Enumeration) módszeré [97].

[24] A CPMw betűszóban a „w” a weighted (súlyozott) szót jelzi



1.20. ábra. Példa a klikk perkolációs módszerrel működő CFinder szoftver használatára. Az ábrán az élesztőgomba Pex13 fehérjéjének a moduljai láthatóak a DIP (Database of Interacting Proteins) adatai alapján. A CFinder szoftver letölthető Windows, Linux és Macintosh operációs rendszerre a http://CFinder.org weboldalról. A letöltött csomag tartalmazza a Linux parancs soros módban futtatható programot 32 és 64 bites architektúrára egyaránt. Az ábra átvétel a CFinder weboldalon található részletes felhasználói leírásból (Manual), és a [T2] publikációnkban található 1. ábra alapján készült.

A szakirodalomban számos további olyan általános (nem speciálisan PPI) hálózati modulkereső módszer ismert, amely a lokális élsűrűség azonosításán alapszik. Ezek közül néhány példa a következő. Raghavan, Albert és Kumara minden csúcsot külön modulként inicializált és az iteratív módszerük minden lépésében mindegyik csúcsot hozzárendelték ahhoz a modulhoz, amelyikhez a legnagyobb számú szomszédja tartozik [98]. Ennek a módszernek a neve „label propagation method”. Érdekes, hogy – Tibély és Kertész eredményei [99] alapján – a „label propagation” módszerrel történő hálózati modulkeresés ekvivalens egy Potts modell energia minimumainak keresésével. Rosvall és Bergstrom a hálózat élein haladó véletlen bolyongások útvonalait tárolta tömörített módon [100]. Blondel és munkatársai szintén minden csúcsot külön modulként inicializáltak és ezután a csúcsokat mozgatták a modulok között a modulokon belüli élsűrűség növelésének irányában [101]. A „label propagation”, Rosvall és

Bergstrom bolyongásos módszere és a csúcsok modulok közötti mozgatása a szükséges számítási idő alapján hatékony megoldások, ellenben a modulok közti átfedéseket nem engedik meg.

Szintén általános (nem speciálisan PPI) hálózatokban Lancichinetti, Fortunato és Kertész a hálózat lokális él sűrűsége alapján definiálta és használta minden egyes csúcs „természetes modulját". A módszerrel kapott eredmény egyszerre azonosított átfedő hálózati modulokat és a hálózat alapján hierarchikus szerkezetet [102]. Ahn, Bagrow és Lehmann a hálózat csúcsai helyett a hálózat éleit csoportosították modulokba [103]. Gregory módszere általánosította Raghavan, Albert és Kumara módszerét olyan módon, hogy minden csúcs tartozhat egynél több modulhoz [104]. Lee és munkatársai egy-egy csúcsból (mint modulból) kiindulva kerestek szomszédos csúcsokat, amelyeket az adott modulhoz hozzáadva a modul sűrű marad [105]. A felsorolt módszereket és továbbiakat is elemez és összehasonlít Fortunato összefoglaló cikke [106]. A hálózati modul keresés szakterületén napjainkra széles körben elfogadottá vált az, hogy (i) a hálózatok moduljai (PPI és más hálózatokban egyaránt) átfedőek és (ii) a matematikailag egzakt módon implementált algoritmusok lelassulása esetén az alkalmazások számára a közelítő (heurisztikus) hálózati modulok is gyakran megfelelő pontosságúak.

Az általános (nem speciálisan PPI) hálózati modulkereső módszereken túl létezik több olyan módszer, amelyik speciálisan PPI hálózatok moduljainak azonosítására lett kidolgozva. Ezek közül néhány példa a következő. A ClusterONE magas belső élsűrűséggel rendelkező csúcs csoportokat keres úgy, hogy a csoportok egymás közötti átfedéseit megengedi [107]. Az MCL („Markov cluster") módszer fehérjéket csoportosít szekvenciáik közötti hasonlóságok alapján, célja fehérje családok azonosítása. Az MCL algoritmusa sztochasztikus (Markov) mátrix-ok segítségével véletlen bolyongásokat értékel ki [108]. Az MCODE („Molecular Complex Detection") magas élsűrűségű kis részgráfokból kiindulva – a részgráfok „növesztésével" – keres modulokat, amelyeknek később az egyedi pontosítását is lehetővé teszi [109]. Yu és munkatársainak algoritmusa olyan részgráfokat keres a PPI hálózatban, amelyeket néhány él hozzáadásával teljes részgráffá (klikk-é) lehet kiegészíteni [110]. A ModuLand kereső 7 különböző módszert – nagyrészt lokális élsűrűség-vizsgálatot – használ átfedő hálózati modulok azonosítására [111]. Az RRW („repeated random walks") módszer az MCL-hez hasonlóan véletlen bolyongásokat értékel ki, de az MCL-lel ellentétben

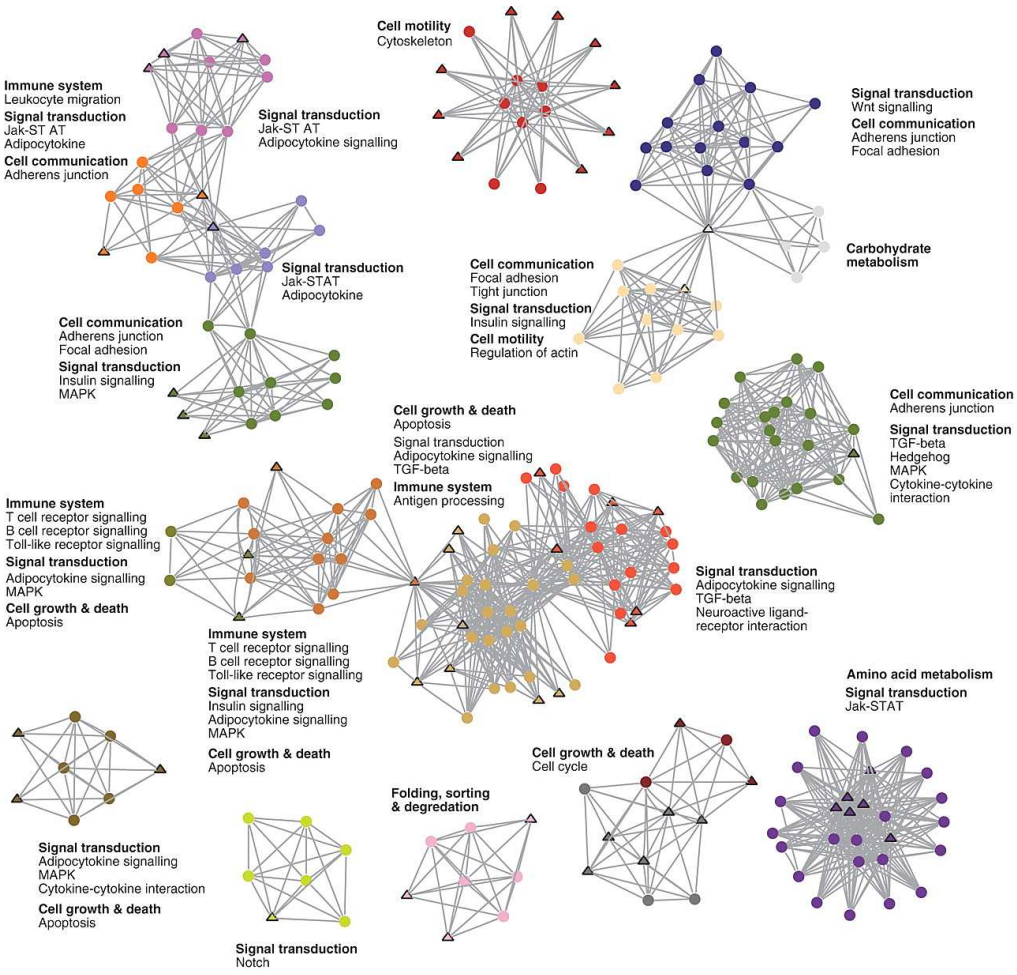
felhasználja az éleken található súlyokat (él erősségeket) is [112]. Jelenleg a ClusterONE és a ModuLand modulkereső algoritmusok egyaránt a Cytoscape szoftvercsomag [113] részeként (moduljaként) érhetőek el[25] A felsoroltakon túl is létezik számos további (bár kevésbé elterjedt) módszer, például a COACH (core-attachment) algoritmus [114] és a korábban már említett DME (Dense Module Enumeration) [97] módszer, amelyek – hasonlóan az MCODE-hoz – sűrű részgráfokból indulnak ki, valamint az MCL módszer egy továbbfejlesztése [115], amely az eredeti MCL-lel szemben megengedi a modulok közötti átfedéseket.

A klikk perkolációs módszert (a CPM-et) és a módszert használó CFinder szoftvert az általunk vizsgált eseteken túl több további fehérje-fehérje kölcsönhatási (PPI) hálózatra alkalmazták. Jonsson és Bates az emberi PPI hálózatban határozott meg modulokat, és vizsgálta azokon belül a daganatos betegségekkel kapcsolatos fehérjék helyét [116] (ld. 1.21. ábra). S. Zhang és munkatársai az élesztőgomba PPI hálózatában talált modulokat összehasonlították kísérletekből ismert fehérje komplex-ekkel és funkcionális csoportokkal [117]. Azt találták, hogy a CPM által kiszámított modulok (i) legtöbbje megfeleltethető egy vagy több funkcionális csoportnak és a kiszámított modulok (ii) nagyjából fele megfeleltethető kísérletekből ismert fehérje komplex-eknek. J. Zhu [118] és munkatársai az élesztőgomba génjeinek transzkripcióját[26] szabályozó kölcsönhatások hálózatát elemezték, és ennek kapcsán a fehérje-fehérje kölcsönhatások hálózatában modulokat azonosítottak a klikk perkolációs módszerrel.

A klikk perkolációs módszert 2005-ben és 2006-ban PPI hálózatokra alkalmazó publikációink [T1, T2] óta a PPI adatok sokat fejlődtek. A frissebb mérési adatok közül kiemelkedőek a közvetlen fehérje-fehérje kölcsönhatásokat és a fehérje komplexeket azonosító mérések. Közvetlen fehérje-fehérje kölcsönhatásokat mérő eredményeket publikáltak élesztőgombában Tarassov és munkatársai 2008-ban [119], korábbi élesztőgomba PPI adatok minőségét elemezték Yu és munkatársai szintén 2008-ban [10], és az *Arabidopsis thaliana* (lúdfű) fehérje-fehérje kölcsönhatásait publikálta egy konzorcium a Vidal csoport vezetésével 2011-ben [48]. A komplex mérések közül néhány példa a következő. Gavin és munkatársai TAP (Tandem Affinity Purification) módszerrel kapott eredményeiket 2006-ban publikálták [120], Krogan és munkatársainak hasonló kísérleti technológiával végzett mérései ugyanebben az évben jelentek

[25] A Cytoscape egy molekuláris biológiai motivációjú szoftvercsomag, amelyet széles körben használnak biológiai hálózatok elemzésére és vizualizációjára.

[26] A transzkripció során a DNS-ben található információ – a DNS szekvencia – alapján messenger RNS készül.



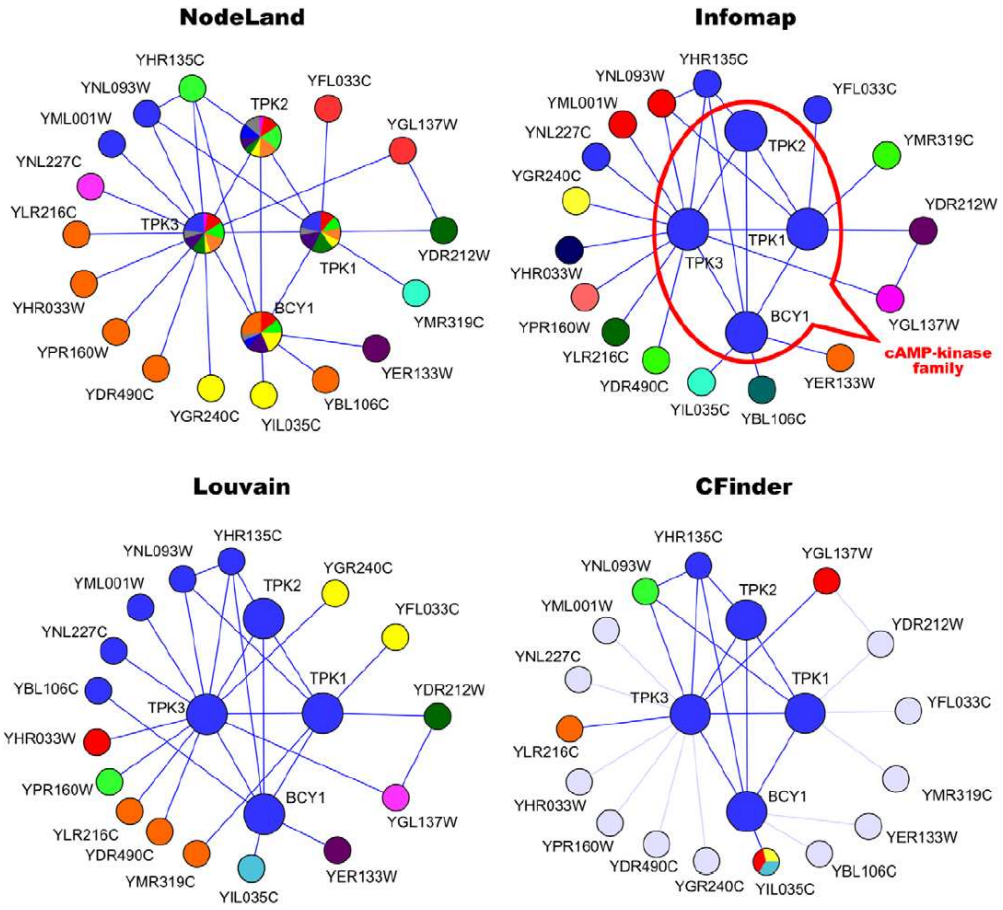
1.21. ábra. Az emberi fehérje-fehérje kölcsönhatások hálózatában Jonsson és Bates által azonosított modulok ($k = 6$ klikk méret paraméterrel). Az ábra átvétel a [116] publikációból. Az átfedő modulok esetén minden modult külön szín jelöl. Az egynél több modulhoz tartozó fehérjék csak az egyik modul színével vannak színezve. A modulok mellett található leírásokat a KEGG adatbázisban az egyes fehérjékhez rendelt funkciók alapján határozták meg a szerzők. A fő funkciókat vastag betűs kiemelés jelzi. A hálózatban a daganatos betegségekkel kapcsolatos fehérjéket háromszög alakú csúcspontok jelölik.

meg [18]. Collins és munkatársai (köztük van Krogan) 2007-es elemzése nagyrészt az előző két mérés sorozat „nyers” eredményeit összesíti részletesebb és frissebb bioinformatikai módszerekkel [121]. Mindhárom publikáció élesztőgombával dolgozott. A Gavin, Krogan és Collins első szerzőségével készült elemzések valamint a BioGrid adatbázis PPI adataira optimalizálva készült a ClusterONE hálózati modulkereső algoritmus, amely a szerzői elemzése szerint ezen adatsorok mindegyikén a rendelkezésre

álló módszerek közül a legjobb eredményt adja [107]. A közvetlenül fehérje-fehérje kölcsönhatásokat valamint a komplex-eket azonosító méréseken túl jelentős eredmények születtek a korábban már említett „genetic interaction” mérésekben is. Például Costanzo és munkatársai 2010-ben az élesztőgomba eddig ismert legrészletesebb genetikai kölcsönhatás térképét publikálták [29], valamint Babu és munkatársainak 2014-es publikációja az *Escherichia coli* fehérje komplex-ek genetikai kölcsönhatásait elemezte mérésekkel [122].

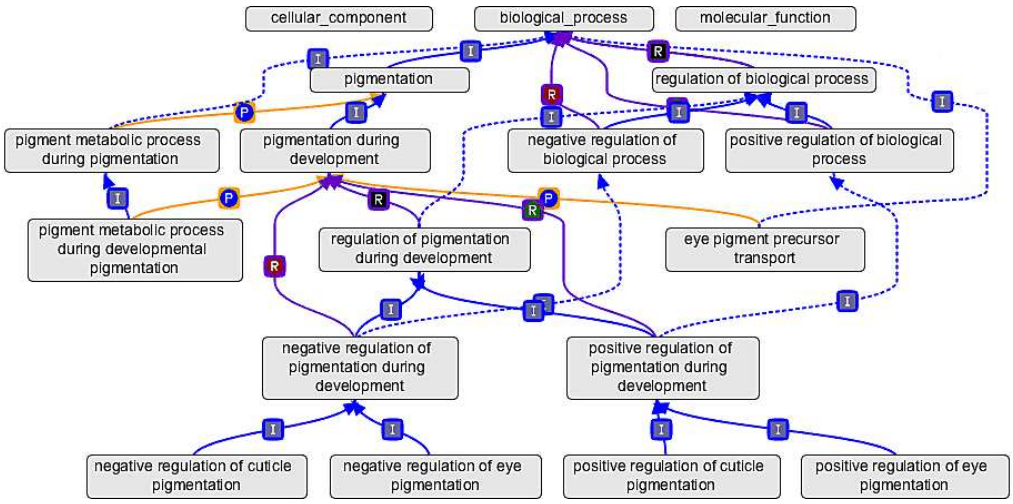
A klikk perkolációs módszer alkalmazásainak rövid leírása és a konkrét módszerek tömör bemutatása után fontos megemlíteni, hogy a PPI hálózatokban (és hasonlóan más molekuláris biológiai és nem biológiai hálózatokban) a magas lokális élsűrűségtől eltérő kritériumokat is lehet használni a modulok keresésére. A további módszerek szükségességét indokolja a sűrű részgráfok keresésének a következő, biológiai szempontból jelentős lehetséges szisztematikus hibája. A sűrű részgráfok keresése – például a klikk perkolációs módszer – biológiai szempontból hibás lehet, ha egy fehérje kevés (például $(k-1)$-nél kevesebb) kapcsolattal rendelkezik. Azért, mert egy ilyen fehérje a módszer szerint egyáltalán nem lehet egy sűrű (például k-klikk perkolációs) modul része. Ennek a lehetséges hibának egy következménye az, hogy a speciális és kevés ismert kölcsönhatással rendelkező fehérjék csak olyan kiegészítő módszerekkel rendelhetőek modulokhoz, amelyek az él sűrűségen kívül további tulajdonságokat is figyelembe vesznek. Egy lehetséges egyszerű megoldás minden ilyen (kevés kölcsönhatással rendelkező) fehérje hozzárendelése ahhoz a modulhoz (vagy azokhoz a modulokhoz), amelynek a fehérjéihez az adott, kevés kapcsolatú fehérjét az ismert néhány kölcsönhatása hozzákapcsolja. Mindez a gyakorlatban úgy valósítható meg, hogy a meglévő sűrű tartományokhoz – több egymás utáni lépésben – mindig hozzáadjuk a velük kapcsolatban lévő, modulhoz még nem besorolt első szomszéd csúcsokat. Egy lehetséges további – hálózati modulokat kereső, de élsűrűséget nem használó – módszer definíciója a következő. Az A és B fehérje azonos csoportba (modulba) tartozik, ha az A-val kölcsönható fehérjék listája hasonló a B-vel kölcsönható fehérjék listájához [27] [123].

[27] Két csúcs kölcsönhatási listájának hasonlóságára a szakirodalomban egy gyakori megnevezés a „topological correlation”, ami mérhető például a két csúcs szomszéd listájának (két halmaznak) a Jaccard-korrelációjával.



1.22. ábra. Az élesztőgomba cAMP-függő protein kináz családjába tartozó fehérjék modul besorolásai 4 különböző PPI hálózati modulkereső módszerrel (a [111] publikáció 2. ábrájának B panelje). A bal felső ábrarész a ModuLand kereső módszerhez tartozó NodeLand algoritmus eredményét mutatja, a jobb felső részen Rosvall és Bergstrom Infomap módszerének eredménye látható [100]. Az alsó két részlet Blondel és munkatársainak „Louvain" algoritmusával [101] valamint a (módosított) CFinder-rel kapott eredményt mutatja [T2]. Jól látható, hogy a módszereket a szerzőik eltérő célok irányában optimalizálták. A NodeLand, az Infomap és a Louvain módszer esetén a hálózat modulokkal való lefedettsége nagy, és az eredmény összetett. A CFinder esetén a lefedettség kisebb, és az eredmény véleményem szerint jobban áttekinthető.

Az alfejezet végén megemlítem, hogy az egymással átfedő modulok által meghatározott hálózatban a [T1] publikációnk definiálta – az irodalomban elsőként – a hálózati modulok fokszámát: egy modul fokszáma a vele átfedő további modulok száma. További, a dolgozatban is felhasznált mennyiségek a modulok mérete (a modulban lévő fehérjék száma) és az egyes fehérjék moduljainak száma (hány modul tartalmazza az adott fehérjét).



1.23. ábra. Példa fehérjék tulajdonságai közötti tartalmazási kapcsolatok hálózatára. A tartalmazási kapcsolatok egy irányított körútmentes fába (directed acyclic graph, DAG) vannak rendezve. Az ábra a Gene Ontology (GO) adatbázis [74] „Biological Process" nevű DAG-jának egy olyan részét mutatja, amelyen a „pigmentation" (pigmentáció) és a „regulation of biological process" nevű címszavak „alatt" található speciálisabb biológiai folyamatok tartalmazási kapcsolatai láthatóak. Az ábrán minden nyíl felfelé mutat, egy speciálisabb funkció felől egy általánosabb („szülő") funkció felé, amelyik a speciálisabb funkciót tartalmazza. Ennek megfelelően a legáltalánosabb címszavak az ábra tetején találhatóak – „cellular component" (CC), „biological process" (BP), „molecular function" (MF) – és a legspeciálisabbak az ábra alján szerepelnek. Mindegyik címszónak lehet egynél több „szülő" címszava, és egy címszótól egy fölötte lévő másik címszóhoz több útvonal is vezethet. A GO projekt honlapjáról letölthető mindhárom teljes táblázat (CC, BP, MF) és számos élőlény összes fehérjéjéhez a (félig) manuálisan előállított GO címszó lista. Az ábra átvétel a Gene Ontology adatbázis által rendszerezett biológiai tulajdonságok leírásából: http://geneontology.org/page/ontology-structure.

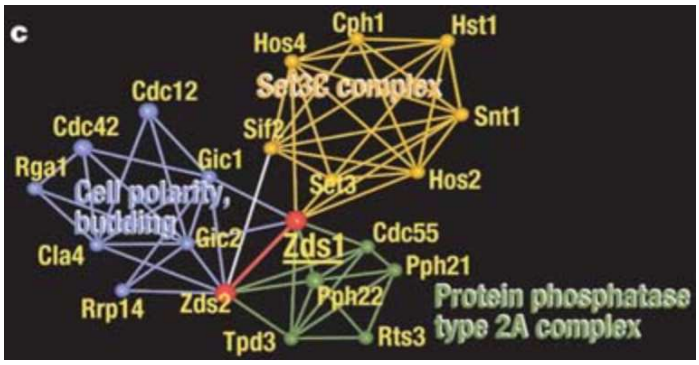
1.2. PPI hálózatok átfedő moduljainak biológiai funkciói [T1, T2]

A fehérje-fehérje kölcsönhatási hálózatokban a hálózat szerkezete („topológiája") alapján azonosított átfedő modulok biológiai értelmezéséhez szükség van a fehérjék biológiai tulajdonságainak és feladatainak szisztematikus felsorolására. Egy ilyen felsorolásban megvizsgálható, hogy egy kiválasztott modulra melyik tulajdonságok a leginkább jellemzőek. Az egyes fehérjék funkcióinak felsorolása (az aktuális ismeretek alapján) évtizedek óta megtalálható a nagy fehérje adatbázisokban, például a UniProt adatbázisban és elődeiben [49]. Viszont idővel nyilvánvalóvá vált, hogy a funkciók egymással erősen kapcsolatosak, egymással átfedhetnek és akár egymás részei lehetnek. Például a „pigmentáció" és a „fejlődés során történő pigmentáció szabályozása"

közül az utóbbi része az elsőnek, de egyben része például a „fejlődés szabályozása” funkciónak is. Emiatt szükséges a funkciók hierarchiába rendezése. Természetesen a fehérjéknek a (molekuláris) biológiai funkcióikon túl számos további tulajdonsága van. Ezek közé a további tulajdonságok közé tartozik például a sejtbeli hely (kompartment belseje, membrán felszíne, stb.), ahol a fehérjék működnek. Az itt felsoroltaknak megfelelően a Gene Ontology (GO) nevű adatbázis [74] az összes megismert fehérje tulajdonságot három irányított aciklikus (körútmentes) fába rendezi. Ezekből néhány részletet mutat példaként az 1.23. ábra.

A GO által használt három aciklikus fa a „Molecular Function” (MF), „Biological Process” (BP) és a „Cellular Component” (CC). Bár a DAG (directed acyclic graph) a neve alapján egy fa, ábrázolni úgy szokás, hogy a „gyökér pont” (a legáltalánosabb címke, amelyik az összes többi jelentését részhalmazként tartalmazza) felül van, és a fa lefelé elágazó ágai tartalmazzák mindig az egyre speciálisabb címkéket. Tehát (i) a DAG rendelkezik egy „gyökér” ponttal, ami a fában „alatta” lévő összes tulajdonságot tartalmazza, (ii) a DAG-ban két pont között lehet egynél több irányított útvonal és (iii) a DAG-ban nem létezik irányított körút (olyan irányított út, ami a saját kezdőpontjába visszavisz). A második megállapítás szemléletesen azt jelenti, hogy a „pigmentáció”-nak (az „A” tulajdonságnak) lehet többféle módon – köztes szinteken keresztül – részhalmaza a „fejlődés során történő pigmentáció szabályozása” („B” tulajdonság). A harmadik megállapítás szemléletes jelentése az, hogy a GO-ban két tulajdonságok között a tartalmazás csak egy irányban lehetséges: ha a „B” tulajdonság részhalmaza (altípusa) az „A” tulajdonságnak, akkor a két tulajdonság között ez a kapcsolat fordított irányban tilos.

Miután egy hálózati modulkereső módszer azonosította a fehérje-fehérje kölcsönhatási hálózat moduljait és a Gene Ontology segítségével sikerült felsorolni a modulok fehérjéinek tulajdonságait, a következő tennivalónk annak megállapítása, hogy egyegy modulban a fehérjéknek mik a statisztikailag legjelentősebb közös tulajdonságai. A [T1] és [T2] publikációnkban az élesztőgomba 2004-es „DIP core” [51] fehérje-fehérje kölcsönhatási hálózatát használtuk. Ebben a CFinder-rel azonosítottunk PPI hálózati modulokat, majd a Gene Ontology alapján felsoroltam a fehérjék „Biological Process” (biológiai folyamat) tulajdonságait. Ezután az egyes modulok által végzett biológiai folyamatok megállapítására a egy, a konkrét feladatra speciálisan kidolgozott statisztikai program csomagot [124] használtuk. A program csomag Perl programozási



1.24. ábra. Az élesztőgomba fehérje-fehérje kölcsönhatási hálózatában [51] a klikk perkolációs módszerrel azonosított modulok közül a ZDS1 fehérjét az ábrán látható három modul tartalmazza. Az ábra a „DIP core” nevű adatsor egy 2004-es verzióját használja. A három modulhoz tartozó fehérjék (hálózati csúcspontok) és kapcsolatok színe sárga, zöld és lila. Az egynél több modul által tartalmazott csúcsok és él színe piros. A három modul statisztikailag legjelentősebb funkcióját a GO::TermFinder szoftver csomaggal állapítottam meg az élesztőgomba fehérjéinek Gene Ontology (GO) annotációi alapján. A három csoporton belül a jelentős közös funkciók: „Set3C complex”, „Cell polarity, budding” és „Protein phosphatase type 2A complex”. Az ábra a [T1] publikációnk 2c ábrájának másolata. A felhasznált CFinder szoftvert Palla Gergely programozta, szintén ő rajzolta meg az ábrát a glay nevű ábrázoló program felhasználásával. Az adatok és az adat átalakítások (kölcsönhatások, fehérje név átalakítások és funkciók) tőlem származnak.

nyelven implementált modulként elérhető, a Gene Ontology adatokat kezelő modulok csoportjába tartozik, a neve GO::TermFinder. A PPI hálózati modul keresés és csoport funkció azonosítás után kapott eredményeinkre az 1.24. ábra mutat egy példát.

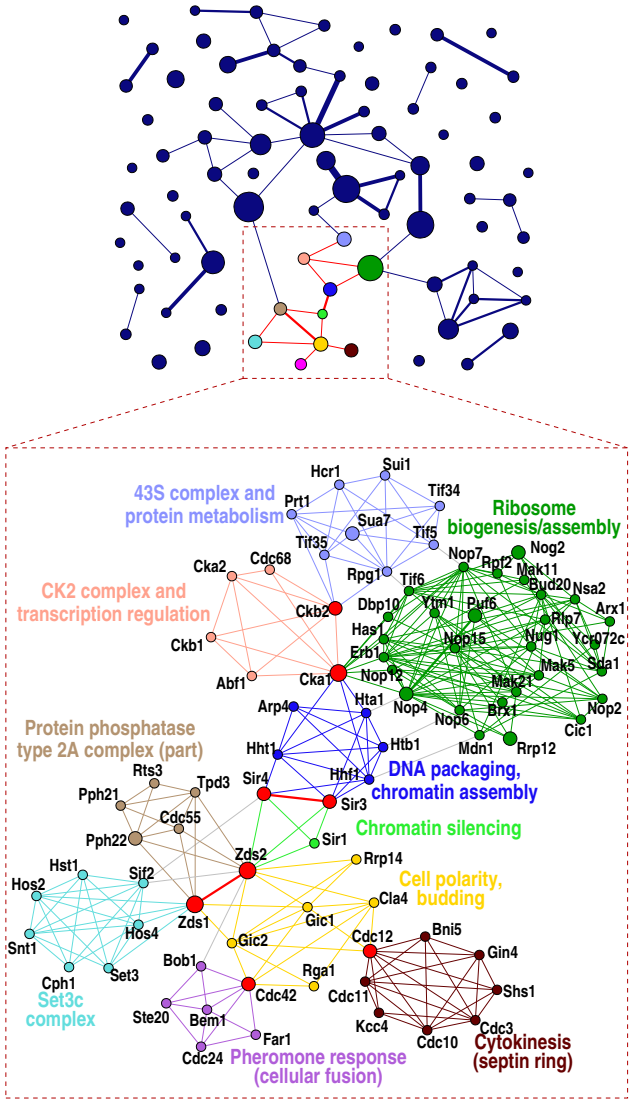
Egy fehérje csoportban (például egy PPI hálózati modulban) egy kiszemelt biológiai tulajdonság előfordulási gyakorisága mikor jelentős (szignifikáns)? Akkor, ha ez a biológiai tulajdonság az adott élőlény összes fehérjéje között lényegesen kisebb gyakorisággal fordul elő, mint a vizsgált fehérje csoportban. Használjuk a következő jelöléseket. A kijelölt biológiai tulajdonság (például a „pigmentáció”) egy n elemű fehérje csoporton belül pontosan k darab fehérjénél fordul elő, valamint ezzel a tulajdonsággal az adott élőlény összes vizsgálható N darab fehérjéje közül pontosan M darab rendelkezik. Első közelítésként vizsgáljunk egy egyszerű – de nem realisztikus – esetet, amikor minden tulajdonságnak a fehérjéken való elosztása független az összes többi tulajdonság elosztásától és minden fehérjének az összes többi fehérjétől független módon lehetnek tulajdonságai. Ebben az esetben az összesen N darab fehérje közül az adott tulajdonsággal rendelkező M darab fehérjét $\binom{N}{M}$ azonos valószínűségű lehetőség közül választhatjuk ki. Ezek közül $\binom{n}{k}\binom{N-n}{M-k}$ olyan lehetőség van, amelyikben

a vizsgált n elemű fehérje csoport k eleme rendelkezik a tulajdonsággal. Így az N, M és n paraméterek ismeretében a k változó 0 és n közötti lehetséges értékeinek $p_{hyp}(k)$ valószínűségét a hipergeometrikus eloszlás adja meg:

$$p_{hyp}(k) = \frac{\binom{n}{k}\binom{N-n}{M-k}}{\binom{N}{M}} \tag{1.3}$$

Mivel a $p_{hyp}(k)$ függvény a nagy k értékeknél (vezető rendben) $1/k!$ szerint csökken, ezért egy adott k érték jelentősen (szignifikánsan) nagyobb a szokásos értékeknél akkor, ha a $p_{hyp}(k)$ „eléggé" alacsony. Részletesebben: ha a k érték növelésével a $p_{hyp}(k)$ függvény elért egy eléggé alacsony értéket, akkor már az annál nagyobb k értékekre vett összes valószínűsége is igen kicsi lesz. Általános esetben az eloszlás nem feltétlenül gyorsan csökkenő, ezért az eloszlásnak egy adott pontban mérhető alacsony értéke helyett a szignifikancia kritérium az szokott lenni, hogy az eloszlásnak az adott ponttól kezdve „felfelé" számolt integrálja elég kicsi legyen. Ez a P érték (P-value) nevű statisztikai mennyiség elve. A P-value segítségével becsülhető, hogy ha egy mérés számszerű eredményének eloszlása egy ismert valószínűségi eloszlás, akkor a mérés során egy konkrét mért x érték mennyire szignifikáns (jelentős). Pontosabban megfogalmazva: ha hipotézisként feltételezzük, hogy a mért x érték jelentős, akkor tesztelhetjük, hogy milyen eséllyel helyes ez a hipotézis.

Vizsgáljuk csak azt az esetet, amikor az ismert eloszlás csak a nemnegatív valós számokon van értelmezve. „felfelé" lévő tartományon vett integrálja, és a P-value azt mondja meg, hogy az ismert eloszlás alapján mi annak a valószínűsége, hogy a mérési eredmény x vagy annál nagyobb szám legyen. Így például az 1.3 egyenletben szereplő hipergeometrikus eloszlás esetén egy mért k értékhez tartozó P-value a következő: $\sum_{i=k\ldots n} p_{hyp}(i)$. Minél kisebb a P-value, annál jelentősebben nagy a mért x érték. A szignifikáns P értékek felső küszöbét α értéknek szokás nevezni. A biológiában elterjedt az $\alpha = 0.01$-nél kisebb P értékek szignifikánsnak tekintése. A PPI hálózati modulok esetében általában egynél több hipotézis tesztelése történik. Ennek figyelembe vétele érdekében az α küszöb értéket szokás módosítani. A legegyszerűbb korrekció az α-nak a hipotézisek (egy modulban tesztelt tulajdonságok) számával történő osztása (Bonferroni korrekció).



1.25. ábra. **Felső részábra.** Az élesztőgomba (*Saccharomyces cerevisiae*) fehérje-fehérje kölcsönhatási (PPI) hálózatának átfedő moduljai. Minden kör egy fehérje modult jelöl, és minden kör területe arányos a modulban lévő fehérjék számával. A köröket összekötő élek vastagsága arányos a két modul által közösen tartalmazott fehérjék számával, azaz a két modul átfedésének méretével. A kölcsönhatási adatok forrása a DIP (Database of Interacting Proteins) [51]. A hálózati modulokat a CFinder szoftver azonosította a dolgozat szövegében leírt klikk perkolációs módszerrel [93, T2], az optimális klikk méret itt $k = 4$. **Alsó részábra.** A felső ábrán színessel jelölt hálózati modulok kinagyítva. A piros szín a modulok közötti átfedéseket jelöli. Mindegyik modul mellett megtalálható a modul fehérjéinek funkciói [74] közül az a legszignifikánsabb közös funkció, amelyet a GO::TermFinder statisztikai programcsomag azonosított [124]. Az ábra átvétel a [T1] publikációnkból. A modul keresést Palla Gergely végezte, az adatok összeállítását és feldolgozását én végeztem. Az ábrát részben Palla Gergely készítette, részben én készítettem.

Azonban még ez a korrekció sem veszi figyelembe, hogy a tesztelt tulajdonságok gyakran egymástól erősen függnek, például gyakori az olyan típusú eset, amikor azonos modulon belül egy fehérjének funkciója a „fejlődés során történő pigmentáció szabályozása” és egy másik fehérjének funkciója az ezt tartalmazó „pigmentáció”. Tehát általában minden modulban több, egy mással kapcsolatos biológiai tulajdonság előfordulási számáról kell megvizsgálni, hogy jelentős-e. Ennek az egyik következménye az, hogy a címkék (funkciók) elosztásának egyenletessége erősen sérül, tehát az N fehérje közül az adott tulajdonsággal rendelkező M darab kiválasztásának $\binom{N}{M}$ lehetősége nem azonos valószínűségű. Annak eldöntésére, hogy a hipergeometrikus eloszlásból kapható P értékek (a Bonferroni korrekcióval) megfelelőek-e, a GO::TermFinder statisztikai programcsomag numerikus elemzést is végez. Megvizsgálja, hogy az N darab fehérje közül 1000-szer függetlenül véletlenszerűen kiválasztott n darab fehérje alapján számolt k eloszlás a konkrét mért k értékre hasonló P-value-t ad-e, mint az elméleti hipergeometrikus eloszlás a Bonferroni korrekcióval. Az elemző program kimenete mindkét (elméleti és numerikus) eredményt megadja. A numerikus szimuláció lényegesen lassabb és az általunk vizsgált moduloknál az elméleti és a numerikusan számított P érték általában 1-3 nagyságrenddel eltért.

A 1.25. ábra az élesztőgomba 2004-es DIP „core” (magas megbízhatóságú) fehérje-fehérje kölcsönhatási hálózata alapján a klikk perkolációs módszerrel kiszámított átfedő modulokat mutatja és a modulokhoz a GO::TermFinder segítségével hozzárendelt funkciókat. Az azonosított modul hálózat lehetővé teszi a fehérje-fehérje kölcsönhatások nagyskálájú áttekintését. A modulok hálózatában egy csúcspont egy modult jelöl, és két csúcspont között van kapcsolat (él), ha a két modul egymással átfed (van olyan fehérje, amelyet mindkét modul tartalmaz). A moduloknak ezt a típusú hálózatát csak olyan hálózati modulkereső módszerrel lehet kiszámítani, amely a megengedi a modulok közötti átfedéseket.

1.3. Fehérjék modul száma PPI hálózatok átfedő moduljaiban [T1]

Közismert, hogy a társadalmi, technológiai, biológiai és más összetett hálózatok jelentős részében a csúcsok fokszámainak eloszlása gyakran hatványfüggvény szerint csökken és a nagy fokszámoknál exponenciális levágásban végződik [66, 125]. A széles (hatványfüggvény) eloszlás fő oka az, hogy a résztvevő csúcsok szempontjából a kapcsolatok alacsony „költséggel” járnak. A fehérje-fehérje kölcsönhatási hálózatokban a magas kapcsolat számok előfordulását a következő (speciálisan PPI hálózati)

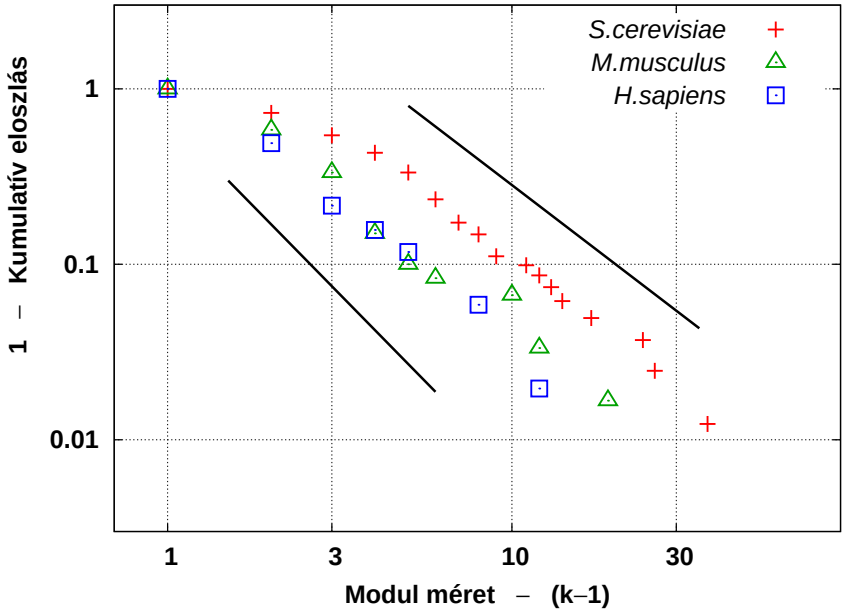
tulajdonság teszi gyakorivá. A PPI hálózat egy csúcspontja nem egyetlen konkrét molekulát jelöl – tehát a fehérje egyetlen, fizikailag rögzített megvalósulását, – hanem összesítve jelöli az adott fehérje típusba tartozó összes molekulát.

Ha egy PPI hálózatban egy csúcspontnak sok szomszédja van – tehát az adott fehérjének sok lehetséges fehérje kölcsönható partnere van, – akkor a szomszédok között általában nagyobb eséllyel találhatóak jelentősen eltérő funkciójúak. Ha egy modulkereső módszer megenged a modulok között átfedéseket, akkor a modulkeresés eredményeként számíthatunk olyan fehérjékre, amelyek több modulhoz tartoznak. Mindez biológiai szempontból azért fontos, mert számos élőlény esetén konkrét példákon keresztül ismert, hogy több génjük (fehérjük) kettő vagy több, lényegesen eltérő funkciót befolyásol. Olyan esetre is vannak részletesen vizsgált példák, hogy egy gén több olyan fenotípust befolyásol, amik látszólag teljesen függetlenek. Ennek az esetnek a neve „pleiotropy” [126]. Emiatt ha egy modulkereső módszer nem enged meg átfedéseket, akkor sok (többféle csoporttal együttműködő) fehérje szerepeinek azonosítását csak jelentős hibákkal tudja elvégezni.

A fehérjék modul számainak megismerése érdekében vizsgáljuk meg, hogy a korábban már említett Database of Interacting Proteins (DIP) adatbázisban a klikk perkolációs módszerrel azonosított PPI hálózati modulok alapján egy fehérje átlagosan hány modulhoz tartozik[28]. Az elemzést a DIP-ből arra a három – széles körben vizsgált élőlényre – végeztem el, amelynél a rendelkezésre álló PPI hálózat csúcsainak (fehérjéinek) száma 1000 felett van: élesztőgomba (2393 csúcs és 5008 él), házi egér (*Mus musculus*, 1879 csúcs és 1982 él) és ember (3708 csúcs és 5376 él). Élőlény-specifikus adatbázisokban mindhárom élőlényhez elérhető ezeknél nagyobb PPI hálózat. A konkrét példában a DIP adatbázis használatának előnye, hogy egységes adatokat biztosít a három élőlényre.

Az élesztőgomba, a házi egér és az ember PPI hálózatában a legalább egy modullal rendelkező fehérjék átlagos modul száma rendre 1.12, 1.09 és 1.10 a klikk perkolációs módszer szerint. Önmagukban ezek az 1.1 körüli értékek nem tűnnek magasnak, ezért az eredménynek az értelmezéséhez fontos hozzátenni, hogy a három szervezetben a legtöbb modulban részt vevő fehérjék rendre 4, 3 és 6 modulban vannak jelen. Ezek a sok kapcsolattal rendelkező fehérjék az élesztőgombában az RPB1 (egy RNS polimeráz alegység), az egérben a BRG1 (egy transzkripciós faktor), a Hdac1 (egy hisztон módosító fehérje) és az Ripk1 (receptorhoz kötött szerin/treonin kináz), valamint

[28] Ezeket az adatokat a DIP 2014. október 1-ei verziójából töltöttem le, és a velük végzett számításokat a doktori dolgozathoz készítettem.

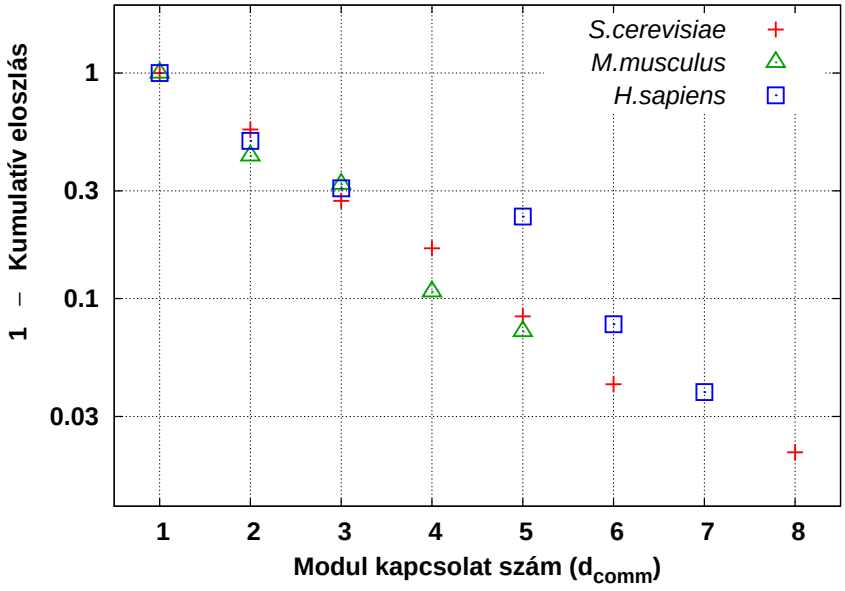


1.26. ábra. A Database of Interacting Proteins fehérje-fehérje kölcsönhatási hálózataiban (2014. október 1-ei verziók) a klikk-perkolációs módszerrel számolt átfedő hálózati modulok méretének eloszlása. A három élőlény, amelyre az ábra eredményeket mutat: *S. cerevisiae* (élesztőgomba), *M. musculus* (házi egér) és *H. sapiens*. A klikk méret paraméter az élesztőgomba és az ember esetén $k = 4$, a házi egérnél $k = 3$. Az ábrán a két folytonos vonal egy-egy hatványfüggvényt mutat. Az adat pontok felett lévő hatványfüggvény kitevője -1.5, az alul lévő hatványfüggvény kitevője -2. Tehát a modul méretek kumulatív eloszlása lassú lecsengésű, és egy -1.5 és -2 közötti exponensű hatványfüggvény szerint konvergál 1-hez (az ábra az 1-ből kivont kumulatív eloszlást mutatja). Emiatt a modul méretek sűrűségfüggvénye (p.d.f.) egy -2.5 és -3 közötti kitevővel lecsengő hatványfüggvényhez hasonlóan csökken a nagy modul méretek tartományában. Az ábrát a doktori dolgozathoz készítettem.

emberben az UBA52. A dolgozat PDF fájljában ezekre a fehérje nevekre kattintva a fehérjék UniProt adatlapjain látható, hogy ezeknek a fehérjéknek – az RPB1 kivételével – a Gene Ontology szerint sok funkciója van, ezért valóban indokolt, hogy az átlagosnál nagyobb számú modulhoz tartoznak. Az RPB1 esetén valószínűleg az RNS polimeráznak a többféle transzkripciós faktor fehérje csoporttal való együttműködése okozza a sok modulban való részvételt.

1.4. Átfedő PPI hálózati modulok mérete [T1]

Az 1.25. ábra felső paneljén mutatott modul hálózat alapján a klikk perkolációs módszer lehetővé teszi a fehérje-fehérje kölcsönhatási hálózatokban megtalálható információ „sűrítését” és ezáltal egy nagyobb, áttekintő térkép felrajzolását. Az eredeti



1.27. ábra. A Database of Interacting Proteins PPI hálózataiban a klikk-perkolációs módszerrel számolt átfedő hálózati modulok modul kapcsolat számainak eloszlása. A modulok hálózatában két modul között pontosan akkor van kapcsolat, ha az eredeti PPI hálózatban a két modul egymással átfed, azaz van legalább egy közös fehérjéjük (hálózati csúcspontjuk). A vizsgált élőlények, az adatbázis verzió és a klikk méret paraméterek az 1.26. ábrán használtakkal egyezőek. Az ábra alapján a modul méretek kumulatív eloszlása gyors, exponenciálishoz hasonló lecsengésű. (Az ábra az 1-ből kivont kumulatív eloszlást mutatja.) Az ábrát a doktori dolgozathoz készítettem.

– fehérje-fehérje kölcsönhatási – hálózattal ellentétben ezen az áttekintő térképen már minden csúcsponthoz tartozik egy méret, a modul által tartalmazott fehérjék száma. Mivel a modulok biológiai feladatokat ellátó egységeknek (komplex-eknek vagy funkcionális moduloknak) felelnek meg, ezért a méreteik eloszlása az elvégzett feladatok összetettségéről is tájékoztathat. Az 1.26. ábra eredményei alapján a DIP-ből elérhető három friss PPI hálózat moduljainak a méret eloszlása (a sűrűségfüggvény, p.d.f.) egy -2.5 és -3 közötti kitevőjű hatványfüggvényként cseng le. Ennek a lassan csökkenő eloszlásnak az egyik oka valószínűleg az, hogy a molekuláris biológiai funkciók között előfordulnak kiugróan nagy összetettségűek.

1.5. PPI hálózatok átfedő moduljainak kapcsolat számai [T1]

Mivel mind a fehérjék kapcsolat számai (a PPI hálózati csúcsok fokszámai), mind a fehérjék által alkotott modulok mérete hatványfüggvény eloszlást követ, tehát nincsen karakterisztikus fokszám a hálózatban. Ezért felmerül a kérdés, hogy vajon a

PPI modulok hálózatában is megfigyelhető-e a hatványfüggvény szerinti fokszám eloszlás. Másként fogalmazva: igaz-e, hogy a fehérje-fehérje kölcsönhatási hálózatban talált modulok által alkotott hálózatban az eredeti hálózathoz hasonlóan a kapcsolatok számának nincsen karakterisztikus értéke (jellemző értéke), csupán egy felső levágási küszöbe? Az 1.27. ábra alapján a vizsgált három élőlényben a PPI hálózat átfedő moduljainak hálózatában nem teljesül a hatványfüggvény szerinti fokszám eloszlás. Ezzel szemben az eloszlás exponenciális (gyors) lecsengést követ. A megjelenő karakterisztikus hosszt a klikk perkolációs módszer által használt k modul sűrűség paraméter okozza. A modul sűrűség paraméter rögzítésével meghatározott modulok fokszámát egy csúcs átlagosan δ-val növeli. Így az 1.27 ábrán megfigyelhető exponenciális függvényekhez tartozó karakterisztikus méret skála a $k\delta$. Fontos, hogy ez a δ érték nem az előző alfejezetben említett átlagos modul számnál 1-gyel kisebb érték, mert ha egy csúcs tagja x darab modulnak, akkor mind az x darab modul fokszámát növeli $(x-1)$-gyel.

Az eredmények összefoglalása

A fejezetben ismertetett eredmények közül a társszerzőségemmel készült publikációk fehérje-fehérje kölcsönhatási hálózatokra vonatkozó eredményei valamint a dolgozathoz külön készült eredmények (számítások és ábrák) a következők:

- Kidolgoztunk egy hálózati modulkereső módszert, amely egymással átfedő és a (hálózati) környezethez képest nagyobb belső élsűrűséggel rendelkező hálózati modulokat azonosít. Javaslatot tettünk a módszerben felhasznált k-klikk méret paraméter lehetséges értékei közül az optimális érték kiválasztására. A modulkereső CFinder alapkutatási program a [T1] publikációink 2005-ös megjelenése óta non-profit célra ingyenesen elérhető, és a felhasználók a weboldalon (Manual, FAQ) és e-mailben támogatást kapnak. A CFinder tartalmazza az általunk kidolgozott súlyozott klikk perkolációs módszert. A CFinder-t 2009 óta 8500 feletti egyedi email című felhasználó töltötte le, (becslésem szerint a 2005-2009 közötti időszakban 2-3000 különböző felhasználó). tehát összesen 10 ezernél több.

- A k-klikk-perkolációs módszerrel azonosítottuk az élesztőgomba (*S. cerevisiae*) fehérje-fehérje kölcsönhatási (PPI) hálózatának átfedő moduljait és azonosítottam a modulok statisztikailag legszignifikánsabb biológiai funkcióit.

- Megállapítottam, hogy három vizsgált élőlény friss PPI hálózataiban a modulokban található csúcsok átlagos modul száma 1.1 és 1.2 közötti, és a három élőlény legtöbb modullal rendelkező fehérjéi az átlagosnál nagyobb számú biológiai funkcióval rendelkeznek.

- A három élőlény fehérje-fehérje kölcsönhatási hálózatában az azonosított modulok méretének eloszlása hatványfüggvény-szerű, széles eloszlás. Ez az eloszlás típus azonos a PPI hálózat csúcsainak fokszámánál ismert eloszlás típusával.

- Az azonosított átfedő PPI hálózati modulok segítségével definiáltunk egy új mennyiséget, a modulok fokszámát: egy modul fokszáma a vele átfedő modulok száma. A modulok közötti átfedések segítségével definiáltuk a modulok hálózatát. A modulok hálózatában egy csúcspont az eredeti hálózat egy modulja, és egy él az eredeti hálózat két kijelölt modulja közötti átfedést jelöli.

- A modulok hálózata az eredeti fehérje-fehérje kölcsönhatási hálózatot jóval tömörebb formában mutatja be. Ebben a tömörített (átskálázott) hálózatban azt találtuk, hogy az eredeti PPI hálózattól eltérően a fokszám eloszlás kezdeti szakasza gyorsan lecsengő, exponenciális típusú, majd az eloszlás vége hatványfüggvény-szerű. Ennek az exponenciális lecsengésnek az oka a klikk méret paraméter, amely egy karakterisztikus csúcs számot okoz, és ezen keresztül egy karakterisztikus modul kapcsolat számot (modul fokszámot) jelent.

Saját hozzájárulásom az eredményekhez

A fejezetben ismertetett eredményekhez a következőkkel járultam hozzá:

- Részt vettem egy hálózati modulkereső módszer kidolgozásában, amely egymással átfedő és a környezethez képest nagy belső élsűrűséggel rendelkező hálózati modulokat azonosít.

- A klikk perkolációs módszer teszteléséhez hálózatokat gyűjtöttem: PPI, szó asszociációs és tudományos társszerzőségi hálózatokat. Az adatokat az eredeti formájukból hálózatokká alakítottam és az adat hibákat szűrtem. Beprogramoztam nagy hálózatokra alkalmazhatóan a(z irányítatlan) csúcs köztiség (node betweenness) számítását és az agglomeratív Girvan-Newman klaszterezést. A [T3] cikkünkben a súlyozott klikk perkolációs módszer vizsgálata érdekében analitikusan levezettem azt a függvényt, amely a perkolációs pontban a módszer két paraméterét egymás függvényeként kifejezi. Ezt az eredményt ugyanebben a cikkben a társszerzők numerikus eredményei igazolták.

- Az adatok feldolgozásához és az eredmények elemezhetőségéhez átalakítottam a PPI hálózatokban szereplő gén és fehérje neveket olyan név típusra, amelyet a funkciókat felsoroló Gene Ontology adatbázis használ. A PPI hálózati modulok kiszámítása után részben automatizáltam a csoportok legszignifikánsabb közös funkcióinak szisztematikus keresését.

- A doktori dolgozathoz kiszámítottam friss adatokkal a fehérjék átlagos modul számát és megvizsgáltam a legtöbb modullal rendelkező fehérjék funkcióit. Szintén a doktori dolgozathoz kiszámítottam ebben a három friss hálózatban a modulok méretének eloszlását és a moduljaik hálózatában a fokszám eloszlást.

- A doktori dolgozathoz végzett számításokkal megállapítottam, hogy a vizsgált három élőlényben (*S. cerevisiae*, *M. musculus*, *H. sapiens*) a PPI hálózati modulok kapcsolat számának eloszlása teljes egészében exponenciális típusú, nincsen hatványfüggvény-szerű szakasza.

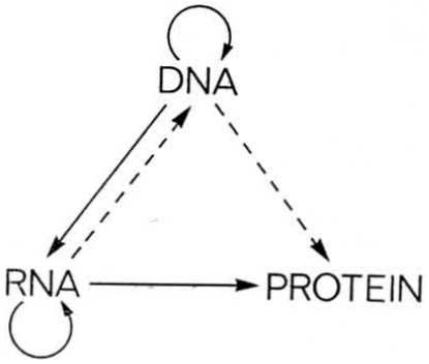
- A klikk perkolációs módszert (Clique Percolation Method, CPM) implementáló CFinder természettudományos szoftver csomag tesztelésében és a felhasználói visszajelzések nyomán történő fejlesztések egyeztetésében részt veszek a szoftver 2005-ös indulásától kezdve folyamatosan. A CFinder kulcsainak frissítését végző, valamint a letöltéseket kezelő és a felhasználói dokumentációt tartalmazó http://CFinder.org website karbantartója vagyok a 2005-ös indítás óta. 2005 óta a felhasználóktól folyamatosan érkeznek kérdések, ezek közül átlagosan heti egy kérdést én kezelek.

2. fejezet

Transzkripció és transzláció szabályozási hálózatok moduljai

A transzkripció során a DNS-ben található nukleotidok sorrendje (szekvenciája) által meghatározott információ alapján RNS molekulák készülnek. A molekuláris biológia fejlődése során a DNS szekvenciában tárolt információ legkorábban ismert formája a fehérjéket kódoló gén volt. A génekből az élő sejtekben először mRNS (messenger RNA, „hírvivő” RNS), majd aminosav lánc (fehérje) készül. Erre és további ismeretekre építette Francis Crick a molekuláris biológia centrális dogmájának nevezett elméletet 1958-ban, amit egy 1970-es cikkében módosítva ismét tárgyalt [127]. Az 1958-as centrális dogma szerint a három fő információ-hordozó biológiai molekula típus – DNS, RNS és fehérje – között az elvileg lehetséges 9-féle információ áramlási irány közül az élő sejtekben nem valósul meg mindegyik. Francis Crick 1958-ban az akkor rendelkezésre álló eredmények alapján felsorolta, hogy biológiai információ áramolhat a DNS→DNS, DNS→RNS és RNS→fehérje irányokban, továbbá (speciálisan RNS vírusok esetén) az RNS→RNS irányban. Crick leírja ebben a cikkben azt is, hogy talán lehetséges információ-áramlás az RNS→DNS és a a DNS→fehérje irányban, továbbá azt, hogy valószínűleg nincsen információ-áramlás a fehérje→fehérje, fehérje→RNS és a fehérje→DNS irányban.

Crick 1958-as írása óta ismertté vált, hogy az akkor még kevésbé vizsgált irányokban is létezik biológiai hatás. Például a centrális dogma által „tiltott” három kölcsönhatási irány esetén napjainkban már részletesen ismert példa a prion fehérjék között megvalósuló fehérje→fehérje irányú hatás [128], valamint az mRNS-t szerkesztő („alternative splicing”) fehérjék és a DNS-t javító fehérjék által megvalósított fehérje→RNS



2.1. ábra. A molekuláris biológia centrális dogmáját bemutató ábra (1958-ig ismert eredmények alapján). A folytonos vonallal jelölt kapcsolatok gyakoriak, a szaggatott vonallal jelölt kapcsolatok speciális esetekben lehetségesek. Az ábra átvétel Francis Crick 1970-es cikkéből [127].

és fehérje→DNS irányú hatás. Napjainkra az is ismertté vált, hogy a korábban csak vírusokban megfigyelt RNS→RNS kölcsönhatás a soksejtű élőlényekben is gyakori. Az RNS→RNS kölcsönhatásokban részt vevő RNS molekulák egyik csoportjának az általános neve „rövid gátló RNS”-nek (short interfering RNA). Ezek a molekulák képesek gátolni a messenger RNS-ből történő fehérje készítést.

Előzmények

Transzkripció szabályozási hálózatok

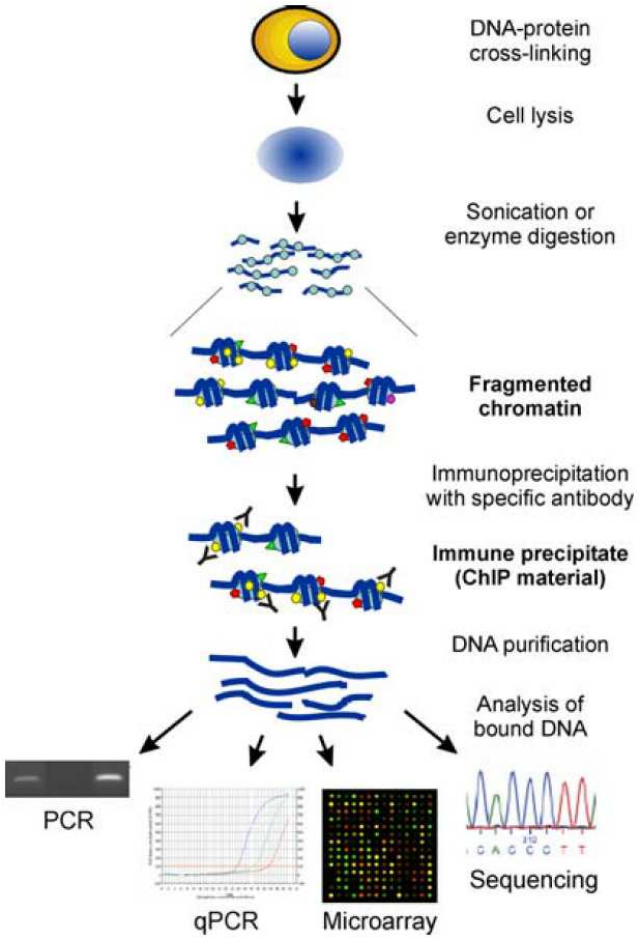
A DNS egy irányított lineáris molekula. A DNS szál irányát – a DNS vázát alkotó cukor típusú molekula rész szénatomjainak számozása alapján – 5’→3’ iránynak nevezik. A DNS szál iránya alapján szokás a „DNS-en egymás utáni” tartományokról beszélni, és ebben az értelemben gyakori az „előtt” (felett) és az „után” (alatt) hely megjelölés. A transzkripciót végző fehérje komplex (transzkripciós komplex) egy gén átírását úgy kezdi el, hogy hozzákapcsolódik a DNS-hez a gén 5’ vége előtt, majd ezután elvégzi az átírást a gén 5’ vége felől a gén 3’ vége felé haladva. A transzkripciós komplex DNS-hez történő kapcsolódását (kötését) számos tényező befolyásolja, ezek közül az egyik legfontosabb a transzkripciós faktor (TF) fehérjék hozzákapcsolódása a DNS-hez a gén 5’ vége előtt (felett). Egyes speciális fehérjék képesek segíteni vagy gátolni azt, hogy a transzkripciós komplex a DNS-hez kössön egy gén feletti szakaszon. Így ezek a speciális fehérjék közvetve módosítani (serkenteni vagy gátolni) tudják annak a génnek az átírási gyakoriságát, amelynek az 5’ vége felett a DNS-hez

kapcsolódik a transzkripciós komplex. Ezeknek a speciális fehérjéknek a neve transzkripciós faktor (TF) fehérje.

Ha egy TF fehérje képes egy „megcélzott” (target) gén előtti DNS szakaszhoz hozzákapcsolódni, és például csökkenti a target gén átírási gyakoriságát, akkor ezt hálózatos formában lehet úgy ábrázolni, hogy a TF fehérjétől a target génhez egy negatív (gátló) kölcsönhatást jelölő irányított hálózati él fut. A transzkripció szabályozási hálózatok elemzésekor egy gyakori egyszerűsítés az a feltételezés, hogy a gének és a fehérjék közötti megfeleltetés kölcsönösen egyértelmű, és ez alapján egy gén és a hozzá tartozó fehérje egyetlen hálózati csúcspontként ábrázolható. Ezzel az egyszerűsítéssel a transzkripciós szabályozási hálózat tekinthető a gének (vagy a fehérjék) közötti szabályozási hálózatnak, amelyben egy irányított él azt mutatja, hogy a kiindulási pontjánál lévő TF gén szabályozza a végpontjánál lévő célpont (target) gén átírásának gyakoriságát. Ezeknek az irányított (és súlyozott) hálózati éleknek a mérését mutatják be a dolgozat következő részei.

Fehérje-DNS kapcsolódást mérő kísérleti módszerek

A ChIP (Chromatin ImmunoPrecipitation, kromatin immunválasz) módszer célja a fehérjék DNS-en történő kötőhelyének meghatározása. Mivel a DNS-hez a transzkripciós faktorokon túl további fehérje típusok is gyakran hozzákapcsolódnak, ezért a ChIP módszer nem csak a TF-eket (és a TF-ek kötési helyeit) azonosítja, hanem például a DNS csomagolt szerkezetét biztosító hiszton fehérjéket és kötési helyeiket szintén. A ChIP módszer első lépése a fehérje-DNS kapcsolatok erős, de visszafordítható (reverzibilis) rögzítése. Ezután a DNS feldarabolása következik néhány száz és ezer bázispár közötti hosszúságú darabokra. A feldarabolás során a DNS darabok mindegyikén rajta maradnak a korábban oda rögzített fehérjék. Az egyes fehérjék – az egyedi szerkezeti elemeikre érzékeny – immunválasz segítségével rögzíthetőek egy lap (chip) előre megtervezett helyein. Ezután mindegyik rögzített fehérjéről leválasztható az addig vele erős kötésben lévő rövid DNS szakasz (fragmentum). A következő lépés a leválasztott DNS szakasz azonosítása (például szekvenálással). Ez a mérési lépés adja meg, hogy a vizsgált élőlény teljes DNS-ében pontosan hova (melyik gén fölé) kapcsolódott az a fehérje, amelyikkel a felismert DNS szakasz kapcsolódott. Az adatok számítógépes szűrése és további feldolgozása után az eredmény egy lista, amelynek minden eleme egy TF→target gén kapcsolat, és mindegyikhez tartozik egy szignifikancia érték.



2.2. ábra. A ChIP (Chromatin ImmunoPrecipitation) módszer bemutatása. A módszer célja a DNS-hez kötődő fehérjék és a kötési helyek azonosítása. A ChIP módszer (megfelelő mérés kiértékelési algoritmusok felhasználásával) alkalmas a transzkripciós faktor fehérjék és az általuk szabályozott gének azonosítására, és ezáltal a transzkripció szabályozási hálózat feltérképezésére. Az ábra átvétel a [129] publikációból.

A ChIP mérési módszerek két fő csoportja a ChIP-chip, amikor a DNS fragmentumok felismerése egy microarray típusú chip-en történik, illetve a ChIP-seq, amikor a DNS fragmentumok felismerése szekvenálással történik. A 2.2. ábra alsó sora mutatja ezt a két esetet „Microarray” és „Sequencing” felirattal. A ChIP módszerek előnye, hogy a mérés sokféle típusú sejtből kiindulva elvégezhető, tehát nem egyszerűen az vizsgálható vele, hogy milyen fehérje-DNS kapcsolatok lehetségesek, hanem sokkal inkább az, hogy az adott körülmények között (a mérés indításakor rögzített állapotú sejtekben) mik a ténylegesen megvalósuló fehérje-DNS kapcsolatok.

Fehérje-DNS kapcsolódás előrejelzése DNS szekvencia alapján

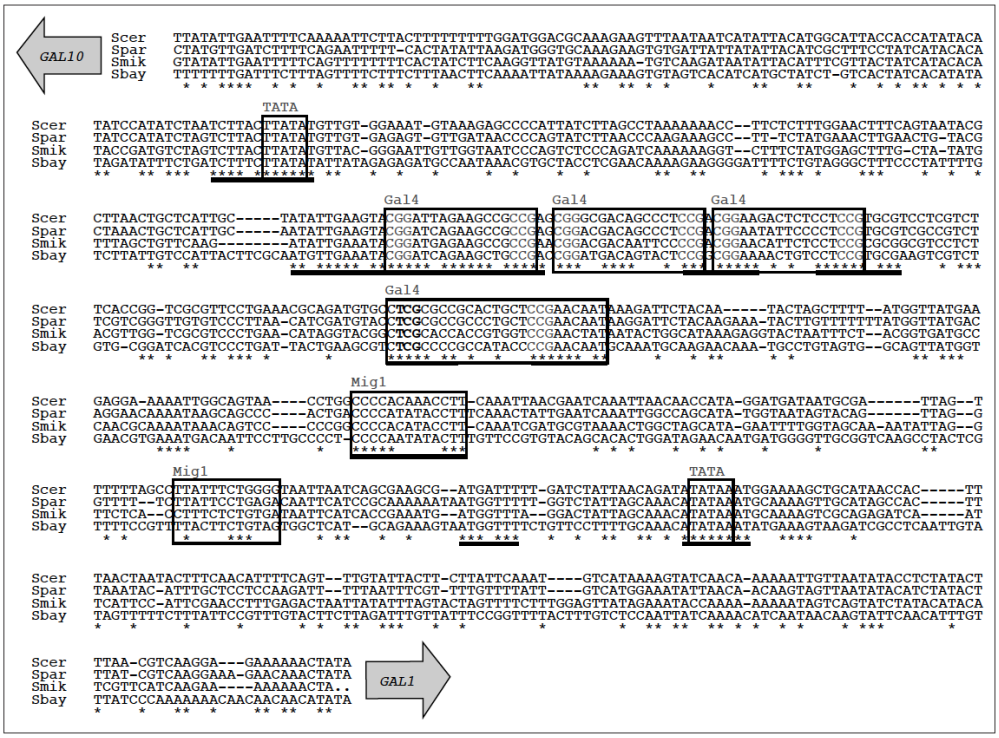
A transzkripciós szabályozási kölcsönhatások DNS alapú számítógépes előrejelzésének két fő eleme a (i) transzkripciós faktor (TF) fehérjék azonosítása és a (ii) TF fehérjék kötőhelyeinek felsorolása.

Egy kijelölt szervezet TF fehérjéinek felsorolásához első lépésként szükség van korábbról ismert és adatbázisokban felsorolt fehérje doménekre. A fehérje domének rövid aminosav lánc részletekből kialakuló, jellegzetes térszerkezetet felvevő molekularészek. Mivel a TF fehérjék mind DNS kötő fehérjék, ezért van bennük DNS-kötő fehérje domén. Az utóbbi évtizedek kutatásai nyomán már sok DNS-kötő fehérje domén ismert, és a szekvenciák hasonlósága segítségével további domének is besorolhatóak a jelenleg ismert nagyobb domén családokba. Ha egy fehérje rendelkezik a DNS-hez kapcsolódni képes doménnel, akkor (további feltételek teljesülése esetén) jó eséllyel TF fehérje lehet. A fehérje doméneket és domén családokat felsoroló két – széles körben használt – adatbázis a Pfam [130] és a SCOP [131]. Mindkettőt használja a transzkripciós faktor fehérjéket előrejelző (felsoroló) DBD (DNA-Binding Domain) adatbázis [132].

A TF fehérjék kötőhelyei (Transcription Factor Binding Site, rövidítve TFBS) jellemzően 5-15 bázispár (bp) hosszúságú szakaszok a DNS-en. Legtöbbször a szabályozott gén 5’ vége előtti (feletti) 500-1000bp hosszúságú tartományban találhatóak, de soksejtű szervezetekben előfordulnak a géneket tartalmazó DNS szakaszokon belül[1] és a génektől távol is. Mivel a TFBS-ek rövidek és sok helyen előfordulhatnak, ezért kísérleti adatok felhasználása nélkül (kizárólag nagyskálájú szekvencia-illesztő módszerekkel, például a BLAST és a ClustalW segítségével) nem lehet megkeresni őket. Szintén a TFBS-ek rövidségének a következménye az, hogy in vivo a fehérje-DNS kapcsolatot a konkrét szekvencián túl erősen befolyásolja számos további hatás, például a DNS szerkezete.[2] A leírtak alapján a TFBS-ek azonosítását szekvencia alapú számítógépes kereséssel csak egyszerűbb DNS-sel rendelkező (például egysejtű) élőlények esetében lehet hatékonyan segíteni, ilyen esetekben is főként kísérletileg ismert eredményekből kiindulva és közeli rokon fajok összehasonlító elemzésével. A 2.3. ábra egy példát mutat négy élesztőgomba faj DNS-ének összehasonlításával.

[1] Az intron a génnek egy olyan szakasza, amelyik a gén mRNS-re történő átírása után az „alternative splicing” nevű mechanizmussal törlődik az mRNS-ből. Emiatt az intron „lefordítása” a fehérjében nem jelenik meg. A soksejtű eukarióta élőlények génjeinek jelentős része intron típusú szakasz. Érdekes, hogy az eukarióta gének esetén az intronokon belül is előfordulnak transzkripciós faktor kötőhelyek.

[2] Az élő sejtekben a kétszálú DNS nagy része „feltekert” állapotban található meg. Ez az állapot a hosszú DNS szál helyett egy kromatinként „becsomagolt” fehérje-DNS komplex-et jelent.



2.3. ábra. A *Saccharomyces cerevisiae* és három további élesztőgomba faj DNS-ének összehasonlítása szekvencia illesztéssel a GAL1 és GAL10 gének közötti tartományon. A csillaggal jelölt nukleotidok azonosak a négy genomban, a hosszabb azonos tartományokat vastag aláhúzás jelöli, és a bekeretezett szakaszok a jósolt transzkripciós faktor kötőhelyek. Az ábra átvétel a [133] publikációból.

Fehérje-DNS kapcsolódás szekvencia és mRNS koncentráció alapján

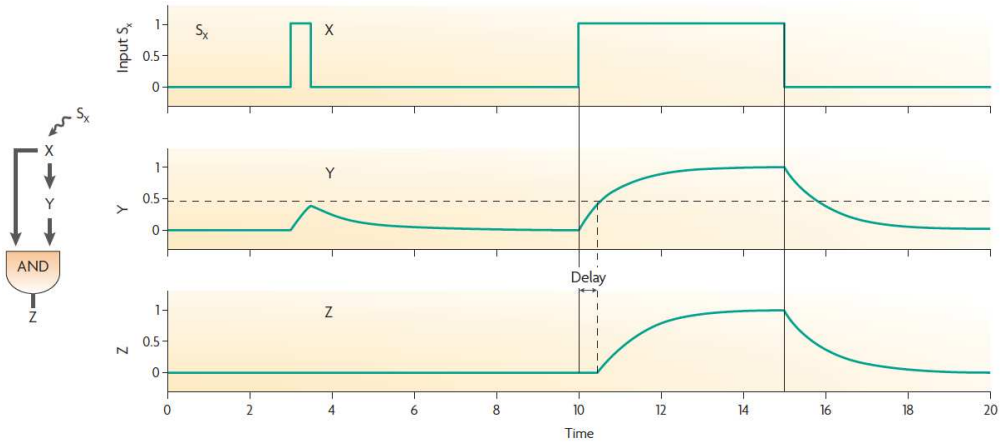
A transzkripció szabályozás közvetlen hatása a szabályozott génből „készülő” mRNS koncentrációjának változása. Ha nem csak a DNS szekvenciát, hanem a kísérletekben mért mRNS koncentrációkat is felhasználjuk, akkor a transzkripciós faktor fehérjék kötőhelyei (TFBS-ek) pontosabban jósolhatóak. Fizikusok dolgozták ki az egyik első olyan statisztikai módszert, amelyik gén expressziós szintek (a génekhez tartozó mRNS koncentrációk) változását összehasonlítja a gének előtt található DNS szakaszokban a vártnál gyakrabban előforduló rövid DNS szakaszokkal. A módszer leírásához egy bevezető megjegyzés, hogy egy gén expresszió változását a génhez tartozó mRNS két különböző esetben mért koncentrációjából számolt hányados logaritmusával szokták mérni. Ha egy 5-8bp hosszúságú DNS szakasz számos gén előtt (a gén „upstream” régiójában) a véletlenszerű esetben vártnál szignifikánsan gyakrabban fordul elő, és minden ilyen gyakori előfordulásához az adott gén magas expressziós

szintje (mRNS koncentrációja) tartozik, akkor valószínűsíthető, hogy ez a DNS szakasz egy aktiváló TF fehérje kötőhelye. Ezt az elvet követték Bussemaker és munkatársai 2001-es publikációjukban [134], amely napjainkra már klasszikusnak számít. A szerzők az élesztőgombában kerestek TFBS-eket minden génhez úgy, hogy a gyakori upstream szekvencia elemek előfordulási gyakorisága (mint vektor) és a gének expresszió változásai (szintén vektor) között egy lineáris transzformációt azonosítottak. Ez a lineáris transzformáció határozta meg, hogy – lineáris közelítésben – melyik upstream szekvencia elemek (TFBS-ek) milyen nagyságú és előjelű expresszió változást okoznak.

A TR hálózat motívumai (felülreprezentált részgráfjai)

A leírtak alapján a transzkripció szabályozási (TR) hálózat egy csúcspontja egyaránt jelent egy gént és az adott gén által kódolt fehérjét. A hálózat élei irányítottak, van előjelük és erősségük, és azt mutatják, hogy két gén között transzkripció szabályozási hatás lehetséges. Az angol nyelvű szakirodalomban a hálózat két gyakori neve TR network vagy Gene Regulatory Network (GRN). Más (nem biológiai) szabályozási hálózatokhoz hasonlóan a TR hálózatban is van szabályozási késleltetés és a hatások összegzése („igazságtáblázata") gyakran közelíthető egyszerű módon, például egy logikai AND vagy OR függvénnyel. Ezért a TR hálózatban is megtalálhatóak alapvető jelfeldolgozási funkciók. Ezek közül a legismertebb az aluláteresztő szűrő, amelyet a feed-forward loop (FFL) csoportba tartozó részgráfok közül néhány megvalósít. Az FFL részgráf három csúcsból (A, B, C) és három élből (A→B, B→C, A→C) áll. Ha mindhárom él a kezdő pontjára érkező jelet késleltetéssel és azonos előjellel továbbítja a végpontjára, és a C pontnál találkozó két él hatása között az összegzés a logikai AND függvény szerint történik, akkor ez az ABC részgráf egy 1-es típusú, AND kaput használó koherens FFL (cFFL-1-AND). Ez a részgráf kis időablakokban nézve integrálást végez, nagy időtartamokon aluláteresztő szűrőként használható (a lassan változó bemeneti jelek a kimenetén erősebben jelennek meg, mint a gyorsan változóak). A részgráfot és mindkét funkcióját a 2.4. ábra mutatja be.

Az FFL részgráfon kívül több más, kis számú csúcspontból álló részgráf képes speciális jelfeldolgozási feladatot ellátni. Például a negatív autoreguláció (NAR) esetében egy fehérje a saját mRNS-ének átírását gátolja. Egy TR hálózatban a kis részgráfok előfordulási száma megmutatja, hogy az általuk végzett jelfeldolgozási funkciók



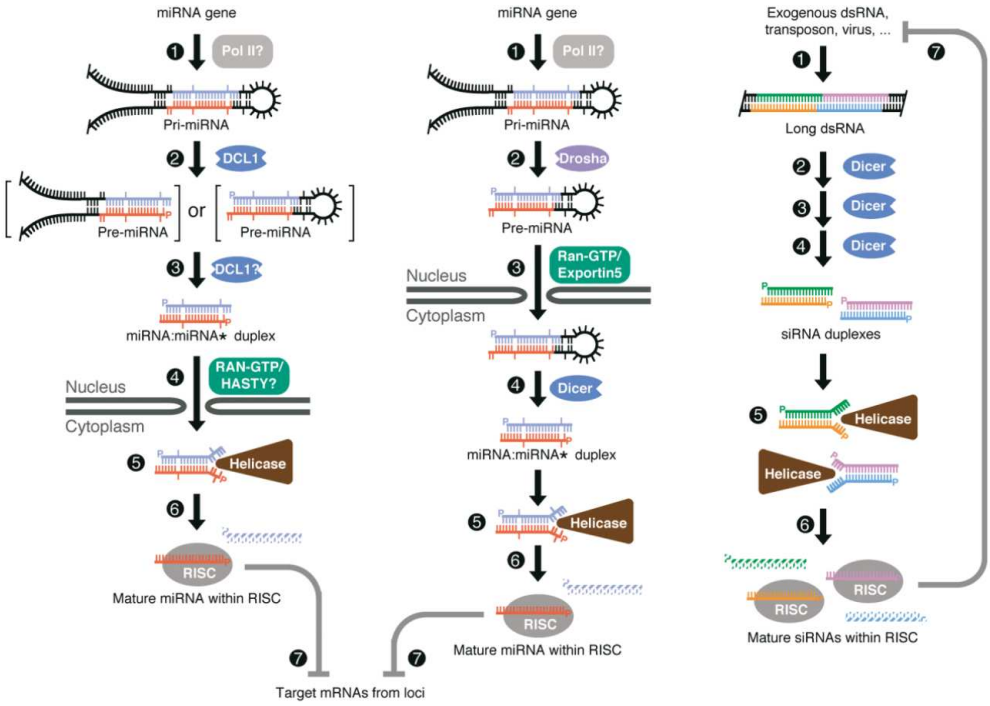
2.4. ábra. A dolgozat szövegében említett cFFL-1-AND (1-es típusú, AND kaput használó koherens FFL) transzkripció szabályozási részgráf. A részgráf késleltetésénél rövidebb S_x bemeneti (input) jel érkezése után a kimeneten – az aluláteresztő jelfeldolgozási funkciónak megfelelően – nincsen jel. Hosszabb bemeneti jel esetén lassú (a jelfeldolgozásban „integráló” típusúnak nevezett) válasz látható. Az ábra átvétel a [135] publikációból.

milyen jelentősek. Mivel mindegyik kis részgráf előfordulhat egy véletlenszerű hálózatban is, ezért a kis részgráfok előfordulási gyakoriságának (hányszor szerepel a TR hálózatban) a szignifikanciáját úgy lehet számszerűsíteni, ha ezt az előfordulási számot összehasonlítjuk a véletlenszerű hálózatokban mérhető előfordulási számmal. Az összehasonlításhoz használt véletlenszerű hálózat típust szokás „null modell”-nek vagy kontroll hálózatnak nevezni. Egy N csúcsból és E élből álló valódi (mért) hálózat esetén a kontroll hálózat általában egy irányított ER hálózat vagy az eredeti hálózatból élkeveréssel előállított hálózat. Az irányított ER hálózat N csúcs között véletlenszerű (korrelálatlan) helyeken tartalmaz E élt. Az élkeveréses kontroll előállítása iteratív, és egy lépésében egy meglévő [(A→B), (C→D)] élpárt kell kicserélni az [(A→D), (B→C)] élpárra. Az irányított ER gráf az eredeti hálózatból csupán a csúcsok és az élek számát tartja meg, míg az élkeverés megtartja minden egyes csúcs bemenő és kimenő fokszámát is. Egy valódi (kísérletekből származó) hálózat valamilyen mért mennyiségét (például az FFL részgráfok számát) úgy lehet összehasonlítani a kontroll hálózatokban mérhető azonos mennyiséggel, hogy a kontroll hálózatot elkészítjük például 100-szor, minden egyes esetben elvégezzük a mérést, és a kapott eredmény átlagát és szórását használjuk.

Egy kis részgráfot hálózati motívumnak nevez a szakirodalom [136], ha a mért hálózatban a kontroll modellhez képest jelentősen többször fordul elő, más szóval: felülreprezentált. Konkrét példaként egy TR hálózatban az FFL részgráfok számának (n_0) jelentősége a következő módon mérhető. Első lépésként készítsük el az eredeti hálózattal azonos számú csúcsot és élt tartalmazó irányított ER hálózatot N_{RND} alkalommal egymástól függetlenül, és minden esetben mérjük meg az FFL-ek számát. A kontroll hálózatokban kapott eredmények átlaga E és szórása σ. Ez alapján az adott típusú kontroll hálózathoz képest az eredeti hálózat FFL-jeinek számát a $Z = (n_0 - E)/\sigma$ mennyiség méri. Ha egy mért hálózatban a Z_{FFL} mennyiség nagy – a gyakorlatban a határ a $Z > 2$ szokott lenni, – akkor azt mondjuk, hogy ebben a hálózatban az FFL részgráf egy hálózati motívum. A használt véletlen gráfok száma általában $N_{RND} = 100$ vagy 1000. Egy gyakorlati szabály, hogy az N_{RND} minta szám akkor elegendően nagy, ha háromszor nagyobb N_{RND} használata esetén nem változik jelentősen a mért mennyiség σ szórása. További érdekesség, hogy egy kis részgráf lehet a kontroll esethez képest alulreprezentált is, tehát a korábban mért Z érték lehet negatív. Ilyenkor a $Z < -2$ esetben a részgráfot anti-motívumnak szokás nevezni.

Az irodalomban gyakran használt TR hálózatok

A [T4] publikációnkhoz a 2004-ben Harbison és kollégái által az élesztőgombában mért adatsort [137] használtuk. Harbison és munkatársai kísérleteket végeztek, majd azok eredményeit kiértékelték. Először három ChIP mérés sorozatot végeztek el azonos módon közeli rokon fajok DNS szekvenciáinak összehasonlításával. Ezután további irodalmi adatok alapján a DNS-en TF fehérjék által felismerhető kötőhelyeket állapítottak meg. A kapott eredményeket összesítve "TF fehérje $\rightarrow$ szabályozott gén" élekből álló irányított hálózat állítottak össze. Az összeállított irányított hálózat szabályozási lehetőségeket sorol fel, de egy-egy konkrét helyzetben (sejt állapotban) a felsorolt kölcsönhatásoknak csak egy része valósul meg. A Harbison *et. al.* publikációban [137] résztvevő Young csoport és Fraenkel csoport több tagja a 2004-es publikáció számítógépes kiértékelését továbbfejlesztette 2006-ban MacIsaac első szerzőségével [138]. Egy szintén 2004-es cikkben Luscombe, Babu és munkatársaik élesztőgomba DNS szekvencia adatai és mRNS koncentrációi alapján [139] többféle sejt állapotban megvalósuló transzkripció szabályozási részhálózatot összegeztek. Később Balaji, Babu és munkatársaik (köztük Luscombe) – szintén az élesztőgombára



2.5. ábra. A transzlációt gátló RNS-ek szabályozási útvonalainak összefoglalása 2004-ből. Az útvonalak számos közös elemének egyike, hogy a rövid szabályozó RNS szakasz (például miRNS vagy siRNS) egy RISC nevű fehérje komplex segítségével kapcsolódik a hírvivő RNS-hez (mRNS-hez). Ezzel gátolja a fehérje szintézist és gyorsíthatja az mRNS lebontását. Az ábra átvétel a [143] publikációból.

– irodalmi adatokat összegezve állítottak össze TR hálózatot [140]. A *C. elegans* fonálféreg transzkripciós faktorainak és transzkripciós szabályozásának általános (nem szövet- vagy állapotspecifikus) kutatásában az egyik kiemelkedő csoport M. Walhout csoportja. A Walhout csoport egyik jelentős eredményét [141] mutatja be a doktori dolgozatban a saját publikációk között megtalálható [T6] preview (kitekintő) cikk. A DNS átírás szabályozásának egy friss áttekintése megtalálható például a [142] publikációban.

A transzláció gátlása rövid nem kódoló RNS-ek segítségével

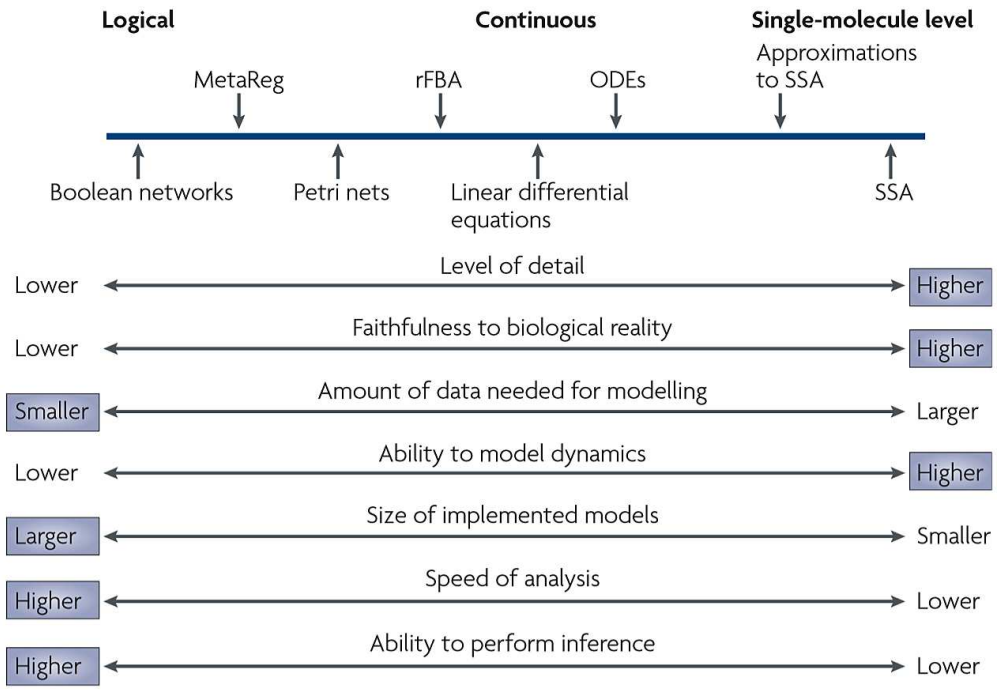
Az élő sejtek a hírvivő (messenger) RNS-ből aminosav láncot tudnak készíteni. Ennek a folyamatnak a neve transzláció (fordítás, mRNS-ről aminosav láncra). A fordítás kiinduló „ABC”-je az RNS-ben található négyféle nukleotid (A, U, C, G) és az eredmény „ABC”-je a 20-féle aminosav. A sikeres transzlációhoz szükséges, hogy a fordítást végző fehérje-RNS komplex (a riboszóma) le tudja olvasni az egyszálú mRNS

nukleotidjait. Ha ez a fordító „gépezet” nem fér hozzá a messenger RNS-hez, akkor a fordítás nem indul el. A rövid gátló RNS-ek (amelyeket összefoglaló névvel gyakran mikroRNS-nek neveznek) a DNS kettős spiráljánál megismert módon képesek templát képződéssel messenger RNS-ekhez (mRNS-ekhez) kapcsolódni. Így hosszú ideig gátolni tudják az mRNS és a transzlációs gépezet kapcsolatát, és ezáltal két fő módon képesek a transzlációt gátolni: (1) az egyszálú RNS a kétszálú DNS-nél jóval kevésbé stabil, gyorsabban elbomlik, és (2) a mikroRNS-hez kapcsolódó fehérje komplex képes gyorsítani a messenger RNS lebontását. Összefoglalva: a mikroRNS-ek (miRNS-ek) képesek egyes, messenger RNS-ek (mRNS-ek) esetén a transzlációt lassítani vagy leállítani. A részletek alapján szokás megkülönböztetni mikroRNS-eket, siRNS-eket és számos további nem kódoló, transzlációt gátoló RNS típust. A jelenség első megfigyelése a *C. elegans* nevű fonálféreg fajban történt [144] és egy áttekintése megtalálható a [143] összefoglaló cikkben. Ennek az összefoglaló cikknek egy ábráját mutatja a 2.5. ábra.

A transzkripció gátlása és a transzláció gátlása esetén a gátlást végző szabályozó molekulák (a transzkripciós faktor fehérjék és a rövid gátló RNS szakaszok) egyaránt a DNS-ben található szekvencia információ alapján készülnek. Viszont a két gátló molekula típus között jelentős különbség, hogy a TF fehérjék előállításának szabályozása lényegesen jobban ismert, mint a rövid gátló RNS-ek előállításának szabályozása. Emiatt a mikroRNS$\rightarrow$mRNS gátlást egy két szintes (bipartit) gráffal lehet modellezni, amelynek első (felső) szintjén a csúcspontok a rövid gátló RNS-ek, és az alsó szint csúcspontjai azok a fehérjék, amelyeknek a készítését a rövid RNS szakaszok gátolják. A TF hálózathoz hasonlóan itt is egy rövid RNS-től olyan erősségű (negatív előjelű) él fut egy fehérjéhez, amilyen erősen az adott RNS annak a fehérjének a transzlációját gátolja.

A molekuláris biológiai szabályozás hálózatos dinamikai modelljei

Az élő sejtekben a transzkripció és transzláció szabályozás egyik legfőbb célja a sejtben adott pillanatban megtalálható fehérjék állapotának és mennyiségének szabályozása. Bár ez a két szabályozó rendszer erős kölcsönhatásban fejlődik [145], jellemző időskálájuk mégis eltérő, ezért a dolgozat külön alfejezeteiben szerepelnek. A molekuláris biológiai szabályozás modellezésére sokféle más hálózat típus és további



2.6. ábra. A molekuláris biológiai szabályozás hálózatos modelljeinek lineáris felsorolása és összehasonlítása. A gyakran használt modell típusok bemutatása szerepel a dolgozat szövegében. Az ábra átvétel a [146] publikációból.

matematikai eszközök is használhatóak, ezeket foglalja össze a [146] publikáció és a 2.6. ábra.

A molekuláris biológiai hálózatokban megfigyelhető dinamika modellezésére az egyik legegyszerűbb matematikai eszköz a Bool típusú hálózat modell, amelyben minden csúcspont egy fehérjét jelöl. Ebben a modellben egy csúcspont állapota 1, ha az adott fehérje jelen van a sejtben és aktív, illetve 0, ha az adott fehérje nincsen jelen a sejtben vagy inaktív. A csúcspontok állapotának frissítése időben diszkrét és párhuzamos, és minden csúcs állapotát az őt befolyásoló csúcspontoknak az előző időpillanatban érvényes állapota határozza meg. Bool-modelleket eredményesen alkalmaztak – többek között – annak modellezésére, hogy egy növényi sejt illetve egy élesztőgomba sejt hogyan éri el egy stacionárius (stabil) állapotát (ld. például [147, 148]).

A Bool-modelleknél valamivel részletesebb leírást tesz lehetővé a fluxus egyensúly elemzés (Flux Balance Analysis, FBA). Ezt a – korábban más területeken már ismert – matematikai módszert biológusok először olyan egysejtűek növekedésének modellezésére alkalmazták, amelyekben ismertek az anyagcsere reakciók (döntő részének)

résztvevői és sztöchiometriai számai, valamint megbecsülhetőek a reakciók minimális és maximális lehetséges fluxusai. A sztöchiometriai számokat egy alulhatározott lineáris algebrai egyenletrendszerben lehet összesíteni. Ehhez az egyenletrendszerhez tegyük hozzá egy további, fiktív egyenletként a biomassza mennyiségének változását mutató egyenletet. A biomassza mennyiségének változását mutató fiktív egyenlethez szükséges sztöchiometriai számokat (egységnyi biomassza keletkezéséhez melyik biológiai molekulából mennyire van szükség) az élő sejteket alkotó molekulák megfigyelt koncentráció arányai alapján rögzítsük. Az így kapott FBA modellekkel – kísérletileg igazolhatóan kis hibával – meghatározható például az, hogy melyik tápanyagok koncentrációjának módosításával hogyan változtatható egysejtű modell szervezetek populációinak növekedési rátája [149].

A Bool-modellekhez és a fluxus egyensúly elemzéshez (FBA) hasonlóan szintén molekula típusok dinamikáját modellezik a differenciálegyenlet rendszerek, amelyek egyszerűbb esetben lineáris kapcsolatokat feltételeznek. A korábbi modell típusoktól eltérően ebben az esetben már gyakori a molekula állapotok (például aktiváltság) megkülönböztetése. A jelenleg széles körben használt molekuláris biológiai modellek közül a legrészletesebbek a sztochasztikus szimulációs algoritmusok (SSA). Ez a módszer család a számolás során minden betöltött molekula állapotot egyenként követ, a betöltetlen állapotokat nem tárolja, és a master egyenleteket számolja numerikusan. A sztochasztikus szimulációs módszerekhez sok paraméter ismerete szükséges, de kis koncentrációk esetén a többi módszernél jóval pontosabb leírást tudnak adni. A 2.6. ábrán követhető, hogy a Bool hálózattól az SSA modellek felé haladva növekszik a modellek összetettsége és paramétereinek száma, és a szükséges számítási idő. Ezzel párhuzamosan az eredmények egyre jobban alkalmazhatóak egyedi rendszerekre.

A dolgozatban bemutatott saját tézispontok dinamikai modellezést nem tartalmaznak. A következő alfejezetek a transzkripció és a transzláció szabályozását is statikus hálózatokon keresztül vizsgálják. A dolgozat mindkét szabályozási típust statikus irányított hálózattal modellezi, amelyben az éleknek súlya (erőssége) és előjele van. Ezek a hálózatok a dinamikai jelenségek során a sejt rendelkezésére álló kölcsönhatásokat jelentik, amelyek közül egy konkrét sejt állapot vagy külső körülmény esetén nem mind aktív.

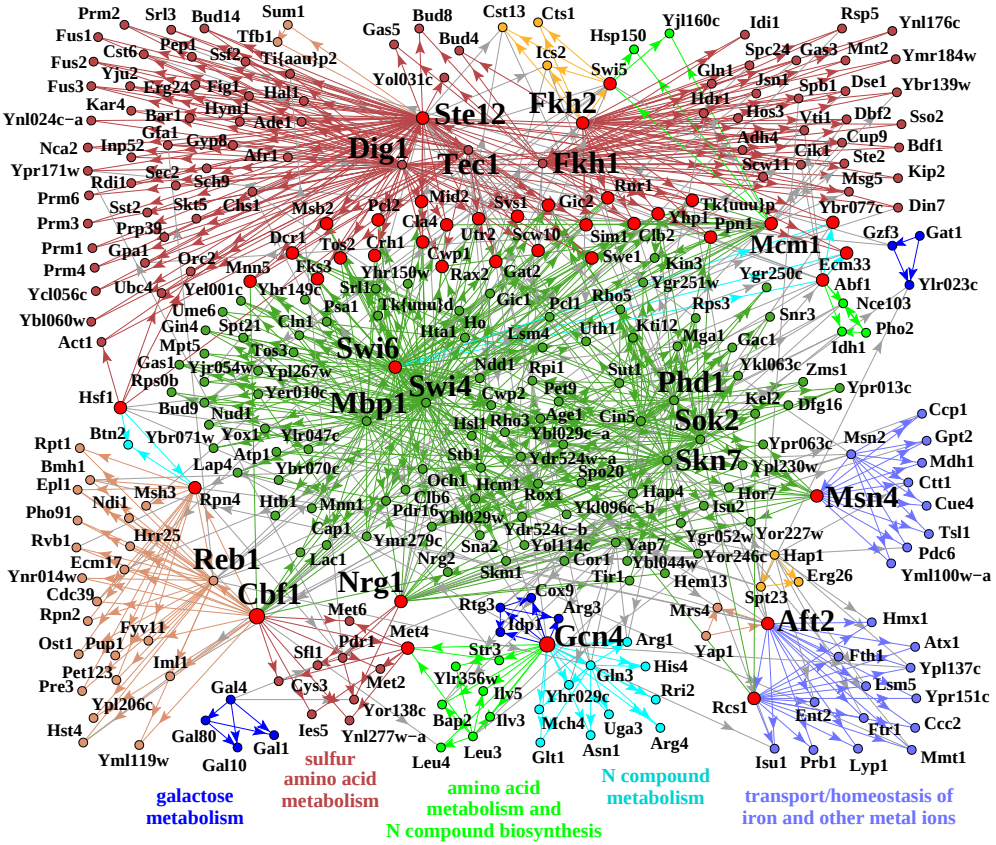
Eredmények

2.1. Irányított hálózati modulok az élesztőgomba transzkripció szabályozási hálózatában [T4, T5, T6]

Ha egy hálózat éleinek nincsen irányítása, akkor a hálózat irányítatlan. Ha az éleknek van irányítása, akkor a hálózat irányított. A transzkripció szabályozási hálózata és a transzláció szabályozási hálózata egyaránt irányított hálózat. Ezzel szemben a dolgozat előző fejezetében használt klikk perkolációs módszer (CPM) irányítatlan hálózatokat elemez. A [T4] publikációnkban az irányított hálózatok – többek között a transzkripció szabályozási hálózat – moduljainak azonosításához bevezettük az irányított k-klikkek perkolációján alapuló CMPd módszert. A CPMd módszer nevében szereplő „d” az irányított (directed) szó rövidítése.

Egy irányítatlan hálózatban a k-klikk egy k csúcspontból álló teljes részgráf. Az irányítatlan esethez képest a [T4] publikációnkban az irányított hálózatokban mérhető irányított k-klikket úgy kívántuk definiálni, hogy a k-klikk tetszőleges két pontja között legyen egyértelmű irány (sorrend). Ez az irány az információ- vagy anyag áramlás irányának is tekinthető, ami a molekuláris biológián kívül több más területen is központi fogalom. További megjegyzés, hogy egy irányított hálózatban két csúcs között kettő kapcsolat is lehet („oda” és „vissza”). Ezek alapján a [T4] cikkben az irányított k-klikk definíciónk a következő volt. Egy irányított hálózatban k darab csúcs irányított k-klikket alkot, ha a k csúcs között egyértelműen felállítható egy sorrend $(1, 2, \ldots, k)$ úgy, hogy tetszőleges két csúcs között van legalább egy olyan él, amelyik a magasabb sorszámú csúcstól az alacsonyabb sorszámú csúcsra mutat. Az így definiált irányított k-klikkek segítségével definiáltuk az irányított k-klikk perkolációs hálózati modulokat a CPMd módszerben. A irányítás nélküli módszerhez hasonlóan az irányított módszer esetén is az volt a k élsűrűség paraméter kiválasztásának a feltétele, hogy a két legnagyobb modul méretének aránya a 2-höz legközelebb legyen. A korábbiakhoz hasonlóan szintén azzal az érveléssel jelöltük ki ezt a 2-es arányt, hogy a modulok méret eloszlása egy hatványfüggvényhez legjobban hasonlítson. A kapott hálózati modulok – az irányítatlan hálózatokban definiált klikk perkolációs modulokhoz hasonlóan – a hálózat egybefüggően sűrű tartományai és egymással átfedhetnek.

Az irányított klikk perkolációs módszerrel meghatározható modulokra a 2.7. ábra mutat példát. Ezen az ábrán a legnagyobb számú kimenő éllel (ki-fokszámmal) rendelkező csúcsok közül sok a modulok átfedésében található, ami biológiailag azt jelenti,

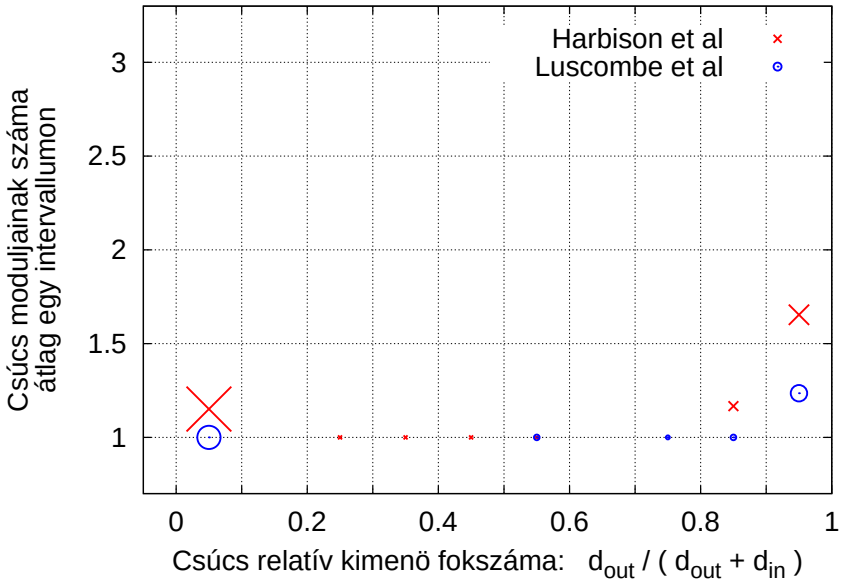


2.7. ábra. Az *S. cerevisiae* (élesztőgomba) transzkripció szabályozási kölcsönhatásainak hálózatában az irányított klikk perkolációs módszerrel (CPMd) $k = 3$ klikk méret paraméterrel meghatározott irányított hálózati modulok a [T4] publikációnkból. A hálózatban minden csúcs egy gént jelöl (és egyben a hozzá tartozó fehérjét), és egy irányított él azt mutatja, hogy az él kezdőpontján található fehérje szabályozza az él végpontján található gén átírását. Azonos modulhoz tartozó csúcspontok azonos – pirostól eltérő – színűek, kivételt jelentenek ez alól az egynél több modulhoz tartozó csúcsok és élek, amelyeknek a színe piros. Az egynél több modulhoz tartozó csúcsok gyakran nagy ki-fokszámúak: sok él vezet ki belőlük, tehát sok gént szabályoznak. Az ábra alján névvel jelölt csoportok szignifikáns funkcióit a GO TermFinder [124] programmal határoztam meg a Gene Ontology adatbázis fehérje funkció listáit felhasználva [74]. Az irányított modulokat azonosító CPMd algoritmus számítógépes programját Palla Gergely írta. Az adatok gyűjtését, feldolgozását és az eredmény ábrázolását én végeztem.

hogy az élesztőgombában ez a kis számú transzkripciós faktor fehérje mind képes irányítani egymástól elkülönülő feladatokat. Az irányított modulok részletesebb vizsgálata érdekében a [T4] cikkünkben az irányított modulok csúcspontjainak (fehérjéknek) két tulajdonságát használtuk fel: a relatív kimenő fokszámot és a csúcsot tartalmazó modulok számát. Ha egy irányított hálózatban egy csúcspontba beérkező és onnan ki-

induló irányított élek száma d_{in} és d_{out}, akkor a csúcspont relatív kimenő fokszáma $d_{out}/(d_{in} + d_{out})$. Az eredményeket mutató 2.8. ábrát a [T4] cikkünk alapján a dolgozathoz készítettem. Az ábrán látható eredmények alapján a modulokban található fehérjék legtöbbje vagy (i) egyetlen modulban van benne és szabályozott fehérje (más fehérjét nem szabályoz), vagy (ii) olyan szabályozó fehérje, ami nem vagy csak alig kap transzkripciós szabályozást és 1, 2, esetleg több modulban vesz részt. Ez a két csoport biológiailag valószínűleg a „fő irányító" TR fehérjéket (csúcspontokat) és az „alközpont" TR fehérjéket jelenti. Érdekes, hogy a modulok fehérjéinek két fő csoportba sorolása a két, különböző projektben készült élesztőgomba TR hálózat [137, 139] esetén nagyon hasonló. A 2.8. ábra kapcsán fontos megjegyezni, hogy mindkét TR hálózatban az irányított k-klikk perkolációs módszerrel meghatározott modulok a hálózat csúcsainak csupán egy kis részét tartalmazzák. Természetesen az élesztőgombán túl több más élőlény esetén is elérhető mérési és számítógépes módszerekkel összeállított transzkripciós faktor fehérje lista (például a DBD adatbázisban [132]) vagy egy-egy speciális sejt állapotra meghatározott transzkripció szabályozási kölcsönhatás lista. De egy élőlény teljes transzkripciós kölcsönhatás listája – az összes lehetséges TR kölcsönhatás térképe – ismereteim szerint eddig csak élesztőgomba esetén készült el jó minőségben.

A 2.7. ábrán és a 2.8. ábrán bemutatott elemzések a TR hálózatok szerkezetét (topológiáját) az irányított klikk perkolációs módszer segítségével vizsgálják. Sok más módszer is lehetséges, például a [T5] publikációnk a hálózati motívumok (kis irányított részgráfok) és nagyobb hálózati csoportok (origons, organizers) alapján rendezi modulokba a TR hálózat csúcsait, és elemzi a kapott csoportokat biológiai funkciók és mRNS expresszió szerint. Az irodalomban a transzkripciós hálózatok moduljainak azonosítására napjainkban használt módszerek már részletesebbek az általunk használt módszereknél. Nagy részük a TR hálózat szerkezetén túl az mRNS szinteket (gén expressziót) is felhasználja. Például Segal és társszerzőinek módszere [150] kétféle bemenő (input) adatot használ. A módszer egyik bemenő adata egy mRNS expressziós mátrix, amelyben az i. sor j. eleme (egy szám) azt mondja meg, hogy a j. expressziós kísérletben vizsgált i. génből készült mRNS milyen koncentrációban van jelen. Ez a koncentráció – a transzláció utáni módosításoktól eltekintve – jelzi, hogy milyen aktív az adott génből készülő fehérje. A módszer másik bemenő adatsora egy fehérje lista: ismert és feltételezett transzkripciós faktorok, és további szabályozó fehérjék.



2.8. ábra. Az élesztőgomba (*S. cerevisiae*) transzkripció szabályozási hálózatában az irányított klikk perkolációs módszerrel (CPMd) azonosított modulok csúcspontjainak (fehérjéknek) a csoportosítása modul szám és relatív kimenő fokszám szerint. A kék körökkel jelölt eredményekhez felhasznált adatok forrása a Harbison *et. al.* publikáció [137]. A piros $\times$ jelek a Luscombe *et. al.* publikáció [139] adatai alapján kapott eredményeket mutatják. Minden mérési pont azoknak a fehérjéknek (TR hálózati csúcspontoknak) az átlagos modul számát mutatja, amelyek a relatív kimenő fokszám adott ($1/10$ szélességű) intervallumában találhatóak. Az ábrán az $\times$ és kör jelek lineáris méretének négyzete (a jelek „területe") arányos az adott relatív kimenő fokszám intervallumba eső fehérjék számával. Az ábra alapján a TR hálózat irányított moduljaiban a csúcspontok két fő csoportba sorolhatóak. Az első csoport tagjai döntően bemenő élekkel rendelkeznek és 1 modulban vannak benne, míg a második csoport tagjai főként kimenő élekkel rendelkeznek és 1 vagy több modulban vannak benne. Az ábrát a doktori dolgozathoz készítettem.

Az adatok alapján Segal és társszerzői meghatároztak gén (fehérje) csoportokat úgy, hogy minden egyes csoporthoz kiválasztották (i) a csoportot szabályozó fehérjéket, (ii) a logikai függvényt, amely alapján a szabályozó fehérjék hatása összegződik és (iii) a kísérleti körülmények listáját, amelyekben az azonosított logikai függvényekkel szerinti szabályozás történik. A módszer előnye, hogy nemcsak azt sorolja fel, milyen szabályozás lehetséges, hanem megmutatja azt is, hogy milyen körülmények között aktívak az egyes szabályozási funkciók. A módszer a szabályozást végző konkrét hálózatot nem tár fel, hanem csak a szabályozási kapcsolatok eredményét írja le az adatokra történő illesztéssel.

Szintén kétféle adatot használ fel Bar-Joseph és társszerzőinek [151] módszere, amelyet a szerzők „GRAM”-nak (Genetic Regulatory Modules) neveztek el. Bar-Joseph és munkatársai szerint a fehérjék közötti szabályozási kapcsolatok megbízható azonosításához nem elegendő a gének mRNS expressziós szintjeinek vizsgálata, amelyen Segal és kollégáinak elemzése alapszik. Ezért a GRAM módszer az mRNS expressziós szinteken túl DNS-fehérje kötési adatokat is felhasznál. Utóbbiak nem tartalmazzák a szabályozási kölcsönhatás előjelét – azaz hogy serkentő vagy gátló a kölcsönhatás, – ezért a GRAM módszer egy hatékony kereső algoritmussal megkeresi a legvalószínűbb szabályozási előjeleket. A hálózatos elemzések szempontjából a GRAM módszer előnye, hogy konkrét kölcsönhatásokat (hálózati éleket) azonosít. A GRAM módszerhez szükséges DNS-fehérje kötési adat, amely a teljes DNS-re megfelelő részletességgel (ismereteim szerint) jelenleg csak az *S. cerevisiae* élesztőgombában áll rendelkezésre. Az irányított CFinder-hez hasonlóan Segal és Bar-Joseph módszere [150, 151] is megengedi a modulok átfedését. Az irányított CFinder-től eltérően mindkét módszer az mRNS expressziós adatok elemzésére készült, nem közvetlen transzkripció szabályozási kölcsönhatás adatok vizsgálatára.

2.2. Emberi mikroRNS-ek szerepeinek összehasonlítása a transzláció csendesítési hálózat moduljai alapján [T7]

A transzláció csendesítési hálózat a mikroRNS→mRNS kölcsönhatásokat súlyozott irányított kapcsolatokkal írja le. Ebben az irányított hálózatban minden él azt mutatja, hogy az él kezdőpontján található mikroRNS az élen található erősséggel gátolja az él végpontján található mRNS transzlációját, és az élekre írt szabályozási erősségek egymáshoz képest értelmezhetőek. A transzláció csendesítésnek ez a hálózata egy bipartit gráf. Ez a szabályozási mód lényegesen eltér a transzkripció szabályozási (TR) hálózat szerkezetétől, amelyben biológiailag azonos típusú elemek (transzkripciós faktor fehérjék) szabályozzák egymás koncentrációját akár 6-7 szintet is tartalmazó „láncban”.

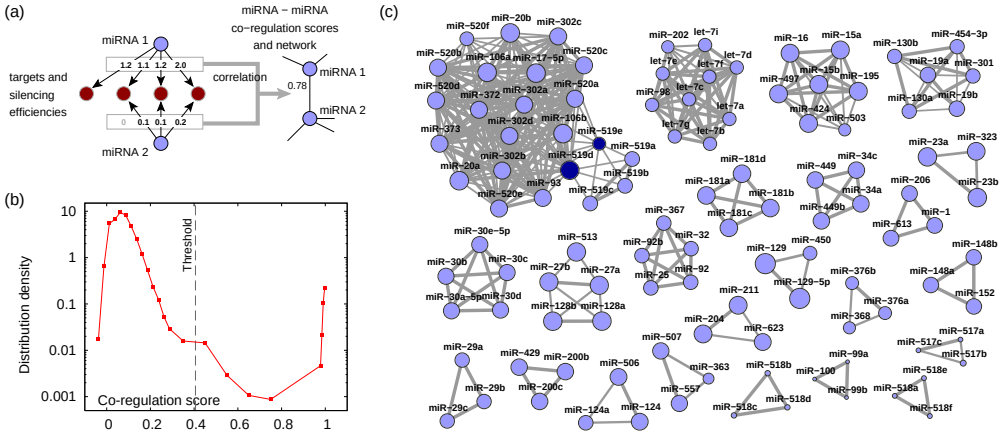
A mikroRNS-ek hatása abban is eltér a transzkripciós faktor fehérjék hatásától, hogy a mikroRNS-ek a fehérje koncentrációkat és a végső fehérje aktivitásokat a TF fehérjéknél jóval kisebb mértékben képesek befolyásolni. További eltérés a TF hálózattól, hogy a mikroRNS→mRNS kölcsönhatások közül jóval kevesebbnek ismert kísérleti igazolása. A [T7] publikációnk időpontjában a TarBase és a miRBase adatbázis [152] a *C. elegans* nevű fonálféreg fajban 33 miRNS és 54 mRNS között 61 kísérletileg

Database	Number of miRNAs	Number of target genes	Number of interactions	Experimentally verified interactions according to TarBase
Experimental				
TarBase	33	54	61	
Computationally predicted				
miRBase	711	22 474	584 403	16 (0.003%)
PicTar	171	6 885	54 947	31 (0.06%)
PITA	470	8 720	152 040	23 (0.02%)
TargetScan	455	15 878	955 644	44 (0.004%)

2.1. táblázat. A *C. elegans* fonálféregben az elméleti (számítógépes szekvencia-elemzéses) módszerek lényegesen több lehetséges mikroRNS-mRNS kölcsönhatást jeleznek, mint amennyire kísérleti igazolás létezik. Ennek legfőbb oka valószínűleg az, hogy a mikroRNS→mRNS kölcsönhatások (a TR kölcsönhatásokhoz képest) gyengék és a kísérletekben egyenként nehezen kimutathatóak. További érdekesség, hogy az elméleti módszerek által jósolt kölcsönhatások száma között is jelentős különbség van. Az ábra átvétel a [T7] publikációnkból, ennek megfelelően 2008-as adatokat tartalmaz. A táblázatot Orosz Katalin és Boross Gábor segítségével készítettem, az adatok gyűjtését és feldolgozását hárman végeztük.

igazolt kölcsönhatást sorolt fel. Viszont a számítógépes módszerekkel (szekvencia-elemzéssel) jósolt kölcsönhatások száma nagyságrendileg 50 ezer és 1 millió között volt (ld. 2.1. táblázat). Mindezek alapján a transzkripciós faktor fehérjékhez képest a mikroRNS-eket szokás finomhangoló irányítóknak ("micromanager"-nek) nevezni. A következők bekezdésekben bemutatom, hogy eredményeink alapján ezek a „finomhangoló” szabályozók a működés során számszerűen mekkora jelentőségűek.

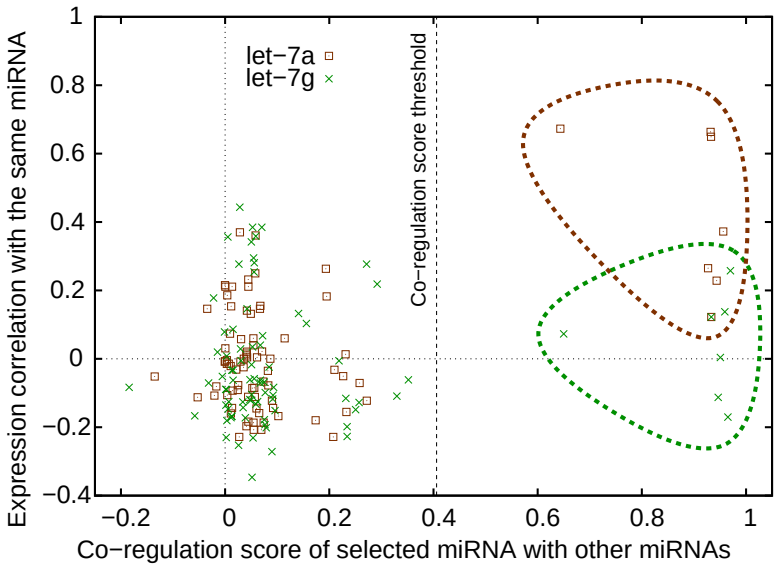
Első lépésként el kell dönteni, hogy a rendelkezésre álló adatsorokat hogyan fésüljük össze egyetlen közös adatsorba. A PPI (vagy PPA) modulok kapcsán a doktori dolgozat korábbi részeiben már szerepelt, hogy ha többféle típusú és minőségű adatsor adatsor áll rendelkezésre, akkor azok összesítésének egy gyakori módja a következő két lépéses módszer. Először nagy pontosságú (pl. kísérleti) „true positive” eredmények segítségével meghatározunk egy minőség mérőszámot minden egyes adatsorhoz külön, majd ezekkel a minőség számokkal – mint súly faktorokkal – összesítjük (összefésüljük) az adatsorokat. Ez a módszer egyszerű, de az RNS csendesítés esetén kevés



2.9. ábra. (a) A mikroRNS-ek szabályozási feladatainak hasonlóságát bemutató mikroRNS-mikroRNS hálózat előállítása. Két mikroRNS között a súly a két mikroRNS szabályozási kölcsönhatásai által definiált vektorok (Pearson) korrelációja. Az így kapott hálózat a súlyozott és irányított mikroRNS → messenger RNS bipartit gráfot vetíti a miRNS gráf komponensre. (b) A kapott mikroRNS-mikroRNS kapcsolati erősségek eloszlása határozottan bimodális, tehát két mikroRNS szabályozási feladatai vagy nagyon hasonlóak, vagy jól megkülönböztethetőek. (c) A (b) részábrán látható bimodális eloszlás következménye, hogy a mikroRNS-ek szabályozási feladataik szerint jól elkülönülő csoportokba rendeződnek. Az ábra átvétel a [T7] publikációnkból, Boross Gábor és én készítettük.

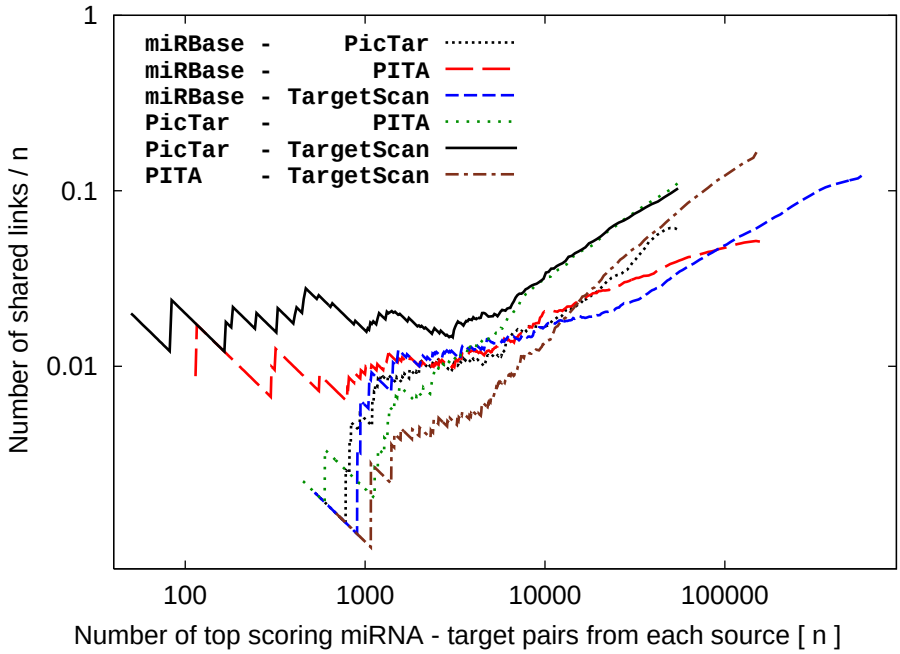
a kísérletileg igazolt kölcsönhatás, ami alapján az egyes adatsorok minőségét becsülni lehetne. Megoldásként – két, a témában megjelent elemzést [153, 154] követve – kiválasztottunk az adatbázisok közül kettőt [T7]: a TargetScan adatbázis [143] mikroRNS-mRNS kölcsönhatásait használtuk, és ezt kiegészítük kontrollként a PicTar adatbázis [155] kölcsönhatásaival.

Mivel nagyságrendileg százezer körüli kölcsönhatásról van szó, ezért érdemes a kölcsönhatásokat csoportokba összevonni. A csoportosításnak egy biológiailag is jól értelmezhető módja a mikroRNS→mRNS bipartit gráf levetítése a mikroRNS-ek partíciójára. Biológiailag ez azt jelenti, hogy a teljes szabályozási hálózat helyett csak azt vizsgáljuk, hogy két tetszőleges mikroRNS csendesítési hatása (az összes mRNS-re gyakorolt hatása) egymáshoz mennyire hasonló. A mikroRNS-ek szabályozási hatásainak hasonlóságát mutató hálózat pontos definícióját a 2.9a ábra mutatja. Az ábra alapján követhető, hogy az együtt szabályozási hálózat („co-silencing network”) egy irányítatlan hálózat. Ebben az irányítatlan hálózatban a CFinder segítségével csoportokat kerestünk. Annak ellenére, hogy a módszer megenged a modulok között átfedéseket, igen kevés átfedést találtunk ([T7]), tehát a mikroRNS-ek az általuk végzett szabályozás szerint jól szétváló modulokba csoportosulnak (ld. 2.9b és c ábra).



2.10. ábra. Két kiválasztott mikroRNS (let-7a és let-7g) expressziós szintjeinek hasonlósága a velük hasonlóan szabályozó mikroRNS-ek expressziós szintjeihez. Az ábrán minden egyes barna négyzet egy mikroRNS-t mutat a let-7a mikroRNS szempontjából és minden egyes zöld kereszt egy mikroRNS-t mutat a let-7g mikroRNS szempontjából. Egy zöld kereszt vízszintes koordinátája azt mutatja, hogy az adott mikroRNS és a let-7a szabályozási kölcsönhatásait két külön listában felsorolva mekkora a két lista (Pearson) korrelációjának abszolút értéke. Ugyanennek a zöld keresztnek a függőleges koordinátája azt mutatja, hogy egy mikroRNS és a let-7a mikroRNS expressziójának mekkora a korrelációja (szintén abszolút értékben értendő). A „co-regulation score threshold"-tól (együtt szabályozási hasonlóságra vonatkozó alsó küszöb értéktől) jobbra lévő együtt szabályozási hasonlóságok azonos modulon belül lévő két mikroRNS-hez tartoznak, tehát a let-7a és a let-7g mikroRNS-ekkel azonos modulban lévő mikroRNS-ekhez tartoznak. Ez a modul a 2.9c ábra felső sorában balról a második. Figyeljük meg, hogy a küszöbértéktől (az ábra közepén lévő függőleges szagatott vonaltól) jobbra található zöld keresztek átlagos függőleges koordinátája közel nulla, viszont a küszöbértéktől jobbra található barna négyzetek átlagosan jóval nulla felett vannak. A két csoportot külön kiemeli két szaggatott vonal (barna és zöld színnel). Ez azt jelenti, hogy a let-7g a hozzá nagyon hasonló módon szabályozó (azonos modulban lévő) mikroRNS-ekkel nagyon hasonló módon van szabályozva, míg mindez a let-7a-ra nem érvényes. Összefoglalva: A let-7a jóval fontosabb („more essential"), mint a let-7g. A mikroRNS expressziós adatok forrása a [156] publikáció. Az ábra a *C. elegans* fonálféregre vonatkozik. Az ábra átvétel a [T7] publikációnkból, Orosz Katalin és én készítettük.

Felmerül a kérdés, hogy ha ilyen határozottan szétválnak a modulok, akkor miért van szükség majdnem mindegyik modulban nagy számú mikroRNS-re? Másképpen megfogalmazva: miért van szüksége a szervezetnek több, egymáshoz igen hasonló szabályozási feladatot végző mikroRNS-re? Erre a kérdésre egy lehetséges válasz az, hogy az egyes szabályozási feladatok elvégzéséhez a mikroRNS-ekből eltérő mennyi-



2.11. ábra. Bármelyik két forrásból (adatbázisból) választjuk ki az 1000 (számítógépes szekvencia-elemzési módszerrel megállapított) legerősebb mikroRNS→mRNS szabályozási kölcsönhatást, a két kiválasztott kölcsönhatás lista relatív átfedése legfeljebb 2-3% körüli, jellemzően 1%. Az adatok 2008-as állapotot mutatnak a *C. elegans* (fonálféreg) élőlényre a következő adatbázisokból: miRBase [152], PicTar [155], PITA [157], TargetScan [143]. Az ábra átvétel a [T7] publikációnkból. Orosz Katalin, Boross Gábor és én gyűjtöttük az adatokat és végeztük a szükséges mRNS név átalakításokat. Az eredményeket én ábrázoltam.

ségre (dózisra) van szükség. Egy másik lehetséges válasz, hogy ha két mikroRNS azonos szabályozást végez, attól még önmaguk állhatnak eltérő szabályozás alatt. Az utóbbi lehetőség szerint az azonos (vagy erősen hasonló) szabályozási feladatot végző mikroRNS-ek között előfordulhatnak olyanok, amelyek a többitől igen eltérő módon szabályozhatóak, például eltérő külső körülmények esetén állnak rendelkezésre.

Ennek a jelenségnek a mérésére bevezettük a „mikroRNS-ek fontossága” [T7] („essentiality of microRNAs”) fogalmat. Egy mikroRNS fontossága egyenesen arányos azzal, hogy a vele azonos szabályozási feladatot végző más mikroRNS-ektől mennyire eltérő szabályozás alatt van. A mikroRNS-ek szabályozása becsülhető abból, hogy milyen koncentrációban vannak jelen (ezt mikroRNS expressziós szintnek is szokás nevezni). A definíciónk felhasználja a korábban meghatározott mikroRNS modulokat és egy konkrét példája látható a 2.10. ábrán.

Az „essentiality of a microRNA” fogalom pontos definíciója a következő. Egy kiválasztott mikroRNS-re számítsuk ki a mikroRNS expresszió korreláció abszolút értékét az összes többi mikroRNS-sel, majd vegyük ezeknek a korrelációknak az átlagát: A_{all}. A kiválasztott mikroRNS és egy tetszőleges másik mikroRNS átlagosan ennyire hasonló szabályozás alatt áll. Ezután vegyük csupán azoknak a korrelációknak az átlagát, amelyek a kiválasztott mikroRNS-sel azonos modulban lévő többi mikroRNS-hez tartoznak: A_{top}. A kiválasztott mikroRNS és a vele azonos modulban egy másik átlagosan ennyire hasonló szabályozás alatt áll. Az A_{all} és A_{top} mennyiségekkel kifejezve a mikroRNS fontosságát következő módon definiáltuk: $E = (1 + A_{all})/(1 + A_{top})$. Ez az érték akkor magas, ha a mikroRNS expressziója erősen eltér a saját moduljában lévő többi mikroRNS expressziójától. A definiált „essentiality of microRNA” (mikroRNS fontossága) mennyiség átfogó kísérleti ellenőrzéséről nem tudok. Azonban – egy nem várt módon – mégis volt a [T7] publikációnknak egy jelentős gyakorlati haszna. A cikkben elemeztük, hogy a különböző számítógépes szekvencia-elemzési módszerekkel felsorolt mikroRNS→mRNS szabályozási kölcsönhatások az egyes adatbázisokban mennyire egyezőek. A 2.11. ábrán látható eredmények alapján az él erősségek (tehát a kölcsönhatások erősség szerinti sorrendjének) egyezése alacsony, és erre az eredményre érkezett később független hivatkozás.

Az eredmények összefoglalása

A fejezetben ismertetett eredmények közül a társszerzőségemmel készült eredmények a következők:

- Irányított hálózatokban definiáltuk az irányított k-klikket, majd az irányított k-klikkekből képezhető k-klikk perkolációs klaszterek segítségével az irányított k-klikk perkolációs hálózati modulkereső algoritmust (CPMd).

- Az irányított TR (transzkripció szabályozási) hálózati modulokban a csúcspontok relatív kimenő fokszáma és modul száma alapján a csúcspontok két nagy csoportját azonosítottam. A relatív kimenő fokszám mennyiség használatát társszerzők javasolták. Ez a két csoport biológiailag valószínűleg a „fő” irányító csúcspontokat (fehérjéket) és az „alközpontokat” jelenti.

- Megmértük a korábban (számítógépes szekvencia-elemzéssel) előrejelzett mikroRNS→mRNS kölcsönhatás listák hasonlóságát. Bármely két adatbázis szerinti 1000 legerősebb kölcsönhatás relatív átfedése legfeljebb néhány százaléknyi.

- Mivel az egyes mikroRNS-mRNS kölcsönhatások száma nagy és önmagában egy-egy kölcsönhatás erőssége kevéssé megbízható, ezért a kölcsönhatások nagyobb csoportokba történő összesítése informatívabb lehet, mint az egyenként történő elemzésük.

- Definiáltuk a mikroRNS-ek hálózatát az általuk végzett szabályozási feladatok hasonlósága alapján és a CFinder modulkereső programmal meghatároztuk a mikroRNS-ek moduljait.

- Miután a mikroRNS-ek moduljait összehasonlítottuk a mikroRNS-ek expressziójával, definiáltuk a modulokban lévő mikroRNS-ek fontosságát („essentiality of a microRNA”). Ha egy mikroRNS (vagy általánosabban: szabályozó egység) feladata több más mikroRNS feladatához hasonló, de ezt a feladatot speciális körülmények között végzi, akkor a definíciónk alapján ez a mikroRNS fontosabb („essential”) a többinél. Biológiailag: az eltávolítása nagyobb eséllyel okoz mérhető fenotípus változást.

Saját hozzájárulásom az eredményekhez

A fejezetben ismertetett eredményekhez a következőkkel járultam hozzá:

- Közös megbeszéléseken részt vettem az irányított k-klikk perkolációs algoritmus kidolgozásában. A CPMd módszerrel meghatároztam több, az irodalomban használt transzkripció szabályozási hálózat irányított moduljait és a GO TermFinder funkció keresővel a modulok legjelentősebb biológiai funkcióit.

- A CMPd algoritmus tesztelésére a transzkripció szabályozási hálózatokon túl a következő adatokat gyűjtöttem és alakítottam hálózatos formába [T4]: irányított szó asszociációk és a Google statikus saját oldalai közötti hiperlinkes kapcsolatok. Az utóbbi hálózat elérhető a http://CFinder.org/data helyen.

- Szintén a CPMd algoritmus tesztelése érdekében numerikusan meghatároztam azt az él sűrűséget (a kritikus él sűrűséget a rendszer méret függvényében), amelynél az irányított Erdős-Rényi hálózatban bekövetkezik az irányított k-klikk perkolációs átalakulás (a [T4] publikáció 3. ábrája).

- Egy, a CPMd-től eltérő TR hálózati modul keresés [T5] során beprogramoztam kis irányított részgráfok keresését és az előfordulási számuk szignifikanciájának meghatározását olyan élkeverési tesztekkel, amelyek változatlanul hagyják a csúcsok ki- és be-fokszámait, és a csúcsok rögzített csoportjainak (a TR hálózat „szintjeinek”) elem számát.

- A fehérje-fehérje kölcsönhatásokhoz és a TR hálózatokhoz hasonlóan a mikroRNS-ek esetén is elvégeztem számos név átalakítást az adatok összehasonlíthatósága érdekében.

- A társszerzőkkel együtt kidolgoztam a mikroRNS-ek moduljainak meghatározását és a mikroRNS-ek fontosságának („essentiality”) mérőszámát. Kiszámítottam az egyes mikroRNS-ek így definiált fontosságát. Teszteltem az eredményeket úgy, hogy a Pearson (kovariancia) korreláció helyett a Spearman (sorrend) korrelációt használtam. Teszteltem több eredményt élkeveréssel (az irányított mikroRNS$\rightarrow$mRNS hálózat csúcsainak be- és ki-fokszámait változatlanul tartva). A tesztek megtalálhatóak a [T7] publikáció kiegészítő anyagában („supplement”-jében).

dc_901_14

3. fejezet

Jelátviteli hálózatok moduljai (útvonalai)

Az előző két fejezettől eltérően ebben a fejezetben az alkalmazásokon van a hangsúly. A molekuláris biológia hálózatok segítségével történő kutatásához szükség van jó minőségű hálózatokra (kölcsönhatás listákra) és további adatokra. A jelátviteli kölcsönhatások esetében több élőlény jelátviteli útvonalait átfogóan és egységesen kezelő adatbázis a 2010-es évek előtt nem volt elérhető, azóta a fejlődés jelentős. Véleményem szerint az itt bemutatott SignaLink adatbázis ebben a folyamatban vesz részt eredményesen[1]. A SignaLink adatai jelenleg elérhetőek adatbázisokat integráló központi adatbázisokban, többek között a következő helyeken (a dolgozat PDF fájljában a hiperlinkek klikkelhetőek):

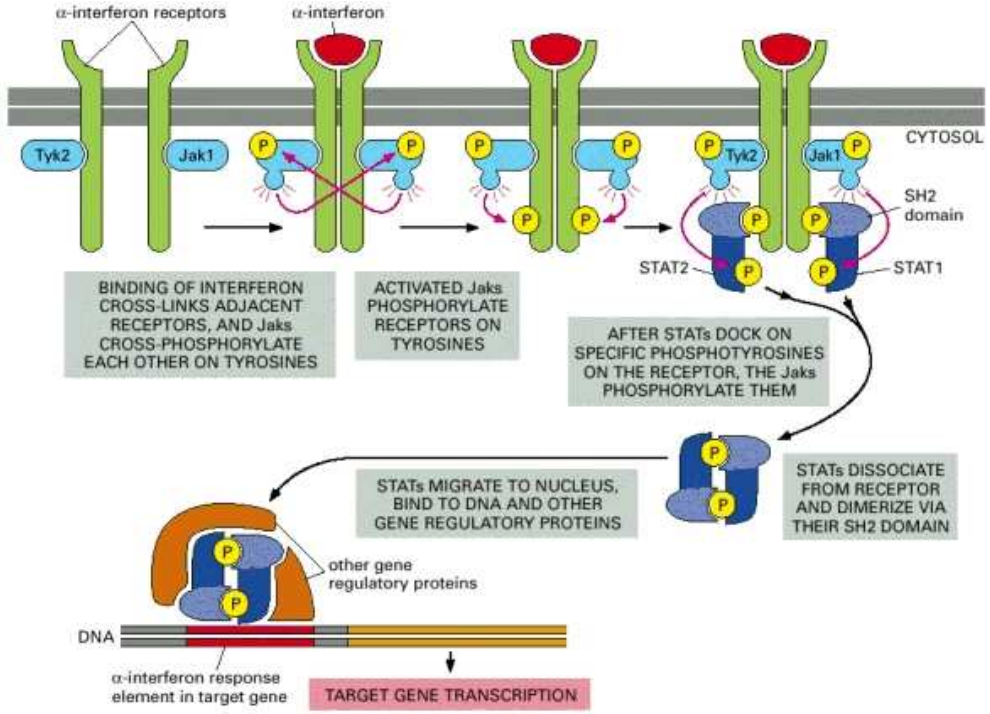
- **UniProt** – a http://www.uniprot.org/uniprot/P46551 adatlapon (*C. elegans* cdk-12 fehérje) az „Enzyme and pathway databases” címszónál szerepel a SignaLink-ből gyűjtött három kölcsönhatás.
- **FlyBase** – a http://flybase.org/reports/FBgn0004360.html adatlapon (*D. melanogaster* Wnt2 fehérje) az „Interactions and Pathways” fejezetben lévő „External Data” címszó alatt leírás és közvetlen hiperlink található a SignaLink adatbázis Wnt2 adatlapjára.
- **WormBase** – a *C. elegans* cwn-1 gén adatlapjának „External Links” részében szintén közvetlen link található a gén SignaLink adatlapjára.

[1] A jelátviteli adatbázisok fejlődési folyamatának részeként a SignaLink-et is bemutatja egy 2015-ben megjelent összefoglaló cikk 1-es ábrája [158] a http://dx.doi.org/10.1093/database/bau126 helyen.

Előzmények

A jelátvitel (signal transduction) szűkebb értelemben a biokémiai jelek továbbítása a sejt külső felszínétől a sejt belsejéig, általában a sejten kívül található jelző fehérjéktől (ligand fehérjéktől) a sejtmagban található transzkripciós faktor fehérjékig [159, 160]. A dolgozat ezt a szűkebb definíciót használja. Tágabb értelemben szokás a jelátvitelhez sorolni a sejt belsejében keletkezett jelek sejtmag felé történő továbbítását is [161]. A „signal transduction” fogalmat olyan fehérjéknek a vizsgálata során kezdték használni, amelyek a sejtmembrán külső oldalán lévő jelet a sejt belsejébe továbbítják. A külső jelet (például egy külső molekula megjelenését) a sejtnek egy, a membránhoz kötött fehérjéje általában továbbítja a sejtbelső felé. Amint egy fehérje megkapja a jelet, általában reverzibilis szerkezeti átalakuláson megy át. Ez a szerkezeti változás lehet nagy, például a fehérje széthasadhat két hasonló méretű részre. De gyakran igen kicsi ez a változás, például egy kis molekularész (foszfát csoport) kovalens kötéssel kapcsolódik a jelátviteli fehérjéhez. Ezek a szerkezeti változások teszik lehetővé, hogy a jelet megkapó fehérje a jel „átvétele” után és csak akkor képes legyen katalizálni azt a biokémiai reakciót, amellyel önmaga továbbítja a jelet. Ha a jel terjedése során két fehérje egymással kölcsönhatásba lép, akkor – egyszerűsítéssel – ezt a két fehérjét a jelátviteli hálózat két csúcspontjaként ábrázolhatjuk és a két fehérje közötti kölcsönhatást a jelátviteli hálózat egy irányított éle jelölheti. Mivel a jelátviteli kölcsönhatások jelentős része két fehérje közötti információ átvitel (gyakran enzimatikus reakcióval), ezért ezeknek a kölcsönhatásoknak a nagy része a fehérje-fehérje kölcsönhatások térképén megtalálható. A fehérje-fehérje kölcsönhatások között a jelátviteli reakciókat a következők alapján szokás azonosítani: (1) a reakció irányított, (2) egy biokémiailag jól körülhatárolható külső jel típus továbbítását végzi, (3) a jel továbbításához a jel belsejében használt reakciók biokémiai típusa (például tirozin kináz reakció) jellemző egy jelátviteli útvonalra. A jelátviteli útvonalak fehérjéit szokás besorolni az útvonalak ligand, receptor, mediátor, ko-faktor és transzkripciós faktor fehérje kategóriáiba.

A jelátviteli fehérjék modulokba (útvonalakba) történő csoportosítása sokáig a felfedezési sorrend alapján haladt. Ebben a folyamatban meghatározóak voltak a jel sejtbe történő „beléptetését” végző fehérjék és a jelet a sejten belül továbbító reakció típus. Például a Jak (Janus acting kinase) és a STAT (Signal Transducer and Activator of Transcription) fehérjék döntően foszforilációs és dimerizációs kölcsönhatásai nyomán



3.1. ábra. A soksejtű élőlényekben gyakran megtalálható Jak-STAT jelátviteli útvonal működésének vázlata (az ábra átvétel a „Molecular Biology of the Cell” c. könyvből [161]). Az útvonal a nevét két meghatározó fehérje típusáról kapta: Jak (Janus acting kinase) és STAT (Signal Transducer and Activator of Transcription). A külső jel molekula (α-interferon, piros színnel) hatására a sejtmembránhoz kötött receptor molekula (zöld színnel) két példánya összekapcsolódik (dimerizálódik). A receptor dimerizálódása után a komplex sejtbelső felőli oldalán lévő Jak típusú molekulák (Tyk2 és Jak1, világoskék színnel) egymással kölcsönhatásba lépnek. Ezután a P-vel (sárga színű kör) jelölt foszfát csoportok segítségével a jel a STAT molekulákra kerül (STAT1 és STAT2, sötétkék), amelyek dimerizálódás után bejutnak a sejtmagba és szabályozzák a transzkripciót.

a sejtmagba vezető jelátviteli kölcsönhatások összefoglaló neve Jak/STAT jelátviteli útvonal (ld. 3.1. ábra). Néhány további hasonló útvonal az RTK (ennek az útvonalnak a **r**eceptorai **t**irozin **k**ináz fehérjék, a Notch, a TGF-β vagy a Wnt(Wingless). A biológiában a jelátviteli útvonalakról alkotott kép sokáig az volt, hogy az útvonalak az eltérő receptoraik és biokémiai átviteli módjaik alapján jelentősen elkülönülnek, és egymáshoz csupán gyengén kapcsolódnak. Ehhez képest az 1990-es évektől kezdődően a jelátvitelt vizsgáló kísérletekben egyre több olyan kölcsönhatás vált ismertté, amelyik útvonalakat köt össze. A jelátvitelben az útvonalak közötti kölcsönhatások neve neve cross-talk („áthallás”).

Eredmények

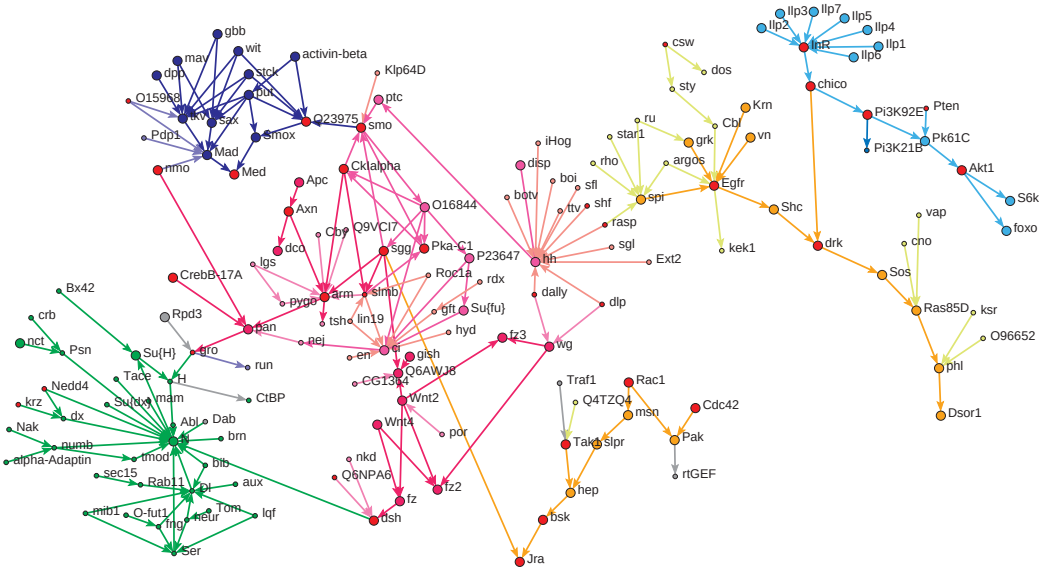
3.1. Átfedő jelátviteli útvonalak összeállítása egységesített gyűjtési kritériumokkal és adatszerkezettel [T8, T9]

A jelátviteli kölcsönhatások vizsgálatához általában a sejten belül speciális helyen lévő vagy speciális állapotú fehérjékhez kell hozzáférni: a legtöbb jelátviteli fehérje működéséhez szükséges, hogy egy membránhoz kapcsolódjon vagy egy, a transzláció utáni speciálisan módosított állapotban[2] legyen. Emiatt az általános fehérje-fehérje kölcsönhatások esetén említett „high-throughput" (sok kölcsönhatást egyszerre tesztelő) kísérleti módszerek a jelátviteli kölcsönhatások feltérképezésére kevésbé használhatóak, és a jelátviteli útvonalak kísérleti kutatása napjainkban is döntően útvonalanként külön-külön történik, egyszerre egy vagy néhány kölcsönhatás vizsgálatával. Az útvonalak nagyrészt egyenként történő kutatása az átfogó, rendszer szintű elemzések – például időbeli folyamatok számítása – szempontjából nehézséget okoz. A jelátviteli adatbázisok még azonos élőlény esetén is gyakran eltérő kritériumok alapján állítják össze a jelátviteli útvonalakat, ezért egy (például dinamikai) elemzés számára lényeges útvonalak gyakran eltérő lefedettséggel és eltérő szempontok alapján összeállítva szerepelnek egy-egy gyűjtésben.

Annak érdekében, hogy számunkra és más kutatók számára a soksejtűek jelátviteli útvonal kölcsönhatásai több élőlényben és több útvonalban azonos lefedettséggel rendelkezésre álljanak, gyűjtést végeztünk, amelynek pontos részletei megtalálhatóak a SignaLink adatbázis első publikációjának kiegészítő anyagában [T8]. A gyűjtésünk első verziója két, a területen elismert összefoglaló cikk [159, 162] alapján az NHR fehérjéken (a sejtmag membránjához kötött hormon receptorok) túl 7 olyan útvonalat tartalmazott (EGF/MAPK, IGF, TGF-β, WNT/Wingless, Hedgehog(Hh) Jak/STAT, Notch), amelyek az egyedfejlődés során folyamatosan aktívak, a kifejlett élőlényekben is fontosak és a soksejtű (metazoa) élőlényekben széles körben megtalálhatóak (konzerváltak). A SignaLink adatbázis első (2010-es) verziója (SignaLink1) elérhető a http://signalink1.netbiol.org web címen.

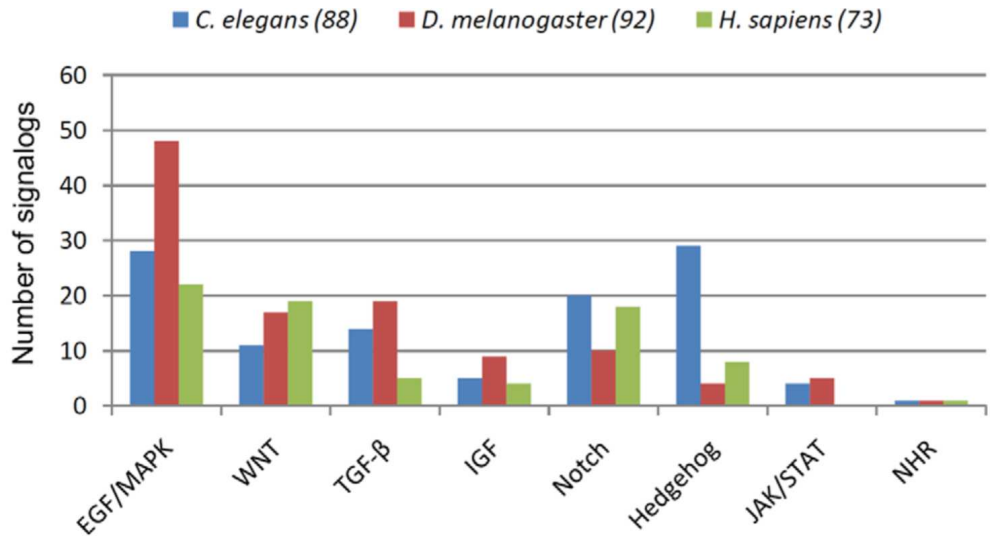
A SignaLink1 előtt elérhető főbb jelátviteli adatbázisokhoz képest – KEGG [164], Reactome [165], NetPath [166] – a SignaLink1 lényegesen több kísérleti publikáció

[2] Ezeknek a módosításoknak a neve Post-Translational Modification (PTM).



3.2. ábra. A SignaLink adatbázis első verziója a *Drosophila melanogaster* (gyümölcslégy) élőlényben 211 kölcsönhatást azonosított publikációk nem automatizált átnézésével. Az adatbázis első verziója a http://signalink1.netbiol.org címen érhető el. Az ábra a jelátviteli hálózat (irányított hálózat) gráf gyengén összekötött komponensei közül a legnagyobbat mutatja. Egy irányított hálózat két csúcspontja azonos gyengén összekötött komponensbe (weakly connected component) tartozik pontosan akkor, ha a két csúcs között van a gráf irányított élein át vezető, és az éleken tetszőleges irányokban haladó irányítatlan út. Egy irányított hálózatban két csúcspont azonos erősen összekötött komponensbe (strongly connected component) tartozik pontosan akkor, ha a két csúcs között mindként irányban van a gráf irányított élein át vezető olyan irányított út, amelyik mindegyik élen az adott él irányával azonos irányban halad. Az ábrán a csúcspontok és élek színei a jelátviteli útvonalakban betöltött szerepeket mutatják. A bal felső sarokban lévő kék színek a TGF-β útvonalat jelölik, a bal alsó sarokban látható zöld színű csúcsok a Notch útvonalba tartoznak. Az ábra közepén található rózsaszín árnyalatok a Wnt és a Hh (Hedgehog) útvonalak tagjait és kölcsönhatásait jelölik. A narancs és az arany szín az EGF, a világoskék szín árnyalatai (főként a jobb felső sarokban) az IGF útvonalat jelölik. A szürke szín a nagy útvonalakon kívüli további jelátviteli fehérjéket mutat. Az egynél több útvonalhoz tartozó fehérjéket jelölő csúcspontok piros színűek. A csúcspontok mérete az útvonalon belüli szerepet mutatja (nagy: központi, kis méret: kisegítő szerep). Az ábrát a doktori dolgozathoz készítettem a SignaLink1 adataiból. Az ábra elrendezését a Pajek hálózatrajzoló szoftverrel készítettem [163].

adatait tartalmazza. Összesen 941 cikk, köztük 170 útvonal alapú áttekintő cikk (review) alapján készültek a kölcsönhatás listák. Ezzel összehasonlítva a KEGG, a Reactome és a NetPath rendre 73, 166 és 351 publikációt használt fel. A [T8] cikkünk megjelenése idejében további előnye volt a SignaLink1-nek, hogy a KEGG adatbázisban csak fehérje komplex-ek és azok kölcsönhatásai voltak elérhetőek, nem egyedi fehérje-



3.3. ábra. A SignaLink adatbázis segítségével előrejelzett jelátviteli fehérjék száma útvonalak és élőlények szerint. Az ábra átvétel a [T10] publikációnkból, Korcsmáros Tamás készítette. Ugyanebben a publikációban 6 darab *C. elegans* fehérjének a Notch útvonalban előrejelzett részvételét társszerzők kísérletekkel ellenőrizték.

fehérje kölcsönhatások. Ezért a KEGG-ből csupán a korábban említett „mátrix" módszerrel (a csoportok összekötéséből adódó minden lehetséges kölcsönhatást feltételezünk) lehetett egyedi fehérje-fehérje kölcsönhatásokat kinyerni [16]. A SignaLink1-ben összegyűjtött jelátviteli kölcsönhatások egy részét mutatja a 3.2. ábra. A SignaLink adatbázis 2. verziója [T9] az első verzióval azonos három élőlényben tartalmaz a korábbinál többféle típusú irányított és irányítatlan molekuláris biológiai hálózatot összesítve. Azonos formátumban elérhetőek jelátviteli, jelátvitel szabályozási, poszttranszlációs módosítási, irányított fehérje-fehérje, transzkripció szabályozási és további kölcsönhatások. A SignaLink adatbázis friss weboldalán (http://SignaLink.org) az adatok letölthetőek és megtekinthetőek a hálózat ábrázolásai.

3.2. Fehérje csoportok jelátviteli elemzése online, valamint részvétel jelátviteli szerepek előrejelzésében és lehetséges gyógyszer célpontok kijelölésében [T10, T11, T12]

Az előző részben leírt jelátviteli kölcsönhatás listákat és útvonal annotációkat[3] elsőként jelátviteli fehérjék előrejelzésére[4] használtuk fel. A SignaLink adatbázisban

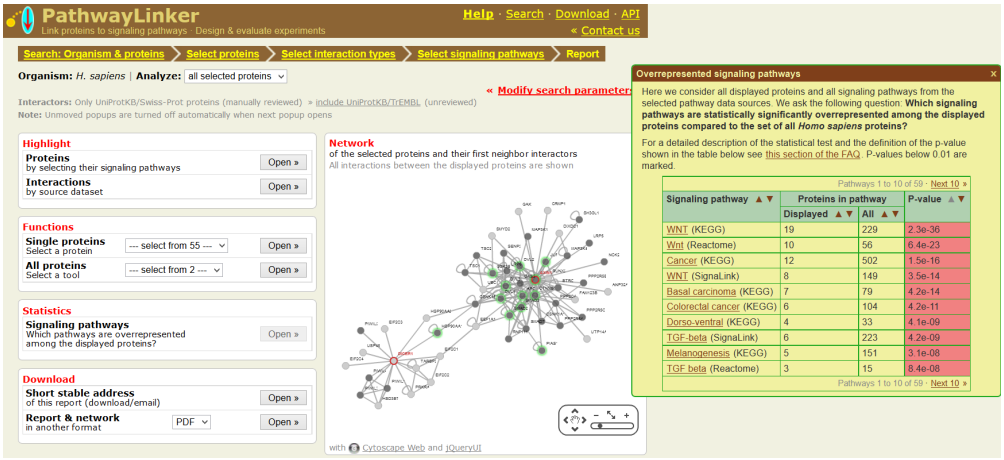
[3] A molekuláris biológiai adatbázisokban egy molekula vagy molekularészlet (fizikai, biokémiai és egyéb) tulajdonságainak általános neve idegen szóval „annotáció".

[4] Az angol nyelvű szakirodalomban a „prediction" szó és a „computational prediction" kifejezés általános. Magyarul a „jóslás" szó kevésbé megfelelő, ehelyett szerepel a dolgozatban az „előrejelzés" szó.

szereplő három élőlény (*C. elegans*, *D. melanogaster* és *H. sapiens*) fehérjéit kódoló DNS szekvenciák (gének) alapján megállapítható, hogy az egyik élőlény egy kiválasztott jelátviteli génjéhez egy másik élőlény melyik ismert génje hasonlít legjobban. A gének hasonlósága kapcsán központi fogalom a homológia. Két gén egymás homológja, ha egy közös ős leszármazottai. Ez a közös ős általában egy olyan gén, ami napjainkban már nem létezik. Ha ez a két gén azonos fajban található, akkor egymás paralógjai. Ha két különböző faj genomjában találhatóak, akkor egymás ortológjai[5]. A SignaLink adatai segítségével megkerestük a három faj mindegyikéből kiindulva az adott fajban felsorolt jelátviteli fehérjék másik két fajban megtalálható ortológjait. A [T10] publikációnkban a szekvencia hasonlóság vizsgálatához egy korábban már kidolgozott szekvencia-hasonlóság alapú módszer eredményeit használtuk [169]. Az előrejelzett jelátviteli gének számát – útvonalanként és élőlényenként külön – a 3.3. ábra mutatja. Az előrejelzett jelátviteli gének közül a *C. elegans* élőlény 6 fehérjéjének Notch útvonalban való működését a [T10] cikkünkben publikált kísérleti eredmények megerősítették. A biológus társszerzők által végzett kísérletek alapja az élőlény (fonálféreg) fenotípusának megfigyelése volt a hat gén funkciójának egyenként történő gátlása esetén.

A Notch útvonal kapcsán kísérletekkel igazolható eredményeket értünk el számítógépes elemzésekkel. Ezzel szemben a következő két elméleti módszer hosszabb távon, a gyógyszerkutatásban hozhat előrelépést. Az első módszer a mellékhatások minimalizálását próbálja elérni, és a nagy számú átfedő jelátviteli fehérje komplex közül kiválasztott egyetlen komplex gátlására ad javaslatot. Az általános fehérje-fehérje kölcsönhatásokhoz hasonlóan a jelátvitelben is számos fehérje komplex vesz részt. Ezek közül az egyik gyakran vizsgált komplex az APC/Axin/GSK-3β [170]. Ebben a komplex-ben a résztvevő fehérjék nagy részének van jelentős funkciója más komplex-ekben is. Tehát ha ennek a komplex-nek a működését egy-egy kiválasztott fehérjéje gátlásával próbáljuk befolyásolni, akkor sok mellékhatásra és esetenként akár a kívánttal ellentétes eredő hatásra is számíthatunk. Viszont ha a komplex mindegyik elemének funkcióját csak kissé gátoljuk – például a komplex-ben részt vevő összes fehérje kódoló génjének

[5] A doktori dolgozat első fejezetében említett „Bidirectional Best Hit” (BBH) nem jelent szükségszerűen egyben ortológ kapcsolatot is. Két friss elemzés szerint prokarióta (sejtmag nélküli egysejtű) élőlények esetén a BBH és az ortológia az esetek néhány százalékától eltekintve azonosak [167], viszont a növények (Viridiplantae) és az állatok (Metazoa) esetében a nagy számú gén duplikáció miatt a BBH-k az ortológok 40-70%-át nem találják meg [168].



3.4. ábra. A PathwayLinker online kereső szolgáltatás eredmény oldala az egyik példa keresés esetén (élőlény: *H. sapiens*, keresés: „Dicer, Axin”). Az eredmény lap két oszlopból áll, a baloldali oszlopban elemzési menüpontok találhatóak, a jobboldali oszlopban a hálózat interaktív ábrája látható. A „Statistics” → „Signaling pathways” menüpontban lévő „Open” gombra klikkelve jelenik meg az ábra jobb szélén látható ablak, amelyben a hipergeometrikus eloszlás alapján felülreprezentált jelátviteli útvonalak vannak felsorolva P érték szerint csökkenő sorrendben. A weboldalon további interaktív funkciók találhatóak. A PathwayLinker-hez részletes összeállítási és használati dokumentáció, valamint API (application programming interface) és PDF kimenet tartozik [T12]. Az ábrát a doktori dolgozathoz készítettem.

átírását gyengén gátoljuk, – akkor célzottan ennek a komplex-nek a működését tudjuk gátolni.

A második módszer a mellékhatások minimalizálása helyett azt próbálja vizsgálni, hogy melyik, jelenleg még nem vizsgált fehérjék gátlásával[6] vagy esetleg aktiválásával lehet kedvező hatást elérni. Ezt a megközelítést mutatja be a [T11] összefoglalónk. Az ismert gyógyszer célpontok különösen nagy számban fordulnak elő a jelátviteli útvonalakban, különösen az útvonalak átfedésében. Tehát az ilyen fehérjék között nagyobb valószínűséggel találhatóak újabb lehetséges gyógyszer célpontok. Ez a gyógyászati (gyógyszer célpont keresési) szempont különösen fontossá teszi a jelátviteli útvonalak részletes feltérképezését, nem csupán emberi sejtekben, hanem – az ortológia segítségével végezhető előrejelzések miatt – más élőlények sejtjeiben is.

A jelátviteli útvonalak könnyű kereshetősége, valamint a jelátviteli fehérjék és hálózati szomszédaik csoportos elemezhetősége érdekében létrehoztuk a PathwayLinker

[6] A genetika eszköztárában egy fehérje (gén, mRNS) gátlása vagy „kiütése” gyakoribb, mint a (túl)aktiválás.

online szolgáltatást a http://PathwayLinker.org címen. A címlapon kiválasztható „Quick search” segítségével az alapértelmezett beállításokat használja a kereső: minden keresett névhez az első megtalált fehérjét használja, és az alapértelmezett adatbázisokat és kölcsönhatás típusokat keresi a kölcsönhatások összegyűjtéséhez. Az „Advanced search”-ben lehetőség van a keresési feltételek részletes beállítására. Mindkét esetben a keresés eredményeként kapott fehérjéket és kölcsönhatásaikat hálózatként megtekintheti a felhasználó, és a fehérjék jelátviteli funkcióiról a fehérjékre összesített elemzést kap (egy példát mutat a 3.4. ábra). A PathwayLinker-ben a SignaLink által vizsgált három szervezet fehérje és gén nevei kereshetőek, valamint ember esetén gyógyszer nevek is.

Az eredmények összefoglalása

A fejezetben ismertetett eredmények közül a társszerzőségemmel készült eredmények a következők:

- Kidolgoztuk a SignaLink adatbázis „manual curation” típusú („kézi”, tehát nem automatizált) adat gyűjtésének kritériumait. A részletes leírás a [T8] publikációnk kiegészítő anyagában található.
- A SignaLink adatai megjelentek a fejezet elején említett központi „aggregált” adatbázisokban. Ez jelentős részben a SignaLink adatbázis minőségének és időszerűségének elismerése.
- A SignaLink adatai segítségével előrejeleztük jelátviteli útvonalak (modulok) új fehérjéit. Hat *C. elegans* fehérje esetén az adatok alapján jósolt eredményeket – a Notch jelátviteli útvonalban való részvételt – kísérletek igazolták.
- Létrehoztuk a PathwayLinker online kereső szolgáltatást, amelynek célja, hogy egy vagy több fehérje hálózati szomszédai és a talált összes fehérje részcsoportjainak vagy egészének jelátviteli funkciói gyorsan kereshetőek legyenek.

Saját hozzájárulásom az eredményekhez

A fejezetben ismertetett eredményekhez a következőkkel járultam hozzá:

- Két fős csapatként együttműködve pontosan leírtuk a SignaLink adatbázis egységesített adatgyűjtési kritériumait.
- Az együttműködés során létrehoztam a SignaLink weboldal és az adat tárolás első verzióját, amely jelenleg a http://signalink1.netbiol.org címen érhető el.
- A SignaLink jelátviteli útvonalaiban lehetséges gyógyszer célpontok kijelöléséhez kereséseket végeztem, és azok összesített eredményeit ábrázoltam.
- Beprogramoztam a PathwayLinker adatbázisát, valamint a kereső és weboldal kb. 80%-át. Ezek során megterveztem és megvalósítottam többek közt egy index-elést, amely a Linux fájlrendszerben egymásba ágyazott könyvtár nevekkel tárolja a kereshető karakterláncokat az egymás utáni karaktereik szerint.
- A PathwayLiner Help oldalain megírtam a szolgáltatás részletes megvalósítási és felhasználási dokumentációját (utóbbit közös megbeszélések alapján), és kb. 80%-ban a PathwayLinker API, PDF és hálózat megjelenítő kimenetét. A társszerzőkkel egyeztetett elvek alapján a megjelenített hálózat és az elemző ablakok („dialogs”) között kommunikációt építettem be a CytoscapeWeb és a jQuery szoftver csomagok felhasználásával.

dc_901_14

Köszönetnyilvánítás

Köszönöm mentoromnak, Vicsek Tamásnak, hogy a tudományos kutatói és személyiségbeli fejlődésemet 1998 óta folyamatosan és személyre szabottan segíti. Tudományos diákköri, MSc és PhD témevezetőm volt, jelenleg munkahelyi vezetőm. 2003 óta az általa vezetett MTA-ELTE kutatócsoportban dolgozom. Ő indított el először a csoportos mozgás vizsgálata irányában, később a véletlen hálózatok (gráfok), majd a molekuláris biológiai hálózatok átfedő moduljainak kutatása felé. A csoportjában nemzetközi szinten kiemelkedő lehetőséget biztosít az önálló kutatói pálya építésére, és ezt a lehetőséget próbálom hatékonyan használni. A tudományos kapcsolatain keresztül diákként és később is számos élvonalbeli külföldi kutatóval dolgozhattam együtt Németországban, az Egyesült Államokban több helyen és itthon. Közülük kiemelve köszönöm Dirk Helbingnek, Barabási Albert-László nak és Oltvai Zoltánnak, hogy az alapkutatás élvonalát sokféle szempontból megismerhettem és ezáltal a saját helyemet a tudomány világában jobban felmérhettem.

Köszönöm az összes társszerzőmnek, hogy együtt dolgozhattam velük és tanulhattam tőlük. Sokukkal jelenleg is kapcsolatban vagyok. Külön kiemelném közülük Palla Gergelyt, Derényi Imrét, Pollner Pétert, Hawoong Jeong-ot és Korcsmáros Tamást, akikkel többször együtt dolgoztunk. A munkámhoz számos technikai ismeretet Farkas Zénótól és Czirók Andrástól tanultam, köszönöm nekik a segítséget. Köszönöm korábbi tudományos diákköri témavezetőmnek, Vincze Imrének, hogy diákként a kutatócsoportjában tanulhattam a Mössbauer-spektroszkópia alapjait. Egyetemi oktatóim közül szeretnék kiemelten köszönetet mondani Rajkovits Zsuzsának és Dávid Gyulának, gimnáziumi tanáraim közül Pósfai Péternek és Pákó Gyulának. Köszönöm a KöMaL-nak, hogy érdekes feladatokat adott, amiken kitartóan kellett gondolkodni.

Nagyon köszönöm szüleimnek és családom többi tagjának, hogy stabil háttérrel és jó példákkal útnak indítottak, folyamatosan segítenek és alapos kérdéseikkel mindig rákényszerítenek arra, hogy átgondoljam, mit csinálok. Nagyon köszönöm feleségemnek, hogy együtt gondolkodunk és hogy folyamatosan biztosítja a nyugodt és vidám családi hátteret.

A tézispontokban használt publikációk társszerzőségemmel a Ph.D. után

[T1] G. Palla, I. Derényi, I. Farkas, T. Vicsek: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
Kivonat, Teljes cikk PDF, Kiegészítő anyagok (Supplement) PDF, Weboldal

[T2] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, T. Vicsek: CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* **22**, 1021–1023 (2006).
Kivonat, Teljes cikk PDF, Weboldal

[T3] I. J. Farkas, D. Ábel, G. Palla, T. Vicsek: Weighted network modules. *New Journal of Physics* **9**, 180:1–18 (2007).
Kivonat, Teljes cikk PDF, Weboldal

[T4] G. Palla, I. J. Farkas, P. Pollner, I. Derényi, T. Vicsek: Directed network modules. *New Journal of Physics* **9**, 186:1–21 (2007).
Kivonat, Teljes cikk PDF, Weboldal

[T5] I. J. Farkas, C. Wu, C. Chennubhotla, I. Bahar, Z. N. Oltvai: Topological basis of signal integration in the transcriptional-regulatory network of the yeast, *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7**, 478:1–12 (2006).
Kivonat, Teljes cikk PDF

[T6] I. J. Farkas, Q. K. Beg, Z. Oltvai: Exploring transcriptional regulatory networks in the worm. *Cell* **125**, 1032–1034 (2006).
Kivonat, Teljes cikk PDF

[T7] G. Boross, K. Orosz, I. Farkas: Human microRNAs co-silence in well-separated groups and have different predicted essentialities. *Bioinformatics* **25**, 1063–1069 (2009).
Kivonat, Teljes cikk PDF, Kiegészítő anyagok (Supplement) PDF

[T8] T. Korcsmáros *, I. J. Farkas *, M. S. Szalay, P. Rovó, D. Fazekas, Z. Spiró, C. Böde, K. Lenti, T. Vellai, P. Csermely: Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Bioinformatics* **26**, 2042–2050 (2010).
*: egyenlő hozzájárulás (megosztott első szerzők)
Kivonat, Teljes cikk PDF, Kiegészítő anyagok (Supplement) PDF, Weboldal

[T9] D. Fazekas *, M. Koltai *, D. Türei *, D. Módos, M. Pálfy, Z. Dúl, L. Zsákai, M. Szalay-Bekő, K. Lenti, I. J. Farkas, T. Vellai, P. Csermely, T. Korcsmáros: SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Systems Biology* **7**, 7:1–15 (2013).
*: egyenlő hozzájárulás (megosztott első szerzők)
Kivonat, Teljes cikk PDF, Weboldal

[T10] T. Korcsmáros, M. S. Szalay, P. Rovó, R. Palotai, D. Fazekas, K. Lenti, I. J. Farkas, P. Csermely, T. Vellai: Signalogs: orthology-based identification of novel signaling pathway components in three metazoans. *PLoS ONE* **6**, e19240:1–13 (2011).
Kivonat, Teljes cikk PDF

[T11] I. J. Farkas, T. Korcsmáros, I. A. Kovács, Á. Mihalik, R. Palotai, G. I. Simkó, K. Z. Szalay, M. Szalay-Bekő, T. Vellai, S. Wang, P. Csermely: Network-based tools for the identification of novel drug targets. *Science Signaling* **4**, pt3:1–12 (2011).
Kivonat, Teljes cikk PDF

[T12] I. J. Farkas, Á. Szántó-Várnagy, T. Korcsmáros: Linking proteins to signaling pathways for experiment design and evaluation. *PLoS ONE* **7**, e36202:1–5 (2012).
Kivonat, Teljes cikk PDF

További publikációk a társszerzőségemmel a Ph.D. fokozat megszerzése után

[M1] G. Palla, I. Farkas, I. Derenyi, A.-L. Barabasi, T. Vicsek: Reverse engineering of linking preferences from network restructuring. *Phys. Rev. E* **70**, 046115 (2004). DOI: 10.1103/PhysRevE.70.046115

[M2] I. J. Farkas, T. Vicsek: Initiating a mexican wave: an instantaneous collective decision with both short and long range interactions. *Physica A* **369**, 830-840 (2006). DOI: 10.1016/j.physa.2006.01.075

[M3] P. Pollner, G. Palla, D. Ábel, A. Vicsek, I. J. Farkas, I. Derényi, T. Vicsek. Centrality properties of directed module members in social networks. *Physica A* **387**, 4959 (2008). DOI: 10.1016/j.physa.2008.04.025

[M4] G. Palla, I. J. Farkas, P. Pollner, I. Derényi, T. Vicsek: Fundamental statistical features and self-similar properties of tagged networks. *New J. Phys.* **10**, 123026 (2008). DOI: 10.1088/1367-2630/10/12/123026

[M5] A. Szanto-Varnagy, P. Pollner, T. Vicsek, I. J. Farkas: Scientometrics: untangling the topics. *National Science Review*, cikk szám: nwu027 (2014). DOI: 10.1093/nsr/nwu027

[M6] M. Xu, Y. Wu, Y. Ye, I. Farkas, H. Jiang, Z. Deng: Collective crowd formation transform with mutual information-based runtime feedback. *Comp. Graph. Forum* (2014). DOI: 10.1111/cgf.12459

[M7] I. J. Farkas, J. Kun, Y. Jin, G. He, M. Xu: Keeping speed and distance for aligned motion. *Phys. Rev. E* **91**, 012807 (2015). DOI: 10.1103/PhysRevE.91.012807

Irodalomjegyzék

[1] C. Mora, C. P. Tittensor, S. Adl, A. G. B. Simpson, B. Worm: How Many Species Are There on Earth and in the Ocean? *PLoS Biol.* **9**, e1001127 (2011). DOI: 10.1371/journal.pbio.1001127

[2] C. Darwin: *On the Origin of Species by Means of Natural Selection*, (Murray, 1871). Link a könyv ingyenes elektronikus verziójára (Gutenberg Project)

[3] L. H. Hartwell, J. J. Hopfield, S. Leibler, A. W. Murray: From molecular to modular cell biology. *Nature* **402**, C47–52 (1999).
Link a cikk ingyenes PDF verziójára

[4] R. Sharan, I. Ulitsky, R. Shamir: Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007). DOI: 10.1038/msb4100129

[5] International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
DOI: 10.1038/35057062. Link a cikk ingyenes PDF verziójára

[6] S. Fields, O. Song: A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989). DOI: 10.1038/340245a0

[7] T. Ito, *et. al.*: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
DOI: 10.1073/pnas.061034498

[8] N. Simonis, *et. al.*: Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nat. Meth.* **6**, 47–54 (2009).
DOI: 10.1038/nmeth.1279

[9] J.-F. Rual, *et. al.*: Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
DOI: 10.1038/nature04209

[10] H. Yu, *et. al.*: High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008). DOI: 10.1126/science.1158684

[11] L. Giot, *et. al.*: A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727–1736 (2003). DOI: 10.1126/science.1090289

[12] S. Li, *et. al.*: A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004). DOI: 10.1126/science.1091403

[13] A. Brückner, C. Polge, N. Lentze, D. Auerbach, Uwe Schlattner: Yeast two-hybrid, a powerful tool for systems biology. *Int. J. Mol. Sci.* **10**, 2763–2788 (2009). DOI: 10.3390/ijms10062763

[14] G. Rigaut, *et. al.*: A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotech.* **17**, 1030–1032 (1999). DOI: 10.1038/13732

[15] O. Puig, *et. al.*: The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218–229 (2001). DOI: 10.1006/meth.2001.1183

[16] G. D. Bader, C. W. V. Hogue: Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotech.* **20**, 991–997 (2002). DOI: 10.1038/nbt1002-991

[17] A.-C. Gavin, *et. al.*: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002). DOI: 10.1038/415141a

[18] N. J. Krogan, *et. al.*: Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006). DOI: 10.1038/nature04670

[19] H. Zhu, *et. al.*: Global analysis of protein activities using proteome chips. *Science* **293**, 2102–2105 (2001). DOI: 10.1126/science.1062191

[20] Y. Ho, *et. al.*: Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002). DOI: 10.1038/415180a

[21] J. Joung, E. Ramm, C. Pabo: A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **97**, 7382–7387 (2000). DOI: 10.1073/pnas.110149297

[22] P. C. Havugimana, *et. al.*: A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012). DOI: 10.1016/j.cell.2012.08.011

[23] A.-L. Barabási, Z. N. Oltvai: Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004). DOI: 10.1038/nrg1272

[24] M. Vidal: A biological atlas of functional maps. *Cell* **104**, 333–339 (2001). DOI: 10.1016/S0092-8674(01)00221-5

[25] C. Sanchez, *et. al.*: Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. *Nucl. Acids Res.* **27**, 89–94 (1999). DOI: 10.1093/nar/27.1.89

[26] T. F. Smith, M. S. Waterman: Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
DOI: 10.1016/0022-2836(81)90087-5

[27] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman: Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
DOI: 10.1016/S0022-2836(05)80360-2

[28] A. J. Enright, I. Iliopoulos, N. C. Kyripides, C. A. Ouzounis: Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999). DOI: 10.1038/47056

[29] M. Costanzo, *et. al.*: The genetic landscape of a cell. *Science* **327**, 425–431 (2010). DOI: 10.1126/science.1180823

[30] B. Snel, G. Lehmann, P. Bork, M. A. Huynen: STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucl. Acids Res.* **15**, 3442–3444 (2000). DOI: 10.1093/nar/28.18.3442

[31] A. H. Y. Tong, *et. al.*: Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
DOI: 10.1126/science.1065810

[32] E. A. Winzeler, *et. al.*: Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
DOI: 10.1126/science.285.5429.901

[33] R. Hoffmann, A. Valencia: A gene network for navigating the literature. *Nat. Genet.* **36**, 664–664 (2004). DOI: 10.1038/ng0704-664

[34] L. Salwinski, D. Eisenberg: Computational methods of analysis of protein-protein interactions. *Curr. Op. Struct. Biol.* **13**, 377–382 (2003).
DOI: 10.1016/S0959-440X(03)00070-8

[35] C. von Mering, *et. al.*: STRING: a database of predicted functional associations between proteins. *Nucl. Acids Res.* **31**, 258–261 (2003).
DOI: 10.1093/nar/gkg034

[36] I. Lee, S. V. Date, A. T. Adai, E. M. Marcotte: A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
DOI: 10.1126/science.1099511

[37] P. Uetz, *et. al.*: A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
DOI: 10.1038/35001009

[38] H. W. Mewes, *et. al.*: MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucl. Acids Res.* **39**, D220–D224 (2011).
DOI: 10.1093/nar/gkq1157

[39] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori: The KEGG resource for deciphering the genome. *Nucl. Acids Res.* **32**, D277–D280 (2004). DOI: 10.1093/nar/gkh063

[40] C. von Mering, *et. al.*: Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002). DOI: 10.1038/nature750

[41] [Editorial]: Searching high and low for interactions. *Nat. Meth.* **4**, 377 (2007). DOI: 10.1038/nmeth0507-377

[42] C. von Mering, *et. al.*: STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucl. Acids Res.* **33**, D433–D437 (2005). DOI: 10.1093/nar/gki005

[43] J. De Las Rivas, C. Fontanillo: Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.* **6**, e1000807 (2010). DOI: 10.1371/journal.pcbi.1000807

[44] R. Goel, H. C. Harsha, A. Pandey, T. S. K. Prasad: Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. *Mol. BioSyst.* **8**, 453–463 (2012). DOI: 10.1039/C1MB05340J

[45] M. Uhlen, *et. al.*: Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010). DOI: 10.1038/nbt1210-1248

[46] T. W. Harris, *et. al.*: WormBase: a comprehensive resource for nematode research. *Nucl. Acids Res.* **38**, D463–D467 (2010). DOI: 10.1093/nar/gkp952

[47] S. E. St.Pierre, *et. al.*: FlyBase 102 – advanced approaches to interrogating FlyBase. *Nucl. Acids Res.* **42**, D780–D788 (2014). DOI: 10.1093/nar/gkt1092

[48] Arabidopsis Interactome Mapping Consortium: Evidence for network evolution in an *Arabidopsis* interactome map. *Science* **333**, 601–607 (2011). DOI: 10.1126/science.1203877

[49] The UniProt Consortium: Activities at the Universal Protein Resource (UniProt). *Nucl. Acids Res.* **42**, D191–D198 (2014). DOI: 10.1093/nar/gkt1140

[50] D. L. Wheeler, *et. al.*: Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* **35**, D5–D12 (2007). DOI: 10.1093/nar/gkl1031

[51] I. Xenarios, *et. al.*: DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* **30**, 303–305 (2002). DOI: 10.1093/nar/30.1.303

[52] S. Orchard, *et. al.*: The MIntAct project – IntAct as a common curation platform for 11 molecular interaction databases. *Nucl. Acids Res.* **42**, D358–D363 (2014). DOI: 10.1093/nar/gkt1115

[53] A. Chatr-Aryamontri, *et. al.*: The BioGRID Interaction Database: 2013 update. *Nucl. Acids Res.* **41**, D816–D823 (2013). DOI: 10.1093/nar/gks1158

[54] G. D. Bader, D. Betel, C. W. V. Hogue: BIND: the Biomolecular Interaction Network Database. *Nucl. Acids Res.* **31**, 248–250 (2003). DOI: 10.1093/nar/gkg056

[55] L. Licata, *et. al.*: MINT, the molecular interaction database: 2012 update. *Nucl. Acids Res.* **40**, D857–D861 (2012). DOI: 10.1093/nar/gkr930

[56] H. Jeong, S. P. Mason, A.-L. Barabási, Z. N. Oltvai: Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001). DOI: 10.1038/35075138

[57] R. Albert, H. Jeong, A.-L. Barabási: Internet: Diameter of the World-Wide Web. *Nature* **401**, 130–131 (1999). DOI: 10.1038/43601

[58] D. J. Watts, S. H. Strogatz: Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998). DOI: 10.1038/30918

[59] R. Albert: Scale-free networks in cell biology. *J. Cell. Sci.* **118**, 4947–4957 (2005). DOI: 10.1242/jcs.02714

[60] S. Maslov, K. Sneppen: Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002). DOI: 10.1126/science.1065103

[61] E. N. Gilbert: Random Graphs. *Ann. Math. Stat.* **30**, 1141–1144 (1959). DOI: 10.1214/aoms/1177706098

[62] P. Erdős, A. Rényi: On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61 (1960).
Link a cikk PDF fájljára (a Rényi Intézet honlapján)

[63] B. Bollobás: *Random Graphs*, 2. kiadás. (Cambridge University Press, 2001, ISBN: 0-521-79722-5). Link a könyv Google Books oldalára

[64] A. Wagner: The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18**, 1283–1292 (2001).
Link a cikk weboldalára és ingyenes PDF fájljára

[65] S.-H. Yook, Z. N. Oltvai, A.-L. Barabási: Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928–942 (2004). DOI: 10.1002/pmic.200300636

[66] A.-L. Barabási, R. Albert: Emergence of scaling in random networks. *Science* **286**, 509–512 (1999). DOI: 10.1126/science.286.5439.509

[67] A. Clauset, C. R. Shalizi, M. E. J. Newman: Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009). DOI: 10.1137/070710111

[68] A.-L. Barabási, R. Albert, H. Jeong: Mean-field theory for scale-free random networks. *Physica A* **272**, 173–187 (1999).
DOI: 10.1016/S0378-4371(99)00291-5

[69] S. N. Dorogovtsev, J. F. F. Mendes, A. N. Samukhin: Structure of growing networks with preferential linking. *Phys. Rev. Lett.* **85**, 4633–4636 (2000).
DOI: 10.1103/PhysRevLett.85.4633

[70] H. A. Simon: On a class of skew distribution functions. *Biometrika* **42**, 425–440 (1955). DOI: 10.1093/biomet/42.3-4.425
Link a teljes cikk ingyenes PDF fájlára

[71] J. Berg, M. Lässig, A. Wagner: Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol.* **4**, 51 (2004). DOI: 10.1186/1471-2148-4-51

[72] P. Holme, B. J. Kim: Growing scale-free networks with tunable clustering. *Phys. Rev. E* **65**, 026107 (2002). DOI: 10.1103/PhysRevE.65.026107

[73] J. S. Bader, A. Chaudhuri, J. M. Rothberg, J. Chant: Gaining confidence in high-throughput protein interaction networks. *Nat. Biotech.* **22**, 78–85 (2004).
DOI: 10.1038/nbt924

[74] M. Ashburner, *et. al.*: Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000). DOI: 10.1038/75556

[75] M. E. J. Newman, M. Girvan: Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).
DOI: 10.1103/PhysRevE.69.026113

[76] M. Girvan, M. E. J. Newman: Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**, 7821–7826 (2002).
DOI: 10.1073/pnas.122653799

[77] S. Tavazoie, *et. al.*: Systematic determination of genetic network architecture *Nat. Genet.* **22**, 281–285 (1999). DOI: 10.1038/10343

[78] P. Tamayo, *et. al.*: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999). DOI: 10.1073/pnas.96.6.2907

[79] J. Khan, *et. al.*: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Medicine* **7**, 673–679 (2001). DOI: 10.1038/89044

[80] O. Alter, P. O. Brown, D. Botstein: Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106 (2000). DOI: 10.1073/pnas.97.18.10101

[81] W. S. Noble: What is a support vector machine? *Nat. Biotech.* **24**, 1565–1567 (2006). DOI: 10.1038/nbt1206-1565

[82] A. W. Rives, T. Galitski: Modular organization of cellular networks. *Proc. Natl. Acad. Sci. USA* **100**, 1128–1133 (2003). DOI: 10.1073/pnas.0237338100

[83] V. Spirin, L. A. Mirny: Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA* **100**, 12123–12128 (2003). DOI: 10.1073/pnas.2032324100

[84] M. Blatt, S. Wiseman, E. Domany: Superparamagnetic clustering of data. *Phys. Rev. Lett.* **76**, 3251–3254 (1996). DOI: 10.1103/PhysRevLett.76.3251

[85] T. Nepusz, A. Petróczi, L. Négyessy, F. Bazsó: Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E* **77**, 016107 (2008). DOI: 10.1103/PhysRevE.77.016107

[86] L. Fu, E. Medico: FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinf.* **8**, 3 (2007). DOI: 10.1186/1471-2105-8-3

[87] U. Güldener, *et. al.*: CYGD: the Comprehensive Yeast Genome Database. *Nucl. Acids Res.* **33**, D364–D368 (2005). DOI: 10.1093/nar/gki053

[88] J. D. Han, *et. al.*: Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004). DOI: 10.1038/nature02555

[89] N. N. Batada, *et. al.* Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol.* **4**, e317 (2006). DOI: 10.1371/journal.pbio.0040317

[90] N. N. Batada, L. D. Hurst, M. Tyers: Evolutionary and physiological importance of hub proteins. *PLoS Comput. Biol.* **2**, e88 (2006). DOI: 10.1371/journal.pcbi.0020088

[91] I. K. Jordan, Y. I. Wolf, E. V. Koonin: No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* **3**, 1 (2003). DOI: 10.1186/1471-2148-3-1

[92] R. Saeed, C. M. Deane: Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinf.* **7**, 128 (2006). DOI: 10.1186/1471-2105-7-128

[93] I. Derényi, G. Palla, T. Vicsek: Clique percolation in random networks. *Phys. Rev. Lett.* **94**, 160202 (2005). DOI: 10.1103/PhysRevLett.94.160202

[94] R. M. Karp: Reducibility Among Combinatorial Problems. In: R. E. Miller, J. W. Thatcher (szerk.): Complexity of Computer Computations. New York, Plenum (1972) 85–103. oldal. DOI: 10.1007/978-1-4684-2001-2_9

[95] J.-P. Onnela, J. Saramäki, J. Kertész, K. Kaski: Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E* **71**, 065103 (2005). DOI: 10.1103/PhysRevE.71.065103

[96] F. Zhang, *et. al.*: Overlapping node discovery for improving classification of lung nodules. *IEEE Eng. Med. Biol. Soc. (EMBC), 35th Ann. Int. Conf. IEEE*, 5461–5464 (2013). DOI: 10.1109/EMBC.2013.6610785

[97] E. Georgii, S. Dietmann, T. Uno, P. Pagel, K. Tsuda: Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics* **25**, 933–940 (2009). DOI: 10.1093/bioinformatics/btp080

[98] U. N. Raghavan, R. Albert, S. Kumara: Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* **76**, 036106 (2007). DOI: 10.1103/PhysRevE.76.036106

[99] G. Tibély, J. Kertész: On the equivalence of the label propagation method of community detection and a Potts model approach. *Physica A* **387**, 4982–4984 (2008). DOI: 10.1016/j.physa.2008.04.024

[100] M. Rosvall, C. T. Bergstrom: Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* **105**, 1118–1123 (2008). DOI: 10.1073/pnas.0706851105

[101] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre: Fast unfolding of communities in large networks. *J.* Stat. Mech. P10008 (2008). DOI: 10.1088/1742-5468/2008/10/P10008

[102] A. Lancichinetti, S. Fortunato, J. Kertész: Detecting the overlapping and hierarchical community structure in complex networks *New J. Phys.* **11**, 033015 (2009). DOI: 10.1088/1367-2630/11/3/033015

[103] Y. Y. Ahn, J. P. Bagrow, S. Lehmann: Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010). DOI: 10.1038/nature09182

[104] S. Gregory: Finding overlapping communities in networks by label propagation. *New J. Phys.* **12**, 103018 (2010). DOI: 10.1088/1367-2630/12/10/103018

[105] C. Lee, F. Reid, A. McDaid, N. Hurley: Seeding for pervasively overlapping communities. *Phys. Rev. E* **8**3, 066107 (2011). DOI: 10.1103/PhysRevE.83.066107

[106] S. Fortunato: Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010). DOI: 10.1016/j.physrep.2009.11.002

[107] T. Nepusz, H. Yu, A. Paccanaro: Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Meth.* **9**, 471–472 (2012). DOI: 10.1038/nmeth.1938

[108] A. J. Enright, S. van Dongen, C. A. Ouzounis: An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* **30**, 1575–1584 (2002). DOI: 10.1093/nar/30.7.1575

[109] G. D. Bader, C. W. V. Hogue: An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf.* **4**, 2 (2003). DOI: 10.1186/1471-2105-4-2

[110] H. Yu, A. Paccanaro, V. Trifonov, M. Gerstein: Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* **22**, 823–829 (2006). DOI: 10.1093/bioinformatics/btl014

[111] I. A. Kovács, R. Palotai, M. S. Szalay, P. Csermely: Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PLoS ONE* **5**, e12528 (2010). DOI: 10.1371/journal.pone.0012528

[112] K. Macropol, T. Can, A. Singh: RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinf.* **10**, 283 (2009). DOI: 10.1186/1471-2105-10-283

[113] P. Shannon, *et. al.*: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003). DOI: 10.1101/gr.1239303

[114] M. Wu, X. Li, C.-K. Kwoh, S.-K. Ng: A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinf.* **10**, 169 (2009). DOI: 10.1186/1471-2105-10-169

[115] Y.-K. Shih, S. Parthasarathy: Identifying functional modules in interaction networks through overlapping Markov clustering. *Bioinformatics* **28**, i473–i479 (2012). DOI: 10.1093/bioinformatics/bts370

[116] P. F. Jonsson, P. A. Bates: Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291–2297 (2006). DOI: 10.1093/bioinformatics/btl390

[117] S. Zhang, X. Ninga, X.-S. Zhang: Identification of functional modules in a PPI network by clique percolation clustering *Comp. Biol. Chem.* **30**, 445–451 (2006). DOI: 10.1016/j.compbiolchem.2006.10.001

[118] J. Zhu, *et. al.*: Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* **40**, 854–861 (2008). DOI: 10.1038/ng.167

[119] K. Tarassov, *et. al.*: An in vivo map of the yeast protein interactome. *Science* **320** 1465–1470 (2008). DOI: 10.1126/science.1153878

[120] A.-C. Gavin, *et. al.*: Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006). DOI: 10.1038/nature04532

[121] S. R. Collins, *et. al.*: Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6**, 439–450 (2007). DOI: 10.1074/mcp.M600381-MCP200

[122] M. Babu, *et. al.*: Quantitative Genome-Wide Genetic Interaction Screens Reveal Global Epistatic Relationships of Protein Complexes in *Escherichia coli*. *PLoS Genet.* **10**, e1004120 (2014). DOI: 10.1371/journal.pgen.1004120

[123] S. Pinkert, J. Schultz, J. Reichardt: Protein interaction networks – more than mere modules. *PLoS Comput. Biol.* **6**, e1000659 (2010). DOI: 10.1371/journal.pcbi.1000659

[124] E. I. Boyle, *et. al.*: GO::TermFinder – open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004). DOI: 10.1093/bioinformatics/bth456

[125] S. N. Dorogovtsev, J. F. F. Mendes: *Evolution of networks: from biological nets to the Internet and WWW*. (Oxford University Press, 2003, ISBN: 978-0199686711). Link a könyv a Google Books oldalára

[126] F. W. Stearns: One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767–773 (2010). DOI: 10.1534/genetics.110.122549

[127] F. Crick: Central dogma of molecular biology. *Nature* **227**, 561–563 (1970). DOI: 10.1038/227561a0

[128] S. B. Prusiner: Prions. *Proc. Natl. Acad. Sci. USA* **95**, 13363–13383 (1998). DOI: 10.1073/pnas.95.23.13363

[129] P. Collas: The current state of chromatin immunoprecipitation. *Mol. Biotechnol.* **45**, 87–100 (2010). DOI: 10.1007/s12033-009-9239-8

[130] R. D. Finn, *et. al.*: Pfam: clans, web tools and services. *Nucl. Acids Res.* **34** (suppl.1), D247–D251 (2006). DOI: 10.1093/nar/gkj149

[131] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, A. G. Murzin: SCOP2 prototype: a new approach to protein structure mining. *Nucl. Acids Res.* **42**, D310–D314 (2014). DOI: 10.1093/nar/gkt1242

[132] D. Wilson, V. Charoensawan, S. K. Kummerfeld, S. A. Teichmann. DBD – taxonomically broad transcription factor predictions: new content and functionality. *Nucl. Acids Res.* **36** (suppl.1), D88–D92 (2008). DOI: 10.1093/nar/gkm964

[133] M. L. Bulyk: Computational prediction of transcription-factor binding site locations. *Genome Biol.* **5**, 201 (2003). DOI: 10.1186/gb-2003-5-1-201

[134] H. J. Bussemaker, H. Li, E. D. Siggia: Regulatory element detection using correlation with expression. *Nat. Genet.* **27**, 167–174 (2001). DOI: 10.1038/84792

[135] U. Alon: Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007). DOI: 10.1038/nrg2102

[136] S. S. Shen-Orr, R. Milo, S. Mangan, U. Alon: Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **64**, 64–68 (2002). DOI: 10.1038/ng881

[137] C. T. Harbison, *et. al.*: Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004). DOI: 10.1038/nature02800

[138] K. D. MacIsaac, *et. al.*: An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinf.* **7**, 113 (2006). DOI: 10.1186/1471-2105-7-113

[139] N. M. Luscombe, *et. al.*: Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308–312 (2004). DOI: 10.1038/nature02782

[140] S. Balaji, M. M. Babu, L. M. Iyer, N. M. Luscombe, L. Aravind: Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.* **360**, 213–227 (2006). DOI: 10.1016/j.jmb.2006.04.029

[141] B. Deplancke, D. Dupuy, M. Vidal, A. J. M. Walhout: A gateway-compatible yeast one-hybrid system. *Genome Res.* **14**, 2169–2175 (2004). DOI: 10.1101/gr.2445504

[142] S. Weingarten-Gabbay, E. Segal: The grammar of transcriptional regulation *Human Genetics* **133**, 701–711 (2014). DOI: 10.1007/s00439-013-1413-1

[143] D. P. Bartel: MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004). DOI: 10.1016/S0092-8674(04)00045-5

[144] A. Fire, *et. al.*: Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998). DOI: 10.1038/35888

[145] K. Chen, N. Rajewsky: The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* **8**, 93–103 (2007). DOI: 10.1038/nrg1990

[146] G. Karlebach, R. Shamir: Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780 (2008). DOI: 10.1038/nrm2503

[147] Z. Sun, X. Jin, R. Albert, S. M. Assmann: Multi-level modeling of light-induced stomatal opening offers new insight into its regulation by drought. *PLOS Comp. Biol.* **10**, e1003930 (2014). DOI: 10.1371/journal.pcbi.1003930

[148] F. Li, *et. al.*: The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. USA* **101**, 4781–4786 (2004). DOI: 10.1073/pnas.0305937101

[149] J. D. Orth, I. Thiele, B. O. Palsson: What is flux balance analysis? *Nat. Biotech.* **28**, 245–248 (2010). DOI: 10.1038/nbt.1614

[150] E. Segal, *et. al.*: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003). DOI: 10.1038/ng1165

[151] Z. Bar-Joseph, *et. al.*: Computational discovery of gene modules and regulatory networks. *Nat. Biotech.* **21**, 1337–1342 (2003). DOI: 10.1038/nbt890

[152] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, A. J. Enright: miRBase: microRNA sequences, targets and gene nomenclature. *Nucl. Acids Res.* **34**, D140–D144 (2006). DOI: 10.1093/nar/gkj112

[153] M. Selbach, *et. al.*: Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008). DOI: 10.1038/nature07228

[154] D. Baek, *et. al.*: The impact of microRNAs on protein output. *Nature* **455**, 64–71 (2008). DOI: 10.1038/nature07242

[155] S. Lall, *et. al.*: A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr. Biol.* **16**, 460–471 (2006). DOI: 10.1016/j.cub.2006.01.050

[156] P. Landgraf, *et. al.*: A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007).
DOI: 10.1016/j.cell.2007.04.040

[157] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, E. Segal: The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39**, 1278–1284 (2007).
DOI: 10.1038/ng2135

[158] S. Chowdhury, R. R. Sarkar: Comparison of human cell signaling pathway databases – evolution, drawbacks and challenges. *Database*, cikk szám: bau126 (2015). DOI: 10.1093/database/bau126

[159] A. Pires-daSilva, R. J. Sommer: The evolution of signalling pathways in animal development. *Nat. Rev. Genet.* **4**, 39–49 (2003). DOI: 10.1038/nrg977

[160] A. Bauer-Mehren, L. I. Furlong, F. Sanz: Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol.* **5**, 290 (2009). DOI: 10.1038/msb.2009.47

[161] B. Alberts, *et. al.*: *Molecular Biology of the Cell* (Garland Science, 2007, ISBN: 978-0815341055).
Link a könyv online kereshető és részenként olvasható verziójára

[162] J. Gerhart: 1998 Warkany lecture: signaling pathways in development *Teratology* **60**, 226–239 (1999). Link a cikk PDF fájljára

[163] W. De Nooy, A. Mrvar, V. Batagelj: *Exploratory social network analysis with Pajek* (Cambridge University Press, 2011, ISBN: 978-0521174800) Link a könyv Google Books oldalára

[164] H. Ogata, *et. al.*: KEGG: Kyoto Encyclopedia of Genes and Genomes *Nucl. Acids Res.* **27**, 29–34 (1999). DOI: 10.1093/nar/27.1.29

[165] G. Joshi-Tope, *et. al.*: Reactome: a knowledgebase of biological pathways. *Nucl. Acids Res.* **33**, D428–D432 (2005). DOI: 10.1093/nar/gki072

[166] K. Kandasamy, *et. al.*: NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* **11**, R3 (2010). DOI: 10.1186/gb-2010-11-1-r3

[167] Y. I. Wolf, E. V. Koonin: A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol. Evol.* **4**, 1286–1294 (2012). DOI: 10.1093/gbe/evs100

[168] D. A. Dalquen, C. Dessimoz: Bidirectional Best Hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biol. Evol.* **5**, 1800–1806 (2013). DOI: 10.1093/gbe/evt132

[169] A. C. Berglund, E. Sjolund, G. Ostlund, E. L. Sonnhammer: InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucl. Acids Res.* **36**, D263–D266 (2008). DOI: 10.1093/nar/gkm1020

[170] S. Ikeda, *et. al.*: Axin, a negative regulator of the Wnt signaling pathway, forms a complex with GSK-3beta and beta-catenin and promotes GSK-3beta-dependent phosphorylation of beta-catenin. *EMBO J.* **17**, 1371–1384 (1998). DOI: 10.1093/emboj/17.5.1371