Biológiai hálózatok átfedő modularizálását végző számítógépes programok és azok alkalmazási területei

Doktori (Ph.D.) értekezés

Szalay-Bekő Máté

Témavezetők:

Prof. Csermely Péter egyetemi tanár

Semmelweis Egyetem, Orvosi Vegytani, Molekuláris Biológiai és
Patobiokémiai Intézet

Dr. Papp Balázs tudományos főmunkatárs

Magyar Tudományos Akadémia, Szegedi Biológiai Központ,
Biokémiai Intézet

Biológus Doktori Iskola

Szegedi Tudományegyetem, Természettudományi és Informatikai Kar

Szeged
2013

Tartalomjegyzék

1	Bevezetés.....	3
1.1	A dolgozat felépítése.....	3
1.2	A hálózatkutatás területe.....	4
1.3	Biológiai hálózatok.....	7
1.4	Hálózatok modularizálása.....	16
2	Átfedő és fuzzy modularizálás.....	27
3	A ModuLand eljáráscsalád definíciója.....	32
3.1	A ModuLand eljáráscsalád főbb lépései.....	32
3.2	A LinkLand centralitás felület számítása.....	35
3.3	Az Arányos modul besoroló algoritmus.....	37
3.4	Magasabb hierarchikus szintek számítása.....	40
3.5	Modularizálás alapú mérőszámok.....	42
4	A ModuLand eljáráscsalád megvalósítása a Cytoscape programba ágyazva.....	44
4.1	A ModuLand plug-in felépítése.....	45
4.2	A ModuLand plug-in által nyújtott szolgáltatások.....	49
4.3	A ModuLand és más modularizáló plug-in-ek összehasonlítása.....	56
5	A ModuLand eljáráscsalád teljesítménye és optimalizálása.....	59
5.1	Gyenge élek számának redukálása a hálózat magasabb szintjein.....	60
5.2	Párhuzamosítási lehetőségek.....	62
6	Alkalmazási területek.....	67
6.1	Fehérje tulajdonságok jóslása.....	67
6.2	Gyógyszerkutatói alkalmazások.....	70
6.3	Modulszerkezetek elemzése, összehasonlítása.....	73
6.4	Hálózat vizualizációs eljárások.....	81
7	Összefoglalás.....	83
8	Summary.....	85
	Köszönetnyilvánítás.....	87
	Saját publikációk jegyzéke.....	88
	A disszertációhoz kapcsolódó közlemények.....	88
	A disszertációtól független közlemények.....	88
	Ábrajegyzék.....	93
	Irodalomjegyzék.....	95

1 Bevezetés

A doktori disszertációm célja a hálózatokkal (gráfokkal) modellezett komplex biológiai rendszerek könnyebb megértését és hatékonyabb elemzését szolgáló átfedő modularizálási módszerek területének bemutatása, valamint az ezen területen elért saját eredményeim ismertetése.

1.1 A dolgozat felépítése

A dolgozat első fejezetében először bevezetem az olvasót a hálózat kutatás interdiszciplináris területébe, bemutatva a hálózatos modellek elterjedtebb típusait. Ezt követően röviden összefoglalom a biológiai és orvostudományi területeken használt hálózatokat, ezek szerkezeti és dinamikai vizsgálatait. Végül bemutatom a hálózatok szerkezeti vizsgálatához használt elterjedt modularizálási módszereket, amelyek csoportokat, egymással szorosabban összefüggő régiókat azonosítanak a hálózatokban.

A második fejezetben definiálom a hálózatok átfedő és fuzzy modularizálásának részterületét, amely a disszertációm szűkebb témáját adja. Bemutatom a fontosabb koncepciókat, illetve a területen használt módszereket.

Ezt követően a harmadik fejezetben a kutatócsoportunk által definiált új átfedő modularizálási módszer, a ModuLand eljárás család elméleti hátterét mutatom be.

A negyedik és ötödik fejezetekben, illetve a hatodik fejezet egyes részeiben a saját tudományos munkámat mutatom be.

A negyedik fejezetben a Cytoscape nevű grafikus hálózatelemző és megjelenítő rendszerbe integrált ModuLand eljárás család megvalósításáról írok. Bemutatom a ModuLand plug-in felépítését, a plug-in által nyújtott szolgáltatásokat, illetve röviden írok a jelenleg elérhető, mások által megvalósított modularizáló Cytoscape plug-inokról.

Az ötödik fejezetben az általam implementált eljárások algoritmikus komplexitásáról írok. Különböző általam bevezetett optimalizációs lehetőségeket mutatok be specializált adatszerkezetekkel, közelítő eljárások használatával, illetve a több számítógépre elosztott futtatási lehetőségekkel kapcsolatban.

A hatodik fejezetben az átfedő modularizálás és a ModuLand plug-in biológiai és

orvosi felhasználási lehetőségeiről írok, részben saját magam által kutatott példákra, részben a szakirodalomból vett egyéb publikációkra támaszkodva.

1.2 A hálózatkutatás területe

A hálózatkutatás¹ fő célja, hogy a különböző természetes vagy mesterséges komplex rendszereket (például sejteket, társadalmi szerveződések, műszaki alkalmazásokat, kommunikációs rendszereket, stb.) hálózatokkal modellezze, illetve ezen modelleket felhasználva megértse a komplex rendszerek működését, sokszor éppen az által, hogy az egyes területeken előálló modelleket egymással hasonlítja össze. A hálózatkutatást körülbelül másfél évtizede kezdték intenzíven alkalmazni a biológia területén.

Nagyon sokféle hálózatos modell létezik, de mindegyik esetében pontok és azokat összekötő élek segítségével írjuk le a komplex rendszereket. A pontok mindig egymástól megkülönböztethető entitásokat jelölnek, míg az élek a pontokkal reprezentált entitások közötti kapcsolatra utalnak. Az alábbi lista a teljesség igénye nélkül sorol fel különböző tudományterületeket, ahol a jelen lévő komplex rendszereket hálózatkutató eszközökkel vizsgálni szokták. Az 1.3. és a 6. fejezetben további példák találhatók a biológiai és orvostudományi területekről.

- A szociológia területén igen régen foglalkoznak társadalmi hálózatokkal, amelyekben különböző közösségek (pl. baráti társaságok, családok, cégek) belső kapcsolatrendszerét vizsgálják. A 4. ábrán egy ilyen hálózatot mutatok be.
- A nyelvészet területén lehetőség van például szóasszociációs hálózatok elemzésére, ahol a hálózat elemei szavak és az őket összekötő asszociációk irányított kapcsolatként jelennek meg.
- Az irodalmi műveket is lehet hálózatos szempontból elemezni, például a regények szereplőinek hálózatainak segítségével. Az újságcikkek hálózatait szintén gyakran elemzik. Itt a cikkeket reprezentáló pontokat vagy társszerzőségi, vagy idézési kapcsolatok kötik össze.
- Az informatikában mind fizikailag valóságos hálózatokat (ahol számítógépek a

1 Angolul: network science

pontok), mind logikai hálózatokat vizsgálnak (ahol weblapok, vagy szoftver modulok közötti kölcsönhatásokat definiálnak). Az utóbbira láthatunk példát a 36. ábrán.

- A telekommunikáció területén szintén létezik a hagyományos értelemben vett fizikai kommunikációs hálózat (telefon készülékek, rádió állomások, műholdak és szerverek között), de gyakran szociális hálózatként is elemezhető például az emberek közötti telefonhívásokból felépített gráf.
- A közlekedési hálózatok szárazföldi, légi vagy vízi útvonalak hálózatait és az azokon való ember vagy anyagáramlás vizsgálatát teszik lehetővé.
- Az ökológiai hálózatok segítségével a táplálékláncokat vagy egyéb fajok közötti kapcsolatokat vizsgálhatjuk.
- Talán a legelterjedtebben kutatott hálózatok a sejt biológia területén vannak, ahol a sejtet felépítő különböző molekulák közötti összefüggéseket vizsgálják. A következő fejezetben számos példát említek erről a területről.

A fenti példák számából és a terület méretéből is jól látszik, hogy mennyire alkalmas a gráfelmélet a komplex rendszerek magas szintű modelljeinek leírására. Az 1. ábrán különböző típusú hálózatos modellek kerülnek szemléltetésre, a legegyszerűbb modellektől a jelentősen bonyolultabb leírásokig.

Az egyszerű gráfok (1A ábra) a legáltalánosabban használt, informatikai szempontból a legkönnyebben kezelhető leírások. Élei szimmetrikus kapcsolatot határoznak meg, vagyis az élek egyik végpontja sem kitüntetett. Az egyszerű gráfok egyik leggyakoribb kiterjesztése, amikor a kapcsolatokat irányokkal és/vagy súlyokkal (1B ábra) ruházzuk fel. Az így nyert modell kétségkívül specifikusabb és több információt tartalmaz, hiszen például egy telekommunikációs hálózat esetén jelölni tudja a hívás irányát (a hívó féltől a hívott fél felé mutatott kapcsolattal) és a beszélgetés hosszát, vagy gyakoriságát (az élhez rendelt súly segítségével). Számos elemzési módszer támogatja a súlyozott vagy irányított kapcsolatokat, de az új információk modellbe való bevezetése mindig körültekintést igényel. Például ha a cél az információ terjedés vizsgálata az adott telekommunikációs hálózaton, akkor a hívó és hívott felek szerepköre nem minden esetben határozza meg a kommunikáció irányát

(felhívhatok valakit azért is mert kérdezni akarok tőle, de azért is, hogy meséljek neki). Hasonlóan a beszélgetés hossza sincs mindig egyenes arányban az átvitt információ mennyiségével vagy fontosságával. Épp ezért hasznos sokszor az egyszerűbb modell elemzéséből kiindulni és csak fokozatosan bővíteni a modellt újabb típusú információkkal.

1. ábra: Elterjedten használt gráf típusok.

A fa gráfok (1C ábra) azért speciálisak, mert nem tartalmaznak köröket. Ez a tulajdonság ritkán teljesül a természetes komplex rendszerek esetén, de azért vannak alapvetően fa jellegű természetes hálózatok (például ökológiai rendszerekben táplálékláncok) és mesterséges hálózatok (rendsztani kategóriákat vagy egyéb ontológiákat leíró gráfok). A páros gráfok (1D ábra) már tartalmazhatnak köröket, viszont az egyszerű hálózatoktól

eltérően a pontokat itt két csoportra lehet bontani úgy, hogy a csoportok között fut él. Érdekes ökológiai példa a páros gráfokra, amikor a virág fajták és méh fajok pontjaiból felépülő hálózatban az élek a beporzást szimbolizálják.

A hipergráfok (1E ábra) a normál gráfok kiterjesztései, a pontokat hiperélekkel kötik össze, amelyek sajátossága hogy nem csak két pontot foglalnak magukba. Valójában a komplex rendszerek többsége hipergráf, hiszen a kölcsönhatás ritkán korlátozódik két résztvevőre. Például általában nem két fehérje hat kölcsön, hanem több fehérje egyidejűleg alakít ki egy fehérjekomplexet. Illetve a beszélgetések vagy a telefonhívások sem feltétlen párok között zajlanak le. Mégis, a hagyományos gráfokat egyszerűbb kezelni és értelmezni, a gráfelméletben is sokkal nagyobb hagyományuk van, ezért egyelőre a hipergráfok kevésbé elterjedtek a hálózatkutatás területén.

Az összetett gráfok (1F ábra) szintén a hagyományos gráfok kiterjesztései, ahol többféle pontot és többféle él típust különböztethetünk meg. Jó példa az összetett gráfokra a jelátviteli hálózatok területe, ahol pontként szerepelhet gén, mRNS, fehérje, transzkripciósfaktor, stb., míg az élek különféle szabályozást (például gátlást vagy serkentést) jelölhetnek.

1.3 Biológiai hálózatok

Számos biológiai rendszert lehet hálózatos eszközökkel modellezni: például ökológiai rendszereket, agyi régiók kapcsolatait, evolúciós leszármazási kapcsolatokat, stb. Azonban a biológia, bioinformatika és a hálózatkutatás területének legtöbbet kutatott és legnagyobb irodalommal rendelkező átfedése a sejt működését leíró biológiai modellek vizsgálata. Ennek egyik oka, hogy az elmúlt évek során hatalmas mennyiségű adat vált elérhetővé a biológus közösség számára, amelyek feldolgozásához és elemzéséhez új bioinformatikai módszerek kidolgozására van szükség. A hálózatkutatás intuitív és hasznos eszköztárat biztosít a sejt különböző alkotóelemei (például fehérjék, gének, metabolitok és különböző RNS típusok) között fennálló kapcsolatrendszer leírására, aminek a segítségével különböző sejt-szinten megjelenő folyamatok, funkciók, illetve ezek változása vizsgálható (például betegség vagy stressz hatás, fejlődési ciklusok vagy jelátviteli folyamatok).

A precíz laboratóriumi kísérletek eredményein és az irodalomban fellelhető adatokon túl az elmúlt évtized alatt ugrásszerűen fejlődött párhuzamosított és automatizált kísérletes

módszerek által generált adatok segítségével egyre egyszerűbb (és olcsóbb) a sejtek működését leíró hálózatok felépítése. Egyszerűsödött a sejtek genetikai állományának feltérképezése (Niedringhaus és mtsai, 2011), vagy a fehérjék egymással (Yu és mtsai, 2008) vagy a DNS láncok adott darabjaival való (Hallikas és Taipale, 2006) fizikai kölcsönhatásainak vizsgálata. Széles körben alkalmazott DNS chip-es (Brown és Botstein, 1999) és RNS szekvenáló (Mortazavi és mtsai, 2008) módszerekkel vizsgálható a sejt adott állapotában expresszálandó mRNS-ek száma, ami által adott szövet régiók akár teljes génállományra kiterjedő expressziós profiljai hozhatók létre, segítve a sejt belső dinamikájának vizsgálatát. Az mRNS expressziós szintek mérése mellett különböző (például tömegspektográfián (Ong és mtsai, 2003; Bantscheff és mtsai, 2012), vagy két dimenziós gél technológián (Friedman, 2007) alapuló) módszerek léteznek a fehérjék aktivitásának (mennyiségének) mérésére is. A fenti adatok felhasználásával különböző fajta információkat tartalmazó hálózattípusokkal írhatjuk le a sejt működését (Chen és mtsai, 2009). A 2. ábrán három különböző, elterjedten használatos megközelítést mutatok be.

A gén regulációs hálózatok pontjai gének, amelyek között indirekt kölcsönhatásokat definiálhatunk. Ilyen kölcsönhatás lehet például két gén között, ha az általuk expresszált mRNS molekulák mennyisége korrelál egymással, vagy ha az egyik gén által kódolt fehérje regulálja (serkenti vagy gátolja) a másik gén aktivitását. Ezen hálózatok előnye, hogy definiálásuk a többi modellhez képest könnyebb, illetve az alapvetően indirekt kölcsönhatásaiknak köszönhetően matematikailag egyszerűbben definiálható és verifikálható dinamikai modellek kialakítását teszik lehetővé.

A fehérje kölcsönhatási hálózatok talán a legszélesebb körben alkalmazott hálózat típusok a biológiai hálózatkutatásban. Pontjaik fehérjék, amelyeket különbözőféle kölcsönhatások kötnek össze. Direkt kölcsönhatások lehetnek a fehérjék összetapadását vizsgáló kísérletek eredményei, vagy például a fizikai kapcsolódás valószínűségét számító (szekvencia, domain szerkezet, lokációs vagy evolúciós adatokra támaszkodó) algoritmusok eredményei. Indirekt kölcsönhatást jelenthetnek például a funkcionális kapcsolatok, amikor a két fehérje ugyanabban a sejtes folyamatban (például egy jelátviteli útban) vesz részt. Szintén indirekt kapcsolatként jelennek meg a regulációs vagy metabolikus folyamatokból származó (DNS, RNS és metabolitokon keresztül lezajlott) kölcsönhatások. A fehérje kölcsönhatási

hálózatokat önmagukban ritkán használják dinamika leírására, viszont kiválóan alkalmasak funkcionális régiók meghatározására, vagy arra, hogy egy adott gén / fehérje funkcióját (vagy akár gyógyszer hatóanyagra való reagálását) valószínűsítsük. A kölcsönhatások általában irányítatlanok, de a kísérletek bizonytalansága, az adatforrások és kapcsolat definíciók diverzitása miatt gyakran rendelnek valószínűségi súlyokat a hálózat éleihez.

2. ábra: Sejtműködés leírására elterjedten alkalmazott biológiai hálózat típusok.

A transzkripció regulációs hálózatok használatakor méginkább arra törekednek hogy megközelítsék a sejt valódi működésének leírását. Ezek olyan összetett hálózatok, amelyekben különböző típusú pontok (például gének, fehérjék, vagy akár mRNS-ek, fehérje komplexek) között definiálunk különböző típusú, lehetőleg direkt kapcsolatot. A cél, hogy a fehérjék és gének közötti regulációs kölcsönhatásokat minél pontosabban definiálják.

Modellezési szempontból ezek a hálózatok (vagy akár az ezeknél bonyolultabb, metabolikus folyamatokkal is kiegészített hálózatok) lennének az ideálisak. Sajnos azonban minél pontosabb leírásra törekszik a modell, annál érzékenyebb az adatforrásokra és az adatok teljességére. Mivel az automatizált és párhuzamosított kísérletes módszerek (a pontos modell meghatározása szempontjából) komoly mennyiségű fals pozitív és negatív kapcsolatokat eredményeznek, a komplex és összetett hálózatok elsősorban csak kisebb (néhány tucat pontból álló) alrendszerekre készíthetők el megfelelő minőségben, kézi és kísérletes módszerek segítségével.

A fenti sejtes hálózat típusok mellett természetesen más hálózatos modelleket is használnak a biológiában. A fehérje szerkezeti hálók² elemei a fehérje aminosavai (ritkább esetben atomjai), és kapcsolatai az aminosavakat összekötő kémiai kötések (például peptid kötések, hidrogén-híd kötések, diszulfid hidak, stb). A kapcsolatokhoz lehet erősséget rendelni, amely meghatározásakor általában figyelembe veszik a kötés erősségét és az aminosavak térbeli elhelyezkedését. A gyógyszerkutatásban hasznosak például a betegségek feltérképezésére szolgáló betegség vagy mellékhatás hálózatok. Ökológiai hálózatokra jó példa a már említett virágok és rovarok pontjait, a beporzást definiáló élek segítségével leíró hálózat.

1.3.1 Biológiai hálózatok topológiája

Az alábbiakban áttekintek röviden néhány egyszerű és hagyományos mérőszámot, amelyek a biológiai (és bármilyen más típusú) hálózat bizonyos strukturális tulajdonságát mérik. Általában a topológia mély ismerete rendszer-specifikus számításokat vagy komplexebb módszerek alkalmazását igényli. Ilyen komplexebb módszer a modularizálás is és az erre épülő egyéb számítások. Azonban sokszor „első megközelítésként” érdemes egy ismeretlen hálózat topológiai elemzését egyszerűbb vizsgálatokkal kezdeni.

A pontok átlagos, minimális és maximális fokszáma, illetve a fokszám eloszlás igen informatív lehet. Ez utóbbit szokás valószínűségi függvényként definiálni, azaz $P(k)$ azt a valószínűséget jelenti, hogy egy véletlenszerűen választott pontnak éppen k szomszédja van. Igen fontos volt az a megfigyelés (Barabasi és Albert, 1999), hogy számos valós rendszer

² Egy konkrét példát a 6.3. fejezetben mutatok be.

fokszám eloszlása bizonyos tartományokban skálafüggetlen, $P(k) = Ak^{-\gamma}$, azaz annak a valószínűsége, hogy találunk egy nagyságrenddel nagyobb fokszámú pontot, megközelítőleg egy nagyságrenddel kisebb. A hálózatnak a legtöbb pontja kevés kapcsolattal rendelkezik, mégis létezik kevés számú, de sok kapcsolattal rendelkező pont („csomópont”, angolul „hub”). A skálafüggetlen jellegű fokszámeloszlást és a csomópontok létezését kimutatták fehérje-fehérje kölcsönhatási, gén regulációs és metabolikus hálózatokban egyaránt (Barabási és Oltvai, 2004). Ennek többek között evolúciós okai is lehetnek, hiszen az újabb és újabb fehérjék / sejtes rendszerek / funkciók tipikusan újrahasznosítják a korábban kialakult folyamatokat, ezáltal kötődnek a régen jelenlévő fehérjékhez, ami miatt kitüntetett csomópontok alakulnak ki (Barabasi és Albert, 1999; Albert, 2005).

A komplex rendszerekre általában jellemző a „kis világ” tulajdonság (Watts és Strogatz, 1998), vagyis két tetszőlegesen kiválasztott pont rövid úton összeköthető (a pontpárokat összekötő legrövidebb utak hossza a hálózat méretével logaritmikusan skálázódik). A pontok lokális környezetének sűrűségét méri a fürtösödési együttható³, amely a pont szomszédai között ténylegesen és lehetségesen futó élek számának arányaként definiálunk:

$$C_i = \frac{e_i}{k_i(k_i - 1)/2}$$

A fürtösödési együttható eloszlás függvényét ($C(k)$) szokás az adott fokszámú pontok átlagos fürtösödési együtthatójaként meghatározni. Így összehasonlítható a kis fokszámú (tipikusan periferiális) pontok és a csomópont jellegű pontok lokális csoportképződési hajlama. Számos valós, hierarchikus elrendeződést mutató komplex rendszer esetén megfigyelhető a $C(k) \approx k^{-1}$ összefüggés, amely szerint minél nagyobb fokszámú a pont, annál kisebb a valószínűsége hogy lokális környezete sűrű.

Metabolikus és jelátviteli utak vizsgálata során kimutatták (Maslov és Sneppen, 2002), hogy az őket leíró hálózatokban az olyan kapcsolatok valószínűsége statisztikailag a vártnál nagyobb mértékű, amelyek egy nagy fokszámú és egy kisebb fokszámú pont között futnak. Később az *S. cerevisiae* négy különböző fehérje-fehérje kölcsönhatási hálózatának vizsgálata során kimutatták a skálafüggetlen jellegű fokszám eloszlást és a hierarchikus modulok

3 Angolul: Clustering Coefficient

jelenlétét (Yook és mtsai, 2004). Ugyanezt később belátták (Li és mtsai, 2006a) *C. elegans*, és *D. melanogaster* esetén is. Mindegyik élőlény fehérje-fehérje kölcsönhatási hálózatára jellemző volt a kis világ tulajdonság, átmérőjük (az összes pontpár között kiszámolt legrövidebb utak közül a leghosszabb út hossza) hasonló 4-5 hosszúságú.

A strukturális elemzések egy érdekes esete, amikor a valamilyen biológiai szempontból érdekes pontok hálózatbeli helyzetét és fontosságát vizsgáljuk. Jó példa erre a fehérje kölcsönhatási hálózatok magas fokszámú pontjainak vizsgálata, amelyeket a kísérletes expressziós adatok elemzése során két eltérő csoportra bonthatunk (Han és mtsai, 2004; Ekman és mtsai, 2006; Kim és mtsai, 2006). Egyfelől *date hub*-ként definiálhatjuk azokat a fehérjéket, amelyek a sejt adott állapotát jellemző expressziós adatokban (például adott időben, adott stressz helyzet esetén, adott szövetben vagy a sejt egy adott területén) egyszerre csak kevés szomszédjukkal expresszálódnak. Azaz a sok lehetséges kapcsolatuk közül általában csak kevés létezik, de ezeket a kapcsolatokat a körülményekhez igazodva változtatják. Másfelől *party hub*-ként definiáljuk az olyan fehérje csomópontokat, amelyek sok kapcsolata jellemzően állandó és mindig közösen expresszálódik a kapcsolódó fehérjékkel. Noha az expressziós kísérletek összevont és kiátlagolt eredményeit tükröző fehérje kölcsönhatási hálózatokban mind a két típusú fehérje egyaránt csomópontként jelenik meg, a hálózatban elfoglalt helyzetük alapján mégis meg lehet őket különböztetni. Egyrészt a hálózat átfedő modularizálása segítségével elkülöníthetőek (lásd: 6.1. fejezet, 32. ábra), mivel a party hub-ok jellemzően a modulok belsejében, míg a date hub-ok különböző modulokat összekötő hídként jelennek meg (Kovács és mtsai, 2010). Másrészt például a date és party hub-ok motívumokban betöltött szerepének vizsgálata során (Jin és mtsai, 2007) kimutatható, hogy a date hub motívumok a hálózattól való törlésének hatására a hálózat a statisztikailag vártnál sokkal gyorsabban, míg a party hub motívumok törlése esetén sokkal lassabban esik szét.

A jelenleg vizsgált komplex biológiai hálózatok szerkezete fokozatosan alakult ki a múltban atominak tekinthető átalakulási lépések sorozatának hatására. Ilyen lépések (Sharan és Ideker, 2006) lehetnek például új kölcsönhatások születései, illetve ezek törlődései (például fehérjéket kódoló génszekvencia mutációk hatására), vagy pontok törlődései (például gének inaktiválódásával) és új pontok születései (génduplikáció esetén). A fenti biológiai folyamatokat számos hálózatépítő eljárás kidolgozása során figyelembe vették, amelyek célja

a valódi hálózatokhoz hasonló tulajdonságú hálózatok „növesztése” (Barabasi és Albert, 1999; Ravasz és mtsai, 2002; Berg és mtsai, 2004). Ezen evolúciós módszerek mindegyikére jellemző, hogy a régebben megjelent pontok tipikusan több kapcsolattal rendelkeznek. Alátámasztva ezen módszerek hitelességét, különböző fajok biológiai hálózatainak összehasonlítása során szintén kimutatták (Eisenberg és Levanon, 2003), hogy a régebben megjelent fehérjéknek több szomszédja van, illetve hogy az evolúció során az új kapcsolatok szerzésének valószínűsége arányos a fehérje pillanatnyi fokszámával (azaz a sok kapcsolattal rendelkező fehérje jellemzően még több kapcsolatot szerez).

1.3.2 Biológiai hálózatok dinamikája

Noha a disszertációm témája a biológiai hálózatok strukturális elemzésének egy új módszerének bemutatása, mégis a teljesség igénye nélkül röviden szeretném megemlíteni a területen az elmúlt években egyre fontosabbá váló hálózat dinamikai vizsgálatokat.

Amikor a biológiai hálózatok dinamikai elemzéseiről beszélünk, számos dologra gondolhatunk. Egyfelől érdekes lehet nyomon követni az egyed biológiai hálózatának időbeli strukturális változását akár rövid távon (például stressz helyzet vagy betegség kialakulásakor (Mihalik és Csermely, 2011)), akár hosszú távon (például az egyed élete során vett minták összehasonlításával (Harris és mtsai, 2009)). Hasonlóan gondolhatunk az előző fejezetben említett, az evolúció során végbement változásokat modellező szerkezeti átalakulásokra, de ide érthetjük akár a modulszerkezet változásának dinamikai elemzését is (Palla és mtsai, 2007; Rosvall és Bergstrom, 2010), amely például a modulok kialakulását, eltűnését, összefolyását és szétszakadását vizsgálja.

Másfelől a biológiai hálózat adott szerkezetének pontosabb megértését segíti, illetve bizonyos szempontból fontos pontok vagy élek meghatározását teszi lehetővé, ha a szerkezet fölött dinamikai jellegű szimulációkat futtatunk. Ilyen dinamikai szimuláció lehet például az adott pontokból indított perturbáció lecsengésének vizsgálata (Szalay és Csermely, 2013), aminek számos gyógyszerkutatói vonatkozása van (Farkas és mtsai, 2011; Csermely és mtsai, 2013), például gyógyszercélpontok meghatározása, vagy mellékhatások előrejelzése. Más jellegű, de szintén dinamikai elemzésként végezhetünk játékelméleti szimulációkat a hálózat pontjai felett, ahol az egyes pontok a szomszédaikkal közösen játékok sorozatában vesznek

részt (Wang és mtsai, 2008; Simko és Csermely, 2013). Ezek a szimulációk és algoritmusok nem szükségszerűen biológiai jellegűek, mégis igen hasznosak a rendszer szerkezetéből adódó dinamikai tulajdonságok feltérképezésére.

A sejt dinamikáját legjobban olyan modellekkel lehet leírni, amelyek a különböző biológiai entitások (fehérjék, mRNS-ek, metabolitok, stb.) számának változását vizsgálják az idő és a környezeti hatások függvényében (Chen és mtsai, 2009). Leggyakrabban az ilyen dinamikát nem is hálózatokkal, hanem különböző típusú egyenletrendszerek felírásával definiálják. Az S-System (Vilela és mtsai, 2008) az egyik legelterjedtebben használt nem-lineáris egyenletrendszerrel megfogalmazott biomolekuláris dinamikai modell, melynek általános felírása:

$$\dot{X}_i = a_i \prod_{j=1}^M X_j^{g_{ij}} - b_i \prod_{j=1}^M X_j^{h_{ij}}, i=1,2,\dots,M \quad ,$$

ahol X_i jelzi az i . metabolit koncentrációját, a_i és b_i a metabolitra jellemző pozitív konstansok, míg a g_{ij} és h_{ij} változók jelzik az adott metabolit keletkezésének és bomlásának az összes többi metabolittól való függőségét. Az egyenletrendszer paramétereinek megfelelő minőségű meghatározása főként kis méretű hálózatok esetén működik. A teljes sejt dinamikai modelljének felépítésére irányuló módszereknek sok nehézséget (Husmeier, 2003b; Zak és mtsai, 2003) kell legyőzniük. Ezek egyike a bemenetként használható adatok bizonytalansága (a rossz minőségű adatok, sok fals pozitív és negatív adat). De ha még az adatok tökéletesen megbízhatóak lennének is, a változók hatalmas számához képest túl kevés mérés áll rendelkezésre. Egy alapos microarray vizsgálat mRNS-ek ezreiről ad információt (ami változók millióit jelenti a modellben), míg az adott chippel az adott körülmények között tipikusan néhány (nagyságrendileg maximum egy tucat) mérést végeznek. Az egyenletrendszereknek ezért rengeteg megoldása létezik. Vagyis túl sok egymástól lényegesen eltérő dinamikai modell felírható, ami a gyakorlati alkalmazást erősen korlátozza. Mégis, az egyre jobb minőségű és olcsóbb mérési technikák, valamint az egyenletrendszerek paramétereit meghatározó módszerek fejlődése (Chowdhury és mtsai, 2013) következtében a dinamikai modellek visszafejtése egyre pontosabb és az általuk kinyert információ egyre használhatóbb lesz.

Az általános többváltozós egyenletrendszeres megfogalmazáson túl a sejt dinamikát lehet hálózatokkal is vizsgálni. Sokszor a különböző típusú felírás csak ugyanannak a modellnek az eltérő reprezentációja és a hálózatokkal felírt dinamika egyenletrendszerekkel ugyanúgy (analitikusan jobban kezelhető módon, bár az ember számára kevésbé áttekinthetően) felírható. A legelterjedtebben használt hálózatos reprezentációk a hálózat pontjaihoz különböző logikai függvényeket rendelnek, amelyek kifejezik hogy például az adott pont által reprezentált mRNS mennyisége hogyan függ a ponttal szomszédos többi mRNS számától. Amennyiben ezek a függvények tisztán bináris logikai függvények⁴, akkor Boole-hálózatokról (Kauffman, 1969) beszélünk. Ebben a matematikai modellben egy mRNS-nek, vagy egy fehérjének csak két állapota létezik, vagy expresszálódik, vagy nem. Ez a reprezentáció jelentősen leegyszerűsített dinamikai leírást tesz csak lehetővé, viszont cserébe gyorsan, nagy hálók esetén is visszafejthető és könnyen áttekinthető leírást ad (Shmulevich és mtsai, 2002; Lähdesmäki és mtsai, 2003; Wang és mtsai, 2012). A 3. ábrán egy Boole-háló szemléltetése látható, a G3 génhez tartozó logikai függvény igazságtáblával⁵ és hagyományos algebrai jelekkel való feltüntetésével.

3. ábra: Biológiai hálók dinamikáját leíró Boole-háló illusztrációja.

4 A bináris logikai függvények olyan matematikai kifejezések, ahol a változónak “igaz” vagy “hamis” értéke lehet és ezen változókat logikai műveletekkel (például “és”, “vagy”) kötjük össze.

5 Az igazságtáblák olyan táblázatok, ahol minden sorban a bemeneti paraméterek egy adott kombinációja mellett a bináris logikai függvény eredménye van feltüntetve.

A pontok dinamikáját leíró függvények az egyszerű logikai függvényeken túl lehetnek lineáris vagy nemlineáris determinisztikus⁶ függvények, de gyakran használnak valószínűségi reprezentációt is, amely jobban illeszkedik a hagyományos reakciókinetikai megfogalmazásokhoz. Létezik a Boole-hálóknak is valószínűségi kiterjesztése (Shmulevich és mtsai, 2002), illetve amennyiben a valószínűségi dinamikai modell nem tartalmaz irányított köröket⁷, akkor alkalmazhatjuk a Bayes-hálós felírási módszert (Husmeier, 2003a; Friedman, 2004), míg irányított körök esetén dinamikus Bayes-hálókat (Kim és mtsai, 2003), vagy például Markov-hálókat is használhatunk (Toh és Horimoto, 2002; Ma és mtsai, 2007).

1.4 Hálózatok modularizálása

A hálózatok statikus elemzésének területén az egyik legintenzívebben vizsgált kérdése a hálózat modulszerkezetének felderítése. A hálózatok moduljai a hálózaton belüli sűrűsödési régiók, amelyeken belül a pontok nagyobb hatással vannak egymásra, mint a hálózat többi részére. Egy szociális hálózatban modulok lehetnek például a szűkebb baráti körök, míg egy fehérje kölcsönhatási hálózatban a sejten belüli különböző funkciókért felelős fehérjék alkothatnak modulokat.

A modulszerkezet ismerete elengedhetetlen a komplex rendszer működésének megértése szempontjából, de segítségével érdekes következtetéseket lehet levonni a hálózat későbbi állapotáról is. Egy sokat idézett tanulmányban (Zachary, 1977) Wayne Zachary felmérte egy amerikai egyetemi karate klub szociális hálózatát, melyben 34 személy között 78 különböző erősségű baráti kapcsolatot írt le. A tanulmány érdekessége, hogy nem sokkal a kapcsolatok föltérképezése után a karate klub felbomlott, annak két befolyásos tagját érintő konfliktusa miatt. A klub kettéválása pont a 4. ábrán látható módon, a szociológiai hálózatban szemmel is jól kivehető modulhatáron történt.

Hasonlóan érdekes lehet a modulszerkezet ismerete egy telefon beszélgetéseket megjelenítő hálózatban. Például a telefonhívások bizonyos paraméterei és a hálózat

6 A determinisztikus függvények eredménye csak a bemenő paraméterektől függ, és annak eredménye minden esetben ugyanaz. Velük ellentétben a valószínűségi modellekben végzett szimulációknál (a szimulációt újra és újra elvégezve) eltérő végeredményt kaphatunk.

7 Irányított körnek nevezzük az irányított gráfokban megtalálható szomszédos élek olyan sorozatát, amely mentén haladva vissza érhetünk a kiinduló élhez. Az irányított köröktől való mentesség számos (gráf bejárást alkalmazó) algoritmus előfeltétele.

modulszerkezetének ismeretében a jövőbeli beszélgetések számos dinamikai jellemzőjét meg lehet jósolni (Onnela és mtsai., 2007; Hidalgo és Rodriguez, 2008). Az egyik ezzel kapcsolatos szabadalmamban az Ericsson cég számára munkatársaimmal közösen egy olyan rendszer alapelvét dolgoztuk ki (Szalay és mtsai., 2010), amely személyre szabott kedvezményeket biztosít a telefonáló ügyfelek számára, így növelve a beszélgetések jövőbeli intenzitását és a hálózat hosszú távú stabilitását. A szabadalmaztatott eljárás részben a modulszerkezet ismeretére is épít.

A 6. fejezetben számos további példát mutatok be, ahol a modularizálás gyakorlati alkalmazhatóságát vizsgálom meg részletesen a biológia és az orvostudomány területein.

4. ábra: A Zachary-féle szociológiai hálózat (Zachary, 1977) egy amerikai egyetemi karate klub baráti kapcsolatait jeleníti meg. A szürke és a fehér pontok a klub szétválása után keletkező két csoportot jelölik.

A modulok meghatározása könnyű akár ránézésre is, amennyiben a hálózat kicsi. Nagy hálózatok esetén azonban koránt sem ilyen könnyű a helyzet, és valamilyen egzakt matematikai definícióra van szükség. A modulok korábbi meghatározásunk szerint a környezetüknél sűrűbben összekötött lokális régiók a hálózatban. Ez a definíció közérthető ugyan, de nagyon magas szintű. Az ennél pontosabb algoritmikus, funkcionális vagy

matematikai definíciók megalkotása viszont számos problémát felvet. Az elmúlt évek során sok száz tudományos publikáció született a különböző modul definíciókról, modularizálási módszerekről, azok összehasonlításáról és használatáról. A területet összefoglaló áttekintő tanulmányokban (Newman, 2004a; Fortunato, 2010; Habib és Paul, 2010; Kovács és mtsai, 2010; Xie és mtsai, 2013; Newman, 2012) összesen több mint 140 egymástól lényegesen eltérő módszert találhatunk. A rendelkezésre álló hely szűke miatt nem tudom a teljes területet kellő részletességgel bemutatni a disszertációmban, de a fejezet további részében röviden áttekintem a modulok meghatározásának legelterjedtebb megközelítéseit. A területen sok matematikában és fizikában elterjedt fogalomra, eljárásra és képletre építenek. Az idegen fogalmakat igyekszem lábjegyzetben közérthetőbbé tenni, további információk pedig az egyes hivatkozott publikációkban találhatóak.

Hierarchikus modularizálás

A hierarchikus módszerek két fő típusba sorolhatók, úgy mint a felépítő algoritmusok (amikor kezdetben a hálózat minden pontja külön modulban helyezkedik el és ezeket fokozatosan összevonjuk), és a lebontó megoldások (kezdetben az egész hálózat egy modul és ezt daraboljuk szét).

5. ábra: Egy példa hálózat hierarchikus modularizálása (A panel) és a modularizáláshoz tartozó dendrogram (B. panel).

A hierarchikus modularizálás célja, hogy a jobban összekötött vagy szerkezeti szempontból hasonlóbb („közelebbi”) elemek kerüljenek egy modulba (Johnson, 1967). A felépítő eljárások egyik legfontosabb eleme a modulok távolságainak kiszámítása, ennek segítségével döntjük el hogy melyek a legközelebbi modulok, amelyeket a következő lépésben össze lehet vonni. Használhatjuk például a két modul pontjai közötti legrövidebb (*single linkage*), leghosszabb (*complete linkage*), vagy átlagos (*average linkage*) úthossz mértékét. Bonyolultabb mérőszámok figyelembe veszik hogy a két adott modul pontjai mennyire hasonló módon kötődnek a hálózat többi részéhez (Burt, 1976; Wasserman és Faust, 1994).

A hierarchikus eljárások eredményét szokás dendrogramon szemléltetni, amely egy fa jellegű struktúra: a legalul elhelyezkedő levél elemei felelnek meg a hálózat pontjainak, és a feljebb lévő elemek mind modulokat reprezentálnak, amelyek az általuk lefedett pontokból állnak. A dendogram legfelső eleme azt a szuper-modult jelképezi, amely a hálózat egészét magába foglalja. A hierarchikus modularizálást az 5. ábrán látható súlyozott példa hálózaton szemléltetem, ahol az élsúlyok a hozzájuk tartozó pontok hasonlóságával arányosak. Az 5A panelen színezéssel jelölt modulszerkezetet az 5B panel dendogramján berajzolt piros szaggatott vonal szerinti vágás jelöli. Ezen az egyszerű példán is jól látható ezeknek a módszereknek egy fontos sajátossága: a kiszámított dendogram alapján számos különböző modulszerkezet meghatározható. Az algoritmusok sokszor a felhasználó döntésére bízzák, hogy melyik szinten állítja meg a modulok összevonását / szétbontását.

A Girvan-Newman eljárás és a modularitás (Q)

A szakirodalomban sokszor GN-ként rövidített módszer (Girvan és Newman, 2002) egy hierarchikus lebontó algoritmus, amely a modulok közötti, hídszerű kapcsolatoknak a felderítésén alapszik. Egy él hídszerűségét a köztiség mérőszám (*betweenness*) segítségével definiálták, amely annál nagyobb, minél több pontpár között húzódó legrövidebb útvonal átmegy az adott élen. Mivel a modulokat sűrűsödési régiókként definiálhatjuk, logikus feltevés hogy a modulokat összekötő éleken sok, a hálózat távoli pontjait összekötő útvonal megy keresztül. Iteratív módon az aktuális hálózatra mindig meghatározva és azután eltávolítva a legnagyobb köztiségű éleket, a hálózat szétesik különálló részekre. Az eljárást

addig folytatva, amíg minden élt el nem távolítunk a hálózathoz, a kapott strukturális információt egy dendrogramként ábrázolhatjuk.

Az eredeti GN módszer (hasonlóan a többi felépítő és lebontó módszerhez) kritikus pontja, a dendrogram alapján kijelölhető lehetséges modularizálások közül az optimális megtalálása. A különböző modulszerkezetek minőségének összehasonlítására Newman bevezette (Newman, 2004b) a modularitás Q -val jelölt mérőszámát, amely az adott modularizálás modulokon belüli éleinek számát hasonlítja össze egy véletlenszerűen felépített azonos méretű hálózat esetén várt modulon belüli élek számával. Pontosabban:

$$Q = \frac{1}{2e} \sum_{ab} \left[S_{ab} - \frac{f_a f_b}{2e} \right] \delta(m_a, m_b) ,$$

ahol S a szomszédsági mátrix (S_{ab} jelöli az a és b pontok között futó élek számát), e a hálózat éleinek száma, és f_a az a pont fokszáma (szomszédos éleinek száma). Az a pont modulját m_a -val és az úgynevezett Kronecker delta függvényt δ -val jelölve, a $\delta(m_a, m_b)$ értéke 1 ha a és b ugyanabba a modulba tartozik, különben 0.

Tegyük fel, hogy készítettünk egy véletlenszerűen összekapcsolt (modulszerkezettel nem rendelkező) hálózatot, amelynek pontszáma, élszáma és fokszámeloszlása megegyezik a vizsgált hálózattal. Ebben a modell hálóban tetszőleges a és b pontok közé eső élek várt száma: $f_a f_b / 2e$. Tehát az $S_{ab} - f_a f_b / 2e$ kifejezés a valós és a modell alapján várt élek számának különbsége, amit a képletben minden modulon belüli pontpárra összegzünk.

A modularitás képlete 0 és 1 közötti valós szám lehet, amely maximális értéke erős modulstruktúrát jelent, míg minimális értéke esetén a hálózat nem rendelkezik szignifikáns modulszerkezettel a nullmodellhez képest.

A lebontó GN eljárás közben figyelve Q értékét, annak maximumhelyei jelentik a hálózat legrelevánsabb modulszerkezeteit. Az algoritmus ilyen módon való használata eredményesnek bizonyult sok valós hálózat elemzése során (Newman, 2004a).

A modularitásra épülő módszerek

A fenti modularitás képlet és a GN módszer hamar népszerű lett a szakirodalomban, és ezek mintájára számos más eljárás került kidolgozásra. Megjelentek a GN módszer gyorsított

változatai, illetve olyan eljárások amelyek különböző informatikai módszereket használva eleve a Q mérőszám maximalizálására törekedve modularizálták a hálózatot (Pujol és mtsai, 2006; Schuetz és Caflisch, 2008; Guimerà és Amaral, 2005; Massen és Doye, 2005; Lehmann és Hansen, 2007; Agarwal és Kempe, 2008). Más tanulmányok a Q mérőszám új definícióit vezették be, eltérő null-modelleket alkalmaztak, azaz más véletlen hálózathoz viszonyították a vizsgált hálózat adott modulszerkezetében az élek számát (Blanchard és mtsai, 2003).

Amennyiben a GN módszerben a híd-jellegű élek azonosítására szolgáló köztiség képletét más hasonló képletekre cseréljük, más modulszerkezetet kapunk, hiszen más sorrendben távolítjuk el a modulok között futó éleket. A véletlen bolyongás köztiség (*random walk betweenness*) egy pontpár mint nyelő és forrás által meghatározott véletlen bolyongás során az adott él forgalmát összegzi fel minden lehetséges pontpárra, míg az áram köztiség (*current-flow betweenness*) a hálózatra mint egy áramkörre gondol és az adott élen átmenő áramot összegzi fel a különböző pontpárookra, mint az áramkör feszültségforrásaira (Newman és Girvan, 2003). Az információs köztiség (*information centrality*) azokra az élekre nagy, amelyek elhagyása a legnagyobb mértékben rontja a pontok közötti információtovábbítás lehetőségét (Fortunato és mtsai, 2004).

Lehetőség van arra is (Tyler és mtsai, 2003), hogy az eredeti köztiséget nem minden pontpár, hanem csak néhány véletlenszerűen kiválasztott pontpár alapján számítsuk. Ilyen módon gyorsítunk a módszeren, illetve többször lefuttatva az eljárást különböző eredményeket kaphatunk, ami alapján meghatározhatjuk az egyes pontokra, hogy melyik modulba mennyire tartoznak. A pontok közötti legrövidebb utakat figyelembe vevő globális köztiség definíciók mellett lehetőség van lokális mérőszámok alkalmazására is. Ilyen például Az él-csomósodási együttható (*edge-clustering coefficient*), amely definíciója (Radicchi és mtsai, 2004):

$$C_{a,b}^{(3)} = \frac{z_{a,b}^{(3)}}{\min[(k_a - 1), (k_b - 1)]} \quad ,$$

ahol $z_{a,b}^{(3)}$ azon háromszögek száma amelyek tartalmazzák az (a,b) élt, míg $\min[(k_a - 1), (k_b - 1)]$ a lehetséges ilyen háromszögek maximális száma. Hosszabb köröket számolva, azaz magasabb rendű csomósodási együtthatókkal ($C_{a,b}^{(g)}$, $g > 3$, azaz nem

háromszögek, hanem négy vagy annál több pontot tartalmazó részhálók vizsgálatával) egyre inkább figyelembe vehető a hálózat globális struktúrája.

Spektrális módszerek

Az összehasonlítható tagokból álló halmazok két részre particionálására szolgáló spektrális módszerek komoly múlttal rendelkeznek az informatika és adatbányászat területein (Kernighan és Lin, 1970; Fiduccia és Mattheyses, 1988). Közös tulajdonságuk, hogy a halmaz (vagy hálózat) pontjainak páronkénti hasonlóságát (azaz a hálózat éleit) leíró mátrixok sajátvektoraiban⁸ az egyes pontokhoz rendelt értékek alapján osztják a pontokat két csoportba. A Laplace mátrix⁹ második nem triviális (azaz legnagyobb nem nulla) sajátértékéhez tartozó sajátvektort előszeretettel alkalmazzák klaszterezésre. Amennyiben a hálózatban például két jól elkülönülő komponens van, akkor ezek a sajátvektor tagjainak előjelei alapján elválaszthatóak egymástól. A hálózat kettévágását ismételve egy lebontó modularizálási módszerhez jutunk (Fiedler, 1973).

6. ábra: Egy négy modult tartalmazó teszt hálózat spektrális modul információinak szemléltetése, egy sajátvektor (a panel) és két sajátvektor (b panel) figyelembevételével. Az ábra forrása: Donetti és Muñoz, 2004

8 A sajátvektorok és sajátértékek definíciójával és számításával kapcsolatban lásd például: http://hu.wikipedia.org/wiki/Saj%C3%A1tvektor_%C3%A9s_saj%C3%A1t%C3%A9rt%C3%A9k

9 A Laplace mátrix a hálózat topológiáját írja le a következő módon: $L_{a,b} = -1$, amennyiben a és b pontok szomszédosak. A főátlóban ($L_{a,a}$) mindig az adott pont fokszáma szerepel. Az összes többi (nem szomszédos) pontpár esetén: $L_{a,b} = 0$.

Lehetőségünk van a vágások ismételtetése helyett több sajátvektort is figyelembe véve egyből kettőnél több modul meghatározására is (Donetti és Muñoz, 2004). A 6. ábrán egy négy modult tartalmazó teszt hálózat pontjai vannak ábrázolva. A pontok koordinátái a Laplace mátrix az első (a panel), illetve az első két (b panel) sajátvektora szerint vannak ábrázolva. A négy színnel jelölt négy eredeti modul jól láthatóan jobban elkülönül két sajátvektor használatával.

A Transzfer mátrixok nem a hálózat szerkezetét írják le, hanem a hálózat felett végzett véletlen bolyongások során várható pontok közötti átmeneti valószínűségeket definiálják. Ezen mátrixok spektrális elemzéséből is értékes következtetéseket lehet levonni a modulok méretére és a modulszerkezetre nézve (Eriksen és mtsai, 2003; Simonsen és mtsai, 2004).

Dinamikai módszerek

A korábban leírt módszerek egyike sem vette figyelembe a hálózat által modellezett valós rendszer tulajdonságait. Az algoritmusok ugyanazokat a modulokat határozták meg függetlenül attól hogy a hálózat pontjai fehérjék, aminosavak, vagy emberek voltak. A dinamikai módszerek közös tulajdonsága, hogy kevésbé általános, de adott esetben pontosabb modulmeghatározást tesznek lehetővé, mert figyelembe veszik a hálózat által reprezentált komplex rendszer időbeli viselkedését és működését is. Természetesen soha nincs lehetőség a komplex rendszer (például sejt, vagy társadalom) teljes dinamikai szimulációjára, ezért általában matematikailag 'egyszerűbb' dinamikai modelleket szokás alkalmazni. Az alábbiakban négy fő elterjedten használt dinamikai megközelítést mutatok példaként.

A spin modellek a statisztikus fizika területéről származnak (Wu, 1982) és az anyagok mágnesezettség változását szimulálva alakítanak ki modulokat a hálózatban. A különböző módszerek a hálózat pontjaihoz több állapotú spineket rendelnek, amely tulajdonságok a gráf élei mentén befolyásolják egymást. A szimulált rendszer minimum energiaszintjén a spin tulajdonságok jellemző módon úgy állnak be, hogy a strukturálisán egymáshoz közel álló pontok azonos spin állapotba kerülnek, amely modulok azonosítását teszi lehetővé (Reichardt és Bornholdt, 2004; Son és mtsai, 2006).

A véletlen bolyongások (Hughes, 1995) a gráfot egy úthálózatként felfogva, az élek

mentén való véletlenszerű 'séták' segítségével definiálnak modulokat. A véletlen bolyongások során bejárt útvonalak segítségével például különböző strukturális hasonlóság mérőszámok definiálhatóak (Zhou, 2003a, 2003b; Zhou és Lipowsky, 2004; Pons és Latapy, 2006) a pontpárok között, amik alapján felépítő vagy lebontó hierarchikus modulkeresési módszerek futtathatóak. Egy másik megközelítés (Delvenne és mtsai, 2010) azt használja ki, hogy egy véletlenszerűen mozgó ágens szignifikánsan több időt tölt a modulokon belül bolyongva, mivel ott (definíció szerint) nagyobb az élek sűrűsége és több a bejárható utak száma, mint a modulok közötti régiókban. Azaz definiálhatjuk egy modul állandóságát az idő függvényében, azáltal hogy meghatározzuk annak a valószínűségét, hogy az ágens adott időn belül elhagyja-e az adott modult. Ezzel lehetőség nyílik az adott modulszerkezet 'stabilitásának' vizsgálatára az idő függvényében. Kimutatható, hogy a korábban megismert Newman-féle modularitás mérőszámra való maximalizálás (Newman, 2004b), valamint a Fiedler-féle spektrális módszer (Fiedler, 1973) is speciális esetei a Delvenne által definiált dinamikai módszernek. Ezen felismerés hatására számos további, modularitásra és dinamikai stabilitásra alapozó módszer született (Arenas és mtsai, 2007; Reichardt és Bornholdt, 2006).

A biológia, fizika vagy a szociológia területén elterjedten vizsgált szinkronizációs folyamatok (Pikovsky és mtsai, 2003) végén jellemző módon az adott komplex rendszer minden tagja azonos állapotba kerül. Az erre építő dinamikai algoritmusok közül az egyik első (Arenas és mtsai, 2006) véletlenszerű kezdő fázissal induló oszcillátorokat helyez el a hálózat pontjaiba. Az oszcillátorok az élek mentén szinkronizálódnak. A szinkronizáció dinamikájából következtetni lehet a modulszerkezetre, mivel először a sűrűbben csatolt modulok kerülnek szinkronba, míg a teljes hálózat szinkronizációjához sokkal több idő szükséges. Hasonló alapelvekre épülve lehetséges lebontó jellegű algoritmus megfogalmazása is (Boccaletti és mtsai, 2007), ahol a rendszer teljes szinkronból indul és fokozatosan esik szét kisebb, egymástól különböző fázisban lévő, belül szinkronizált részekre.

Információ elméleti és egyéb megközelítések

Matematikailag elegáns, de aránylag számításigényes módja a modulok keresésének a hálózat kódelméleti tömörítésén alapuló eljárás (Rosvall és Bergstrom, 2007). Ennek alap gondolata a 7. ábrán látható. A modularizálást úgy fogja fel, mint a hálózat 'egyszerűsített'

leírását, ahol ahelyett hogy minden különálló pontot definiálnánk, a hálózat modulszerkezete segítségével megadjuk az átfogó szerkezetet. A modulszerkezet sokkal kevesebb adattal leírható, mint a hálózat, elegendő ha a modulok méretét (pontok és élek száma), valamint a modulok közötti kapcsolatokat definiáljuk. Viszont a tömörítés veszteséges, azaz a modulszerkezetből többféle hálózat is visszaállítható. Így további kiegészítő információkat kell tárolunk, hogy a kódból visszaállítható hálózatok közül kiválaszthassuk az eredeti hálózatot. A kódolás (és ezzel együtt a modularizálás) annál optimálisabb, minél kisebb a modulszerkezetet és a kiegészítő információkat leíró kód hosszának összege.

7. ábra: A hálózat tömörítési módszer alapgondolata.
Az ábrát közlő eredeti cikk: Rosvall és Bergstrom, 2007

Az úgynevezett címke terjesztési módszerben (Raghavan és mtsai, 2007) minden ponthoz kezdetben egy egyedi címkét rendelünk, például a pont nevét. Ezek után véletlenszerű sorrendben végigmenve a pontokon, minden pont címkéjét beállítjuk arra az értékre, amely a legtöbb szomszédos pontjainál szerepel a címkén. Ha több egyformán nagy számú szomszédos címke van, akkor véletlenszerűen választunk egyet. Ezt ismételve sok címke végleg el fog tűnni, némelyek pedig elterjednek. Addig ismételjük ezt, amíg már nem találunk olyan pontot, amely címkéjét változtatni kellene. Ha a modulokat az azonos címkével rendelkező pontok csoportjaival azonosítjuk, akkor elmondhatjuk, hogy az algoritmus végén minden pontnak több modulon belüli kapcsolata lesz, mint modulon kívüli, ami megegyezik a modulok általánosan elfogadott definíciójával.

Az adatbányászat területén szintén fontos és régen kutatott kérdés a hasonló elemek klaszterekbe való csoportosítása (Jain és mtsai, 1999; Berkhin, 2006). Első ránézésre ez a problémakör nagyon hasonló az itt bemutatott modularizálási feladathoz. Mégis a két terület lényegesen eltér egymástól abban, hogy az adatbányászat esetében jellemzően a vizsgált egyedek tulajdonságairól vannak információink, és ezen tulajdonságok alapján szeretnénk jellemző szegmenseget találni, a hasonló elemeket klaszterekbe sorolni. A hálózatkutatás jellemzően komplex kölcsönható rendszereket vizsgál, ahol a modellben az egyedek tulajdonságai háttérbe szorulnak, viszont az egyedek közötti kapcsolatrendszer előtérbe kerül. A hálózatkutatás területén pont a kölcsönhatások alapján alakítjuk ki a modulokat.

A fejezetben összefoglaltam számos, elterjedten használt modularizálási megközelítést. Látható, hogy az különböző definíciók más és más eredményt adnak, de az hogy melyik módszer alkalmazása a legcélravezetőbb, az egyaránt függ a hálózattal modellezett komplex rendszer dinamikájától, a hálózatépítés során használt eljárások tulajdonságaitól és attól, hogy milyen jelegű kérdéseket kívánunk megvizsgálni a modulszerkezet ismeretében.

2 Átfedő és fuzzy modularizálás

Az előző fejezetben bemutatott modularizálási eljárások főként diszkrét modulokat eredményeztek, azaz az adott pontokhoz egyetlen modult rendeltek hozzá. Átfedő modularizálásról akkor beszélünk, ha az eljárás eredményeképpen kialakuló modulok 'egymásba érhetnek', azaz egy pont több modulba is tartozhat. Az ilyen modulszerkezetek megengedése a legtöbb biológiai rendszer esetén célszerű. Például egyáltalán nem ritka és adott esetben elvárható, hogy egy fehérje több funkcióval is rendelkezzen (egy fehérje hálózat modularizálása után), vagy több jelátviteli útba is beletartozzon (ha egy jelátviteli hálózatot vizsgálunk).

8. ábra: A k -klikk közösségek módszerének szemléltetése egy példa hálózaton (4-es klikk mérettel). Az egyes modulok átfedésében lévő pontokat piros szín jelöli. Forrás: Palla és mtsai, 2005.

Az egyik első, kifejezetten átfedő modulok meghatározására kidolgozott módszer az Eötvös Lóránt Tudományegyetemen kidolgozott és a Nature folyóiratban megjelent k -klikk közösségek módszere¹⁰ (Palla és mtsai, 2005). A k -klikk az a hálózat k elemű részhalmaza, amely teljes gráfot alkot. Azaz 4 tetszőlegesen kiválasztott pont 4-klikk, amennyiben mind a négy pontra igaz, hogy az adott pont a klikk mind a három másik tagjával össze van kötve. Két k -klikk akkor szomszédos, ha legalább $k-1$ pontjuk közös. A k -klikk közösségek által

¹⁰ Angolul: Clique Percolation Method, CPM

definiált modulok pedig úgy jönnek létre, hogy egy adott k -klikket addig bővítünk szomszédos k -klikkekkel, amíg csak lehetséges. A 8. ábra egy példa hálózat 4-klikk közösségeit szemlélteti, kiemelve az eljárás során keletkező modul átfedéseket.

Egy másik átfedő módszer (Baumes és mtsai, 2005) első lépésben a modulok központjait jelöli ki, majd körük lokális modulokat definiál pontok hozzáadása és elvétele által, amelyet addig folytat, amíg az adott modul sűrűségét leíró függvény lokális maximumát el nem éri. Az így keletkezett modulok természetes módon szintén átfedéseket képezhetnek. Az eljárás a modulok közepét különféle centralitás értékek számítása után jelöli ki. Az eljárás egyszerűsített formájában (Lancichinetti és mtsai, 2009) új modul közepe bármely olyan véletlenszerűen választott pont lehet, amely eddig még nem volt egy modulba sem besorolva. Ezen módszer további módosításaként előálló eljárás (Lee és mtsai, 2010) már nem véletlenszerűen kiválasztott különálló pontokat, hanem a legnagyobb elérhető klikkeket használta modulközepeknek, így elődeinél még több átfedést azonosított.

Átfedő modulokat kapunk akkor is, ha a hálózat pontjai helyett első lépésben a hálózat éleit soroljuk be modulokba (Kovács és mtsai, 2010; Ahn és mtsai, 2010; Kim és Jeong, 2011). Ilyen esetekben a pontok modulbatartozása az élek modulbatartozása alapján számítható. A pontok ekkor modulok átfedéseibe kerülhetnek még abban az esetben is ha az éleket eredetileg diszjunkt módszerrel modularizáltuk, hiszen egy pont szomszédos élei különböző modulokba kerülhetnek.

Számos, az előző fejezetben bemutatott, eredetileg diszjunkt modulokat eredményező módszert lehet módosítani (Gregory, 2009b) úgy, hogy modulátfedéseket is azonosítson. A CONGA¹¹ névre keresztelt (Gregory, 2007) eljárás a szétbontó hierarchikus GN algoritmust (Girvan és Newman, 2002) alakította át úgy, hogy a hálózat kettévágása közben bizonyos (a két modulhoz egyaránt erősen kötődő) pontokat duplikálva, azokat mindkét modulba egyszerre behelyezte, ezzel hozva létre átfedéseket. A COPRA¹² algoritmus (Gregory, 2009a) pedig a címke terjesztéses eljárásra (Raghavan és mtsai, 2007) építve határoz meg átfedő modulokat. A módosítás lényege, hogy az algoritmus futása során egy pont számára több modul címke egyidejű használatát engedélyezi.

11 CONGA: Cluster Overlap Newman Girvan-Algorithm

12 COPRA: Community Overlap Propagation Algorithm

Fuzzy algoritmusok

A 'fuzzy' modulszerkezetet eredményező algoritmusok (Gregory, 2011) annyival általánosabbak az átfedő módszereknél, hogy ezek az adott pont adott modulba való tartozásához erősség értékeket is rendelnek. Azaz nem csak arra a kérdésre adnak választ, hogy az adott pont mely modul(ok)ba tartozik, hanem azt is vizsgálják, hogy melyikbe mennyire. A disszertációmban részletesen vizsgált ModuLand eljáráscsalád is ebbe a kategóriába esik. Ezek az algoritmusok a hagyományos diszjunkt vagy átfedő módszereknél sokkal pontosabban térképezik fel a komplex biológiai rendszereket. Igen bő a felhasználási területük mind a biológia, mind az orvostudomány területén, ahogy azt a 6. fejezetben részletesen ismertetem. Az alábbi ábrán (9. ábra) a diszkrét, az átfedő és a fuzzy módszerek különbségét szemléltetem. A diszkrét módszer a modulok átfedésében lévő pontot teljes egészében ahhoz a modulhoz rendeli, amelyikhez a leginkább tartozik. Átfedő módszerek képesek detektálni, hogy a pont több modulba is tartozik, míg a fuzzy módszerek azt is megadják, hogy melyik modulba mennyire.

9. ábra: A diszkrét, az átfedő és a fuzzy modularizálási módszerek által nyújtott információ összehasonlítása.

Nepusz Tamás és munkatársai (Nepusz és mtsai, 2008) olyan átfedő fuzzy modularizálási módszert fejlesztettek ki, amely pontosan meghatározza a pontok modulokba tartozásának mértékét, amennyiben előre megadjuk a modulok számát, illetve a pontok egymáshoz viszonyított hasonlóságát. Azaz a módszer különösen akkor hatékony, ha valamilyen módon (például kísérletes megfigyelésekkel vagy dinamikai szimulációkkal) előre tudomásunk van a hálózat hasonlóan viselkedő, egyszerre aktív elemeiről. A szerzők a

módszert sikeresen alkalmazták például agyi régiók hálózatának modularizálására, ahol az eredményeket felhasználták többek között eddig nem detektált kapcsolatok jóslására.

Az FCM klaszterezés¹³ (Dunn, 1973; Bezdek, 1981) eredetileg az adatbányászat területén került kidolgozásra. Lényegében egy sok dimenziós adathalmaz átfedő csoportosítását oldja meg egy (az adatok távolságán alapuló) célfüggvény iteratív optimalizációja segítségével. Az informatikában gyakran alkalmazzák például mintaillesztési feladatok megoldása során. Az algoritmus átültetésre került (Zhang és mtsai, 2007) a hálózat kutatás területébe is, ahol a hálózat pontjait először egy Euklidészi térbe vetítik ki, majd az FCM célfüggvény és egy (a Newman-féle Q mérőszámot általánosító) átfedő modularitás függvényt kombinálva iteratív optimalizációval határozzák meg az átfedő modulokat.

Egy másik megközelítés (Psorakis és mtsai, 2011) során nem-negatív mátrix faktorizációt végzünk a gráf topológiáját leíró mátrixon (azaz algebrai úton felbontjuk a mátrixot több részmátrixra), úgy optimalizálva az így kialakuló átfedő modulszerkezetet, hogy az minél jobban megmagyarázza a hálózat éleit. Az optimalizáció során a modulszerkezetből valószínűségi függvényeket származtatunk, amelyek megadják hogy két pont mekkora valószínűséggel kerül összekötésre. A modulszerkezetet annak alapján módosítjuk, hogy a hálózat valódi élei minél inkább fedésbe kerüljenek a valószínűségi függvények által jóslott élekkel.

Markovi (más néven emlékezetmentes) folyamatoknak nevezzük azokat a dinamikai folyamatokat, ahol a rendszer következő állapota csak az aktuális állapot függvénye és nem függ attól, hogy korábban milyen állapotokban volt a rendszer. A Markov-láncos modularizálási módszerek alapötlete az, hogy a hálózat felett végzett véletlen bolyongás folyamatát egy diszkrét idejű Markov-lánccal reprezentáljuk, hiszen az hogy a bolyongás során hová lépünk legközelebb, mindig csak attól függ, hogy most mely pontban állunk. Az így nyert Markov-lánc pontjai és élei megegyeznek az eredeti hálózattal, míg az élekhez irányítottan rendelt „állapot átmeneti valószínűség” a hálózat eredeti élsúlyaitól és a pont összes szomszédos élettől függ. A Markov-lánc alapú modularizálási módszer (Weinan és mtsai, 2008) lényege, hogy a markovi állapotokat úgy vonja össze, hogy az új, redukált

13 Fuzzy C-means Clustering

Markov-háló dinamikáját tekintve minél inkább megegyezzen az eredetivel. A Markov-hálózatos modularizálási módszereknek szintén létezik fuzzy változata (Liu, 2010), amely a markovi állapotok összevonásakor az adott állapotot nem teljes egészében egy másik állapottal vonja össze, hanem egybeolvasztja több lehetséges szomszédos állapottal.

3 A ModuLand eljáráscsalád definíciója

Az általunk kifejlesztett ModuLand eljáráscsalád célja, hogy hálózatok pontjait és éleit átfedő módon folytonos erősséggel modulokba sorolja. Az eljáráscsalád alapötlete Kovács István fizikustól származik, aki az eljárások mögött álló legtöbb képletet definiálta. Az eljáráscsalád számítógépes megvalósításában rajtam kívül Palotai Robin, Szuromi Gábor és Zalányi Balázs informatikusok vettek részt a Semmelweis Egyetem LinkGroup hálózatkutató csoportjából (Kovács és mtsai, 2010). A saját munkám során főként az algoritmusok optimalizációjával (Szalay-Bekő, 2010), a felhasználóbarát kezelőfelület megalkotásával és számos biológiai példa elemzésével foglalkoztam (Szalay-Bekő és mtsai, 2012).

3.1 A ModuLand eljáráscsalád főbb lépései

A ModuLand eljáráscsalád (Kovács és mtsai, 2010) képes akár irányított vagy irányítatlan, súlyozott vagy súlyozatlan hálózatok átfedő moduljainak hatékony azonosítására. Az eljáráscsalád egy olyan keretet nyújt, amelybe könnyen beilleszthető sok más, a szakirodalomból ismert eljárás. A ModuLand eljáráscsalád működését az *E. coli* baktériumban található Met-tRNS szintetáz rendszer hálózatos modelljének modularizálásán szemléltetem (10. ábra). Az eljáráscsalád főbb lépései a következők:

- **Indirekt hatás régiók definiálása:** A módszer először a hálózat minden éléhez vagy pontjához definiál egy régiót a hálózatban, amelyre az adott él vagy pont indirekt módon valamekkora hatással van. Például egy aminosav háló esetén ez jelentheti a fehérjének azt a részét, amely valamilyen módon reagál egy adott aminosav elmozdulására. (Nyilván az adott aminosavhoz térben közel eső aminosavak jobban elmozdulnak, míg a hatás egy bizonyos távolság után lecseng a fehérjében.) A ModuLand eljárás során általában a hálózat éleit ért indirekt hatást vizsgáljuk. Az a -val jelölt pont vagy él által keltett hatást általános esetben egy olyan h függvénnyel írhatjuk le, amely a hálózat valamennyi (i,j) éléhez egy valós $h_{(i,j)}^a$ számot rendel amely szám arányos az a forrás által keltett indirekt hatással. A 10. ábrán látható három példa a három aminosav kapcsolat által keltett három indirekt hatás régiót szemlélteti a hálózaton.

- ModuLand centralitás meghatározása:** A következő lépésben a módszer egy centralitás mérőszámot rendel a hálózat valamennyi éléhez úgy, hogy az adott (i,j) élhez tartozó centralitás egyenlő lesz a valamennyi a forrásból indított az (i,j) élt ért indirekt hatások összegével: $c_{(i,j)} = \sum_a h_{(i,j)}^a$. Azaz, egy él annál centrálisabb, minél jobban hat rá a hálózat egésze indirekt módon. Ez a definíció logikusan kiemeli a központi elemeket, hiszen például egy jelátviteli szerepet betöltő fehérje akkor lesz központi a jelátviteli hálózatban, ha minél inkább érzékeny a sejten belüli jelátviteli folyamatokra, ha minél több 'kommunikáció' rajta megy keresztül.
- A Modulok azonosítása:** Ebben a lépésben kerülnek a modulok meghatározásra, amihez először a korábbi lépésben megkapott centralitás értékeket mint képzeletbeli 'magasság értékeket' értelmezzük, így egy centralitás felületet kapunk a hálózat élei felett (10. ábra), amely felület legmagasabb részei lesznek a legcentrálisabb élek. A korábban megfogalmazott (ld: 1.4 fejezet) definíció szerint a modul az egymással a környezeténél erősebben kölcsönható pontok halmaza. Az ilyen pontokra jellemző, hogy több forrás által okozott indirekt kölcsönhatási régióban is közösen szerepelnek, ami miatt a modulok mint hegyek emelkednek ki a centralitás felületen. A modulok közepei lesznek a legcentrálisabb élek, a felület hegycsúcsai. A hálózat éleit és pontjait ezekhez a modulközepekhez soroljuk hozzá átfedő módon, például úgy, mintha a hegycsúcsokból különböző színű festéket folytatnánk le és a festékek keveredéseivel arányosan határoznánk meg az adott él vagy pont modulba tartozásait. A későbbiekben az m_i^A az i pont, míg az $m_{(i,j)}^A$ az (i,j) él A modulba való tartozásának mértékét fogja jelölni.
- Magasabb hálózati szint létrehozása:** Az átfedő modularizálás segítségével olyan metaháló (azaz származtatott hálózat) hozható létre, amelynek pontjai az eredeti hálózat moduljainak felelnek meg, míg metaélei az eredeti modulok közötti átfedéseket jelenítik meg. Ilyen módon egy 'tömörített', egyszerűsített képét kapjuk meg a hálózatnak, amelyen további szerkezeti elemzéseket hajthatunk végre.

10. ábra: A ModuLand eljárás fogalmainak szemléltetése a Met-tRNS szintetáz fehérje aminosav hálóján

A ModuLand eljáráscsalád mindegyik lépése többféleképpen is megvalósítható. Sőt, számos esetben előnyös az adott komplex rendszer dinamikai tulajdonságainak ismeretében definiálni az egyes lépéseket. Például az indirekt hatásokat kiszámító lépésben igen erősen különbözhet egy fehérje expressziójának gátlására bekövetkező változás terjedése a sejtben (egy fehérje hálózat esetén) és egy pletyka elterjedése (egy szociális hálózatban). Az eljáráscsalád nagy erőssége mégis az, hogy ezeket a komplex problémákat nem kell minden bonyolult rendszer esetében külön megoldani, hanem már általános közelítő megoldásokkal is kielégítően jó eredményeket lehet elérni.

A következő fejezetekben az eljáráscsaládnak azokat az algoritmusait fogom definiálni, amelyeket felhasználtam az általam megvalósított grafikus modularizáló program (Szalay-Bekő és mtsai, 2012) fejlesztése során. Emiatt irányított hálózatok modularizálására ebben a fejezetben nem térek ki.

3.2 A LinkLand centralitás felület számítása

A LinkLand eljárás egy gyors, közelítő megoldás egy él indirekt kölcsönhatásának a feltérképezésére. Az eljárás kezdetén az adott él csupán a két végpontjára van hatással. Ezután újabb és újabb szomszédos pontokkal bővítjük az indirekt módon befolyásolt pontok halmazát a következő kritériumok alapján:

- a hatás (irányát tekintve) mindig a sűrűbb régiók felé terjed a hálózatban, illetve
- a hatás mindaddig tovább terjed, amíg a hatás alatt álló pontok által meghatározott részgráf sűrűsége nem csökken.

11. ábra: Példák a LinkLand algoritmus által definiált indirekt hatás terjedésére.

Az eljárást legjobban úgy lehet elképzelni, mintha egy elit klubbot kívánnánk tagokkal bővíteni. Kezdetben a klub két emberből áll, akik jól ismerik egymást. Ezután (majd később is

Az algoritmus alapvetően súlyozott hálózatokra van definiálva: $s_{(i,j)} < 0$. A súlyozatlan hálózatok esetén minden élsúlyt azonos értékűnek veszünk, pl: $s_{(i,j)} := 1$. Az adott részgráf sűrűségét a részgráfban szereplő élek összsúlyának és a benne lévő pontok darabszámának hányadosaként definiáljuk.

Az adott (a,b) élből induló, az (i,j) élt ért indirekt hatás értéke megegyezik az él kezdeti súlyával: $h_{(i,j)}^{(a,b)} = s_{(i,j)}$, amennyiben a hatás elérte az (i,j) élt, különben pedig nulla.

Az algoritmust minden kezdő élre lefuttatva megkapjuk a ModuLand centralitás értékeket, amennyiben az élekre összegezzük az összes forrás élből őket ért hatásokat:

$$c_{(i,j)} = \sum_{(a,b)} h_{(i,j)}^{(a,b)} .$$

Ezeket a centralitás értékeket a 12. ábrán mutatom be. A két modul

középső régiója lényegesen magasabb centralitás értéket kapott a modulok szélén lévő régiók éleinél.

A LinkLand algoritmust a ModuLand eljáráscsaládot bemutató publikációnkban annak pszeudo kódjával pontosan definiáltuk (Kovács és mtsai, 2010).

3.3 Az Arányos modul besoroló algoritmus

A ModuLand centralitások meghatározása után következik a centralitási felület hegyeinek modulként való azonosítása. Ehhez először az él felületen megtalálható lokális hegycsúcsokat kell megkeresni, amelyek a modulok közepét fogják adni. A ModuLand eljáráscsaládban a modulok közepét tehát élek definiálják, noha lehetséges a legcentrálisabb pontok meghatározása is a modulokban. A 12. ábrán látható centralitás felületen például a (G,I) él egy lokális hegycsúcs, hiszen ehhez az élhez nem csatlakozik nála magasabb él. A másik lokális hegycsúcsot a B , C és D pontok közötti három él definiálja, amelyek mivel egyenlő magasságúak, ezért együtt alkotnak egy hegycsúcsot.

Az arányos besoroló algoritmus a hálózat éleit rendeli modulokba, a nála nagyobb modul centralitású (a centralitás felületen magasabban elhelyezkedő) szomszédos élek modulbatartozásai alapján. Jelölje $m_{(i,j)}^A$ az (i,j) él A modulba való tartozásának mértékét. Az algoritmus gyakorlatilag az adott él centralitását 'szétosztja' a modulok között. Ha az adott (x,y) él az A modul közepe, úgy ő teljes centralitásával az A modulba fog tartozni, azaz $m_{(x,y)}^A = c_{(i,j)}$. Az összes többi esetben, amikor az (x,y) él nem közepe egyik modulnak sem, akkor pedig az él különböző modulokhoz való tartozásának összege együtt adja ki az él modul centralitását, azaz: $\sum_A m_{(x,y)}^A = c_{(i,j)}$.

Az élek modulba sorolásánál a magasabb szomszédok modulba tartozása alapján dől el az alacsonyabb élek besorolása, mégpedig arányos mértékben a szomszédok modul centralitás értékeivel.

Az alábbi képlet adja meg, hogy az adott nem modulközép (i,j) él mennyire fog az A modulba tartozni.

$$m_{(i,j)}^A = c_{(i,j)} \frac{\sum_{k, k \neq j} m_{(i,k)}^A + \sum_{l, l \neq i} m_{(l,j)}^A}{\sum_{k, k \neq j} c_{(i,k)} + \sum_{l, l \neq i} c_{(l,j)}}, \text{ ahol } k \text{ és } l \text{ indexekre teljesül a}$$

$$c_{(i,j)} < c_{(i,k)} \text{ és } c_{(i,j)} < c_{(l,j)} \text{ feltétel.}$$

Az 13. ábra szemlélteti a besorolási képlet működését egy példa hálózat részletein. Amikor az algoritmus az (A,B) él besorolását számítja ki, akkor valamennyi nála nagyobb centralitású szomszédos él besorolását figyelembe veszi. Az (A,E) és a (B,G) élek nem magasabbak, ezért az ő modulokhoz való hozzárendelésük nem számít. Az (A,B) él az összes többi él modulba tartozásának mértéke és az adott élek centralitás értékének függvényében a fent leírt képlet alapján kerül besorolásra.

Az **A-B** él modulba tartozási vektorának kiszámítása a szomszédos élek alapján.
Élek modul centralitása: **piros**
Élek modulba tartozási vektora: **[1-es modul, 2-es modul]**

13. ábra: Az arányos besoroló működésének szemléltetése egy példa hálózat részletén.

A besoroló képletek alapján az összes él modulba tartozása megadható egyetlen nagy

lineáris egyenletrendszer megoldásával, azonban ennek a futásideje $O(n^4)$. Az eljáráscsalád megvalósításában egy másik, ennél optimalizáltabb eljárást fejlesztettem ki, amelynek a futásideje $O(n^3)$ és ugyanazt az eredményt adja mint a fenti képlet lineáris egyenletrendszerben való alkalmazása. Az optimalizált algoritmust, amely egy több kezdőpontból indított párhuzamos gráf bejáráshoz hasonlít, a ModuLand eljáráscsaládot bemutató publikációnkban definiáltuk (Kovács és mtsai, 2010).

14. ábra: Az arányos besoroló algoritmus eredménye a példa hálózat élein.

A pontok modulba tartozásait a velük szomszédos élek modulba tartozásai alapján számíthatjuk ki, egyszerűen összegezve az élek besorolásait: $m_{(i)}^A = \sum_j m_{(i,j)}^A$.

15. ábra: Az arányos besoroló algoritmus eredménye a példa hálózat pontjain.

A 14. és 15. ábra a korábban is használt példa hálózaton mutatja meg a LinkLand modul centralitás számláló és az arányos besoroló algoritmusok használata után keletkezett él és pont besorolásokat. A két modul jól láthatólag elkülönül egymástól, míg a közöttük elhelyezkedő pontok és élek átfedő módon mind a két modulhoz egyaránt tartoznak, különböző erősséggel. A pontok és élek színezése a ModuLand Cytoscape plug-in segítségével történt, a színek arányosak a modulokba való beletartozásokkal.

3.4 Magasabb hierarchikus szintek számítása

Nagy előnye az eljáráscsaládnak, hogy a részletes átfedések ismeretében kézenfekvő módon előállítható a hálózat egy 'magasabb szintjét' reprezentáló meta hálózat, amelynek pontjai az eredeti hálózat moduljai, és a meta pontok között húzódó kapcsolatokat a hálózat moduljai közötti átfedések definiálják. A magasabb szintű hálózaton a modularizálást újból el lehet végezni és így egyre magasabb és magasabb szintű hálózatokat kapunk. A magasabb szintű modulokat visszavetítve az eredeti hálózatra, láthatjuk hogyan szerveződnek egyre nagyobb egységekbe az eredeti hálózat pontjai. A 16. ábrán a Met-tRNS szintetáz fehérje térszerkezeti hálózatán mutatom be a három hierarchikus szintet, amelyet a ModuLand Cytoscape plug-in (Szalay-Bekő és mtsai, 2012) segítségével határoztam meg, és ábrázoltam.

16. ábra: A Met-tRNS szintetáz fehérje szerkezeti hálózat modul hierarchiája.

A magasabb szintű hálózatok az alacsonyabb szintek modulszerkezetét reprezentálják és megőrzik a hálózat szerkezetének alapvető tulajdonságait, persze egyre kisebb és kisebb részletességgel. Ilyen módon az eredeti hálózat veszteségesen tömörített változataiként is felfoghatóak, mintha a hálózatot egyre messzebből vizsgálnánk.

Irányítatlan esetben a magasabb szint meta éleinek súlyát (azaz a két modul közötti átfedés mértékét) az alábbi képlettel definiáljuk:

$$s'_{(A,B)} = \sum_i \frac{2 m_i^A m_i^B}{c_i}$$

A magasabb szinten megjelenő élsúlyt ($s'_{(A,B)}$) jelölésben egy aposztróf segítségével különböztetem meg az eredeti hálózatban használt élsúly jelöléstől ($s_{(A,B)}$). A képlet szerint minden egyes pontra megnézzük hogy mennyire jelent átfedést az adott két modulra nézve, és ezen átfedés értékeket összegezzük. A pont centralitásával való osztás oka, hogy a képletben szereplő $m_i^A m_i^B$ kifejezés négyzetesen skálázódik az adott pont modul centralitásával. A kettes faktornak nincs gyakorlati jelentősége, a bevezetése csupán az irányított hálózatokra való áttérést egyszerűsíti le. A magasabb szintű hálózat tehát nullánál nagyobb élsúlyokkal fog rendelkezni. Definíció szerint a meta él nem jelenik meg a magasabb szintű hálózatban, amennyiben $s'_{(A,B)} = 0$.

Ha a korábbi fejezetekben definiált LinkLand kupacképző és arányos modulbesoroló algoritmusokat irányítatlan hálózatokon használjuk, akkor a magasabb szintekben sosem növekszik sem a pontok sem a komponensek száma, és végül minden egybefüggő komponens egyetlen ponttá zsugorodik. Ezen állítás belátása egyszerű, amennyiben figyelembe vesszük hogy a modulok száma mindig kisebb a pontok számánál. Ennek oka, hogy az élfelületen megjelenő lokális maximumok száma korlátos a pontok számával, hiszen egy pont szomszédságában nem lehet két lokális maximum.

A magasabb szintek visszavetítése triviálisan kiszámítható a különböző szinteken kiszámított modulbesorolást tartalmazó mátrixok szorzataként (Kovács és mtsai, 2010).

3.5 Modularizálás alapú mérőszámok

Az átfedő modulok szerkezetének ismeretében számos hasznos mérőszámot definiálhatunk, amelyek a hálózat pontjait, moduljait, vagy az egész modulszerkezetét jellemzik. Az alábbiakban azokat a mérőszámokat definiáljuk, amelyek az általam implementált ModuLand Cytoscape plug-in (Szalay-Bekő és mtsai, 2012) segítségével könnyen meghatározhatóak.

ModuLand Centralitás

A LinkLand algoritmus által megadott modul centralitás értékek önmagukban is fontos és használható mérőszámok, hiszen megadják, hogy egy adott pont mennyire helyezkedik el központi helyen a hálózat egészét tekintve.

Lényegi darabszámok

Egy számhalmaz *lényegi darabszáma* a halmazban lévő szignifikáns elemek számáról ad információt. A lényegi darabszám képlete a valószínűségi változók *hozzájárulásának lényegi mértékén*¹⁴ alapul (Grendar, 2006). A 17. ábrán látható példában három darab, egyaránt három elemű halmaz lényegi darabszáma látható. Az érték egyhez közeli, amennyiben a legnagyobb elem nagyságrendekkel megelőzi a másik kettőt. A hármas számot pedig csak akkor veszi fel, amikor mind a három elem egyforma.

17. ábra: Lényegi darabszám jelentésének illusztrálása.

A lényegi darabszám látható módon nem diszkrét módon számolja meg a halmaz elemeit, hanem folytonosan, nagyobb súlyt adva a relatív módon lényegesebb elemeknek. A V

¹⁴ Angolul: effective size of support

halmaz lényegi darabszámát az $n(V) = \exp(-\sum_i p_i \log p_i)$ képlettel definiáljuk, ahol p_i

az i -edik tag relatív súlya, azaz $p_i = \frac{V_i}{\sum_j V_j}$.

A lényegi darabszám képletét számos mérőszám esetén alkalmazhatjuk. Attól függően, hogy mi az a V halmaz aminek a tagjait megszámloljuk, megadhatjuk például:

- az adott a pont átfedés mérőszámát, amely azt fejezi ki hogy az adott pont hány modulba tartozik szignifikánsan: $V_i = m_a^i$
- a modulok lényegi darabszámát, amely kifejezi hogy nagyságrendileg hány szignifikáns méretű modult tartalmaz a hálózat: $V_i = \sum_j m_j^i$
- az A meta pont éleinek lényegi darabszámát, ami megadja hogy az adott meta pont által reprezentált A modul hány másik modullal fed át szignifikánsan: $V_i = s'_{(i,A)}$

Hídság mérőszám

A híd fogalom alatt a gráfelméletben hagyományosan olyan élt értünk, melynek törlése növelné a gráf komponenseinek számát. A szociológiában hídelem alatt olyan pontot értenek, amely két, egyébként elkülönülő csoportot köt össze. Bevezethetjük a pont vagy él hídságának mértékét, ami definíciónk szerint azt fejezi ki, hogy az adott pont vagy él modulba tartozása alapján milyen mértékben átfedő adott két vagy több modul között, relatíve a többi ponthoz vagy élhez képest.

Az i pont hídsága A és B modul között: $B_i^{A,B} = \frac{T_i^{A,B}}{\sum_j T_j^{A,B}}$, ahol $T_i^{A,B}$ az i pont

'közös területe', azaz $T_i^{A,B} = \min(m_i^A, m_i^B)$. Az i pont összesített hídsága ez alapján pedig kifejezi, hogy mennyire tölt be híd funkciót az adott pont az összes modulpár között:

$$B_i = \sum_{a=1}^m \sum_{b=1, b \neq a}^m B_i^{A,B}.$$

A híd mérőszám tehát 0, vagy annál nagyobb pozitív szám, amely akkor magas, ha a pont nagy átfedést jelent sok modulpár között.

4 A ModuLand eljáráscsalád megvalósítása a Cytoscape programba ágyazva

A Cytoscape program (Shannon és mtsai, 2003) egy nyílt forráskódú szoftver, amely célja hogy egyetlen keretbe integrálja a biológiában vizsgált molekuláris interakciós hálózatokat, az expressziós és egyéb biológiai adatbázisokkal¹⁵. Az elsődlegesen a biológiai hálózatok és molekuláris adatbázisok vizualizálására készült keretprogramba szabványos interfészekon keresztül tetszőleges elemző és megjelenítő modul beépíthető. A felhasználó miután feltelepíti a Cytoscape programot, plug-in formájában telepítheti a más fejlesztők által készített program modulokat, amelyek között szerepel például számos modularizáló eljárás is.

18. ábra: A ModuLand plug-in működése a Met-tRNS szintetáz protein aminosav hálózatának modularizálása közben.

A ModuLand eljárás Cytoscape programba épülő változatának¹⁶ elkészítése a doktori témám egyik legfőbb eredménye, amelyet munkatársaimmal közösen a Bioinformatics

¹⁵ A Cytoscape program letölthető: <http://www.cytoscape.org/download.php>

¹⁶ A ModuLand Cytoscape plug-in letölthető: <http://www.linkgroup.hu/modules.php>

folyóiratban közzétünk (Szalay-Bekő és mtsai, 2012). Ebben a fejezetben a ModuLand plug-in működését és az általa nyújtott szolgáltatásokat ismertetem. A 18. ábra a ModuLand plug-in működését demonstrálja a Met-tRNS szintetáz fehérje aminosav hálózatának modularizálása közben.

4.1 A ModuLand plug-in felépítése

A ModuLand Cytoscape plug-in egy alapvetően Java környezetben futó program modul, amely beépül a Cytoscape programba. A szoftver felépítése a 19. ábrán látható.

19. ábra: A ModuLand Cytoscape plug-in felépítése (a lényegesebb belső komponensek).

A plug-in alapvetően három fő részből tevődik össze, de a felhasználó számára egyetlen jar¹⁷ fájlként letölthető.

Bináris modularizáló programok

A modularizálás mögötti képleteket és algoritmusokat a Java nyelvénél gyorsabb futást eredményező C/C++ nyelveken valósítottam meg. A modularizálásért felelős programok

¹⁷ jar: Java Archive – a Java nyelven írt programok tárolása számára szabványosított fájlformátum

maguk is moduláris felépítésűek, abban az értelemben hogy az eljárás család minden tagja külön bináris programként került implementálásra. Az egyes bináris programok a részeredményeiket fájlokba mentik, amelyeket részben a többi modularizáló program használ további számításokhoz, részben pedig a ModuLand plug-in használja fel az eredmények vizualizálására és a modularizálással kapcsolatos jelentések generálására.

Mindegyik bináris programot három különböző operációs rendszerre lefordítottam, így a plug-in egyaránt működik Microsoft Windows, Linux és Mac OS felett. Mindegyik program a plug-in-t tartalmazó telepítő fájlba van csomagolva, így nem kell különböző plug-in verziókat letölteni a különböző operációs rendszerek esetén.

Plug-in logika

A plug-in alkalmazáslogikája Java nyelven íródott. A fő plug-in osztály neve¹⁸ a plug-in telepítések eltárolására kerül a Cytoscape-ben, így azután a Cytoscape minden indításakor betöltődik. Ez az osztály regisztrálja az új funkcióknak megfelelő menüpontokat az eszköztáron és a menüben.

Amikor a felhasználó a Cytoscape program indulása óta először használja a ModuLand plug-in-t, a plug-in kicsomagolja az operációs rendszernek megfelelő bináris modularizáló programokat egy átmeneti könyvtárba. A Cytoscape leállása után ezek a bináris programok automatikusan törlődnek.

A plug-in képes több modularizációs projektet is futtatni párhuzamosan. Egy modularizálási projekt egy alap hálózatot és az arra épülő magasabb szintű modul metahálózatokat tartalmazza. A különböző modul hierarchikus szintek (ld. 18. ábra) automatikusan, külön hálózatként töltődnek be a Cytoscape programba.

Minden projekt indulásakor a felhasználónak meg kell adnia egy könyvtárat, ahová a plug-in a modularizációval kapcsolatos adatokat tárolja. Ilyen adatok például a különböző hierarchia szinten megjelenő hálózatok, modul centralitás értékek, modulbatartozási mártixok, mérőszám fájlok és projekt leírók. A Cytoscape program újraindítása után lehetőség van korábbi modularizálási projektek betöltésére. Ilyenkor a plug-in megvizsgálja a projekt

¹⁸ A fő osztály neve: `hu.linkgroup.moduland.cytoscape.ModuLandPlugin`

könyvtár tartalmát, és az alapján rekonstruálja a korábbi projekt során korábban meghatározott hálózatokat, mérőszámokat és egyéb adatokat.

Felhasználói felület

A ModuLand plug-in tervezése és megvalósítása közben lényeges szempont volt a könnyen kezelhető felület kialakítása. A legtöbb funkciót a bináris programok futtatásával is el lehet érni, de kétség kívül könnyebb és gyorsabb felhasználást tesz lehetővé a grafikus felület.

Két fő dialógus ablak került bevezetésre. Az első (20. ábra) arra szolgál, hogy a felhasználó egy adott modularizálás elkezdése előtt kiválaszthassa az élsúlyként alkalmazni kívánt Cytoscape attribútumot. Erre azért van szükség, mert a Cytoscape-ben tetszőleges számú tulajdonság tárolható minden egyes élhez, így könnyen előfordulhat hogy egy adott biológiai hálózathoz többféle tulajdonság is definiálva van (például egy aminosav hálózathoz az élekhez tárolva lehet a két aminosav közötti távolság, a kötés erőssége, vagy a kötés típusa egyaránt). Amennyiben nincs minden él esetén definiálva a kívánt élsúly, a ModuLand plug-in lehetőséget ad alapértelmezett élsúlyok megadására is.

20. ábra: Az élsúlyok kiválasztására szolgáló dialógus ablak.

A második fő dialógus ablak szolgál a modularizálás vezérlésére, különböző modul szintek létrehozására és törlésére, a modularizálással kapcsolatos jelentések és mérőszámok generálására valamint a Cytoscape program utasítására a hálózatok adott (modulba tartozás vagy mérőszám alapján történő) színezésre. Az ablak felépítését a 21. ábrán mutatom be. Az ablak tetején egy összefoglaló táblázat található, amely tartalmazza az eredeti hálózat (legfelső sor) és az egyes meta-hálózatok (további sorok) legfontosabb tulajdonságait: például pontok, élek vagy modulok számát. A lista adott sorára kattintva kiválaszthatjuk hogy melyik hálózaton kívánunk műveletet végezni. A lista alatt négy fül található, amelyek a főbb funkciókat csoportosítják.

21. ábra: A modularizálást irányító fő dialógus ablak.

A felhasználói felület több pontján a felhasználónak lehetősége van ('help' feliratú gombokkal) az adott funkció leírásának megtekintésére. Ezen kívül létrehoztam egy online

felületet, amely az esetleges javaslatokat, észrevételeket és hibaesetek bejelentését teszi lehetővé. Ez a felület a fő dialógus ablakból egy gombnyomásra elérhető.

4.2 A ModuLand plug-in által nyújtott szolgáltatások

A ModuLand plug-in számos hasznos szolgáltatást nyújt a felhasználók számára a modulszerkezet értelmezéséhez. Az alábbiakban a három leglényegesebb szolgáltatást ismertetem.

Összefoglaló jelentések és adatok exportálása

A plug-in lehetőséget ad valamennyi modularizálás közben előállt adat rendezett exportálására. A modulszerkezet bináris fájlokban tárolódik, de a felhasználó szöveges állományokba mentheti bármely modulhierarchia szint információit. Minden modulba tartozási adat elérhető az adott hierarchia szinten és az eredeti hálózatra visszavetített formában (16. ábra). Például ha egy magasabb szintű modulszerkezetet választ ki a felhasználó, akkor előbbi esetben a meta hálózat pontjainak, utóbbi esetben az eredeti hálózat pontjainak besorolását menti ki a program. Mind a két esetben lehetőség van áttekintő és részletesebb kimenet exportálására. Legegyszerűbb esetben a program minden ponthoz azt a modult listázza, amelyikbe a pont leginkább tartozik. Ennél több információt nyújt, de még mindig kényelmesen átlátható a 22. ábra A paneljén bemutatott formátum, ahol minden pontnál a legjellemzőbb 10 (vagy annál kevesebb) modult listázza ki a program, a modulba tartozási értékek szerint csökkenő sorrendben. Ez a formátum kimondottan alkalmas arra hogy az adott pont (a példában adott aminosav) moduláris tulajdonságairól áttekintő képet kapjon a felhasználó. Az emberi szem számára legkevésbé értelmezhető, de legtöbb információt tartalmazó formátum egy táblázat (a modul besorolási mátrix), amely táblázatnak sorai az egyes modulokat reprezentálják és az adott sor adott oszlopában az adott pontnak adott modulba való tartozásának erősség értéke található (22. ábra, B panel). Ez a formátum elsősorban további programok vagy részletesebb kézi elemzések számára ajánlott.

A

	A	B
1	the 10 modules where the node mostly belongs	Node id: ALA-4
2	moduleid	module assignment value
3	HIS-28	10.482884
4	TYR-490	0.338982
5	TRP-204	0.051856
6	ASN-382	0.010028
7		
8	the 10 modules where the node mostly belongs	Node id: LYS-5
9	moduleid	module assignment value
10	HIS-28	17.749777
11	TYR-490	0.197307
12	TRP-204	0.120052
13	ASN-382	0.005784
14		

B

```

Programmer's Notepad - [E-coli-Met-tRNA-synthase-protein-structure-network-1_moduleBe
File Edit Search View Tools Window Help
Plain Text
E-coli-Met-tRNA-synthase-protein-structure-network-1_moduleBelongs_proj.txt
nodes of the original network: ,ALA-4,LYS-5,LYS-6,ILE-7,LEU-8,VAL-9,THR-
node belong values in module HIS-28: ,10.482884,17.749777,21.216524,35.3
node belong values in module TRP-204: ,0.051856,0.120052,0.129472,0.2577
node belong values in module TYR-490: ,0.338982,0.197307,0.043930,0.1660
node belong values in module ASN-382: ,0.010028,0.005784,0.001245,0.0047
node belong values in module LYS-142: ,0.000000,0.000000,0.000000,0.0000

```

22. ábra: Példák modulszerkezeti adatok exportálására. A pontok modulbatartozásáról adott áttekintő (A panel) és részletes (B panel) információk.

A modulszerkezet adatai mellett lehetőség van a hálózat pontjaihoz tartozó különböző mérőszámok kiszámítására és ezen adatok exportálására is. A ModuLand plug-in az alábbi mérőszámok meghatározásával segíti a hálózatok szerkezeti elemzéseit:

- pontok éleinek élsúly által súlyozott lényegi darabszáma (lényegi foksám érték)
- a modulok meghatározása során is használt modul centralitás értékek
- a pontok köztiség centralitás értéke¹⁹
- a pontok átfedés értékei
- a pontok hídság értékei

¹⁹ Betweenness Centrality: a szociológia területén bevezetésre került jól ismert mérőszám, amelyet elterjedten használnak a hálózatok kutatásban (Freeman, 1977). Lásd még: 1.4. fejezet, Girvan-Newman módszer.

A köztiség mérőszám kivételével valamennyi mérőszámot a ModuLand eljárás család kapcsán került definiálásra a 3.5. fejezetben. A 23. ábrán különböző mérőszámokat tartalmazó exportált dokumentumokra mutatok példát. A példán a Met-tRNS szintetáz fehérje aminosav hálózatának moduljaiból álló első metahálózatán kiszámolt mérőszámokat mutatom be. Az első három mérőszám a 49 meta ponthoz (azaz az eredeti hálózat aminosav moduljaihoz) rendel értékeket, míg a másik két mérőszám az eredeti hálózat 547 pontjához (a fehérje egyes aminosavaihoz) tartozik.

Mérőszámok a modulok metahálózatán

	A	B	C	D
1	nodeID	effective degree	ModuLand community centrality	betweenness centrality
2	TYR-490	11.109523	111810.3	0
3	TYR-260	13.84704	343673.1	19.014596
4	LEU-473	6.772878	28723.26	0
5	HIS-28	11.013364	191996.3	150.999807
6	LEU-447	8.107162	52083.33	25.472464
7	PHE-505	14.347497	37409.7	24.143319
8	LEU-113	11.465927	138411.4	48.6538
9	TRP-204	8.23693	103052	127.482324
10	ALA-459	3.584986	1655.871	0
47	SER-178	3.048574	199.5889	0
48	LEU-210	4.654849	677.1175	1.981793
49	GLN-78	5.885966	568.5015	0
50	THR-511	5.923211	249.7736	0
51				

Az eredeti hálózatra visszavetített mérőszámok

	A	B	C
1	nodeID	ModuLand overlap	ModuLand bridgeness
2	ALA-4	1.192564	0.000086
3	LYS-5	1.108403	0.002206
4	LYS-6	1.053712	0.00127
5	ILE-7	1.076355	0.008426
6	LEU-8	1.047959	0.00239
7	VAL-9	1.133096	0.00607
8	THR-10	1.127601	0.001333
9	CYS-11	1.172533	0.004857
10	ALA-12	1.198701	0.005899
545	LYS-547	2.58736	0.004042
546	GLU-548	2.584141	0.003815
547	GLU-549	2.190928	0.001626
548	VAL-550	2.459823	0.000168
549			

23. ábra: Mérőszámok exportálása.

Általánosságban az első három mérőszám különböző centralitás értékeket rendel a felhasználó által kért hierarchia szinten megjelenő hálózat pontjaihoz, azaz vagy az eredeti hálózat pontjaihoz, vagy a metahálózatok (az eredeti hálózat moduljait reprezentáló) pontjaihoz. A centralitás értékek az adott pont központi szerepét határozzák meg. A lényegi fokszám egy lokális centralitás érték, amely annál magasabb, minél több lokálisan erősnek számító kapcsolattal rendelkezik egy adott pont. A köztiség mérőszám (Girvan és Newman, 2002) egy globális centralitás mérőszám, amely annál nagyobb, minél több pontpár között húzódó legrövidebb útvonal megy át az adott ponton. A modul centralitás érték sem a pont közvetlen környezete alapján (mint a lokális centralitás értékek), sem a teljes hálózat figyelembe vétele alapján (mint a globális centralitás értékek) súlyozott centralitás, hanem a

pontra hatással lévő lokális környezete alapján meghatározott érték. A három centralitás érték egyaránt nagyon jól használható biológiai hálózatok elemzéseiben (ld. 6. fejezet), de más jellegű információt hordoz.

A pontok átfedés és hídság mérőszámai mindig a kiválasztott hierarchia szinten megjelenő visszavetített modulszerkezet alapján kerülnek kiszámításra, így ezen értékek mindig az eredeti hálózat pontjaira vonatkoznak. Így összehasonlítható például, hogy az adott pont inkább a kisebb (alacsonyabb hierarchia szinten megjelenő) helyi modulok között játszik fontos híd szerepet, vagy a nagyobb (magasabb szintű) közösségek között közvetít.

A plug-in igény esetén átfogó jelentést készít az adott szinten megjelenő modulszerkezetéről, amely igen hasznos az adott modulszerkezet értelmezéséhez. Az átfogó jelentés (24. ábra) által tartalmazott legfontosabb információk:

- pontok és élek száma
- modulok száma és lényegi darabszáma (az utóbbihoz ld: 3.5. fejezet)
- modulok méretei (a modulokba tartozó pontok darabszáma, lényegi darabszáma, illetve a modul pontjai modulbatartozási értékeinek összege)
- modulok központi pontjai (a legkiemelkedőbb pont, illetve a legkiemelkedőbb 10 pont listája).

Az átfogó jelentés egyaránt elkészíthető az adott hierarchia szintre, illetve az adott szinten megjelenő modulok visszavetített szerkezetére.

A különböző jelentéseket és adatokat tartalmazó fájlok helytakarékosági okokból csak a felhasználó kérésére generálódnak. Valamennyi exportált dokumentum a projekt könyvtárába kerül lementésre és megnyitható a plug-in felületén keresztül egyetlen kattintással. A legtöbb jelentés és adatfájl CSV formátumban tárolódik, amit valamennyi elterjedten használt táblázatkezelő program értelmezni tud. A nagyobb (alkalomadtán több száz megabájtos) részletes modulba tartozási mátrixok szöveges TXT fájlokba kerülnek mentésre, amelyek elsősorban nem emberi, hanem gépi feldolgozás céljából érhetőek el.

	A	B	C	D	E	F	G	H	I	J
1	project:			E-coli-Met-tRNA-synthase-protein-structure-network						
2	level:			1						
3	nodes on this level:			49						
4	links on this level:			490						
5	nodes on the original network:			547						
6	links on the original network:			2153						
7	are the modules projected to the original network:			YES						
8	number of modules:			5						
9	effective number of modules:			3.266426932						
10	module merge correlation threshold:			1						
11	original number of modules before merge:			5						
12										
13	Detailed module info:									
14	moduleId	number of nodes	eff. number of nodes	module size (sum of module assignment values)	maximum module assignment value	module center nodes (nodes with maximum assignment value)				
15	HIS-28	547	378.02792	8598.344	51.718563	TYR-260				
16	TRP-204	547	103.0783	760.66644	25.86719	TRP-221				
17	TYR-490	540	178.74016	3506.9573	54.27747	TYR-490				
18	ASN-382	466	75.772125	815.1889	33.32877	PHE-377				
19	LYS-142	242	65.36767	1339.2533	34.267094	CYM-145				
20										
21	the 10 core nodes (decreasingly ordered by the module assignment values) of the module HIS-28									
22	TYR-260	HIS-28	TRP-281	ASN-102	ARG-356	PHE-84	PHE-264	GLN-30	TRP-34	TYR-280
23	the 10 core nodes (decreasingly ordered by the module assignment values) of the module TRP-204									
24	TRP-221	TRP-204	LEU-201	ASN-216	MET-218	GLU-220	GLN-202	PHE-197	PHE-222	SER-198
25	the 10 core nodes (decreasingly ordered by the module assignment values) of the module TYR-490									
26	TYR-490	TYR-418	MET-479	ILE-400	ARG-485	GLN-474	LEU-473	PHE-484	CYS-477	SER-478
27	the 10 core nodes (decreasingly ordered by the module assignment values) of the module ASN-382									
28	PHE-377	ASN-382	VAL-386	ASP-384	ILE-385	ASN-387	VAL-381	VAL-378	VAL-390	ALA-383
29	the 10 core nodes (decreasingly ordered by the module assignment values) of the module LYS-142									
30	CYM-145	TYR-128	TYR-165	LEU-170	GLN-126	SER-175	VAL-141	LYS-142	PRO-167	GLN-153
31										

24. ábra: Átfogó jelentés a modulszerkezet legfontosabb tulajdonságairól.

Vizualizációs funkciók

A plug-in-ba számos vizualizációs lehetőséget építtem be. A hálózat modulszerkezetét a modul-hierarchia tetszőleges szintjén megjeleníthetjük, a pontok és élek színezésével. A visszavetített modulokat szintén megtekinthetjük, ekkor nem az adott hierarchiaszint, hanem az eredeti hálózat pontjai és élei kerülnek kiszínezésre. A 16. ábrán mind a két típusú vizualizációt bemutatom az *E. coli* baktériumban található Met-tRNS szintetáz fehérje aminosav hálózatán.

A modulszerkezet ábrázolásánál lehetőség van választani diszkrét és átfedő vizualizáció között. Mindkét esetben a plug-in először egymástól távoli színeket definiál az összes modulhoz. A diszkrét modulszerkezet ábrázolásánál a pontok annak a modulnak a színével lesznek kiszínezve, amelyik modulba a pont legnagyobb mértékben tartozik. Ilyen módon az alapvetően átfedő modul információkból egy diszkrét (átfedéseket nem tartalmazó)

szerkezeti ábra keletkezik. Az átfedő modulszerkezeti ábrázolás (25. ábra, A panel) esetében a pontok színe a modulok színének keveréséből jön létre, ahol a keverés a pont modulba tartozási értékeivel arányos. Az élek színe a végpontok színátlagaként kerül definiálásra.

A modulszerkezet ábrázolása mellett a plug-in által számolt mérőszámok ábrázolására (25. ábra, B panel) is lehetőség van. A plug-in minden mérőszám kiszámolása után a Cytoscape felületen elérhető pont tulajdonságként²⁰ letárolja az adott értékeket, és a ModuLand dialógus ablakon egyetlen kattintással képes ezen értékek alapján egy kék-piros lineáris színskála alapján pontszíneket rendelni.

25. ábra: Az ábra a ModuLand plug-in által kínált vizualizációs megoldások bemutatása az *E. coli* Met-tRNS szintetáz protein aminosav hálózatán. A panel: a pontok színezése átfedő modulbatartozásuk szerint; B panel: a pontok színezése a modul átfedés mérőszám alapján.

A mérőszámokat azért is jelenítettem meg a Cytoscape program pontokhoz rendelt tulajdonságai között, mert ilyen módon nem csak a plug-in által automatikusan definiált lineáris kék-piros színátmenettel lehet dolgozni, hanem be lehet állítani más típusú vizualizációt is a Cytoscape megfelelő vizualizációs funkcióit használva. A plug-in alkalmazási területeiről szóló fejezetben a kiemelt fehérjék azonosításával kapcsolatos 31. ábrán a Cytoscape programban kézzel definiált nem lineáris színskálát alkalmazok a modul centralitás megjelenítésére. A modul centralitás, a hídság és a köztiség centralitás mérőszámok alapvetően nem lineárisan skálázódnak, de az esztétikai élmény és az információ tartalom maximalizálása szempontjából legelőnyösebb színskála mindig az adott hálózat szerkezetétől függ.

²⁰ Node attribute

Hasonló modulok összevonása

Az egyik igen fontos funkció a modulok összevonását teszi lehetővé (ld. 26. ábra). A ModuLand eljáráscsaládban a modulokat a modulcentralitás felület hegyeiként azonosítjuk, ahol a felület lokális maximumai (hegycsúcsai) határozzák meg a modulok közepeit. A „zajos” / „recés” centralitás felület azt eredményezheti, hogy az egyes nagyobb modulok több egymással majdnem teljesen átfedő modulként kerülnek azonosításra. Ez bizonyos esetekben nem jelent problémát, hiszen ezek a modulok a következő meta-hálózat modularizálásánál nagy valószínűséggel összevonásra kerülnek a nagy átfedettségük miatt. Némely esetekben viszont a következő hierarchia szinten azonosított modulszerkezet már túlságosan nagy léptékű az adott felhasználáshoz. Ezekben az esetekben lehet hasznos a 26. ábrán bemutatott funkció, amely egy hisztogram segítségével ábrázolja az adott szinten előforduló modulpárok korrelációját.

26. ábra: Adott szintnél jobban korreláló modulok összevonása.

A modulpárok korrelációját a modulok hasonlóságaként értelmezhetjük. Értékét a két adott modulhoz definiált modulbatartozási vektor Spearman korrelációja adja (lásd még: 6.3. fejezet). A modulbatartozási vektor az összes pont adott modulba való besorolásának erősségét tartalmazza, és elemszáma megegyezik a hálózat pontjainak számával. A 26. ábrán bemutatott példán a Met-tRNS szintetáz fehérje aminosavainak hálózatán azonosított modulok korrelációs histogramját látjuk. A legtöbb modulpár korrelációja negatív, ami azt jelenti, hogy alapvetően nem hasonlítanak egymásra, a hálózat más-más pontjai tartoznak beléjük. A histogram nullánál nagyobb részén azt várnánk, hogy végig monoton csökkenjen, hiszen még az intenzív átfedések figyelembe vételével is egyre kisebb a valószínűsége, hogy egyre hasonlóbb aminosav régiókat lefedő modulokat találjunk. Azonban az előbb említett „zajos” centralitás felület miatt a 90% fölött korreláló modulpárok száma nagyobb a vártnál. Ilyen esetben előnyös lehet például a 0.9-es szint felett korreláló modulpárokat összevonni. A plug-in a felhasználó belátására bízta ezt a lépést, de segít a döntés meghozatalában a korrelációs értékek kiszámításával és ábrázolásával.

4.3 A ModuLand és más modularizáló plug-in-ek összehasonlítása

Számos átfedő modularizálási módszer került publikálásra az elmúlt években (ld. 2. fejezet) és számos modularizáló plug-in létezik, amelyek a Cytoscape programmal alkalmazva diszkrét, vagy éppen átfedő modulokat határoznak meg. Némelyik program (pl. a clusterMaker és a GLay) több, az irodalomból jól ismert modularizálási algoritmust is implementál, amelyeket egységes felhasználói felületbe integrál, és kiegészít különböző vizualizációs módszerekkel. Az MCODE és MINE plug-in-ek lokális hálózat sűrűségi definíciókat használnak, amelyek segítségével igen nagy sebességgel működnek, ami lehetőséget ad kimondottan nagy méretű hálózatok gyors, közelítő elemzésére. Ezen módszerek gyorsabban futnak a ModuLand programnál, és némelyikük modul átfedéseket is képes megmutatni, de tudomásom szerint a Cytoscape programban nem létezik jelenleg a ModuLand plug-in mellett fuzzy modularizálást végző plug-in, azaz olyan módszer, amely hozzá hasonló részletességgel tárná fel a modul hierarchiát és az átfedések finomszerkezetét azáltal, hogy a pontokat különböző erősséggel képes a különböző hierarchia szinten megjelenő valamennyi modulhoz hozzárendelni., valamint számos modulszerkezeten alapuló mérőszámot definiálni.

Az alábbiakban röviden jellemzem az általam ismert Cytoscape modularizáló plug-in-eket.

GLay plug-in (<http://brainarray.mbni.med.umich.edu/glay>)

A Glay plug-in (Su és mtsai, 2010) többféle modularizálási és kiterítési algoritmust tartalmaz. Többek között elérhető benne a modularitás mérőszámon (Newman, 2004b) alapuló Clauset-Newman-Moore algoritmus (Clauset és mtsai, 2004) többféle módon továbbfejlesztett, Java nyelven implementált változata (Wakita és Tsurumi, 2007). Ezek a modularizálási algoritmusok tetszőleges platformon elérhetőek. Számos további, az iGraph²¹ függvénykönyvtárban elérhető, C nyelven implementált algoritmus (például: *Walk Trap (Pons és Latapy, 2006)*, *Label Propagation* (Raghavan és mtsai, 2007), *Spin Glass* (Reichardt és Bornholdt, 2006), *Leading Eigenvector* (Newman, 2006)) szintén támogatott, azonban ezek csak 64 bites Windows operációs rendszeren futtathatóak.

ClusterMaker plug-in

(<http://www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.html>)

A ClusterMaker plug-in (Morris és mtsai, 2011) szintén számos klaszterező technikát integrál egy közös felhasználói felületen. A jelen verzió többek között támogatja a *K-mediod* (Sheng és Liu, 2006) és *K-means* (Bishop, 1995) hierarchikus algoritmusokat, amelyek eredményét pontok hierarchikus csoportjai, vagy hő térképek segítségével jeleníti meg. A plug-in további hasonlósági vagy távolsági mérőszámokon alapuló algoritmusokat is támogat: például a *Markov clustering* (Enright és mtsai, 2002), *transitivity clustering* (Wittkop és mtsai, 2010), *affinity propagation* (Frey és Dueck, 2007), *MCODE* (Bader és Hogue, 2003), *community clustering* (a GLay plug-in-ből ismert CNM változat), *SCPS* (Nepusz és mtsai, 2010) és az *AutoSOME* (Newman és Cooper, 2010) algoritmusokat.

²¹ Igraph, nyílt forráskódú függvénykönyvtár: <http://igraph.sourceforge.net/>

MCODE plug-in (<http://baderlab.org/Software/MCODE>)

Az MCODE algoritmus (Bader és Hogue, 2003) a többi klaszterező algoritmushoz képest gyors megvalósítást ad, amely a pontok lokális sűrűség szerinti súlyozásán, illetve a lokálisan sűrű régiók kiterjesztésén alapul. A felhasználó egy szabad paraméter szerint állíthatja, hogy milyen sűrűségű részek legyenek még bevonva az adott modulba, így akár a modularizálás után is lehetőség van a modulok méretének finomhangolására, anélkül hogy az egész hálózatot újra kellene modularizálni. Az algoritmus jellemző módon nem sorolja be a hálózat valamennyi pontját modulokba.

MINE plug-in

(http://chianti.ucsd.edu/cyto_web/plugins/displayplugininfo.php?name=MINE)

A MINE plug-in kimondottan sűrű, nagy összekötöttségű fehérje (gén) hálózatokban történő funkcionális modulok jó minőségű meghatározása céljából készült. A MINE algoritmus (Rhissorakrai és Gunsalus, 2011) egy, az MCODE-hoz hasonló felépítő jellegű modularizációs algoritmus, amely némileg más pont súlyozási szabályokat alkalmaz. Az algoritmus ezen felül egy hálózat modularitás mérőszámot is figyelembe vesz, annak érdekében, hogy elkerülje a sűrű fehérje hálózatokra jellemző túlzott modulméret növekedést.

NeMo plug-in (<http://128.220.136.46/wiki/baderlab/index.php/NeMo>)

A NeMo egy súlyozatlan hálózatok modularizálására szolgáló Cytoscape plug-in. A modularizáló algoritmus (Rivera és mtsai., 2010) egy speciális (közös szomszédok súlyozásán alapuló) lokális sűrűségi mérőszámot kombinál hierarchikus felépítő jellegű klaszterezési módszerekkel. A sűrűségi mérőszám lényege, hogy összeveti az adott pont pár közös szomszédainak valós számát a véletlenszerűen várható közös szomszédok számával.

5 A ModuLand eljáráscsalád teljesítménye és optimalizálása

Számos lehetőség kínálkozik a ModuLand eljáráscsalád optimalizált megvalósítására. A ModuLand Cytoscape plug-in fontosabb algoritmikus lépéseit és ezek futásidejét a következő táblázat foglalja össze. Az egyes algoritmusok mindegyike polinomiális futásidejű.

Az algoritmus leírása	Optimalizált egy szálon futó algoritmus komplexitása
LinkLand eljárás, az élek indirekt hatásainak szimulálása	$O(e(n+e))$
Az élfelület lokális maximumainak detektálása	$O(n+e)$
Arányos besoroló algoritmus	$O(edm)$
Modul korrelációk számítása	$O(m^2 n \log(n))$
Modulok összevonása korreláció alapján	$O(nm)$
Modulszerkezet levetítése az eredeti hálózatra	$O(nm^2)$
Átfedés mérőszám meghatározása a hálózat összes pontjára	$O(nm)$
Hídság mérőszám meghatározása a hálózat összes pontjára	$O(nm^2)$

1. táblázat: A ModuLand plug-in egyes számítási lépéseinek komplexitása. A képletekben az alábbi betűkkel jelöltem a hálózat különböző szerkezeti paramétereit: n : pontok száma, e : élek száma, d : átlagos fokszám ($d=e/n$), m : modulok száma.

Az algoritmusok gyorsítása és a tárhely igények csökkentése szempontjából fontosak a megfelelően definiált adatszerkezetek. Egy nagyobb (több százezer) pontot tartalmazó hálózat esetén előfordulhat, hogy a hagyományos két dimenziós mátrix struktúrák nem férnek el a memóriában, és a merevlemezre kiírt adatszerkezetekben is több gigabájt tárhelyet foglalnak. A programok megvalósítása során legtöbb esetben úgynevezett ritka vektor és ritka mátrix adatszerkezeteket használtam a modulba tartozási vektorok és mátrixok reprezentálására. Amennyiben minden pont minden modulba számottevő mértékben tartozik, akkor a ritka mátrixok (az index adatok miatt) több helyet foglalnak a hagyományos struktúráknál. Azonban a gyakorlat szerint mégis előnyösek, mivel például a modulba tartozási mátrixok

esetén a modulok átlagosan kisebbek a hálózatonál, illetve a pontok nem tartoznak valamennyi modulba. Gyakorlati tapasztalataink szerint egy fél millió pontos hálózat esetén nem ritka, hogy az optimalizált adatszerkezetek több mint 90-95%-os memória csökkenést okoztak.

Másrészt a ritka mátrixok bejárása (például az adatok kiolvasása, vagy épp a mátrixok szorzása esetén) sokkal kevesebb időt vesz igénybe. Ennek hatása különösen a mérőszákok generálása, az új hálózati szintek létrehozása és a modulszerkezetek szintek közötti vetítései közben jelentkezik. Komoly méretű (sok százezer pontos) hálózatok esetében ez napokról akár percekre redukálhatja a futás időt.

A LinkLand algoritmus megvalósítása közben az indirekt hatás által már elért pontokkal szomszédos pontthalmaz egy kupac adatszerkezetben kerül tárolásra. A szakirodalomban elérhető kupac szerkezetek közül a Fibonacci-kupac (Fredman és Tarjan, 1987) került implementálásra a kiemelkedő gyorsasága miatt. A modul besoroló programok számára szükséges az élek centralitás szerinti csökkenő rendezése, amelyet a QuickSort algoritmus (Hoare, 1962) segítségével végeztem el.

A fenti példák mindegyike olyan implementációs részleteket mutat be, amelyek nem csökkentik az algoritmusok elméleti komplexitását, mégis jelentős mértékben csökkentik a gyakorlatban mért futásidőt vagy memória szükségleteket.

A jelenleg publikált Cytoscape plug-in programjai egy számítógépen, egyetlen processzor mag használatával működnek, ami erősen behatárolja a vizsgálható hálózatok méretét. Valójában valamennyi lépés igény esetén jól párhuzamosítható, így a feladat több gépre szétosztva gyorsabban megoldható. Más részről alkalmazhatunk közelítő megoldásokat, amelyek a gyakorlatban nem változtatják meg az eredményeket, mégis sokkal gyorsabb futtatásokat tesznek lehetővé. A következő két fejezetben bemutatok egy párhuzamosítás és egy közelítés segítségével végrehajtott optimalizálási lépést, illetve közlöm az ezekre támaszkodó mérési eredményeket.

5.1 Gyenge élek számának redukálása a hálózat magasabb szintjein

A magasabb szintű meta-hálózatok pontjai az alacsonyabb szint moduljait reprezentálják, míg az élek a pontok közötti átfedések alapján kerülnek definiálásra. Az

alkalmazott képletek (ld. 3.4. fejezet) szerint, ha egy pont akármilyen kis részben is beletartozik több modulba, akkor ő önmagában átfedést okoz valamennyi modul között, amelybe tartozik. Amennyiben tehát x modulba tartozik az adott pont, akkor $\frac{x(x-1)}{2}$ él jelenik meg miatta a magasabb szintű meta-hálózatban.

A ModuLand eljáráscsalád egyik előnye, hogy feltárja a modulok finomszerkezetét. Ez azonban igen sokszor azt jelenti, hogy minden modul átfedésbe kerül a legtöbb másik modullal és a magasabb szintű hálózat arányaiban nagyon sok élt tartalmaz, akár teljes hálózatot alkothat. Természetesen a távoli modulok átfedésének értéke igen kicsi, és a nekik megfelelő magasabb szintű élek súlya sokszor 10-15 nagyságrenddel kisebb a valóban jelentős átfedéssel rendelkező modulok közötti meta-élek súlyánál. Az ilyen kis élsúlyok a magasabb szint modularizálása során gyakorlatilag nem játszanak szerepet, mert ezek indirekt hatása elenyésző. Azonban a LinkLand és a modulbesoroló algoritmusok futásideje az élszámmal skálázódik, így a rengeteg jelentéktelen magasabb szintű él komoly időtöbbletet okoz.

Triviális megoldás lenne a magasabb szintű hálózatok alacsony éleinek eldobása. Ez azonban két okból sem jó megoldás. Egyrészt a magasabb szintet előállító program lassulását nem oldja meg, hiszen csak ezen program futása után alkalmazható az optimalizáció. Másrészt a határérték definiálása nem triviális, hiszen a magasabb szinten megjelenő élsúlyok abszolút értéke és skálázódása a modulszerkezettől függ.

Az általam alkalmazott optimalizációban a magasabb szintű hálózatokat előállító programba építettem egy kapcsolót, amely beolvasás után minden pont esetén eldobja az adott pont számára relatív nagyon kicsi modulba tartozási értékeket. Pontosabban, a program az adott pont modulba tartozási erősségeinek összegének 1%-ánál kisebb modulba tartozási értékeket már nem veszi figyelembe. Ilyen módon valóban csak az elenyésző modulba tartozási értékek kerülnek nullázásra, és ez radikálisan csökkenti egyrészt a magasabb szinten megjelenő élek számát, másrészt magának a magasabb szintet előállító programnak a futásidejét.

A következő táblázatban a ModuLand plug-in futásidejét mutatom be különböző méretű példahálózatok segítségével. A példahálózatok jellemzése és a méréshez használt

számítógép hardver és szoftver tulajdonságai megtalálhatóak a ModuLand plug-in-t bemutató publikáció letölthető háttéranyagában²².

Hálózat	Pontok száma	Élek száma	Modulok száma	Opt. nélküli futásidő [mp]	Opt. futásidő [mp]
<i>B. aphidicola</i> metabolikus hálózat	190	563	8	0,82	0,78
<i>E. coli</i> metabolikus hálózat	294	730	23	0,92	0,84
<i>E. coli</i> Met-tRNS fehérje térszerkezeti hálózata	547	2153	96	2,19	2,04
Élesztő fehérje kölcsönhatási hálózat	2444	6271	55	9,6	8,91
Iskolai baráti hálózat	1127	5096	236	14,91	8,88
Nagyfeszültségű áram elosztó hálózat	4941	6594	207	23,61	22,98
Szó asszociációs hálózat	10617	63788	994	4935,51	702,97

2. táblázat: A ModuLand plug-in (másodpercben kifejezett) futásidejének összehasonlítása különböző méretű példahálózatokon, optimalizáció nélküli és a hálózat magasabb szintjén előálló élek számának redukálásából adódó optimalizáció bekapcsolásával.

A szóasszociációs hálózat esetében például a hálózat 994 modulja között összesen 493.521 modulpár esetén detektált a ModuLand plug-in valamilyen nem nulla átfedés értéket. Azonban ezen átfedések 78,077 százaléka annyira minimális volt, hogy az átfedés kialakításában résztvevő pontok mindegyikénél az adott modulpárba való közös relatív modulbatartozás nem érte el az 1 százalékot. Ezen nagyon apró átfedések eltávolítása után a magasabb szintű hálózatban 6321 él maradt, és a futásidő az egyhetedére csökkent.

5.2 Párhuzamosítási lehetőségek

A ModuLand algoritmus család számos tagja kiválóan párhuzamosítható. Az MSc diplomamunkám részeként (Szalay-Bekő, 2010) elkészítettem egy elosztott Java keretrendszert, amelyben gráfok felett futó algoritmusokat, illetve azok lépéseit lehet definiálni. Az algoritmus jellegében hasonlít a Google által implementált MapReduce

²² A példa hálózatok és a háttéranyag letölthetőek a plug-in honlapjáról: <http://www.linkgroup.hu/modules.php>

programozási modellhez (Dean és Ghemawat, 2004), mivel az adott feladatokat két meghatározott lépés ismételtetésére bontja: az egyes gráfelméleti algoritmikus feladatokat egy központi egység előbb szétosztja a kiszolgáló számítógépek között, majd a részműveletek eredményeinek összefűzése történik meg. A keretrendszerbe az egyes gráf algoritmusok plug-in-ként ékelhetők be, amely plug-in-ekbe a felhasználók a központi szerver egységben és az elosztott kliens egységben futó számításokat implementálják.

27. ábra: Az elosztott gráf algoritmusok futtatására szolgáló keretrendszer felépítése.

A központi egység és az ebbe épített plug-in felelős az algoritmus lépéseinek szinkronizálásáért, valamint a részfeladatok kiosztásáért és az elosztott adatbázisban replikált adatok csoportosításáért. A grafikus felhasználói felületen keresztül a felhasználó algoritmusokat indíthat el a központi egységen, illetve megfigyelheti az egész elosztott rendszer állapotát. Mivel a rendszer növekedésével a központi egységnek egyre több klienssel kell tartania a kapcsolatot, ezért a rendszer skálázhatóságának szempontjából fontos, hogy az általa végzett feladatok gyorsak legyenek.

A kliens egységekbe illesztett plug-in-ek végzik a jelentősen időigényesebb számítási munkákat, valamint a kliens egységek rendelkeznek az elosztott adatbázis egy-egy példányával. Minden osztottan tárolt adatelem (például gráfok pontjai, élei, vagy az ezekhez rendelt adat) rendelkezik egy elsődleges adatbázissal, amely felelős az adott adat tárolásáért. Ha a többi kliensnek szüksége van az adott adatra, akkor ezt az adott adatelem elsődleges adatbázisától kéri el. A kommunikáció csökkentése érdekében az adatbázisok nem csak a

saját adataikat mentik el, hanem tárolják az elsődlegesen nem hozzájuk tartozó, de általuk már valamikor használt adatokat is.

A rendszer megbízhatóságának növelése érdekében minden egység összehangolt módon menteni képes az állapotát. Ha egy-egy számítógép meghibásodik a hosszabb programok futtatása közben, a rendszer képes automatikusan visszatölteni a legutóbbi mentett állapotot minden egyes gépen, majd újra elindítani a végrehajtást. A biztonság fokozása érdekében két fázisú állapotmentés gondoskodik arról, hogy a mentések készítése közben bekövetkező hibák se tegyék működésképtelenné a rendszert.

Az egyes elosztott komponensek közötti kommunikációt asszinkron üzenetváltásokkal valósítottam meg. A fontosabb üzenet típusok közé tartoznak a plug-in-ek és a keretrendszer életciklusát vezérlő üzenetek, a komponensek állapotát monitorozó (úgynevezett *health-check*) üzenetek és az elosztott adatbázis szinkronban tartásáért felelős kliensek közötti üzenetváltások.

28. ábra: Az elosztott gráf algoritmusok futásának illusztrálása.

Az algoritmusok MapReduce (Dean és Ghemawat, 2004) jellegű megfogalmazása azt jelenti, hogy a megoldást felbontjuk több olyan részfeladatra, amelyeken az elosztott számítást végző kliensek anélkül tudnak dolgozni, hogy a globális hálózat teljes egészére rálátnának. Miután az adott részfeladattal valamennyi kliens végzett, egyrészt jelzi ezt a szervernek, másrészt elvégzi részeredményeinek beírását az elosztott gráf adatbázisba. A szerver a részeredmények alapján kijelöli a következő részfeladatot. Az algoritmusok elosztott futását a 28. ábrán szemléltetem.

29. ábra: Az elosztott LinkLand algoritmus futásidejének összehasonlítása az egy gépes változattal, a futtatást végző kliensek számának függvényében.

Az elosztott működés egy-két gép használata esetén nem térül meg, hiszen a szerver és a kliens kommunikációja, valamint az algoritmus részfeladatokra való bontása többlet időt vesz igénybe. Viszont egy nagy hálózaton futtatott hosszú algoritmus esetén a több gép használata jelentősen megtérül. A 29. ábrán látható a LinkLand algoritmus MapReduce jellegű átíratának futása egy körülbelül 20.000 élt tartalmazó hálózaton. Az eredeti (Cytoscape ModuLand plug-in-ban is megtalálható, a Java nyelvű elosztott programmal szemben a gyorsabb C++ nyelven íródott) egy gépes implementáció futásideje viszonyításképpen piros szaggatott vonallal van jelezve. Látható, hogy még két gép esetén sem éri meg elosztottan futtatni, azonban például öt gép használatánál már a futásidő negyedére esik. Az elosztott programok általában ritkán skálázódnak lineárisan, és ha mégis, akkor azt csak egy adott

tartományon belül teszik. Jól látható, hogy az algoritmus skálázhatóságából és az adott hálózat szerkezetéből kifolyólag (illetve a keretrendszer által hozzáadott futásidő többlet miatt) jelen esetben öt gépnél több hozzáadása nem gyorsítja a LinkLand algoritmust. Természetesen nagyobb hálózatok elemzésénél valószínűleg ennél jóval több számítógép használata is kifizetődő lehet.

30. ábra: Az elosztott LinkLand algoritmus futásidejének összehasonlítása az egy gépes változattal, a vizsgált hálózat méretének függvényében.

A kliensek száma mellett a hálózat mérete is befolyásolja az elosztott keretrendszer hatékonyságát. Kis hálózatok esetén nem érdemes elosztott algoritmusokat futtatni, mivel a hálózat szétterítésének és a munka szétosztásának ideje összemérhető a számítási feladatok által igényelt idővel. A 30. ábrán az egy számítógépen futtatott optimalizált LinkLand program futásidejét hasonlítom össze az 5 számítógépen futtatott elosztott verzió futásidejével. Látható hogy körülbelül 5-6000 él alatt nincs számottevő különbség a két verzió között (sőt, az adatokból kiderül, hogy néhol az egy gépes verzió gyorsabb is). Azonban az élek számának növekedésével a két program futásideje közötti különbség egyre inkább nő.

6 Alkalmazási területek

Ebben a fejezetben áttekintem a fuzzy modularizálási módszerek alkalmazhatóságát a biológia különböző területein. Számos példát mutatok a ModuLand eljáráscsalád használatára, amelyeket részben én, részben a kutatócsoportom tagjai, részben pedig független szerzők publikáltak. Emellett igyekszem olyan új alkalmazási területeket is megemlíteni, amelyeken tudomásom szerint még nem alkalmaztak fuzzy modularizálást, de valószínűleg érdekes eredményeket lehetne elérni annak segítségével.

6.1 Fehérje tulajdonságok jóslása

A fehérje-fehérje interakciós és metabolikus hálózatok esetében széles körben elterjedt az a vélemény, hogy az itt azonosított modulok funkcionális egységeket, vagy akár adott esetben molekula komplexeket alkotnak a sejtben (Ravasz és mtsai, 2002; Chen és Yuan, 2006; Spirin és Mirny, 2003). Azaz a modulokon belül vizsgálva a fehérjék funkcióit, bizonyos fehérje funkciók szignifikánsan feldúsulnak az adott modulban, míg a többi funkció a statisztikailag vártnál ritkábban fordul elő. Sőt, amennyiben az ismert fehérje komplexeket eltávolítjuk (vagy összevonjuk) a molekuláris hálózatokban, az így kialakult hálózatok továbbra is erősen modulárisak (Wang és Zhang, 2007) és a modulok funkcionálisan értelmezhetőek maradnak. Amennyiben pedig a modulokhoz sejtes funkciókat rendelhetünk, úgy azok a fehérjék és egyéb metabolitok, amik szintén az adott modulban vannak, valószínűleg részt vesznek az adott funkció megvalósításában.

A fuzzy modularizációs módszerek a hagyományos klaszterezési eljárásoknál önmagukban is alkalmasabbak a fehérje funkció jóslására (Szalay-Bekő és mtsai, 2012; Kovács és mtsai, 2010; Ma és mtsai, 2010). Egyrészt detektálni tudják hogy az adott pont milyen erősen tartozik az adott funkcionális modulba. Ennek felhasználásával valószínűségi értéket rendelhetünk az adott fehérje adott sejtes funkcióban való részvételére. Másrészt a modulok átfedésében lévő pontok természetes módon egyszerre több sejtes funkcióhoz is hozzárendelhetőek, ami a valóságot jobban megközelíti a diszkrét klaszterezés által adott eredményeknél. Az utóbbi megállapítást számszerűen is alátámasztották egy tanulmányban (Becker és mtsai, 2012), ahol a klaszterek átfedéseiben található fehérjék között

szignifikánsan feldúsultak az olyan fehérjék, amelyeknek több eltérő funkciója ismert. A 31. ábrán a ModuLand plug-in segítségével egy élesztő fehérje-fehérje kölcsönhatási hálózaton meghatározott funkcionális modulok jelenlétét szemléltetem.

31. ábra: A ModuLand Cytoscape plug-in segítségével kiszámított modul centralitás értékek vizualizálása kézzel beállított színskála alkalmazásával egy élesztő fehérje hálózaton. A fekete ellipszisek modul központokat jelölnek, a következő jellemző funkcionális tulajdonságokkal: (a): sejtvázas alkotók, (b):sejtmag membrán pórus komplexek, (c): riboszóma képződés folyamata, (d): RNS kivágódás (splicing)

A fehérje-fehérje kölcsönhatási hálózat topológiája alapján a fehérje funkció jóslás minősége gyenge lehet, pusztán a bemeneti adatok hiányossága és zajossága (Suthram és mtsai, 2006) miatt. Wang és munkatársai (Wang és mtsai, 2007) a fehérje párok funkcionális hasonlóságának jóslására komplex statisztikai keretrendszert fejlesztettek ki, amely figyelembe veszi a fehérje párok közötti legrövidebb utak hosszát, a közös szomszédok létezését és a domainszerkezetbeli hasonlóságot. Zhao és munkatársai (Zhao és mtsai, 2008) pedig egy általános gépi tanulást használó keretrendszert alkottak hasonló célra. A fenti rendszerekbe az átfedő modularizálás eredményeit szintén be lehetne csatornázni a jobb minőségű jóslások elérése érdekében.

Az 1.3.1. fejezetben említettem a date és party hub-ok elkülönítésének problémáját. A ModuLand eljárás család használatával (Kovács és mtsai, 2010) lehetőség nyílik a két csomópont típus jóslására, tisztán a hálózat topológiájának ismeretében.

32. ábra: Az élesztő fehérje kölcsönhatási hálózatában definiált date és party hub-ok elkülönítése az átfedő modularizálás eredményeinek felhasználásával.
Forrás: Kovács és mtsai, 2010

A 32. ábrán látható, hogy a date és party hub-ok csoportja hogyan különül el egymástól az Ekman és munkatársai által közölt (Ekman és mtsai, 2006) adatok ModuLand elemzése után. A használt hálózatban 2633 fehérje között 6379 kölcsönhatást definiáltak, a hálózat legnagyobb komponense 2445 pontból állt, melyek közül 318 date és 201 party hub-ot jelölt meg a cikk, kísérleti eredmények alapján. Az átfedő modularizálásra támaszkodó elemzés alapötlete az, hogy a party hub-okra elvárásaink szerint jobban jellemző, hogy az egy sejt alkotóban és egyidejűleg kialakított kapcsolataik végpontjai velük egy modulban

tartózkodnak, mint a kapcsolatait különböző helyen vagy időben kialakító date hub-okra. Az elemzés elvégzéséhez a LinkLand kupacképző és a totális besoroló eljárást használtuk (Kovács és mtsai, 2010). Az egyes fehérjék hídság és modul centralitás mérőszámainak meghatározása után (lásd: 3.5. fejezet) a 32. ábrán látható nem-lineáris szeparációs függvény használatával helyesen azonosítottunk 84 party és 307 date hub-ot. Az eredmény még meggyőzőbb, ha figyelembe vesszük, hogy a hibásan jósolt date hub-ok 91%-a, illetve hibásan jósolt party hub-ok 76%-a feltehetőleg az Ekman cikkben helytelenül (de legalábbis vitatott módon) volt besorolva, összevetve az eredeti adatsort a később publikált (Bertin és mtsai, 2007; Han és mtsai, 2004; Kim és mtsai, 2006; Komurov és White, 2007) adatokkal.

6.2 Gyógyszerkutatási alkalmazások

Egy új, átfogó, a gyógyszerkutatásban használt hálózatkutató eszközöket bemutató tanulmány (Csermely és mtsai, 2013) szerint hálózatos szempontból alapvetően két gyógyszercélpon keresési stratégiát használnak ma. Az egyik, amikor szelektív módon roncsolni akarják a hálózatot, tipikusan központi és esszenciális célpontokat megtámadva, így iktatva ki a fertőzést okozó organizmust vagy a rákos sejteket. A másik stratégia a sejt működését befolyásolja, tipikusan több ponton módosítja a sejt molekuláris hálózatát azzal a céllal, hogy helyreállítsa a hálózat egészséges egyensúlyi állapotát. Összefoglalásul elmondhatjuk, hogy a fuzzy modularizálási módszerek és a ModuLand plug-in ezeket a stratégiákat egyrészt támogatják a biomolekuláris hálózatok funkcionális moduljainak feltérképezésével és a rendszer struktúrájuk megértésével, másrészt pedig segítenek azonosítani az esszenciális központi, vagy épp az átfedésekben lévő kulcsszerepet betöltő fehérjéket.

Az elmúlt években a hálózat kutatás által nyújtott rendszerszintű nézőpont és a folyamatosan bővülő elemzési lehetőségek (Csermely és mtsai, 2013) egyre népszerűbb kiegészítő eszközei az egyébként komoly kihívásokkal szembesülő (Begley és Ellis, 2012; Bunnage, 2011) gyógyszerkutatói szektornak. A modularizálás lehetőséget ad betegség modulok azonosítására, azaz olyan funkcionális modulok detektálására a fehérje-fehérje kölcsönhatási hálózatban, amelyek egy-egy adott betegséghez köthetőek. Az ilyen módon jósolt fehérje funkciók segítik a gyógyszer célpontok kiválasztását (Loscalzo és Barabasi,

2011; Cho és mtsai, 2012). Az átfedő modularizálási módszerek felfedik emellett a sejten belüli molekuláris hálózatok (a többi természetes komplex rendszerhez képest is) igen sűrű, átfedésekben gazdag modulszerkezetét (Ahn és mtsai, 2010; Kovács és mtsai, 2010; Palla és mtsai, 2005), ahol néha az átfedés maga sűrűbb a funkcionális modulok központi részénél (Yang és Leskovec, 2012). A sejt molekuláris hálózatának sűrűsége jelzi hogy a természet „nem pazarol”, azaz egy fehérje, vagy egyéb kémiai anyag számos különböző funkcióban is részt vesz.

33. ábra: Biomolekuláris hálózatokban funkcionális modulok között elhelyezkedő, gyógyszerkutatási szempontból érdekes elemek szemléltetése.

Az előzőekben említett okok miatt a fuzzy modularizálási módszerekkel azonosítható, modulok között elhelyezkedő és azokat összekötő „híd” fehérjék vonzó gyógyszer-célpontok (Hwang és mtsai, 2008). Hasonlóan érdekesek a távoli modulokat összekötő csomópontok (date hub-ok), amelyek szintén felfedhetők a modulok finomszerkezetének ismeretében (Kovács és mtsai, 2010). A date hub-oknak általában regulációs szerepük van (Fox és mtsai, 2011) és rákos sejtekben gyakran mutálódnak (Taylor és mtsai, 2009). A modulok közötti átfedő régióknak szintén kiemelt szerepük van a gyógyszer-célpontok választásakor. A sejt stressz válasz esetén éppen a funkcionális moduljai közötti átfedések mértékének

változtatásával reagál (Mihalik és Csermely, 2011). A 33. ábrán az itt említett, gyógyszerkutatás szempontjából is fontos elemek elhelyezkedését szemléltetem.

A fertőzések vagy rák ellenes gyógyszerek célja sokszor a paraziták, vagy beteg sejtek elpusztítása, ami hálózatos szempontból a hálózatuk szétroncsolásának felel meg. Azaz a hatékony gyógyszerek itt esszenciális (a sejt életben maradásához elengedhetetlen) fehérjéket céloznak meg, amelyeket azonosításához szintén hasznos segítséget ad a fuzzy modularizálás. Más típusú gyógyszerek esetben épp a toxicitás minimalizálása érdekében hasznos az esszenciális fehérjék ismerete és elkerülése (Csermely és mtsai, 2013). Számos lokális és globális tulajdonság (Jeong és mtsai, 2001; Estrada, 2006; Yu és mtsai, 2007; Song és Singh, 2013) figyelembevételével mellett a modulszerkezet ismeretében szintén hatásos módon lehet esszenciális fehérjéket detektálni, például a ModuLand eljárás család által meghatározott modul centralitás (Kovács és mtsai, 2010) mérőszámokkal.

A sejten belüli jelátviteli hálózatok feltérképezése kiemelten fontos mind a sejt szabályozó folyamatainak megértése, mind a hatékony gyógyszerkutatás szempontjából. (Csermely és mtsai, 2013) A jelátviteli hálózatok a hagyományosan vizsgált komplex jelátviteli útvonalakból és az őket összekötő, utóbbi évtized folyamán egyre nagyobb hangsúlyt kapó (Papin és mtsai, 2005; Fraser és Germain, 2009) kereszt-csatlakozásokból (*cross-talk*) állnak. Ha rendszer szinten vizsgáljuk a sejt jelátviteli folyamatait, akkor gyakorlatilag a jelátviteli útvonalak modulként, a *cross-talk* régiók pedig átfedéseként jelennek meg a hálózatban. Összehasonlítva az ember, a *D. melanogaster* (muslinca), és a *C. elegans* (fonalféreg) jelátviteli hálózatát, látható hogy az ember esetében sokkal nagyobb az útvonalak között található átfedések mértéke és a híd jellegű fehérjék száma (Korcsmáros és mtsai, 2010). Más részről viszont az átfedések nem léphetnek át egy adott mértéket, és szükség van a jól definiált modulszerkezetre a jelátvitelben, különben a jelzések túl könnyen átkerülnének egyik jelátviteli funkcióból a másikba, túl nagy lenne a „zaj” a jelátvitelben és a szabályozás nem lenne hatékony (Nussinov, 2012). A ModuLand-hez hasonló eszközök képesek az átfedések pontos detektálására, amik számos már jelenlévő hatékony gyógyszer által megcélzott fehérje azonosítását teszik lehetővé, ezért ígéretesek lehetnek a jövő gyógyszereinek fejlesztésében (Farkas és mtsai, 2011).

6.3 Modulszerkezetek elemzése, összehasonlítása

A fuzzy modularizálási eljárások sokkal több adatot generálnak diszkrét társaiknál. Míg diszkrét esetben minden ponthoz egy modul sorszámát rendeljük, addig fuzzy esetben minden ponthoz 1-1 vektor társul, amely leírja hogy melyik modulba mennyire tartozik az adott pont. A több adat nem csak finomabb elemzéseket tesz lehetővé, de a különböző hálózatok modulszerkezetének sokkal finomabb összehasonlítását is eredményezi.

A ModuLand eljárás család segítségével végzett elemzéseinkben (Kovács és mtsai, 2010; Szalay-Bekő és mtsai, 2012) gyakran érdekes eredményekre vezettek a különböző modulszerkezetek korrelációs összehasonlításai. Az elemzéseinkben az egyes modulokhoz tartozó pont erősség vektorok Spearman-rangkorrelációját számítjuk ki, ami egy modulpárra -1 és 1 közötti valós számot ad. A korreláció értéke 1-hez közeli, ha a két modul gyakorlatilag ugyanazokat a pontokat fedi le és arányaiban ugyanolyan súllyal. A -1-hez közeli érték a negatív korrelációt fejezi ki, azaz ha egy pont az egyik modulba nagyon beletartozik az a másik modulba szinte egyáltalán nem, és fordítva. A 0-hoz közeli korrelációs értékek azt jelölik, hogy a két modul között nincs szignifikáns összefüggés. A mérőszámot önmagában is lehet használni arra, hogy a nagyon hasonló modulokat összevonjuk, ezzel egyszerűsítve a hálózat modulszerkezetét²³.

A fenti módszer akkor is hasznos lehet, ha az adott rendszer különböző állapotait leíró hálózatok struktúráját kell összehasonlítani. Ilyenkor a hagyományos globális mérőszámok (például átmérő, foksám eloszlás, stb – lásd 1.3.1. fejezet) mellett egyrészt érdemes összehasonlítani a modulok számát, átlagos méretét, átfedéseik mértékét, illetve az egyes modul vagy pont tulajdonságok eloszlásait. Másrészt a korrelációs számításokkal meg lehet határozni azt is, hogy melyek voltak azok a modulok, amelyek majdnem érintetlenül megmaradtak, melyek olvadtak össze, vagy váltak szét egymástól.

Élesztő biomolekuláris hálózatainak átalakulása stressz hatására

Egy érdekes tanulmány (Mihalik és Csermely, 2011) a ModuLand eljárás család használatával megvizsgálta az élesztő (*S. cerevisiae*) mRNS expressziós adatok segítségével

²³ Lásd: 4.2. fejezet, a ModuLand plug-in által kínált szolgáltatás hasonló modulok összevonására.

súlyozott fehérje-fehérje kölcsönhatási hálózatának hőstressz hatására végbement szerkezeti változását. A stresszelt hálózat pontjai megfeleltek a stressz mentes hálózat esetében használt pontoknak, de az éleket és azok súlyát az mRNS szintek változása befolyásolta. Az elemzés egyik kulcsfontosságú eredménye, hogy az élesztő fehérje moduljai a stressz hatására szétválnak egymástól, átfedéseik mértéke és éleik sűrűsége is csökken. Az ismert, a hő sokk hatását javító fehérjék közül sok a modulok központi részébe kerül (praktikusan körük épülnek az új modulok), vagy pont a stressz hatására átalakult modulok összekötésében kulcs szerepet játszó híd elemmé válik.

Egy másik kutatócsoport nemrég megerősítette és tovább vitte a fenti elemzést és egy eltérő élesztő típust használva, mind annak fehérje interakciós hálózatát, mind a biomolekulák közös expresszáldást modellező hálózatát (*co-expression network*) megvizsgálta nyugalmi állapotban, majd oxidatív stressz alatt (Lehtinen és mtsai, 2013). A vizsgálatokhoz többek között a ModuLand plug-in-t használták. Az eredményeik alátámasztották a korábban megfogalmazottakat, továbbá számos új és érdekes megfigyelést tartalmaztak az oxidatív stressz élesztőre gyakorolt rendszer szintű hatásaival kapcsolatban.

Fehérje térszerkezeti háló elemzése és összehasonlítása a domén szerkezettel

A fehérjék térszerkezetét leíró hálózatok pontjai aminosavak (vagy atomok), amelyeket kémiai kötéseknek megfeleltetett irányítatlan élek kötnek össze. Amennyiben ezeket a hálózatokat modularizáljuk, a modulstruktúrát össze lehet hasonlítani a doménstruktúrával, hiszen ez utóbbi felfogható az aminosavak feletti diszkrét modularizálásként. Sokszor a doménszerkezet és a modulszerkezet erősen korrelál (Xu és mtsai, 2000; Delmotte és mtsai, 2011; Szalay-Bekő és mtsai, 2012). A térszerkezeti hálók olyan centrális pontjai (csomópontok, magas *betweenness* vagy *closeness centrality* mérőszámmal rendelkező elemek), amelyek szomszédai között kevés kapcsolat fut, gyakran részt vesznek például heme-csoportok megkötésében (Xie és mtsai, 2011). A pontos modulszerkezet meghatározása a doménszerkezet meghatározásán túl segíthet a fehérje konformációs változásainak megértésében épp úgy, mint a flexibilis (rendezetlen) régiók detektálásában vagy a kitüntetett szerepű szerkezeti részek meghatározásában (Csermely és mtsai, 2011).

A ModuLand plug-in-t bemutató cikkünkben (Szalay-Bekő és mtsai, 2012) elemeztem az *E. coli* baktérium Met-tRNS szintetáz fehérje 547 aminosavból álló hálózatát. Egyrészt megállapítottam, hogy az első modulszinten megjelenő 47 modul további összevonása után a modul hierarchia következő szintjén megjelenő 5 modul erősen korrelál a fehérje 3 doménjével (amelyek közül az első domén három al-doménre oszlik). Az 3. táblázatban bemutatom a modulok és a domének Spearman-rangkorrelációs értékeit, amelyekből jól látható, hogy az ötből három modul erősen korrelál a három nagy doménnel. A maradék két (ezidáig még be nem sorolt) modul pedig bizonyos domén és al-domén párokkal korrelál (4. táblázat).

	N-terminális, katalitikus domén	Peptid kötő domén	C-terminális, antikodon kötő domén
1-es modul	0,68	-0,19	-0,4
2-es modul	0,33	0,28	-0,58
3-as modul	-0,11	0,61	0,77
4-es modul	0,08	-0,68	0,59
5-ös modul	-0,24	0,8	-0,53

3. táblázat: Az *E. coli* Met-tRNS szintetáz fehérje második szintű moduljainak korrelációja három fő doménnel.

	N-terminális, katalitikus domén			Peptid kötő domén	C-terminális, antikodon kötő domén
	Rossmann-fold 1 aldomén (katalitikus)	Rossmann-fold 2 aldomén (katalitikus)	Stem contact fold domén		
2-es modul	0,13	0,22	0,13	0,28	-0,58
4-es modul	-0,02	-0,27	0,42	-0,68	0,59

4. táblázat: Az *E. coli* Met-tRNS szintetáz fehérje két második szintű moduljának korrelációja a három al-doménnel és a két maradék fő-doménnel.

Az eredeti adatokat közlő publikációban (Ghosh és Vishveshwara, 2007) a szerzők 43 aminosavat jelöltek meg, amelyeket ebben a fehérjében a területen eddig megjelent cikkek különböző kísérletes és analitikus eszközök felhasználásával jelátviteli aminosavként azonosítottak. Ezen aminosavak felelősek az enzim katalitikus központja és az antikodon

kötőhely közötti konformációs változások továbbításáért. A cikkünkben ezt a hálózatot példaként használtuk, mert jól látható hogy a jelátviteli aminosavak azonosítását elősegítheti a ModuLand eljáráscsalád által meghatározott átfedő modulszerkezet elemzése. A 34. ábrán látható az aminosav hálózat ModuLand programmal meghatározott átfedés és hídság mérőszámainak eloszlása. A 43 jelzés átvitelért felelős aminosav hídság és átfedés értéke szignifikánsan²⁴ nagyobb a hálózat többi pontjára számított értékeknél, ami azt sugallja, hogy a jelterjedés megkülönböztetetten használja a modulok között található és azokat összekötő régiókat.

34. ábra: A Met-tRNS szintetáz fehérje aminosav hálózatában a pontok modul átfedés és hídság mérőszámainak eloszlása (grafikonok) és átlag értéke (nyilak), megkülönböztetve a konformációs változás továbbításáért felelős pontokat (piros) a hálózat többi pontjától (szürke).

24 A szignifikancia meghatározására a Welch-féle kétmintás t-próbát alkalmaztuk: $p < 0,0008$

A Met-tRNS szintetáz fehérje térszerkezeti hálójának további elemzése megtalálható a ModuLand plug-in leírását közlő cikk (Szalay-Bekő és mtsai, 2012) letölthető háttéranyagában²⁵.

Evolúciós és filogenetikus elemzések

Az evolúció egyik alap jellemzője a korábban már bevált struktúrák megőrzése és későbbi felhasználása. Ezért – nem meglepő módon – a biomolekuláris hálózatok funkcionális moduljai evolúciósan őrződnek (Wang és Zhang, 2007). Persze minden élőlény hálózata eltérő, ezért annak a felismeréséhez hogy mely funkcionális részek őrződtek, szükséges lehet a modulok finomszerkezetének ismerete amit a fuzzy modularizálással tudunk meghatározni.

A különböző fajok fehérje kölcsönhatási hálózatainak egymáshoz való illesztése – tisztán informatikai problémaként kezelve – igen nehéz, NP-teljes probléma (Zhang és mtsai, 2008). Szerencsére azonban az összevetésben segít, hogy számos fehérje ortológ párosítása ismert, habár ez a párosítás sokszor nem egyértelmű. A fehérje hálózatokat illesztő algoritmusok a hálózat topológiája mellett kezdetektől felhasználták a szekvencia adatokat és a fehérje ortológ adatbázisokat, (Ogata és mtsai, 2000; Kelley és mtsai, 2003; Berg és Lässig, 2004). A későbbi kifinomultabb algoritmusok valószínűségi modellekkel (Berg és Lässig, 2006), matematikai optimalizációs eljárásokkal (Li és mtsai, 2006b), vagy például a Google PageRank algoritmusához hasonló iteratív módszerekkel (Singh és mtsai, 2008) illesztették a fehérje hálózat egészét, vagy kerestek egymáshoz illeszkedő funkcionális modulokat és jelátviteli utakat. A különböző fajok egymással illesztett biomolekuláris hálózatainak összehasonlítása és a modulszerkezet evolúciójának vizsgálata érdekes kérdés, aminek a jövőbeli vizsgálatához a fuzzy modularizációs módszerek igen hasznos segítséget tudnak nyújtani. Egy nemrégiben megjelent tanulmányban (Wang és Gao, 2012) éppen a ModuLand eljáráscsalád segítségével illesztettek különböző fajokból származó fehérje-fehérje kölcsönhatási hálózatokat, olyan módon, hogy az illesztés során a funkcionális modulok szerkezete minél kevésbé sérüljön.

25 <http://www.linkgroup.hu/modules.php>; http://www.linkgroup.hu/moduland_plugin/latest/ModuLand-Cytoscape-plugin-Suppl.pdf

A hálózatos modellezés és a modularizálás egy igen érdekes felhasználása (Andrade és mtsai, 2011), amikor nem egy faj, hanem egy adott fehérje evolúciós fejlődését szemlélítve egy hálózattal, a pontoknak az adott fehérje különböző fajokhoz tartozó ortológ megjelenéseit választjuk, és az élek súlyai fordítottan arányosak az ortológ fehérje szekvenciák között számolt különbségekkel. Ebben a hálózatban a modulok gyakorlatilag olyan faj csoportokat jelölnek, amelyekben az adott fehérje szerkezete és működése nagyon hasonló, evolúciós szempontból stabilan konzerválódott. Noha az eredeti cikkben a szerzők hagyományos, diszkrét klaszterezést végeztek, érdekes lenne feltárni a modulszerkezet finomabb tulajdonságait is, például a ModuLand eljárás családjának használatával.

Az *Escherichia coli* és *Buchnera aphidicola* baktériumok metabolikus hálózatainak összehasonlítása

A ModuLand plug-in közlésekor (Szalay-Bekő és mtsai, 2012) példaként részletesen összehasonlítottuk az *Escherichia coli* és *Buchnera aphidicola* baktériumok metabolikus hálózatainak (Feist és mtsai, 2007; Thomas és mtsai, 2009) modulszerkezetét. A két baktérium filogenetikailag közel áll egymáshoz, génkészletük elemzése alapján (Moran és Mira, 2001) valószínűleg közös ősbaktériumból fejlődtek ki. A két élőlény metabolikus hálózata azért is érdekes, mert ezek a hasonló baktériumok teljesen eltérő életmódot folytatnak (Moran és Wernegreen, 2000) és metabolikus folyamataik ennek megfelelően specializálódtak. Míg az *E. coli* önállóan, változó körülmények között él, addig a *B. aphidicola* egy szimbióta baktérium aminek a környezete meglehetősen stabil.

A 35. ábrán bemutatom a két baktérium metabolikus hálózatát, az 5. táblázatban pedig feltüntettem a hálózatok és azok modulszerkezeteinek fontosabb tulajdonságait. Mivel az *E. coli* metabolikus hálózata 294 pontot tartalmaz, ami lényegesen több a *B. aphidicola* 190 pontjánál, ezért a méretkülönbségből adódó eltérések kiszűrése érdekében 1000-1000 különböző véletlenszerű részhálózatot készítettem az *E. coli* eredeti hálózatából. Ezen hálózatok készítése során folyamatosan véletlenszerű pontokat hagytam el az eredeti *E. coli* hálózatból, amíg annak pontjainak vagy éleinek száma el nem érte a *B. aphidicola* hálózat pont vagy él számát.²⁶

²⁶ A 2000 véletlenszerű részháló és a mintavételezést végző program forráskódja szabadon elérhető és

35. ábra: Az *E. coli* és *B. aphidicola* baktériumok metabolikus hálózata. A pontok a ModuLand program által meghatározott moduloknak megfelelően vannak színezve, a modulok meta-hálózatai a kis részabrákon találhatók.

A két hálózatban a pontok szomszédainak átlagos száma egyaránt alacsony (5 és 6), viszont a két háló szerkezete jelentősen eltér egymástól. A *B. aphidicola* esetében egy jóval centralizáltabb hálózatot láthatunk két domináns, nagy méretű modullal, amelyek középpontjában az ATP szintézisért, illetve a D-glükóz transzportfunkciókért felelős metabolikus elemek helyezkednek el. Az *E. coli* esetében a modulok nagyobb száma (23, szemben a *B. aphidicola* 8 moduljával), nagyobb lényegi darabszáma (6,2 szemben 2,1-el) és a modulok átlagosan kisebb mérete (*E. coli*: 12,78; *B. aphidicola*: 23,75) mind

letölthető: <http://www.linkgroup.hu/modules.php>

alátámasztották azt a megfigyelésünket, hogy az *E. coli* hálózata jóval tagoltabb, itt több, mind arányait, mind abszolút méretét tekintve kisebb modul figyelhető meg. Ez a megállapítás összhangban van azokkal a tanulmányokkal (Kreimer és mtsai, 2008; Parter és mtsai, 2007; Samal és mtsai, 2011), amelyek szerint a környezeti variabilitás nagyobb fokú tagoltságot okoz az élőlények metabolikus hálózataiban.

	<i>E. coli</i>	<i>E. coli</i> részháló (pont limit: 190)	<i>E. coli</i> részháló (él limit: 563)	<i>B. aphidicola</i>
Pontok száma:	294	190	258 ±5,4	190
Élek száma:	730	328 ±22,9	559 ±4,04	563
Globális fürtösödési együttható:	0,54	0,58 ±0,04	0,59 ±0,02	0,6
Pontok átlagos fokszáma:	4,97	3,45 ±0,24	4,34 ±0,1	5,93
Pontpárok közötti legrövidebb utak hosszának átlaga:	5,74	8,86 ±1,16	6,56 ±0,35	4,2
A hálózat átmérője:	15	24,61 ±3,95	18,13 ±2,2	11
Modulok száma:	23	21,97 ±2,67	23,66 ±2,4	8
Modulok lényegi darabszáma:	6,16	5,5 ±0,95	5,61 ±0,63	2,11
Átlagos modul méret:	12,78	8,78 ±1,1	11 ±1,1	23,75
Átlagos lényegi modul méret:	8,07	5,81 ±0,46	7,14 ±0,49	11,34
A pontok átlagos modul átfedés értéke:	1,69	1,49 ±0,	1,28 ±0,08	1,4

5. táblázat: Az *E. coli* és *B. aphidicola* baktériumok metabolikus hálózatának fontosabb strukturális és modulszerkezeti tulajdonságai, kiegészítve a méretbeli különbségek ellensúlyozására készült 1000-1000 véletlenszerű részháló hasonló paramétereivel (az utóbbi oszlopoknál az átlag és a szórás értékek szerepelnek).

A ModuLand plug-in segítségével meghatározott modulszerkezetet összevetettük egy korábbi, az *E. coli* metabolikus hálózatának modul szerkezetét vizsgáló cikk (Guimerà és Amaral, 2005) eredményeivel. A két módszer által azonosított modul szerkezet szignifikánsan átfedett egymással²⁷. Szintén érdekes kérdés, hogy hogyan oszlanak el a metabolikus funkciók a modulok között, és hogy van-e különbség ilyen szempontból a két hálózat között. Az általunk végzett számítások eredményei azt mutatták, hogy a *B. aphidicola* egy átlagos

²⁷ A szignifikancia meghatározására Fisher-féle egzakt próbát alkalmaztunk: $p=1,4 \times 10^{-7}$

modulja szignifikánsan több²⁸ metabolikus funkcióban vesz részt, mint egy átlagos *E. coli* modul. Az állítás igaz maradt akkor is, amikor a *B. aphidicola* moduljait a véletlenszerűen előállított *E. coli* részhálókkal hasonlítottuk össze, azaz a különbség nem a hálózatok méretéből adódott. Ez a megállapításunk szintén összhangban volt a korábbi publikációkkal (Guimerà és Amaral, 2005; Parter és mtsai, 2007), amelyek a két élőlény moduljainak tisztaságát vizsgálták a KEGG²⁹ útvonal adatbázis (Kanehisa és Goto, 2000) segítségével.

6.4 Hálózat vizualizációs eljárások

A gráfok két vagy három dimenziós ábrázolása (a pontok koordinátáinak meghatározása) informatikai szempontból igen bonyolult kihívás, akár csak néhány ezer pontot tartalmazó hálózat esetében is. Az emberi szem számára szépen strukturált ábrák automatikus generálására sok tucat lényegesen különböző algoritmikus megközelítés létezik (Tamassia, 2013).

Az egyik kézenfekvő igény, hogy a valamilyen módon egybe tartozó pontok (például funkcionális modulok tagjai) kompakt módon, egymáshoz közel kerüljenek (Bourqui és mtsai, 2007). Egy másik megközelítés, hogy az élek hosszát vagy az egymást keresztező élek számát minimalizáljuk. Biológiai hálózatok ábrázolásánál mind a két előbb említett kritérium miatt hasznos lehet a hálózat modulszerkezetéből kiindulni. Egyrészt ahogyan a 6. fejezet számos példáján láttuk, a modulok sokszor valóban funkcionális egységeknek feleltethetők meg. Másrészt pedig a modulok intuitív definíciója szerint pont a lokálisan sűrűn összekötött régiókat jelölik, vagyis a modulok tagjait egymáshoz közel rajzolva rendezettebben ábrázolhatjuk a hálózatot.

A ModuLand plug-in által is kiszámolt átfedő modul-hierarchia különösen alkalmas a gyors és optimális vizualizáció támogatására, hiszen először a felsőbb modulszintek hálózatát ábrázolva optimálisan el lehet helyezni térben a modulokat, míg később az alacsonyabb szint moduljait, vagy a hálózat pontjait a magasabb szinten szereplő modulok pozícióihoz lehet igazítani. Ezt az algoritmikus megközelítést használják a több szintű vizualizációs gráfkiterítő algoritmusok is (Walshaw, 2001; Balzer és Deussen, 2007; Bartel és mtsai, 2010; Dürr és Brandenburg, 2012). A 36. ábrán egy szoftver hálózat segítségével szemléltetem a

28 Bootstrap-módszer, átlagosan 0,53 és 0,67 funkció az *E.coli* és a *B.aphidicola* moduljai esetében. $p=0,0392$

29 A KEGG (Kyoto Encyclopedia of Genes and Genomes) adatbázis honlapja: <http://www.genome.jp/kegg/>

hierarchikus modularizálás segítségével nyerhető hálózatabbrázolás eredményét. Az ábrát Michael Balzer és Oliver Deussen gráf megjelenítési algoritmust leíró tanulmányából másoltam (Balzer és Deussen, 2007).

36. ábra: Egy modularizálás adatait felhasználó több szintű gráfkitérítési eljárás alkalmazása egy szoftverrendszer forráskódját ábrázoló hálózaton. A hálózat több mint 1500 pontja a forráskódban definiált osztályokat, míg a több mint 1800 él az őket jellemző öröklődési kapcsolatokat jelöli. A 126 modult azonosító diszkrét modularizálás segítségével háromdimenziós (**A és B panel**) és kétdimenziós (**C panel**) ábrák készültek. Az ábrák Michael Balzer és Oliver Deussen eredeti közleményéből (Balzer és Deussen, 2007) származnak.

Az előbb említett eljárásokban használt klaszterező algoritmusok nem fedik fel a modulok finomszerkezetét. A ModuLand eljárás család azonban nem csak az átfedő modulhierarchiát építi fel, hanem meghatározza a modulba való tartozás erősségének mértékét is. Ezt az információt a pontok térben való elhelyezésénél kiválóan fel lehet használni, hiszen arányosan közelebb kerülhetnek a pontok azokhoz a modulokhoz, amelyekbe jobban beletartoznak. Érdekes lenne a jövőben kipróbálni a ModuLand és a hozzá hasonló fuzzy modularizálási eljárások használatát a gráf ábrázoló algoritmusokban.

7 Összefoglalás

Az elmúlt másfél évtizedben a biológiai kutatások területénban ugrásszerűen megnőtt a mérési lehetőségek és adatok száma, utat nyitva ezzel a biológia informatikai módszerekkel támogatott rendszerszintű elemzése számára. A hálózatkutatás érdekes és új nézőpontok megjelenését hozta be a biológiába, azáltal hogy a biológiai rendszereket (például a sejt jelátviteli vagy regulációs folyamatait) gráfokkal modellezi, ahol a gráf pontjai biológiai entitásokat (például fehérjéket vagy géneket), míg a gráf élei ezen entitások közötti fizikai vagy logikai kölcsönhatásokat fejeznek ki.

A természetes rendszerek túl fejlettek és ellenállóak ahhoz, hogy szerkezeti felépítésük véletlenszerű legyen. Ezért a komplex biológiai, társadalmi vagy akár üzleti folyamatokat modellező hálózatok vizsgálata során az egyik legfontosabb feladat a hálózat szerkezetének feltérképezése. Ezen hálózatokra általában jellemző az egymással szorosan együttműködő pontok modulokba, sűrű hálózati régiókba való csoportosulása. Például egy sejt fehérje-fehérje kölcsönhatási hálózatában ilyen csoport lehet a riboszómát alakító, vagy az RNS poszttranszkripció módosulásáért felelős fehérjék funkcionális csoportja. A nagy, akár több ezer vagy több tízezer pontot tartalmazó komplex hálózatokban a csoportok megtalálása bonyolult feladat, amelyre több mint száz eltérő megoldást adnak a szakirodalomban (Fortunato, 2010). Ezen megoldások túlnyomó többsége a hálózat egy adott pontját csak egyetlen modulba sorolja be. Ennél több információt nyújtanak az átfedő modularizálási algoritmusok (ahol a pontok több modul tagjai lehetnek egyszerre), illetve a fuzzy eljárások (ahol rendelkezésre áll az az információ, hogy az adott pont melyik modulba milyen mértékben tartozik bele).

A disszertációm alapját képző önálló tudományos munkámban a ModuLand fuzzy modularizációs eljárás családját (Kovács és mtsai., 2010) implementálásával és annak alkalmazásával foglalkoztam. A kutatócsoportom tagjaival közösen definiált módszert egy Cytoscape nevű, hálózat ábrázoló és elemző keretrendszerbe építettem³⁰. Az így nyert, biológusok számára kényelmesen használható programot egy nemzetközi bioinformatikai

30 A ModuLand plug-in letölthető a következő honlapról: <http://www.linkgroup.hu/modules.php>

folyóiratban megjelent megosztott első szerzős cikként publikáltam (Szalay-Bekő *et al.*, 2012). Az implementáció során C++ és Java programozási nyelveket használva több optimalizációt is megvizsgáltam és bevezettem a programok futásidejének csökkentése érdekében.

A plug-in-t számos biológiai kutatás során felhasználtam részben én, részben a kutatócsoportom tagjai és részben más nemzetközi kutatócsoportok. Például az *E.coli* baktérium Met-tRNS szintetáz fehérjéjének térszerkezetét modellező aminosav hálózat esetében megvizsgáltam az enzim katalitikus központja és antikodon kötőhelye közötti konformációs változások továbbadásáért felelős aminosavak modulszerkezetben betöltött pozícióját és ezek sajátosságait. A ModuLand plug-in használatával munkatársaimmal közösen összehasonlítottuk a *B. aphidicola* és *E. coli* baktériumok metabolikus folyamatait leíró hálózatokat, kimutatva az előbbi élőlénynél a szabadon élő, illetve az utóbbi esetében a szimbióta életmódból fakadó különbségeket a metabolikus hálók modulszerkezeteiben.

Az általam megvalósított ModuLand plug-in-t 2012-es publikálása óta több mint 150 kutató töltötte le és több nemzetközi publikációban közölt kutatásnál felhasználásra került. Például a ModuLand plug-in segítségével vizsgálták (Lehtinen és mtsai., 2013) az oxidatív stressz élesztőre gyakorolt rendszerszintű hatását, míg egy másik tanulmány pedig különböző fajok fehérje-fehérje kölcsönhatási hálózatának illesztését oldotta meg a ModuLand plug-in használatával, olyan módon hogy a modulszerkezet minél kevésbé sérüljön a fajok között (Wang és Gao, 2012).

Meggyőződésem, hogy a ModuLand Cytoscape plug-in, illetve a jövőben elkészülő hozzá hasonló, átfedő modulok finomszerkezetét feltáró programok a jövőben több érdekes biológiai és gyógyszerkutatási kérdést segítenek megválaszolni. A korábban említett konkrét példákon túl a plug-in a jövőben felhasználható lehet például fehérjék funkció jóslásánál, jelátviteli utak keresztcsatlakozásainak (cross-talk) detektálásánál, gyógyszercélpont fehérjék azonosításánál, de akár komplex biológiai rendszerek áttekinthető ábrázolásánál is.

8 Summary

In the last fifteen years, the rapid development of high-throughput technologies and the increased amount of available experimental data enabled the emergence of systematic informatics algorithms and methods analyzing complex biological systems. Network science brings interesting and new aspects to the field of biology by constructing graph models for many of these systems, like signaling or regulation processes of the cells. In these graphs the nodes represent biological entities (such as proteins or genes), while the links are modeling physical interactions or logical associations between these entities.

The real natural systems (like biological, social or even economical systems) are too successful and robust to have a random structure. Therefore, one of the key challenges of network science is to analyze and understand the structure and topology of these networks. In case of the most of natural networks, the interacting nodes tend to form modules, dense regions. For example, in the protein interaction network of a cell such module can be the functional group of proteins responsible for forming ribosomes, or involved in RNA splicing. It is a hard algorithmic task to find these modules in large networks containing thousands or ten thousands of nodes. More than a hundred significantly different methods have been published to solve this problem (Fortunato, 2010). However, most of these methods assign a given node into only one given module. More details of the module structure are revealed by the overlapping modularization methods, which are capable to assign a given node into multiple modules. Fuzzy methods are producing even more information by also determining the strength of module assignment.

In my PhD research I was working on the implementation and application of the ModuLand fuzzy modularization method family (Kovács *et al.*, 2010), defined by the members of my research group. I integrated³¹ the method into the widely used open source Cytoscape program, which provides a network visualization and analysis framework running on Windows, Linux and Mac operating systems. The Cytoscape ModuLand plug-in was published in the international Bioinformatics journal (Szalay-Bekő *et al.*, 2012). During the

31 The ModuLand plug-in can be downloaded from the following site: <http://www.linkgroup.hu/modules.php>

implementation I used C++ and Java programming languages, I introduced several optimizations to reduce the run-time of the modularization and implemented several measure calculation and visualization features to help the analysis of biological networks.

The plug-in was successfully used for bioinformatics research projects many times by me, by the members of my research group and also by independent groups. For example, I used the plug-in to analyze the protein structure network of the Met-tRNA synthetase protein of *E. coli* bacteria. The ModuLand plug-in revealed the interesting modular positions of the amino acids responsible for transferring the conformational changes between the catalytic center and the anticodon binding site of the enzyme. Also with the help of the plug-in, together with my colleagues we compared the metabolic network of the symbiont *B. aphidicola* and free-living *E. coli* bacteria and showed the differences of the module structures caused by the different levels of environmental stability.

Since I developed the ModuLand plug-in and the related publication in 2012, more than 150 researchers downloaded the program and used it for several internationally published research projects. For example, the ModuLand plug-in was used to analyze the system level effects of oxidative stress on yeast cells (Lehtinen *et al.*, 2013), while an other study used the plug-in to align protein interaction networks of different species based on their module structures (Wang and Gao, 2012).

Based on these initial verifications of the use of the program, the ModuLand Cytoscape plug-in (and the other fuzzy modularization methods) will help to answer several interesting questions in the fields of biology and pharmacology in the future. Beside the previously mentioned examples, the plug-in also can be used to predict cross-talk regions in signaling networks, to identify possible drug target proteins, or even to help in the structured visualization of complex biological systems.

Köszönetnyilvánítás

Köszönöm a Szegedi Tudományegyetemnek, a Biológus Doktori Iskolának és a disszertációm bírálóinak a fokozatszerzési eljárás során nyújtott segítséget.

Ezúton köszönöm témavezetőim segítségét! Különösen hálás vagyok **Csermely Péter professzor úrnak**, aki az elmúlt tíz évben sok különböző (elnöki, mentori, főnöki, témavezetői és különösen baráti) minőségben segített. Szintén köszönöm **Dr. Papp Balázs** segítségét a fokozatszerzéssel és publikációkkal kapcsolatos tudományos és adminisztratív munkákban.

Köszönettel tartozom **Dr. Korcsmáros Tamásnak** a kutatási munkámmal és a fokozatszerzési eljárással kapcsolatos folyamatos és szakszerű tanácsaiért. Szeretném megköszönni a Csermely Péter professzor úr által vezetett **LinkGroup hálózatkutató csoport** tagjainak a sokéves közös munkát és együttgondolkodást. A PhD disszertációmban bemutatott biológiai elemzéseket és informatikai munkák egy részét a csoport számos tagja segítette. Ezúton köszönöm **Dr. Kovács Istvánnak**, a ModuLand eljáráscsalád kitalálójának segítségét és a fuzzy modularizálás területéről folytatott számos érdekes beszélgetést. Szeretném megköszönni **Palotai Robin**, **Szuromi Gábor** és **Zalányi Balázs** informatikusok segítségét, akik részt vettek a ModuLand programcsomag eredeti, linux operációs rendszeren futó parancssoros verziójának megvalósításában. Szintén köszönöm **Szappanos Balásznak** a *B. aphidicola* és *E. coli* baktériumok metabolikus hálózatainak elemzésében nyújtott segítségét.

Köszönettel tartozom az **Ericsson Magyarország Kft-nek** amiért a fokozatszerzésem tanulmányi szabadsággal és az eljárási díj egy részének átvállalásával segítette.

Végül és főként köszönöm családomnak, hogy az elmúlt években támogattak a bioinformatikusi tevékenységemet és elviselték az informatikai munkáim mellett kutatásra fordított túlórák sokaságát. Köszönöm szüleimnek, testvéreimnek és főként feleségemnek és fiamnak, **Melindának** és **Márknak** a támogatást!

Saját publikációk jegyzéke

2013. november 2.³²

Összesített impakt: 50

Független idézetek száma: 187

Házasságkötés előtt használt publikációs név: **Szalay, M.S.**

A disszertációhoz kapcsolódó közlemények

1. Szalay-Bekő, M., Palotai, R., Szappanos, B., Kovács, I.A., Papp, B., Csermely, P. (2012) ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping modules and community centrality. *Bioinformatics*, **28**, 2202-2204, IF: 5,3 <http://arxiv.org/abs/1111.3033> – 6 független idézet
2. Kovács, I.A., Palotai, R., Szalay, M.S., Csermely, P. (2010) Community landscapes: a novel, integrative approach for the determination of overlapping network modules. *PloS ONE* **7**, e12528, IF: 3,7 www.arxiv.org/abs/0912.0161-- 26 független idézet

A disszertációtól független közlemények

Eredeti tudományos közlemények

1. Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálffy, M., Dúl, Z., Zsákai, L., Szalay-Bekő, M., Lenti, K., Farkas, I.J., Vellai, T., Csermely, P., Korcsmáros, T. (2013) SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Systems Biology* **7**(1):7, IF: 3 – 1 független idézet
2. Korcsmáros, T. Szalay, M.S., Rovó, P., Palotai, R., Fazekas, D., Lenti, K., Farkas, I.J. Csermely, P., Vellai, T. (2011) Signalogs: orthology-based identification of novel signaling pathway components in three metazoans. *PLoS ONE* **8**, e19240, IF: 3,7 – 2 független idézet

³² A független idézetek és az *Impact Factor* adatok kigyűjtéséhez a *Web of Science* online eszközt használtam. (<http://www.webofknowledge.com>)

3. Korcsmáros, T., Farkas, I.J., Szalay, M. S., Rovó, P., Fazekas, D., Spiró, Z., Böde, C., Lenti, K., Vellai, T., Csermely, P. (2010) Uniformly curated signaling pathways reveal tissue-specific cross-talks, novel pathway components, and drug target candidates. *Bioinformatics* **26**, 2042-2050, IF: 5,3 www.signalink.org – 10 független idézet
4. Wang, S., Szalay, M.S., Zhang, C., Csermely, P. (2008) Learning and innovative elements of strategy update rules expand cooperative network topologies. *PLoS ONE* **3**, e1917, IF: 3,7 www.arxiv.org/0708.2707 -- 21 független idézet

Szabadalmak

1. Korcsmáros T., Szalay-Bekő M., Palotai R., Szuromi G., Fazekas D., Dunai Zs. (2011) Eljárás és számítógépes rendszer gyógyszerhatóanyagok hatásmechanizmusának szimulációjára. Magyarországi szabadalmi bejelentés, P1100368
2. Szalay, M., Stanojevic, O., Farkas, L. (2010) Automatic use of behavioral information for promotional purposes in communications system. Nemzetközi Ericsson PCT szabadalmi bejelentés, PCT/SE2010/051312
3. Kovács, I.A., Csermely, P., Szalay, M.S., Korcsmáros, T. (2006) Method for analyzing the fine structure of networks. Nemzetközi PCT szabadalmi bejelentés, PCT/IB2007/05047

Összefoglaló munkák

1. Farkas, I.J., Korcsmáros, T., Kovács, I.A., Mihalik, Á., Palotai, R., Simkó, G.I., Szalay, K.Z., Szalay-Bekő, M., Vellai, T., Wang, S., Csermely, P. (2011) Network-based tools in the identification of novel drug-targets. *Science Signaling* **4**, pt3, IF: 7,6 -- 3 független idézet
2. Palotai, R. Szalay, M.S., Csermely, P. (2008) Chaperones as integrators of cellular networks: changes of cellular integrity in stress and diseases. *IUBMB Life* **60**, 10-18, arxiv.org/0710.1622, IF: 2,8 -- 24 független idézet
3. Korcsmáros, T., Szalay, M.S., Böde, C., Kovács, I.A., Csermely, P. (2007) How to

design multi-target drugs: Target-search options in cellular networks. *Expert Op. Drug Discov.* **2**, 799-808, arxiv.org/q-bio.MN/0703010, IF: 2,3 -- 20 független idézet

4. Böde. C., Kovács, I.A., Szalay, M.S., Palotai, R. Korcsmáros, T., Csermely, P. (2007) Network analysis of protein dynamics. *FEBS Lett.* **581**, 2776-2782, arxiv.org/q-bio.BM/0703025, IF: 3,6 -- 51 független idézet
5. Szalay, M.S., Kovács, I.A., Korcsmáros, T., Böde. C., Csermely, P. (2007) Stress-induced rearrangements of cellular networks: consequences for protection and drug design. *FEBS Lett.* **581**, 3675-3680, arxiv.org/q-bio.MN/0702006, IF: 3,6 -- 18 független idézet
6. Korcsmáros, T., Kovács, I.A., Szalay, M.S., Csermely, P. (2007) Molecular chaperones: the modular evolution of cellular networks. *J. Biosci.* **32**, 441-446, arxiv.org/q-bio.MN/0701030, IF: 1,8 -- 14 független idézet
7. Kovacs, I.A., Szalay, M.S., Csermely, P. (2005) Water and molecular chaperones act as weak links of protein folding networks: energy landscape és punctuated equilibrium changes point towards a game theory of proteins. <http://arxiv.org/abs/q-bio.BM/0409030>, *FEBS Lett.* **579**, 2254-2260, IF: 3,6 -- 21 független idézet

Konferencia előadások és poszterek

1. Fazekas, D., Dénes Türei, Módos, D., Mihály Koltai, Pálffy, M, Máté Szalay-Bekő, Lenti, K., Farkas, I. J., Vellai, T., Csermely, P., Korcsmáros, T. (2012) Extending signaling pathways – regulatory layers of signaling networks in *C. elegans*, *D. melanogaster* and humans; Poster, EMBO Conference Series From Functional Genomics to Systems Biology, Heidelberg, Germany
2. Fazekas, D., Koltai, M., Dúl, Z., Borda, J., Pálffy, M, Módos, D., Szalay-Bekő, M., Vellai, T., Csermely, P., Farkas, I. J., Korcsmáros, T. (2011) Regulating, an integrated database of the regulation of signaling; Poster, NetSci - The International School and Conference on Network Science, Budapest, Hungary
3. Farkas, I. J., Korcsmáros, T., Fazekas, D., Szalay-Bekő, M., Petra, R., Spiró, Z., Böde,

- C., Lenti, K., Vellai, T., Csermely, P. (2011) A resource and a tool for signaling pathway analysis; Poster, International Union of Pure and Applied Physics, San Diego, USA
4. Koltai, M., Fazekas, D., Dúl, Z., Borda, J., Pálffy, M, Módos, D., Szalay-Bekő, M., Vellai, T., Csermely, P., Farkas, I. J., Korcsmáros, T. (2011) Regulink, an integrated database of the regulation of signaling; Poster, IX. Hungarian Conference on Biometry, Biomathematics, and Bioinformatics, Budapest, Hungary
 5. Fazekas, D., Koltai, M., Dúl, Z., Borda, J., Pálffy, M, Módos, D., Szalay-Bekő, M., Vellai, T., Csermely, P., Farkas, I. J., Korcsmáros, T. (2011) Regulink, an integrated database of the regulation of signaling; Poster, 4th European Conference on Chemistry for Life Sciences, Budapest, Hungary
 6. Farkas, I. J., Korcsmáros, T., Fazekas, D., Szalay, M.S., Petra, R., Spiró, Z., Böde, C., Lenti, K., Vellai, T., Csermely, P. (2010) Uniform curation of metazoan signaling pathways – resource and tool for experiment design and evaluation; Poster, Conference on multi-scale dynamics and evolvability of biological networks, Leipzig, Germany
 7. Korcsmáros, T., Farkas, I. J., Fazekas, D., Szalay, M.S., Petra, R., Spiró, Z., Böde, C., Lenti, K., Vellai, T., Csermely, P. (2010) Uniform curation of metazoan signaling pathways – resource and tool for experiment design and evaluation; ICSB 2010, Edinburgh, UK
 8. Korcsmáros, T., Szalay, M.S., Fazekas, D., Petra, R., Zoltan Spiro, Vellai, T., Csermely, P. (2009) SignaLink, a signaling pathway database: theoretical and practical applications; Poster, FEBS SysBio Workshop, Alpbach, Austria
 9. Korcsmáros, T., Farkas, I. J., Szalay, M.S., Petra, R., Fazekas, D., Spiró, Z., Böde, C., Lenti, K., Vellai, T., Csermely, P. (2009) A signaling pathway resource for cross-talk analysis in *C.elegans*, *Drosophila*, and humans; Poster, ICSB 2009, Stanford, USA
 10. Korcsmáros, T., Farkas, I. J., Szalay, M.S., Petra, R., Fazekas, D., Spiró, Z., Böde, C., Lenti, K., Vellai, T., Csermely, P. (2009) A signaling pathway resource for cross-talk analysis in *C.elegans*, *Drosophila*, and humans; Poster, FEBS Protein Modules

Workshop, Seefeld in Tirol, Austria

11. Csermely, P., Korcsmáros, T., Kovács, I.A., Szalay M.S., Sőti, C. (2008) Systems biology of molecular chaperone networks. In: The biology of extracellular molecular chaperones. *Novartis Foundation Symposium Series* Vol. **291**, pp. 45-58 -- 5 független idézet
12. Palotai, R., Wang, S., Zhang, C., Szalay, M.S., Csermely, P. (2007) Learning and innovation expand cooperative network topologies; Poster, Third Yamada Symposium, Japan
13. Palotai R., Kovács, I.A., Szalay, M.S., Csermely P. (2007) ModuLand, a powerful and extensive family of methods to uncover the overlapping modules and hierarchy of complex networks; Poster, Third Yamada Symposium, Japan
14. Csermely P., Kovács, I.A., Szalay, M.S. (2007) Modular structure, learning és innovation of networks; Conference talk: International Workshop on Complex Systems és Networks, Sovata, Romania

Ábrajegyzék

1. ábra: Elterjedten használt gráf típusok.....	6
2. ábra: Sejtműködés leírására elterjedten alkalmazott biológiai hálózat típusok.....	9
3. ábra: Biológiai hálók dinamikáját leíró Boole-háló illusztrációja.....	15
4. ábra: A Zachary-féle szociológiai hálózat (Zachary, 1977).....	17
5. ábra: Egy példa hálózat hierarchikus modularizálása és a modularizáláshoz tartozó dendrogram.....	18
6. ábra: Egy négy modult tartalmazó teszt hálózat spektrális modul információinak szemléltetése. Az ábra forrása: Donetti és Muñoz, 2004.....	22
7. ábra: A hálózat tömörítéses módszer alapgondolata. Az ábrát közlő eredeti cikk: Rosvall és Bergstrom, 2007.....	25
8. ábra: A k-klikk közösségek módszerének szemléltetése egy példa hálózaton. Forrás:Palla és mtsai, 2005.....	27
9. ábra: A diszkrét, az átfedő és a fuzzy modularizálási módszerek által nyújtott információ összehasonlítása.....	29
10. ábra: A ModuLand eljárás fogalmainak szemléltetése az <i>E. coli</i> baktériumban található Met-tRNS szintetáz fehérje aminosav hálózatán.....	34
11. ábra: Példák a LinkLand algoritmus által definiált indirekt hatás terjedésére.....	35
12. ábra: A teszhálózat éleinek LinkLand algoritmus által meghatározott modulcentralitás értékei.....	36
13. ábra: Az arányos besoroló működésének szemléltetése egy példa hálózat részletén.....	38
14. ábra: Az arányos besoroló algoritmus eredménye a példa hálózat élein.....	39
15. ábra: Az arányos besoroló algoritmus eredménye a példa hálózat pontjain.....	39
16. ábra: A Met-tRNS szintetáz fehérje szerkezeti hálózat modul hierarchiája.....	40
17. ábra: Lényegi darabszám jelentésének illusztrálása.....	42
18. ábra: A ModuLand plug-in működése a Met-tRNS szintetáz protein aminosav hálózatának modularizálása közben.....	44
19. ábra: A ModuLand Cytoscape plug-in felépítése.....	45
20. ábra: Az élsúlyok kiválasztására szolgáló dialógus ablak.....	47
21. ábra: A modularizálást irányító fő dialógus ablak.....	48
22. ábra: Példák modulszerkezeti adatok exportálására.....	50

23. ábra: Mérőszámok exportálása.....	51
24. ábra: Átfogó jelentés a modulszerkezet legfontosabb tulajdonságairól.....	53
25. ábra: AModuLand plug-in által kínált vizualizációs megoldások bemutatása a Met-tRNS szintetáz protein aminosav hálóján.....	54
26. ábra: Adott szintnél jobban korreláló modulok összevonása.....	55
27. ábra: Az elosztott gráf algoritmusok futtatására szolgáló keretrendszer felépítése.....	63
28. ábra: Az elosztott gráf algoritmusok futásának illusztrálása.....	64
29. ábra: Az elosztott LinkLand algoritmus futásidejének összehasonlítása az egy gépes változattal, a futtatást végző kliensek számának függvényében.....	65
30. ábra: Az elosztott LinkLand algoritmus futásidejének összehasonlítása az egy gépes változattal, a vizsgált hálózat méretének függvényében.....	66
31. ábra: A ModuLand Cytoscape plug-in segítségével kiszámított modul centralitás értékek vizualizálása kézzel beállított színskála alkalmazásával egy élesztő fehérje hálóján.....	68
32. ábra: Az élesztő fehérje kölcsönhatási hálójában definiált date és party hub-ok elkülönítése az átfedő modularizálás eredményeinek felhasználásával. Forrás: Kovács és mtsai, 2010.....	69
33. ábra: Biomolekuláris hálózatokban funkcionális modulok között elhelyezkedő, gyógyszerkutatási szempontból érdekes elemek szemléltetése.....	71
34. ábra: A Met-tRNS szintetáz fehérje aminosav hálójában a pontok modul átfedés és hídság mérőszámainak eloszlása.....	76
35. ábra: Az <i>E. coli</i> és <i>B. aphidicola</i> baktériumok metabolikus hálójata.....	79
36. ábra: Egy modularizálás adatait felhasználó több szintű gráfkitérítési eljárás alkalmazása egy szoftverrendszer forráskódját ábrázoló hálóján. Forrás: Balzer és Deussen, 2007.....	82

Irodalomjegyzék

1. Agarwal,G., Kempe,D. (2008) Modularity-maximizing graph communities via mathematical programming. *The European Physical Journal B*, **66**, 409–418.
2. Ahn,Y.-Y., Bagrow,J.P., Lehmann,S. (2010) Link communities reveal multiscale complexity in networks. *Nature*, **466**, 761–764.
3. Albert,R. (2005) Scale-free networks in cell biology. *Journal of Cell Science*, **118**, 4947–4957.
4. Albert,R., Barabasi,A.-L. (2001) Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**, 78.
5. Andrade,R.F.S., Rocha-Neto,I.C., Santos,L.B.L., De Santana,C.N., Diniz,M.V.C., Lobão,T.P., Goés-Neto,A., Pinho,S.T.R., El-Hani,C.N. (2011) Detecting network communities: an application to phylogenetic analysis. *PLoS Computational Biology*, **7**, e1001131.
6. Arenas,A., Diaz-Guilera,A., Perez-Vicente,C.J. (2006) Synchronization reveals topological scales in complex networks. *Physical Review Letters*, **96**, 114102.
7. Arenas,A., Fernandez,A., Gomez,S. (2007) Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, **10**, 23.
8. Bader,G.D., Hogue,C.W. V (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
9. Balzer,M., Deussen,O. (2007) Level-of-detail visualization of clustered graph layouts. In: *APVIS 2007 - 6th International Asia-Pacific Symposium on Visualization*, 133-140.
10. Bantscheff,M., Lemeer,S., Savitski,M.M., Kuster,B. (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*, **404**, 939–965.
11. Barabasi,A.-L., Albert,R. (1999) Emergence of scaling in random networks. *Science*,

286, 11.

12. Barabási,A.-L., Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**, 101–113.
13. Bartel,G., Gutwenger,C., Klein,K., Mutzel,P. (2010) An experimental evaluation of multilevel layout methods. In: *Graph Drawing'10.*, pp. 81–90.
14. Baumes,J., Goldberg,M., Magdon-Ismail,M. (2005) Efficient identification of overlapping communities. In: *IEEE International Conference on Intelligence and Security Informatics ISI 2005*, **3495**, 27–36.
15. Becker,E., Robisson,B., Chapple,C.E., Guénoche,A., Brun,C. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, **28**, 84–90.
16. Begley,C.G., Ellis,L.M. (2012) Drug development: Raise standards for preclinical cancer research. *Nature*, **483**, 531–533.
17. Berg,J., Lässig,M. (2006) Cross-species analysis of biological networks by Bayesian alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 10967–10972.
18. Berg,J., Lässig,M. (2004) Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14689–14694.
19. Berg,J., Lässig,M., Wagner,A. (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and. *BMC Evolutionary Biology*, **4**, 51.
20. Berkhin,P. (2006) A Survey of Clustering Data Mining Techniques. *Grouping Multidimensional Data*, **C1**, 25–71.
21. Bertin,N., Simonis,N., Dupuy,D., Cusick,M.E., Han,J.-D.J., Fraser,H.B., Roth,F.P., Vidal,M. (2007) Confirmation of organized modularity in the yeast interactome. *PLoS Biology*, **5**, e153.
22. Bezdek,J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*.

Kluwer Academic Publishers Norwell, MA, USA.

23. Bishop,C.M. (1995) *Neural networks for pattern recognition*. Oxford University Press, UK.
24. Blanchard,P., Chang,C.-H., Krüger,T. (2003) Epidemic thresholds on scale-free graphs: the interplay between exponent and preferential choice. *Annales Henri Poincaré*, **4**, 957–970.
25. Boccaletti,S., Ivanchenko,M., Latora,V., Pluchino,A., Rapisarda,A. (2007) Detecting complex network modularity by dynamical clustering. *Physical Review E*, **75**, 045102.
26. Bourqui,R., Auber,D., Mary,P. (2007) How to draw clustered weighted graphs using a multilevel force-directed graph drawing algorithm. *11th International Conference Information Visualization IV 07*, 757–764.
27. Brown,P.O., Botstein,D. (1999) Exploring the new world of genome with DNA microarrays. *Nature Genetics*, **21**, 33–37.
28. Bunnage,M.E. (2011) Getting pharmaceutical R&D back on target. *Nature Chemical Biology*, **7**, 335–339.
29. Burt,R.S. (1976) Positions in networks. *Social Forces*, **55**, 93–122.
30. Chen,J., Yuan,B. (2006) Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, **22**, 2283–2290.
31. Chen,L., Wang,R.-S., Zhang,X.-S. (2009) *Biomolecular Networks*. John Wiley & Sons, Inc., Hoboken, New Jersey
32. Cho,D.-Y., Kim,Y.-A., Przytycka,T.M. (2012) Chapter 5: Network Biology Approach to Complex Diseases. *PLoS Computational Biology*, **8**, e1002820.
33. Chowdhury,A.R., Chetty,M., Vinh,N.X. (2013) Incorporating time-delays in S-System model for reverse engineering genetic networks. *BMC Bioinformatics*, **14**, 196.
34. Clauset,A., Newman,M.E.J., Moore,C. (2004) Finding community structure in very large networks. *Physical Review E*, **70**, 066111.

35. Csermely,P., Korcsmáros,T., Kiss,H.J.M., London,G., Nussinov,R. (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & Therapeutics*, **138**, 333–408.
36. Csermely,P., Sandhu,K.S., Hazai,E., Hoksza,Z., Kiss,H.J.M., Miozzo,F., Veres,D. V, Piazza,F., Nussinov,R. (2011) Disordered proteins and network disorder in network descriptions of protein structure, dynamics and function. Hypotheses and a comprehensive review. *Current Protein Peptide Science*, **13**, 19-33.
37. Dean,J., Ghemawat,S. (2004) MapReduce : Simplified data processing on large clusters. In: *6th Symposium on Operating System Design and Implementation (OSDI'04)*., pp. 137–149.
38. Delmotte,A., Tate,E.W., Yaliraki,S.N., Barahona,M. (2011) Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction. *Physical Biology*, **8**, 055010.
39. Delvenne,J.-C., Yaliraki,S.N., Barahona,M. (2010) Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 12755–12760.
40. Donetti,L., Muñoz,M.A. (2004) Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, **2004**, P10012.
41. Dunn,J.C. (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybernetics and Systems*, **3**, 32–57.
42. Dürr,O., Brandenburg,A. (2012) Using community structure for complex network layout. <http://arxiv.org/abs/1207.6282>
43. Eisenberg,E., Levanon,E.Y. (2003) Preferential attachment in the protein network evolution. *Physical Review Letters*, **91**, 138701.
44. Ekman,D., Light,S., Björklund,Å.K., Elofsson,A. (2006) What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? *Genome Biology*, **7**, R45.

45. Enright,A.J., Van Dongen,S., Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**, 1575–1584.
46. Estrada,E. (2006) Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, **6**, 35–40.
47. Farkas,I.J., Korcsmáros,T., Kovács,I.A., Mihalik,Á., Palotai,R., Simkó,G.I., Szalay,K.Z., Szalay-Bekő,M., Vellai,T., Wang,S., Csermely,P. (2011) Network-based tools for the identification of novel drug targets. *Science Signaling*, **4**, pt3.
48. Feist,A.M., Henry,C.S., Reed,J.L., Krummenacker,M., Joyce,A.R., Karp,P.D., Broadbelt,L.J., Hatzimanikatis,V., Palsson,B.Ø. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, **3**, 121.
49. Fiduccia,C.M., Mattheyses,R.M. (1988) A linear-time heuristic for improving network partitions. In: *Papers on Twenty-five years of electronic design automation - 25 years of DAC*. ACM Press, New York, USA, pp. 241–247.
50. Fiedler,M. (1973) Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, **23**, 298–305.
51. Fortunato,S. (2010) Community detection in graphs. *Physics Reports*, **486**, 75–174.
52. Fortunato,S., Latora,V., Marchiori,M. (2004) Method to find community structures based on information centrality. *Physical Review E*, **70**, 056104.
53. Fox,A.D., Hescott,B.J., Blumer,A.C., Slonim,D.K. (2011) Connectedness of PPI network neighborhoods identifies regulatory hub proteins. *Bioinformatics*, **27**, 1135–1142.
54. Fraser,I.D., Germain,R.N. (2009) Navigating the network: signaling cross-talk in hematopoietic cells. *Nature Immunology*, **10**, 327–331.
55. Fredman,M.L., Tarjan,R.E. (1987) Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM*, **34**, 596–615.
56. Freeman,L. (1977) A set of measures of centrality based on betweenness. *Sociometry*, **40**, 35–41.

57. Frey,B.J., Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.
58. Friedman,D.B. (2007) Quantitative proteomics for two-dimensional gels using difference gel electrophoresis. *Methods in Molecular Biology*, **367**, 219–239.
59. Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.
60. Ghosh,A., Vishveshwara,S. (2007) A study of communication pathways in methionyl-tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 15711–15716.
61. Girvan,M., Newman,M.E.J. (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7821–7826.
62. Gregory,S. (2007) An algorithm to find overlapping community structure in networks. *Knowledge Discovery in Databases PKDD 2007*, **4702**, 91–102.
63. Gregory,S. (2009a) Finding overlapping communities in networks by label propagation. *New Journal of Physics*, **12**, 103018.
64. Gregory,S. (2009b) Finding overlapping communities using disjoint community detection algorithms. *Complex Networks*, **207**, 47–61.
65. Gregory,S. (2011) Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2011**, P02017.
66. Grendar,M. (2006) Entropy and effective support size. *Entropy*, **8**, 169–174.
67. Guimerà,R., Amaral,N.L.A. (2005) Functional cartography of complex metabolic networks. *Nature*, **433**, 895–900.
68. Habib,M., Paul,C. (2010) A survey on algorithmic aspects of modular decomposition. *Computer Science Review*, **4**, 41–59.
69. Hallikas,O., Taipale,J. (2006) High-throughput assay for determining specificity and

affinity of protein-DNA binding interactions. *Nature Protocols*, **1**, 215–222.

70. Han,J.-D.J., Bertin,N., Hao,T., Goldberg,D.S., Berriz,G.F., Zhang,L. V, Dupuy,D., Walhout,A.J.M., Cusick,M.E., Roth,F.P., Vidal,M. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.
71. Harris,L.W., Lockstone,H.E., Khaitovich,P., Weickert,C.S., Webster,M.J., Bahn,S. (2009) Gene expression in the prefrontal cortex during adolescence: implications for the onset of schizophrenia. *BMC Medical Genomics*, **2**, 28.
72. Hidalgo,C.A., Rodriguez-Sickert,C. (2008) The dynamics of a mobile phone network, *Physica A*, **387**, 3017–3024.
73. Hoare,C.A.R. (1962) Quicksort. *The Computer Journal*, **5**, 10–16.
74. Hughes,B.D. (1995) *Random walks and random environments: random walks Vol 1*. Clarendon Press, Oxford, UK.
75. Husmeier,D. (2003) Reverse engineering of genetic networks with Bayesian networks. *Biochemical Society Transactions*, **31**, 1516–1518.
76. Husmeier,D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.
77. Hwang,W.-C., Zhang,A., Ramanathan,M. (2008) Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clinical Pharmacology & Therapeutics*, **84**, 563–572.
78. Jain,A.K., Murty,M.N., Flynn,P.J. (1999) Data clustering: a review. *ACM Computing Surveys*, **31**, 264–323.
79. Jeong,H., Mason,S.P., Barabási,A.L., Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
80. Jin,G., Zhang,S., Zhang,X.-S., Chen,L. (2007) Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast. *PLoS ONE*, **2**, e1207.

81. Johnson,S. (1967) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.
82. Kanehisa,M., Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**, 27-30.
83. Kauffman,S. (1969) Homeostasis and differentiation in random genetic control networks. *Nature*, **224**, 177–178.
84. Kelley,B.P., Sharan,R., Karp,R.M., Sittler,T., Root,D.E., Stockwell,B.R., Ideker,T. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 11394–11399.
85. Kernighan,B.W., Lin,S. (1970) An efficient heuristic procedure for partitioning graphs. *Bell Sys. Tech. J.*, **49**, 291–308.
86. Kim,P.M., Lu,L.J., Xia,Y., Gerstein,M.B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, **314**, 1938–1941.
87. Kim,S.Y., Imoto,S., Miyano,S. (2003) Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Briefings in Bioinformatics*, **4**, 228–235.
88. Kim,Y., Jeong,H. (2011) Map equation for link communities. *Physical Review E*, **84**, 026110.
89. Komurov,K., White,M. (2007) Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Molecular Systems Biology*, **3**, 110.
90. Korcsmáros,T., Farkas,I.J., Szalay,M.S., Rovó,P., Fazekas,D., Spiró,Z., Böde,C., Lenti,K., Vellai,T., Csermely,P. (2010) Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Bioinformatics*, **26**, 2042–2050.
91. Kovács,I.A., Palotai,R., Szalay,M.S., Csermely,P. (2010) Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PLoS ONE*, **5**, e12528.
92. Kreimer,A., Borenstein,E., Gophna,U., Ruppín,E. (2008) The evolution of modularity

- in bacterial metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 6976–6981.
93. Lähdesmäki,H., Shmulevich,I., Yli-Harja,O. (2003) On learning gene regulatory networks under the Boolean network model. *Machine Learning*, **52**, 147–167.
 94. Lancichinetti,A., Fortunato,S., Kertész,J. (2009) Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, **11**, 033015.
 95. Lee,C., Reid,F., McDaid,A., Hurley,N. (2010) Detecting highly overlapping community structure by greedy clique expansion. <http://arxiv.org/abs/1002.1827>
 96. Lehmann,S., Hansen,L.K. (2007) Deterministic Modularity Optimization. *The European Physical Journal B*, **60**, 83-88.
 97. Lehtinen,S., Marsellach,F.X., Codlin,S., Schmidt,A., Clément-Ziza,M., Beyer,A., Bähler,J., Orengo,C., Pancaldi,V. (2013) Stress induces remodelling of yeast interaction and co-expression networks. *Molecular Biosystems*, **9**, 1697–1707.
 98. Li,D., Li,J., Ouyang,S., Wang,J., Wu,S., Wan,P., Zhu,Y., Xu,X., He,F. (2006a) Protein interaction networks of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*: large-scale organization and robustness. *Proteomics*, **6**, 456–461.
 99. Li,Z.L.Z., Wang,Y.W.Y., Zhang,S.Z.S., Zhang,X.-S.Z.X.-S., Chen,L.C.L. (2006b) Alignment of protein interaction networks by integer quadratic programming. *Bioinformatics*, **1**, 5527–5530.
 100. Liu,J. (2010) Detecting the fuzzy clusters of complex networks. *Pattern Recognition*, **43**, 1334–1345.
 101. Loscalzo,J., Barabasi,A.-L. (2011) Systems biology and the future of medicine. *Wiley Interdisciplinary Reviews Systems Biology and Medicine*, **3**, 619–627.
 102. Ma,S., Gong,Q., Bohnert,H.J. (2007) An Arabidopsis gene network based on the graphical Gaussian model. *Genome Research*, **17**, 1614–1625.
 103. Ma,X., Gao,L., Yong,X. (2010) Eigenspaces of networks reveal the overlapping and

- hierarchical community structure more precisely. *Journal of Statistical Mechanics: Theory and Experiment*, **2010**, P08012.
104. Maslov,S., Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.
 105. Massen,C., Doye,J. (2005) Identifying communities within energy landscapes. *Physical Review E*, **71**, 046101.
 106. Mihalik,Á., Csermely,P. (2011) Heat shock partially dissociates the overlapping modules of the yeast protein-protein interaction network: a systems level model of adaptation. *PLoS Computational Biology*, **7**, e1002187.
 107. Moran,N.A., Mira,A. (2001) The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biology*, **2**, R54.
 108. Moran,N.A., Wernegreen,J.J. (2000) Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends in Ecology & Evolution*, **15**, 321-326.
 109. Morris,J.H., Apeltsin,L., Newman,A.M., Baumbach,J., Wittkop,T., Su,G., Bader,G.D., Ferrin,T.E. (2011) ClusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, **12**, 436.
 110. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L., Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621–628.
 111. Nepusz,T., Négyessy,L., Bazsó,F. (2008) Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, **77**, 016107.
 112. Nepusz,T., Sasidharan,R., Paccanaro,A. (2010) SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. *BMC Bioinformatics*, **11**, 120.
 113. Newman,A.M., Cooper,J.B. (2010) AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. *BMC Bioinformatics*, **11**, 117.
 114. Newman,M.E.J. (2012) Communities, modules and large-scale structure in networks. *Nature Physics*, **8**, 25–31.

115. Newman,M.E.J. (2004a) Detecting community structure in networks. *The European Physical Journal B - Condensed Matter*, **38**, 321–330.
116. Newman,M.E.J. (2004b) Fast algorithm for detecting community structure in networks. *Physical Review E*, **69**, e066133.
117. Newman,M.E.J. (2006) Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, **74**, 036104.
118. Newman,M.E.J., Girvan,M. (2003) Finding and evaluating community structure in networks. *Physical Review E*, **69**, 026113.
119. Niedringhaus,T.P., Milanova,D., Kerby,M.B., Snyder,M.P., Barron,A.E. (2011) Landscape of next-generation sequencing technologies. *Analytical Chemistry*, **83**, 4327–4341.
120. Nussinov,R. (2012) How do dynamic cellular signals travel long distances? *Molecular Biosystems*, **8**, 22–26.
121. Ogata,H., Fujibuchi,W., Goto,S., Kanehisa,M. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Research*, **28**, 4021–4028.
122. Ong,S.-E., Foster,L.J., Mann,M. (2003) Mass spectrometric-based approaches in quantitative proteomics. *Methods San Diego Calif*, **29**, 124–130.
123. Onnela,J.P., Saramaki,J., Hyvönen,J., Szabó,G., Lazer,D., Kaski,K., Kertész,J., Barabási,A.L. (2007) Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 7332-7336.
124. Palla,G., Barabási,A.-L., Vicsek,T. (2007) Quantifying social group evolution. *Nature*, **446**, 664–667.
125. Palla,G., Derényi,I., Farkas,I., Vicsek,T. (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**, 814–818.
126. Papin,J.A., Hunter,T., Palsson,B.O., Subramaniam,S. (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell*

Biology, **6**, 99–111.

127. Parter,M., Kashtan,N., Alon,U. (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology*, **7**, 169.
128. Pikovsky,A., Rosenblum,M., Kurths,J. (2003) *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press.
129. Pons,P., Latapy,M. (2006) Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, **10**, 191-218.
130. Psorakis,I., Roberts,S., Ebden,M., Sheldon,B. (2011) Overlapping community detection using Bayesian non-negative matrix factorization. *Physical Review E*, **83**, 066114.
131. Pujol,J., Béjar,J., Delgado,J. (2006) Clustering algorithm for determining community structure in large networks. *Physical Review E*, **74**, 016107.
132. Radicchi,F., Castellano,C., Cecconi,F., Loreto,V., Parisi,D. (2004) Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 2658–2663.
133. Raghavan,U.N., Albert,R., Kumara,S. (2007) Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, **76**, 036106.
134. Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N., Barabasi,A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.
135. Reichardt,J., Bornholdt,S. (2004) Detecting fuzzy community structures in complex networks with a Potts model. *Physical Review Letters*, **93**, 19–22.
136. Reichardt,J., Bornholdt,S. (2006) Statistical mechanics of community detection. *Physical Review E* , **74**, 016110.
137. Rhrissorrakrai,K., Gunsalus,K.C. (2011) MINE: module identification in networks. *BMC Bioinformatics*, **12**, 192.
138. Rivera,C.G., Vakil,R., Bader,J.S. (2010) NeMo: Network module identification in cytoscape. *BMC Bioinformatics*, **11**, S61.
139. Rosvall,M., Bergstrom,C.T. (2007) An information-theoretic framework for resolving

- community structure in complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 7327-7331.
140. Rosvall,M., Bergstrom,C.T. (2010) Mapping change in large networks. *PLoS ONE*, **5**, e8694.
 141. Samal,A., Wagner,A., Martin,O.C. (2011) Environmental versatility promotes modularity in genome-scale metabolic networks. *BMC Systems Biology*, **5**, 34.
 142. Schuetz,P., Caflisch,A. (2008) Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Physical Review E*, **77**, 046112.
 143. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B., Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.
 144. Sharan,R., Ideker,T. (2006) Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, **24**, 427–433.
 145. Sheng,W., Liu,X. (2006) A Genetic K-medoids Clustering Algorithm. *Journal of Heuristics*, **12**, 447–466.
 146. Shmulevich,I., Dougherty,E.R., Kim,S., Zhang,W. (2002) Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**, 261–274.
 147. Simko,G.I., Csermely,P. (2013) Nodes having a major influence to break cooperation define a novel centrality measure: game centrality. *PLoS ONE*, **8**, e67159.
 148. Singh,R., Xu,J., Berger,B. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 12763–12768.
 149. Son,S.-W., Jeong,H., Noh,J.D. (2006) Random field Ising model and community structure in complex networks. *The European Physical Journal B*, **50**, 431–437.
 150. Song,J., Singh,M. (2013) From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interactome at different scales of organization. *PLoS Computational Biology*, **9**, e1002910.

151. Spirin,V., Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 12123–12128.
152. Su,G., Kuchinsky,A., Morris,J.H., States,D.J., Meng,F. (2010) GLay: community structure analysis of biological networks. *Bioinformatics*, **26**, 3135–3137.
153. Suthram,S., Shlomi,T., Ruppin,E., Sharan,R. and Ideker,T. (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, **7**, 360.
154. Szalay,M., Stanojevic,O., Farkas,L. (2010) Automatic use of behavioral information for promotional purposes in communications system. International Ericsson PCT patent application, PCT/SE2010/051312
155. Szalay,K.Z., Csermely,P. (2013) Perturbation centrality: a novel centrality measure obtained by the general network dynamics tool, Turbine. <http://arxiv.org/abs/1305.2496>
156. Szalay-Bekő,M. (2010) Distributed graph clustering engine. Master's thesis: http://linkgroup.hu/docs/thesis_szalay.pdf
157. Szalay-Bekő,M., Palotai,R., Szappanos,B., Kovács,I.A., Papp,B., Csermely,P. (2012) ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping network modules and community centrality. *Bioinformatics*, **28**, 2202–2204.
158. Tamassia,R. (2013) *Handbook of Graph Drawing and Visualization*. Chapman & Hall/CRC.
159. Taylor,I.W., Linding,R., Warde-Farley,D., Liu,Y., Pesquita,C., Faria,D., Bull,S., Pawson,T., Morris,Q., Wrana,J.L. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, **27**, 199–204.
160. Thomas,G.H., Zucker,J., Macdonald,S.J., Sorokin,A., Goryanin,I., Douglas,A.E. (2009) A fragile metabolic network adapted for cooperation in the symbiotic bacterium *Buchnera aphidicola*. *BMC Systems Biology*, **3**, 24.
161. Toh,H., Horimoto,K. (2002) Inference of a genetic network by a combined approach of

- cluster analysis and graphical Gaussian modeling. *Bioinformatics*, **18**, 287–297.
162. Tyler,J.R., Wilkinson,D.M., Huberman,B.A. (2003) Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. *The Information Society*, **21**, 143–153.
 163. Vilela,M., Chou,I.-C., Vinga,S., Vasconcelos,A.T.R., Voit,E.O., Almeida,J.S. (2008) Parameter optimization in S-system models. *BMC Systems Biology*, **2**, 35.
 164. Wakita,K., Tsurumi,T. (2007) Finding community structure in mega-scale social networks. In: *Proceedings of the 16th international conference on World Wide Web - WWW '07*, pp. 1275–1276.
 165. Walshaw,C. (2001) A Multilevel Algorithm for Force-Directed Graph Drawing. *Graph Drawing*, **1984**, 171–182.
 166. Wang,B., Gao,L. (2012) Seed selection strategy in global network alignment without destroying the entire structures of functional modules. *Proteome Science*, **10**, S16.
 167. Wang,R.-S., Saadatpour,A., Albert,R. (2012) Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*, **9**, 055001.
 168. Wang,S., Szalay,M.S., Zhang,C., Csermely,P. (2008) Learning and innovative elements of strategy adoption rules expand cooperative network topologies. *PLoS ONE*, **3**, e1917.
 169. Wang,Y., Wang,R.S., Zhang,X.S., Chen,L. (2007) Establishing protein functional linkage in a systematic way. *Lect. Notes Oper. Res.*, **7**, 75–88.
 170. Wang,Z., Zhang,J. (2007) In search of the biological significance of modular structures in protein networks. *PLoS Computational Biology*, **3**, e107.
 171. Watts, D. J.; Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**: 440–442.
 172. Wasserman,S., Faust,K. (1994) *Social network analysis*. Cambridge University Press.
 173. Weinan,E., Li,T., Vanden-Eijnden,E. (2008) Optimal partition and effective dynamics of complex networks. *Proceedings of the National Academy of Sciences of the United*

States of America, **105**, 7907–7912.

174. Wittkop,T., Emig,D., Lange,S., Rahmann,S., Albrecht,M., Morris,J.H., Böcker,S., Stoye,J., Baumbach,J. (2010) Partitioning biological data with transitivity clustering. *Nature Methods*, **7**, 419–420.
175. Wu,F.Y. (1982) The Potts model. *Reviews of Modern Physics*, **54**, 235–268.
176. Xie,J., Kelley,S., Szymanski,B. (2013) Overlapping community detection in networks: the state of the art and comparative study. *ACM Computing Surveys*, **45**, 43.
177. Xie,J., Szymanski,B.K., Liu,X. (2011) SLPA: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: *11th IEEE International Conference on Data Mining - ICDM 2011*, pp. 344-349.
178. Xu,Y., Xu,D., Gabow,H.N. (2000) Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, **16**, 1091–1104.
179. Yang,J., Leskovec,J. (2012) Structure and Overlaps of Communities in Networks. <http://arxiv.org/abs/1205.6228>
180. Yook,S.-H., Oltvai,Z.N., Barabási,A.-L. (2004) Functional and topological characterization of protein interaction networks. *Proteomics*, **4**, 928–942.
181. Yu,H., Braun,P., Yildirim,M.A., Lemmens,I., Venkatesan,K., Sahalie,J., Hirozane-Kishikawa,T., Gebreab,F., Li,N., Simonis,N., Hao,T., Rual,J.-F., Dricot,A., Vazquez,A., Murray,R.R., *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.
182. Yu,H., Kim,P.M., Sprecher,E., Trifonov,V., Gerstein,M. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, **3**, e59.
183. Zachary, W.W. (1977) An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33**, 452–473.
184. Zak,D.E., Gonye,G.E., Schwaber,J.S., Doyle,F.J. (2003) Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network.

Genome Research, **13**, 2396–2405.

185. Zhang,S., Wang,R.-S., Zhang,X.-S. (2007) Identification of overlapping community structure in complex networks using fuzzy -means clustering. *Physica A: Statistical Mechanics and its Applications*, **374**, 483–490.
186. Zhang,S., Zhang,X.-S., Chen,L. (2008) Biomolecular network querying: a promising approach in systems biology. *BMC Systems Biology*, **2**, 5.
187. Zhao,X.-M., Wang,Y., Chen,L., Aihara,K. (2008) Gene function prediction using labeled and unlabeled data. *BMC Bioinformatics*, **9**, 57.
188. Zhou,H. (2003a) Distance, dissimilarity index, and network community structure. *Physical Review E*, **67**, 061901.
189. Zhou,H. (2003b) Network landscape from a Brownian particle's perspective. *Physical Review E*, **67**, 041908.
190. Zhou,H., Lipowsky,R. (2004) Network Brownian motion: A new method to measure vertex-vertex Proximity and to identify communities and subcommunities. In: *4th International Conference on Computational Science – ICCS 2004*, pp. 1062–1069.