Szakmai beszámoló

Komplex hálózatok vizsgálata statisztikus fizikai módszerekkel

OTKA posztdoktori kutatás, PD 48422
Farkas Illés, ELTE Biológiai Fizika tanszék

1. Bevezető

A természettudományos kutatás legjelentősebb eredményei közé tartozik a természetben megfigyelhető jelenségek elemi alkotórészeinek feltárása. A legtöbb természeti jelenség sok elemi egység kölcsönhatása nyomán alakul ki, és az alkotóelemek együttes (csoportos, kollektív) viselkedésének törvényszerűségeit vizsgálja a statisztikus fizika. Az elvégzett kutatómunka során kollégákkal együttműködve elméleti és számítógépes statisztikus fizikai eszközökkel kölcsönható sokrészecske-rendszereket vizsgáltam: elemeztem a természetben és a társadalomban előforduló komplex hálózatok nagyskálájú tulajdonságait és moduláris szerkezetét.

2. Kutatási célkitűzések

2.1. A kutatás kezdetén (a pályázatban) kitűzött célok

- A hálózati csoportosulások felismerésében és elemzésében egy egyszerű elvekre épülő, gyors és könnyen kezelhető módszer kidolgozása.
- Fehérjék funkciójának jóslása, pontosítása és fehérje-fehérje kölcsönhatási térképek készítése hálózati modulok és expressziós adatok segítségével.
- Komplex hálózatok vizsgálata hálózati motívumok segítségével.

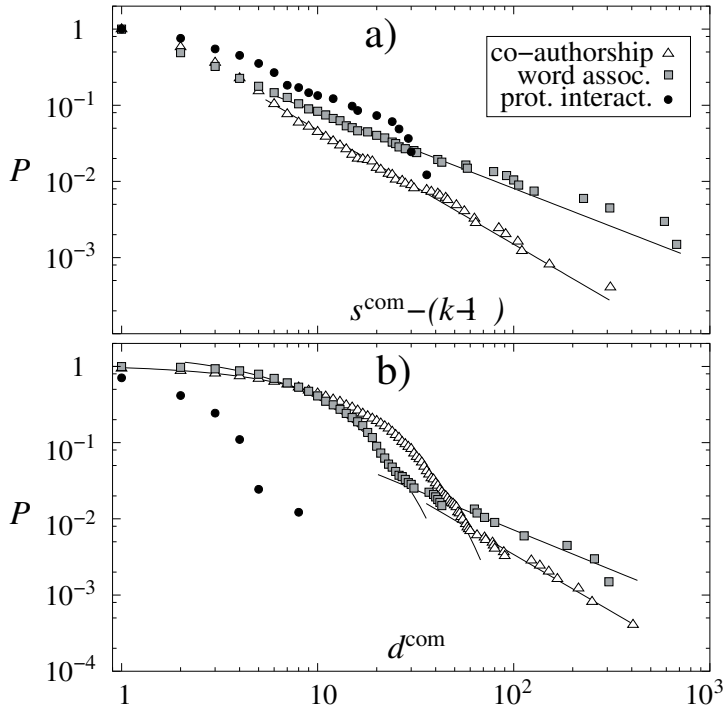
2.2. A kutatás során kitűzött további célok

- Irányított és súlyozott hálózatok átfedő moduljainak definiálása és a hálózatok osztályozása átfedő moduljaik szerkezete alapján.

3. Kutatási eredmények

3.1. Szociális, molekuláris biológiai és kognitív hálózatok átfedő csoportosulásainak vizsgálata

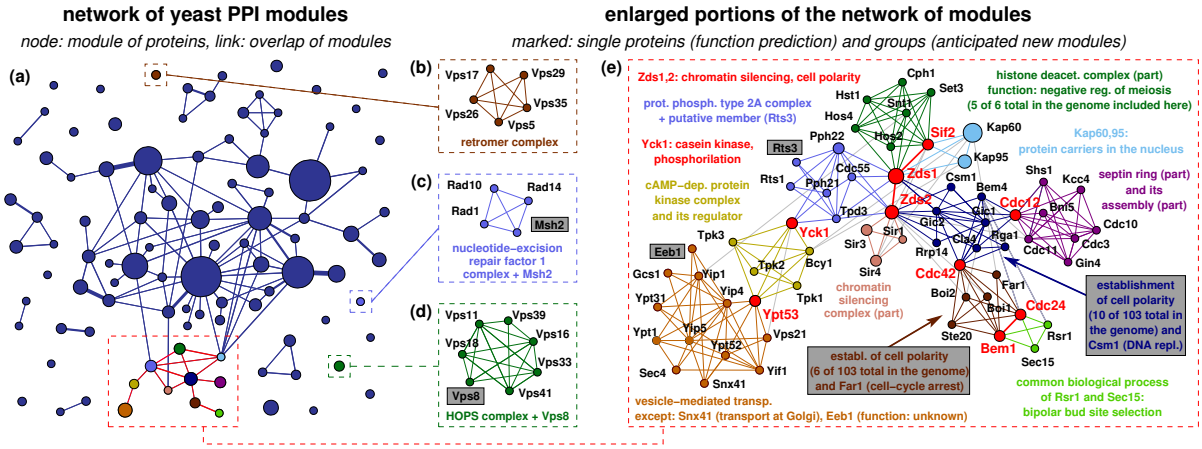
A természet és a társadalom komplex hálózatainak leírásakor az egyik kulcsfontosságú kérdés, hogy milyen módon lehet ezeket a hálózatokat modulok (csoportosulások, részrendszerek) segítségével felbontani és értelmezni. A modulok egyik leggyakrabban használt definíciója szerint minden egyes modul a hálózat olyan csúcspontjait tartalmazza,



1. ábra. A k-klikk perkolációs módszerrel kiszámított átfedő modulok statisztikai tulajdonságai három nagy méretű hálózatban. A három hálózat: a Los Alamos cond-mat preprint archívumban található szerzők társszerzőségi kapcsolatai (háromszögek, 30 739 szerző és 136 065 kapcsolat), a „South Florida Free Association norms" szó asszociációs hálózat (négyzetek, 10 617 szó és 63 788 asszociációs kapcsolat) sarjadzó élesztő fehérje-fehérje kölcsönhatási hálózata Database of Interacting Proteins forrásból (körök, 2 609 fehérje és 6 355 kölcsönhatás). (a) Egy csoportosulás mérete az általa tartalmazott hálózati csúcspontok (szerzők, szavak, fehérjék) száma. A csoportosulások méretének kumulatív eloszlása hatványfüggvényt követ, amelynek kitevője -1 (felső vonal) és -1.6 közötti (alsó vonal). (b) Egy csoportosulás fokszáma a vele legalább egy közös csúcspontot tartalmazó további csoportosulások száma. A csoportosulások fokszámának kumulatív eloszlása exponenciálisan indul, majd hatványfüggvényre vált (a csoportosulások méret-eloszlásánál látottal azonos kitevővel). Az ábra átvétel a [1] publikációból.

amelyen belül a csúcsok egymással az átlagosnál több (erősebb) kapcsolattal rendelkeznek. A hálózatok szerkezetének és működésének megismerése szempontjából alapvető jelentőségű a modulok azonosítása (ezek például az ipari termelés részterületei, egymással szorosan együttműködő fehérjék, vagy emberek csoportjai) és megfelelő értelmezése. A korábban ismert, nagy hálózatokban használt determinisztikus módszerek egymástól különálló csoportosulásokat azonosítottak, ezzel szemben a legtöbb megfigyelt hálózatról ismert, hogy a szorosan összefüggő csúcspontok csoportjai egymással jelentősen átfednek.

Bevezettünk egy hálózati modulkereső módszert [1], amellyel elemeztük a komplex hálózatok erősen kapcsolt, átfedő moduláris szerkezetét. Az elemzéshez használt statisztikus mennyiségek egy része korábbi munkákban nem vizsgált, új szerkezeti jellemzőket mér (1. ábra). A módszert megvalósító hatékony algoritmussal azonosítottuk nagy hálózatok csoportosulásait. Eredményeink szerint a hálózati modulok átfedései jelentősek és a bevezetett eloszlásfüggvények alakjai a komplex hálózatok igen tág csoportjában érvényesek. Az együttműködési, szó-asszociációs és biológiai gráfokban kapott eredményeink alapján a modulok hálózata nemtriviális korrelációkat és speciális skálázási tulajdonságokat mutat. A publikációkhoz létrehozott CFinder hálózati modulkereső szoftver non-profit kutatási célokra ingyenesen letölthető a http://www.CFinder.org weboldal-



2. ábra. (a) Az átfedő hálózati modulokat kereső k-klikk perkolációs módszerrel azonosított fehérje csoportosulások sarjadzó élesztő (*S.cerevisiae*) fehérje-fehérje kölcsönhatási hálózatában. Az adatok forrása: Database of Interacting Proteins [2]. (b-d) Az ismert fehérje komplex-ek azonosításán túl a k-klikk perckolációs módszert használó CFinder keresőprogram gyakran tesz hozzá egy ismert csoporthoz egy még ismeretlen funkciójú fehérjét (Msh2, Vps8), amelynek így a funkciója jósolható. (e) A fehérje-fehérje kölcsönhatási hálózat moduljainak részletes elemzése az azonos modulokba sorolt fehérjék ismert funckióinak vizsgálatával. Az ábrán mutatott sötétkék és sötétbarna csoport fehérjéienk legjelentősebb közös funckiója a sejtpolaritás létrehozása. Mivel ezen a feladaton az élesztő sejtben száz feletti fehérje dologzik, ezért a két kiemelt csoport valószínűleg a sejtpolaritás létrehozásánál speciálisabb, eddig még ismeretlen részfeladatot végez. Az ábra átvétel a [3] publikációból.

ról. A szoftver letöltőinek száma jelenleg 1 500 felett van. A programot használják például statisztikus fizikai, szociometriai, szervezetfejlesztési, szövegelemzési, "data mining" (adatbányászati) és bioinformatikai kérdések hálózati alapú vizsgálatára.

3.2. Fehérjék funkciójának jóslása és fehérje-fehérje kölcsönhatási térképek készítése hálózati modulok segítségével

Az elmúlt néhány évtizedben a molekuláris biológiai mérési módszerek gyors fejlődésen mentek keresztül. Korábban egy-egy mérésben általában a vizsgált folyamat néhány résztvevőjének (például fehérjének) az azonosítása és követése volt a cél. Napjainkban már van lehetőség több ezer különböző "építőelem" (gén, fehérje) viselkedésének együttes vizsgálatára is. Egy adott sejttípusban a fehérjék (csúcspontok) és az összes vizsgált kísérleti körülmény során regisztrált fizikai/kémiai kapcsolat (él) definiálja a fehérje-fehérje kölcsönhatási gráfot. Ebben a gráfban a modulok egymáshoz sűrűn kapcsolt pontok csoportjaiként jelennek meg. Érdekes, hogy ezek a modulok nem függetlenek egymástól. Gyakori, hogy két modulnak számos fehérjéje vagy akár nagyobb alegysége azonos.

A bevezetett k-klikk perkolációs módszer segítségével azonosítottunk a sarjadzó élesztő (S. cerevisiae) egysejtű szervezet fehérje-fehérje kölcsönhatási hálózatának moduljait és a modulok biológiai szerepét (ld. 2. ábra). A modulok jelentős segítséget nyújtanak A hálózati modulok segítségével megjósoltuk számos, ismeretlen funkciójú fehérje szerepét. Továbbá azonosítottunk korábban ismeretlen hálózati modulokat, amelyeknek a résztvevői (fehérjék) valószínűleg azonos, speciális biológiai feladat elvégzése során együttműködnek. A talált modulok az alapkutatás számára új működési elveket tárhatnak fel, és a gyógyászati alkalmazások számára a potenciális célpont fehérjéken túl célpontként

modulokat is kijelölhetnek.

3.3. Hálózati motívumok és jelfeldolgozás élesztő transzkripció szabályozási hálózatában

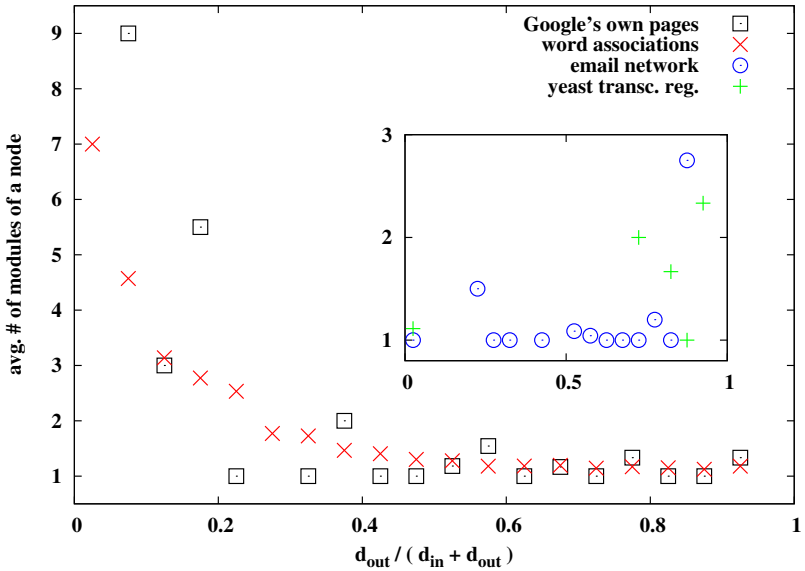
Az elmúlt évtizedben a molekuláris biológia egyik leggyorsabban fejlődő területe a sejtekben felismert alkotóelemek közötti szabályozási kapcsolatok feltárása. Az folyamatos működés fenntartása valamint a sokféle belső és külső környezeti változásra adandó válasz (alkalmazkodás) kialakítása minden sejtben összetett jelátviteli és szabályozási útvonalak használatát teszi szükségessé. A szabályozási rendszer egyik eleme a transzkripció reguláció, amelynek során a szabályozó fehérjék (transzkripciós faktorok) serkentik ill. gátolják számos gén átírását. A serkentés és a gátlás által megvalósított finom szabályozás lehetővé teszi, hogy a sejt belső állapota az érzékelt környezetnek megfelelően változzon.

A sarjadzó élesztő modell szervezetben ismert mérési adatok alapján kollégákkal együttműködve megvizsgáltam a transzkripció szabályozási hálózat hálózati motívumait. Ezek a néhány csúcspontól álló részgráfok gyakran jellegzetes információ-feldolgozási feladatot képesek elvégezni, például a Feed-Forward Loop (FFL) aluláteresztő szűrőként használható. Az elemi jelfeldolgozási feladatok súlyának összehasonlítása érdekében azonosítottuk a statisztikailag kiugró számban előforduló hálózati motívumokat és az egy környezeti jel által befolyásolt fehérjék listáját. Feltérképeztük a transzkripció szabályozási hálózat moduljait, azonosítottuk a modulok funkcióját [4]. Az adatok összegyűjtéséhez, rendszerezéséhez és minőségének kontrolljához további módszerekre javaslatot tettünk [5].

3.4. Irányított és súlyozott hálózatok átfedő moduláris szerkezete

Definiáltunk hatékony algoritmusokat irányított és súlyozott hálózatok átfedő modulok azonosítására. A súlyozott hálózati modulokat meghatározó algoritmus optimális paramétereinek becslése illetve a korrelálatlan kontroll elemzése érdekében két közelítő módszerrel kiszámítottam a súlyozott k-klikk perkolációs átalakulás kritikus pontját (súlyozott) Erdős-Rényi gráfokon. Szintén az átfedő hálózati modulok segítségével kapott eredmény, hogy a szomszédos élek hármas illetve nagyobb csoportjaiban a nagy élsúlyok preferenciálisan kapcsolódnak egymáshoz. A szakirodalomban ismert korábbi eredmények szomszédos élek 2 elemű csoportjai esetén ilyen korrelációt nem találtak. Jelentős molekuláris biológiai alkalmazás a fehérje-fehérje kölcsönhatási hálózatok átfedő moduljainak a korábbi, súlyozatlan kereső módszerhez képest pontosabb azonosítása [7].

Az irányított hálózati modulok segítségével az irányított hálózatokat két csoportba soroltuk aszerint, hogy az egymással átfedő modulok az átfedésekből kifelé ill. azokba befelé mutatnak (ld. 3 ábra). Ez a különbség technológiai, kognitív, szociális és molekuláris biológiai hálózatokban egyaránt az információ ill. anyag áramlásának eltérő irányára utal. A sarjadzó élesztő (S. cerevisiae) transzkripció szabályozási hálózatában és egy elektronikus levelezési gráfban az irányított hálózati modulok kifelé irányulóak, míg egy technológiai és egy kognitív hálózatban a modulok befelé, az átfedő tartományok felé mutatnak. A sarjadzó élesztő esetén azonosított irányított hálózati modulok esetén



3. ábra. Az irányított k-klikk perkolációs módszerrel megtalalált átfedő irányított hálózati modulok két típusának azonosítása. A vízszintes tengelyen szereplő $d_{out}/d_{in}+d_{out}$ hányados megmutatja, hogy a hálózat egy pontjának kapcsolatai közül mekkora rész kimenő kapcsolat. Például ha a csúcspont csak kifelé mutató kapcsolattal rendelkezik, akkor $d_{out}/d_{in}+d_{out}=1$, ha csak bejövő éllel rendelkezik, akkor $d_{out}/d_{in}+d_{out}=0$. A függőleges tengelyen egy csúcs moduljainak számát mértük. (Nagy ábra) A http://www.google.com web tartományban található oldalak a köztük futó irányított élekkel (hyperlinkek) egy olyan hálózatot alkotnak, amelyben az átfedő irányított modulok az átfedések felé mutatnak. Ehhez hasonló, befelé mutató modulokat találtunk az angol szavak asszociációs kapcsolatait mutató hálózatban. Mindkét esetben a mért függvény csökkenő, mert a modulokat összekötő pontok (az átfedések) sok befelé mutató ($d_{out}/d_{in}+d_{out}=0$) és kevés kifelé mutató éllel rendelkeznek. (Kis ábra) A modulok átfedéseiből kifelé mutató irányított modulokat találtunk egy elektronikus levelezési hálózatban és a sarjadzó élesztő transzkripció szabályozási hálózatában. Az ábra átvétel a [6] publikációból.

statisztikai tesztek segítségével megkerestük az egyes modulok funkcióját (a modulban található fehérjék szignifikáns közös funkcióját). A kifejlesztett irányított és súlyozott hálózati modulkereső módszerek segítségével lehetőség nyílik súlyozott fehérje-fehérje kölcsönhatási hálózatok és irányított (pl. szignál transzdukciós) hálózatok részletesebb vizsgálatára [6].

4. Publikációs tevékenység

A kutatási időszak alatt az OTKA posztdoktori kutatás által támogatott témában a posztdoktori kutató társszerzőségével referált nemzetközi folyóiratban megjelent 6 cikk (5 kutatási, 1 összefoglaló), amelyeknek a kumulatív impakt faktora az ISI szerint 84. Ezekre a cikkekre az ISI alapján 2008 júniusig 141 független hivatkozás érkezett. A Google Scholar kereső által talált hivatkozások száma 293. További adatok: 5 konferencia előadás és 4 konferencia poszter nemzetközi konferencián, 5 szemináriumi előadás egyetemeken és kutatóintézetekben.

A támogatott kutatás témájában a posztdoktori kutató egy magyar nyelvű ismeretterjesztő cikknek szerzője volt (a Fizikai Szemlében) és egy továbbinak társszerzője (a Magyar Tudományban).

5. Az OTKA posztdoktori kutató által elvégzett kutatási feladatok részletezése

Az együttműködésben végzett kutatások során az OTKA posztdoktori kutató végezte el a következő kutatási feladatokat. Hálózatok összeállítása a használt összes hálózat típusban: az adatok felkutatása, hálózatok elkészítése a nyers adatokból, automatizált feldolgozás [1, 3, 4, 6, 7]. A molekuláris biológiai hálózatokban fehérje nevek automatizált átalakítása a megfelelő név típusra [1, 3, 4, 6]. Részvétel a modulkereső algoritmusok kifejlesztésében [1, 6, 7], a kereső algoritmusok alkalmazásainak kidolgozása és a számítógépes programban megírt algoritmusok használata különböző hálózat típusokban [3, 4, 6, 7]. Fehérje csoportok legjelentősebb közös funkcióját megtaláló statisztikai tesztek elvégzése [4, 3, 6], majd ez alapján fehérje funkció jóslás és korábban ismeretlen feladatot végző fehérje csoportok előrejelzése [3]. Részvétel az eredmények statisztikus fizikai értelmezésében és modellezésében [1, 3, 4, 6, 7]. Az irányított k-klikk perkolációs átalakulás pontjának numerikus kiszámítása Palla Gergellyel közösen [6]. A súlyozott k-klikk csoportosulások perkolációs küszöbének közelítő számítása kétféle analitikus módszerrel [7]. Ábrák készítése és kéziratok írása, szerkesztése [1, 3, 5, 4, 6, 7]. Az OTKA posztdoktori kutató a [3, 6, 7] projektekben a kutatás közös kezdeményezője volt.

Hivatkozások

[1] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814-818 (2005).

[2] Salwinski L, Miller C S, Smith A J, Pettit F K, Bowie J U, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucl. Acids Res.* **32**, Database issue: D449-51 (2004).

[3] Adamcsek B, Palla G, Farkas I J, Derényi I, Vicsek T. CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics* **22**, 1021-1023 (2006).

[4] Farkas I J, Wu C, Chennubhotla C, Bahar I, Oltvai Z N. Topological basis of signal integration in the transcriptional-regulatory network of the yeast, Saccharomyces cerevisiae. *BMC Bioinformatics* **7**, 478 (2006).

[5] Farkas I J, Beg Q K, Oltvai Z N. Exploring transcriptional regulatory networks in the worm. *Cell* **125**, 1032-1034 (2006). Preview article.

[6] Palla G, Farkas I J, Pollner P, Derényi I, Vicsek T. Directed network modules. *New J. Phys.* **9**, 186 (2007).

[7] Farkas I J, Abel D, Palla G, Vicsek T. Weighted network modules. *New J. Phys.* **9**, 180 (2007).