

Válasz ÓDOR GÉZÁNAK

„Átfedő modulok molekuláris biológiai kölcsönhatási hálózatokban”

című MTA doktori értekezésem bírálatára

Köszönöm Ódor Gézának a dolgozat alapos átolvasását, és a dolgozat kapcsán megfogalmazott pozitív véleményt. A bírálat végén felsorolt kérdésekre a következő válaszokat adom.

(1) A 31. oldal lábjegyzete szerint az általánosított kisvilág modell definíciója az alacsony átmérő és a magas klaszterezettség. Ettől eltérő definíciót ismerek és használtunk saját publikációinkban, mely reguláris rácshoz hozzáadott hosszú élekkel van meghatározva. Mi az eltérő definíciók oka?

Watts és Strogatz a kisvilág modellt definiáló 1998-as cikkük 2. ábráján a kisvilág hálózatot annak az általános elvnek a felhasználásával jelölik ki, hogy a hálózatban legyen az átlagos úthossz alacsony és a klaszterezettség magas [1]. Watts és Strogatz ugyanennek a cikknek az 1. ábráján az általános elv megvalósítására egy konkrét lehetőséget mutat. Továbbá az átlagos úthossz és az átmérő egymással arányos mennyiségek: az átmérő gyakoribb definíciója a maximális úthossz, a ritkább az átlagos úthossz, és a két mennyiség egymással arányos. Ezek alapján szerepel a dolgozat 31. oldalán lévő 13. számú lábjegyzetében az, hogy az általánosított kisvilág modell definíciója az alacsony átmérő és a magas klaszterezettség.

Fontos, hogy a **reguláris rácso**k közül nem mindegyik megfelelő, például 2 dimenzióban az egyszerű négyzetrács klaszterezettségi együtthatója nulla, mert ebben a rácso**n** egy pont szomszédai között nincsen él. A kisvilág modell Watts és Strogatz által bevezetett eredeti verziójában a kiindulási hálózat egy olyan 1 dimenziós periodikus rács, amelyben minden csúcspont hozzá van kapcsolva a k darab (k páros szám) legközelebbi szomszédjához. Például a $k=2$ esetben ennek az egydimenziós rácsnak a klaszterezettsége $C=1/2$. Egy további példa két dimenzióban: a háromszögrács klaszterezettsége $C=2/5$. A magas klaszterezettségű kiindulási hálózathoz **hozzáadott élek** pontos definíciója az, hogy ezek az új élek korrelálatlan módon választott pont párokat kötnek össze. Ez a kis számú véletlenszerű él a kiindulási hálózat szempontjából általában valóban „hosszú”. Végül még egy technikai jellegű megjegyzés: az itt használt „klaszterezettség” kifejezés teljes neve **átlagos lokális klaszterezettség** együttható.

(2) A 32. oldal alján Erdős Rényi véletlen gráfra az az állítás van megfogalmazva, hogy ha az átlagos fokszám nagyobb mint 1, akkor a végtelen limeszben a gráf összes csúcspontja egyetlen komponensbe fog tartozni. Nem lehetséges, hogy a véletlenszerűség miatt kimarad néhány csúcs az óriás komponensből?

Köszönöm a fontos észrevételt. A dolgozat 32. oldalán lévő utolsó mondat és a 33. oldalon lévő első **mondat hibás**. A két mondat javított változata a következő: *Ha az ER gráfban az átlagos fokszámra $\langle k \rangle = 2E/N > 1$ teljesül, akkor az $N \rightarrow \infty$ határesetben a gráfban megjelenik egy (óriás) komponens, amelybe az összes csúcspont véges hányada tartozik. Ebben az óriás komponensben tetszőleges két pont között van kapcsolat és – egy rövid érvelés¹⁴ alapján – az ilyen kapcsolatok (pont-pont utak) jellemző hossza $O(\log N)$ szerint növekszik.*

Szintén ehhez a részhez tartozó megjegyzés (ami kimaradt a dolgozatból), hogy az itt tárgyalt fehérje-fehérje kölcsönhatási (PPI) hálózatok szerkezetének elemzésekor a szakirodalomban a teljes hálózat helyett **szokás csak az óriás komponens**t vizsgálni.

(3) Az 54. oldal alján az olvasható, hogy a „CPM módszer ROC görbéje jobb (magasabban halad), mint a CPM módszeré”. Kérem fejtse ki, hogy ez alatt mit kell érteni pontosan.

A kérdésben idézett mondatrész első szavának a végén a dolgozat szövegéhez képest lemaradt a „w” betű. A dolgozat 54. oldalának alján három különböző hálózati modulkereső algoritmus összehasonlítása szerepel: CPMw, CPM és DME. Az összehasonlítás és a DME algoritmus Georgii és munkatársainak eredménye [2]. A szerzők a három algoritmus összehasonlítását az ROC (Receiver Operating Characteristic) nevű statisztikai módszer segítségével végezték el. Az összehasonlítás eredménye az, hogy a vizsgált konkrét feladatokra a legjobb a DME algoritmus, ezután következik a CPMw és azt követi a CPM.

Az **ROC módszer** leírásához tekintsünk először egy n elemű halmazt, amelynek mindegyik eleme rendelkezik egy darab címkével. A címke értéke kétféle lehet: Igaz vagy Hamis. Az ROC ismeri az összes elem címkéjét. Tekintsünk továbbá egy A algoritmust, amelyik nem ismeri a címkéket, és az n elem mindegyikére megmondja, hogy szerinte az adott elem címkéje Igaz vagy Hamis. Az A algoritmus által kijelölt Igaz elemek halmazának neve az A algoritmus „Positive” (P) halmaza. Ha egy elem tényleges címkéje és az A algoritmus által mondott címkéje egyaránt Igaz, akkor ez az elem az A algoritmus egy „True Positive” (TP) eredménye. Ha egy elem tényleges címkéje Hamis, viszont az A algoritmus által mondott címkéje Igaz, akkor ez az elem az A algoritmus egy „False Positive” (FP) eredménye. (Szöveg vagy képlet olvasásakor a TP, FP, stb. betűszavakat gyakran angolul kiolvassák.) Hasonlóan definiálhatóak az A algoritmus „Negative” (N), „True Negative” (TN) és „False Negative” (FN) eredményei. Ezután jelöljük mindegyik halmaz (P, TP, FP, N, TN, FN) elemeinek a számát az adott halmaz nevével, például a True Positive elemek száma TP.

Az n elemű halmazban található Hamis címkéjű elemek mindegyike esetén az A algoritmus kétféle dolgot tehet: „Igaz” címkét mond (ezek a FP eredmények) vagy „Hamis” címkét mond (ezek a TN eredmények). Tehát a vizsgált n darab elem között szereplő Hamis címkéjű elemek száma FP+TN. Az itt leírtak segítségével a False Positive Rate (FPR) nevű mennyiség azt méri, hogy az A algoritmus az n elem között található Hamis címkéjű elemek mekkora hányadára mondja, hogy Igaz: $FPR = FP / (FP + TN)$. A fehérje-fehérje kölcsönhatási hálózatok modulkereső algoritmusai kapcsán az FPR mennyiségre egy **konkrét példa** az, hogy az A modulkereső módszer **az azonos modulba nem tartozó fehérje párok (FP+TN) mekkora hányadát (FPR)** teszi hibásan azonos modulba. Az FPR-hez hasonlóan definiálható a True Positive Rate (TPR): $TPR = TP / (TP + FN)$.

Egy tökéletes A algoritmus esetén az algoritmus paramétereitől függetlenül $FP = FN = 0$, és emiatt $FPR = 0$ és $TPR = 1$. Tehát ideális esetben a TPR az FPR függvényében ábrázolva egyetlen pont: $(FPR, TPR) = (0, 1)$. Valós A algoritmusok esetén a TPR az FPR-nek monoton növekedő függvénye, és ennek a (numerikusan mért) függvénynek a neve ROC görbe. Szemléletesen: egy algoritmus paramétereinek „hangolásával” általában csak úgy növelhetjük az Igaz elemekből megtalált hányadot, hogy ezzel egyidejűleg a Hamis elemek hibásan Igaz-nak azonosított hányada is nő. A fentiek szerint az ideális esetben már a lehetséges legalacsonyabb FPR érték (0) mellett elérjük a lehetséges legmagasabb TPR értéket (1). Valós A algoritmusok összehasonlításakor az ROC módszerrel azt az algoritmust tekintjük a legjobbnak, amelyik az ideális esethez legközelebb halad. Azaz, már alacsony FPR mellett elér magas TPR értéket. A 2 dimenziós ROC ábrán ez úgy jelenik meg, hogy a legjobb algoritmus ROC görbéje (egy szakaszon vagy akár a teljes 0-1 intervallumon) a **többi vizsgált algoritmus ROC görbéje felett halad**.

Megjegyzés: az ROC (Receiver Operating Characteristic) betűszó mérés technikai eredetű.

(4) A 72. oldalon az szerepel, hogy a fehérje modulok foksám eloszlása exponenciális típusú, majd az eloszlás vége hatványfüggvényszerű. Melyik ábra mutatja az utóbbi komplex eloszlási görbét? Ha az exponenciális lecsengésnek az oka az alkalmazott klikk paraméter, akkor ez a modulkeresés jellemzőjének tűnik és nem a hálózatra jellemző?

Az első kérdésre adott válasz: ezt az eloszlást a dolgozatban T1. jelzéssel hivatkozott publikáció 4.b ábrája mutatja [3]. A második kérdésre adott válasz: igen, ez a modulkeresés jellemzője. Ismert, hogy a valós hálózatok foksám eloszlása (nem csak fehérje-fehérje kölcsönhatási, hanem más hálózatok esetén is) gyakran hatványfüggvényhez igen közeli [4,5]. Ehhez képest a modulok méretének eloszlására az eredményünk azt mutatja, hogy a modulkeresési módszer élsűrűség paramétere egy foksám (csúcs szám) skálájú mennyiséget vezet be, amelyik karakterisztikus mennyiségként megjelenik az azonosított modulok méret (csúcs szám) eloszlásának kezdeti szakaszán, és így azon a szakaszon exponenciális eloszlást eredményez.

(5) A fehérje hálózatok kapcsolatán túl léteznek olyan tulajdonságok, melyek ezen hálózatokon definiált elágazó, bolyongó és annihiláló terjedési modellekkel jellemezhetőek? Ismertek-e eredmények ilyen modellekben a dinamikus viselkedésről?

Speciálisan a fehérje-fehérje kölcsönhatási (PPI), a transzkripció szabályozási (TR) vagy a jelátviteli (ST) hálózatok esetén nem ismerek ilyen tulajdonságokat és ilyen tulajdonságok statisztikus fizikai értelemben vett dinamikus viselkedéséről szóló eredményeket. Véleményem szerint a kutatások fő iránya ezen a területen az lehet, hogy **a sejt hogyan választja ki** a rendelkezésére álló néhány lehetőség közül a számára megfelelőt. Ez a nagyon aktív kutatási irány például a „cell fate control” kifejezéssel található meg.

A dolgozatban tárgyalt mindhárom hálózat típusban (PPI, TR, ST) az élő sejt célja az, hogy a kölcsönhatásokon keresztül terjedő információkat a véletlenszerű bolyongás helyett irányítsa. Ennek során a sejt a számára lehetséges biológiai „programok” közül határozottan kiválaszt egyet vagy esetleg néhányat. Minden egyes kiválasztott program határozott kölcsönhatási mintázattal jár együtt. Tehát a sejt működése során a (biokémiaiilag lehetséges összes kölcsönhatást felsoroló) teljes kölcsönhatási hálózattól főként a sejt aktuális állapotára/folyamatára jellemző részgráf aktív, és a hálózat többi része jóval kevésbé áll rendelkezésre.

Általános hálózatokon (például PPI hálózaton) történő bolyongást használ a dolgozat 100. számú hivatkozásában szereplő modulkereső módszer [6]. A terjedési jelenségek kapcsán egy további biológiai vonatkozású modell csoport (a dolgozattól eltérő biológiai témában) a fertőzések terjedésének modellezése. A hálózatos fertőzési modellek dinamikus viselkedésének egyik központi eredménye, hogy ha a hálózat növekedésével a foksámok második momentumuma divergál, akkor a Susceptible-Infected-Susceptible (SIS) fertőzési modellben tetszőleges kis (pozitív) terjedési ráta esetén a fertőzés kiterjed a kezdeti fertőzést tartalmazó gráf komponens összes csúcspontjára. Egy gyakran vizsgált eset a Skálafüggetlen (SF) hálózati modell, amely egybefüggő (egy gráf komponensből áll) és a foksámainak második momentumuma divergál, tehát a SF modellben is igaz, hogy az SIS modell járványterjedési küszöbe nulla [7].

Budapest, 2016. április 14.

Farkas Illés

Hivatkozott irodalom

Minden hivatkozásnál a DOI szám a publikációra mutató hiperlinket tartalmaz.

- [1] D. J. Watts, S. H. Strogatz: Collective dynamics of 'small-world' networks. *Nature* 393, 440-442 (1998) doi:[10.1038/30918](https://doi.org/10.1038/30918).
- [2] E. Georgii, S. Dietmann, T. Uno, P. Pagel, K. Tsuda: Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics* 25, 933-940 (2009) doi:[10.1093/bioinformatics/btp080](https://doi.org/10.1093/bioinformatics/btp080).
- [3] G. Palla, I. Derényi, I. Farkas, T. Vicsek: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814-818 (2005) doi:[10.1038/nature03607](https://doi.org/10.1038/nature03607).
- [4] R. Albert, A.-L. Barabási: Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47-97 (2002) doi:[10.1103/RevModPhys.74.47](https://doi.org/10.1103/RevModPhys.74.47).
- [5] S.N. Dorogovtsev, J.F.F. Mendes: Evolution of networks. *Adv. Phys.* 51, 1079-1187 (2002) doi:[10.1080/00018730110112519](https://doi.org/10.1080/00018730110112519).
- [6] M. Rosvall, C. Bergstrom: Maps of random walks on complex networks reveal community structure. *PNAS* 105, 1118-1123 (2008) doi:[10.1073/pnas.0706851105](https://doi.org/10.1073/pnas.0706851105).
- [7] R. Pastor-Satorras, A. Vespignani: Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86, 3200-3203 (2001) doi:[10.1103/PhysRevLett.86.3200](https://doi.org/10.1103/PhysRevLett.86.3200).