Fehérjehálózatok gráfelméleti elemzése
A szakdolgozat kivonata

Tóthmérész Lilla
Matematika BSc, Matematikus szakirány

Témavezető: Grolmusz Vince, professzor
Számítógéptudományi Tanszék

A dolgozatban gráfelméleti hasonlóság-mértékeket vizsgálunk, illetve alkalmazunk fehérje interakciós hálózatokra. A motiváció egy biológiai eredetű probléma: Hasonló funkciójú fehérjéket szeretnénk találni egy élőlény fehérjéinek kölcsönhatási hálózatának struktúrája alapján. A biológiában az utóbbi években elterjedtek bizonyos úgynevezett nagy áteresztő képességű mérési módszerek, amik igen sok mérési adat kinyerésére alkalmasak. Az így nyert adatok rengeteg információt rejtenek, de az ebben rejlő lehetőségek kihasználásához hatékony elemző módszerek szükségesek. Mivel a probléma még igen fiatal, nem sok kialakult módszer áll rendelkezésre. Sok biológiai jellegű probléma viszont jellegében hasonló más területeken felvetődő problémákhoz, amikre viszont már dolgoztak ki módszereket. A fehérjék egymással való kölcsönhatásainak rendszerét például formalizálhatjuk egy gráfként: Egy élőlény fehérje interakciós hálózata egy irányítatlan gráf, ahol a csúcsok a fehérjéknek felelnek meg és két csúcsot akkor köt össze él, ha a megfelelő fehérjék kölcsönhatásba léphetnek. A fehérje interakciós gráf segítségével a fehérjék hasonlóságának kérdését gráf csúcsainak hasonlóságaként tudjuk megfogalmazni. Sok más a gyakorlatban felvetődő probléma is formalizálható hasonló módon: Például kereshetünk egy adott cikkhez hasonló témájú cikket. Itt a cikkeket feleltethetjük meg csúcsoknak, és irányított élet húzunk egy cikknek megfelelő csúcsból az általa hivatkozott cikkeknek megfelelő csúcsokba. Szintén felvetődik az igény hogy egy weboldalhoz hasonló lapokat találjunk. A web esetén az oldalakat feleltethetjük meg csúcsoknak és a hiperlinkeket irányított éleknek. Kihasználva hogy ezekben a gráfokban az éllel való összekötöttségnek fontos intuitív jelentése van, erre az intuitív jelentésre alapozva definiálhatunk valamilyen hasonlósági mértéket a gráfok csúcsain. Ha az éllel való összekötöttség intuitív jelentését jól értettük meg, remélhetjük, hogy az így kapott hasonlósági mérték jól ki fogja fejezni a csúcsoknak megfelelő objektumok hasonlóságát. A fentebb említett két problémával kapcsolatban a szakirodalomban javasoltak különféle gráfelméleti hasonlóság-mértékeket [4],[1]: ezek közül én a ko-citációval és a SimRank-el foglalkoztam (illetve a ko-citáció bizonyos normált változataival). Ezek közül a ko-citáció a gyakorlatban is elterjedt, és a SimRank-nek is találni gyakorlati alkalmazását [3],[2]. Azt gondoljuk, hogy a fehérje interakciós hálózatokban az éllel való összekötöttség jelentése alapjaiban véve hasonló ezekhez az esetekhez, így ezek a hasonlósági mértékek a fehérjék hasonlóságának mértékéről is hasznos információt szolgáltathatnak. A dolgozatban ezért azt vizsgáljuk, vajon ezek a más alkalmazásokban már hasznosnak bizonyult gráfelméleti hasonlóság-mértékek alkalmasak-e a fehérjék hasonlóságának mérésére.

Ahhoz hogy egy gráfelméleti hasonlóság-mérték által adott hasonlóság-fogalom relevanciáját meg tudjuk ítélni, választanunk kell egy ettől független mértéket az

adott objektumok hasonlóságára. Ezek után megvizsgálhatjuk hogy a gráfelméleti mérték által adott hasonlóság mennyire egyezik ezzel. Mi a fehérjehálózatok esetén az EC szám első három blokkjában való megegyezést vettük a hasonlóság indikátorának. (Az EC szám egy enzimfehérjéhez az általa katalizált reakció alapján rendelt 4 blokkból álló azonosító. Mivel az egymást követő blokkok egyre finomabb osztályozását adják a reakciónak, az EC szám első 3 blokkjában való egyezés kifejez valamilyen fajta hasonlóságot.) Ezek után vizsgáltuk hogy a gráfelméleti hasonlóság-mérték szerint nagy hasonlóságú csúcsok milyen arányban hasonlóak az EC-számuk alapján, illetve hogy az EC-szám alapján hasonló fehérjék átlagosan nagyobb hasonlóságot kapnak-e a gráfelméleti hasonlóság-mérték szerint, mint az EC-szám szerint nem hasonló fehérjék.

Egy hasonlóság-mértékkel kapcsolatban nem csak az az egyedüli elvárás hogy az általa kifejezett hasonlóság-fogalom minél inkább egybevágjon az intuitív vagy valamilyen más alapon vett hasonlóságfogalommal. A jó alkalmazhatóság szempontjából nagyon fontos például hogy egy ilyen mérték mennyire stabil a gráf perturbációival szemben. A mi esetünkben ez különösen fontos, mert a rendelkezésre álló fehérje interakciós hálózatok még korántsem teljesek, és a fehérjék kölcsönhatásainak kimérésére használt módszereknek is vannak hibalehetőségei. Ezért vizsgáltam ezeknek a gráfelméleti hasonlóság-mértékeknek a stabilitását. Feltételt adtam a csúcspárok relatív hasonlóságainak stabilitására a SimRank esetén, és ellenpéldákat arra az estre ha ezek a feltételek nem teljesülnek. A ko-citáció esetén a következő eredményre jutottam: Általános vélekedés, hogy a ko-citáció nem tükrözi vissza jól a csúcsok hasonlóságát ha a gráfban a fokszámok nagyon szóródnak. Ennek a problémának az orvoslásaképp a ko-citációt le szokták normálni a fokszámok egy függvényével. A szokásos normálások esetén viszont a normált hasonlósági mérték elveszti a stabilitását. Ennek hatása a kísérleti eredményeken is jól megfigyelhető volt. Hogy ezen a problémán segítsek, javasoltam egy más fajta normálást, ami mellett lehet valamilyen relatív stabilitást garantálni.

Végül szintén fontos kérdés hogy a gráfelméleti hasonlóság-mérték értékeit könnyen ki lehessen számolni. A ko-citáció gyorsan kiszámítható, a dolgozatban egy $O(md_{max})$ lépésszámú algoritmust javasolok a kiszámítására. (m a gráf éleinek számát jelöli, d_{max} pedig a maximális fokszámot.) A SimRank iteratív algoritmikus kiszámításával is foglalkozok. Ezzel kapcsolatban javaslok egy más megközelítést az algoritmikus kiszámításra, ami a gyakorlatban az operációk egy más csoportosításának felel meg. Ezt a mószert a [2] cikk (ami szintén javasol gyorsításokat a SimRank kiszámítására) eredményeivel hasonlítom össze, illetve fel is használom a cikkben javasolt módszerek némelyikét az eredeti algoritmus javítására.

A teljes dolgozat a cél megjelölésével elkérhető Grolmusz Vincétől.

Hivatkozások

[1] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, New York, NY, USA, 2002. ACM.

[2] Dimitry Lizorkin, Pavel Velikhov, Maxim Grinev, and Denis Turdakov. Accuracy estimate and optimization techniques for simrank computation. *The VLDB Journal*, 19(1):45–66, 2010.

[3] James Pitkow and Peter Pirolli. Life, death, and lawfulness on the electronic frontier. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 383–390, New York, NY, USA, 1997. ACM.

[4] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, Vol. 24(No. 4), July-August 1973.