Ágoston Vilmos:

Hálózati modellek alkalmazása a molekuláris biológia néhány problémájára

Ph.D. dolgozat
Készült a Szegedi Tudományegyetem
Biológia Doktori Iskolájában

Témavezető:
Dr. Pongor Sándor

Benyújtva:

2007. április 5.

MTA Szegedi Biológiai Központ

Tartalom:

Bevezetés

Dolgozatom 2002 és 2005 közötti kutatásokat foglal össze. Ennek az időszaknak egyik jellemző tudományos momentuma volt az a felismerés, hogy a természettudomány és a technika sok területén megismert kölcsönhatási hálózatok struktúrája, a hálózatok topológiája közös vonásokat mutat. Jóllehet a hálózatok alaptulajdonságait, és ezen belül például az úgynevezett skálamentes hálózati topológiát már régebben ismerték, az Internet szerkezetével kapcsolatos vizsgálatok (1), (2) óriási feltűnést keltettek, amelyet tovább növeltek azok a felismerések, melyek szerint a technikai hálózatokhoz hasonló topológia a biológiai rendszerekben is megtalálható.

Munkám célja az volt, hogy az újonnan megjelent hálózati modellek általános tulajdonságait megismerjem, és megkíséreljem alkalmazni azokat a biológiai problémák néhány újabb területén. Az egyik kérdés az volt, hogy egyszerű, topológiai leírásokra korlátozott hálózati modellek milyen választ adhatnak egyes problémákra, vagyis a felfedezett „nagy párhuzamok" valóban olyan mélyreható analógiákat mutatnak-e a különböző rendszerek között, mint az eleinte sejthető volt. Ebben a kérdésben igen jogos az óvatosság, bár kétségtelen, hogy a hálózati modellek nagyon megkönnyítették az egyes tudományterületek közötti kommunikációt, minek következtében egyes alapfogalmakat, megközelítéseket több tudományterületen is ki lehetett próbálni.

Munkám során három területen próbálkoztam meg a hálózati modellek alkalmazásával. Az első a hálózatok stabilitásának, robusztusságának területe. Itt azt a kérdést vizsgáltam, hogy vajon a több ponton való gyengébb támadás hatékonyabb lehet-e, mint az egy ponton való erős támadás. Ez a kérdés a gyógyszer-alkalmazási stratégiáknál fontos, hiszen a modern gyógyszerek általában egy ponton történő specifikus támadásra alkalmasak, ugyanakkor a természetben és a gyógyításban egyaránt ismeretes, hogy a több irányú hatással rendelkező gyógyszerek illetve a koktél-terápiák nagyon sikeresek. Modellkísérleteim arra utalnak, hogy a többpontos támadás hatékonysága általános, és a hálózatok topológiájától független. Feltehető tehát, hogy a gyakorlatban is érdemes ilyen beavatkozásokat tervezni, és ebben a hálózati modellek is segítséget nyújthatnak.

Második alkalmazási területem a fehérjék konformációjának kialakulásával, konkrétan a diszulfid-hidak képződésével (oxidatív folding) kapcsolatos. Ismeretes, hogy az oxidatív folding intermedierjei az ún. diszulfid-kicserélési reakciók során alakulhatnak egymásba. Ezen reakciók rendszerét szisztematikusan fel lehet írni, és az elképzelhető termékek így egy sokpontos hálózatot alkotnak. Ezen hálózatok feltérképezése és megjelenítése pedig lehetővé teszi, hogy a kísérletileg megismert intermedierek segítségével felírjuk a natív állapot kialakulásának pontos útvonalát.

Harmadik alkalmazási területem egy leíró jellegű munka. Kutatócsoportunk egyik programja a fehérjék hasonlósági viszonyainak leírásával foglalkozik, és ezek megjelenítésére már viszonylag régóta alkalmazzák a „hasonlósági hálózat” fogalmát. Munkám célja itt az volt, hogy az általános, kvalitatív jellegű leíráson túlmenően

kvantitatíve is próbáljam meghatározni a fehérjék hasonlósági hálózatainak jellemzőit, felhasználva mindazokat a módszertani eszközöket, amelyeket a hálózati modellek új irányzatai megteremtettek.

Dolgozatomban e három kutatási terület eredményeit foglalom össze. A dolgozat első fejezete a hálózati modellek alaptulajdonságait és a választott alkalmazási területek szükséges ismereteit foglalja össze. Ezen összefoglalók egy része – a hasonlóságokkal kapcsolatos fejezet – önálló áttekintő jellegű közleményként is megjelent. A dolgozat eredményeit taglaló második rész három fejezetre oszlik, mindegyik a fent említett kutatási témák egyikéttekinti át. A dolgozat az eredmények összefoglalásával és az irodalomjegyzékkel fejeződik be, függelékben csatolom a témáról megjelent közleményeimet is.

Irodalmi áttekintés

Gráfok és hálózatok

A gráfelmélet kezdetei

A gráfok fogalmát Euler alkotta meg. 1736-ban ő írta azt a cikket, amelyet az első gráfelméleti témájú publikációként emlegetnek (3). Azt vizsgálta, hogy lehet-e tenni egy olyan sétát Königsbergben, amely során minden hídon pontosan egyszer haladunk keresztül.

1. ábra Königsbergi hidak

Ugyanannak a struktúrának a különböző megfogalmazásai. Minél kevesebb a részlet, annál egyértelműbb a struktúra.

A kérdés természetesen nem bírt semmilyen gyakorlati jelentőséggel, ebből kifolyólag a rá adott nemleges válasz sem. A cikk jelentősége sem a probléma megoldásában rejlik. Sokkal inkább abban, hogy újfajta szemléletmódot és terminológiát használt. Ez a bizonyos újfajta szemlélet abból állt, hogy Königsberg négy különálló részét

csomópontoknak tekintette, a hidakat pedig a csomópontok közötti kapcsolatoknak. Már csak arra a nyilvánvaló megfigyelésre volt szükség, hogy a bejárás során a séta kiindulópontjának és végpontjának kivételével minden csomópontra igaz az, hogy minden alkalommal két hídra van szükség egy városrész meglátogatásához. Mivel Königsbergnek több mint két olyan városrésze volt, amely páratlan számú kapcsolattal rendelkezik, lehetetlen volt olyan útvonalat találni, amely minden hídon pontosan egyszer halad át.

Euler kora óta a gráfelmélet a matematika egyik jól kiművelt ágává vált, melynek klasszikus tankönyvei és monográfiái vannak. Ezt a fejlődést áttekinteni itt nem áll módunkban, csak a szakterület néhány fontos tankönyvére utalhatunk. (4), (5), (6), (7), (8) Érdekességként szeretném megemlíteni, hogy ez a terület már a korai időszakban összefonódott a molekulastruktúrák fogalmával, a XIX. század gráfelméletében kiemelkedő Cayley munkásságát például éppen a kémiai képletek megjelenése inspirálta. Jóllehet a kémiai gráfelmélet ma fontos gyakorlati alkalmazási területnek számít, a gráf-modellek mai reneszánsza azonban mégsem ennek, hanem az Internet, a kommunikációs hálózatok és a biológiai hálózatok iránti óriási érdeklődésnek köszönhető. Mielőtt áttekintjük ennek az alapvetően nem-matematikai indíttatású fejlődésnek néhány állomását, le kell írnunk néhány alapfogalmat.

Gráfok és hálózatok jellemzése

Kezdjük egy terminológiai megjegyzéssel: a gráf fogalma a matematikai irodalomban alakult ki. Az újabb fizikai, számítástechnikai, biológiai alkalmazások során a nagyobb gráfokra inkább a hálózat nevet alkalmazzák.

A mai matematika a következőképpen fogalmaz a gráfokkal kapcsolatban: Adott egy V halmaz, ez tartalmazza a csúcsokat, vagy csomópontokat, és adott egy $E \subseteq V \times V$, vagyis csúcspárokat tartalmazó halmaz, ami az élek, vagy kapcsolatok halmaza. A V és E halmaz együtt leír egy gráfot, vagyis elmondja mik a csúcsai, és mely csúcsai között van él. Ha a kapcsolatok minden esetben kölcsönösek, a gráfot irányítatlannak, ellenkező esetben irányítottnak nevezzük. Súlyozott gráfok esetén az $E \subseteq V \times V \times \Re$, vagyis minden élhez tartozik egy valós szám is, ami leírhatja a kapcsolat erősségét, vagy a két csomópont közötti távolságot. Hurokélnek nevezzük azokat az éleket, amelyek egy csúcsot saját magával kötnek össze.

A jelző nélkül alkalmazott „gráf" megjelölés általában az irányítatlan, hurokélektől mentes, nem súlyozott gráfra vonatkozik, jelzővel megjelölni az ettől való eltérést szokás (súlyozott gráf, irányított gráf, stb.).

Abban az esetben, ha minden csúcspár között csak egy él lehetséges, egyszerű gráfról beszélünk. Ebben az esetben a hálózatot alkotó kapcsolatokat mátrix alakban is megadhatjuk, amikor is az *i*-ik sor *j*-ik eleme mutatja meg, hogy az *i*-ik csúcs és a *j*-ik csúcs között van-e kapcsolat. Súlyozatlan esetben 0 jelzi, hogy nincs kapcsolat, 1 pedig,

azt hogy van. Súlyozott esetben az élek súlyai kerülnek a mátrixba. Ez a mátrix a szomszédsági mátrix. Irányítatlan gráf esetén szimmetrikus. A szomszédsági mátrix a csúcsok sorszámozásától eltekintve egyértelműen meghatározza a hálózatot, de egy hálózathoz tartózó szomszédsági mátrixban sor és oszlopcseréket eredményezhet a csúcsok átszámozása.

A gráf csúcsaihoz, éleihez esetleg egyszerre mindkettőhöz rendelhetünk színeket is. (Tulajdonképpen a megfelelő pontot vagy vonalat beszínezzük.) Ezt általában számozással szokás megvalósítani, de részint hagyományból, részint pedig a súlyozástól való megkülönböztetés kedvéért a gráf színezésének nevezzük. Egy színezés akkor „helyes" színezés, ha nincs két szomszédos csúcs, azonos színnel. Élek esetén hasonlóan, nincs két él, aminek azonos a színe, és ugyanabból a csúcsból indulnak. Gyakran a helyes színezést egyszerűen csak színezésként említik.

Egy olyan súlyozott irányított gráfot, amelyben van egy pont, amiből csak kifele indulnak élek (forrás), és egy olyan, ahová csak befutnak (nyelő), folyamnak szokás nevezni. Ennél a megközelítésnél az alapkérdés az, hogy ha a súlyokat áteresztő-képességként értelmezzük, mekkora a teljes folyam áteresztőképessége.

Mint arra a hasonlóságokról szóló fejezetben bővebben is kitérek, a gyakorlatban a gráfok összehasonlítása többnyire nem-strukturált, összetételszerű leírásokon keresztül történik. Az alábbiakban szinte csak felsorolásszerűen áttekintem azokat a mennyiségeket, amelyek a hálózatok jellemzésére és összehasonlítására ma a leggyakrabban használatosak.

A legegyszerűbbek talán a csúcsok száma (jelölés: $|V|$) azaz hálózatot alkotó csomópontok száma, és az élek száma (jelölés: $|E|$) azaz a csomópontok közötti kapcsolatok száma. Egy csúcsnak szomszédja egy másik, ha éllel össze vannak kötve.

Egy csúcs fokszáma („degree”, jelölés: d(v), vagy k_i) a belőle induló élek száma, vagyis hogy hány kapcsolattal rendelkezik. Az átlagos fokszám, (jelölés: <k>) azaz $|V|/(2*|E|)$ példáján szeretném bemutatni, hogy általában, ha egy x mennyiséget minden élre vagy csúcsra ki lehet számolni, akkor x átlagát szokás <x> jelöléssel ellátni. Irányított gráfok esetén megkülönböztethetjük egy csúcs éleit aszerint, hogy azok a csúcsból indulnak, vagy oda érkeznek. Ez alapján beszélhetünk kifokról és befokról („in degree, out degree”). A kettő összege adja a teljes fokszámot. Súlyozott gráfok esetében a fokszám mellett figyelembe vehető a csúcsból induló élek összsúlya, vagyis a csúcs ereje („strength”, jelölés: s_i). Irányított súlyozott gráfok esetén kifelé- és befelé mutató élekből számított erő is definiálható, bár ritkán van gyakorlati jelentősége.

Az első jelentősebb, gyakorlatban is sokat használt leírás a fokszámeloszlás. A fokszámok elméleti eloszlását csak akkor tudjuk teljes bizonyossággal megállapítani, ha ismerjük a keletkezés hátterét, de egy empirikus becslést adhatunk a fokszámok gyakoriságának (hisztogram) ábrázolásával. Súlyozott hálózatok esetén ábrázolhatjuk, hogy egyes fokszámokkal rendelkező csúcsokhoz tartozó éleknek mekkora az átlagos súlya, vagyis a fokszám függvényében az átlagos erőt. Ha ez jelentősen eltér a k*<w>/<k> értéktől (<w> a hálózat éleinek átlagos súlya), vagyis attól, amit véletlenszerű súlykiosztás esetén várunk, az valamilyen struktúrát feltételez.

Egy hálózat egy tulajdonság szempontjából lehet asszortatív, vagy dizasszortatív (9), (10). Ennek megállapítására a csúcsok fokszámának függvényében kell ábrázolni az illető tulajdonságnak a csúcsokszomszédaira számított átlagos értékeit. Ha ez növekvő görbe, akkor a hálózat asszortatív, ha csökkenő trendet mutat, (ekkor általában negatív kitevőjű exponenciális függvény illeszthető rá,) akkor dizasszortatív.

Lehetséges, hogy egy hálózatban jellemzően nem kapcsolódnak egymáshoz a nagy fokszámú csúcsok, avagy éppen fordítva. Ezt próbálja leírni a fokszám-asszortativitás, vagyis a szomszédok átlagos fokszáma. A fokszám-asszortativitás segítségével képet alkothatunk arról, hogy milyen irányban és mennyire jelentős az eltérés. Súlyozott hálózat esetén a csúcs affinitásának nevezzük a szomszédai összerejét (9).

A szomszédsági mátrix sajátértékei adják a hálózat spektrumát. Gráfelméleti jelentősége igen nagy, azonban a hálózatok világában kevésbé népszerű. Igaz ugyan, hogy a szomszédsági mátrix sajátértékeinek és sajátvektorainak alapján egyértelműen rekonstruálható egy gráf, azonban a spektrum tulajdonságai és a gráf strukturális tulajdonságai között az összefüggések nem egyértelműek.

Ha az egyik csúcsból élek sorozatán el tudunk jutni egy másik csúcsba, akkor ez a két csúcs között egy út. Egy út hossza az őt alkotó élek száma, illetve súlyozott esetben az élek súlyainak összege. Két csúcs távolsága (jelölés: d(u,v)) az őket összekötő legrövidebb út hossza. Ha nincs olyan élsorozat amin el lehetne jutni egyik csúcsból a

másikba, akkor a távolságuk végtelen. Amennyiben minden csúcspárra véges a távolság, értelmezhetjük az átlagos távolságot (jelölés: *l*), mint ezek számtani közepét.

Abban az esetben, ha a hálózat több olyan darabból áll, amelyek egymással nincsenek összeköttetésben az átlagos távolság helyett számolhatunk efficienciát. Ekkor nem a legrövidebb utakat, hanem azok reciprokait átlagoljuk. Ha nincs út két pont között, akkor annak hozzájárulását nullának értelmezzük. Képletszerűen:

$$E = 2\frac{\sum_{i,j}\frac{1}{d(i,j)}}{N(N-1)} \qquad [1]$$

Így mindig egy 0 és 1 közötti mennyiséget kapunk (11) (12).

A klaszterezettségi együttható (jelölés: c(v)) egy csomópont esetében definíció szerint azoknak a három élből álló utaknak a száma, amelyek ebből a csomópontból indulnak, és ide is térnek vissza, osztva azzal, ahány ilyen út lenne abban az esetben, ha a csúcs bármely két szomszédja közt lenne él.

$$C(i) = \frac{\sum_{j,h} a_{ij} a_{jh} a_{hi}}{k_i(k_i - 1)} \qquad [2]$$

Ahol *A* a szomszédsági mátrix, k_i pedig az *i* csúcs fokszáma. Súlyozott esetben a

$$C^w(i) = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{jh})}{2} a_{ij} a_{jh} a_{hi} \qquad [3]$$

képlettel általánosíthatjuk az előbbi definíciót. Itt *W* a súlymátrix, s_i pedig az *i* csúcs ereje. Ez utóbbi mennyiség szemléletes jelentése az, hogy mennyire sűrűn helyezkednek el az élek egy csúcs szomszédai között, vagy mekkora eséllyel ismeri egymást két ismerőse.

A hálózat állapotának felmérésére szokás használni a klaszterek számát, illetve a klaszterek méreteinek eloszlását. Ez a mennyiség nem csak összefüggő hálózatokra informatív, és nem tesz különbséget összefüggő és nem-összefüggő hálózatok között.

Véletlen gráfok

A gráfelméletben az 1950-es években bekövetkezett szemléletváltás Erdős Pál és Rényi Alfréd nevéhez köthető (13), (14). Erdős kiemelkedő alakja volt a modern matematikának, 1996-ban bekövetkezett haláláig több mint 1500 cikket publikált. Nagyon sok matematikussal jegyez közös cikket, a matematika számos területén. Részint ennek a ténynek köszönhetően, részint pedig tiszteletadásképpen őt jelölték meg a matematika origójának azzal, hogy megalkották az Erdős-szám fogalmát. Erdős Pál Erdős-száma 0. Minden társszerzőjének 1 az Erdős-száma, ebbe a körbe ma már sajnos lehetetlen bekerülni. Ha valakinek van olyan társszerzője, akinek véges az Erdős-száma, akkor az ő Erdős-száma egyel nagyobb, mint a társszerzői Erdős-számainak minimuma. Nagy valószínűséggel, aki már publikált tudományos folyóiratban véges Erdős-számmal rendelkezik. Elképzelhetőek elszigetelt egyének, vagy kis csoportok, akik nem kapcsolódnak társszerzők sorozatán keresztül Erdős Pálhoz, azonban nem ez a jellemző. Az én Erdős-számom legfeljebb 5, mivel egy láncot tudok képezni Erdős Páltól úgy, hogy az egymást követő személyek mind társszerzők, és a lánc 6-ik helyén én magam állok. (Erdős Pál, Totik Vilmos, Csirik János, Kocsor András, Pongor Sándor, Ágoston Vilmos) Lehetséges, hogy létezik ennél rövidebb lánc is, csak nem sikerült megtalálnom, másrészt az idő múlásával remélhetőleg csökkeni fog ez a szám.

Erdős és Rényi újítása a véletlen gráf volt. Az általuk alkotott modellt máig Erdős-Rényi gráfként említik. A modellben minden él meglétének azonos a valószínűsége. Egy konkrét megvalósulást kaphatunk azáltal, hogy ténylegesen felrajzolunk 100 csúcsot, és minden lehetséges élnél (4950 darab) sorsolunk 1 és 100 között egy számot, és csak akkor rajzolunk fel egy élet, ha a hozzá sorsolt szám pont 1volt. Ezáltal nagyjából 50 éle lesz az elkészült gráfnak, vagyis átlagosan egy csúcsból egy él indul ki. Erdős és Rényi arra a megállapításra jutott, hogy azokban a gráfokban, ahol átlagosan egy csúcsra egy kapcsolat jut, már kialakul egy nagy csoport úgy, hogy a csoporton belül bárhonnan bárhova el lehet jutni kapcsolatok sorozatán keresztül. Igazából egy ennél általánosabb megállapításra jutottak. A gráfoknak sok olyan, az élek számával monoton növekvő valószínűséggel előforduló tulajdonsága van, amiknél jelentkezik egyfajta fázisátmenet. Van egy küszöbérték, aminél kevesebb éllel rendelkező gráfokra egy tulajdonság valószínűtlen, de alig valamivel több él esetén ugyanez a tulajdonság valószínűvé, sőt majdnem biztossá válik. Természetesen a küszöb értéke egyaránt függ a csúcsok számától, és az adott tulajdonságtól. Azt, hogy szinte mindenki, aki rendelkezik publikációval, tudna találni egy társszerzői láncolatot, amin keresztül összekötheti magát Erdőssel, megmagyarázza a fent említett tény, hiszen a legtöbb kutató több mint egy társszerzővel rendelkezik, ezért nagyon kicsi annak a valószínűsége, hogy ebben a rendszerben ne alakuljon ki egy nagy csoport, amin belül mindenkitől mindenkihez el lehet jutni a társszerzői kapcsolatokon keresztül.

A véletlen gráfokkal kapcsolatos kutatások mára a gráfelmélet önálló ágává váltak, ennek a szakterületnek önálló monográfiái (5) és szakfolyóiratai (pl. Random Structures)

vannak. A valóságban azonban a legtöbb kapcsolatrendszer nem véletlenszerű, és nem csak azért lehet ezt kimondani, mert sok esetben ismerjük azt az elvet, ami alapján a kapcsolatok keletkeznek, egyszerűen a gráf struktúrája lényegesen különbözik attól, mint amilyet akkor kapunk, ha véletlenszerűen húzzuk be az éleket.

Szociális hálózatok

Karinthy Frigyes 1929-ben írta meg Láncszemek című novelláját. Ebben a társaság egyik tagja azt a nézetét hangoztatja, hogy a Földön élő bármely emberhez tud létesíteni egy olyan ismeretségi láncolatot, amelynek egyik végén ő áll, a másik végén pedig a kijelölt illető, és sosem lesz szükség ötnél több közbenső láncszemre. Ezzel nem önmagát akarta magasztalni, hanem azt akarta igazolni, hogy a Föld lakossága sokkal közelebb van egymáshoz, mint korábban bármikor. A novellában ez sikerül is, legyen szó akár a Nobel díj egyik nyerteséről, vagy a Ford gyár egy szegecselő munkásáról. Bármilyen érdekes elképzelés volt is ez, majdnem 40 évnek kellett eltelnie hozzá, hogy kísérletileg is megpróbálják igazolni. Stanley Milgram 1967-es kísérletében (15) arra kérte véletlenszerűen kiválasztott alanyait, hogy juttassanak el egy adott emberhez, egy adott címre egy küldeményt, de nem közvetlenül, hanem úgy, hogy továbbküldik egy ismerősüknek, akiről úgy gondolják, közelebb áll az ismeretségi láncolatban az illetőhöz, mint ők maguk. (Az egyik problémás rész, az ismeretség megfogalmazása. Általában úgy szokták definiálni, hogy ismerős az, akit keresztnéven szólítunk. Ez a megfogalmazás többé-kevésbé fedi az ismeretség fogalmát, elég egzakt, és elég egyszerű. Természetesen vannak hibái, de nehéz jobbat találni.) Milgram szinte biztos nem olvasta Karinthy novelláját, legfeljebb hallomásból ismerhette, de ez is valószínűtlen. A kísérlet

eredménye viszont meglepően egybeesik Karinthy sejtésével, ugyanis a csomagok döntő többségét sikerült célba juttatni, méghozzá legfeljebb hat postázással. Ez pont azt jelenti, hogy általában legfeljebb öt köztes láncszemre volt szükség. Milgram kísérletében csak az Egyesült Államokon belüli kapcsolatokat vizsgált, azonban nyilvánvaló, hogy léteznek kontinenseken átívelő ismeretségek, viszont az ezek közötti postázás jelentősen megnövelte volna a költségeket. Milgram bírálói két kifogással éltek. Az egyik pont az volt, hogy magasabb jövedelemmel rendelkező alanyokat választott, akik könnyen megengedhették maguknak a postaköltséget, a másik pedig az, hogy következtetéseit a célba érkezett küldemények alapján vonta le, mivel nem tudta figyelembe venni azokat, amik nem érték el a céljukat. Ha ezt a jelenséget véletlen gráfok segítségével próbálnánk modellezni, azaz vennénk egy akkora véletlen gráfot, amelynek csúcsszáma összemérhető a Föld (vagy az Egyesült Államok) lakosságával, átlagos fokszáma pedig annyi, mint amennyi ismerőse van átlagosan az embereknek, és ebben a gráfban keresnénk a legrövidebb utat két véletlenszerűen kiválasztott csúcs között, nagyságrendekkel nagyobb számokat kapnánk, mint Milgram kísérletében. A kísérletben résztvevők egy nagyságrenddel nagyobb eredményre számítottak, legtöbbjük több mint száz köztes lépést várt. Ebben nem az a meglepő, hogy az emberek ismeretségei nem teljesen véletlenszerűek, ezt korábban is tudták. Nyilván nagyobb valószínűséggel ismeri meg egymást két ember, ha van közös ismerősük, és ez önmagában kizárja a véletlenszerűséget. A kísérlet igazi érdekessége sokkal inkább az, hogy az emberek ismeretségi hálózata strukturált, és ez a struktúra azt eredményezi, hogy az emberek ismerőseiken keresztül kevés lépésben szinte bármelyik másik emberhez el tudnak jutni.

A szociális hálózatok kutatása ma népszerű kutatási irány, melyet szakkönyvek sora foglal össze (16). Jellemző alkalmazási területei közé tartozik a társszerzői hálózatok vizsgálata, az Internet felhasználói közösségeinek, levelezőlistáinak tanulmányozása, de az állat-közösségek és a vállalatszervezés problémáinak leírása is.

Kis világok

A későbbiekben a hálózatok vizsgálata során egyre inkább körvonalazódott az az elképzelés, hogy kifejezetten ritkák a valóságban a véletlenszerű hálózatok, a legtöbb rendelkezik valamiféle belső rendezettséggel. Watts és Strogatz 1998-ban (17) állt elő egy matematikai modellel, amely megmagyarázhatja a jelenséget. Ők a hangsúlyt arra helyezték, hogy a hálózatokban tapasztalt ”kis világ ” jelenségre adjanak magyarázatot. Az elnevezés Milgramtől származik, és arra utal, hogy az átlagos úthossz lényegesen kisebb, mint a hasonló méretű és sűrűségű véletlen gráfban. A modell dióhéjban abból áll, hogy egy szabályos szerkezetű, rács struktúrájú gráfból kiindulva az élek egy részét véletlenszerűen áthelyezzük. Ha az áthelyezett élek számát változtatjuk a modellben, változik az átlagos úthossz is. Ha elég sok élet helyezünk át, teljesen megszűnik a gráf struktúrája, és véletlen gráfot kapunk. A köztes állapotban azonban jelentősen csökken az átlagos úthossz.

Skálamentes fokszámeloszlás

Barabási Albert-Lászó publikációjában (18) a hálózatok egy másik tulajdonságára fókuszál, nevezetesen arra, hogy jelentős részüknek a fokszám-eloszlására jól illeszthető valamilyen hatványfüggvény. Ezt Barabási skálamentességnek nevezi, mivel a hatványfüggvény grafikonja változatlan marad akkor is, ha a koordinátarendszert átskálázzuk. Ebből ered az a jelenség, hogy ha mindkét tengelyen logaritmikus skálázást alkalmazunk, a hatványfüggvény képe egyenes lesz. A véletlen gráfok fokszám-eloszlása binomiális, ami exponenciális lecsengésű, de természetesen nem az a legfontosabb konklúzió, hogy a hálózatok nem teljesen véletlenszerűek, hanem az, hogy nagyon soknak hasonló a struktúrája, legalábbis van valami hasonlóság bennük. A modellezések alapján azt mondhatjuk, hogy a gráfok hatványfüggvény lecsengésű fokszám eloszlása kis világot eredményez, vagyis jelentősen csökken az átlagos úthossz. Ez azonban felveti azt a kérdést, hogy mi okozza az eloszlásfüggvények hasonlóságát. Barabási modelljében a főszerepet az általa preferenciális kötődésnek nevezett jelenség játssza. Eszerint a hálózat épülése, esetleg változása során az újonnan beépülő csúcsok szívesebben kötődnek olyan csúcsokhoz, amelyeknek sok kapcsolata, avagy nagy fokszáma van.

A fokszámeloszlás, a gráfparaméterek és a topológia összefüggéseire több mint 50 év elteltével sem sikerült általános magyarázatot adni, a mai napig születnek új, speciális modellek, napról napra finomodik a kép. Valószínű, hogy nem lesz olyan egységes modell, ami minden hálózat keletkezését, dinamikáját megfelelően le tudná írni, mivel ismereteink szerint ezek igen eltérőek. Ezek az eltérések bizonyos paraméterekben is

megfigyelhetőek. Példaként említeném, hogy az eloszlás-függvényekre illeszthető hatványfüggvényben a kitevő általában -2 és -3 között mozog fizikai, vagy ember alkotta hálózatok esetében, biológiai hálózatok esetében viszont jellemzően -1 és -2 közé esik.

Az utóbbi években szélesedett a vizsgált jelenségek köre is. Új fogalomként jelent meg például a hálózat *támadhatósága*, vagyis annak számszerűsítése, hogy a hálózat milyen kis részének eltávolításával lehet tönkretenni, funkcióképtelenné tenni a hálózatot. Ehhez a jelenségkörhöz kapcsolatos mennyiség a *hibatűrés*, melynek vizsgálatakor a hálózatban véletlenszerű hibákat idézünk elő. Az Internettel kapcsolatban fontossá vált annak vizsgálata is, hogy a hálózatok milyen struktúrája segíti, hogy a hasznos információk terjedhessenek benne, ugyanakkor korlátozni tudjuk a zavaró információk, esetlegesen a vírusok terjedését, illetve, hogy ugyanezt egy adott struktúra mellett milyen külső behatásokkal lehet elérni.

Biológiai modellek

Kölcsönhatási hálózatok

A biológiai hálózatok fogalmán ma leginkább a génszabályzási kölcsönhatások, a metabolikus útvonalak, a fehérje-kölcsönhatások hálózatait értjük, mivel ezekre vonatkoznak a molekuláris biológia és a genomkutatás legfontosabb adatai. A genom-elemzések során olyan mennyiségű kölcsönhatási adat keletkezik, hogy azok hálózatai papíron ceruzával, vizuális kiértékeléssel nem is vizsgálhatók. A főbb adattípusok a következők:

Hálózat	Csúcsok	Élek
Génszabályozás	Gének	Szabályzó hatás
Metabolizmus	Metabolitok (szubsztrátok)	Reakciók (enzimek)
Fehérjekölcsönhatás	Fehérjék	Fizikai kötődés

Már első pillantásra is látható, hogy az adatok rendkívül heterogének. A génszabályozás adatait például részben hagyományos, alaposan ellenőrzött genetikai vizsgálatok szolgáltatják, részben pedig nagyteljesítményű microarray vizsgálatok. A fehérje-kölcsönhatási adatokat általában nagyteljesítményű proteomikai módszerek, pl. tömegspektrometria vagy kettőshibrid vizsgálatok szolgáltatják. A nagyteljesítményű módszerek adattömegei sok újdonságot ígérnek – ugyanakkor nagy hibával terheltek, a fehérje-kölcsönhatási adatoknál például feltételezhetően a kölcsönhatások mintegy 10-

20%-a műtermék. A spektrum másik végén vannak a metabolikus hálózatok, ezen kölcsönhatások mindegyikét részletes laborvizsgálatokkal igazolják, ezek azonban nagyon időigényesek, így valójában csak nagyon kevés organizmus metabolikus hálózatai megbízhatóak. Mindebből többféle probléma is adódhat. Fel szoktuk tételezni például, hogy az újonnan analizált organizmusok ugyanúgy metabolizálnak, ezért lehetséges, hogy sok egyéni metabolikus megoldásai egyszerűen kimaradnak vizsgálatainkból. De problémát jelenthet az is, hogy a metabolizmus klasszikus, részben akár száz éves eredményei nem tartalmaznak minden kölcsönhatási információt. Valójában tehát itt sem rendelkezünk teljesen megbízható adatokkal.

A biológiai modellek vizsgálatában alapvető volt Barabási Albert László, Oltvai Nagy Zoltán és munkatársaik munkássága, akik először hívták fel a figyelmet arra, hogy a biológiai modellek között is általános a skálamentes fokszámeloszlás, és ezt a jelenséget azzal magyarázták, hogy a skálamentes hálózatok stabilisak a véletlenszerű mutációval szemben. Nagy érdeklődésre tartott számot az a felismerés is, hogy minél centrálisabb szerepet tölt be egy gén vagy fehérje a kölcsönhatási hálózatokban, annál esszenciálisabb, annál valószínűbb, hogy mutációja letális hatású (19).

Mindezek a kutatások a hálózati modellek topológiájára vonatkoztak, emellett egyre inkább növekszik a dinamikus leírások szerepe is. Ezekben az idő (lehet folyamatos vagy diszkrét) múlásával a hálózat bizonyos szabályok szerint megváltozik, de minden rögzített pillanatban a fenti modellekkel leírható. Ezeknek a leírásoknak a körébe tartoznak az igen széles körben alkalmazott fluxus-modellek, melyekben a hálózatokon átfolyó „áramok” (anyag, energia stb. áramok) intenzitása változik meg. Ez a terület mind

eszközeit, mind pedig tárgyalásmódját tekintve távolabb esik disszertációm témakörétől, tárgyalására ezért nem térek ki.

Kölcsönhatási hálózatok és gyógyszerhatás-vizsgálatok

Dolgozatom egyik célja, hogy megkísérelje a hálózati modelleket alkalmazni a gyógyszertervezési stratégiák területén. A gyógyszerhatás-vizsgálatoknál mindig a teljes organizmus válaszát vizsgálják, s ebben tehát szükségszerűen benne van az élő rendszerek teljes komplexitása. Adekvát vizsgálati rendszert jelenthetnek-e mindehhez a hálózati modellek? Lehetséges, hogy nem, de a hálózati modellek mindenesetre a komplexitás egy új fokát képviselik, és sok olyan, matematikailag nehezen leírható tulajdonsággal rendelkeznek, amelyek az élő rendszerek sajátságai. Ilyen fogalom például a robusztusság, amelyen azt értjük, hogy a rendszer ellen tud állni a véletlenszerű vagy célzott támadásoknak.

Konkrétan egy olyan jelenségre kerestük a választ, hogy mi az oka annak, hogy a nagy specifitású, de egyetlen molekuláris célpontra ható gyógyszerek sokszor kevésbé hatékonyak, mint a több célpontra ható kezelések. Kiindulópontunkat részletesen a következőképpen fejthetjük ki. A hatóanyag-fejlesztési stratégiákat alapvetően befolyásolta az a tény, hogy a genom-projektek révén nagyszámú új potenciális gyógyszercélpontot ismertünk meg. A jelenlegi módszerek alapelve a következő : 1) Találnunk kell egy megfelelő funkciójú gyógyszercélpontot; 2) Azonosítanunk kell a hozzá legerősebben kötődő hatóanyagot (best binder), ezt általában nagyméretű kombinatorikus kémiai könyvtárak „high throughput” tesztelésével, és/vagy a célpont

(fehérje vagy nukleinsav) háromdimenziós struktúrája alapján történő hatóanyag tervezéssel oldják meg; 3) Fel kell tudni mutatni néhány bizonyító erejű alapkísérletet (proof of principle); 4) Ki kell fejleszteni egy technológiai platformot, ami egyúttal előrevetíti a lehetséges klinikai alkalmazásokat is. Bárhogy is legyen, az összes alapos tanulmány és a gyógyszerfejlesztés területén tett figyelemre méltó erőfeszítések ellenére, a hatásos gyógyszerek száma nem növekedett számottevő mértékben az elmúlt évtizedben. Ez igen nagy ellentmondásban van azzal a közismert ténnyel, hogy a gyógyszeripari kutatások költsége nagymértékben növekedett, és ezen felül azt sincs okunk feltételezni, hogy a korszerű módszerek valóban megtalálják kiszemelt molekuláris célpontjaik hatékony gátlószereit.

Ugyanakkor több jel mutat arra, hogy a hatékony gyógyszereknek nem csak egy hatása kell legyen, mint azt számos igen hatásos gyógyszer példája mutatja: a nem szteroid alapú gyulladáscsökkentők (NSAID-ok), a szalicilát, a metformin vagy a Gleevec viszont több célpontot befolyásol párhuzamosan. Szintén több célpontra irányuló kombinatorikus (koktél) terápiát is egyre szélesebb körben alkalmazzák sokfajta betegség elleni küzdelemben, például az AIDS, a rák, vagy az atherosclerosis. A kígyók és pókok mérgei egyaránt többkomponensű rendszerek, és a növényvilág kórokozók elleni eszköztárában is előfordulnak többkomponensű rendszerek. Ennek alapján a molekulák kombinációjának használata evolúciós sikertörténetnek tűnik. Végül pedig, a hagyományos (természet)gyógyászatban is gyakori a természetes összetevők többkomponensű kivonatainak használata. Ezek alapján tehát megfontolandó, hogy a szisztematikus gyógyszerfejlesztési stratégiának több célpontot kellene figyelembe vennie, és hogy ez az új gyógyszerfejlesztési paradigma esetleg több és hatékonyabb

molekulát eredményezne a fejlesztés során, mint a jelenleg támogatott egyszeres hatású gyógyszerek.

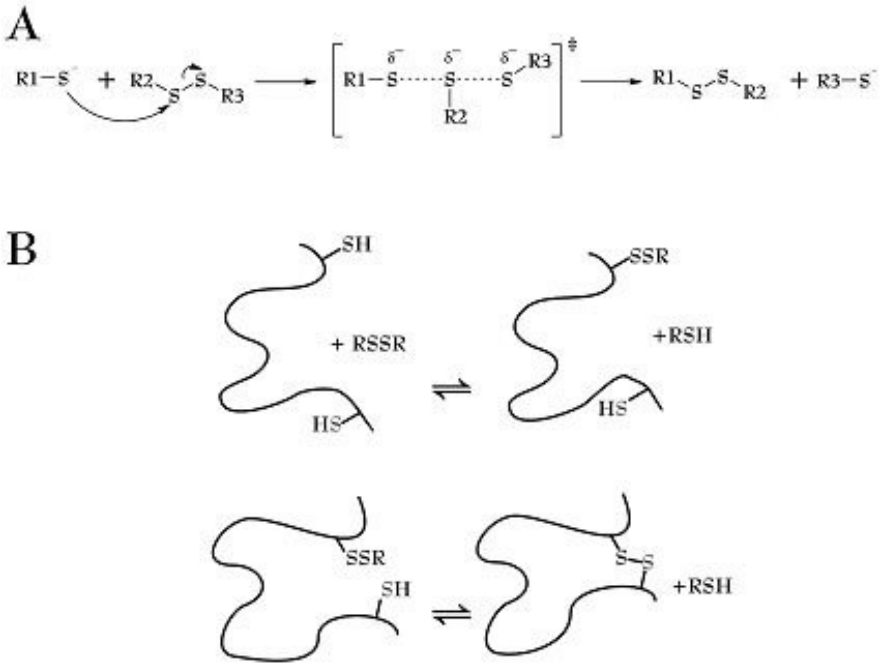
Munkám során egyik feladatom az volt, hogy modellszámításokat végezzek arra, mennyire reális elvárni azt, hogy a többpontos támadás hatékonyabb lehet, mint az egyetlen ponton való beavatkozás.

Hálózati modellek a fehérje-folding kutatásában.

A fehérjehajtogatódás folyamata, vagyis az a folyamat, amely során egy lineáris fehérjeláncnak kialakul a natív szerkezete, az elmúlt 50 év egyik legintenzívebben kutatott biomolekuláris problémája (20), (21), (22). A fehérje hajtogatódását első közelítésben úgy képzelhetjük el, mint a biológiai aktivitást hordozó natív konformációs állapot keresését egy, az összes konformációs állapotot magába foglaló konformációs térben, ahol minden konformációs állapotot számos paraméter jellemez. Habár a legtöbb konformációs állapot (azaz az ezeknek megfelelő molekulák) fizikailag nem izolálhatók, a potenciális energia felületének kvalitatív, grafikus ábrázolása mégis kulcsszerepet játszott annak megértésében, hogy a konformációs tér hogyan szűkül be fokozatosan a hajtogatódás folyamán (23). Olyan kulcsfogalmak, mint például a hajtogatódási útvonal (folding pathway) (24) szintén grafikus ábrázolások segítségével magyarázhatóak el a legkönnyebben.

A hajtogatódás témakörén belül az oxidatív hajtogatódásra szeretnék részletesebben kitérni, ami a natív diszulfid kötések kialakulását és a konformációs hajtogatódást

egyaránt magába foglalja. Ezt a komplex folyamatot alapvetően kétfajta kölcsönhatás irányítja: egyrészről a nem kovalens kölcsönhatások létrehozzák a másodlagos és harmadlagos struktúrát, másrészről kovalens kötések alakulnak ki a cisztein oldalláncok között, melynek eredményeként végül kialakulnak a natív diszulfid hidak. A diszulfid kialakulása egyszerű kémiai reakció, amiben két SH csoport egyesül, és így alakul ki a diszulfid kötés. (2 A ábra) Ha az SH csoport egy polipeptid láncon helyezkedik el, akkor az *in vitro* reakció előidézhető külső redox rendszer hozzáadásával, például oxidált és redukált glutation, vagy cisztein és cisztin keverékével. In vivo az oxidatív hatás különböző speciális mechanizmusoknak köszönhetően jön létre, például olyan molekuláris chaperonoknak, dajkafehérjéknek, mint a diszulfid izomeráz (25).



2. ábra Diszulfid kötés

A) A diszulfid kötés kialakulása B) Kicserélődési reakció fehérjelánc és egy oldatban lévő redox rendszer (RSH, RSSR, pl. cisztein-cisztin) között. Az eredmény a kinetikusan stabilis, intramolekuláris SS-kötés (jobbra lent)

A jelenség alapjául szolgáló kémiai mechanizmus a diszulfid kicserélődési reakció (2 B ábra). A vázlat kétfajta kicserélődési reakciót ábrázol: 1) a redox reakcióban egy diszulfid kötés keletkezik (vagy szűnik meg) vagyis a polipeptid oxidatív állapota változik. Ez az eset akkor áll fenn, amikor a reakció egyik résztvevője (például az RSH) egy külső reagens, tehát nem a fehérje tartalmazza. 2) A keverési reakcióban a reakció mindkét résztvevője a fehérjéhez kötött, vagyis a polipeptid oxidatív állapota a kicserélődés során nem változik. Ezeket a lehetőségeket figyelembe véve nyilvánvaló, hogy a hajtogatódási folyamat során nagyon sokféleképpen kialakulhatnak és átrendeződhetnek a diszulfid hidak. Manapság általánosan elfogadott, hogy a nem kovalens kötések irányítják a hajtogatódási folyamatot, és a diszulfid hidak rögzítik a fehérjét a helyes konformációban. Az oxidatív hajtogatódás kutatásának előnye, az általános hajtogatódás vizsgálatával szemben az, hogy diszulfid intermedierek kémiailag izolálhatóak és tanulmányozhatóak, olyan egyszerű kémiai technikák segítségével, mint pl. az intermedierek savas pH-n történő „csapdába ejtése" (acid trapping). Az intermedierek analízise pl. enzimatikus hasítások és tömegspektrometria kombinált alkalmazásával történik. Az irodalomban egyre nagyobb számban találhatóak diszulfid intermedierek segítségével leírt oxidatív útvonalak (26), (27), (28), a célom az, hogy megmutassam, hogyan lehet ezeket vizualizálni a gráfelméleti eszközök segítségével.

A gráfelméletet a fehérjekutatás számos területén alkalmazták. (Egy máig aktuális áttekintés (29) jelent meg 2000-ben.) Az alkalmazások két fő iránya a következő:

1) A fehérje szerkezete egy gráfnak tekinthető, amelynek éleit különböző kölcsönhatások (kovalens kötések, hidrogén hidak, térbeli érintkezések, stb.)

alkotják, a csúcsai pedig atomok, vagy aminosavak. Az egyik klasszikus definíciója a fehérje másodlagos szerkezetének például a fő és oldalláncok közötti hidrogénkötések leírásán alapszik (30). Strukturális hálózati leírásokat alkalmaztak a hajtogatódás kutatásában is. Többek közt olyan megállapításra jutottak, hogy az ún. „kontakt sorrend" (contact order), vagyis a 3D szerkezetben közel lévő aminosavak átlagos szekvencia-távolsága, a hajtogatódás sebességét alapvetően meghatározza (31). A kutatások egy másik iránya az atomközi kapcsolatok olyan speciális hálózataival foglalkozik, amelyek stabilizációs centrumokat alkothatnak a fehérje szerkezetében (32), (33). Molekuláris dinamikai számításokban azt is észrevették, hogy bizonyos aminosavak jellegzetes lokális hálózatokat hoznak létre a szimuláció során (34), (35).

2) A hajtogatódás konformációs terének („folding space") hálózatos leírásában a hajtogatódási állapotok alkotják a csúcsokat, a lehetséges átalakulások pedig az éleket. A hálózati leírás implicite jelen volt már a hajtogatódás kutatásainak kezdeteitől, hiszen a „folding pathway" már említett fogalma – és így maga a jól ismert Lewenstein-paradoxon is – egy ilyen hálózati modellen belül értelmezhető. A további munkát nagyban inspirálta Barabási már említett felvetése, amely szerint a hálózatok robosztussága és stabilitása a kapcsolatok topológiájával magyarázható (36). Az ezt követő években a fehérjeszerkezet kialakulásának kutatásában is megkísérelték a gráfelméleti módszerek alkalmazását. Scala és munkatársai rácsmodellen végzett Monte Carlo szimulációval írta le rövid peptidek hajtogatódási állapotait (37). Azt találták, hogy a hálózat geometriai tulajdonságai hasonlóak a „kis világ" hálózatokhoz, azaz a hálózat átmérője az

állapotok számának logaritmusával arányosan növekszik, miközben lokálisan a hálózat alacsony dimenziójú marad. Shakhnovich munkatársaival analizálta a fehérjék mulekuláris dinamikai szimuláció által kapott hajtogatódási állapotait. Azt találták, hogy a hajtogatódási tér skálamentes hálózatra emlékeztet, azáltal, hogy a tér nagy részében alig található hajtogatódási állapot, eközben néhány kis részen pedig sűrűn helyezkednek el, ami a más rendszerekben található „csomópontokra” (hub-okra) emlékeztet (38).

Hasonlósággal kapcsolatos matematikai fogalmak

Hasonlósági mértékek, hasonlósági hálózatok

A hétköznapi értelemben vett hasonlóság definiálásának, és számszerűsítésének nagy jelentősége van a tudomány sok területén. Egy példán keresztül szeretném szemléltetni: A fehérjék esetében megállapítható, hogy a szekvencia hasonlósága, és a térszerkezet hasonlósága erős korrelációt mutat. Ezáltal lehetővé válik, hogy egy új, csak szekvenciájában ismert fehérjének a térszerkezetéről adjunk becslést oly módon, hogy megkeressük az ismert térszerkezetű fehérjék közül azt, amelyikkel a legnagyobb szekvencia-hasonlóságot mutat, és feltételezzük, hogy az ennek megfelelő ismert térszerkezet tartozik a vizsgált szekvenciához. Ezáltal nyerhetünk képet arról, hogy az új fehérjénk nagyjából milyen funkcióval rendelkezik, vagy ez a térszerkezet egy molekulamodellezés esetén számítások kiindulópontjaként szolgálhat. Ahhoz a megállapításhoz azonban, hogy ha két fehérjének hasonló a szekvenciája, akkor nagy valószínűséggel hasonló a térszerkezete, elengedhetetlen, hogy tudjuk mit értünk hasonlóság alatt, és ezt számszerűsíteni is tudjuk.

A dolgozatban három modell játszik alapvető szerepet. A szekvenciák, a 3D szerkezetek és a hálózatok. Ezek közötti hasonlóságokról lesz szó az alábbiakban.

Ekvivalenciák

Az ekvivalencia relációk (jelölés: "≅") az általánosan használt azonosság fogalom megfelelői. Szigorú értelemben véve a fizikai világban minden csak önmagával azonos; mi azzal az esettel fogunk foglalkozni, amikor azonos a matematikai leírás.

Az ekvivalencia relációk matematikai definíciója három tulajdonságot követel: reflexivitás, szimmetria, tranzitivitás. Egy reláció reflexív, ha A ≅ A minden A esetén, szimmetrikus, ha A ≅ B-ből következik B ≅ A, és tranzitív, ha A ≅ B-ből és B ≅ C-ből következik A ≅ C. Jelölje [A] a ≅ szempontjából A-val ekvivalens elemek halmazát. Ha B egy másik elem, akkor matematikailag bizonyítható, hogy [A] és [B] vagy az elemek ugyanazon halmazát jelöli, vagy a két halmaznak nincs közös eleme. Az [A] halmazt ekvivalencia-osztálynak nevezik. Például két molekulát egy lehetséges definíció szerint akkor és csak akkor tekintünk azonosnak, ha az aminosav-sorrendjük megegyezik. Megjegyzendő, hogy az "azonosság" egy adott leírásra vonatkozik; ebben a példában például figyelmen kívül hagyjuk az esetleges poszttranszlációs módosítások által okozott különbségeket.

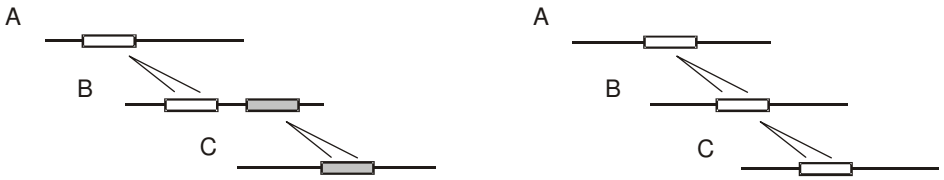
Részbenrendezések

A részbenrendezés relációk a “részstruktúrája valaminek”, “része valaminek” kifejezések megfelelői. Egy ”≤” reláció részbenrendezés ha reflexív, antiszimmetrikus és tranzitív. A reflexivitás és tranzitivitás tulajdonságok definíciója már korábban szerepelt. Egy reláció antiszimmetrikus ha A ≤ B-ből és B ≤ A-ból következik, hogy A és B megegyezik. Például ha A ≤ B azt jelenti, hogy A részszekvenciája B-nek, akkor ≤ részbenrendezés.

Tolerancia relációk, általános és specifikus hasonlóság

A tolerancia-relációk azt fejezik ki, hogy két dolognak közös része vagy jellegzetessége van, vagy két struktúrának van egy közös részstruktúrája. Egy ~ reláció tolerancia-reláció, ha reflexív, szimmetrikus, de – ellentétben az ekvivalencia relációval – nem feltétlenül tranzitív. Azaz A~A, A~B-ből következik B~A. A tolerancia-reláció közelebb áll a hasonlóság fogalom általános értelmezéséhez, mégis egy fontos megkülönböztetést kell tennünk. Goldmeier (39) pszichológiai koncepcióját alapul véve akkor nevezhetünk két struktúrát hasonlónak, ha van közös részstruktúrájuk (lásd: 3**.** ábra).

Ez az *általános hasonlóság* nem tranzitív, ahogy a 3**.** ábra is mutatja, valójában ez egy tolerancia-reláció. Másrészről használhatjuk a *specifikus hasonlóság* kifejezést, ha két struktúrának van egy előre rögzített közös részstruktúrája (jellegzetessége). A közös részstruktúra rögzítése tranzitívvá teszi a relációt, ezért a specifikus hasonlóság ekvivalencia reláció (3. ábra).



3. ábra Általános és specifikus hasonlóság

Az *általános hasonlóság* (bal) esetében valamilyen közös részstruktúrát várunk el (ami nem biztos, hogy egyezik A,B és C között), a *specifikus hasonlóság* (jobb) esetében ez a részstruktúra előre rögzített.

Ha a BLAST hasonlóságot talál néhány biológiai szekvencia között, az egy általános hasonlóság, vagyis nem feltétlenül igaz, hogy mindegyik tartalmazza ugyanazt a részsorozatot, vagyis egy adott fehérje domént. Mindenesetre azok a fehérjék, amiknek van egy közös részsorozatuk, ekvivalencia-osztályt alkotnak. Megjegyzendő, hogy a közös részsorozatot gyakran empirikus alapon definiálják: a biológusok a korábbi ismereteik alapján döntik el, hogy egy fehérje részsorozata valóban tagja-e egy domén családnak (mint az EGF domének). Ha egyszer egy pozitív döntés születik, a fehérje-szekvencia bekerül az EGF-t tartalmazó fehérjék ekvivalencia osztályába. Azt mondhatjuk, hogy a BLAST keresések kiértékelése az út az általános hasonlóságtól a specifikus hasonlósághoz.

A fenti fogalmakhoz kapcsoló, de a kémiai struktúrák elemzésére használatos relációkról összefoglaló található (40)-ben.

Proximitási mértékek

A proximitás mértékek két molekuláris leírás hasonlóságának vagy különbözőségének kifejezésére szolgáló mennyiségek. A proximitás mértékeknek két általános típusa használatos. A hasonlósági mérték annál magasabb értéket vesz fel, minél inkább hasonlít egymásra két molekula. Például a Jaccard együttható:

$$J(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} B_i^2 - \sum_{i=1}^{n} A_i B_i} \qquad [4]$$

A távolság mértékek ezzel szemben 0 értéket vesznek fel azonos molekulák esetén, és magas értéket egymástól eltérő molekulák esetén. Például az Euklideszi távolság:

$$\sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \qquad [5]$$

A továbbiakban a proximitás mérték, távolság mérték, hasonlósági mérték fogalmakat használjuk.

A proximitás mértékeket a legkülönbözőbb összefüggésekben lehet használni. Hasznos megkülönböztetni két esetet, amelyek közösek a bioinformatikai alkalmazásokban. A) Egyszerű közelség két objektum között egy egyértelmű algoritmussal számítható. Két csillag távolsága az űrben jó példa, mint ahogy azok a mértékek is, amiket nem strukturált leírások (pl. vektorok) között számítanak. Definiálhatóak egyszerű proximitás

mértékek strukturális leírásokra is, feltéve, hogy az alkotók közti ekvivalenciák *a priori* adottak. Ilyen mérték pl. az klasszikus informatikában a karaktersorozatok között számított Hamming távolság (4. ábra) és a molekulák 3D struktúráinál alkalmazott *root mean square distance*.

$$rmsd = \sqrt{\frac{\sum_i d_i^2}{N}} \qquad [6]$$

(Az összehasonlítandó struktúrák N elemet (aminosavat) tartalmaznak, d_i a két struktúra i-edik elemének távolságát jelöli.)

Ezeknél egyszerű algoritmusok segítségével számítunk távolságot két objektum között. Az ilyen távolságokat gyakran egy nagyobb objektumnak csak a töredékére számítják ki, ezért töredék távolságként is szokták említeni. B) *Részstruktúra proximitás* mértékeket strukturált leírások elemei között számítanak. Szükséges egy egyszerű mérték, valamint egy algoritmus, amely kiválasztja az „optimális részstruktúrákat" a két objektumból. Például két galaxis távolsága definiálható az egymáshoz legközelebbi csillagaik távolságával. Ebben az esetben ki kell számítanunk az összes pár közötti távolságot (egyszerű közelség), és kiválasztani ezek közül a legkisebbet. A bioinformatika két központi problémája – a szekvencia illesztés és a 3D struktúraillesztés – részstruktúra proximitás problémák.

A két molekula-leírás között számítható proximitási értékeket könnyen általánosíthatjuk molekula csoportokra. Ha adott egy S leírás, és egy $[A]=\{A_1, A_2, \ldots A_n\}$ leírás-csoport, akkor az S és $[A]$ közötti $P(X,Y)$ mértéket a $P(S,A_i)$ páronkénti összehasonlításokból vezethetjük le. Használhatjuk pl. a minimális, maximális vagy átlagos $P(S, A_i)$ értéket arra, hogy S és a $[A]$ csoport hasonlóságát kifejezzük. Egy másik lehetőség, hogy az A_i

leírásokból egy <*A*> konszenzus-leírást számolunk ki, melyet néha az *[A]* centroidjának szoktunk nevezni. Ha a leírások szimpla számértékek vagy vektorok, akkor <*A*> például lehet az átlaguk, vektorok esetében a vektoriális átlag stb. Ennek birtokában *S* és *[A]* között a *P(S,<A>)* mértéket számolhatjuk ki, vagyis az átlagleírást, mint egyetlen leírást tekintjük, és páros összehasonlítást végzünk.

Két csoportra (*[A]* és *[B])* nagyon hasonlóan definiálhatunk proximitási mértékeket: vehetjük a $P(A_i,B_j)$ mértékek minimumát, maximumát, átlagát, vagy meghatározhatjuk a két centroid *P(<A>,<B>)* proximitását.

Szokás figyelembe venni a változók eloszlását is a csoporton belül. Tegyük fel, hogy egy S objektumot és az *[A]* csoportot egy bizonyos *f* tulajdonság alapján hasonlítunk össze és *f* normális eloszlást mutat az *[A]* csoportban, átlaga *m*, standard deviációja *sd*. Akkor az egyszerű különbség helyett használhatunk egy skálázott $\left|\frac{f-m}{sd}\right|$ különbséget is, amely kifejezi az S távolságát a csoport átlagától. Analóg módon, két csoport-centroid távolságát (a csoportokat a felső 1-es és 2-es index jelöli) $\left|\frac{m^1-m^2}{\sqrt{(sd^1)^2+(sd^2)^2}}\right|$ alakban fejezhetjük ki. Ezek a távolságok tehát a felhasznált változót skálázzák a csoportokon belüli eloszlása szerint. Más értékeket kapunk, ha az eloszlások szorosak (kicsi a variancia) és mást, ha a csoportok lazák (nagy variancia). Megjegyezzük, hogy kromatográfiánál a csúcsok szeparációját szokták ilyen képlettel kifejezni. Ezt a skálázást általánosíthatjuk arra az esetre, ha az objektumok az $f_1, f_2 \ldots f_n$ tulajdonságok

vektoraival vannak leírva, és ismerjük a tulajdonságok egymástól való függését leíró *C* kovariancia-mátrixot. Ebben az esetben az ún. Mahalanobis-távolságot számíthatjuk ki:

$$MD = (m^1 - m^2)' C\hat{}\, (m^1 - m^2) \qquad [7]$$

ahol m^1 és m^2 az 1 és 2 csoport átlagai $(m^1 - m^2)'$ az $(m^1 - m^2)$ transzponáltja és *C*^ pedig a *C* kovariancia-mátrix inverze. *MD* egy olyan euklideszi távolságnak tekinthető, amelyet az összes változó varianciája és kovarianciája szerint skáláztunk, ez utóbbiakat egyenlőnek tekintjük mindkét csoportra.

Távolság mértékek (Distance measures)

A vektorok közötti távolságok alkotják talán a legegyszerűbb osztályát a proximitás mértékeknek, köszönhetően geometriai jelentésüknek. A legelterjedtebb az euklideszi távolság:

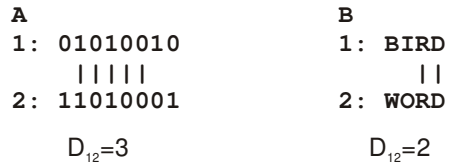
$$\sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \qquad [8]$$

A vektorok távolságmértékének metrika tulajdonságai fontos szerepet játszanak a klaszterezésben és az evolúciós vizsgálatokban. Ahhoz, hogy *M* egy metrika (távolság) legyen, a következőknek kell eleget tennie, minden *A B C* elemre *X*-ből: *1)* $M(A,B) \geq 0$, egyenlőség akkor és csak akkor áll fenn, ha $A = B$; *2)* $M(A,B) = M(B,A)$ (szimmetria); *3)* $M(A,B) + M(B,C) \geq M(A,C)$ (háromszög egyenlőtlenség). A metrikák tulajdonságai alapvetőek, amennyiben klaszterezéshez használják őket. Egy S karakterlánc-hasonlósági mérték alapján akkor lehet klaszterezni, ha létezik egy hozzá tartozó távolság

mérték, $M = f(S)$, ami rendelkezik a metrika tulajdonságaival, és ahol f egy monoton függvény. Az 1-konstans*S típusú távolságmértékeket rutinszerűen használják a klaszterezési eljárásokban.

Biológiai szekvenciák hasonlósági mértékei

A biológiai szekvenciák összerendezéséből adódó hasonlóságok számszerűsítésére a proximitási mértékek egy speciális csoportját használják. Ezek hátterében egy jól ismert matematikai fogalom áll, a karakterlánc-távolság (string distance). Vegyünk először példának nulla és egyes értékekből álló bit-sorozatokat. Két ilyen sorozat között kiszámítható az ún. Hamming távolság, (4. ábra) a két sorozat közötti különbözőségek



4. ábra Hamming távolság

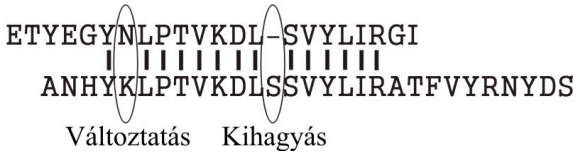
Két string azonos pozícióiban szereplő különbözőségek (helyettesítési operációk) száma

száma, másként mondva azon csere-lépések száma, melyekkel az 1-es sorozatot átváltoztathatjuk a 2-essé (ill. fordítva, mert ez a távolság szimmetrikus). Ugyanezt a trükköt természetesen karakterek sorozatára is alkalmazhatjuk, feltéve, hogy a két sorozat egyenlő hosszú, tehát nincs szükség összerendezésre, pl. beszúrásokra stb. A távolságot súlyozhatjuk is, például egy táblázatba felírunk súlyfaktorokat, hogy mennyibe kerül mondjuk az A karaktert B-re cserélni. De képezhetünk egy részletesebb táblázatot, három- vagy többkarakteres szavakból is. A Hamming távolság ún. egyszerű távolság,

mert egyértelműen kiszámítható két karaktersorozat között. Megmutatható, hogy a Hamming távolság rendelkezik a metrikus tulajdonságokkal.

Bonyolultabb a helyzet, ha tetszőleges és különböző hosszúságú karaktersorozatokat hasonlítunk össze és megengedünk beszúrásokat is. Az ilyen távolságfüggvényt leírt szavak összehasonlítására vezették be és "string edit distance"-nek hívják. Azon lépések számát jelenti, amellyel egyik szót a másikba változtathatjuk át, a beszúrás, kihagyás és a csere egy-egy lépésnek számít. Mivel számos lehetséges úton érkezhetünk el egyik szóból a másikba, a string-edit távolság a lehető legrövidebb ilyen útnak a hossza. Természetesen ezt a string-edit távolságot is súlyozhatjuk egy egyszerű táblázat segítségével. A string-edit távolság is metrikus.

A biológiai szekvenciák összehasonlításánál lényegében egy súlyozott string-edit távolságnak megfelelő hasonlósági mértéket alkalmaznak, amelyik azonos karaktersorozatok esetén maximális (a string-edit distance ezekre nulla lenne). Az általánosan használt hasonlósági mértéket az 5. ábra, ill. a [9] egyenlet mutatja be.



5. ábra String edit distance

Azon lépések száma, amellyel egyik szót a másikba változtathatjuk át, beszúrások, kihagyások és a cserék segítségével.

$$S_{1,2} = \sum \text{költség}_{egyezések, változtatások} - \sum \text{költség}_{kihagyások} \qquad [9]$$

Sokféleképpen súlyozhatunk, aminosavaknál pl. az ún. Dayhoff vagy BLOSUM mátrixokat alkalmazzuk, és sokféle mód van a beszúrások meghatározására is (hossz-függő, másodlagos-szerkezettől függő stb.).

Ezek a módszerek és speciális szekvencia-analitikai alkalmazásaik megtalálhatók tankönyvekben is (41), (42). Ez a fejezet a továbbiakban arra próbál választ adni, hogy miként interpretáljuk az eredményeket, és főként, mi a valószínűségi interpretálás alapja.

Egy egyszerű felosztást fogunk alkalmazni. Az *általános módszerek* azok, amelyek a véletlen hasonlóságok statisztikus leírásán alapulnak. Ilyenek a leggyakrabban használt heurisztikus módszerek, tehát a BLAST (43) és a FASTA (44), valamint a globális illesztést kereső Needleman-Wunsch (45) és a Smith-Waterman (46) is. A *specifikus módszerek* pedig azok, amelyek egyes biológiailag fontos szekvencia-csoportok leírására koncentrálnak, valamely módszert célcsopontonként parametrizálnak. Ezek a csoportok gyakran túl kicsik ahhoz, hogy jó statisztikát tudjunk rajtuk készíteni, itt inkább az előzetes biológiai tudásunk segíthet. Specifikus módszerek tipikusan a doménfelismerő programok, melyek sokszor külön adatbázisokkal, speciális ábrázolásokkal dolgoznak.

A szekvencia-összehasonlítás általános módszerei

Az általános módszerek közül a legnépszerűbbek azok, amelyek lokális összerendezéseket adnak (BLAST, FASTA). Jellemzőjük, hogy az így nyert hasonlósági érdemjegyek a) nem metrikus tulajdonságúak (nem lehet őket egyszerű aritmetikával távolságfüggvénnyé alakítani). Ezért pl. klaszterezésre csak közelítőleges pontossággal

használhatók. Másik jellemzőjük viszont, hogy b) a véletlenszerű hasonlóságokra jól leírható statisztika van, tehát könnyű valószínűségi értelmezést adni egy érdemjegynek. Ezeket a módszereket szokták felhasználni arra, hogy az egyre növekvő szekvencia-adatbázist klaszterekre osszák fel. Ehhez elvben egy metrikus távolság kell. (Mint amilyen a Needleman-Wunsch módszer érdemjegyeiből elvben képezhető.) A valóságban azonban minden hasonlóságkereső módszert fel szoktak használni, különböző heurisztikus megoldásokkal megfejelve. Erre az adatok természete miatt van lehetőség. A szekvenciák által alkotott tér ugyanis nagyon ritkán van betöltve, az elméletileg elképzelhető szekvenciáknak csak egy igen kis töredéke jelenik meg a természetben, és azok is szoros csoportokat alkotnak. Ezért aztán sok módszer alkalmas a csoportok felderítésére. Krause és Vingron (47), (48) például az igen gyors BLAST program sorozatos futtatásával iteratívan keresett új és új tagokat egy-egy szekvenciához. Bizonyos P-küszöbnél szignifikánsabb hasonlóságokat fogadtak csak el, és egy szekvenciához mindig kiválasztották a legalacsonyabb S értékű, de még mindig szignifikáns szekvenciát, azzal újra kerestek, stb. Így olyan klaszterek álltak össze, amelyek egy idő után nem növekedtek tovább. Ez az ún. SYSTERS adatbázis, amelynek automatikusan előállított klaszterei vannak, de azok természetesen nem mindig egyeznek meg a biológusok által felállított csoportokkal. A helyzet ugyanis korántsem ilyen könnyű. Ha olyan küszöbértéket veszünk, ahol a klaszterek nem folynak össze, akkor a biológiailag legérdekesebb távoli hasonlóságokat elveszítjük. Hosszabb szekvenciáknál ez a veszély kevésbé áll fent, de a rövidebb doménszekvenciákat már nem könnyű csoportosítani ezen az alapon.

A szekvencia-összehasonlítás speciális módszerei

Nincs matematikailag definiálható éles határ a biológiailag jelentős és a véletlen hasonlóságok között, az ilyen döntésekhez mindig biológiai ismeretek szükségesek. Például a gyenge szekvencia-hasonlóságokat is hajlamosak vagyunk elfogadni, ha a fehérje doménstruktúrája egyébként megfelelő, az exon-intron eloszlás is megfelelő stb. Az ún. speciális módszerek az ilyen, biológiai tudás alapján képzett csoportokat tanulmányozzák, és ezek speciális leírásaival foglalkoznak. Ezek a módszerek nemcsak egy algoritmusból állnak, hanem tartozik hozzájuk egy speciális adatbázis is. A fehérje-domén-adatbázisok általában ilyenek.

A leírások egy nagy csoportjával itt nem foglalkozunk: Ezek azok az un. konszenzus-leírások, mikoris a fehérjeszekvenciák (általában a doménszekvenciák) egy csoportjának valamilyen közös leírási formát választunk. A gyakorlatban használt doménkeresőket és domén-adatbázisokat ennek alapján különböztethetjük meg. Ilyenek az ún. reguláris kifejezések (PROSITE), a HMM - Hidden Markov models - (SMART, PFAM-A), ilyenek a szekvencia-profilok (PROSITE) – ezekről az irodalomban illetve tankönyvekben jó összefoglalókat találunk (49), (42).

Most azokkal a speciális módszerekkel foglalkozunk, amelyek adatbázis-keresésen alapulnak. Az egyik megoldás, amelyiket az SBASE alkalmaz, a domén-adatbázis teljes összehasonlításán alapul, minden szekvenciát minden szekvenciával összehasonlítunk a BLAST programmal (50). Az így előálló értékeket, mint egy hatalmas hálózatot képzelhetjük el, ahol a pontok a szekvenciák, az élek pedig a szignifikáns BLAST score-

nak megfelelő vonalak. Ezt a megközelítést memória-alapú osztályozásnak is hívjuk, a rendszer tudásbázisa a hasonlóságok hálózata (51), (52) Ez lényegében egy gráf-leírás, és azt az általános képet kapjuk, hogy az ismert doménszekvenciák valóban sűrűn össze vannak egymással kötve, de nincs olyan éles elválás, mint a SYSTERS adatbázis fenti példájában, hanem a csoportok elég tetemes hasonlóságot mutatnak más, irreleváns szekvenciákkal is.

A másik példa az ún. COG adatbázis, ahol a teljes genomok ortológjainak szekvenciáit gyűjtik, és ez főként a prokariótáknál működik jól (53). Itt a válogatás alapja egy háromszög alakú klikk: ha van három szekvencia, melyek egymás legközelebbi BLAST szomszédjai, azok egy egységbe tartoznak. Ennek alapján meg lehetett állapítani az ortológ klasztereket, és az új genomokból pedig nagyrészt be lehetett sorolni a fehérjéket a régiekbe, illetve időnként új klasztereket lehetett indítani. Ez egy biológiai szempontból nagyon fontos megközelítés, mert a teljes genomok ún. természetes adatbázist jelentenek, szemben a SWISSPROT v. PIR adatbázisok ad hoc gyűjteményeivel. A COG-gal való összehasonlítás tehát BLAST alapú, és a szekvenciákat besorolja azt ortológ csoportba, szerencsés esetben pedig a fehérje funkciója is megállapítható.

A mai módszerek sok tekintetben a fenti elvek kombinációinak tekinthetők. A modern megközelítések alapjait úgy lehet összefoglalni, hogy az adatbázis-keresés lényegében tartalmazza a távoli homológiák kereséséhez szükséges információt is, és ezt valamiféle ügyes válogatással meg is tudjuk találni. A hagyományos keresések csak a score szerint ill. ami azzal egyértelmű, a *P* értékek szerint rangsorolnak. De lehet újabb rangsorolási dimenziókat is bevezetni, például el lehet osztani a találatokat a Query szekvencia

mentén, sőt lehet válogatni aszerint is, hogy van-e bennük közös mintázat. Mindezek nem új módszerek, de kombinációjukkal nagyon hatékony kereső programok alakultak ki. A PSI-BLAST például a pozició-specifikus scoring módszerét alkalmazza (54), (55), (56). A talált hasonló szekvenciákból összegző adatstruktúrát, szekvencia-PROFILE-t épít (Attwood összefoglalójában ez is megtalálható, részletezve), és a következő iterációban már ezzel az érzékeny kereső-eszközzel vizsgálja végig az adatbázist. Az új szekvenciákat ismét csak valamiféle empirikus – és véleményem szerint nem elég finoman beállítható – küszöb alapján veszi be a keresésbe. Optimális esetben már néhány kezdeti ciklus után világosan látszik, vannak-e az adatbázisban olyan szekvenciák, melyek érdekesek a vizsgálat szempontjából. Ez a módszer tehát egyrészt a BLAST gyorsaságára, másrészt az itt nem részletezett PROFILE konszenzus leírás érzékenységére épít. Némileg hasonló módon azt is megoldották, hogy a BLAST a keresésnél csak azokat a szekvenciákat vegye elő, amelyek egy adott mintázatot tartalmaznak.

A szekvencia-összehasonlítás közelítő gyorsmódszerei

A szekvencia-keresés (BLAST) gyorsasága miatt tulajdonképpen csak nagyon kevés helyen van szükség még gyorsabb módszerekre. Ilyen alkalmazások például a teljes genom, teljes proteom vizsgálatok. A közelítő módszerek csak kvalitatíve jelzik, hogy egy illető szekvencia beletartozik-e egy csoportba, illesztést nem adnak.

A közelítő módszerek alapja valamiféle strukturálatlan leírás, pl. vektor-ábrázolás, amelyet gyorsan, összerendezés nélkül lehet összehasonlítani egymással. Tipikusan ilyenek az aminosav-összetétel, amelynek mintájára beszélhetünk dipeptid, tripeptid összetételről is. Ezeket valami egyszerű távolsággal – pl. Euklideszi távolsággal – hasonlítjuk össze. Régebben sokat használtak Fourier transzformált jellegű leírásokat, mikoris egy fehérje hidrofóbicitási diagramja helyett, annak Fourier-transzformáltját hasonlították össze, mint eloszlást. Ezeknek a módszereknek azonban ma már viszonylag kisebb a jelentősége (57).

Fehérjék 3D szerkezeteinek gyors összehasonlítása a PRIDE algoritmussal

A PRIDE program (58) a fehérjéket eloszlások formájában ábrázolja, az ábrázolás alapja a C_α atomok közötti távolság. Vegyük példának az egymástól 5 aminosav távolságra lévő C_α atomokat. Ezeket a C_α (i)- C_α (i+5) távolságok hisztogramja jellemzi (pl. fél Angströmös felbontásban felvesszük, hogy hány ilyen távolság van, és az egészet 100%-ra normáljuk). Két fehérje ilyen hisztogramját egy standard módszerrel, az ún. kontingencia-táblázat módszerével vetjük össze. Ez egy 0 és 1 közötti értéket ad eredményül, amit valószínűségi mérőszámként értelmezünk. Valójában nem egyféle hisztogramot, hanem minimum 3, maximum 30 aminosav távolságra vesszük fel a hisztogramokat, ez összesen 28 hisztogramot jelent egy fehérjére. Két fehérje úgy hasonlítható össze, hogy a hisztogramokat páronként vetjük össze a kontingencia-táblázat módszerével a 28 érték átlaga a végső valószínűség, amely azt adja meg, mennyire

valószínű, hogy a két fehérje hasonlít egymáshoz (Probability of IDEntity). Ez egy igen gyors számítás, és a statisztikai módszer is elég biztonságosnak mondható. A hasonlóságok 99,5%-ban visszaigazolhatóak, ami igen jónak mondható. Előnye, hogy adatbázis keresésre is felhasználható, ilyenkor az adatbázist már előre hisztogramok formájában tárolják, az összehasonlítás tehát villámgyors. A módszer fold-előrejelzésre is alkalmas, ahol mindig a hozzá leginkább hasonló fold típusát rendeljük az ismeretlen szerkezethez. De végezhető vele klaszterezés is – a módszer gyorsasága miatt on-line is alkalmazható. Az osztályozást a CATH fold adatbázis segítségével végzi, a fold típusa a hozzá leginkább hasonló CATH-beli szerkezet típusa lesz.

Hasonló, kissé bonyolult elv az, ha a fehérjék C_α sorozatát mint térgörbét tekintjük, és annak topológiai invariánsait számitjuk ki (59). Ennek matematikája igen bonyolult, a lényeg az, hogy a C_α-k sorozatához, tehát a fold-hoz egy 30 komponenses vektort rendelünk, amelynek komponensei az említett topológiai mérőszámok. A vektorok között Euklideszi távolságokat számítva választhatjuk ki a leghasonlóbbat egy adatbázisból. Ennek a módszernek még nincs on-line változata, pontosságát azonban jónak tartják.

Genomok, proteómok, hálózatok összehasonlítása

A genomok, proteómok és génhálózatok ábrázolása a bioinformatika legújabb, és legnehezebb feladatát jelenti. Általában igaz, hogy ezekre is az ER (entity-relationship) leírásokat alkalmazzuk, de a struktúrák nagysága, az adatok hiányos és bizonytalan volta sok nehézség forrása.

Mindennek a hátterében elsősorban az áll, hogy a szekvenciákkal, és a térszerkezeti leírásokkal szemben a gráfok esetében általában nem elegendő lineáris eltolásokat végezni, és gap-eket beilleszteni két gráf összeillesztése során. Abban az esetben, ha nincs a priori megfeleltetésünk a két gráf csúcsai közt, akkor N! (különböző méret esetén: M!/((M-N)!*N!)) féleképpen végezhető el az illesztés. Ezért még azt a kérdést is igen nehéz megválaszolni, hogy egy gráfnak része-e egy másik gráf. (Ezt a problémát szokás részgráfizomorfizmusként említeni.) Mivel ez a kérdés a gyakorlatban megválaszolhatatlan, két gráf összehasonlítása származtatott tulajdonságaik, paramétereik által történik. Ehhez azonban lényeges, hogy olyan paraméterekkel írjuk le a gráfokat, amelyek nem változnak meg a csúcsok átsorszámozásától, és a gráf „kis" megváltoztatására „kis" mértékben változnak. Ennek eldöntése a problémától is erősen függ. Példaként említeném a legrövidebb utak átlaga és az efficiencia közti különbséget. Példánkban egy gráf egy élének eltávolítása után megszűnik összefüggő lenni. Ekkor *l* (a legrövidebb utak átlaga) egy véges értékről végtelenre változik, attól függetlenül, hogy egyetlen csúcs szakadt le, vagy két nagy rész között szűnt meg az összeköttetés. Ezzel ellenben E (efficiencia, ld. [1] képlet) alig változik, ha csak egy csúcs válik elérhetetlenné, és nagymértékben, ha kettészakad a hálózat, de önmagában nem elég annak eldöntésére, összefüggő-e a hálózat. Visszatérve a részgráfizomorfizmusra: ebben a konkrét esetben nem mondana többet, mint hogy a két gráf legjobban ilyen csúcsillesztéssel feleltethető meg egymásnak, és ebben az illesztésben egy él nem illeszkedik, de nem mondana semmit a strukturális változásokról.

A gyakorlatban nagyon sok háttér-információ áll a rendelkezésünkre, amelyek lényegesek a modell felállításakor, vagy a kérdés megfogalmazásánál. Vegyük például a genomokat. Ezeket tekinthetjük lineáris struktúráknak, ahol az elemeket – a géneket – a kísérleti úton nyert genomszekvenciák alapján (génpredikciós programokkal) határozzák meg. A gének mellett ismerhetünk promótereket, enhancereket, melyeknek szekvenciális helyét szintén ismerjük. Ezek között az elemek között sokféle relációt definiálhatunk, mint például a szekvenciális szomszédság, regulációs kapcsolat, stb. Ezen relációkat az elemeket összekötő vonalakkal ábrázolva egy óriási hálózatot kapunk. Ez egy címkézett gráf, ahol az elemek és relációk címkéit éppúgy egy zárt definíció-rendszer (un. ontológia) tartalmazza, akárcsak a fehérjéknél vagy a szerves vegyületeknél.

A proteómok leírása lényegileg hasonló, a fehérjéket funkcionális, biokémiai vagy szerkezeti tulajdonságaik alapján címkézzük, a köztük lévő relációk pedig lehetnek metabolikus kapcsolatok, közös szubsztrátok és kofaktorok, vagy esetleg hasonlóságok (szekvencia vagy szerkezeti hasonlóságok).

Ebből a vázlatos leírásból is látszik, hogy ezeknél az adatoknál az elemek és kapcsolatok sokfélesége miatt egy újfajta komplexitás jelentkezik. Ezt a komplexitást igyekszik megragadni a hálózatokkal foglalkozó új megközelítés, amelyik az Internet, a társadalmi kapcsolatok, az úthálózatok, az elektromos hálózatok közös tulajdonságait igyekszik leírni. Ezeket az eredményeket sikerrel alkalmazzák ma a genomokra, sőt a tudományos bibliográfia hálózataira is.

A következőkben azokra a genomok közötti összehasonlításokra fogok koncentrálni, ahol a genomot egyetlen struktúrával írjuk le. A teljes genomok összehasonlítására vannak más, egyes szekvenciák filogenetikus összehasonlításán alapuló módszerek is, de ezekre itt nem térek ki, jó összefoglalók (60), (61). keletkeztek a témában. Azt is meg kell jegyeznem, hogy a genomok összehasonlítása alatt néha a proteómok összehasonlítását értik, a terminológia tehát kissé laza.

Számítástechnikai szempontból a genomok és egyéb hálózatok hatalmas gráfok, melyek egymástól méretben nagyon különbözhetnek (más a csúcsok és élek száma), ráadásul az adatok egy része bizonytalan is (gondoljunk arra, hogy a publikált genomokban még évek múltán is fedeznek fel új géneket). Ezekre a nagy és bizonytalan hálózatszerű modellekre főként azokat az összegző leírásokat tudjuk alkalmazni, amelyeket a fentiekben nem-strukturális leírásoknak neveztünk, ezek közül is leginkább az összetétel-szerű leírásokat alkalmazzák a leggyakrabban.

Az összehasonlítás alapja tehát általában valamilyen összetételszerű vagy jelenlét/hiány adatokat összegző leírás, és a módszereket aszerint érdemes csoportosítani, hogy a leírás komponensei eleve adottak-e, vagy pedig az analízis során definiáljuk őket.

Az előre definiált komponensek példája a már említett COG adatbázis ortológ fehérjéinek táblázata. Ezeket a csoportokat funkció szerint ismerjük, és a funkciók nagyobb osztályokba (metabolizmus, információfeldolgozás, stb.) is be vannak osztva. Ez a majdnem háromezer fehérjetípus olyan, mint egy sokdimenziós koordinátarendszer, minden genomról eldönthetjük, hogy egy illető COG csoport (ortológ fehérjék) jelen van-

e az illető genomban. Az ilyen jelenlét/hiány adatokból aztán a [4] egyenlet szerinti Jaccard koefficienssel dönthetjük el, hogy két genom mennyire hasonlít. Az (1 – Jaccard koefficiens) mennyiség pedig metrikus tulajdonságokkal rendelkező távolságfüggvény, ennek alapján a genomok klaszterezhetők is. Ilyen egyszerű módszerekkel döntötték el pl., hogy az archeabaktériumok bizonyos szempontból az eukariótákhoz is közel állnak. Maga a leírás tehát nagyon egyszerű, külön beállítandó paraméterek sincsenek benne, de a komponensek maguk a probléma elegendően finom felbontását adják. Teljesen analóg módon a fehérjék foldtípusai is felhasználhatók a genomok osztályozására. Itt azt kell eldönteni, hogy egy adott foldtípus jelen van-e egy genomban (proteomban), amit általában szekvencia-kereséssel lehet eldönteni. Az ilyen adatokat aztán ugyanúgy használhatjuk fel, mint a COG jelenlét/hiány adatokat (62).

A metabolikus adatok ugyancsak felhasználhatók, mint egy előre definiált koordinátarendszer. A proteom ebben a rendszerben mint enzimek, szubsztrátok és ezek komplexeinek hálózata irható le (63) (19), ilyen adatokat az un. WIT adatbázisból szerezhetünk (64). Az egyes organizmusok, vagy pl. azok metabolikus enzimjeinek adatait azután vektorokká alakíthatjuk. Pl. az *E. coli* metabolikus enzimjeinek vektora annyi komponensből áll, ahány enzim szerepel benne, és a komponensek azt írják le, hogy az illető enzim hányszor szerepel a metabolikus leírásokban. Az ilyen vektorokat aztán bármilyen hasonlósági vagy távolságfüggvénnyel összehasonlíthatunk. Podani és munkatársai ezzel a módszerrel mutatták ki, hogy az archeabaktériumok és az eukarióták metabolikus szervezettsége hasonlít egymáshoz. Ez az összehasonlítás részben a Jaccard koefficiensre, részben pedig a statisztikából ismert rang-korrelációra alapult (65).

Az analízis során definiált komponensekre a számítás sokban hasonló az előzőkhöz, itt is a Jaccard koefficiens a csodafegyver. A legjobb példa a közös gének (fehérjék) összeszámlálása. Két genom hasonló fehérjéit megkereshetjük BLAST-tal (66) vagy a pontosabb Smith-Waterman programmal (67), (68), itt egyszerűen egy bizonyos küszöbértéknél szignifikánsabb hasonlóságokat fogadunk el „közös elemnek". Így megállapítható, hogy a két genom hány közös fehérjét tartalmaz, illetve összesen hányféle fehérjét tartalmaz. E kettő hányadosa a Jaccard koefficiens. Azért hívjuk ezt analízis során definiált komponenseknek, mert a komponensek száma az összehasonlított pároktól függ, más genom párok esetén ez változni fog.

Ennek a módszernek egy érdekes kiterjesztése a közös génpárok módszere, amelyik egymás melletti génekből származó fehérje-párokat használ, amelyeknél a transzkripció iránya is megegyezik (69), (70), (71). Az ilyen közös génpárok száma a számláló, míg a két genomban fellelhető összes génpár száma a nevező. Ez az analízis legalább olyan gyors, mint az előző (a közös gének számlálásán alapuló), és ráadásul a leírás gazdagabb, megragad valamit a génsorrendből is.

A fenti egyszerű összehasonlítások mellett meg kell említeni a genom-motívumok területét is. Megjegyezzük, hogy a szekvencia-motívumok és a szerkezeti motívumok gyűjteményei sokkal fejlettebbek, mint a genom-motívumoké, amelyeket éppen ezért érdemes röviden áttekinteni. A génsorrendek számítógépes vizsgálata először is olyan konzervált gén-elrendezéseket eredményezett, amelyek között a már ismert operonokat lehetett felfedezni. Ezeken felül lehetett találni olyan transzkripciós irány, génfunkció és génsorrend szerint hasonló nagyobb csoportokat, amelyeket űber-operonnak vagy szuper-

operonnak szoktak nevezni (72), (73). Ezekről feltehető, hogy az evolúció során együtt mozognak. A transzkripciós irányok és a regulációs kapcsolatok ábrázolása miatt itt már irányított gráfokról beszélünk, az operonok ilyen gráfok.

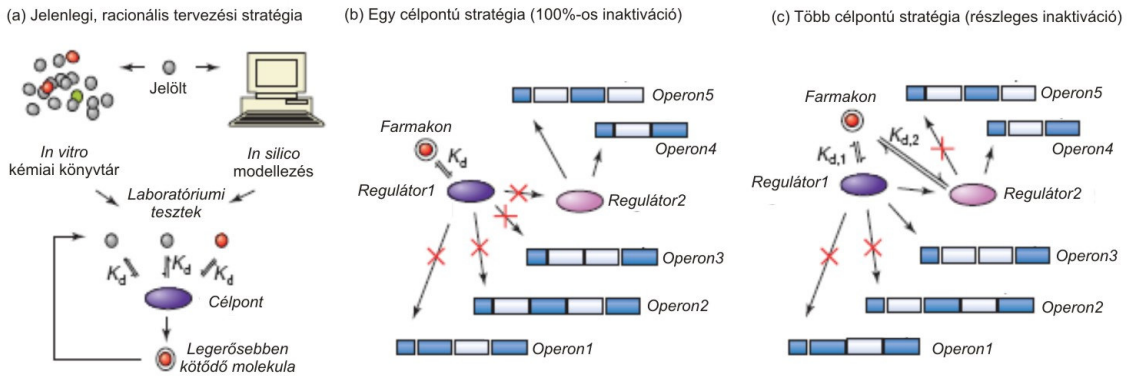
A metabolikus útvonalak szintén irányított gráfok, ahol a csúcsokat a fehérjék, az éleket pedig a biokémiai reakciók illetve szubsztrátok adják.

Eredmények

Biológiai hálózatok többpontos támadása: egy gyógyszertervezési stratégia modellje

Kérdésfelvetés

A probléma tanulmányozásánál fő célunk az volt, hogy a hálózati modellekkel adjunk választ egy gyógyszertervezési alapkérdésre. Bár a hatóanyagokat ma számítógéppel tervezik és genomikai-proteomikai alapozású módszerekkel vizsgálják, az engedélyezett hatóanyagok száma nem növekszik. Munkánk kiindulópontja az a hipotézis volt, hogy az egyetlen biokémiai célpontot támadó „egycélpontú" gyógyszerek nem optimálisak, néhány többhatású hatóanyag sokkal hatásosabb. A fejlesztési stratégiákat az alábbi ábra foglalja vázlatosan össze:



6. ábra Gyógyszertervezési stratégiák sematikus összehason1ítása

A jelenlegi megközelítéssel egy fehérjét blokkolnak, lehetetlenné téve, hogy részt vegyen bármilyen kölcsönhatásban. Az új megközelítéssel elegendő a lényeges kölcsönhatások lokális hálózatának csak egy részét gátolni.

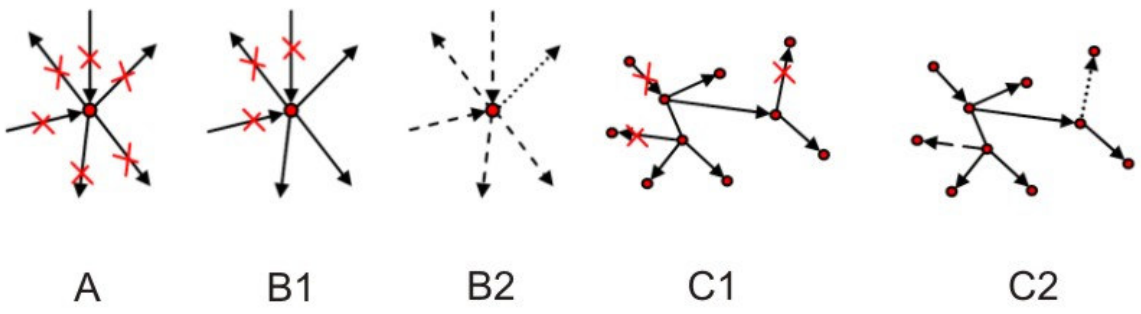
A baloldali panel a ma általános racionális tervezést vázolja. Ezeknél végeredményben egyetlen biokémiai, laboratóriumi vizsgálatban legjobban szerepelő speciális hatóanyagot választanak ki, melyik egy, a tesztelés alapjául választott speciális célpontra hat. Hálózatként ábrázolva a kérdéses fehérjéket, géneket (középső panel), ez a hatás olyan lesz, mintha egyetlen fehérjét iktatnánk ki a hálózatból (mondjuk egy patogén szervezet egyetlen enzimét gátoljuk). A hálózati modell azonban sugallja a másikfajta megközelítést is, amit az ábra jobb oldalán mutatok be, a hálózatot több ponton is meg lehet támadni. Milyen legyen a többpontos támadás? Elegendő-e csak gyengíteni a hálózatot, és nem teljes gátlásra törekedni? Ez utóbbi kérdés azért érdekes, mert részleges gátlást sokkal könnyebben – pl. kevesebb hatóanyaggal - lehet elérni, mint „teljes” gátlást.

Azok a hatóanyagok, amelyek csak egyetlen célpontra hatnak (egypontos támadás) néha nem változtatnak meg egy komplex rendszert a kívánt módon, még akkor sem, ha a célpontjukat teljesen meg is változtatták. Például egy célponthoz tartozhatnak „tartalék”

rendszerek, kerülőutak, amik sokszor eléggé különböznek a célponttól ahhoz, hogy ne hasson rájuk ugyanaz a hatóanyag. Többek között ennek a következménye az a jól ismert tény is, hogy számos sejten belüli molekuláris hálózat robosztus, és még az alkotóelemeit érő drasztikus hatások ellenére sincs nagyobb ingadozás a végtermékek (metabolitok, kimenőjelek, stb.) szintjén. Ezek a megállapítások egyaránt érvényesek arra az esetre, ha a hatóanyag gátolja, vagy éppen serkenti a célpontját.

A vizsgálati modell

Az előzőekben vázolt hálózati modell értelmében a hatóanyag hatást a hálózatok támadhatóságával modellezzük. Egy elemi támadás egy hálózatnak egy csúcsát, vagy egy élét távolítja el. Munkánk során ezeket a módszereket általánosítottuk, tehát olyan támadási stratégiákat vizsgáltunk, amelynek során csúcsokat, vagy éleket támadunk, illetve gyengítünk. A csúcsok támadása például egy gén vagy fehérje gátlását jelenti. Az élek támadása pedig egyes kölcsönhatások speciális gátlásának felel meg. Az alkalmazott kombinációkat a következő ábra szemlélteti:



7. ábra Hálózati beavatkozások

Hálózatok támadási stratégiái. A: Egy csomópont teljes gátlása B: Egy csomópont részleges gátlása egyes kölcsönhatásainak blokkolásával (B1) illetve a csomópont körüli kölcsönhatások gyengítésével (B2). Csomóponttól független támadás a hálózat egyes kölcsönhatásainak teljes (C1) illetve részleges (C2) blokkolásával.

A támadások modellezésekor a hálózatok egy vagy néhány jellemző tulajdonságának a támadások hatására bekövetkezett változását vizsgáljuk. Szokás használni a legnagyobb összekötött komponens méretét, a csúcsok átlagos távolságát (a legrövidebb utak átlagát), stb. A dolgozatomban ismertetett vizsgálatok szempontjából a hálózat kommunikációs hatékonyság, az efficiencia ([1] képlet) mérése tűnt a legcélravezetőbbnek, mivel ez éppúgy jellemzi a csúcsok közötti kommunikációt, mint pl. a csúcsok átlagos távolsága, de ugyanakkor szélesebb körben számolható és általánosítható. A legrövidebb utakat használó mennyiségek fragmentált hálózatokra nem számolhatóak, az efficiencia viszont igen, ugyanakkor számítása könnyen alkalmazható irányított és súlyozott gráfokra is.

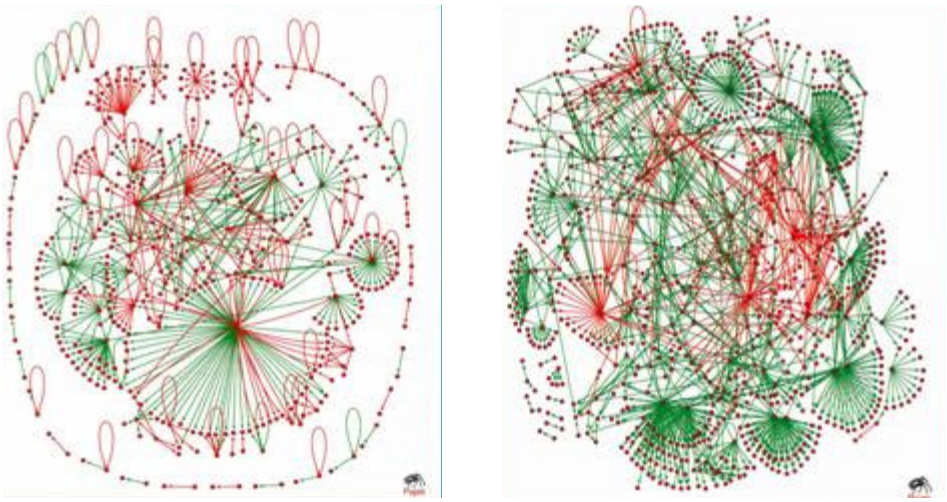
Többpontos támadás modellezése mohó algoritmussal

A támadások hatásosságának összevetése az efficiencia segítségével történt oly módon, hogy minden elem támadásakor (pl. csúcs eltávolításakor, él gyengítésekor) kiválasztottam azt a pontot, ahol az efficiencia a legnagyobb mértékben csökkent, tehát a kár maximális volt. A soron következő eltávolítandó elem kiválasztásakor is mindig azt vettem figyelembe, hogy melyik elem eltávolítása okozza a legnagyobb kárt, vagyis csökkenti legnagyobb mértékben az efficienciát. Megjegyzendő, hogy a hálózatok támadásánál gyakran a legnagyobb fokszámú elemet szokták eltávolítani. Az általam alkalmazott mohó algoritmus ugyan több számítási kapacitást igényelt, mintha csak a fokszámot vettem volna figyelembe, de rávilágított, hogy nem mindig a legnagyobb fokszámú csúcs eltávolítása okozza a legnagyobb kárt. Meg kell azt is jegyezni, hogy a módszer valójában csak alsó becslést ad a több elem eltávolításával okozható kár mértékére. De ahhoz, hogy megállapíthassuk, egy 1000 csúcsot, esetleg 1000 élet tartalmazó hálózatban melyik az az öt csúcs (él), amelyek eltávolítása a legnagyobb kárt okozná, minden lehetőség esetén ki kellene számolnunk az efficienciát. Ez a számítás pl. egy 1000 csúcsot tartalmazó hálózat esetében például kb. $8*10^{11}$ esetet jelent, ami a jelenlegi számítási kapacitás mellett elfogadhatatlanul hosszú időt jelentene.

Modellszámítások a kólibaktérium és az élesztő regulációs hálózatán

A kérdés tanulmányozására első modellnek az *E. coli* illetve az *S. cerevisiae* regulációs hálózatát választottam, mivel a regulációban résztvevő fehérjék megfelelőek a gyógyszerhatások vizsgálatára. Egyrészről a regulációs rendszerek központi és nagyon érzékeny részét képezik a sejtműködésnek, kis módosításuk is hatalmas változásokat okozhat a legváltozatosabb életfunkciókban. Másrészről hálózati paramétereik alapján a biológiai rendszerek egy igen széles rétegébe sorolhatjuk őket, elsősorban a skálamentes fokszámeloszlásuk alapján. Ezek a hálózatok irányított hálózatok, a kóli regulációs hálózata 424 csúcsot és 520 élt, az élesztőé 689 csúcsot és 1080 élt tartalmaz.

A támadási modellezést az általánosabb eredmények érdekében irányítatlan hálózatokon végeztem el. Az eredmények ábrázolásánál a hálózat kiindulási efficienciáját 100%-nak vettem, és ehhez képest ábrázoltam az efficiencia csökkenését. Az összehasonlítás alapja a teljes gátlásnak megfelelő stratégia volt, amelyet úgy modelleztem, hogy – a 7. A ábrának megfelelően – a hálózatból sorra távolítottam el a legnagyobb kárt (efficiencia-csökkenést) okozó csúcsokat. Így egy fokozatosan ereszkedő görbét kaptam (9. ábra, A-D panel, kék görbe). Ha egy stratégia ennél hatásosabb, annak görbéje ez alatt fog haladni. Ha kevésbé hatásos, akkor fölötte.

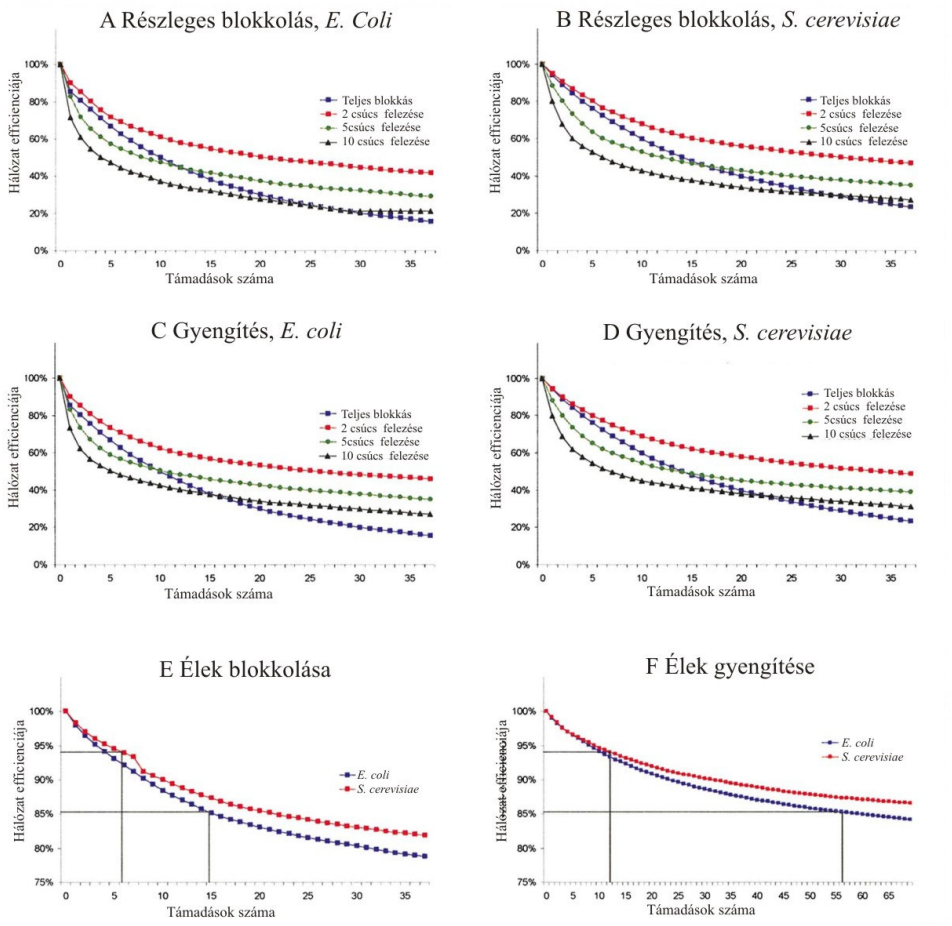


8. ábra A kólibaktérium és a sütőélesztő regulációs hálózatának képe (74), (75) alapján.

A nyilak a regulátortól a regulált génre mutatnak, a zöld a serkentő, a piros a gátló hatást jelzi.

Az 9. ábrából látható, hogy az egypontos támadás hatásához képes 2 csúcs gyengítése, vagy kapcsolatainak részleges törlése még kevés, de 5 csúcs esetében már több lépésen keresztül tartható ill. meghaladható a kár mértéke (9. ábra A és B panel). Meg kell említeni, hogy az 5 pontos részleges támadás hatékonysága stratégiától független, tehát fennáll akkor is, ha a csúcs kapcsolatainak felét töröljük (7. ábra B1stratégia, 9. ábra, A-B panel, zöld görbe), de akkor is, ha a csúcs körüli kölcsönhatásokat felére gyengítjük (7. ábra B2 stratégia, 9. ábra, C-D panel, zöld görbe), ami alapján bármelyik modell is bizonyul később a valóságot jobban megközelítőnek, az eredmények érvényesek maradnak. Vagyis akár így, akár úgy szeretnénk modellezni egy elem részleges gátlását, a szimulációban ugyanolyan hatékonynak bizonyul a két módszer. Az élek egyesével történő eltávolítás vagy súlyozása (7. ábra C stratégia) meglepően hatékonynak bizonyult. Az E. coli hálózatában mindössze 15 él eltávolításával, illetve 56 lépésben egy-egy él súlyának megduplázásával sikerült azt a hatást elérni, amit egy 72 éllel rendelkező csúcs

törlése jelentett (9. ábra, E panel, kék görbe, F panel, kék görbe). Az S. cerevisiae esetében pedig 6 él eltávolításával, illetve 13 lépésben történő súlyozással volt elérhető az a hatás, amit a legnagyobb kárt okozó, 18 éllel rendelkező csúcs törlése jelentett (9. ábra, E panel, piros görbe, F panel, piros görbe), (1. táblázat). A jelenségek a kólibaktérium és az élesztő esetén azonos jellegűek, csak számszerű jellemzőikben térnek el.



9. ábra Okozott károk összevetése grafikonokon

A teljes hálózat efficienciájához viszonyított relatív efficienciák csökkenése a megtámadott pontok (gének) számának függvényében, különböző támadási stratégiák esetén. Az alsó (E,F) paneleken az élek ellen irányuló támadásokat ábrázoltam, a bejelölt pontok azt jelzik, hogy egy központi gén kiiktatásával eltűnő kapcsolatok hatását disztributív támadásnál hány él eltávolítása pótolja. Ez a szám mindig kisebb, mint a központi gén kölcsönhatásainak száma.

Az első táblázatból a 9. ábrán található görbék számszerű adatain kívül név szerint is szerepelnek azok a gének, amelyek a különböző stratégiák által érintettek.

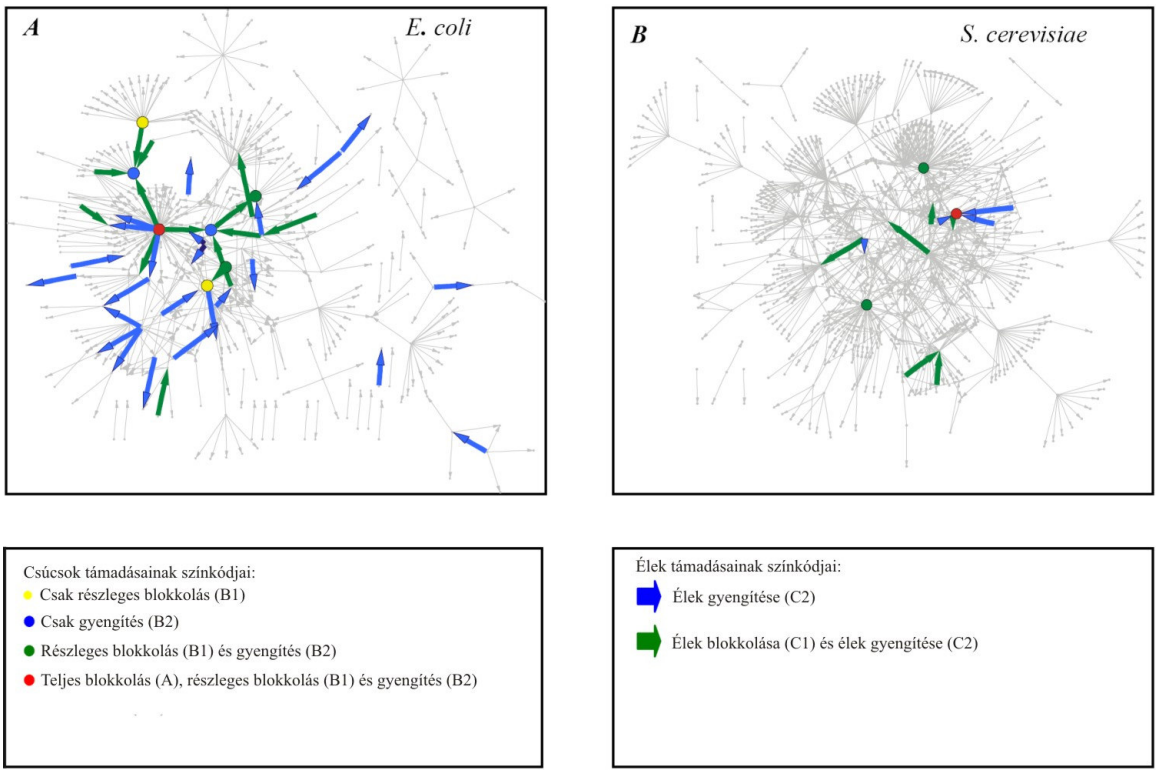
A 10. ábrán mutatom be, hogy a hálózatban egymáshoz képest hogy helyezkedtek el a különböző stratégiák által érintett elemek. Látható, hogy bár jellemzően a hálózat központi, sok összeköttetéssel rendelkező részei érintettek, ezek egymással ritkán vannak közvetlen összeköttetésben. Vagyis a támadás hozzávetőlegesen akkor okozza a legérzékenyebb kárt, ha a hálózat különböző alcentrumait éri egyidejű támadás. Figyelemreméltó viszont, hogy a különböző stratégiák a hálózatoknak többé-kevésbé ugyanazt a törzs-állományát támadják, vagyis az eredmények látszólag nem függenek az alkalmazott modellezési stratégiától.

Az egy célpontú blokkolás kvantitatív összehasonlítása a többcélpontú stratégiákkal

Hálózat	A) Egycélpontú blokkolás			B) Több célpont részleges inaktiválása				C) Elosztott hatások			
				B1) Az élek felének törlése[m]		B2) Minden él gyengítése		C1) Élek elosztott törlése		C2) Élek elosztott gyengítése	
	Törölt csúcsok száma	Törölt élek száma	Kár (%-os csökkenés az E-ban)	Csúcsok ekvivalens száma	Érintett élek száma	Csúcsok ekvivalens száma	Érintett élek száma	Érintett élek ekvivalens száma	Érintett csúcsok száma (élek %-a)[a]	Érintett élek ekvivalens száma	Érintett csúcsok száma (élek %-a)[a]
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
E. coli regulációs hálózata(N=424, E=521)	1	72[b]	15%	4.2	64.8[c]	5	129[d]	15	19 (5.8%)[e]	38[f]	53 (10.5%)[g]
S. cerevisiae regulációs hálózata (N=689, E=1080)	1	18[h]	6%	2.8	61.0[i]	3	142[j]	6	11 (3.1%)[f]	10[l]	16 (5.4%)[l]
Random irányított hálózat (N=424, E=521)[m]	1	6.0	20%	2.0	5.8	4.0	19.4	2.0	4.0 (19.7%)	5.0	8.2 (10.24%)
Random, irányított hálózat (N=689, E=1080)[m]	1	8.2	7%	2.0	6.4	2.0	7.6	2.0	4.0 (10.1%)	3.0	6.0 (9.84%)

[a]Pl. a 15 megtámadott él az *E. coli* hálózatában 5.8%-át teszi ki annak a 328 élnek, ami ahhoz a 19 csúcshoz tartozik, amiket a támadás érintett. (Ebben az esetben 11 olyan csúcs volt a maximálisan lehetséges 30 közül, amelyik több él végpontjaként is előfordult.)
[b]Érintett operonok (élek száma): crp(72)
[c]Érintett operonok (élek száma): crp(72) , rpoH (14), fliAZY (14), fnr (22), arcA (21), rpoE_rseABC (24)
[d]Érintett operonok (élek száma): crp (72), rpoH (14), fnr (22), fliAZY (14), flhDC (10)
[e]Érintett operonok (élek száma): arcA (21), cpxAR(10), crp (72), cspA (2), cytR (7), dnaA (2), flhDC (10), fliAZY (14), fnr (22), fur (10), hns (8), malt (7), mlc (4), nlpD_rpoS (14), ompR_envZ (7), rpoE_rseABC (24), rpoH (14), soxR (1), soxS (7)
[f]A támadások száma (38) azért lehet nagyobb, mint a megtámadott élek száma (56), mivel minden él kétszer támadható.
[g]Érintett operonok (élek száma): arcA (21), cpxAR(10), crp (72), cspA (2), cytR (7), dnaA (2), flhDC (10), fliAZY (14), fnr (22), fur (10), hns (8), malt (7), mlc (4), nlpD_rpoS (14), ompR_envZ (7), rpoE_rseABC (24), rpoH (14), soxR (1), soxS (7), acrAB (1), acrR (1), ada_alkB (2), adiA (1), adiA_adiY (1), aidB (3), alkA (2), appCBA (2), appY (3), atoC (3), betIBA (2), caiF (6), caiTABCDE (3), exuR (3), fadR (5), fecABCDE (1), fecIR (2), fhlA (4), fixABCX (2), fpr (2), GalR (2), gals (3), glnALG (4), himA (21), hypABCDE (3), iclMR (3), marRAB (6), metJ (4), metR (4), nac (4), nagBACD (4), rpoN (13), rtcR (2), uxuABR (2)
[h]Érintett operonok (élek száma): IME1 (18)
[i]Érintett operonok (élek száma): IME1 (18), STE12 (71), GCN4 (53)
[j]Érintett operonok (élek száma): IME1 (18), STE12 (71), GCN4 (53)
[k]Érintett operonok (élek száma): SNF2_SWI1 (20), SIN3 (13) SWI5 (11), MCM1 (13), HAP2_3_4_5 (26), MIG1 (26), DAL80 (20), DAL80_GZF3 (5), GAT1 (6), HSF1 (15), UME6 (38)

1. táblázat



10. ábra Támadási stratégiák grafikus összevetése

A különböző támadási stratégiák ugyanazt a központi törzsállományt jelölik ki.

Modellkísérletek egyéb hálózat-típusokon

A fenti eredmények viszonylag hosszú előkísérletek után születtek, melyekben arra is kíváncsiak voltunk, hogy az észlelt különbségek – vagyis hogy a több ponton történő részleges támadás hatásosabb az egy pontra összpontosított támadásnál – vajon a vizsgált irányított hálózat jellemzője csupán, vagy fennáll más topológiájú és irányítatlan hálózatokra is. Ezért a modellezést elvégeztük irányítatlan véletlen (Erdős-Rényi) gráfokra, és a kár mértékét nemcsak az efficiencia, hanem más jellemzők csökkenésével

is követtük. Az 1. táblázat utolsó két sora ebből mutat be néhány eredményt, a teljesség igénye nélkül.

Az eredmények hálózatonként igen változatos képeket mutatnak, ezért csak néhány általános jellemzőjüket emelem ki.

Az általunk vizsgált összes rendszerre (tehát a regulációs hálózatok irányított hálózati modelljeire, a velük megegyező topológiájú irányítatlan hálózatokra, és a velük megegyező csúcsot és élt tartalmazó véletlen hálózatokra) igaz az, hogy a célzott támadásra érzékenyebbek, mint a véletlenszerű támadásra. Ez megerősíti Barabási és Albert eredményeit.

Az összes általunk vizsgált rendszerre igaz, hogy néhány elem részleges támadása nagyobb kárt okoz, mint egyetlen elem teljes inaktiválása. Ugyanezt a kvalitatív eredményt kaptuk az efficiencia illetve a legnagyobb összekötött komponens vizsgálatával is.

Az eredmények diszkussziója

Eredményeinket érdemes megvizsgálni az eredeti kérdésfeltevés, vagyis a gyógyszertervezés kontextusában. Eredményeink megerősíteni látszanak azt a feltevést, hogy a jelenkori racionális gyógyszertervezés alap-paradigmája, a biológiai rendszerek egypontos támadása nem a leghatásosabb stratégia.

A többszörös hatású gyógyszerek fejlesztése várhatóan olyan termékeket eredményezne, amik kisebb affinitással hatnak kölcsön, mint az egyszeres hatású gyógyszerek, mivel valószínűtlen, hogy egy kis molekula ugyanolyan erősen kötődjön a számos különböző célponthoz. Bárhogy történjen is a dolog, az alacsony affinitású kötődés nyilvánvalóan nem hátrányos. Például a memantine (egy az Alzheimer-kór elleni küzdelemben használt gyógyszer) és más többszörös hatású szer a példa arra, hogy az alacsony affinitású, több ponton ható gyógyszereknél ritkábban és kisebb számban jelentkeznek mellékhatások, mint a nagy affinitású, egyetlen célponttal rendelkező gyógyszereknél. Vajon az alacsony affinitású kötődés azt eredményezi, hogy a gyógyszer kölcsönhatása a célpontjával eredménytelen? Nem feltétlenül. A legtöbb komponens a sejtek fehérje-, jeltovábbító- és transzkripciós hálózataiban amúgy is csak „gyenge kapcsolatban" áll egymással, azaz csak alacsony affinitással, vagy csak ideiglenesen kötődik egymáshoz, illetve közöttük alacsony-fluxusú metabolikus kapcsolat van. Mivel a legtöbb kapcsolat egy sejtes hálózatban gyenge, egy alacsony affinitású, több célponttal rendelkező gyógyszer is elegendő jelentős változások kiváltásához.

Fontos azonban megjegyezni, hogy következtetésünk nem tartalmaz kikötést arra, hogy a hatást egyetlen hatóanyaggal, vagy több hatóanyag kombinációjával érjük el. Következtetésünk tehát azt a jól ismert farmakológiai elvet is sugallja, hogy érdemes hatóanyagok kombinációival kísérletezni. Ezt a pontot azért érdemes hangsúlyozni, mert a többpontos hatás legismertebb példái esetében a nem tervezett, talán nem is tervezhető hatások összege adja a kedvező összhatást. Feltehető, hogy többhatású farmakonokat a

jövőben is nehéz lesz tervezni. Lehetőség van azonban arra, hogy a racionálisan tervezett molekulákat kombináljuk. Modellkísérleteink egyik tanulsága tehát az lehet, hogy ésszerű gyógyszerfejlesztési megközelítésnek tűnik, hogy a racionálisan tervezett molekulák hatásait adjuk össze, azaz koktélterápiát alkalmazunk.

A felhasznált algoritmusok vázlata

A maximális hálózati kár becslésére szolgáló mohó algoritmus pszeudokódjának vázlata:

Input:G hálózat, N a támadások száma

```
For ( i = 1 to N )
Begin
        MinEff = A G efficiencája                                   (1)
        For j a G minden csúcsára
        Begin
                G2 = G
                A j csúcs eltávolítása G2-ből                       (2)
                G2Eff  = A G2 efficienciája                         (3)
                If (G2Eff <  MinEff) MinEff = G2Eff, MinCsúcs = j   (4)
        End

        Output (MinCsúcs, MinEff)

        MinCsúcs törlése G-ből                                      (5)
End
```

A kárt egy adott hálózati paraméter csökkenésével jellemeztem, a fenti példában a hálózati efficienciát használtam, az [1] egyenlet szerint. Ez a változó helyettesíthető tetszőleges más integritás-mértékkel, ekkor az (1), (3), (4) sorok változtatása szügséges. A fenti példában a támadási stratégia egy csúcs és az összes abba futó él eliminálásán alapul. Ez helyettesíthető bármely más támadási stratégiával, ld. a 7. ábra, ezekben az esetekben a (2) és az (5) sor értelemszerűen módosul. A dolgozatban bemutatott eredmények mindegyikét a fenti vázlat szerint számítottam ki.

A diszulfidhidak keletkezésének hálózati modellezése

Kérdésfelvetés

Második kutatási témám annak vizsgálata volt, hogy a hálózati modellek hogyan alkalmazhatóak az oxidatív hajtogatódás konformációs terének leírására és megjelenítésére. A feladat érdekes, mivel a „közönséges" hajtogatódással ellentétben a diszulfid kapcsolatok által meghatározott állapotok száma nem túlzottan nagy, mitöbb a valódi diszulfid-intermedierek kísérletesen izolálhatók és tanulmányozhatóak. A probléma megközelítésének két fázisát különböztethetjük meg. 1) a diszulfid intermedierek gráfelméleti leírása, és a hajtogatódási tér állapotainak sorszámokkal történő megjelölése és 2) a hajtogatódási tér minden állapotának megjelenítése egy hálózaton (gráfon).

Egy fehérje diszulfid topológiáját egyértelműen meghatározza az, hogy melyik ciszteinek kapcsolódnak egymáshoz. Pl. az '1-3, 2-4' leírás jelentése, hogy a fehérjében két diszulfid híd alakult ki, méghozzá egy az első és harmadik cisztein között, és egy a második és a negyedik között. A ciszteinek számozása történhet a szekvenciában elfoglalt pozíciójuk alapján, vagy ahogy az előző példában is, az N terminális végtől sorszámozva őket.

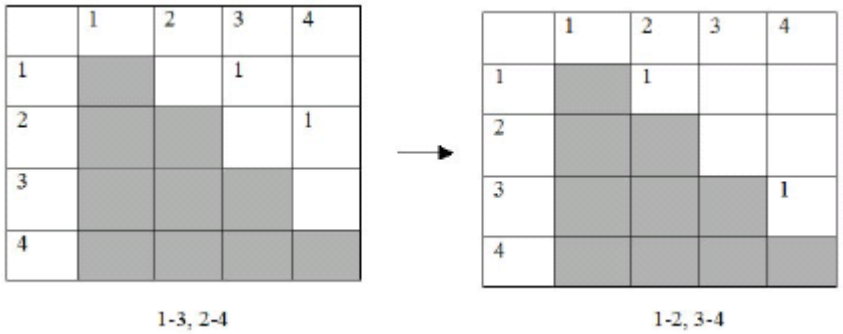
Gráfelméleti modell

Egy teljesen oxidált, n diszulfid hidat (2n ciszteint) tartalmazó fehérjének (2n)! /(n!*2^n) izomerje van. Ez 3 diszulfid híd esetén 15, 4 diszulfid híd esetén 105 izomert jelent, vagyis a kötést alkotó ciszteinek függvényében igen gyorsan növekszik az izomerek száma.

Ciszteinek száma	Intermedierek száma (csúcsok)	Redox átmenetek	Felcserélődési átmenetek	Összes átmenetek száma (élek)	Klaszterezettségi együttható	Átlagos úthossz
1	1	0	0	0	1.000	0.000
2	2	1	0	1	1.000	1.000
3	4	3	3	6	1.000	1.000
4	10	12	12	24	0.400	1.467
5	26	40	60	100	0.410	1.810
6	76	150	240	390	0.247	2.293
7	232	546	1050	1596	0.253	2.640
8	764	2128	8736	10864	0.181	3.149
9	2620	8352	19152	27504	0.182	3.550
10	9496	34380	83520	117900	0.142	3.977

2. táblázat

A hajtogatódási folyamat leírásához figyelembe kell vennünk az oxidációs folyamat köztes állapotait is. Ennek érdekében vezessük be a következő formális leírást rájuk: Olyan gráfok, amelyeknek a csúcsai ciszteinek, és két csúcs között akkor van él, ha a két cisztein diszulfid hidat alkot. Az ilyen gráfok szomszédsági mátrixa (11. ábra) szimmetrikus, ezért elegendő a jobb felső háromszöget tekinteni. Mivel minden cisztein csak egy kötésben vehet részt, minden sorban és oszlopban legfeljebb egy 1-es lehet. Példaként ld. a 11. ábrát.



11. ábra Ciszteinkötések mátrixreprezentációja

Az ábra egy két diszulfid hidat tartalmazó hipotetikus peptid két állapota közötti átmenetet ábrázolja. A baloldali mátrix az 1-3, 2-4 diszulfid hidaknak felel meg, a jobboldalon az 1-2, 3-4 állapot található.

A köztes állapotok közötti átmenetek egyértelműen leírhatóak a szomszédsági mátrixaik összevetésével. Ehhez vezessünk be néhány jelölést. Legyen NB a cisztein hidak száma, vagyis:

$$NB = \sum_{i,j} A_{ij} \qquad [10]$$

Az elemek összege az i-ik sorban és az i-ik oszlopban

$$S_i = \sum_j A_{ij} + \sum_j A_{ji} \qquad [11]$$

1, ha az i-ik cisztein része egy diszulfid hídnak, és 0 egyébként. A szomszédsági mátrixok közötti eltérések összege,

$$SD = \sum_i \Delta S_i \qquad [12]$$

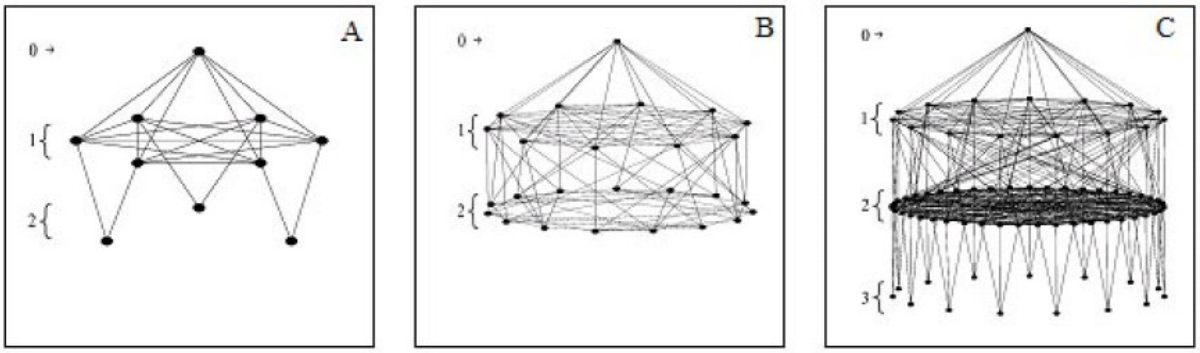
megmutatja, hogy hány cisztein lépett ki egy kötésből, vagy be egy kötésbe amíg a fehérje az egyik állapotból a másikba jutott. Számunkra csak az a két alapreakció érdekes, amit a bevezető 2. ábráján bemutattunk. Felcserélődési reakcióban NB, vagyis a cisztein hidak száma nem változik, és SD pontosan 2. Redox reakcióban NB 1-el változik, SD pedig 2.

Ezek alapján könnyen létrehozható az a hálózat, ahol az összes lehetséges köztes állapot szerepel mint csúcs, a lehetséges átmenetek pedig élek. A 2. táblázatban összefoglalt eredményekből látható, hogy a ciszteinek számának növekedésével a klaszterezettségi együttható csökken, az átlagos legrövidebb út pedig nő. Ezek összhangban vannak azzal az intuitív képpel, hogy a sok ciszteinnel rendelkező peptidek hajtogatódási tere túl komplexé válhat, így az ilyen rendszerek hajtogatódása nagymértékben lelassulhat.

Grafikus megjelenítés

Ennek a hálózatnak a 3D megjelenítésében külön síkokra helyezve a különböző számú diszulfid hidakkal rendelkező állapotokat, egy a probléma méretéhez képest átlátható szerkezetet kapunk (12. ábra). Az egyes síkokon reguláris gráfok helyezkednek el. Két szomszédos sík között pedig olyan élek futnak, amelyek cisztein híd kialakulását, vagy megszűnését reprezentálják. A legfelső sík reprezentálja a teljesen redukált állapotot, a legalsó pedig a teljesen oxidált állapotokat. A 12. ábra B. panelje egy olyan fehérje hajtogatódási terét mutatja, aminek páratlan számú ciszteinje van, például a granulocyte-colony stimulating-factor. Az ilyen fehérjék esetében még a natív állapotban is van

szabad cisztein, így még a legalsó síkon is találhatóak felcserélődési élek. Ha a fehérje páros számú ciszteint tartalmaz, akkor a teljesen oxidált állapotok közötti átmenet csak több lépésben valósulhat meg. A gyakorlatban előfordul, hogy az in vitro reakció felgyorsítására egy hozzáadott ciszteint használnak.



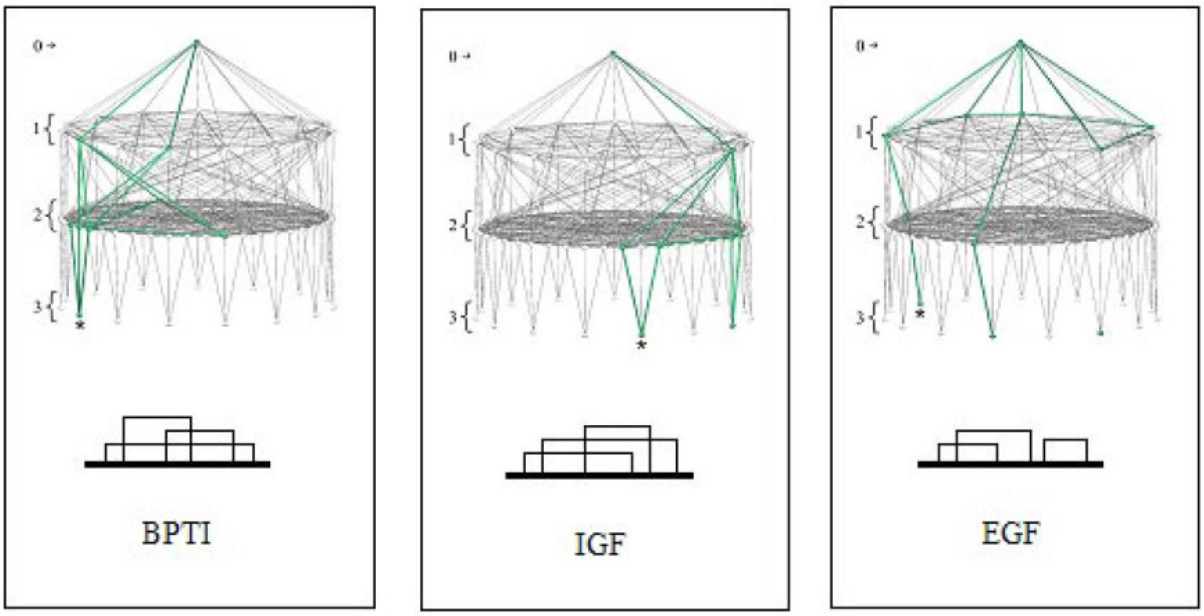
12. ábra 4, 5 és 6 ciszteint tartalmazó fehérjék hajtogatódási tere

Gráfként ábrázolhatjuk a hajtogatódási teret, minden csomópont egy köztes állapotnak felel meg, minden él egy lehetséges átmenet két állapot között. A legfelső szinten található a teljesen redukált állapot, az azonos szinteken pedig azonos oxidációs szintűek a köztes állapotok.

Az oxidatív hajtogatódás útvonalai ebben a hálózatban olyan utakként ábrázolhatóak, amik a teljesen redukált állapotból indulnak és a natív állapotban érnek véget. Az irodalomban csak néhány alaposan tanulmányozott esetben vannak leírva az intermedierek. A kísérletesen igazolt intermedierek három példája a bovine pancreatic trypsin inhibitor, insulin-like growth factor és az epidermal growth factor, ezek oxidatív foldingjából alkotott hálózatainak ábrázolásai a 13. ábrán láthatóak, a számszerű adatok pedig 3. táblázatban vannak összefoglalva. A kísérletes módszerek nem feltétlenül fedik fel az összes intermediert, lehetnek köztük olyanok, amelyek túl rövid életűek, vagy nem jelennek meg elég nagy mennyiségben ahhoz, hogy ki lehessen őket mutatni. Egy hajtogatódás ábrázolása során tapasztalt izolált pont, mint például az EGF esetében látható, olyan intermedierekre utal, amelyeket nem sikerült kísérletesen feltárni.

Fehérje	Diszulfid intermedierek	Ref.
Bovine pancreatic trypsin inhibitor (BPTI)	3-5; 1-6; 3-5, 1-2; 3-5, 1-4; 3-5, 2-4; 1-6, 2-4; 3-5, 1-6; **1-6, 3-5, 2-4**;	(76), (77)
Insulin-like growth factor (IGF)	2-6; 2-6, 3-5; 2-6, 1-4; 2-6, 4-5; 2-6, 1-3; 2-6, 1-3, 4-5; **1-4, 2-6, 3-5;**	(78). (79), (80)
Epidermal growth factor (EGF)	2-3; 1-2; 4-6; 5-6; 3-4; 2-4, 5-6; 2-5, 3-4; 1-6, 2-5, 3-4; 1-2, 3-4, 5-6; **1-3, 2-4, 5-6;**	(81)

3. táblázat



13. ábra 3 fehérje hajtogatódási hálózata

A bovine pancreatic trypsin inhibitor, az insulin-like growth factor és az epidermal growth factor hajtogatódásának megfeleltetett hálózatok. A *-al jelölt csúcsok felelnek meg a natív állapotnak, amelyeknek a sematikus ábrázolása a hálózat alatt látható.

Az eredmények diszkussziója

A szerkezetkutatás területén különleges szerepet játszanak az állapot-gráfok, melyekben a csúcspontokat intermedierek, az éleket pedig a köztük lévő átalakulások jelentik. Tágabb értelemben ilyenek a metabolikus útvonalak hálózatai, melyek az élő szervezeteket alkotó vegyületeket és azok enzim-katalizált átalakulásait ábrázolják. A fehérjék foldingjára nehéz ilyen hálózatokat felvázolni, mivel az intermedierek nagy többségét nem ismerjük, sőt kevés remény látszik arra, hogy kísérletesen megismerjük őket. A diszulfid-hidak intermedierjei ebből a szempontból kivételesek, mert ezeket – a fenti eredmények szerint - tételesen fel lehet sorolni, és meg lehet határozni az összes lehetséges átalakulást is. Ezekből állnak össze az általam konstruált elméleti állapot-gráfok. Az állapotok és az átmenetek száma meredeken növekszik a cisztein csoportok függvényében. Akár a konformációs, nem-oxidativ folding esetén, itt is feltehető, hogy a rendszernek nincs elegendő ideje bejárni az összes állapotot. Akárcsak a konformációs folding esetén, itt is fel kell tételeznünk, hogy a folyamat preferált útvonalakon halad, tehát az összes állapot közül csak néhány, egymáshoz közel elhelyezkedő állapot játszik benne szerepet. Pontosan ezt a képet kapjuk, ha az elméleti állapot-hálózatra rávetítjük az irodalomban található diszulfid-intermediereket: a folding útvonalak jól látható, összefüggő vonalat alkotnak. Ha találunk is izolált intermediereket, akkor is feltételezhetjük, hogy néhány további intermedier analízisével kialakulnak az összefüggő hálózatok. Az oxidatív folding tanulmányozásakor ugyanis gyakran csak a legnagyobb mennyiségben megjelenő intermediereket szokták meghatározni, és mellettük számos kisebb intermedier is található. Az itt javasolt ábrázolási módszer megkönnyíti, hogy ne csak a gyakori intermediereket, hanem az összefüggő folding-utakat is jellemezni tudjuk.

A hálózat felépítésének vázlata

A C darab ciszteint tartalmazó hálózat felépítésére használt algoritmus vázlata:

Input: C, a ciszteinek száma

```
N=1 //Csúcsok száma
For(i = 1 to [C/2])
Begin
        N = N+(C!/(2^i*i!*(n-2*i-2))
End
For (i = 1 to N)
Begin
        For (j = 1 to i)
        Begin
                If( |SD| = 2 AND ( |NB(i)-NB(j)| <= 1)) AddEdge(i, j)
        End
End
```

Fehérjék hasonlósági hálózatai

Kérdésfelvetés

Harmadik kutatási témám a fehérjék hasonlósági hálózatainak vizsgálata volt. A hálózati modellek sok új fogalom és jellemzési módszer keletkezéséhez járultak hozzá, ezeket kívántam alkalmazni a fehérjék hasonlósági hálózataira. E vizsgálatnak gyakorlati jelentősége is van, a fehérjék hasonlósági hálózatait ugyanis szívesen, és napjainkban különösen gyakran alkalmazzák a fehérjék funkciójának predikciójánál és a fehérje evolúció leírására is. Egyre inkább terjed ugyanis az a felfogás, hogy a biológiai ismereteket egymással összeköttetésben álló adathálózatok formájában képzeljük el, az információkeresés is leginkább egy hálózaton való bolyongásként írható le. Ez a megállapítás több mint metafora, hiszen a keresési módszerek új algoritmusainak egy jellemző irányvonala az ún. hálózati propagáció (pl. belief networks, message-passing stb.), a Google által is alkalmazott PageRank algoritmust (82) például RankProp néven sikerrel alkalmazták a szekvencia-hasonlóságok keresésére is (83).

Mai adatbázisaink közül a fehérjeadatbázisok a leggazdagabbak, mert bennük nem csak a molekuláris biológia, de a biokémia és a szerkezetmeghatározás eredményei egyaránt összpontosulnak. Célom az volt, hogy a hálózatok kutatásában felmerült módszereket alkalmazzam néhány jól definiált fehérje-adatbázis vizsgálatára, valamint leíró jelleggel jellemezzem a fehérjék hasonlósági hálózatainak alapvető topológiáját, és hogy

eldöntsem, a többi biológiai hálózat modularitására vonatkozó jelek észrevehetők-e a fehérjék hasonlósági hálózatain.

Alapösszefüggések

Ha adott elemek egy halmaza, és egy rajtuk értelmezett hasonlóság, akkor ezt az együttest matematikai szempontból súlyozott teljes gráfnak foghatjuk fel, ugyanis minden elemet valamilyen fokú hasonlóság köt össze bármelyik másik elemmel. Ha ebből a teljes gráfból elhagyjuk azokat az éleket, amelyek gyengébbek egy adott küszöbértéknél, az eredmény egy a biológiai gyakorlatban használtakra emlékeztető hasonlósági hálózat lesz, amely már csak a valamilyen (biológiai) szempontból fontos hasonlóságokat tartalmazza. Több olyan biológiai adatbázis is ismert, ahol nemcsak az elemek közötti hasonlóság értelmezett, hanem egy osztályozás is. Az ilyen strukturált rendszer esetében érdemes megkülönböztetni belső és külső hasonlóságokat. A belső hasonlóságok az azonos osztályba tartozó elemek közöttiek, ezek tehát például a biológiailag releváns hasonlóságok, az ezeknek megfeleltethető élek egy csoportba tartozó csúcsokat kötnek össze. A külső hasonlóságok pedig az egymással nem-rokon csoportok elemei között állnak fenn, melyeket az adott biológiai osztályozás szempontjából irreleváns, véletlenszerű jelenségeknek tekinthetünk.

Ennek a fejezetnek a nagyobbik része leíró jelleggel összefoglalja azt, ami általában megállapítható a hasonlósági hálózatokról. Mivel több ilyen rendszer vizsgálatát

végeztem el, az eredmények felsorolása mellett, ahol lehetőség nyílik rá, ott az összehasonlításból levonható következtetésekre is szeretnék kitérni.

A fehérje-adatbázisok egyik alapproblémája, hogy nagyon nagy méretűek, „zajosak" (azaz bizonytalannak tekinthető adatokat is tartalmaznak), heterogének (többféle technikával meghatározott adatot tartalmaznak), és sok bennük a redundancia is. A fehérje-adatbázisok másik jellemzője, hogy a fehérjék a genom-annotáció során eleve osztályba vannak sorolva – ezért vizsgálni kívántam osztályok külső és belső hasonlóságait. A fehérje-osztályok (csoportok) maguk is igen heterogének, mind a tagszám, mind az alkotó fehérjék lánchosszúsága, mind pedig a közöttük lévő hasonlóság erőssége szempontjából. A vizsgálathoz ezért néhány jól-gondozott adatbázist választottam ki, amelyeknél a redundancia kiszűrhető, és a csoportokat szakértő kutatók egyedileg validálták. A fehérjék evolúciójának alapköveinek tekintett fehérjedomének információit nyilvántartó adatbázisokat választottam vizsgálataimhoz, mivel a köztük lévő hasonlóságokat nem terhelik a teljes (többdoménes) fehérjékre jellemző problémák.

Igyekeztem olyan adatbázisokat kiválasztani, amelyekre mind a térszerkezet, mind a fehérje-szekvencia adatai rendelkezésre állnak. A kiválasztást befolyásolta, hogy a gyakorlatban alkalmazott adatbázisok már túlságosan nagyok és redundánsak, és az osztályhasonlóságok jellege szerint is igen nagy a változatosság. A doménszekvencia-adatbázisok (pl. PFAM (84), SBASE (85)) hasonlósági csoportjai például részben a különböző multidomén fehérjecsaládokban is előforduló doméneket tartalmazzák, részben pedig egydoménes fehérjecsaládok tagjait. Ezeket az adatbázisokat a szekvencia-

hasonlóságok alapján fejlesztik, így a csoportok tagjai között nagy a szekvencia-hasonlóság.

Más a helyzet a szerkezeti tulajdonságok alapján csoportosított domén-gyűjteményeknél. Itt a csoportosítás alapja a háromdimenziós architektúra, és az azonos felépítésű fehérjék szekvenciái között nem is mindig észlelhető hasonlóság. Ilyen adatbázisok például a CATH (86), és a SCOP (87), az előbbit számítógépes módszerekkel, az utóbbit többségében emberi értékelés alapján hozták létre. Ezen adatbázisok fehérjéi között a szerkezeti hasonlóság a kapocs, a szekvenciális hasonlóság néha a csoport tagjai között is elég laza, a távoli rokonságot mutató doméncsoportok között pedig sokszor alig észlelhető.

Végül harmadik típusként említhetők az ortológ csoportokat – azaz evolúciós szempontból rokon és azonos funkciót ellátó fehérjék csoportjait gyűjtő funkciós adatbázisok, melynek archetípusa a Eugene Koonin csoportja által létrehozott COG, a Clusters of Orthologous Sequences (88). A COG csoportok tagjai között határozott szekvenciális hasonlóság van (a válogatás alapja a BLAST, tehát egy viszonylag nem túl érzékeny algoritmus, amit ráadásul magas küszöbértékkel futtatnak), ugyanakkor a csoportok között többnyire csak gyenge, véletlenszerű hasonlóságok vannak.

A vizsgálatokat a 4. táblázatban felsorolt adatbázisokon végeztem.

Adatbázis		Elemek száma	Csoportok száma
Neve	Rövid leírása		
Pfam seed	A PFAM adatbázis (84) kézzel-rendezett törzsállománya, amely doménszekvenciákat tartalmaz, hasonlóság szerint csoportosítva.	128732	8250
COG	A COG (83) mikrobiális ortológ fehérjecsoportokat tartalmaz funkció szerint csoportosítva.	102464	4873
CATH	A CATH adatbázis (86) térszerkezeti doméneket tartalmaz, térszerkezeti felépítés szerint hierarchikusan csoportosítva. A megfelelő szekvenciákat és térszerkezeteket a hierarchia ún. family szintjén csoportosítottam.	17844	1459

4. táblázat A felhasznált adatbázisok főbb jellemzői

Az adatbázisok 2003-as verziójából származó adatokat analizáltam. A szekvencia-hasonlóságot BLAST-tal számítottam ki, a raw score küszöbértéknek 40-et választva (43) a térszerkezeti hasonlóságokat a PRIDE algoritmussal (58) számítottam ki.

A hasonlósági mérőszámok

Igyekeztem azokat a hasonlósági mérőszámokat alkalmazni, amelyek a molekuláris biológiában a legelfogadottabbak. Szekvencia-összehasonlításra a BLAST program (43) *S* hasonlósági mérőszáma (raw similarity score) a legelterjedtebb, ezt használtam a számításokhoz. Az *S* mérőszám pozitív érték, melynek általában csak egy bizonyos tapasztalati szignifikancia-küszöb fölötti értéket adó hasonlóságokat szokták kiértékelni. A számításoknál én is az általánosan elfogadott S=40 küszöböt alkalmaztam.

A szerkezeti összehasonlításra használt algoritmusok nagyon időigényesek, a csoportunknál kifejlesztett gyorsmódszer, a PRIDE fő előnye relatív gyorsasága, ezért ezt használtam (58). Ennek hasonlósági valószínűség-mérőszáma nulla és 1 közé esik, és a gyakorlatban a 0.5-ös érték felett szokták szignifikánsnak tekinteni az eredményeket.

A hálózati ábrázolásoknál a fehérjék alkották a csúcsokat, az élek a köztük fennálló hasonlóságok voltak, az élek súlya pedig a megfelelő mérőszám (BLAST vagy PRIDE score) volt. A számításokhoz szükséges programokat C nyelven írtam meg és Linux operációs rendszer alatt működő, 6 processzoros PC hálózaton futtattam le. Az alábbiakban csak a lényegesebb megállapításokhoz szükséges eredményekre térek ki, a számítások teljes eredményeit a http://aladar.szbk.u-szeged.hu/figures.htm weboldalon helyeztem el. Módszertani szempontból megjegyzem, hogy a hálózatok fokszám- és egyéb paraméter-eloszlásait szokás a jobb megjelenés érdekében simítani. Véleményem szerint egy újszerű jelenség megismerését jobban segíti, ha a nyers adatokat vesszük szemügyre, így az ábrázolásokon nem alkalmaztam simítást.

Teljes hálózatok

Az előzetes értékeléshez a következő mennyiségeket számítottam ki:

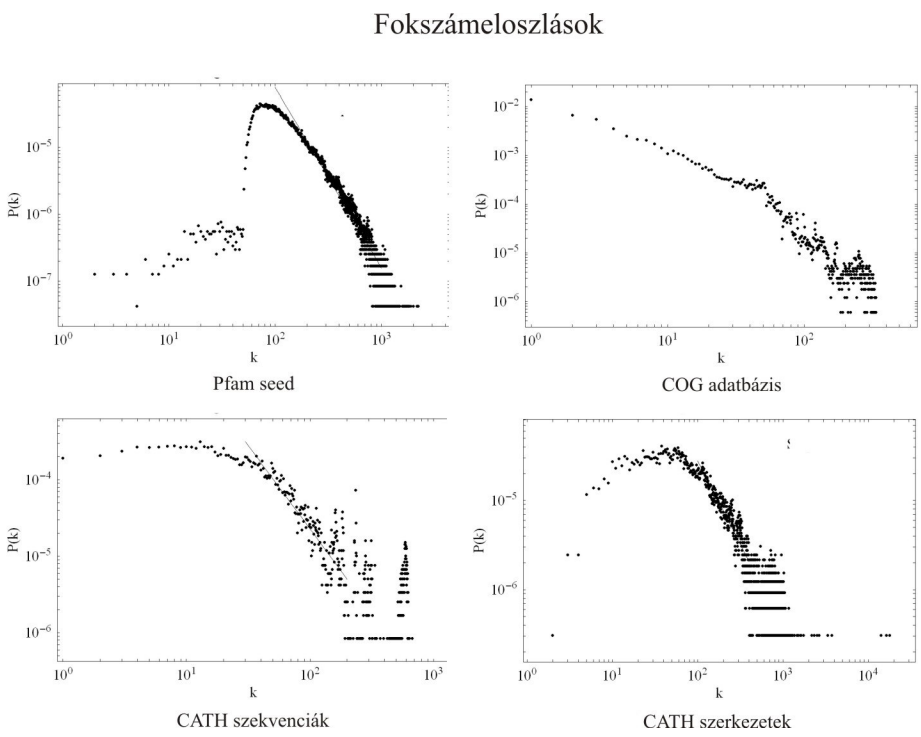
- Csúcsok és élek száma
- A csúcsok fokszáma, fokszámeloszlás
- Klaszterezettségi együttható (Ez kifejezi, hogy a csúcsok szomszédai mennyire vannak összeköttetésben.)
- Csúcsok átlagos ereje (Ez az egy csúcsba futó élek súlyainak összege.).
- Csúcsok affinitása. (Ez a szomszédos csúcsok átlagos ereje.)

A számított értékeket az 5**.** ábra foglalja röviden össze. Megállapítható, hogy a hasonlósági hálózatok átlagos paraméterei feltűnően nem különböznek az eddig analizált technikai hálózatokétól. Mind a szekvencia-, mind a térszerkezeti hasonlóságok hálózatai ritka hálózatoknak tekinthetők (a csúcsok átlagos fokszáma viszonylag alacsony), ugyanakkor a klaszterezettségi együttható viszonylagosan magas értéke jelzi, hogy a csúcsok többé-kevésbé összekötött szomszédságokat, hasonlósági klasztereket alkotnak a hálózaton belül. Ez nem meglepő, hiszen ha a biológiai osztályozás releváns, akkor a csoportokon belül több hasonlósági élnek kell lennie, mint a csoportok között. Formai szempontból viszont úgy tűnik, hogy a hasonlósági hálózatok is az ún. „kis-világ" hálózatok (17) közé tartoznak, melyeket az alacsony élszám és a magas klaszterezettségi együttható jellemez.

Adatbázis		Csúcsok (élek) száma	Átlagos paraméter (minimum-maximum)			
			Fokszám	Klaszterezettségi együttható	Erő (strength)	Affinitás
Pfam seed	teljes[1]	128732 (11843517)	184,0 (2/2216)	0,20 (0/1)	13546,8 (71/164292)	241,5 (12,15/868,5)
	külső[1]	128720 (6458865)	100,4 (1/2208)	0,009 (0/1)	5027,2 (30/144243)	199,8 (12/2163)
COG	teljes[1]	102464 (838765)	16,4 (1/338)	0,71 (0/1)	14537,9 (501/236144)	17,7 (1/337)
	külső[1]	77024 (1907931)	49,5 (1/1285)	0,15 (0/1)	11402,5 (101/470216)	105,8 (1/752,2)
CATH szekvencia	teljes[1]	17706 (594139)	67,1 (1/674)	0,70 (0/1)	16756,7 (51/140304)	74,2 (1/583,1)
	külső[1]	17659 (578524)	65,5 (1/664)	0,64 (0/1)	15887,5 (139995)	73,9 (1/577,0)
CATH szerkezet	teljes[1]	17844 (1631443)	182.9 (2/17843)	0,59 (0,01/1)	110,9 (1,126/10192,6)	1,5 (0,18/17,4)
	külső[1]	17844 (1048999)	117,6 (2/17843)	0,46 (0,007/1)	64,5 (1,0/10192,6)	2,4 (0,12/17,1)

5. táblázat A vizsgált hálózatok főbb átlagos paraméterei (részletes adatok)

[1]A teljes hálózat adatai megegyeznek a 4. táblázatban írottakkal, a külső élek a csoportok közötti hasonlóságokat jelentik, ebből a hálózatból tehát elhagytam az azonos csoporton belüli fehérjéket összekötő éleket.



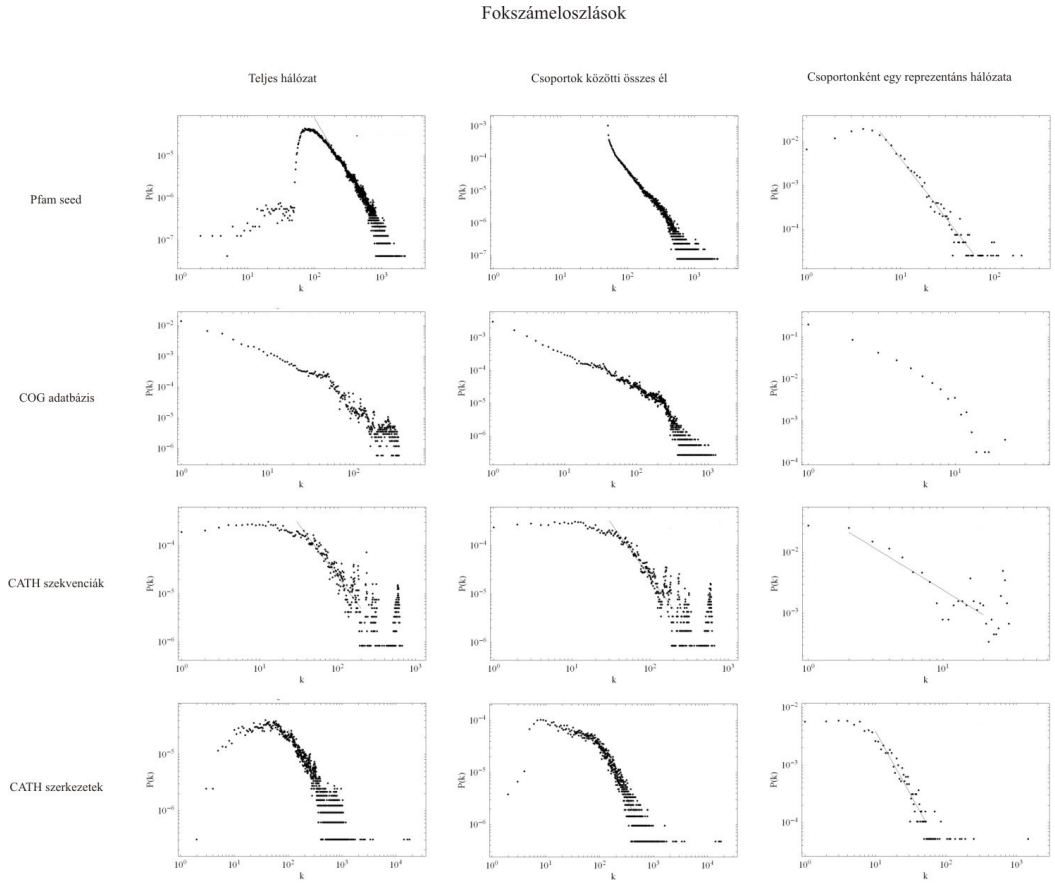
14. ábra A fokszámok eloszlása a különböző hasonlósági hálózatokon
Az ábra a teljes fokszámot mutatja, melyben benne vannak a csoporton belüli és a csoporton kívüli élek is.

Részletesebb információt ad a fokszám-eloszlás vizsgálata, amelyet a 14. ábra mutat be. Ezt az ábrázolást Barabási alapvető felismerése óta log-log diagrammal szokták végezni, mert ebben az ún. hatvány-függvények megfelelő skálamentes eloszlás negatív meredekségű egyenes szakasz formájában jelenik meg, melynek meredeksége a hatványfüggvény kitevője. Barabási és munkatársai az általuk vizsgált hálózatoknál többnyire -2-körüli kitevőket találtak. Egyenes szakaszt jobb-rosszabb közelítéssel az itt vizsgált hálózatoknál is fel lehet fedezni, tehát mondhatjuk, hogy a hasonlósági hálózatok fokszám-eloszlása bizonyos fokig megközelíti a skálamentes eloszlást, a kitevő értéke pedig -1,3 és -2,8 között van. Mindez köznapi nyelven azt jelenti, hogy ritkán, de előfordulnak nagyon nagy fokszámú csúcsok is, ugyanakkor a csúcsok többségének nagyon kevés éle van. Itt népszerű példaként a repülőjáratokat szokták emlegetni, a távoli vidékek kis repterei általában csak egyetlen másikkal, a legközelebbi nagyobb reptérrel vannak összekötve. A hasonlósági hálózatoknál más a helyzet, mert itt már eleve olyan csoportokat gyűjtenek, melyeknek több tagja van, tehát a „magányos fehérjék” (singleton-ok) nincsenek képviselve ezekben a gyűjteményekben. Az adatgyűjtés hatása tehát mutatkozik a fokszámeloszláson is.

Csoportok közötti élek fokszámeloszlása

Érdemes áttekinteni, hogy mennyiben különböznek a csoportok közötti élek hálózatai, ezek átlagos adatait a 5. táblázat „külső élek” sora foglalja össze. A várakozással ellentétben a külső élek klaszterezettsége nem gyengébb, mint a teljes hálózaté, jóllehet a csoportok önmagukban gyakran erősen klaszterezettnek tűnnek (ld. lejjebb).

A csoportok közötti élek fokszámeloszlását kétféleképpen vizsgáltam meg. Az első módszernél elhagytam minden olyan élt, amely azonos csoporton belüli fehérjéket köt össze. A második módszernél minden csoportból kiválasztottam egy reprezentánst, tehát így olyan – kevesebb, pontosabban csoportonként egyetlen csúcsot tartalmazó – hálózat maradt, melyben csak a nagy hálózat csoportok közötti élei szerepelnek. Mindkét fokszámeloszlást ábrázoltam, az eredményeket a 15. ábra foglalja össze. Összehasonlításképpen az ábrán megjelenítem a teljes hálózat fokszámeloszlását is. Megállapítható, hogy ugyanarra az adatbázisra mindhárom ábrázolás többé-kevésbé hasonló eloszlást ad, amit értelmezhetünk úgy, hogy a teljes hálózatok eloszlását (baloldali oszlop) dominálja a csoportok közötti eloszlás (középső oszlop). A lineáris – lineárisra emlékeztető – szakaszok itt is megjelennek, a meredekségük az illesztés hibahatárán belül nem nevezhető szignifikánsan eltérőnek.



15. ábra A teljes hálózat és a csoportok közötti fokszámok eloszlásának összehasonlítása

Az eloszlások jellege lényegében hasonló az összes adatbázison.

A csoportok közötti hasonlóságokat a biológiában véletlenszerűnek szokták nevezni, és ez logikus is, hiszen ha a csoportokat eltérő, egymással nem rokon fehérjecsaládok alkotják, akkor a közöttük fennálló szekvencia-hasonlóságok a józan ész alapján véletlennek vélhetjük. Ezt a képet persze finomítani kell, hiszen ha strukturális hasonlóságokat ábrázolunk, akkor például minden alfa-hélix hasonló lesz egymáshoz, s így az alfa-hélixet tartalmazó fehérjék csoportjai között már lesznek hasonlóságok. Ezt a gondolatkísérletet folytatva megjegyezhetjük, hogy az alfa-hélixekben léteznek bizonyos preferált, rövid (pl. 4-5 aminosavat magukba foglaló) szekvenciarészletek, tehát *elvben* ez megmutatkozhatna a csoportok közötti szekvencia-hasonlóságokban is (gyakorlatban

nem mutatkozik meg, mivel az ilyen rövid hasonlóságok mérőszámai az értékelési küszöb alatt szoktak maradni). Ebből következik, hogy a biológiai szempontból irreleváns hasonlóságok hátterében vannak nem-véletlenszerű elemek, tehát érdemes megvizsgálni, hogy kapott eloszlásaink tükrözik-e vajon ezeket.

Az első eldöntendő kérdés, milyen eloszlást kapunk akkor, ha a hasonlóságokat leíró súlyok eloszlása véletlenszerű. Legyen adva egy súlyozott teljes gráf, melyben a súlyok nagyságának eloszlása tetszőleges, és a súlyok véletlenszerűen vannak kiosztva ez élek között. Ekkor a súlyok minimuma és maximuma között akárhogy választunk egy küszöb értéket, aminél kisebb súlyokkal rendelkező éleket töröljük. Triviálisan belátható, hogy a megmaradó gráf fokszámeloszlása binomiális lesz. Ebből a szempontból érdekes megjegyezni, hogy a csoportok közötti élek eloszlása a 14. ábra szerint viszont nem binomiális, a hálózat – a súlyeloszlása alapján- nem véletlenszerű. Ennek a jelenségnek a részletes magyarázatára tulajdonképpen nem vállalkozhatunk – még a közelítő-magyarázathoz is valamiféle modellt kellene alkotni, és azt a kísérleti adatokhoz illesztve is maximum csak érvelhetnénk a magyarázat valószínűsége mellett. A fenti érvek (pl. az alapvető strukturális elemek közötti hasonlóságok) azonban sejtetik, hogy léteznek olyan faktorok, amelyek miatt a csoportok közötti hasonlóságok eltérhetnek a véletlenszerűtől.

Kérdés, hogy a jelenleg biológiailag irrelevánsnak tekintett hasonlóságok hordoznak-e valamilyen fontos információt például a hasonlóság-keresés szempontjából. Erre indirekt választ adhatunk. Munkámmal szinte egyidőben jelentek meg a hasonlósági hálózatokon működő propagációs algoritmusok, elsősorban a RankProp (82), mely a hasonlóság-

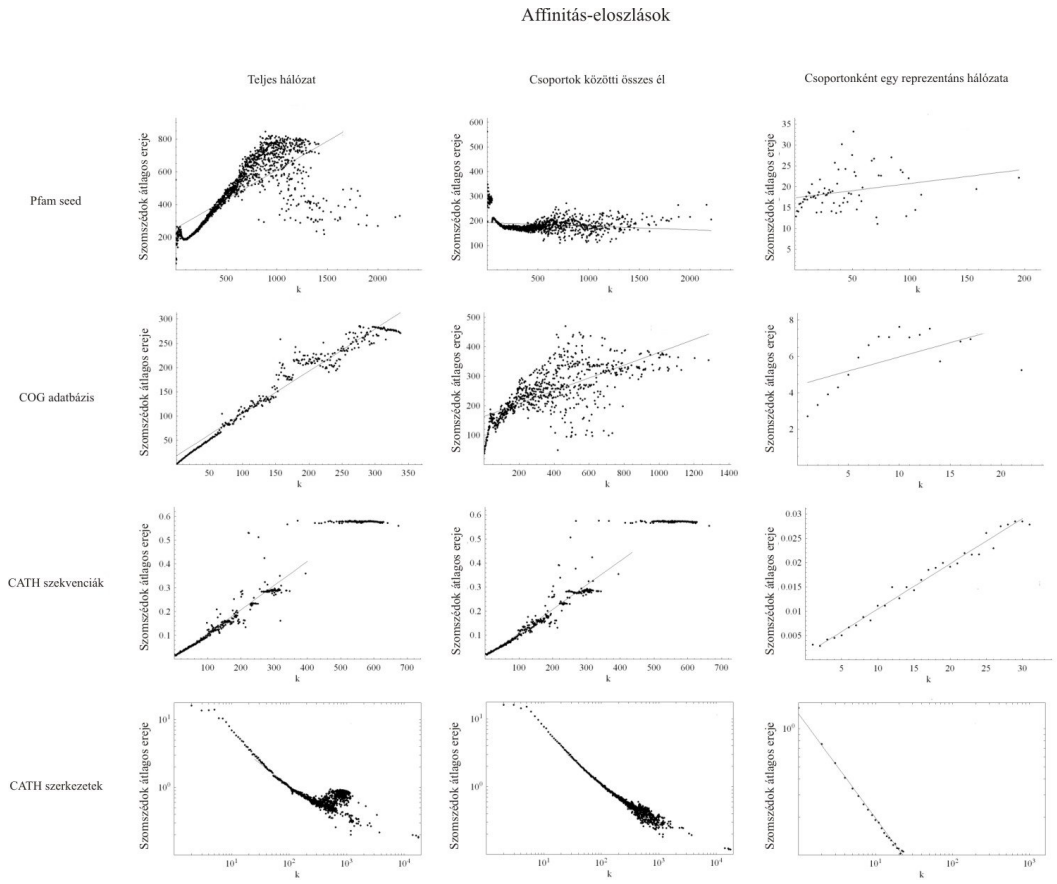
keresés hatásfokát úgy javítja meg, hogy a hálózat élein át (csoporton belüli és kívüli éleken) propagáltatja a súlyokat, hasonlóan a Google által alkalmazott PageRank algoritmushoz. Ez az eljárás javítja a keresés hatékonyságát, ami indirekt bizonyíték arra, hogy a csoportok közötti élek is hordoznak hasznos információt. (Például: ha egy csoport tagjai nagyrészt ugyanazokkal a csoporton kívüli elemekkel vannak összekötve, akkor a csoporttagok a „propagáció" során segíteni tudják egymást, hogy feljebb kerüljenek a keresés eredményeként megjelenő találati rangsorban. Ez pedig valójában csak akkor lehetséges, ha a csoportból kifele mutató élek nem véletlenszerűek, hanem többé-kevésbé egymással összekötött elemek egy adott csoportjára mutatnak).

A hálózatok hierarchikus jellege

A különböző paraméterek (elsősorban a fokszám) eloszlásából szoktak a hálózatok asszortatív-dizasszortatív voltára következtetni. Amennyiben egy hálózatban a magas fokszámú tagok preferálják egymást, pl. a szomszédok fokszáma nő az adott csúcs fokszámának függvényében, akkor a hálózatot asszortatívnak szokták nevezni. Dizasszortatív, vagyis széttagolt hálózatról pedig akkor beszélünk, ha a szomszédok fokszám-eloszlása csökkenő tendenciát mutat a saját fokszám függvényében. Ilyenkor a csoportokon belüli kapcsolatok dominálnak. Súlyozott hálózatoknál ezt a vizsgálatot affinitás-vizsgálatnak szokták nevezni, azt szokás megvizsgálni, hogy egy adott fokszámú csúcs szomszédai milyen súlyúak. Az értékelés módja ugyanaz, vagyis a növekedő tendencia a hierarchikus felépítés jele, míg a csökkenő tendencia a széttagolt hálózatot jellemzi.

A 16. ábra a hálózatok asszortativitási vizsgálatának eredményeit foglalja össze. A teljes hálózatok összképét tekintve, a szekvencia-hálózatoknál (PFAM-seed, CATH-szekvencia, COG) általában asszortativitást tapasztalunk, azaz az illesztett egyenes meredeksége pozitív. Érdekes megjegyezni, hogy a teljes hálózatok ábráinak meredeksége (baloldali oszlop) mindig nagyobb, mint a csoportközi élek meredeksége (középső oszlop), tehát úgy tűnik, hogy az asszortativitási tulajdonság főként a csoportokon belüli éleknek köszönhető.

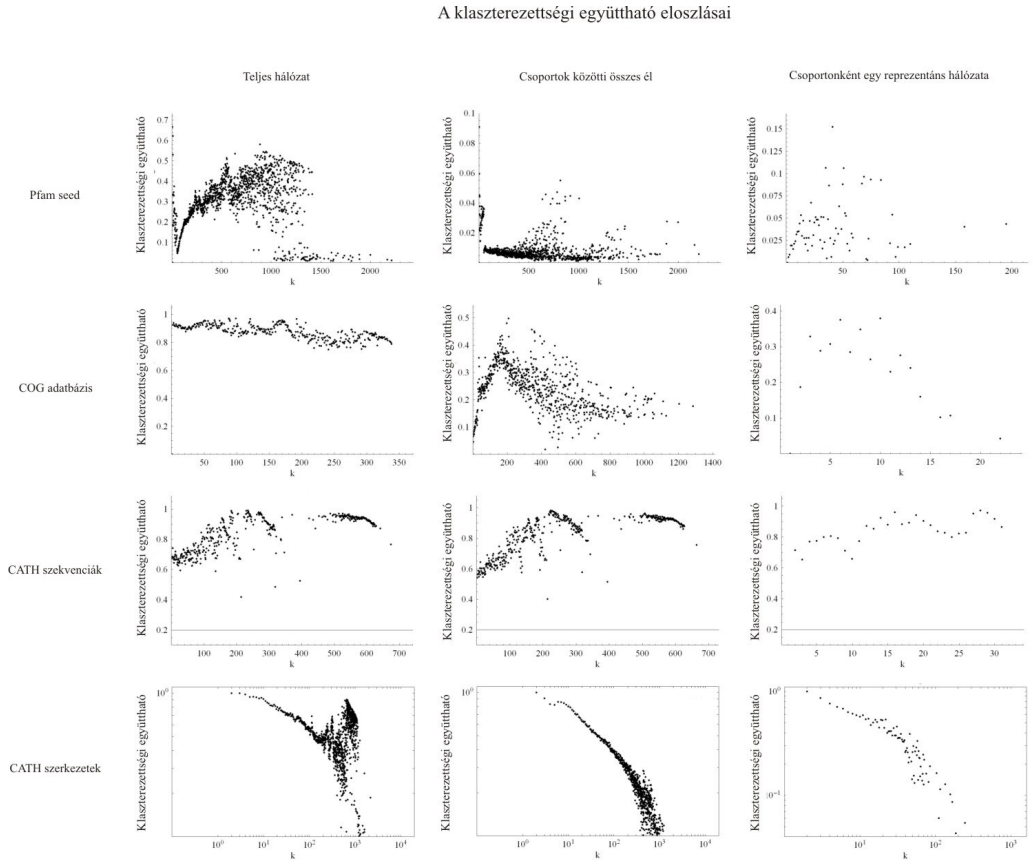
Az ábrán egyetlen kivétel van az asszortativitási tendencia tekintetében, ez a CATH adatbázis szerkezeti hasonlóságainak ábrája. Ez jellemzően csökkenő tendenciát mutat, tehát a szerkezeti hasonlóságok hálózata dizasszortatív, miközben ugyanezen fehérjék szekvenciális hasonlóságai asszortatív tendenciát mutatnak (pozitív meredekség). A jelenségre pillanatnyilag nehéz konzisztens magyarázatot találni, lehetséges, hogy az eltérés oka az alapvetően szerkezeti elveken alapuló CATH csoportbeosztás.



16. ábra Az affinitás eloszlásának vizsgálata különböző adatbázisokon
A szekvencia-hasonlóságok eloszlásainak (fölső három sor) némileg emelkedő tendenciája asszortativitásra utal. Ezzel markánsan ellentétes a szerkezeti hasonlóságok affinitásának eloszlása (alsó sor) amelyik dizasszortatív tulajdonságot mutat.

Ugyanezt a különbséget tükrözi a klaszterezettségi együttható vizsgálata (17. ábra). Itt a negatív meredekség a hierarchikus elrendezés jele. Az egyetlen hálózat, amely esetünkben jellemzően negatív tendenciát mutat, a szerkezeti hasonlóságokon alapuló CATH hálózat. Mivel a CATH adatbázis szerkezeti hierarchián alapul, nagyon valószínűnek tűnik, hogy ez az osztályozási elv mutatkozik meg a szerkezeti hasonlóságok hierarchikus voltában is. Ugyanezeknek a fehérjéknek a szekvencia-hasonlóságai viszont nem hierarchikusak. Közismert, hogy a szerkezeti alapon definiált csoportok között nem lehet szignifikáns szekvencia-hasonlóságokat felfedezni, ezért nem

meglepő, hogy a CATH csoportok szekvencia-eloszlásai nem tükrözik a hierarchikus szerkezet jeleit. Ebből a szempontból érdekes, hogy a szekvencia-hasonlósági hálózatok egyike sem tükrözi a hierarchikus elrendezést. Bár ezt a jelenséget nem lehet kimerítően megmagyarázni a jelen adatok alapján, úgy tűnik, hogy a hierarchikus szerkezeti osztályozás és a szerkezeti hasonlóságokban megnyilvánuló hierarchikus tendencia együttes előfordulása nem lehet véletlen.

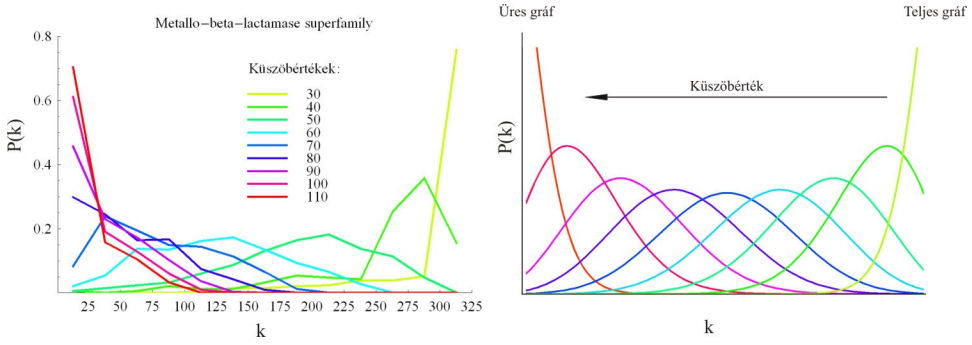


17. ábra A klaszterezettségi együttható eloszlásának vizsgálata különböző adatbázisokon

Ebben az ábrázolásban egyedül a szerkezeti hasonlóságok eloszlása (alsó sor) mutat határozott tendenciát, a csökkenő tendencia hierarchikus hálózatra utal. A többi panelen bemutatott szekvencia-hasonlóságok tendenciái nagyon gyengék (és rendszertelenek), de jellegükben – amennyiben ezt egyáltalán érdemes értékelni – ezzel inkább ellentétesek.

A csoportokon belüli fokszámeloszlások

Az általam vizsgált adatbázisok mindegyike több száz csoportból áll, melyek erősen különböznek a csoportok tagszámának és a közöttük lévő hasonlóságok erősségének tekintetében. Ezek egyenkénti vizsgálata túlmenne a dolgozat keretein, itt csak néhány kvalitatív tendenciára hívnám fel a figyelmet.



18. ábra Csoporton belüli eloszlások

Egyes csoportokon belüli fokszámeloszlások tipikus példája. Bal: A Metallo-béta-laktamáz szekvencia-csoport fokszámeloszlásának változása a BLAST score küszöb függvényében. Jobb: Különböző paraméterű binomiális eloszlások sematikus ábrázolása.

Először is, a teljes adatbázis fokszámainak eloszlásával ellentétben a csoportokon belüli fokszámok eloszlása általában maximumos eloszlást követ. Ez a maximumos eloszlás azonban csak bizonyos küszöb-tartományban jelenik meg. A 18. ábra egy tipikus helyzetet mutat be, itt egy adott szekvencia-csoport fokszámeloszlását ábrázoltam a hasonlósági küszöb függvényében. Az ábra jobb oldalán a változás kvalitatív értelmezését ábrázoltam. Amennyiben nincs küszöb, minden csúcs össze van kötve egymással, így a fokszámeloszlás egyetlen értékre korlátozódik. Amennyiben a küszöb értékét nagyon nagyra választjuk (nagyobbra, mint a fehérjék között lehetséges

hasonlóság mérőszámának értéke), akkor viszont minden fokszám nulla lesz. A két szélsőérték között jelenik meg a maximumos eloszlás.

Itt ismét emlékeztetnék az előzőekben már említett modellre, melyben adott egy teljes súlyozott gráf, melyben a súlyok nagysága akármilyen eloszlást követhet, és a súlyok véletlenszerűen vannak kiosztva ez élek között. Ekkor akárhogy is választunk a súlyok minimuma és maximuma között egy küszöb értéket, aminél kisebb súlyokkal rendelkező éleket töröljük, a megmaradó gráf fokszámeloszlása binomiális lesz. Ennek az eloszlásnak az n,p paraméterei közül *n* állandó, és a gráf csúcsszámával azonos, *p* pedig 0 és 1 között változik, ahogy a küszöb értéke a minimumtól a maximum felé halad. Ez a változás kvalitatíve leírja, ahogyan ez az átmenet a két ábrán megjelenik. Azoknál a csoportoknál kapunk ilyen vagy ehhez hasonló képet, melyekben elég nagyok és elég erősek az elemek közötti hasonlóságok ahhoz, hogy még magas küszöbök mellett is kivehető legyen az eloszlás. Hiba lenne ebből azt a következtetést levonni, hogy minden csoporton belül, (vagy akár csak egyetlen csoporton belül) a hasonlóságok véletlenszerűek, rendezetlenül helyezkednek el. Azonban ennek a feltételezésnek nem mond ellent a fokszámeloszlás viselkedése. A csoportok száma túl nagy ahhoz, hogy mindet teljes alapossággal megvizsgáljuk.

Az eredmények diszkussziója

A fehérjehasonlóságok hálózatai alapvetően fontosak mind a gyakorlat (fehérjék osztályozása, genomok annotációja), mind pedig az elméleti kutatások (fehérje-evolúció) számára. A hálózatkutatás egyik alapvető eredménye az, hogy a természetben talált hálózatok skálamentes eloszlást követnek, és ehhez kapcsolódik a preferenciális kötődés modellje, vagyis egy konkrét folyamat-leírás, amely révén az ilyen hálózatok létrejöhetnek.

A preferenciális kötődés modellje a fehérjék hasonlósági hálózatainak esetében különösen logikusnak látszik. Ha ugyanis feltételezzük, hogy minden (fehérjét kódoló) gén egyforma valószínűséggel duplikálódhat, akkor világos, hogy a már eleve nagy kópiaszámban jelenlévő géneknek több duplikált utódja lesz, így elvárható, hogy a hasonlósági hálózat valóban a preferenciális kötődés alapján fog fejlődni, azaz fokszámeloszlása skálamentes lesz. Ezt az elvárást többféle hasonlósági hálózaton is teszteltem, és úgy találtam, hogy csak tendenciaszerűen érvényesül, tehát az ábrázolásnak vannak többé-kevésbé lineáris szakaszai, de az egész eloszlás nem nevezhető tisztán skálamentesnek, amennyire azt például az Internet esetében kimutatták. Leginkább tehát az az interpretáció tűnik jogosnak, hogy a hasonlósági hálózatok, és általában a biológiai hálózatok a „hálózati tér” egy olyan régiójába esnek, amelyhez a skálamentes eloszlás is tartozik, de azzal nem azonosak. Dolgozatom beadásával szinte egyidőben jelent meg az evolúció-kutató Evelyn Fox Keller tanulmánya (89) a Nature-ben, melyben így ír: „A növekedés és a preferenciális kötődés csak egyike a sokféle módnak, ahogy az ilyen eloszlások létrejöhetnek, viszonylagos gyenge teljesítménye miatt azonban nem

valószínű, hogy az evolúció éppen ezt eredményezte volna.”

A hálózatok modularitásával, hierarchikus voltával kapcsolatos megállapításokat is próbáltam tesztelni. A hierarchikus szervezettségű szerkezeti adatbázisok esetén meg is jelent a hierarchikus hálózatokra jellemző eloszlásforma, ha a szerkezeti hasonlóságokat használtam fel az analízisben. Ha ugyanezen fehérjék szekvencia-hasonlóságait analizáltam, a hierarchikusság nem volt észlelhető. Ez azonban magyarázható a mintavétellel, hiszen a szerkezeti adatbázisokat eleve úgy építik fel, hogy a minták között nincs nagy szekvencia-hasonlóság (a hasonló szekvenciájú fehérjék szerkezetét legtöbbször nem is határozzák meg fizikai méréssel, hanem csak modellezéssel). Annyi következtetés tehát levonható, hogy az eloszlások vizsgálatánál nyert képet mind a mintavétel, mind pedig az alkalmazott hasonlósági mérőszám befolyásolja.

Mindezek alapján tehát úgy tűnik, hogy fehérje-hasonlósági hálózatoknál bár bizonyos fokig észlelhető a skálamentes fokszámeloszlás és a hierarchikusság tendenciája, a technikai problémák, pl. a mintavétel azt elfedhetik. Tehát még ebben az egy, viszonylag egyszerű esetben sem mondható ki, hogy a tendenciák kétséget kizáróan érvényesülnének.

Felmerül tehát a kérdés, igazak-e a skálamentes eloszlás illetve a hierarchikusság elvei a biológiai hálózatoknál. Fox Keller erről így ír: „Jó ötlet-e hogy a biológiai rendszereket mindent magukba foglaló törvényekkel próbáljuk leírni? ... A kémia és fizika törvényei természetesen alapvetőek. De ezeken túlmenően, a biológiai általánosításoknak mindig

feltételesnek kell lenniük, az evolúció és a történeti kötöttségek miatt, amelyektől az élet megjelenése és fejlődése alapvetően függ.” Vagyis az alapvető összefüggések túlzott általánosításától óvakodni kell. Ennek megfelelően tehát a hálózati modellek alaptulajdonságai, tendenciái nagyon fontos felismeréseket szolgáltathatnak, de nem szabad részletekbemenően kiterjeszteni őket.

Felhasznált programok

A felhasznált paraméterek leírásai a bevezetőben megtalálhatóak, kiszámításukhoz a boost progkönyvtár 1.32.0 verziójának objektumait használtam fel, az adatok feldolgozása és ábrázolása pedig a Mathematica 5.0 verziójú programmal történt.

Hálózati eredmények áttekintése

Dolgozatomban a hálózati modellek alkalmazásával foglalkoztam három biológiai probléma, az oxidatív folding, a fehérjehasonlóságok és a többpontos támadással szembeni stabilitás vizsgálatában. A hálózati modellek robbanásszerű elterjedése új jelenség, mely saját vizsgálataim időszakával esik egybe. Az igen nagyszámú hálózatos publikáció áttekintésére ezért nem vállalkozhattam, de mégis fontosnak tartom, hogy legalább megkíséreljem áttekinteni, milyen kérdésekre adhatnak, illetve nem adhatnak választ a hálózati modellek, és ezeken belül milyen jellegűek voltak saját vizsgálataim.

Az irodalom áttekintése és itteni eredményeim is azt sugallják, hogy a jelenleg rendelkezésre álló adatbázisok már akkora méretet és olyan komplexitást értek el, hogy az elszigetelt elemek, egyes interakciók vizsgálata már sokszor nem tekinthető át, és sokan keresik azokat a módszereket, amelyekkel legalább a már megismert elemek együttes viselkedését lehet vizsgálni. Ennek az egyre erősödő igénynek felelnek meg a hálózati modellek, melyek valójában a legegyszerűbb leírásai egy többszereplős komplex vizsgálati rendszernek. Ezek vizsgálatától a kutatók olyan információkat remélnek, melyek a hálózatot alkotó individuális objektumoknak, illetve az azok közötti egyedi kölcsönhatásoknak a vizsgálatával nem érhetők el.

A hálózati rendszerek leírásának fogalomtára ma még nem túlságosan nagy, és egyaránt merít a matematika, a technikai- és a biológiai rendszerek fogalmi készleteiből. A legjellemzőbb fogalmak például a hálózati topológiák (skálamentes, rácsszerű,

csillagpontos, “kis-világ”, hierarchikus és elosztott hálózatok stb.), a hálózaton belüli szerepek (centrum, periféria, hub), a stabilitás (támadhatóság, zavartűrés, robosztusság) köréből kerülnek ki. Az egyik nehézség abból adódik, hogy a komplex rendszerek leírására nincsenek hatékony fogalmaink.

A rendszerbiológia (systems biology) fogalmának megjelenése is mutatja, hogy a biológiai entitások komplex rendszerként való vizsgálata mára egyértelműen a tudományos érdeklődés előterébe került. Ezen belül kétféle irányvonalat is megkülönböztethetünk, az egyik a “bottom-up” megközelítés, a sokféle adat kombinációjából létrejövő komplexitás vizsgálata, a másik a “top-down” megközelítés, melynél először egy komplexen viselkedő absztrakt rendszert próbálunk építeni, majd azt igyekszünk hozzákötni a mérhető adatokhoz.

Hálózati modelleket mindkét megközelítésnél alkalmazhatunk. A bottom-up segítségével a meglévő adatok és az azok közötti viszonyok (kölcsönhatások, átmenetek) tekinthetőek át. Az oxidatív folding esetében egy állapot-hálózat elemeit és a közöttük lévő kölcsönhatásokat állítottam elő. A fehérjehálózatok is bottom-up vizsgálatok körébe tartoznak, ott a hálózat elemei a fehérjék, kapcsolataik a hasonlósági viszonyok. Vizsgálataim itt a topológiára vonatkoztak, és kimutattam, hogy hasonló ábrázolásmódok hasonló topológiát eredményeznek, ugyanakkor egy adattömeg kétfajta ábrázolása különböző struktúrát eredményezhet.

A többpontos támadhatóságra vonatkozó modellvizsgálataim – véleményem szerint – mind a bottm-up, mind a top-down vizsgálatok elemeit ötvözik. Ugyanis az eredeti kérdésfeltevés úgy hangzott, tudjuk-e modellezni a több ponton ható gyógyszerek hatását. Ehhez először egy biológiai szempontból reális, minimális rendszert kellett keresni (ezek voltak a regulációs hálózatok), majd ezek kontextusában alakítottam ki a támadási stratégiákat, amelyekkel az egyes gyógyszerhatóanyagok viselkedését igyekeztem modellezni. Ez a vizsgálat tehát bár létező biológiai hálózatokat használt fel, mégis lényegében a felülről lefele tervezett vizsgálatok elveire hasonlít.

Izgalmas, és dolgozatom konkrét célján messze túlmutat az a kérdés, hogy valójában mennyire jogos, és milyen jellegű következtetésére jogosíthat fel bennünket a hálózati rendszerek alkalmazása. Mindenekelőtt azt kell figyelembe vennünk, hogy a hálózati leírások minimalisták, például két elem között legtöbbször egyetlen (irányított vagy irányítatlan) kapcsolatot tételezünk fel. Sok esetben a hálózat elemeit kvalitatíve sem különböztetjük meg (pl. a topológiai statisztikák esetében). Mindez olyan fokú absztrakció, ami a konkrét alkalmazások realitását komolyan megkérdőjelezi. Ezzel a korláttal akkor szembesültem, mikor meg kellett kérdeznem, mennyire várható el, hogy a többpontos támadhatóság elve a gyakorlatban is érvényesüljön. Erre a kérdésre úgy kíséreltem meg válaszolni, hogy a ma ismert főbb – az ismert biológiai hálózatok valamennyi példáját felölelő – topológia-típusra egyaránt elvégeztem a vizsgálatot, és mivel minden esetben ugyanazt a választ kaptam, valószínűsíthető, hogy a jelenség a gyakorlatban is megfigyelhető, de ezt bizonyossággal állítani csak kísérletes eredmények alapján lehet.

Összefoglalás

A hálózati modellek megjelenése napjaink egyik meghatározó tudományos eseménye, melynek révén megismerhetjük különböző technikai, számítógépes és biológiai rendszerek kölcsönhatási viszonyainak közös vonásait. Ez a folyamat elsősorban azért jelentős, mert általa új alapokon tudjuk rendszerezni a különböző szakterületek rokon-jelenségeit, másrészt nagymértékben fel is gyorsította a szakterületek közötti információcserét. A hálózati modellek topológiája, a hálózatokon átterjedő jel- és anyagáramok ugyanis sok közös tulajdonsággal rendelkeznek, a hálózati rendszerek topológiai és fluxus-plaszticitása magyarázatul szolgálhat a biológiai rendszerek stabilitásának megértéséhez.

Dolgozatomban a hálózati modellek biológiai alkalmazását tűztem ki célul, három területen.

1) Vizsgáltam a biológiai (gén-regulációs) hálózatok stabilitását többpontos támadással szemben. Vizsgálataim során tudtommal elsőként hasonlítottam össze csúcsok és élek eltávolításainak hatását. Számítógépes modelleket dolgoztam ki a hálózatok pontjainak (génjeinek) és a közöttük lévő kapcsolatoknak a kiiktatására illetve részleges gátlására. Megállapítottam, hogy a rendszer több ponton való gyenge perturbációja révén ugyanaz a hatás könnyebben érhető el, mint a hálózat egy elemének (például egy központi génjének) kiiktatásával. Ez az eredmény felhívja a figyelmet arra, hogy a többhatású, több célponton ható gyógyszerek,

koktélterápiák hatásosabbak lehetnek, mint a mai racionális gyógyszertervezés által fejlesztett nagyspecifitású, egy adott génre vagy fehérjére ható szerek.

2) Hálózati modellt állítottam fel a fehérjék diszulfid hídjainak keletkezésére, az oxidatív folding folyamatának ábrázolására, ez egy szabályos (rácsszerű) hálózatot eredményezett. Az oxidatív folding folyamata nagyszámú, kísérletesen is tanulmányozható intermedieren keresztül zajlik, az általam kifejlesztett hálózati modell révén az intermedierek és a lehetséges átalakulások hálózati térképe megrajzolható, és a kísérletesen meghatározott intermedierek révén megjeleníthetőek rajta a folding-útvonalak. Amennyiben a kísérletesen meghatározott folding-útvonalak nem összefüggőek, az felhívja a figyelmet arra, hogy egyes intermediereket nem sikerült izolálni.

3) Elvégeztem a fehérjék (szekvenciák, térszerkezetek) hasonlósági viszonyait megjelenítő hasonlósági hálózatok topológiájának vizsgálatát. Ezek a hálózatok viszonylag magas klaszterezettséggel bíró ún. kis-világ hálózatok, amelyekben a hasonlósági csoportok sűrűbben összekötött együttesekként jelennek meg. Több paraméter tekintetében nagy hasonlóságot mutatnak egymással azok a hálózatok, amelyek szekvenciális hasonlóságok alapján épülnek fel, ugyanakkor a fehérjék szerkezeti adatbázisainak hierarchikus volta megnyilvánul a szerkezeti hasonlóságok topológiájában, de ugyanez nem mutatkozik meg a szekvenciák

közötti hasonlóságokban. A vizsgált hasonlósági hálózatok fokszám-eloszlása közel áll a skálamentes modelléhez, de azzal nem egyezik meg.

Ezek alapján az eredmények alapján megállapítható, hogy egyrészt a biológiai hálózatok méretük alapján alkalmasak arra, hogy a más területek vizsgálatára kifejlesztett módszerek alkalmazhatóak legyenek, másrészt a biológia számára a kapott válaszok értelmezhetőek, és hasznosak lehetnek.

Summary

The success of network models is one of the most conspicuous scientific phenomena of the past few years. Importantly, network models can help us to recognize the common features of technical, computational as well as biological systems as they highlight the similar features in and the deep analogies between different fields. As a result they help to speed up sharing information between fields. For instance, network topologies, propagation of signals and material flows have many interesting common characteristics in vastly different fields, and it is widely believed that topological as well as flux-plasticity of network models may explain the robustness of biological systems.

My work included the application of network models to three biological applications.

1) I examined the stability of biological (gene regulation) networks against multiple attacks. I worked out computational models for simulating the elimination or partial inactivation of network nodes (genes) and/or the contacts between them. It was found that a few weak multiple perturbations can cause as much damage to a network as the elimination of a central object. This result draws attention to the possibilities of multitarget drug strategies as opposed to rationally designed, high specificity drugs.

2) I designed a network model to visualize the states corresponding to disulfide intermediates that can form during the oxidative folding of proteins. The resulting

networks are regular (grid-like) networks, and experimentally studied intermediates can be mapped onto then whereby the folding pathways appear as continuous paths connecting the reduced and the native state.

3) Similarity networks of protein sequences and structures were analyzed from a topological point of view. These networks have relatively high clustering coefficient characteristic of the called small world networks. In this kind of networks the similarity groups of proteins (such as domain types of protein families) appears as highly connected clusters. The topologies were found both database and method dependent. For instance, similarity networks built of the 3D structural similarities between members of protein domain data showed a hierarchical structure. On the other hand, the sequence similarity networks of the same data did not appear hierarchical. The degree distribution of analyzed similarity networks is reminiscent of the scale-free distribution but not identical with it.

So we can conclude that on one hand, biological networks are suitable subjects for the analytical approaches developed in other fields, and on the other hand, the answers can be relevant and useful for biologists.

Hivatkozásjegyzék

1. Karagiannis, T., Molle, M. and Faloutsos, M. (2004) *Ieee Internet Computing*, **8,** 57-64.
2. Yook, S.H., Jeong, H. and Barabasi, A.L. (2002) *Proc Natl Acad Sci U S A*, **99,** 13382-13386.
3. Euler, L. (1736) *Opera Omnia*, **7,** 128-140.
4. Bollobás, B. (1998) *Modern Graph Theory*. Springer-Verlag, New York.
5. Bollobás, B. (2001) *Random Graphs*. 2 ed. Cambridge University Press, Cambridge.
6. Andrásfai, B. (1997) *Gráfelmélet*. Polygon, Szeged.
7. Hajnal, P. (1997) *Gráfelmélet*. Polygon, Szeged.
8. Cormen, T.H., Leiserson, C.E. and Rivest, R.L. (1997) *Algoritmusok*. Mûszaki könyvkiadó, Budapest.
9. Barthelemy, M., Barrat, A., Pastor-Satorras, R. and Vespignani, A. (2005) *Physica a-Statistical Mechanics and Its Applications*, **346,** 34-43.
10. Barabasi, A.L. and Oltvai, Z.N. (2004) *Nat Rev Genet*, **5,** 101-113.
11. Latora, V. and Marchiori, M. (2001) *Phys Rev Lett*, **87,** 198701.
12. Høvik, T. and Gleditsch, N.P. (1970) *Quality and Quantity*, **4,** 193-209.
13. Erdös, P. and Rényi, A. (1959) *Publicationes Mathematicae Debrecen*, **6,** 290-297.
14. Erdös, P. and Rényi, A. (1960) *Publ. Math. Inst. Hungar. Acad. Sci.*, **5,** 17-61.
15. Milgram, S. (1967) *Phsychology Today***,** 60-67.
16. Wasserman, S., Faust, K. and Iacobucci, D. (1994) *Social Network Analysis*. Cambridge University Press, New York.
17. Watts, D.J. and Strogatz, S.H. (1998) *Nature*, **393,** 440-442.
18. Barabasi, A.L. and Albert, R. (1999) *Science*, **286,** 509-512.
19. Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) *Nature*, **411,** 41-42.
20. Dobson, C.M. and Karplus, M. (1999) *Curr Opin Struct Biol*, **9,** 92-101.
21. Dinner, A.R., Sali, A., Smith, L.J., Dobson, C.M. and Karplus, M. (2000) *Trends Biochem Sci*, **25,** 331-339.
22. Pain, R.H. (2000) *Mechanisms of Protein Folding*. 2nd ed. Oxford University Press, Oxford, New York.
23. Onuchic, J.N., Socci, N.D., Luthey-Schulten, Z. and Wolynes, P.G. (1996) *Fold Des*, **1,** 441-450.
24. Levinthal, C. (1968) *J. Chim. Phys.*, **65,** 44-45.
25. Tu, B.P. and Weissman, J.S. (2004) *J Cell Biol*, **164,** 341-346.
26. Chang, J.Y. (2004) *Biochemistry*, **43,** 4522-4529.
27. Wedemeyer, W.J., Welker, E. and Scheraga, H.A. (2002) *Biochemistry*, **41,** 14637-14644.
28. Welker, E., Wedemeyer, W.J., Narayan, M. and Scheraga, H.A. (2001) *Biochemistry*, **40,** 9059-9064.

29. Vishveshwara, S., Brinda, K.V. and Kannan, N. (2002) *Journal of Theoretical and Computational Chemistry*, **1,** 187-211.
30. Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22,** 2577-2637.
31. Plaxco, K.W., Simons, K.T. and Baker, D. (1998) *J Mol Biol*, **277,** 985-994.
32. Magyar, C., Tudos, E. and Simon, I. (2004) *FEBS Lett*, **567,** 239-242.
33. Selvaraj, S. and Gromiha, M.M. (2004) *Proteins*, **55,** 1023-1035.
34. Vendruscolo, M., Paci, E., Dobson, C.M. and Karplus, M. (2001) *Nature*, **409,** 641-645.
35. Vendruscolo, M., Paci, E., Karplus, M. and Dobson, C.M. (2003) *Proc Natl Acad Sci U S A*, **100,** 14817-14821.
36. Albert, R., Jeong, H. and Barabasi, A.L. (2000) *Nature*, **406,** 378-382.
37. Scala, A., Amaral, L.A.N. and Barthelemy, M. (2001) *Europhysics Letters*, **55,** 594-600.
38. Dokholyan, N.V., Shakhnovich, B. and Shakhnovich, E.I. (2002) *Proc Natl Acad Sci U S A*, **99,** 14132-14136.
39. Goldmeier, E. (1972) *Similarity in visually percieved forms*. 1 ed. International Universities Press, Inc., New Yorik, N.Y.
40. Johnson, M.A. (1989) *Journal of Mathematical Chemistry*, **3,** 117-145.
41. Koonin, E.V. and Galperin, M.Y. (2003) *Sequence, evolution, function*. Kluver Academic Publishers, Boston, Dordrecht, London.
42. Higgins, D. and Taylor, W.R. (2000) *Bioinformatics, Sequence, structure, and databanks*. Oxford University Press, Oxford, New York.
43. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J Mol Biol*, **215,** 403-410.
44. Pearson, W.R. and Lipman, D.J. (1988) *Proc Natl Acad Sci U S A*, **85,** 2444-2448.
45. Needleman, S.B. and Wunsch, C.D. (1970) *J Mol Biol*, **48,** 443-453.
46. Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.*, **147,** 195-197.
47. Krause, A. and Vingron, M. (1998) *Bioinformatics*, **14,** 430-438.
48. Krause, A., Stoye, J. and Vingron, M. (2000) *Nucleic Acids Res*, **28,** 270-272.
49. Attwood, T.K. (2000) *Brief Bioinform*, **1,** 45-59.
50. Murvai, J., Vlahovicek, K., Barta, E. and Pongor, S. (2001) *Nucleic Acids Res*, **29,** 58-60.
51. Murvai, J., Vlahovicek, K. and Pongor, S. (2001) In Pifat-Mrzljak, G. (ed.), *Supramolecular Structure and Function*. New York, Plenum Press, pp. 155-166.
52. Vlahovicek, K., Carugo, O., Murvai, J. and Pongor, S. (2001) In Gromiha, M. and Selvaraj, S. (eds.), *Recent Research Developments in Protein Folding Stability and Design*. Research Signpost, Trivandrum, India, pp. 141-150.
53. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. (2001) *Nucleic Acids Res*, **29,** 22-28.
54. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucleic Acids Res*, **25,** 3389-3402.
55. Schaffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L. and Altschul, S.F. (1999) *Bioinformatics*, **15,** 1000-1011.
56. Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) *Nucleic Acids Res*, **29,** 2994-3005.

57. Pongor, S. (1987) *Methods Enzymol*, **154,** 450-473.
58. Carugo, O. and Pongor, S. (2002) *J Mol Biol*, **315,** 887-898.
59. Røgen, P. and Fain, B. (2003) *Proc Natl Acad Sci U S A*, **100,** 119-124.
60. Wolf, Y.I., Rogozin, I.B., Grishin, N.V. and Koonin, E.V. (2002) *Trends Genet*, **18,** 472-479.
61. Mirkin, B.G., Fenner, T.I., Galperin, M.Y. and Koonin, E.V. (2003) *BMC Evol Biol*, **3,** 2.
62. Gerstein, M. and Hegyi, H. (1998) *FEMS Microbiol Rev*, **22,** 277-304.
63. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) *Nature*, **407,** 651-654.
64. Overbeek, R., Larsen, N., Pusch, G.D., D'Souza, M., Selkov, E., Jr., Kyrpides, N., Fonstein, M., Maltsev, N. and Selkov, E. (2000) *Nucleic Acids Res*, **28,** 123-125.
65. Podani, J., Oltvai, Z.N., Jeong, H., Tombor, B., Barabasi, A.L. and Szathmary, E. (2001) *Nat Genet*, **29,** 54-56.
66. Tekaia, F., Lazcano, A. and Dujon, B. (1999) *Genome Res*, **9,** 550-557.
67. Snel, B., Bork, P. and Huynen, M.A. (1999) *Nat Genet*, **21,** 108-110.
68. Korbel, J.O., Snel, B., Huynen, M.A. and Bork, P. (2002) *Trends Genet*, **18,** 158-162.
69. Snel, B., Bork, P. and Huynen, M. (2000) *Trends Genet*, **16,** 9-11.
70. Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000) *Curr Opin Struct Biol*, **10,** 366-370.
71. Huynen, M.A. and Snel, B. (2000) *Adv Protein Chem*, **54,** 345-379.
72. Lathe, W.C., 3rd, Snel, B. and Bork, P. (2000) *Trends Biochem Sci*, **25,** 474-479.
73. Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A. and Koonin, E.V. (2002) *Nucleic Acids Res*, **30,** 2212-2223.
74. Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002) *Nat Genet*, **31,** 64-68.
75. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) *Science*, **298,** 824-827.
76. Creighton, T.E. (1992) *Science*, **256,** 111-114.
77. Weissman, J.S. and Kim, P.S. (1991) *Science*, **253,** 1386-1393.
78. Hober, S., Uhlen, M. and Nilsson, B. (1997) *Biochemistry*, **36,** 4616-4622.
79. Milner, S.J., Carver, J.A., Ballard, F.J. and Francis, G.L. (1999) *Biotechnol Bioeng*, **62,** 693-703.
80. Yang, Y., Wu, J. and Watson, J.T. (1999) *J Biol Chem*, **274,** 37598-37604.
81. Chang, J.Y., Li, L. and Lai, P.H. (2001) *J Biol Chem*, **276,** 4845-4852.
82. Page, L. (2001) USA patent no. 09/895,174.
83. Weston, J., Elisseeff, A., Zhou, D., Leslie, C.S. and Noble, W.S. (2004) *Proc Natl Acad Sci U S A*, **101,** 6559-6563.
84. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L. *et al.* (2004) *Nucleic Acids Res*, **32 Database issue,** D138-141.
85. Vlahovicek, K., Murvai, J., Barta, E. and Pongor, S. (2002) *Nucleic Acids Res*, **30,** 273-275.
86. Pearl, F.M., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J. and Orengo, C.A. (2003) *Nucleic Acids Res*, **31,** 452-455.

87. Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2004) *Nucleic Acids Res*, **32,** D226-229.
88. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N. *et al.* (2003) *BMC Bioinformatics*, **4,** 41.
89. Keller, E.F. (2007) *Nature*, **445,** 603.

Köszönetnyilvánítás

Ez a dolgozat a Szegedi Tudományegyetem Biológia Doktori Iskola keretei között készült, 2002 és 2006 között. Ösztöndíjamat az MTA Szegedi Biológiai Központ biztosította. Témavezetőm Pongor Sándor, a biológiai tudományok doktora volt.

Szeretnék köszönetet mondani a Doktori Iskola oktatóinak, elsősorban Boross Imre és Maróy Péter egyetemi tanároknak, hogy irányításukkal megismerkedhettem a molekuláris biológia és a genetika néhány fejezetével.

Köszönöm az MTA Szegedi Biológiai Központ Bioinformatikai csoport munkatársainak, hogy munkámhoz szakmai támogatást és segítő munkahelyi légkört biztosítottak.

Köszönöm az International Centre for Genetic Engineering and Biotechnology-nak, hogy munkámat egy három hónapos „sandwich Ph.D.” ösztöndíjjal támogatta.

Szeretnék köszönetet mondani Csermely Péternek amiért rámutatott gyógyszerfejlesztési alkalmazásokra, és az együttműködési lehetőségért.

Végül köszönöm témavezetőmnek a témaválasztásban és a téma kidolgozásában nyújtott segítségét és irányítását.