KRÉSZ MIKLÓS – TÓTH ATTILA

Gráf alapú adatbányászat és vizualizáció: egy esettanulmány

ALKALMAZOTT TERMÉSZETTUDOMÁNYI INTÉZET

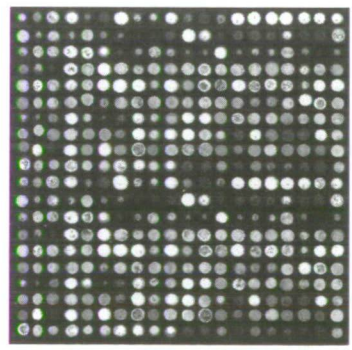
adatbányászat, adatvizualizáció, klaszterezés, genomikai kutatások, DNS-chip technológia

1. Bevezetés

Az információs rendszerek utóbbi évtizedekben végbement dinamikus fejlődésének eredményeként napjainkra olyan mennyiségű adat halmozódott fel az adatbázisok többségében, melyeknek hagyományos úton történő elemzése, analitikája igen bonyolult, az ezen elveken alapuló technológiák (relációs alapú lekérdezés, statisztikai elemzés) az információk kinyerése szempontjából gyakran hosszadalmasak, vagy pontatlan eredményt adnak. A fenti típusú problémák megoldására fejlesztették ki az 1980–90-es években azokat a mesterséges intelligencia alapú úgynevezett adatbányászati módszereket, melyek automatikus eljárások révén hatalmas méretű, több millió sorral rendelkező táblázatokból is gyorsan és igen hatékonyan nyerik ki a keresett információkat.

Az adatbányászat ([8]) eleinte az üzleti életben nyert sikeres alkalmazást, azonban a kilencvenes évektől kezdődően nyilvánvalóvá vált, hogy a technológiai fejlődésnek köszönhetően a tudományos kísérletek is olyan méretű adathalmazokat eredményeznek, melyek hatékony elemzéséhez ezen új terület módszerei sikeresen felhasználhatóak. A genomikai kutatásokban a DNS-chip technológia felfedezése új távlatokat nyitott. A DNS adatbázisokban tárolt statikus információkkal szemben, a DNS-chip kísérletek több ezer gén expressziójának dinamikus változásairól szolgáltatnak hatalmas adattömeget. Ezen adatokban rejlő információk kinyerése új kihívásokat jelent a bioinformatika számára. Ugyanakkor az információtechnológia fejlődése lehetővé tette, hogy a biológiai témájú publikációk összefoglalói nyílt adatbázisokban váljanak hozzáférhetővé. Így természetszerűleg vetődik fel a kérdés, hogy milyen módszerek segítségével lehet a kísérleti eredmények kiértékelését az összefoglalókban fellelhető információk kinyerésével segíteni.

A genomikai kutatások számára a legfontosabb információtartalmat az szolgáltatja, hogy az egyes gének és fehérjék különböző állapotokban milyen expressziós szintet mutatnak. A fentiek alapján ugyanis következtetések vonhatóak le a gének egymáshoz és az adott fehérjékhez való kapcsolatáról. Az összefoglalókban található erre vonatkozó információk alapján a gének struktúrált adathalmazba rendezhetőek, ahol az elemek közötti kapcsolódás hálózattal (gráffal) reprezentálható.



1. ábra
DNS-chip

Az elmúlt évek során az úgynevezett biológiai szövegbányászat ([2]) dinamikusan fejlődő önálló területté vált, ugyanakkor a gráfstruktúrán alapuló adatbányászat ([16]) is az érdeklődés középpontjába került. Ezen két módszertan összekapcsolásából született a genomikai kutatásokat segítő BiblioGraph Explorer elnevezésű

rendszer, melynek adatelemző és adatvizualizációs moduljának kifejlesztése során szerzett tapasztalatok bemutatásával szeretnénk bepillantást adni a gráf alapú adatbányászat alkalmazásaiba.

2. Genomikai adatok elemzése

A microarray technológia (DNS-síkmátrix, DNS-chip) megjelenése óta módunk van időben nyomon követni egyes organizmusok akár összes génjének működését, tehát megtudhatjuk, hogy a szervezet miként válaszol környezeti hatásokra, sőt össze tudjuk hasonlítani a beteg és egészséges szöveteket, a rezisztens és szenzitív növényeket. A DNS-chip lényegében egy üveglapkára integrált nagy számú oligonukleotid, cDNS, fehérje vagy gyógyszer-jellegű vegyület. (1. ábra). Az új eszköz forradalmi távlatokat nyitott a funkcionális molekuláris biológiában, lehetővé téve gyors és széleskörű elemzési módszerek kifejlesztését a genom különböző mutációinak, ismert és ismeretlen gének és fehérjék expressziós szintjének szimultán megfigyelésével.

Tehát az eddigi leíró jellegű genom-kutatások statikus információi mellett ma már a genom-működés dinamikus adatait is tudnunk kell kezelni. A bioinformatika fiatal tudománya számára óriási feladat a sokféle adat kvantitatív elemzése ([1]), ugyanis az adatokban rejlő újszerű mintázatok felismeréséhez szükséges a genomkutatás teljes adatállományának és számítástechnikai eszköztárának, valamint az irodalmi adatbankok tartalmának kezelése is. Mindehhez modern elemző-technikákra (mintázat felismerés, adatbányászat) van szükség.

A génexpressziós kísérletek eredményeként olyan komplex adathalmaz keletkezik, melyet csak időigényes irodalomkutatás és előzetes információk segítségével lehet rendszerezni és megfelelően értékelni. A biológiai kutatások egyik legidőigényesebb része ezért az irodalomkutatás, melyet szövegalapú adatbányászati technikákkal lehet felgyorsítani. Az így kinyert információ adatbányászati módszerekkel történő elemzése azonban nem csak a kutatások felgyorsítását, hanem új összefüggések feltárását is lehetővé teszi. A jelen cikk keretében ismertetett esettanulmány a széles körben elérhető MedLine adatbázisban található összefoglalókat (absztraktokat) használta fel az információkinyerésre. A MedLine összefoglalók szövegbányászati feldolgozásával kapcsolatban a projekt során megvalósított fejlesztéseket a [4] és [6] közlemények ismertetik. A továbbiakban bemutatjuk a fenti módon előállt adatok közötti összefüggések feltárását támogató gráf alapú adatbányászati és vizualizációs módszerek segítségével elért eredményeinket.

3. Gráfstruktúra építése

Az adatbányászati eljárások megvalósíthatósága és hatékonysága nagymértékben függ az elemzendő adatok és a köztük fennálló kapcsolatok tárolására szolgáló úgynevezett analitikus infrastuktúra felépítésétől. A gyakorlatban felmerülő problémák kapcsán számos esetben megfigyelhető, hogy a feladathoz kapcsolódó adatok mint csúcspontok egy olyan komplex hálózatot alkotnak, melyben az élek az elemzési szempontok által specifikált relációkat reprezentálják. A fenti komplex hálózatoknak megfelelő analitikus infrastruktúra formális leírására olyan gráfok szolgálnak, melyek élei az adatok közötti relációk típusainak megfelelően címkézettek, vagy az élek súlyai az adatok közötti kapcsolat erősségét reprezentálják.

A szövegbányászati módszerek alapján lényegében a MEDLINE adatbázis egy olyan struktúrált tartalomjegyzéke áll elő, amely a kísérletekből nyert génexpressziós adatok által definiált. A genomikai kutatások esetén a törvényszerűségek felismeréséhez a legfontosabb alapot a gének közötti kapcsolatrendszer minél pontosabb feltárása szolgáltatja. A gének közötti kapcsolatokat a szövegalapú adatbányászat három szinten elemezte. Első szintű kapcsolatnak azt tekintettük, amikor két gén azonos absztraktban szerepel. Második szinten már azt vizsgáltuk, hogy a két gén szerepel-e azonos mondatban. Ugyanakkor fontos, hogy az irodalomból kinyert adatok sokrétű elemzést tegyenek lehetővé azt illetően, hogy az egyes gének miként hatnak egymásra különböző állapotokban. Az állapotok az absztraktokból előálló strukturális táblázat információkategóriái lehetnek, pl. egy adott szövet, kezelés vagy betegség, stb. Így harmadik szinten a csúcspontok közötti relációk paramétereit az szolgáltatja, hogy a vizsgálandó gének expressziós szintje miként változik egy adott állapotban (pl. mindkettőé nő). Ez utóbbiak alapján következtethető, hogy az adott absztraktban található információ utal-e a két gén interakciójára. Az analitikus infrastuktúra alapját képező gráfban ezért a csúcspontok a gének, és két gén akkor lesz éllel összekötve az adott szinten, ha szerepelnek legalább egy azonos absztraktban, azonos mondatban, illetve interakcióban állnak egymással. Ezeknek a mértéke határozza meg a kapcsolat erősségét.

A génhálózatok vizsgálatára használt más módszerekhez képest azonban új elemzési szempontokat tesz lehetővé a gráfstruktúra egy további genomikai kategóriája, melyet az úgynevezett funkcionális csoportok alkotnak. Egy funkcionális csoportot az adott speciális funkcióhoz szükséges gének alkotják. Egy gén tetszőleges számú funkcionális csoportba sorolható, amelyek ezen felül még hierarchiába is szervezhetőek. Egy adott kapcsolat esetén a gén a funkcionális csoportnak a hierarchia szerinti összes leszármazottjával kapcsolatban áll, ezért a gén és a funkcionális csoport közötti kapcsolat megadása minden esetben a lehetséges legmagasabb szintű funkcionális csoportokkal történik. A funkcionális csoportokon keresztül történő kapcsolódás az irodalmi absztraktokból kinyert adatok mellett egy további paramétert kínál a gének közötti kapcsolat erősségének meghatározására.

Összefoglalva: a MedLine adatbázisból kinyert különböző szintű adatok (közös absztrakt, közös mondat, interakció), valamint a funkcionális csoportokon keresztül történő kapcsolódás együttesen határozza meg a struktúrában a két gén kapcsolatának erősségét, mely a gráfban a géneknek megfelelő csúcsokat összekötő él súlyában nyer kifejezést. A feladat tehát egy súlyozott gráffal reprezentált struktúrált adathalmaz elemzése.

4. Adatbányászat és vizualizáció

Az adatbányászat valójában egy gyűjtőfogalom, mely olyan különböző eljárásokat, technológiákat tartalmaz, melyek mindegyike alkalmas arra, hogy hatalmas méretű adatbázisokban is nagy hatékonysággal keresse meg az adatok között fennálló összefüggéseket. Azonban abból adódóan, hogy a szövegalapú adatbányászattal előálló adatok is tartalmazhatnak hibákat (pl. a szinonimák nagy számára való tekintettel az elemzés rendkívül komplex feladat), nem tűnt célszerűnek az automatikus következtetések levonására alkalmas módszerek integrálása az esettanulmányban ismertetett projekt során kifejlesztett rendszerbe. Ezért a szintén az adatbányászat témakörébe tartozó olyan eljárások alkalmazását és kifejlesztését tűztük ki célul, melyek az eredmények értelmezését könnyítik meg.

Két módszertan képezi a fentiek alapját, az automatikus szegmentáció (klaszterezés) és a vizualizáció. Az automatikus szegmentáció célja, hogy csoportosítsa azon géneket, melyek erősen kapcsolódnak egymáshoz. Mivel az adatok egy súlyozott gráfba rendezettek, ezért a gráfok csúcspontjait kell olyan módon diszjunkt csoportokba (klaszterekbe) rendezni, hogy az egy-egy klaszter által kifeszített részgráf „sűrű" legyen (a csúcsok számához viszonyítva az élek összsúlya nagy), míg a klaszterek között futó élek „ritka" gráfot határozzanak meg.

A vizualizáció lehetőséget kínál a vizsgálandó gráfstruktúra „logikájának" megjelenítésére, azaz egy olyan egyszerűsített ábrázolását kapjuk a kapcsolatrendszernek, mely az aktuális szempontok tekintetében lényeges összefüggéseket emeli ki, és ez által az eredmények könnyebben értelmezhetővé válnak. A klaszterezés azonban a vizualizáció során is döntő jelentőséggel bír, hiszen egy nagyobb gráfban az összefüggések átláthatatlanokká válnak, amennyiben a csúcspontokat nem rendezzük a klaszterek szerinti csoportokba. A fentiek miatt a kifejlesztett klaszterezési eljárásokkal kapcsolatban alapvető elvárás volt, hogy a vizulizációba integrálhatóak legyenek.

A vizualizációba integrált klaszterezés lényegében két fő elv alapján valósulhat meg. Az egyik megközelítés előbb kialakítja a klasztereket és a megjelenítéskor elhelyezi őket optimálisan a térben, másrészt a klasztereken belül próbálja a csúcspontokat az egymáshoz való kapcsolatuk alapján vizualizálni.(lsd. [9]) A másik megközelítés olyan metrikákat próbál alkalmazni a klaszterezésre és a vizualizációra is, amelyek könnyen transzformálhatóak egymásba. Mi az utóbbi megközelítést választottuk, így a gráfok klaszterezésére a gráfstruktúrán alapuló kombinatorikus módszereket nem is vizsgáltuk, csak olyan eljárásokat teszteltünk, amelyek egy alkalmas metrikát definiálnak a gráfon. Azonban a szakirodalomban található metrikán alapuló gráfklaszterezési eljárásokhoz kapcsolódóan vagy nem található vizualizációs alkalmazás (lsd. pl. [17]), vagy pedig az adott metrikához kötődő a megvalósítás ([10]). Mivel a különböző lekérdezések változó struktúrájú gráfokat eredményezhetnek, ezért az volt a koncepciónk, hogy több módszer kombinációjával valósítsuk meg az analitikát, illetve, hogy több klaszterezési eljárás közül választhasson a felhasználó. Ez azt jelentette, hogy több metrikát is alkalmaznunk kellett, amely szükségessé tette, hogy a vizualizáció tartalmazzon egy általános metrikamegőrző leképezést ahelyett, hogy minden esetben külön módszer kerüljön kidolgozásra a csúcspontok térben történő elhelyezésére.

A gráfok vizualizációjánál azt is definiálni kell, hogy a rajzolás milyen szempontokat vegyen figyelembe. A jelen esetben természetes elvárás az, hogy a csúcspontok közötti távolság a hasonlóság mértékét fejezze ki,

hiszen a gráf csúcsai a géneket reprezentálják. Mivel a vizualizáció megvalósítása metrikamegőrző, ezért a gráfok csúcspontjain definiált mértéknek a gének közötti hasonlóságot kell kifejezniük.

Összefoglalva: a gráfokon definiáltunk egy metrikát, amely a csúcspontok közötti hasonlóságot/különbözőséget reprezentálja, azaz egy szimmetrikus távolságfüggvényt hoztunk létre, amely bármely pontpáron értelmezett és érvényes rá a háromszögegyenlőtlenség. Erre a függvényre alkalmazva egy metrika-megőrző leképezést kapjuk a csúcsok térbeli elhelyezését. Ugyanezt a metrikát alapul véve pedig valamely klaszterezési eljárást alkalmazva a vizualizáció konzisztens lesz a szegmentációval.

4.1 Metrikák

A kialakított távolságfüggvény különböző metrikák súlyozott kombinációjaként adódott. Az alábbiakban ezeket a metrikákat ismertetjük.

Euklideszi távolság: Vektortérben a vektorok közötti távolságra alkalmazzák a két vektor különbségvektorának a hosszát, azaz a különbségvektor önmagával vett skaláris szorzatának a négyzetgyökét. A gráfok esetében ez nem jelent mást, mint az illeszkedési mátrix sorvektoraira alkalmazzuk a sorok különbözőségének kifejezésére. Minél hosszabb különbségvektort kapunk, annál nagyobb a soroknak megfelelő pontok (gének) közötti távolság. A gráfok klaszterezésére használt metrikák között nem ajánlja a szakirodalom ([11]), azonban a csúcsok közötti erős kapcsolatok feltérképezésére alkalmas, ezért a kombinált metrikánknak ez is a részét képezi.

$$D(i,j)=\sqrt{\sum_{k\neq i,j}^{n}(a(i,k)-a(j,k))^2}$$

ahol *n* a csúcsok számát, az *a(i,j)* pedig az illeszkedési mátrix *(i,j)*-ik elemét jelöli, azaz az *i*-ik és *j*-ik csúcsot összekötő él súlyát.

Véletlen bolyongáson alapuló metrika: Súlyozott gráfokban a kapcsolatok erejének a feltérképezésére gyakran alkalmazzák a véletlen sétákat, melynek tipikus alkalmazásai például a webes keresők ([3]). Az elgondolás az, hogy amennyiben két pont között nagyobb valószínűséggel terjed az információ, akkor azok erősebb kapcsolatban állnak. Ezen módszerek az úgynevezett gyenge kapcsolatok feltérképezésére is alkalmasak szemben a fenti klasszikus módszerrel. Jelen eljárás kiindulópontja a gráf csúcsain vett véletlen bolyongás, az úgynevezett Brown-mozgás. A véletlen bolyongás azt feltételezi, hogy egy adott *v* csúcsból történő továbblépés esetén, annak a valószínűsége, hogy ez a *v*-re illeszkedő adott *e* élen történik meg, az *e* súlyának és a *v*-re illeszkedő élek összsúlyának aránya által adódik. Ezeket az értékeket az úgynevezett *valószínűségi átmenetmátrix*ban tároljuk. Két csúcs között az alapértéket a véletlen bolyongás lépésszámának várható értéke szolgáltatja ([17]):

$$B(i,j)=\sum_{l=1}^{n}\left(\frac{1}{I-A(j)}\right)_{il}$$

ahol *I* az egységmátrix, *A(j)*-t pedig a valószínűségi átmenetmátrixból képezzük úgy, hogy a *j*-ik oszlop konstans nulla. Ez a hozzárendelés nem teljesíti a metrikához szükséges feltételeket (sem a szimmetriát, sem a háromszög-egyenlőtlenséget), viszont az euklideszi metrikához hasonló transzformáció már valódi metrikát biztosít ([18]):

$$D(i,j)=\frac{\sqrt{\sum_{k\neq i,j}^{n}(B(i,k)-B(j,k))^2}}{n-2}$$

Diffúziós metrika: Két csúcs között a diffúziós távolságot ([14]) azon alapmetrika segítségével definiáljuk, amely azt fejezi k, hogy egy adott csúcsból indulva egy másik adott csúcsba milyen valószínűséggel kerülünk pontosan *t* lépést követően. Ezt követően ezekre az alapértékekre alkalmazzuk az euklideszi metrikát olyan módon, hogy a különbségvektor egyes komponenseit az adott csúcs fokszámának megfelelően normáljuk. Ennél a módszernél a *t* paraméter jó választása döntő jelentőségű, a tesztjeink azt mutatták, hogy a *3* és *5* közötti érték az optimális.

$$D(i,j)=\sqrt{\sum_{k=1}^{n}\frac{\left(P_{ik}^{t}-P_{jk}^{t}\right)^2}{d(k)}}$$

ahol P^t a valószínűségi átmenetmátrix *t*-ik hatványa, *d(k)* pedig a *k*-ik csúcsra illeszkedő élek összsúlya.

4.2 Metrikamegőrző leképezések

A szóba jöhető általános metrikamegőrző leképezések kapcsán a dimenzió redukciós módszerek adódnak, amelyek egy tetszőleges olyan *M* sokaságot, melynek pontjai között távolság definiálható egy *n*-dimenziós euklideszi térbe helyez el olyan módon, hogy a sokaságon értelmezett távolságot a pontok térbeli elhelyezése a lehető legjobban megőrizze. A legismertebb módszerek a Locally Linear Embedding ([15]), a Principal Component Analysis ([7]) és a Multidimensional Scaling ([5]). Mi a Multidimensional Scaling (MDS) módszerét alkalmaztuk, amely az elvégzett tesztek alapján a három eljárás közül numerikusan a legstabilabbnak bizonyult. Az MDS módszerek statisztikai alapú technikák gyűjteményei, ahol a célfüggvény a pontok közötti eredeti távolságok és a leképezés után kapott távolságok négyzetösszege által meghatározott hibafüggvény:

$$E_M = \sum_{k \neq l} [d(k,l) - d'(k,l)]^2$$

ahol *d(k,l)* a *k*-ik és *l*-ik pont közötti eredeti távolság, a *d'(k,l)* pedig ezen pontoknak a leképezés utáni euklideszi távolsága. A cél a fenti hibafüggvény minimalizálása, azaz olyan leképezés meghatározása, amelyre a fenti érték minimális.

4.3 Klaszterezés

A megfelelő metrika meghatározásával a gráf csúcspontjain a klasszikus klaszterezési módszerek alkalmazhatóak ([13]). Alapvetően két megközelítés alapján lehetséges elvégezni a klaszterezést: partícionáló és hierarchikus módszerek. A partícionáló módszerek az adatokat *k* osztályba sorolják, ahol a *k* előre adott, míg a hierarchikus módszerek az adathalmaz hierarchikus dekompozícióját állítják elő. A partícionáló módszerek hátránya, hogy a *k* értéket előre meg kell adni, azonban az adott *k* esetén egy jó közelítést nyújt az optimumra. A hierarchikus módszerek esetében minden lépésben az aktuális klaszterezést két klaszter egyesítésével vagy egy adott klaszter felosztásával igyekszünk javítani, azaz vagy kiindulunk egy klaszterből és minden lépésben felosztunk egy választott klasztert két részre, vagy *n* klaszterből indulunk ki és minden lépésben összevonunk két kiválasztott klasztert. Ez hatékonyabb implementációt eredményez, de amennyiben egy adott ágra kerülünk a hierarchikus fában, akkor onnan már nem tudunk visszalépni. A gráfok klaszterezésénél általában az egyesítő hierarchikus klaszterezést szokás alkalmazni, azonban mi kihasználva a metrikus teret, partícionáló módszereket is megvalósítottunk.

A partícionáló módszerek esetében a *k* paramétert végig kellett futtatni több lehetséges értékre. Az alkalmazott módszerek a *k*-átlag és a *k*-medoid eljárásokra épültek A *k*-átlag eljárás esetében a hasonlóságot a klaszterbeli elemek átlagához mérjük, míg a *k*-medoid esetében a klaszterbeli elemek közül a medoidhoz. Mindkét módszer szerint kiválasztunk *k* darab klasztert középpont, illetve medoid alapján, majd besoroljuk a többi elemet a legközelebbi klaszterbe. Ezután meghatározzuk az így kapott új középpontokat, illetve medoidokat. Ezt az eljárást addig folytatjuk, amíg változik a klaszterek összetétele.

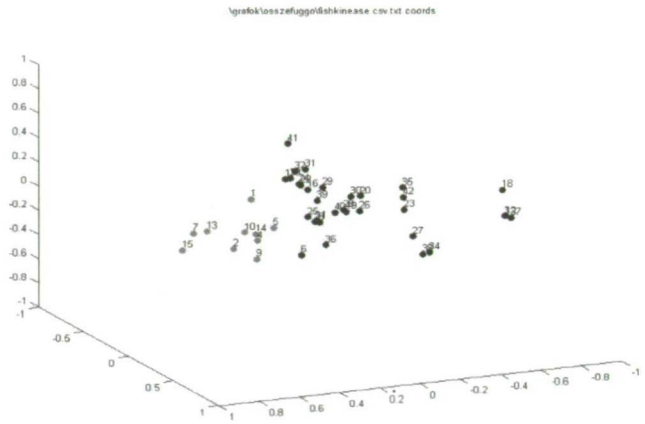
A hierarchikus módszerek esetében az egyesítő eljárásokat valósítottuk meg. Itt a fő kérdés azon alapszik, hogy a klaszterek közötti távolságot milyen alapon határozzuk meg, illetve, hogy ennek alapján miként választunk a klaszterek között az összevonásra. Klasztertávolságként elemeztük a legközelebbi, a legtávolabbi és a medoid csúcsok távolságát, valamint az átlagos távolságot. Az összehasonlítások jelentős eltérést nem mutattak, így a teszteredményeink alapján a legközelebbi csúcsok alapján történő távolság lett az optimális választás. A klaszterek közötti választás esetében több mérőszám szerint is elemeztük a kapott klaszterezéseket, de jelentős eltérést ezek sem mutattak, így itt is a leghatékonyabb, a legközelebbi klaszterek választása bizonyult optimálisnak.

4.4 Kiértékelés

Mindkét módszertan esetében a biológiai jelentésből adódó elemzés mellett fontos szempont volt, hogy a klaszterezés minőségét matematikailag is kiértékelhetővé tegyük. Több mérőszámot alkalmaztunk ennek a vizsgálatára, végül a biológiai elemzéseket és a hatékonyságot alapul véve az úgynevezett modularitás ([12]) bizonyult optimálisnak. A modularitás képlete hatékonyan számítható és azt adja vissza, hogy a klasztereken belüli élsúlyok összegének az aránya miként viszonyul egy véletlen klaszterezéshez.

$$Q = \sum_i e_{ii} - \sum_{ijk} e_{ij} e_{ki},$$

ahol e_{ij} jelenti az *i*-ik és *j*-ik klaszter közötti élsúlyok összegének a gráfban levő összélsúlyhoz viszonyított arányát.



2. ábra
FishKinease klaszterezése Matlab környezetben

A magasabb érték jelent jobb minőségű klaszterezést. Azonban a modularitás mellett a klaszterek számát is figyelembe kellett venni az algoritmusok testre szabásánál, melyet a mellékelt ábrán szemléltetünk (2. ábra). Ezen teszteredmény még az algoritmusfejlesztés fázisából való, a Matlab programcsomag segítségével végrehajtott egyszerűsített (élek és génnevek nélküli) megjelenítést használva a Fishkinease mintaállományon. Azonban jól látszik, hogy mindössze két klasztert állít elő a módszer, amely bár magas modularitású, a biológiai kiértékelés nem tartotta megfelelőnek a szegmentációt. A fő ok abban keresendő, hogy túl kevés vagy túl sok klaszter biológiailag nem hordoz megfelelően értékelhető információt. A fenti mintaállománynak a BiblioGraph Explorer rendszerébe beépített klaszterezését és vizualizációját a következő részben ismertetjük.

5. Alkalmazás

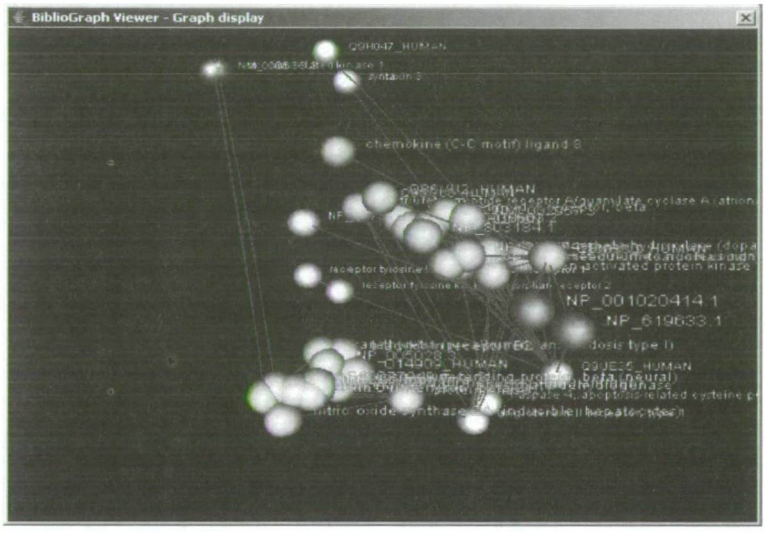
A szöveg alapú és a gráf alapú adatbányászati módszerek alapján kifejlesztésre került a BiblioGraph Explorer elnevezésű alkalmazás. A rendszer kifejlesztésében az MTA SZBK Funkcionális Genomika Laboratórium (DNS chip kísérletek, biológiai adatelemzés), az SZTE Informatika Tanszékcsoport (szöveg alapú adatbányászat), az SZTE JGYPK Számítástechnika Tanszék (gráf alapú adatbányászat és vizualizáció), valamint a Data Explorer Kft. (rendszer megvalósítása) vett részt. A rendszer használatát az alábbi teszteredménnyel szemléltetjük.

A kísérlet során az MTA SZBK Funkcionális Genomika Laboratórium munkatársai azt vizsgálták, hogy bizonyos többszörösen telített zsírsavakkal kezelt rákos sejtekben milyen génexpressziós változások következnek be. A kapott információ nagyon értékes a zsírsavakkal kapcsolatos kutatások területén, mivel több tanulmány is foglalkozik ezeknek a vegyületeknek a daganatellenes hatásaival, de a pontos hatásmechanizmus még nem ismert. A DNS chip technológiát felhasználva megállapították, hogy több gén kifejeződése is megváltozott. A gének közötti kapcsolatot a BiblioGraph Explorer szoftverrel határozták meg. A kapott eredményeket szoftver nélkül is leellenőrizték a PubMed biológiai publikációs adatbázis alkalmazásával. A publikációk összefoglalóit, amelyekben a vizsgált gének szerepeltek, a szoftver minden esetben megtalálta. Az ennek alapján felépített gráfstruktúrának (a gének kapcsolatának) a szoftverrel történő megjelenítését a 3. ábrán láthatjuk, ahol a klasztereket az eltérő színek jelenítik meg. A kapcsolatrendszer felderítésével közelebb juthatunk a génexpresszió szintű változások pontosabb megértéséhez.

Az ábrából, illetve a klaszterezésből a következők kerültek megállapításra. A változást mutató gének együttes előfordulása 6 jól elkülöníthető csoportot határozott meg, melyek mindegyike biológiailag jól jellemezhető, a két legnagyobb csoport a főként sejtciklussal kapcsolatos gének (jobboldalon felül) illetve a főként gyulladásos folyamatokban szereplő gének (baloldalon alul).

A klaszterezési eredményből kitűnik, hogy a központi helyen a Q9UE35 gén szerepel (baloldalon középen), amelynek kitüntetett szerepe van többek között a sejtciklusban. Szintén kapcsolatot fedezhetünk fel ezen gén és a gyulladásspecifikus gének között. Ezen felismerések újdonságnak számítanak a politelítetlen zsírsavak sejtburjánzásra gyakorolt gátló hatásában.

A fenti példa jól mutatja, hogy az alkalmazás segítségével nemcsak a gének közötti kapcsolat ábrázolható, hanem a csoportosításukkal lehetőség nyílik új összefüggések felismerésére is.



3. ábra
BiblioGraph Explorer

6. Összefoglalás

A gráf alapú adatbányászat napjainkban a nagy méretű nyílt adatbázisokban szereplő adatok közötti strukturális kapcsolatok felderítésével a tudásfeltárás egyik leghatékonyabb eszköze. A biológiai kísérletekből származó eredmények mennyiségének rohamos növekedéséből fakadólag a bioinformatikának is fontos részévé kezd válni ez a modern technológia. A cikkben ismertetett esettanulmányban bemutattuk, hogy a DNS-chip kísérletek eredményéből származó adatok alapján a MEDLINE adatbázisból kinyert információk segítségével miként lehet olyan szoftverbe is integrálható gráf alapú adatbányászati és vizualizációs módszereket kifejleszteni, amelyek segítségével új funkcionális genomikai összefüggések felfedezésére is lehetőség nyílik.

Irodalomjegyzék

[1] T. Aittokallio, M. Kurki, O. Nevalainen, T. Nikula, A. West, R. Lahesmaa, Computational Strategies for Analyzing Data in Gene Expression Microarray Experiments, *Journal of Bioinformatics and Computational Biology* **1** (3): 541–586 (2003).

[2] S. Ananiadou, J. Mcnaught, Text Mining for Biology and Biomedicine, *Artech House Publishers*, 2005.

[3] S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks* **30** (1–7): 107–117 (1998).

[4] R. Busa-Fekete, A. Kocsor, Extracting Human Protein Information from MEDLINE Using a Full.Sentence Parser, *Acta Cybernetica*, megjelenés alatt.

[5] M.F. Cox, M.A. A. Cox, *Multidimensional Scaling*, Chapman and Hall, 2001.

[6] D. Csendes, Z. Alexin, R. Busa-Fekete, K. Kovács, New, Linguistics-based, Ontology-enabled Approaches, in *Biological Information Management, in the Proceedings of the e-Challenges 2006 Conference, October 25–27*, pp. 1352–1359, Barcelona, Spain (2006).

[7] B. S. Everitt, G. Dunn, *Applied Multivariate Data Analysis*, Arnold, 1991.

[8] J. Han, M. Kamber, *Data Mining: concepts and techniques*, Morgan Kaufmann Publishers Inc, 2000. (Magyar fordítás: Panem, 2004.)

[9] I. Herman, G. Melancon, M. S. Marschall, Graph Visualization and Navigation in Information Visualization: a Survey, *IEEE Transactions on Visualization and Computer Graphics*, **6** (1): 23–42 (2000)

[10] S. Lafon, A. B. Lee, Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28** (9): 1393- 1403 (2006).

[11] M. E. J. Newman, Detecting community structure in networks, *Eur. Phys. J. B* **38**, 321–330 (2004).

[12] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, *Phys. Rev. E* **69** (2) 026113 (2004).

[13] G. Pfister, *In Search of Clusters*, Prentice Hall, 2nd ed., 1997.

[14] P. Pons, M. Latapy, Computing Communities in Large Networks Using Random Walks, *Lecture Notes in Computer Science* **3733**, pp. 284–293 (2005).

[15] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**: 2323–2326 (2000).

[16] T. Washio, H. Motoda, State of the Art of Graph-based Data Mining, *SIGKDD Explorations* **5** (1): 59–68 (2003)

[17] H. Zhou, Network landscape from a Brownian particle's perspective, *Phys. Rev. E* **67** 041908 (2003).

[18] H. Zhou, Distance, dissimilarity index and network community structure, *Phys. Rev. E* **67** 061901 (2003).

MIKLÓS KRÉSZ – ATTILA TÓTH

Graph-based data search and visualisation: a case study

Questions arisen by DNS-chip technology pose new challanges to bioinformatics. In contrast to the information stored in static DNS databases, DNS-chip experiments provide large amount of information about dynamic changes in expressions of several thousand genes simultaneously. It is a natural goal to exploit both of these information sources, obtaining new results and dependencies which opens new horizon in bioinformatics branch of genomic research. Since structural relationships play an important role in modern data analysis, graph theoretic models and algorithms are popular tools in this field. In this paper we present our experiences about graph clustering and graph visualizing methods developed in the project „Natural Language Processing, Information Extraction and Development of a Graph Based Analytic Infrastructure for Genomic Research"[1]

[1] A projekt a Magyar Köztársaság és az Európai Unió társfinanszírozásával a GVOP AKF program keretében jött létre (projekt azonosító száma: GVOP–3.1.1–2004–05–0119/3.0)