Eötvös Loránd Tudományegyetem
Informatika Doktori Iskola

Szalkai Balázs

Algorithmic Problems In Bioinformatics
Algoritmikus kérdések a bioinformatika területén

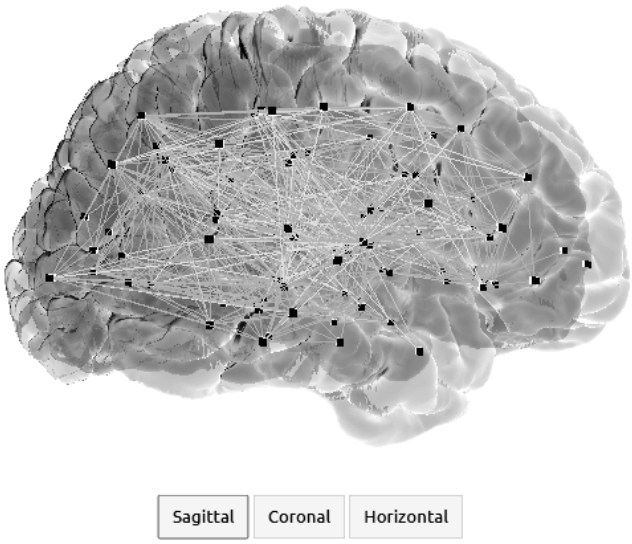
Doktori értekezés tézisei

Eötvös Loránd Tudományegyetem
Informatika Doktori Iskola (vezető: Dr. Csuhaj Varjú Erzsébet)
Információs rendszerek doktori program (vezető: Dr. Benczúr András)

Témavezető: Dr. Grolmusz Vince
Számítógéptudományi Tanszék

Budapest, 2018.

1. Motiváció



1. ábra. A Budapest Reference Connectome szerver

A biológiai kutatás gyakran nagy adathalmazokat eredményez, melyeket aztán számítógéppel tudunk feldolgozni. Ez a „big data” sokféle természetű lehet, pl. több ezer emberen végzett gyógyszerkísérletek eredményei táblázatokba szedve, több Gbp-nyi (gigabázispár) szekvenált genom, MR-felvételek, protein–protein interakciós gráfok, adatok fehérjék 3D struktúrájáról, avagy neurológiai hálózatok, melyek egy organizmus idegrendszerének mikro- vagy makroszerkezetét írják le.

Egy jól ismert példa a *Caenorhabditis elegans* fonálféreg „agygráfja” [14], azaz neuronjainak és azok összeköttetéseinek diagramja. Ezt sokan tanulmányozták, és az OpenWorm [12] projekt keretén belül már részben szimulálták is. Bár ennek az egyszerű féregnek az idegrendszeri gráfja csak pár száz neuronból és az azok közötti szinapszisokból áll, mégis jelenleg olyan komplexnek tűnik, hogy emberi ésszel még csak nagyon korlátozottan sikerült dekódolni a működési elvét. Az emberi agyról is készíthetünk gráfokat, bár itt még nem neuronok, hanem nagyobb területek (ROI-k, regions of interest) között. Hatékony számítógépes algoritmusok nélkül ezeket a gráfokat nem tudnánk elemezni.

Manapság könnyen találhatunk szabadon hozzáférhető bioinformatikai adathalmazokat az interneten. Rengeteg erőforrást invesztáltak ezeknek az adatoknak az összegyűjtésébe, de hiszek abban, hogy ezekből az adathalmazokból messze nem minden információt nyertek ki az összeállítóik. Másszóval, további érdekes felfedezéseket tehetünk azáltal, hogy közzétett adatokat tovább elemzünk. Ezek az adatbázisok általában egy adott hipotézis tesztelésére készültek (pl. hatékony-e egy gyógyszer), de jelentős eredményeket érhetünk el azzal, hogy ezekre az adatokra más perspektívából tekintünk, megpróbáljuk belőlük a legtöbbet kihozni, és ezt jelenti számomra a bioinformatika.

2. Módszerek és eredmények

2.1. Nem-euklideszi k-means

A tézis első fejezete a k-means klaszterezési algoritmus általam végzett általánosításáról szól [6]. Az eredeti k-means algoritmus csak egy euklideszi tér pontjaira működik. Ez egy nagyon fontos limitáció, mivel gyakran az adatpontok valami absztrakt tér elemei, melyeken egy nem-euklideszi távolságfüggvényt definiáltunk. Megmutattam, hogy a klasszikus k-means algoritmus általánosítható nem-euklideszi esetekre is, amikor egy tetszőleges távolságmátrixunk van, amelynek nem is kell metrikának lennie. Az algoritmusomat azóta számos más kutató [1–3] alkalmazta.

2.2. Asszociációs szabályok és az Alzheimer-kór

Az első eredményeim közé tartozott, hogy egy gyógyszerkísérletekből származó adatbázist használtam arra, hogy asszociációs szabályokat találjak az Alzheimer-kórral kapcsolatban. Ez a betegség egy nagy teher a fejlett társadalmak számára, ahol az emberek viszonylag sokáig élnek a mai fejlett orvostudománynak köszönhetően. Habár most már könnyen megérhetjük a 80–90 éves kort is, nem igazán tudunk még hatékonyan küzdeni a demencia különböző formái ellen. Van pár gyógyszerünk az Alzheimer-kórra, melyek lelassíthatják a kór előrehaladtát, de a legtöbb esetben a diagnózis már túl későn következik be, mikor az agysejteknek egy jelentős része már elpusztult. Ezért biomarkereket akartunk találni, melyek akár jóval előrejelezhetik az Alzheimer-kórt.

Egy, a CAMD (Coalition Against Major Diseases) által összeállított adatbázist használtunk, amely 11 gyógyszerkísérlet alanyainak adatait tartalmazza: demográfiai adatokat, vérképet, egyéb laboreredményeket, mentális egészségről és kognitív státuszról szóló felmérések adatait beleértve. A célunk az volt, hogy *kombinatorikai asszociációs szabályokat* találjunk az Alzheimerrel és általánosságban véve a demenciával kapcsolatban. Azaz, logikai következtetéseket kerestünk, ahol a bal oldal attribútum-érték pároknak egy ÉS/VAGY kombinációja, és a jobb oldal valahogy összefügg a demenciával. Például, a következő kifejezés egy kombinatorikai asszociációs szabály:

$$sodium = high \wedge (protein = high \vee age \geq 60) \implies mmse\_total \leq 15 \tag{1}$$

Az általam kifejlesztett program hasonló kifejezéseket generál, és megvizsgálja ezeknek a szabályoknak az igazságtartalmát. Az eredmények önmagukban is érdekesek, viszont az újonnan kifejlesztett adatbányász programomat publikáltam az interneten is, telepíthető formában és nyílt hozzáférésű webszerver formájában is. A webszevernek a SCARF nevet adtam, amely a Simple Combinatorial Association Rule Finder rövidítése. A webszerver lehetővé teszi más kutatók számára, hogy kombinatorikai asszociációs szabályokat bányásszanak anélkül, hogy egy adatbányász szoftvert kelljen telepíteniük. A [7] cikkünk írja le ezeket az eredményeket.

Tapasztalatunk szerint minél könnyebben használható egy bioinformatikai szoftver, annál nép-

szerűbbé válik a kutatók körében. Ez nem meglepő, mivel a felhasználók nagy része biológiai és nem programozói háttérrel rendelkezik. Kiváló példa a webszerverek esete. Az online alkalmazások gyakran azért jobbak a telepítendő programoknál, mivel néha a legnehezebb dolog letölteni, lefordítani, beállítani és működésre bírni egy szoftvert. Bármilyen könnyűnek is hangzik, fáradságos munka lehet. Azt gondolom, hogy a valóban gyors és jól implementált algoritmusok mellett a felhasználóbarát interfész az oka többek között a jól ismert BLAST (Basic Local Alignment Search Tool) és MG-RAST bioinformatikai eszközök népszerűségének. Ezért tartottam nagyon fontosnak azt is, hogy a SCARF-ot online webszerverként is megvalósítsam.

2.3. A metagenomikai teleszkóp

A teljes humán genom szekvenálása kétségkívül mérföldkő volt a bioinformatika történetében. A Humán Genom Projekt, amelynek ez volt a deklarált célja, 1990-től 2003-ig tartott, és 3 milliárd dollárt emésztett fel. Manapság a teljes genom szekvenálás (whole genome sequencing, WGS) mintánként kevesebbe kerül, mint $10,000, sőt, a gyakorlatban megközelíti az ezer dolláros határt.

Az új generációs szekvenálás rengeteg adattal lát el minket. Kérdés, hogy mit tudunk kezdeni mindezzel az adattal. A legfőbb kihívás már nem igazán a szekvenálás költségének csökkentése, nem is az összeállítás (assembly) és az utófeldolgozás felgyorsítása, hanem az, hogy alkalmazásokat találjunk, mivel maga a szekvenálási folyamat már eleve hatékonynak mondható. A teljes genom szekvenálással pl. új kapcsolatokat találhatunk betegséget és genetikai mutációk között, öröklletes betegségeket diagnosztizálhatunk, és eddig nem ismert, viszont hasznos géneket találhatunk baktériumokban és archaeákban.

Egy hasonló technológia megjelenése hívta életre a metagenomika tudományágát. Az új generációs szekvenálás megjelenése előtt egy környezeti minta bakteriális összetételének megállapítása tenyésztéssel történt, aztán mikroszkóp alatt a kolóniák megszámolásával. Bár ez egy egyszerű és alacsony anyagköltségű művelet, az emberi beavatkozás szükségessége miatt lassú és nehezen automatizálható. De a legfőbb probléma az, hogy csak az organizmusoknak egy kis része tenyészthető laboratóriumban: azok, amelyek számára megfelelő az adott táptalaj. Pl. az extremofilek nem élnek túl klasszikus laboratóriumi körülmények között, éppen azért, mert extrém körülményekhez alkalmazkodtak. Továbbá, ez a módszer nem alkalmas új fajok felfedezésére, mert azokat nem igazán tudjuk, hogy kellene tenyészteni.

Ebből az ördögi körből úgy léphetünk ki, hogy új generációs szekvenálási eljárásokat használunk ahhoz, hogy pontos információt nyerjünk környezeti mintákról. A mintának először kivonják a DNS-tartalmát (amit *metagenom*nak hívunk), aztán a DNS-t feldarabolják pár száz bázispár hosszú darabokra. Ezután ezeket a szekvenciákat további biokémiai módszerekkel (pl. PCR) dolgozzák fel, utána pedig egy szenzorchipre kerülnek, ahol minden egység egy-egy DNS-darab nukleotidsorrendjét határozza meg. Az eredmény több millió rövid DNS-szekvencia, azaz *read* lesz. Ennek a megközelítésnek nagy előnye, hogy a rövid darabokat egyenletesen választjuk a mintában eredetileg

élő organizmusok genomjából. Bár a kiolvasott DNS-darabok csak egy kis töredékét alkotják az organizmusok genomjai uniójának, mégis elég sokat olvasunk ki belőlük ahhoz, hogy statisztikailag helyes következtetéseket vonhassunk le, másrészt minden DNS-darab egyformán valószínűen kerül bele az eredmény-adathalmazba.

Az eredményként kapott adathalmaz ezután az eredeti környezeti minta tulajdonságainak meghatározására használható. A readeket ismert fajokhoz vagy nagyobb taxonómiai csoportokhoz rendelhetjük, így a minta taxonómiai összetételét megbecsülhetjük a metagenom alapján. Vigyáznunk kell, hogy korrigáljunk a különböző baktériumok genomjának méretével, mivel egy arányosan nagyobb genommal rendelkező organizmusból arányosan több read kerül kiolvasásra.

A taxonómiai analízis mellett, a readeket egy ismert proteinkódoló géneket tartalmazó adatbázishoz illeszthetjük, és megállapíthatjuk a metagenom *funkcionális összetételé*t. Pl. megállapíthatjuk azt, hogy a mintában élő organizmusoknak sok szénhidrogén-metabolizmussal kapcsolatos génjük van, és ebből az következhet, hogy a környezetben több szénhidrogén található. Egy ilyen mintából pedig olyan organizmusokat nyerhetünk, amelyek segíthetnek nekünk a szennyezett területeken az olaj lebontásában.

Egy új, *metagenomikai teleszkóp* [4] nevű eszközt fejlesztettünk ki arra, hogy meghatározzuk magasabbrendű organizmusokban több, eddig ismeretlen funkciójú gén szerepét. Először egy HMM-et (Hidden Markov Model) építettünk ismert DNS-javító génekre. Ezután DNS-javító géneket kerestünk extrém körülmények között élő mikroorganizmusokban ezzel a modellel. Savas bányaiszapból, gejzírekből és egy szennyvíztisztító iszapjából származó minták metagenomjait használtuk ehhez, mivel feltételeztük, hogy az ilyen extrém körülmények között élő organizmusok hatékonyabb DNS-javító enzimekkel rendelkeznek. Ezután egy új, kibővített HMM-et építettünk az eredményekből és az eredeti szekvenciákból. Ezt az új HMM-et használtuk arra, hogy bizonyos élőlények (ember, kutya, csirke, szarvasmarha, stb.) genomjában DNS-javító enzimeket kódoló géneket keressünk. Az eredeti HMM-et is lefuttattuk kontrollként. A kibővített HMM-ek több szekvenciát találtak. Az új találatok által kódolt proteinek 3D szerkezetét megvizsgálva azt láttuk, hogy egyes esetekben ezek a szekvenciálisan hasonló fehérjék térszerkezete is hasonlított a DNS-javító enzimekéhez, sőt, hasonló multimereket is képeznek. Ez akár jelentheti azt is, hogy ezek is DNS-javító enzimek, vagy esetleg közös fehérjékből evolválódtak. Ezzel igazoltuk a metagenomikai teleszkóp hasznosságát.

2.4. 9-merek gyakorisága

Nem feltétlenül kell a HMM-hez hasonló viszonylag komplikált módszereket alkalmazni a metagenomikai analízisben. Ha egyszerűen csak megszámoljuk egy rövid nukleotid-szekvencia darabszámát, az is érdekes felfedezésekhez vezethet. Pl. már régóta ismert, hogy a GC-tartalom (a G és C nukleotidok száma) egy genomban baktériumfajokra specifikus. De mi történik, ha hosszabb szekvenciákat számolunk meg? Talán ezek is segíthetnek genomok vagy metagenomok

osztályozásában. Persze a kb. 1000 nukleotid hosszú szekvenciák biztosan specifikusak és „értelmesek”, hiszen ezek kb. 333 aminosavból álló fehérjéket kódoló géneknek felelhetnek meg. Ez a két triviális extrém esetünk van: az 1-hosszú és a több száz hosszú szekvenciákról is tudjuk, hogy az előfordulási gyakoriságuk jellemző a genomra. Nemrég már 50-hosszú szekvenciákról is megmutatták, hogy metagenomokban baktériumok markerei lehetnek [13], viszont a kb. 10-hosszú DNS-szekvenciák gyakorisága és a genom vagy metagenom tulajdonságai között eddig még nem mutattak ki összefüggést. Azonban demonstráltam, hogy bizonyos legfeljebb 9 hosszú szekvenciák gyakorisága bélflóra-metagenomokban összefügg a diabétesszel (ld. [1]).

2.5. Agygráf

Újabb eredményeim azt mutatják, hogy nők és férfiak konnektómja (azaz agygráfja) különböző gráfelméleti tulajdonságokat mutat [2]. Ha adott egy diffúziós MR-felvétel egy alany agyáról, abból gráfot építhetünk, amely a különböző agyterületek összekötöttségét írja le. Először az agy képét fehér- és szürkeállományra bontjuk a frakcionális anizotrópia alapján, anatómiai szabályszerűségek figyelembe vételével. A frakcionális anizotrópia (FA) egy 0 és 1 közötti mennyiség. Ha egy voxelben (3D pixel) erős diffúzió mutatkozik egy kitüntetett tengely mentén, akkor közel lesz 1-hez, a 0 pedig azt jelenti, hogy minden irányban azonos intenzitású a diffúzió. A szürkeállományban általában kisebb az FA, mint a fehérállományban, mivel a fehérállományt idegrostok kötegei alkotják, jól definiált diffúziós irányokkal.

Miután az agyat szürke- és fehérállományra bontottuk, a parcellációs algoritmus azonosítja a ROI-kat (regions of interest, agyterületek). Ez úgy történik, hogy a képet egy referencia agyra diffeomorfáljuk, amit már előre kézzel parcelláltunk. A traktográfiai algoritmus párhuzamosan fut: ez a modul felelős az axonok lekövetéséért. A parcelláció és a traktográfia eredményeinek összekombinálásával egy gráfot kapunk, melynek csúcsai a ROI-k, élei pedig az ezeket a régiókat összekötő idegpályáknak felelnek meg. Többféle felbontást (agyatlaszt) is használhatunk a parcellációs fázisban, és ezáltal kevesebb vagy több csúcsa lesz az agygráfnak. A megfelelő atlasz kiválasztása mindig kompromisszum, hiszen a kevesebb ROI anatómiailag „értelmesebb”, viszont több ROI-val nagyobb gráfot kapunk, ami több információt tartalmaz, és matematikai szempontból érdekesebb lehet. A Connectome Mapping Toolkit nevű szoftvert használtuk az MRI képekből történő gráfépítésre. Ezután összehasonlítottuk a nők és férfiak agygráfjait.

Bár a nők és férfiak konnektómjai elég hasonlónak mutatkoznak, azt láttuk, hogy a női agygráfnak szignifikánsan több éle van, sőt, az agyféltekék közti élek aránya is nagyobb, mint a férfiaknál. Matematikai elemzés után azt láttuk, hogy a női agy jobb expander, és több feszítőfája van (ezek természetesen a nagyobb élszámmal is összefüggenek). Hangsúlyozzuk, hogy ezek a különbségek csupán az agy *összeköttetési struktúrájával* függenek össze, és nem mondanak semmit annak *működéséről*. A [2] cikkünk írja le részletesen ezeket az eredményeket.

Egy online 3D agygráf-modellt is készítettem, amit Budapest Reference Connectome-nak [3]

neveztem el. Ez egy konszenzus agygráfot jelenít meg, ami azt jelenti, hogy átlagoltam több száz ember agygráfját egyetlen gráffá, és ezt a gráfot egy interaktív webalkalmazásommal megjelenítettem.

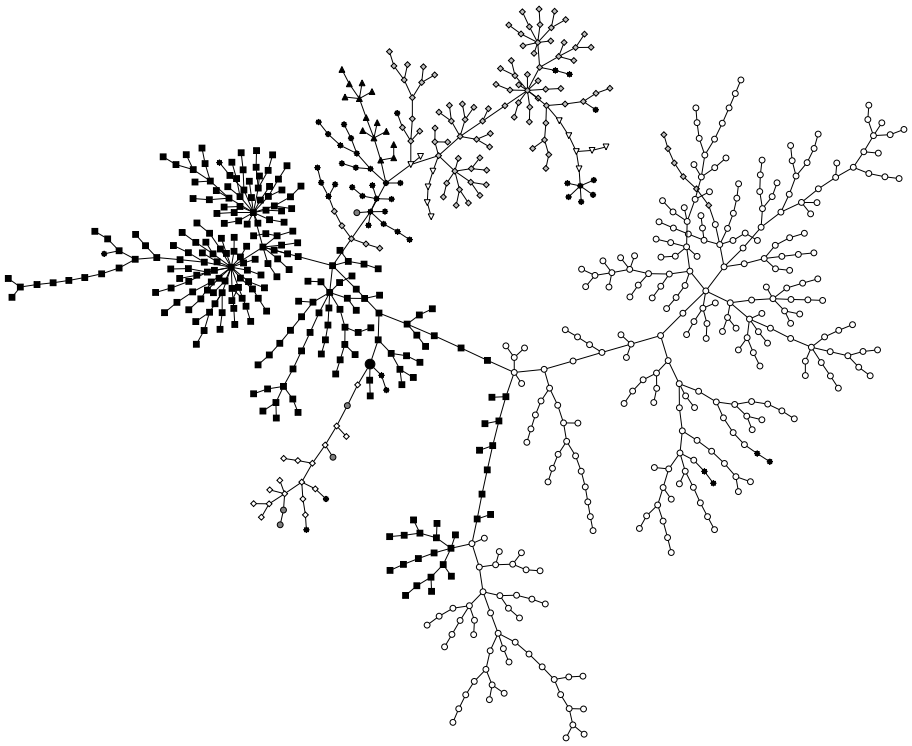
A Human Connectome Project adatbázisa lehetővé tette számunkra azt is, hogy érdekes felfedezéseket tegyünk a ROI-k méretével, pszichológiai tesztek eredményével, és kognitív pontszámokkal kapcsolatban. Ez az adattábla 527 sort tartalmazott (egyet a kísérlet minden alanyáról) és 451 oszlopot (attribútumot). Minden alanyra sok adat állt rendelkezésre, többek között demográfiai adatok, kognitív és mentális egészségi tesztek eredményei, pl. MMSE-pontszám (Mini Mental State Exam), NIH toolbox (pszichológiai és kogníciós tesztek), és NEO-FFI (ötdimenziós személyiségteszt). Az adatbázisba belevették a parcellációval kapott agyterületek méretét is: az adatbázis összeállítói a FreeSurfer szoftver parcellációs modulját futtatták le az MR-felvételekre. Kérgi régiókra a vastagság és a terület, kéreg alatti magvakra a térfogat szerepelt a táblában, tehát minden ROI-ra ismert volt annak térfogata. Ezt az adatbázist kibővítettük az általunk számolt gráfparaméterekkel.

Azt volt a célunk, hogy érdekes korrelációkat találjunk az attribútumok között. Másszóval, az attribútum-pároknak egy olyan részhalmazát akartuk kiválasztani, azaz egy gráfot az attribútumokon mint csúcsokon, amelynek élei bizonyos szempontból fontos korrelációknak felelnek meg. Ezt úgy értük el, hogy maximális súlyú feszítőfát számoltunk az attribútumokon, ahol minden él súlya a megfelelő korreláció abszolút értéke volt. Azért választottam ezt a módszert, mert ez egy hierarchikus klaszterezést is jelent az attribútumokon, egyúttal szűri a korrelációs mátrixban található információt. El akartuk kerülni, hogy a kapott gráfban körök legyenek, mivel a jelentős korreláció sokszor tranzitív. Ez a maximális súlyú feszítőfa megközelítés hasonló ahhoz, amit Mantegna et al. [4] használtak részvények hozamai közötti összefüggések felderítésére.

A feszítőfa pár olyan élet is tartalmazott, amikre számítottunk, de sok új és érdekes összefüggést is feltárt [11]. Egyrészt, az NIH Toolbox Emotion Domain elemei részfát alkottak, ami várható volt, de az érdekes dolog éppen az, ahogy ezek a csúcsok össze lettek kötve. Lényegében az történt, hogy a matematika segítségével klasztereztük az érzelmeket. Információt nyertünk arról, hogy hogyan függenek össze egymással. Pl. a stressz egy központi csúcs volt ebben a részfában, ami az élettel való elégedettséggel, az érzékelt önhatékonysággal, a szomorúsággal és a haraggal/-ellenségességgel volt összekötve. A fából az is kiderül, hogy a barátság elsősorban az érzelmi és nem az instrumentális támogatáshoz kapcsolódik. A ROI-méretek részfája pedig két jelentősebb központot tartalmazott: nem meglepő módon a bal és jobb agyfélteke térfogata volt ez a kettő, amelyek az egész agy térfogatától függenek, továbbá meghatározzák az egyes ROI-k méretét.

3. Összegzés

Remélem, hogy a kutatásom hozzájárult ahhoz, hogy az elérhető nagy mennyiségű bioinformatikai adatot valahogy képesek legyünk az emberiség javára fordítani. Ahhoz, hogy javíthassunk az



❍ Gráfparaméter ■ Agyterület mérete ♦ fMRI feladat pontszáma
▽ NIH Toolbox Cognition Domain ◊ NIH Toolbox Emotion Domain
● NEO Five-Factor Inventory
▲ Delayed Discounting Score ● Nem (nő/férfi) ✸ Egyéb

2. ábra. A korrelációs fa

emberek életminőségén, jobban meg kell értenünk testünk működését és a betegségeket. Hiszek abban, hogy a rengeteg szabadon elérhető adatot más perspektívából nézve, a matematikát és az informatikát felhasználva közelebb kerülhetünk ennek a célnak az eléréséhez.

Hivatkozások

[1] I. Alsmadi and I. Alhami, *Clustering and classification of email contents*, Journal of King Saud University - Computer and Information Sciences, 27 (2015), pp. 46 – 57.

[2] C. Bauckhage, *Numpy / scipy recipes for data science: k-medoids clustering*, 2009.

[3] J. P. González-Brenes and J. Mostow, *What and when do students learn? fully data-driven joint estimation of cognitive and student models.*

[4] R. N. Mantegna, *Hierarchical structure in financial markets*, The European Physical Journal B-Condensed Matter and Complex Systems, 11 (1999), pp. 193–197.

[5] B. Szalkai, *Generalizing k-means for an arbitrary distance matrix*, ArXiv e-prints http://arxiv.org/abs/1303.6001, (2013).

[6] B. Szalkai and V. Grolmusz, *Nucleotide 9-mers characterize the type ii diabetic gut metagenome*, Genomics, (2016), pp. –.

[7] B. Szalkai, V. K. Grolmusz, V. I. Grolmusz, and C. Against Major Diseases, *Identifying Combinatorial Biomarkers by Association Rule Mining in the CAMD Alzheimer's Database*, ArXiv e-prints, (2013).

[8] B. Szalkai, C. Kerepesi, B. Varga, and V. Grolmusz, *The budapest reference connectome server v2.0*, Neuroscience Letters, 595 (2015), pp. 60 – 62.

[9] B. Szalkai, I. Scheer, K. Nagy, B. G. Vértessy, and V. Grolmusz, *The metagenomic telescope*, PLoS ONE, 9 (2014), pp. 1–9.

[10] B. Szalkai, B. Varga, and V. Grolmusz, *Graph theoretical analysis reveals: Women's brains are better connected than men's*, PLoS ONE, 10 (2015), p. e0130045.

[11] B. Szalkai, B. Varga, and V. Grolmusz, *Mapping Correlations of Psychological and Connectomical Properties of the Dataset of the Human Connectome Project with the Maximum Spanning Tree Method*, ArXiv e-prints, (2016).

[12] B. Szigeti, P. Gleeson, M. Vella, S. Khayrulin, A. Palyanov, J. Hokanson, M. Currie, M. Cantarelli, G. Idili, and S. Larson, *Openworm: an open-science approach to modelling caenorhabditis elegans*, Frontiers in Computational Neuroscience, 8 (2014).

[13] Q. Tu, Z. He, and J. Zhou, *Strain/species identification in metagenomes using genome-specific markers.*, Nucleic Acids Res, 42 (2014), p. e67.

[14] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, *The structure of the nervous system of the nematode caenorhabditis elegans*, Philosophical Transactions of the Royal Society of London B: Biological Sciences, 314 (1986), pp. 1–340.

Szalkai Balázs publikációi

[1] B. Szalkai and V. Grolmusz, *Nucleotide 9-mers characterize the type {II} diabetic gut metagenome*, Genomics, (2016), pp. –.

[2] B. Szalkai, B. Varga, and V. Grolmusz, *Graph theoretical analysis reveals: Women's brains are better connected than men's*, PLoS ONE, 10 (2015), p. e0130045.

[3] B. Szalkai, C. Kerepesi, B. Varga, and V. Grolmusz, *The Budapest Reference Connectome Server v2.0*, Neuroscience Letters, 595 (2015), pp. 60 – 62.

[4] B. Szalkai, I. Scheer, K. Nagy, B. G. Vértessy, and V. Grolmusz, *The metagenomic telescope*, PLoS ONE, 9 (2014), pp. 1–9.

[5] B. Szalkai, *An implementation of the relational k-means algorithm*, ArXiv e-prints http://arxiv.org/abs/1304.6899, (2013).

[6] B. Szalkai, *Generalizing k-means for an arbitrary distance matrix*, ArXiv e-prints http://arxiv.org/abs/1303.6001, (2013).

[7] B. Szalkai, V. K. Grolmusz, and V. I. Grolmusz, *Identifying Combinatorial Biomarkers by Association Rule Mining in the CAMD Alzheimer's Database*, ArXiv e-prints http://arxiv.org/abs/1312.1876, (2013).

A tézisben nem szereplő publikációk

[1] B. Szalkai, B. Varga, and V. Grolmusz, *The graph of our mind*, ArXiv e-prints http://arxiv.org/abs/1603.00904, (2016).

[2] Cs. Kerepesi, B. Szalkai, B. Varga, and V. Grolmusz, *Does the Budapest Reference Connectome Server Shed Light to the Development of the Connections of the Human Brain?*, ArXiv e-prints http://arxiv.org/abs/1509.05703, (2015).

[3] Cs. Kerepesi, B. Szalkai, B. Varga, and V. Grolmusz, *Comparative Connectomics: Mapping the Inter-Individual Variability of Connections within the Regions of the Human Brain*, ArXiv e-prints http://arxiv.org/abs/1507.00327, (2015).

[4] C. Kerepesi, B. Szalkai, and V. Grolmusz, *Visual analysis of the quantitative composition of metagenomic communities: the AmphoraVizu webserver.*, Microb Ecol, (2014).

[5] D. Bánky, B. Szalkai, and V. Grolmusz, *An intuitive graphical webserver for multiple-choice protein sequence search*, Gene, 1 (2014), pp. 152–153.