DEBRECENI EGYETEM, INFORMATIKAI KAR

4028 Debrecen, Kassai út 26., 4002 Debrecen, Pf. 400.
☎ 52/512-700, email: hivatal@inf.unideb.hu

Bírálói vélemény

Miklós István: *Computational complexity of counting and sampling problems in bioinformatics / Leszámlálások és mintavételezések bonyolultságelmélete a bioinformatikában*

című doktori munkájáról

1. Témaválasztás

A génszerkezetek vizsgálata, illetve tágabb értelemben a bioinformatika napjaink egyik különösen fontos kutatási területe, amit jól illusztrál, hogy a Google Scholar keresője a „bioinformatics" kulcsszóra csak az idei évre 27 700 találatot ad. Az egyes problémák esetén azonban, még napjaink ugrásszerű számítási kapacitásnövekedése mellett is, nagyon nem mindegy azok számításigénye. Ebben segítenek Miklós István bonyolultságelméleti eredményei, amelyek a bioinformatikában és a tágabb értelemben vett élettudományokban felmerülő leszámlálási és főleg MCMC alapú mintavételezési feladatok egy széles spektrumát fedik le. Ennek kapcsán jegyezném meg, hogy a Google Scholar a „computational complexity" kulcsszóra is 26 100 darab 2024 évi találatot ad, így a fentiek alapján nehezen kérdőjelezhető meg a kutatási téma korszerűsége és fontossága.

2. A doktori mű felépítése, tartalma, értékelése

Az értekezés felépítése, tartalma

Miklós István MTA doktori értekezését angol nyelven készítette el, annak terjedelme 266 számozott oldal, amiket megelőz egy 3 oldalas Előszó és a Tartalomjegyzék. A doktori mű összesen 12 fejezetre tagolódik, amiből az 1. fejezet tartalmazza a jelöltnek a maradék 11 fejezetben leírt eredményei megértéséhez szükséges alapvető fogalmakat és állításokat. Ez utóbbi 11 fejezet a vizsgált problémakörök alapján 4 különálló részbe van csoportosítva. Ezt követi a 12 oldalas, összesen 126 tételből álló Irodalomjegyzék, ahol külön blokkban szerepelnek a szerzőnek a dolgozat alapját adó publikációi. A művet egy 3 oldalas Tárgymutató zárja.

Nem lévén járatos sem az elméleti számítástudományban, sem pedig a gráfelméletnek a disszertációban tárgyalt részeiben, számomra az 1. fejezet különösen hasznosnak bizonyult, gyakran lapoztam ide vissza a későbbi fejezetek olvasásakor. A genomátrendeződések kapcsán felmerülő struktúrák jobb megértésében például sokat segítettek a jól megválasztott példák. Ugyancsak erénye a fejezetnek, hogy a végén a szerző röviden összefoglalja a későbbi fejezetekben tárgyalt 4 nagy problémakört. Megjegyezném azonban,

hogy ebből a fejezetből, hacsak azok nem a szerző önálló és új eredményei, teljes mértékben kihagytam volna a bizonyításokat, különös tekintettel a teljesen triviális 26. tétel igazolására.

Az első tárgyalt témakör a polinom időben megoldható leszámlálási problémákról szól. Ezen belül a 2. fejezetben a szerző először ismertet egy univerzális dinamikus programozási algoritmust (algebrai dinamikus programozás), és megmutatja, hogyan származtatható belőle az RNS térszerkezetek Boltzman eloszlása momentumainak kiszámításra alkalmas algoritmus, illetve másik speciális esetként megadja a Baum-Welch tréning egy memóriahatékony változatát. A két bemutatott és korábban külön-külön publikált ([121], illetve [120]) példa jól mutatja az általános technika alkalmazhatóságát, de én eltekintettem volna a 49. és 50. tétel, valamint az 52. lemma meglehetősen triviális bizonyításainak közlésétől. A [126] és [122] cikkek egyes részein alapuló 3. fejezetben a szerző a legtakarékosabb szcenáriók és a legtakarékosabb medián problémakörével foglalkozik a szimpla vágás vagy kötés (SCJ) genomátrendeződési modellben. Felhasználva, hogy az előbbiek az alternáló permutációkkal, az utóbbiak pedig legfeljebb 2 fokú egyszerű gráfok párosításával állnak kapcsolatban, igazolja, hogy mindkét probléma FP-beli.

A következő rész sztochasztikus leszámlálási kérdésekkel foglalkozik, ahol a szerző speciális problémákhoz tartozó Markov láncok gyors konvergenciáját igazolja. A [106, 107, 109] és [111] cikkek eredményeit összesítő 4. fejezet a páros fokszámsorozatok realizációin értelmezett ún. switch Markov láncról szól. Itt a jelölt elsőként azt bizonyítja, hogy páros fokszámsorozatok Tyshkevich dekompozíciója esetén a switch Markov lánc gyors konvergenciájának igazolásához elegendő bizonyítani a komponensekhez tartozó láncok gyors konvergenciáját. Ezek után megmutatja, hogy a switch Markov lánc gyorsan konvergál P-stabil páros fokszámsorozatok realizációin is, valamint igazolja, hogy a lineárisan korlátozott fokszámsorozatok P-stabilak. Végezetül megad egy speciális nem P-stabil fokszámsorozat osztályt is, amin igazolható a gyors konvergencia, valamint ismerteti a fejezetben leírt eredmények néhány további általánosítását. Az 5. fejezet a [104] és [110] dolgozatok eredményeit tartalmazza. A szerző először azt mutatja meg, hogy ha egy Markov lánc állapottere faktorizálható, a komponensekre megszorított láncok gyorsan konvergálnak, valamint a komponensek között lépkedő Markov lánc is gyorsan konvergál, akkor ez a tulajdonság teljesül az eredeti Markov láncra is (74. tétel). Ezt használja majd a fejezet fő tételének (78. tétel) igazolására, ami kimondja, hogy tetszőleges grafikus közös fokszám-mátrix kiegyensúlyozott realizációin az RSO (restricted switch operation) Markov lánc gyorsan konvergál, bizonyítva persze azt is, hogy az említett kiegyensúlyozott realizáció minden grafikus közös fokszám-mátrix esetén létezik. A disszertáció II. részét a [125] cikken alapuló 6. fejezet zárja, melynek 3 lépésben bizonyított fő eredménye, hogy a dupla vágás és kötés (DCJ) genomátrendeződési modellben a legtakarékosabb szcenárióra létezik FPRAS és FPAUS algoritmus.

A III. rész negatív eredményekre koncentrál, olyan genomátrendezésekhez kapcsolódó Markov láncokat mutat be, amelyek lassan konvergálnak. A [119] eredményeit bemutató 7. fejezet a reverziós (REV) modellel foglalkozik, ahol a szerző ezt a tulajdonságot igazolja a

legtakarékosabb REV szcenáriókon értelmezett egyetlen ismert irreducibilis Markov láncra. A fejezetet egy értékes kitekintés zárja, egyrészt magyarázatot adva a lassú konvergencia okára, másrészt lehetséges irányokat sorolva fel a probléma kezelésére. A 8. fejezet a [123] cikken és a [126] cikk 4. fejezetén alapul. Itt a szerző megmutatja, hogy az SCJ modell mellett az evolúciós fán vizsgált legtakarékosabb szcenáriók és ugyanezen genomátrendeződési séma mellett a legtakarékosabb medián szcenáriók #P-teljes problémák. A bizonyítások azon az ismert állításon alapulnak, hogy a 3 konjuktív normál formákra vonatkozó #3SAT probléma #P-teljes, és polinom időben visszavezethető mindkét itt vizsgált leszámlálási problémára.

A IV. rész első fejezetében (9. fejezet) a jelölt egy a [103] cikkben is leírt eredményt ismertet. Megfogalmaz egy sejtést a végtelen hely modellből származó genomok REV modell alatti legtakarékosabb szcenárióiról, amiből, ha igaz, következik, hogy megadható ezen szcenáriókon egy olyan irreducibilis Markov lánc, ami esetén a Metropolis-Hastings hányados reciprokára van polinom felső korlát. A sejtést rekurzív módon igazolja olyan permutációkra, melyek átfedési gráfja lineáris. A [124] cikken alapuló 10. fejezet unikromoszomális lineáris genompárok közötti legtakarékosabb REV szcenáriók mintavételezését a legtakarékosabb DCJ szcenáriók mintavételezésére vezeti vissza egy speciális energiafüggvény segítségével. A szerző egy olyan párhuzamosan temperált, egyszerre kevés DCJ transzformációt módosító Markov láncokon alapuló MCMC technikát mutat be, ahol a kiinduló (legmelegebb) Markov láncban a cél eloszlás a DCJ modellben legtakarékosabb szcenáriókon vett egyenletes eloszlás (amire van FPAUS), a leghidegebb lánc cél eloszlásában pedig a legtakarékosabb REV szcenáriók legalább 1/2 valószínűséggel jelennek meg. A 11. fejezetben a jelölt egy Gibbs mintavételezőt mutat be a legtakarékosabb fa címkézés problémára az SCJ modell mellett, melyről igazolja, hogy egy irreducibilis Markov lánc. Bár ezen lánc gyors konvergenciáját nem sikerült bizonyítani, valós példákon tesztelve (8 faj genomjai, lásd [122]) gyorsan csökkenő autokorrelációs diagrammokat eredményezett. Véleményem szerint a disszertációban is szerencsés lett volna bemutatni a [122] példáit és az ott közölt 2. ábrát. Az utolsó, 12. fejezet első részében, ami a [102] eredményeit ismerteti, a szerző előbb szükséges és elégséges feltételt ad arra, hogy egy félig reguláris páros fokszámsorozatmátrixnak mikor létezik realizációja. Ezután megmutatja, hogy tetszőleges két realizáció átvihető egymásba olyan realizációk sorozatával, amik legfeljebb 3 színt változtatnak meg. Ezen észrevétel adja az alapját az 1. algoritmussal definiált Markov láncnak, amihez tartozó Metropolis-Hastings hányados reciprokára (egy generált állapot elfogadásnak várható várakozási ideje) megadható polinom felső korlát. Végezetül, a 12. fejezet hátralevő része páros gráfok *k*-élszínezésével foglalkozik és egy a disszertáció benyújtása után megjelent cikket (Hong, L., Miklós, I., A Markov chain on the solution space of edge colorings of bipartite graphs. *Discrete Applied Mathematics* **332** (2023), 7-22, lásd még [112]) ismertet. A jelölt itt egy olyan irreducibilis Markov láncot konstruál, melynek mind az átmérője, mind pedig a kapcsolódó Metropolis-Hastings hányados reciproka felülről a gráf méretének és *k*-nak polinomjával korlátozható.

4028 Debrecen, Kassai út 26., 4002 Debrecen, Pf. 400.
☎ 52/512-700, email: hivatal@inf.unideb.hu

Nyelvezet, formai követelmények

A kiváló angolsággal megírt értekezés formálisan teljes mértékben megfelel az MTA III. Osztálya által meghatározott követelményeknek, de azért jó lett volna, ha a jelölt figyelembe veszi az alábbi ajánlást. „Kerülendő, hogy az értekezés a pályázó összegyűjtött műveivé váljon. Sokkal inkább arra kell törekedni, hogy egységes témát öleljen fel, és csak a fontos eredményeket tartalmazza (akár egy monográfia). Bőven elegendő kb. 100 oldal összterjedelem." A mindent egybevetve 275 oldalas mű 2005 és 2023 között megjelent 16 cikk eredményeit foglalja össze, az Irodalomjegyzékében a szerző összesen 25 munkája szerepel, beleértve a [115] monográfiát. Pozitívum, hogy mivel a fent említett 16 cikk mindegyike társszerzős, a jelölt minden fejezet elején egyértelműsíti, mi volt az ő hozzájárulása a közölt eredményekhez. Negatívumként említeném, hogy a doktori mű nem tekinthető gondosan szerkesztettnek. Csak pár észrevétel a teljesség igénye nélkül:

1. Az 1. fejezet sztochasztikus részekben találtam némi inkonzisztenciát a jelölésekben (pl. a 30. definícióban nem világos, hogy a θ a stacionárius eloszlást jelöli, ami a 32. definícióban már π, bár ez utóbbi explicit nem lett kimondva). Ugyanitt nem minden fogalom (pl. reverzibilitás) lett definiálva.

2. 102 definíció 4. pontjánál nem világos mik a p_j valószínűségek.

3. A 125. tételből hiányzik az állítás. Az eredeti cikkben „can be defined" szerepel.

4. A 202. oldalon levő 139. definíció tulajdonképpen szerepel már a 25. oldalon is.

5. A [126] cikk a folyóirat honlapján az itt megadottaktól eltérő adatokkal szerepel (más a szerzők sorrendje, a cikk címe és az oldalszám).

Emellett engem komoly kihívás elé állított a 12.4 fejezet, ahol a Markov lánc átmeneteit leíró 2.5 oldalas 157. definíció láttán azért megremegett picit a lábam. Lehet, hogy az ilyen esetekben (pl. az 1. algoritmusnál is) egy folyamatábra sokat segített volna. Természetesen azonban sem az említett hiányosságok, sem az egyéb elírások, sem pedig a hivatkozások meglehetősen kevert stílusa nem csökkentik a mű tudományos értékét.

A disszertáció hitelessége

Miklós István publikációs tevékenysége figyelemreméltó, az MTMT oldala 67 nemzetközi folyóiratban megjelent cikket mutat, melyekre összesen 1823 független hivatkozás érkezett. A disszertációban bemutatott eredmények mindegyike rangos és a téma szempontjából releváns nemzetközi folyóiratban jelent meg (SJR D1: 8; Q1: 6; Q2: 8; egyéb: 2), hitelességükhöz kétség nem férhet.

4028 Debrecen, Kassai út 26., 4002 Debrecen, Pf. 400.
☎ 52/512-700, email: hivatal@inf.unideb.hu

3. A doktori műhöz benyújtott tézisfüzet értékelése

A tézisfüzetben az eredmények megértéséhez szükséges alapvető fogalmak ismertetése után a doktori mű 4 fejezetének megfelelő csoportosításban a jelölt összesen 16 tézist fogalmaz meg. Véleményem szerint mind a 16 tézispont megfelelően igazolt és új tudományos eredményként fogadható el.

4. Kérdések

1. Számomra nem érthető a 65. tétel bizonyítása utáni példa, de lehet, hogy ennek csak a témában való járatlanságom az oka. A szerző azt szeretné illusztrálni, hogy a (4.8) egyenlőtlenség éles. Példaként az $c_1 = n/4$, $c_2 = 3n/4$, $d_1 = m/4$ és $d_2 = 3m/4$ határokat veszi és megmutatja, hogy pl. $d_2 = (3/4 + \varepsilon)n$ esetén van olyan páros fokszámsorozat, ami nem P-stabil (d_1 és d_2 esetén gondolom a dolgozatban szereplő n szorzó csak elírás). Mi garantálja, hogy az eredetileg megadott határok mellett a fokszámsorozatok P-stabilak, hiszen ezekre a (4.9) egyenlőtlenség nem feltétlenül teljesül (pl. ha $n=m=8$)?

2. Lehet, hogy ez a kérdés is triviális. Miért létezik mindig a 148. lemmában szereplő V' halmaz? Ugyanez a kérdés merül fel a 151. definícióban szereplő megadott tulajdonságú c_s és c_e csúcsok (színek) létezésével kapcsolatban.

3. Milyen gyakorlati bioinformatikai problémákhoz köthető a 12. fejezetben ismertetett páros gráfok k-élszínezése?

5. Javaslat

A fent leírtak alapján Miklós István értekezésében bemutatott kutatómunkát nemzetközi viszonylatban is magas szintűnek tartom, eredményei messzemenően elegendőek az MTA doktora cím megszerzéséhez. Javaslom a doktori mű nyilvános vitára bocsátását és sikeres védés esetén az MTA doktora cím odaítélését.

Debrecen, 2024. augusztus 12.

Baran Sándor
egyetemi tanár
MTA doktora