Leszámlálások és mintavételezések bonyolultságelmélete a bioinformatikában

MTA doktori értekezés tézisei

Miklós István

Rényi Alfréd Matematikai Kutatóintézet

Budapest, 2022

1. Bevezetés

Minden „Létezik-e...?”, „Van-e olyan...?” kérdés természetes velejárói a ”Hány darab van...?”, „Hány olyan létezik...?” kérdések. Például ha adott nem-negatív egész számok sorozata, akkor megkérdezhetjük, hogy létezik-e egyszerű gráf, amelynek a fokai a megadott számok. Ugyanígy megkérdezhetjük, hogy hány olyan gráf van, amelynek a fokai a megadott számok. Beszélhetünk ezen *döntési* és *leszámlálási* kérdések bonyolultságelméletéről, azaz arról, hogy ezen kérdéseket megválaszoló számítógépes algoritmusnak a futási ideje hogyan növekszik a bemenő adatok mennyiségével. A leszámlálási kérdések mindig legalább annyira bonyolultak, mint a döntési kérdések. Valóban, az egy trivialitás, hogy a „Létezik-e...?” kérdésre a válasz „Igen” akkor és csak akkor, ha a hozzá tartozó „Hány darab van...?” kérdésre a válasz egy 0-nál nagyobb szám. Sok esetben a triviális döntési kérdésekhez tartozó leszámlálási kérdések is lehetnek nehezek.

Sokszor a fő cél nem az, hogy megmondjuk azt, hogy hány matematikai objektum (pl. a fenti példában: egyszerű gráf) létezik amely teljesít adott feltételeket, hanem ezen objektumokat akarjuk mintavételezni adott eloszlás szerint. Az adott eloszlás sokszor az egyenletes, de statisztikai kérdések esetén más eloszlások, pl. Boltzmann, Bayes statisztikai, stb., is szóba jönnek. Pl. ökológiai állapotfelméréseknél készíthetünk egy páros gráfot, amely gráf egyik pontosztálya az adott élőhelyek (pl. a Vanuatu szigetvilág szigetei), a másik pontosztály pedig a fajok (pl. a Vanuatu szigetvilág madárfajai) [29]. Két pont között akkor van él, ha az adott faj az adott élőhelyen megtalalálható. Ezekben a gráfokban mintázatokat kereshetünk. Az egyik ilyen minta a kölcsönös elkerülés, ami egy 2×2-es feszített részgráfon egy 1-faktor. A kölcsönös elkerülés a fajok közötti kompetíciót jelzi, és az ökológusok kíváncsiak arra, hogy egy ökoszisztémában milyen mértékű a fajok közötti kompetíció. A kölcsönös elkerülési mintázatok száma egy ökoszisztémában önmagában nem elég informatív. Így klasszikus statisztikai hipotézisvizsgálat eszközéhez szeretnénk nyúlni. A H_0 hipotézis az, hogy az ökoszisztémában nincs kompetíció, a kölcsönös elkerülési mintázatok száma véletlenszerű. Ehhez egy háttéreloszlást akarunk gyártani, azaz olyan random páros gráfokat, amelyben minden faj annyira elterjedt, mint az adott ökoszisztémában, és minden élőhely annyira fajgazdag, mint az adott ökoszisztémában. Azaz adott fokszámsorozatú páros gráfokat akarunk véletlen generálni. Ezen véletlen gráfok segítségével legyárthatjuk a kölcsönös elkerülési mintázatok számának a H_0 szerinti eloszlását. A H_0 hipotézist akkor

tartjuk meg, ha a felmért ökoszisztémában a kölcsönös elkerülési mintázatok száma nem extrém nagy, különben a H_0 hipotézist elvetjük és a H_1 alternatív hipotézist támogatjuk mely szerint a fajok elterjedése nem véletlenszerű, van közöttük kompetíció. Egy ilyen statisztikai vizsgálatot akkor lehet elvégezni, ha van hatékony algoritmus a háttéreloszlás legenerálására, azaz van olyan gyors algoritmus, amely adott fokszámsorozatú páros gráfokat mintavételez majdnem egyenletesen. A majdnem egyenletes itt azt jelenti, hogy az egyenletes eloszlástól való eltérés szisztematikus hibája elhanyagolhatóan kicsi a mintavételezési hibához képest. Valóban, egy ilyen mintavételezés a gyakorlat számára tökéletesen megfelel. Fontos kérdés, hogy mely esetekben van majdnem egyenletes mintavételezésre gyors algoritmus. Kiderül, hogy standard bonyolultságelméleti feltételezések mellett néha közelítőleg egyenletes mintavételezésre sincs gyors algoritmus.

Mint látható, a leszámlálások és mintavételezések bonyolultságelmélete változatos. Vannak olyan problémák, amelyek polinom időben egzaktul megoldhatóak, vannak, amelyek egzakt megoldása nehéz (bonyolultságelméleti értelemben), de random algoritmusokkal jól approximálhatóak, és vannak olyan nehéz problémák, amelyek még csak jól sem approximálhatóak (standard bonyolultságelméleti feltételezésekkel élve, nevezetesen, feltéve, hogy RP $\neq$ NP) [20]. Valójában a problémakör bonyolultságelméleti klasszifikációja még ennél is gazdagabb. A polinom időben megoldható problémák között vannak olyanok, amelyek *monoton komputációval* megoldhatóak (azaz, csak összeadások és szorzások segítségével), vannak olyanok, amelyek bár polinom időben megoldhatóak, pl. kivonás műveleteket is megengedve, de exponenciális idő kell a kiszámolásukhoz ha csak az összeadás és szorzás műveletek megengedettek [16]. Vannak olyan leszámlálási problémák, amelyek csak bonyolult tenzoralgebrai számításokkal, ún. *holografikus redukcióval* oldhatóak meg polinom időben [34]. Ismerünk olyan nehéz leszámlálási problémát, amely detreminisztikusan jól approximálható. Dyers és munkatársai bevezették a #BIS-teljes bonyolultságelméleti osztályt, amely teljes abban az értelemben, hogy vagy minden #BIS-teljes probléma sztochasztikusan jól approximálható vagy egyik sem. Az a sejtés, hogy ezek nem jól approximálható problémák, de a nem-approximálhatóság nem bizonyítható a #3SAT nem-approximálhatóságából kiindulva, approximációtartó polinom redukcióval [5]. És akkor még nem beszéltünk arról, hogy a döntési problémákhoz hasonlóan leszámlálási problémáknál is lehetséges ún. köztes bonyolultságelméleti osztályok létezése, azaz olyan problémák, amelyek szubexponenciális, de szuprapolinomiális időben oldhatóak meg. A

gráfok automorfizmusainak a száma gyaníthatóan egy ilyen probléma [3].

A kutatásaim és így az értekezésem fő célja nem új bonyolultságelméleti osztályok bevezetése, hanem a bioinformatikában felmerülő leszámlálási és mintavételezési problémák bonyolultságának a meghatározása. Az esetek többségében ez annak az eldöntését jelenti, hogy a probléma polinom időben megoldható vagy #P-teljes és ha #P-teljes, akkor jól approximálható-e vagy nem is approximálható jól, bár mint az előző paragrafusban láthattuk, a teljes kép ennél bonyolultabb is lehet. A Jerrum-Vailiant-Vazirani tétel értelmében nagyon sok esetben a majdnem egyenletes mintavételezésből már következik a jól approximálhatóság is [17]. Ezért ismerten vagy gyanítottan #P-teljes problémák esetében az elsődleges cél a majdnem egyenletes mintavételezés elérése. Ezt nagyon sokszor gyorsan keveredő Markov láncokkal lehet elérni, így a munkám tekintélyes részét Markov láncok tervezése és azok gyors konvergenciájának a bizonyítása teszi ki.

Az értekezésem egy bevezető után négy részből áll. Az első részben polinom időben megoldható leszámlálási problémákat tárgyalok. A második részben jól approximálható problémákat mutatok be. A harmadik rész negatív eredményeknek van szentelve. Megadok #P-teljességi bizonyítást, bizonyítok egy nem-approximálhatóságot valamint bebizonyítom egy széles körben használt Markov lánc lassú keveredését. Végül az utolsó részben Markov láncok tervezésében elért részeredményeket mutatok be. A harmadik részben bemutatott negatív eredményeket figyelembe véve válik érthetővé, hogy ezek a részeredmények miért jelentenek előrelépést.

2. Alapok

2.1. Rövid történeti áttekintés

Indiában Fibonacci korát jóval megelőzően, már az időszámításunk előtti V.–III. században felfedezték a Fibonacci számokat. Nevezetesen, észrevették, hogy 1 és 2 hosszúságú szótagokból F_n féleképpen lehet n hosszúságú prozódiákat készíteni. Spóradikus enumeratív kombinatorikai eredmények születtek egészen a modern korig (pl. a Striling számok a XVIII. században, a Catalan számok a XIX. század elején), de az első bonyolultságelméleti kérdés a mátrix permanensek kiszámolásánál vetődött fel. Míg a determináns kiszámolására ismert volt a Gauss elimináció gyors algoritmusa, a permanens kiszámolására nem találtak hatékony módszert. Pólya tette fel a kérdést,

hogy lehetséges-e egy M mátrix előjeleit úgy megváltoztatni, hogy az így kapott M' mátrix determinánsa megegyezzen M permanensével [30], és erre a negatív választ Szegő adta meg [31]. Valiant vezette be a #P-teljes bonyolultságelméleti osztályt, amelyről tudjuk, hogy legalább olyan nehéz, mint az NP-nehéz, és mutatta meg, hogy a permanens kiszámolása #P-teljes [32]. Egy következő cikkben pedig számos természetes leszámlálási problémáról mutatta meg Valiant, hogy #P-teljesek [33].

Jerrum, Valiant és Vazirani bizonyította be, hogy bármely önhasonló leszámlálási probléma sztochasztikusan jól approximálható akkor és csak akkor, ha a leszámolandó matematikai objektumok polinom időben majdnem egyenletesen mintavételezhetőek [17]. Karzanov és Khaciyan adott meg egy gyorsan keveredő Markov láncot, amely segítségével egy $\mathcal{P} = (A, \preceq)$ véges részbenrendezett halmaz lineáris kiterjesztései (teljes rendezései) majdnem egyenletesen mintavételezhetőek [18]. Mivel a lineáris kiterjesztések száma önhasonló leszámlálási probléma, így a Jerrum-Valiant-Vazirani tétel értelmében sztochasztikusan jó approximálható, azaz az ún. FPRAS osztályban is van. Brightwell és Winkler mutatta meg, hogy a lineáris kiterjesztések száma #P-teljes leszámlálási probléma [4]. Ez meglepő annak a tekintetében, hogy a hozzá tartozó döntési kérdés triviális. Valóban, minden véges részbenrendezett halmaz kiterjeszthető teljes rendezéssé. Így a „Létezik-e a $\mathcal{P} = (A, \preceq)$ véges részbenrendezett halmaznak teljes rendezéssé kiterjesztése?" kérdésre a válasz „Igen", és ehhez be sem kell olvasni a $\mathcal{P} = (A, \preceq)$-t leíró adatokat. A lineáris kiterjesztések száma volt az első leszámlálási probléma, amelyről megmutatták, hogy a #P-teljes és az FPRAS osztályok metszetében van.

Jerrum és Sinclair adott meg metódusokat, amelyekkel Markov láncok gyors konvergenciája bizonyítható [14]. Ennek segítségével számos #P-nehéz problémáról mutatták meg, hogy úgyszintén FPRAS-ban vannak, ezek közül talán a nem-negatív mátrixok permanensének a kiszámítása a legnevezetesebb [15].

Markov láncokat régóta használnak tudományos számításokban. Metropolis és munkatársai adtak meg egy általános algoritmust, amely segítségével olyan Markov lánc konstruálható, amely adott eloszláshoz konvergál [19]. Hastings népszerűsítette ezt a metódust biostatisztikai alkalmazásokban [12]. Azóta ez a módszer Metropolis-Hastings algoritmus néven ismert. Bioinformatikai problémákban is gyakran használt a Metopolis-Hastings algoritmus, de lényegében nem vizsgálták elméleti úton a Metropolis-Hastings algoritmussal konstruált Markov láncok konvergenciasebességét. Helyette csak di-

agnosztikus módszereket alkalmaztak, pl. loglikelihood nyom vagy autokorrelációs diagram alapján mauálisan állapítottak meg konvergenciasebességet. Ezek a módszerek csak a lassú konvergenciát képesek empirikusan bizonyítani a gyors konvergenciát nem.

Jerrum, Valiant és Vazirani is észrevette már a következőt. Ha élünk azzal a standard bonyolultságelméleti feltételezéssel, hogy RP≠ NP, abból már következik az, hogy még csak nagyon közelítőleg egyenletesen sem lehet irányított gráfokból köröket mintavételezni [17]. Ez meglepő lehet, mivel polinom időben eldönthető, hogy létezik-e kör irányított gráfban. Papadimitriou teteléből következik az is, hogy nincs BPP algoritmus a Hamilton kör eldöntésére irányított gráfokban (még mindig feltéve, hogy RP ≠ NP), és ebből következik az is, hogy FPRAS algoritmus sincs a körök számának a becslésére irányított gráfokban. A '90-es évektől kezdődően számos leszámlálási problémáról mutatták meg, hogy nem jól approximálható (feltéve, hogy RP≠NP), ezek közül talán az egyik legmeglepőbb, hogy a monoton 2CNF formulák kielégítéseinek a száma sem jól approximálható [5].

A '90-es évektől kezdődően a leszámlálások és mintavételezések bonyolultságelmélete sokat fejlődött, de a bioinformatikai problémákban felmerülő leszámlálási és mintavételezési problémák bonyolultságelméleti vizsgálata a mai napig gyerekcipőben jár.

2.2. Bonyolultságelméleti fogalmak, tételek

A tézisekben feltételezzük, hogy az olvasó ismeri a döntési problémák főbb osztályait, nevezetesen a P, NP, NP-nehéz, NP-teljes és RP osztályokat, így ezek definiálásától eltekintünk.

A továbbiakban leszámlálási és mintavételezési problémákról beszélünk.

1. Definíció. *A* #P *osztályt úgy definiáljuk, mint azon leszámlálási problémák osztálya, amely NP-beli problémák valamely polinom időben ellenőrizhető tanuinak a számát kérdezi meg. Egy probléma* #P-nehéz*, ha minden #P-beli probléma polinom időben visszavezethető rá. A* #P-teljes *problémák osztálya a #P-nehéz és #P osztályok metszete. Azon #P-beli pronlémák, amelyek polinom időben megoldhatóak* FP*-ben (*Function Polynomial*) vannak.*

Valiant megmutatta, hogy a páros gráfok teljes párosításainak a száma #P-teljes probléma. Vannak olyan #P-teljes problémák, amelyek jól approximálhatóak és majdnem egyenletesen mintavételezhetőek. Ennek a definícióit adjuk meg alább.

2. Definíció. *Egy #A leszámlálási probléma az* FPRAS (Fully Polynomial Randomized Approximation Scheme) *osztályban van, ha létezik olyan algoritmus, amely minden #A-beli x feladatra és $\varepsilon, \delta > 0$ számokra olyan $\hat{f}$ becslést ad az x feladat tényleges f eredményére, amelyre teljesül, hogy*

$$P\left(\frac{f}{1+\varepsilon} \leq \hat{f} \leq f(1+\varepsilon)\right) \geq 1-\delta,$$

és az algoritmus futási ideje $O(poly(|x|, 1/\varepsilon, -\log(\delta)))$. Az ezen tulajdonságokkal rendelkező algoritmust úgyszintén FPRAS algoritmusnak hívjuk.

Az FPAUS osztály definiálásához először a teljes variációs távolságot definiáljuk.

3. Definíció. *Az X (diszkrét) téren adott p és π eloszlások közötti* teljes variációs távolság

$$d_{TV}(p, \pi) := \frac{1}{2} \sum_{x \in X} |p(x) - \pi(x)|.$$

4. Definíció. *Egy #A leszámlálási probléma az* FPAUS (Fully Polynomial Almost Uniform Sampler) *osztályban van, ha létezik olyan algoritmus, amely minden #A-beli x feladatra és $\varepsilon > 0$ számra az x feladat egy véletlen megoldását generálja egy p eloszlásból, amely p eloszlásra teljesül a*

$$d_{TV}(p, U) \leq \varepsilon,$$

ahol U a megoldások egyenletes eloszlása, továbbá az algoritmus futási ideje $O(poly(|x|, -\log(\varepsilon)))$. Az ezen tulajdonságokkal rendelkező algoritmust úgyszintén FPAUS algoritmusnak hívjuk.

A leszámlálások és mintavételezések bonyolultságelmélete nagyon sokszor összefügg. Ezt tételszerűen az önhasonló leszámlálási problémákra lehet megadni. Nehéz, technikai definíciója miatt eltekintünk az önhasonlóság definíciójától, csak annyit említünk meg, hogy számos természetes leszámlálási probléma, mint pl. a páros gráfok teljes párosításainak a száma önhasonló. A Jerrum-Valiant-Vazirani tétel a következő

1. Tétel ([17])**.** *Minden önhasonló #A leszámlálási problémára*

$$\#A \in FPRAS \Leftrightarrow \#A \in FPAUS.$$

FPAUS algoritmusokat gyakran gyorsan keveredő Markov láncokon keresztül adunk meg.

5. Definíció. *Egy véges X halmazon megadott M* diszkrét idejű Markov lánc *az $X_0, X_1, X_2, \ldots$ véletlen változók ($\forall X_i \in X$) olyan sorozata, amelyre*

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots X_0 = x_0) = P(X_t = x_t | X_{t-1} = x_{t-1}).$$

Egyszerűség kedvéért, $P(X_t = x_t | X_{t-1} = x_{t-1})$ jelölést $P(x_t|x_{t-1})$-nek rövidítjük. Az $x_0, x_1, x_2 \ldots$ elemeket a Markov lánc állapot*ainak hívjuk. Legyen $\vec{G} = (V, E)$ az az irányított gráf, melyre $V = X$, és $(u, v) \in E$ akkor és csak akkor, ha $P(v|u) \neq 0$. A $\vec{G}$ gráfot a Markov lánc* átmeneti gráfjának *hívjuk. Ekkor M-et* irreducibilis*nek hívjuk, ha $\vec{G}$ (erősen) összefüggő, és* aperiodikus*nak, ha $\vec{G}$ irányított körei hosszainak legnagyobb közös osztója* 1.

M gyengén reverzibilis *ha minden i, j-re*

$$P(x_j|x_i) \neq 0 \Leftrightarrow P(x_i|x_j) \neq 0.$$

M reverzibilis egy π eloszlásra nézve*, ha minden i, j-re*

$$P(x_j|x_i)\pi(x_i) = P(x_i|x_j)\pi(x_j).$$

Sokszor az átmeneti valószínűségeket egy mátrixba írjuk, és a $T(\cdot|\cdot)$ jelölést használjuk rá (T az angol *transition* szóból jön). Ismert tétel, hogy ha egy Markov lánc irreducibilis, aperiodikus és reverzibilis egy π eloszlásra vézve, akkor tetszőleges x_0 pontból indulva a π eloszláshoz fog konvergálni. A konvergencia sebességére definiáljuk a relaxációs időt.

6. Definíció. *Egy M Markov lánc relaxációs idejét minden $\varepsilon > 0$-ra a következőképpen definiáljuk:*

$$\tau_i(\varepsilon) := \inf \{n_0 | \forall n \geq n_0 \; d_{TV}(T^n \mathbb{1}_i, \pi) \leq \varepsilon\},$$

$$\tau(\varepsilon) := \max_i \tau_i(\varepsilon),$$

ahol $\mathbb{1}_i$ az a vektor, amely mindenhol 0 *kivéve az i-edik koordinátája* 1*.*

Ezután definiáljuk a Markov láncok gyors konvergenciáját.

7. Definíció. *Legyen #A egy leszámlálási probléma, és legyen $\mathcal{M}$ Markov láncok egy osztálya, amely minden #A-beli y feladatra tartalmaz egy olyan Markov láncot, amelynek az állapottere y megoldásai. A $\mathcal{M}$ Markov lánc osztályt* gyorsan konvergáló*nak hívjuk, ha minden y feladatra és $\varepsilon > 0$-ra, az y-hoz tartozó Markov lánc relaxációs idejére teljesül, hogy*

$$\tau(\varepsilon) = O(poly(|y|, -\log(\varepsilon))).$$

A $\mathcal{M}$ Markov lánc osztályt lassan konvergáló*nak hívjuk, ha létezik olyan $c > 1$, $\varepsilon > 0$ és feladatok $y_1, y_2, \ldots$ végtelen sorozta, amelyre a hozzá tartozó $M_1, M_2, \ldots$ Markov láncok $\tau_1, \tau_2 \ldots$ relaxációs idejére*

$$\lim_{i\to\infty} \frac{\tau_i(\varepsilon)}{c^{|y_i|}} = \infty.$$

Gyorsan keveredő Markov láncokra megadható az alábbi könnyen belátható tétel.

2. Tétel. *Legyen #A egy leszámlálási probléma, és legyen $\mathcal{M}$ egy #A-hoz tartozó gyorsan konvergáló Markov lánc osztály. Ekkor ha #A minden y feladatára egy megoldás konstruálható polinom időben, valamint az y-hoz tartozó Markov láncban egy lépés polinom időben megtehető, akkor #A FPAUS-ban van.*

Adott eloszláshoz konvergáló Markov láncokat a Metropolis-Hastings algoritmussal konstruálhatunk.

8. Definíció. *A Metropolis-Hastings algoritmus bemenetele egy véges X téren adott, sehol nem eltűnő π eloszlás és egy X-en $T(\cdot|\cdot)$ átmeneti valószínűségekkel megadott, irreducibilis, aperiodikus, gyengén reverzibilis Markov lánc. A Metropolis-Hastings algoritmus a következő két lépéssel állít elő egy módosított Markov láncot.*

- *Ha x_t a Markov lánc aktuális állapota, akkor generáljunk egy y random állapotot, amely a $T(\cdot|x_t)$ eloszlást követi.*
- *Generáljunk egy u véletlen számot, amely a $[0, 1]$ intervallumon egyenletes eloszlást követi. A módosított Markov lánc x_{t+1} állapota y ha*

$$u \leq \frac{T(x_t|y)\pi(y)}{T(y|x_t)\pi(x_t)}, \tag{1}$$

és $x_{t+1} = x_t$ ha az egyenlőtlenség nem teljesül.

Az (1) képletben megadott hányadost Metropolis-Hastings hányados*nak hívjuk.*

A Metropolis-Hastings algoritmusra igaz a következő tétel.

3. Tétel. *A Metropolis-Hastings algoritmussal módosított Markov lánc reverzibilis, és tetszőleges kezdeti állapotból indulva a π eloszláshoz konvergál.*

2.3. Az értekezésben tárgyalt bioinformatikai problémák matematikai modelljei

Három fő témával foglalkozunk: gráfok és diszkrét tomográfia, genomátrendeződések, biológiai szekvenciák és szerkezetpredikciójuk.

2.3.1. Gráfok és diszkrét tomográfia

9. Definíció. *Egy $D = d_1, d_2, \ldots, d_n$* fokszámsorozat *nem-negatív egészek sorozata. Egy fokszámsorozat* grafikus *ha léteik olyan (pontcimkézett) $G = (V, E)$ gráf, amelyre minden i-re $d_i = d(v_i)$, ahol $d(v)$ a v pont foka. Egy ilyen G gráfot D* realizáció*jának hívjuk.*

Hasonlóan definiálhatjuk a $D_b = (d_{1,1}, d_{1,2}, \ldots, d_{1,n}), (d_{2,1}, d_{2,2}, \ldots, d_{2,m})$ páros fokszámsorozatok grafikusságát és páros gráf realizációit.

A switch *operáció egy G gráfból kitörli a (v_1, v_2) és (v_3, v_4) éleket és beteszi a (v_1, v_4) és (v_2, v_3) éleket. Az operáció feltétele, hogy a berakandó élek ne legyenek a gráfban az operáció megkezdése előtt.*

A switch operáció nem változtatja meg a gráf fokszámsorozatát. Ismert, hogy egy tetszőleges D fokszámsorozat vagy egy tetszőleges D_b páros fokszámsorozat esetén a fokszámsorozatok tetszőleges realizációjából switch operációkkal eljuthatunk a fokszámsorozat egy tetszőleges másik realizációjába. A switch operációkat Markov láncokban alkalmazzuk, amelyek adott fokszámsorozat realizációiból mintavételeznek. Ezen mintákat statisztikai hipotézisvizsgálatokban szereplő H_0 hipotézisek háttéreolszlásainak a generálásához használjuk fel.

Van, amikor egy H_0 hipotézis háttéreloszlásához csak olyan relizációkat tekintünk, amelyekről többet szeretnénk kikötni, mint a fokszámok. Például, arra is megszorításokat akarunk tenni, hogy mely fokú pontok között mennek élek. Erre ad lehetőséget a közös fokszám-mátrix, amelyet alább definiálunk.

10. Definíció. *A $\mathcal{J}(G) = [\mathcal{J}_{ij}]$ $\{1 \leq i, j \leq k\}$ mátrix a G gráf* közös fokszám-mátrix*a, ha minden i-re $J_{i,i}$ az i fokú pontok feszített részgráfjának*

az összfoka, és minden i-re és j-re $J_{i,j}$ az i és a j fokú pontok közötti feszített páros gráf összfoka. Egy szimmetrikus $k \times k$ M mátrix realizációja *G, ha $M = \mathcal{J}(G)$. Ha M-nek van realizációja, akkor M* grafikus *közös fokszám-mátrix.*

A diszkrét tomográfiai problémák körében olyan feladatokkal foglalkozunk, amelyek adott fokszámsorozatok együttes realizációinak a megkonstruálását keresik. Ezt definiáljuk alább.

11. Definíció. *Egy $k \times n$-es D és egy $k \times m$-es F nem-negatív egész mátrixok párját páros fokszámsorozatmátrixnak hívjuk. Egy k színnel élszínezett $G = (U, V, E)$ páros gráf a (D, F) pár* realizációja*, ha minden c_i színre az adott színű részgráf fokai a D és az F mátrixok i-edik sorából vett fokszámsorozattal egyeznek meg.*

2.3.2. Genomátrendeződések

A genomok hosszú DNS láncokból állnak, amelyek beazonosítható blokkokból tevődnek össze. A DNS kettős spirál irányítottsága miatt ezeknek a blokkoknak irányuk is van. Így van értelme a következő definíciónak.

12. Definíció. *Egy* genom *egy élcimkézett, irányított gráf, $\vec{G} = (V, E)$, amely minden komponense (nem szükségképpen irányított) út vagy kör, és minden cimke egyedi. Hurkok megengedettek, minden hurok egy* 1 *hosszú körként értendő. $\vec{G}$ komponenseit* kromoszómáknak *hívjuk. A kromoszómák lehetnek lineárisak és cirkulárisak. Az egy komponensű genom* unikromoszómális*, a több komponensű genom* multikromoszómális*. $\vec{G}$ éleit* blokk*oknak hívjuk (angolul:* synteny block*), megkülönböztetjük az élek csúcsait és farkait, amiket együttesen* extremitás*nak hívunk. A kétfokú pontok nevei* szomszédság*, az egyfokú pontok a* telomer*ek. Minden pont egyértelműen azonosítható azzal, hogy mely extremitások illeszkednek rá.*

A klasszikus gráfelmélettel szemben a genomok építőkövei az élek (hiszen azok cimkézettek). A genomátrendeződések során az éleket tartjuk meg és a pontokat változtatjuk meg. Az alábbi genomátrendezési operációkat definiáljuk.

13. Definíció. *A* dupla vágás és kötés *modell legfeljebb kettő pontot vesz, és azokat kombinálja össze legfeljebb kettő új pontba. Ez lehet:*

1. *Két szomszédság szétvágásából a keletkező négy extremitásból készít két új szomszédságot.*

2. *Egy szomszédság szétvágásából származó kettő extremitás és egy telomer extremitásából készít egy új telomert és egy új szomszédságot.*

3. *Egy szomszédság kettévágásával készít két új telomert.*

4. *Két telomer összekötésével készít egy új szomszédságot.*

A szimpla vágás vagy kötés *model vagy egy extremitást vág szét két telomerré vagy két telomert köt össze egy szomszédságba.*

A reverzió *olyan dupla vágás és kötés operáció, amely nem változtatja meg a kromoszóbák számát és minőségét. Azaz egy lineáris vagy cirkuláris kromoszómában egy vagy több blokkot fordít meg.*

A genomátrendeződési modellek abban különböznek, hogy mely oprációkat engedjük meg. A DCJ modell*ben a dupla vágás és kötés operációk megengedettek. Az* SCJ modell*ben a szimpla vágás és kötés operációk megengedettek. A* REV modell*ben a reverziók a megengedett operációk.*

A genomátrendeződések alapkérdése az, hogy adott két genom, $\vec{G}_1$ és $\vec{G}_2$, valamint egy genomátrendeződési modell, és azt kérdezzük, hogy mi az a minimális számú operáció az adott modellben, amivel $\vec{G}_1$ $\vec{G}_2$-be alakítható át. Ezt *legtakarékosabb scenárió*nak hívjuk. A szükséges műveletek száma a két genom *távolság*a, és ezt $d_M(\vec{G}_1, \vec{G}_2)$-vel jelöljük, ahol $M \in \{DCJ, SCJ, REV\}$.

Az alábbi öt leszámlálási kérdést lehet definiálni tetszőleges $M \in \{DCJ, SCJ, REV\}$ modellre:

1. **Legtakarékosabb scenáriók** Két genom, $\vec{G}_1$ és $\vec{G}_2$, közötti legtakarékosabb scenárióknak a számát kérdezzük meg az M modellben. Ezt $n_M(\vec{G}_1, \vec{G}_2)$-vel jelöljük.

2. **Legtakarékosabb mediánok** Adottak $\vec{G}_1, \vec{G}_2, \ldots, \vec{G}_k$ genomok, azon $\vec{G}_m$ medián genomok számát kérdezzük meg, amelyek minimalizálják a

$$\sum_{i=1}^{k} d_M(\vec{G}_i, \vec{G}_m)$$

összeget. A legtakarékosabb mediánok halmazát $\mathcal{O}_M(\vec{G}_1, \vec{G}_2, \ldots, \vec{G}_k)$ jelöli.

3. **Legtakarékosabb medián scenáriók** Adottak $\vec{G}_1, \vec{G}_2, \ldots, \vec{G}_k$ genomok, a

$$\sum_{\vec{G}_m \in \mathcal{O}_M(\vec{G}_1, \vec{G}_2, \ldots, \vec{G}_k)} \prod_{i=1}^{k} n_M(\vec{G}_i, \vec{G}_m)$$

összeget keressük.

4. **Legtakarékosabb fa cimkézés** Adott egy $T = (V, E)$ gyökerezett bináris fa és egy $f : L \to \mathcal{G}$ függvény, ahol L T leveleinek a halmaza, $\mathcal{G}$ pedig az M modellben lehetséges genomok halmaza. Azt kérdezzük meg, hogy hány olyan $g : V \to \mathcal{G}$ függvény van, ami az f függvény kiterjesztése T összes pontjára és minimalizálja a

$$\sum_{(u,v) \in E} d_M(g(u), g(v))$$

függvényt. Ezen függvények halmazát $\mathcal{O}'(T, f)$ jelöli.

5. **Legtakarékosabb scenáriók evolúciós fán** Adott egy $T = (V, E)$ gyökerezett bináris fa és egy $f : L \to \mathcal{G}$ függvény, a

$$\sum_{g \in \mathcal{O}'(T,f)} \prod_{(u,v) \in E} n_M(g(u), g(v))$$

összeget keressük.

2.3.3. Biológiai szekvenciák és térszerkezetpredikciójuk

A leggyakoribb biológiai szekvenciák a DNS, RNS és fehérjeláncok. A DNS és az RNS szekvenciák 4 különböző nukleotidból épülnek fel, és ezekből tetszőleges hosszú láncok készíthetőek. Így a DNS és RNS szekvenciák 4 elemű ABC feltti szekvenciaként modellezhetőek. A fehérjék 20 különböző aminosavból felépülő makromolekulák, így 20 elemű ABC feletti szekvenciaként modellezhetőek. Mindegyik fajta makromolekulának lehet funkcionális és térbeli szerkezete, ezeket lehet sztochasztikus nyelvtanokkal és energiamodellekkel is modellezni.

14. Definíció. *Egy* sztochasztikus reguláris nyelvtan *egy olyan* (T, N, S, R, P) *ötös, melyben* T *a terminális karakterek egy véges halmaza,* N *a nem-terminális karakterek egy véges halmaza,* $T \cap N = \emptyset$*,* $S \in N$

a kezdő nem-terminális, R átírási szabályok véges halmaza, melyben minden szabály $W \to aW'$ vagy $W \to a$ formátumú, ahol $W, W' \in N$, $a \in T$, P pedig egy függvény $R \to \mathbb{R}^{>0}$, úgy, hogy minden $W \in N$-re

$$\sum_{a|(W\to a)\in R} P(W \to a) + \sum_{W',a|(W\to aW')\in R} P(W \to aW') = 1.$$

15. Definíció. *Egy* sztochasztikus környezetfüggetlen nyelvtan *egy olyan (T, N, S, R, P) ötös, melyben T a terminális karakterek egy véges halmaza, N a nem-terminális karakterek egy véges halmaza, $T \cap N = \emptyset$, $S \in N$ a kezdő nem-terminális, R átírási szabályok véges halmaza, melyben minden szabály $W \to \beta$ formátumú, ahol $W \in N$, $\beta \in (T \cup N)^* \setminus \{\varepsilon\}$, ahol ε az üres szekvencia, P pedig egy függvény $R \to \mathbb{R}^{>0}$, úgy, hogy minden $W \in N$-re*

$$\sum_{\beta|(W\to\beta)\in R} P(W \to \beta) = 1.$$

Sztochasztikus nyelvtanokban egy levezetés alatt szekvenciák olyan

$$S = A_0, A_2, \ldots, A_k \in T^*$$

sorozatát értjük, melyben minden A_{i+1} szekvencia úgy keletkezik A_i-ből, hogy A_i egy W karakterét átírjuk egy szekvenciára valamely átírási szabály alapján. A levezetés valószínűsége a benne levő átírási szabályok valószínűségeinek a szorzata (multiplicitással). Biológiai szerkezetpredikcióban T a biológiai szekvenciát felépítő ABC, a levezetési szabályok valamely szerkezeti tulajdonsággal azonosíthatóak, és a szerkezetpredikció pedig a legvalószínűbb levezetést keresi. A sztochasztikus nyelvtanok átírási szabályainak a valószínűségeit a Baum-Welch tréninggel is meg lehet határozni. Ez egy iteráció, amely valamely kezdeti P valószínűségekből ad meg egy $\hat{P}$ becslést, majd P-nek a $\hat{P}$ értékeket adva a becslést lehet iterálni. A Baum-Welch tréning tulajdonsága az, hogy az iteráció során a likelihood, azaz az adott szekvencia teljes levezetési valószínűsége monoton növekszik. Jelölje $\mathcal{G}$ egy X szekvencia lehetséges levezetéseinek a halmazát, $P(g)$ egy $g \in \mathcal{G}$ levezetés valószínűségét, $n(g, r)$ pedig azt, hogy a g levezetésben az $r \in R$ szabályt hányszor használtuk. Ekkor a Baum-Welch tréning képlete

$$\hat{P}(W \to \beta) := \frac{\sum_{g\in\mathcal{G}} P(g) n(g, W \to \beta)}{\sum_{g\in\mathcal{G}} \sum_{\beta'|(W\to\beta')\in R} P(g) n(g, W \to \beta')}.$$

16. Definíció. *Legyen X egy RNS szekvencia. Ekkor X egy* RNS térszerkezete *olyan S rendezett index párok halmaza, melyre minden $(i,j) \in S$-re $1 \leq i < j \leq |X|$, továbbá minden index legfeljebb egy párban szerepel, és minden (i,j), $(i',j') \in S$, $i < i'$-re vagy $j < i'$ vagy $j' < j$.*

Ha S egy X RNS szekvencia RNS térszerkezete, akkor készíthetünk egy olyan $H = (V,E)$ gráfot, amely pontjai X karakterei, és $(x_i, x_j) \in E$, ha $(i,j) \in S$ vagy $j = i+1$ vagy $i = 1$ és $j = |X|$. H egyszerű körei olyan körök, amely pontjai által feszített részgráfok nem tartalmaznak további köröket. Belátható, hogy H egyértelműen bontható fel egyszerű körökre, és minden S-ből származó él pontosan két körben van benne.

17. Definíció. *A* Zucker-Tinoco energiamodell *a H gráfok lehetséges egyszerű köreihez definiál szabadenergiákat az egyszerű körök hossza, az S-beli élek száma, és a pontokhoz tartozó nukleotidok típusai szerint. Egy X RNS szekvencia S térszerkezetének a szabadenergiája a hozzá tartozó H gráf egyszerű körei szabadenergiáinak az összegeként van definiálva. Az X szekvencia lehetséges RNS térszerkezetei* Boltzmann eloszlás*a egy T hőmérsékleten az az eloszlás, amelyben az S térszerkezet valószínűsége*

$$\pi_T(S) \propto e^{\frac{-\Delta G(S)}{RT}}, \tag{2}$$

ahol $\Delta G(S)$ az S térszerkezet szabadenergiája, R az egyetemes gázállandó ($R \approx 8.314 \frac{J}{K \times mol}$), a $\propto$ pedig arányosságot jelöl.

Egy X RNS szekvencia térszerkezetére a becslés a minimális szabadenergiájú térszerkezet. Vizsgálhatjuk a térszerkezetnek a Boltzmann eloszlásban vett valószínűségét, ekkor a 2 képletben implicite megadott arányossági tényezőt is ki akarjuk számolni, azaz a

$$\sum_S e^{\frac{-\Delta G(S)}{RT}}$$

összeget, ahol az összegzés egy adott RNS szekvencia összes lehetséges térszerkezetén megy. A Zucker-Tinoco energiamodellben a lehetséges egyszerű körök szabadenergiáira olyan szabályok vannak megadva, hogy mind a minimális szabadenergiájú térszerkezet megkeresésére, mind a Boltzmann elszolásban szereplő arányossági tényező kiszámolására lehetséges legyen polinom időben futó dinamikus programozási algoritmus. Ez a dinamikus programozási algoritmus lényegében ekvivalens a sztochasztikus környezetfüggetlen nyelvtanok legvalószínűbb levezetésének a megkeresésére valamit a teljes levezetés valószínűségére adott dinamikus programozási algoritmussal.

3. Tudományos eredmények rövid összefoglalása

3.1. Polinom időben megoldható problémák

Megadható egy olyan S félgyűrű, amelyik únió és kompozíció műveleteket definiál RNS térszerkezetek (multi)halmazain. Továbbá megadható egy olyan dinamikus programozási algoritmus, amely egy X RNS szekvenciához tartozó lehetséges térszerkezetet halmazait állítja elő az előbbi félgyűrű műveletek segítségével. Ez a dinamikus programozási algoritmus univerzálisnak tekinthető abban az értelemben, hogy az RNS térszerkezetek Zucker-Tinoco modelljében a minimális szabadenergiájú térszerkezet szabadenergiáját kiszámoló dinamikus programozási algoritmus valamint a Boltzmann eloszlásban szereplő arányossági tényező kiszámolására megadott dinamikus programozási algoritmus az univerzális dinamikus programozási algoritmusból származtatható az S félgyűrűnek a tropikus félgyűrűbe illetve a nem-negatív valós számok félgyűrűjébe vett homomorfizmusai segítségével. Megadtunk további félgyűrűket, amelyek segítségével a Boltzmann eloszlás momentumai is kiszámolhatóak.

1. Tézis ([23]). *A Zucker-Tinoco energiamodellben egy n hosszúságú RNS szekvencia térszerkezetei Boltzmann eloszlásának a k-adik momentuma kiszámítható $O(n^3k^2)$ időben.*

Egy másik félgyűrű segítségével pedig memória-hatékony változatát adtuk meg a Baum-Welch tréningnek.

2. Tézis ([22]). *Egy sztochasztikus reguláris nyelvtan Baum-Welch tréningje egy n hosszúságú szekvencián elvégezhető $O(n)$ időben és $O(1)$ memóriában.*

Az SCJ genomátrendeződési modellben a legtakarékosabb scenáriók a zig-zag permutációkkal állnak kapcsolatban, a legtakarékosabb mediánok pedig a legfeljebb kettő fokú egyszerű gráfok (nem feltétlenül teljes) párosításaival. Mivel mind az adott hosszúságú zig-zag permutációkat, mind a $\Delta = 2$ egyszerű gráfokban a párosításokat meg lehet számolni polinom időben egyszerű dinamikus programozási algoritmusokkal, adódik a következő tézis.

3. Tézis ([24, 28]). *Az SCJ genomátrendeződési modellben a Legtakarékosabb scenáriók és a Legtakarékosabb mediánok FP-ben vannak.*

3.2. Gyorsan konvergáló Markov láncok

A switch Markov lánc gyors konvergenciájáról számos eredményünk van. Először a Tyshkevich kompozíciót definiáljuk.

18. Definíció. *A* $D_b = (d_{1,1}, d_{1,2}, \ldots, d_{1,n}), (d_{2,1}, d_{2,2}, \ldots, d_{2,m})$ *és* $D'_b = (d'_{1,1}, d'_{1,2}, \ldots, d'_{1,n'}), (d'_{2,1}, d'_{2,2}, \ldots, d'_{2,m'})$ *páros fokszámsorozatok Tyshkevich kompozíciója a*

$$\begin{aligned} D_b \circ D'_b \;\; := \;\; & (d_{1,1}, d_{1,2}, \ldots, d_{1,n}, d'_{1,1}+m, d'_{1,2}+m, \ldots, d'_{1,n'}+m), \\ & (d_{2,1}+n', d_{2,2}+n', \ldots, d_{2,m}+n', d'_{2,1}, d'_{2,2}, \ldots, d'_{2,m'}) \end{aligned}$$

fokszámsorozat.

Tyshkevich kompozíciókról a következő tételt bizonyítottuk.

4. Tézis ([11])**.** *Legyen* $\mathcal{D}_b$ *páros fokszámsoroztok egy osztálya,* $\mathcal{D}'_b$ *pedig azon páros fokszámsorozatok osztálya, amelyeket a* $\mathcal{D}_b$*-beli fokszámsorozatok (akár többszörös) Tyshkevich dekompozícióival kapunk. Ha a* $\mathcal{D}_b$ *realizációihoz tartozó* $\mathcal{M}$ *switch Markov láncok osztálya gyorsan konvergál, akkor a* $\mathcal{D}'_b$ *realizációihoz tartozó* $\mathcal{M}'$ *switch Markov láncok osztálya is gyorsan konvergál.*

A P-stabil fokszámsorozat osztályokon is bizonyítottuk a switch Markov lánc gyors keveredését. Először a P-stabilitást definiáljuk.

19. Definíció. *Legyen* $\mathcal{D}_b$ *páros fokszámsorozatok egy osztálya. A* $\mathcal{D}_b$ *osztály* P-stabil *ha van olyan* $p \in \mathbb{R}[x]$ *polinom, hogy minden* x*-re és páros fokszámsorozat* $D_b = (D_1, D_2) \in \mathcal{D}_b$*,* $n + m = x$ *esetén*

$$\left| \mathbb{G}(D_b) \cup \left(\bigcup_{i \in [n],\ j \in [m]} \mathbb{G}(D_1 + \mathbb{1}_i, D_2 + \mathbb{1}_j) \right) \right| \leq p(x) |\mathbb{G}(D_b)|, \qquad (3)$$

ahol $\mathbb{G}(D_s)$ D_b *realizációinak a halmaza , és* $\mathbb{1}_i$ *az a vektor, ami* 1*-et tartalmaz az* i*-edik koordinátán, és mindenhol máshol* 0*-t.*

5. Tézis ([9, 6])**.** *P-stabil páros fokszámsorozatok realizációin a switch Markov lánc gyorsan konvergál.*

Bizonyítottunk egy tételt arról is, hogy lineárisan korlátozott fokszámsorozatok P-stabilak.

6. Tézis ([9])**.** *Lgyen $\mathcal{D}_b$ páros fokszámsorozatok egy osztálya, úgy, hogy minden $D_b = (D_1, D_2) = ((d_{1,1}, d_{1,2}, \ldots, d_{1,n}), (d_{2,1}, d_{2,2}, \ldots, d_{2,m})) \in \mathcal{D}_b$ fokszámsorozatra a következő teljesül. Léteznek $0 < c_1 \leq c_2 < n$ és $0 < d_1 \leq d_2 < m$ konstansok a következő tulajdonságokkal:*

$$\begin{aligned} c_1 \leq d_{1,i} \leq c_2, \qquad & \forall i \in [n] \\ d_1 \leq d_{2,j} \leq d_2, \qquad & \forall j \in [m]. \end{aligned} \tag{4}$$

Továbbá a

$$(c_2 - c_1 - 1) \cdot (d_2 - d_1 - 1) < 1 + \max\{c_1(m - d_2), d_1(n - c_2)\} \tag{5}$$

egyenlőtlenség áll. Ekkor $\mathcal{D}_b$ P-stabil.

Egy nem-P-stabil fokszámsorozat osztályon is bizonyítottunk gyors konvergenciát. Ennek az osztálynak az érdekessége, hogy minden nem-P-stabil fokszámsorozat osztályban megjelenik valamely realizáció feszített részgráfjának a sorozataként.

7. Tézis. *Legyen $\mathcal{D}_b$ nem-P-stabil fokszámsorozat osztály. Ekkor minden $c \in \mathbb{N}$-re, létezik $D_b \in \mathcal{D}_b$ és egy $G(U, V, E)$ realizációja D_b-nek, hogy $G(U, V, E)$ tartalmaz egy C kört úgy, hogy C feszített részgráfjának a fokszámsorozata $((1, 1, 2, 3, \ldots, \ell - 1), (1, 1, 2, 3, \ldots, \ell - 1))$, ahol $\ell \geq c$.*

Ugyanakkor a $((1, 1, 2, 3, \ldots, \ell-1), (1, 1, 2, 3, \ldots, \ell-1))$ fokszámsorozatok realizációin a switch Markov lánc gyorsan konvergál.

Közös fokszám-mátrixok realizációin bolyongó Markov láncokról is bizonyítottunk gyors konvergenciát. Először a kiegyensúlyozott realizációkat definiáljuk.

20. Definíció. *Egy közös fokszám-mátrix realizációja* kiegyensúlyozott*, ha minden i-re és j-re, az i és j fokú pontok közötti feszített páros gráf fokszámsorozatára teljesül, hogy valamely k-ra az egyik osztály fokai mind k vagy $k+1$ és valamely ℓ-re a másik osztály fokai mind ℓ vagy $\ell+1$, továbbá minden i-re az i fokú pontok által feszített részgráf fokszámsorozatára teljesül, hogy valamely k-ra a fokok k vagy $k + 1$.*

8. Tézis ([10])**.** *Grafikus közös fokszám-mátrixoknak mindig létezik kiegyensúlyozott realizációja is. Bármely közös fokszám-mátrix kiegyensúlyozott realizációin a switch Markov lánc irreducibilis és gyorsan konvergál.*

Egy gyorsan konvergáló Markov lánc és egy projekciós lemma segítségével bizonyítottuk a legtakarékosabb DCJ scenáriók bonyolultságáról a következőket.

9. Tézis ([27]). *A DCJ genomátrendeződési modellben a Legtakarékosabb scenáriókra van FPRAS és FPAUS algoritmus.*

3.3. Negatív eredmények

Genomátrendeződési modellekben a legtakarékosabb scenáriók mintavételezésére egy ésszerű Markov lánc az, amely az aktuális scenárióból kivág egy darabot valamely g és g' genomok között, és egy random legtakarékosabb scenárióra cseréli le. Ezt a random legtakarékosabb scenáriót lépésről lépésre generálja le, minden lépésben g' felé a távolságot eggyel csökkentő átrendezések közül egyenletes eloszlásból választva. A REV modellben nem is ismerünk más Markov láncot, amely irreducibilis lenne. Azonban ez a Markov lánc lassan konvergál.

10. Tézis ([21]). *Legyen $\mathcal{M}$ azon Markov láncok osztálya, amely a REV modellben két genom közötti legtakarékosabb scenáriókon bolyong az alábbiak szerint. Az adott scenárióból valamely p eloszlás szerint kivág egy ablakot valamely g és g' genomok közt, és egy új részscenáriót javasol g és g' között, minden lépésben a g' felé a távolságot eggyel csökkentő átrendezések közül egyenletes eloszlásból választva. A javasolt új scenárióra a Metropolis-Hastings algoritmust alkalmazzuk. Ekkor az $\mathcal{M}$ Markov lánc osztály lassan konvergál, bármi is a p eloszlás.*

Az SCJ modellben az alábbi negatív eredményeket bizonyítottuk.

11. Tézis ([25, 28]). *Az SCJ modellben a legtakarékosabb scenáriók evolúciós fán #P-teljes probléma, és ha RP≠NP, akkor nem létezik rá FPRAS approximáció.*

Az SCJ modellben a legtakarékosabb medián scenáriók #P-teljes probléma.

3.4. Markov láncok nem triviális kernelekkel, kis átmérővel és nagy Metropolis-Hastings hányadossal

A REV modellben a legtakarékosabb scenáriók mintavételezésére az egyetlen ismert irreducibilis Markov lánc lassan konvergál. A lassú konvergencia oka,

hogy a Metropolis-Hastings hányados nagyon pici lehet. Ezért kutatásaink egyik célja olyan irreducibilis Markov lánc megadása, amely Metropolis-Hastings hányadosa nem lehet túl pici, azaz a reciprokára létezik polinom felső korlát. Mivel tudjuk, hogy egyenletes eloszlás esetén a konvergencia sebessége nem lehet az átmérő szublineáris függvénye, ezért arra is törekedni kell, hogy az átmérőre is legyen polinom felső korlát. Ezzel kapcsolatban fogalmaztunk meg egy sejtést.

1. Sejtés. *Adott G_1 és G_2 genomok REV modell alatti legtakarékosabb scenárióit írjuk le az átrendeződések szekvenciáival, amely szekvenciában minden karakter azt adja meg, hogy mely extremitások hogyan változtak. Készítsünk egy $G = (V, E)$ gráfot, amelynek a pontjai a legtakarékosabb scenáriók, és $(v_1, v_2) \in E$ ha a v_1-et és a v_2-t leíró szekvenciák leghosszabb közös részszekvenciájának a hossza legalább a szekvenciák közös hossza mínusz négy. A sejtés az, hogy G összefüggő és az átmérője felülről korlátos G_1 és G_2 méretének egy polinom függvényével.*

Könnyen belátható, hogy a hosszú közös részszekvencia lehetőséget ad egy olyan Markov láncra, amelynek az átmeneti gráfja G irányított változata (G minden élere mindkét irányban teszünk irányatott éleket), és ebben a Markov láncban meg lehet adni olyan átmeneti valószínűségeket, hogy a Metropolis-Hastings hányados reciprokára legyen polinom felső korlát.

A G_1 és G_2 genomok közötti különbség leírható az ún. *kívánalom és valóság gráfjával* (definíciót ld. az értekezésben), amely körökre bontható. Ha a körök csak 2 és 4 hopsszúak, akkor a körök átfedéseiből készíthető egy *átfedési gráf* (definíciót ld. szintén az értekezésben). Az ilyen típusú G_1 és G_2 párok biológiai szempontból akkor a legérdekesebbek, ha az ún. *végtelen hely (angolul: infinite site)* modellből jönnek. A következő tételt bizonyítottuk.

12. Tézis ([2])**.** *Legyen G_1 és G_2 a végtelen hely modellből származó genomok úgy, hogy a kívánalom és valóság gráfjukban a körök csak 2 és 4 hosszúak, és az átfedési gráf komponensei utak. Ekkor a Sejtés 1 igaz, és igaz marad akkor is, ha két pont között akkor megy él, ha az őket reprezentáló szekvenciák leghosszabb közös részszekvenciájának a hossza legalább a szekvenciák közös hosszai mínusz 2.*

A végtelen hely modellből származó G_1 és G_2 genompárokra az is igaz, hogy a REV modellben legtakarékosabb scenáriók halmaza a DCJ modellben legtakarékosabb scenáriók halmazának egy részhalmaza. Ez motiválta

a következő munkát. Legyen G_1 és G_2 a végtelen modellből származó unikromoszómális genompár lineáris kromoszómákal. Ekkor a DCJ modellben minden legtakarékosabb

$$G_1 = H_1, H_2, \ldots, H_k = G_2$$

scenárióra definiálhatunk egy hipotetikus energiát a

$$\sum_{i=1}^{k} c(H_i) \tag{6}$$

képlettel, ahol $c(H_i)$ a H_i genomban a cirkuláris kromoszómák száma. Egy scenárió pontosan akkor lesz a REV modellben is legtakarékosabb scenárió, ha az energiája nulla. Ez lehetőséget ad, egy ún. párhuzamos temperálás metódusra, amelyben párhuzamos Markov láncok vannak megadva. Mindegyik Markov lánc a fenti energia alapján egy adott hőmérséklethez tartozó Boltzmann eloszláshoz konvergál. Továbbá a Markov láncok néha állapotokat cserélnek, ezekre az állapotcserékre úgyszintén meg lehet adni egy Metropolis-Hastings algoritmust úgy, hogy minden lánc megtartsa az eedeti célelolszlását. Ezzel kapcsolatban a következőt bizonyítottuk.

13. Tézis ([26])**.** *Legyen G_1 és G_2 a végtelen modellből származó unikromoszómális genompár lineáris kromoszómákal. Ekkor megadható a (6) képletben definiált energia segítségével egy párhuzamos temperálás a következő tulajdonságokkal.*

1. *Minden Markov láncban egy lépés legfeljebb három konszekutív transzformációt változtat meg.*
2. *A párhuzamos temperálás Markov láncai irreducibilisek lesznek, polinom korlátú átmérővel.*
3. *A Boltzmann eloszlásokban a fenti változtatások által definiált topológia mellett minden lokális maximum globális.*
4. *A párhuzamos Markov láncok száma polinomiálisan nő a genomok méretével.*
5. *A legmelegebb Markov láncban a céleloszlás a DCJ modellben legtakarékosabb scenáriók egyenletes eloszlása.*

6. *A leghidegebb Markov láncban a céleloszlásban a REV modellben legtakarékosabb scenáriók összvlószínűsége legalább* 1/2.

7. *Két szomszédos Markov lánc állapotainak a kicserélésére a Metropolis-Hastings hányados legalább* 1/2.

Egy ilyen párhuzamos temperálás metódus azért biztató jelölt a REV modellben legtakarékosabb scenáriók FPAUS mintavételezésére, mert

- A DCJ modellben legtakarékosabb scenáriókra van FPAUS.
- A legmelegebb láncból polinom várható idő alatt 'lefiltrálódik' egy scenárió a leghidegebb láncba.
- A leghidegebb láncban a konvergencia elérése után legalább a minták fele a REV modellben legtakarékosabb scenáriókból származik.

Ezen kedvező tulajdonságok ellenére a gyors konvergenciát ezidáig nem sikerült bizonyítani.

Az SCJ modellben a legtakarékosabb fa cimkézés prombléma bonyolultsága nem ismert. A sejtés az, hogy ez már #P-teljes, de jól approximálható. Egy irreducibilis Markov láncot sikerült megadnunk erre a problémára, ami ráadásul egy Gibbs mintavételező. Egy Markov lánc Gibbs mintavételező, ha az állapottér leírható d dimenziós $\mathbf{x} := (x_1, x_2, \ldots, x_d)$ vektorokkal (nem feltétlenül Euklideszi térben), és minden lépésben a Markov lánc választ egy random i koordinátát, és arra egy új x_i' értéket a

$$\pi(\cdot|\mathbf{x}_{[-i]})$$

eloszlásból, ahol $\mathbf{x}_{[-i]}$ a $\mathbf{x}$ vektor összes koordinátája, kivéve az i-edik koordinátát. A Gibbs mintavételező tulajdonsága, hogy a Metropolis-Hastings hányadosa mindig 1. Bár a Gibbs mintavételezőnek ez egy kedvező tulajdonsága, a Gibbs mintavételező nem szükségképpen gyorsan konvergáló Markov lánc, sőt, még csak nem is szükségképpen irreducibilis. Az alábbi tételt bizonyítottuk.

14. Tézis ([24])**.** *Minden legtakarékosabb fa cimkézés problémára létezik egy Gibbs mintavételező, amely irreducibilis Markov lánc. A Markov láncban a*

lehetséges genomok a lehetséges szomszédságok prezencia/abszencia 1/0 *vektoraival vannak ábrázolva. A Gibbs mintavételező a fa összes cimkéjén az i-edik szomszédsághoz tartozó koordinátát mintavételezi újra, a lehetséges koordinátaértékek egyenletes eloszlásából mintavételezve. A Markov lánc átmérője polinomiálisan nő a feladat méretével.*

A Gibbs mintavételezőről nem sikerült bebizonyítani hogy gyorsan konvergál bár biológiai adatokon tesztelve kedvező autokorrelációs diagrammokat kaptunk.

Végezetül két irreducibilis Markov lánc osztályt adunk meg, amelyek átmérőjéről és a hozzá tartozó Metropolis-Hastings hányadosai reciprokáról tudtunk polinom felső korlátokat megadni.

21. Definíció. *Egy* (D, F) *páros fokszámsorozatmátrix* félig reguláris*, ha* D *miden sora konstans.*

15. Tézis ([1])**.** *Egy* (D, F) *félig reguláris páros fokszámsorozatmátrixnak létezik realizációja akkor és csak akkor ha minden* i*-re az* i*-edik sorok egy grafikus páros fokszámsorozatot adnak, és az oszlopösszegek által adott páros fokszámsorozat is grafikus.*

A félig reguláris páros fokszámsorozatmátrixok realizációin létezik olyan irreducibilis Markov lánc, amely egy lépésben legfeljebb három színt változtat meg, az átmérője polinomiális függvénye (D, F) *méretének, és ha a Metropolis-Hastings algoritmust alkalmazzuk rá a* π *egyenletes eloszlással, akkor a Metropolis-Hastings hányados reciprokára létezik polinom felső korlát.*

A Kőnig-tétel értelmében egy Δ maximális fokú páros gráfnak mindig létezik k-élszínezése, ha $k \geq \Delta$. Ezen élszínezésekre is megadtunk egy Markov láncot kis átmérővel és nagy Metropolis-Hastings hányadossal.

16. Tézis ([13])**.** *Páros gráfok* k*-élszínezéseire létezik olyan irreducibilis Markov lánc, amely egy lépésben legfeljebb három színhez tartozó éleket változtat meg, az átmérője polinom függvénye* k*-nak és a gráf méretének, és ha a Metropolis-Hastings algoritmust alkalmazzuk rá a* π *egyenletes eloszlással, akkor a Metropolis-Hastings hányados reciprokára létezik felső korlát, amely polinomja mind* k*-nak, mind a gráf méretének.*

Hivatkozások

[1] Aksen, M., **Miklós, I.**, Zhu, K. (2017) Half-regular factorizations of the complete bipartite graph. Discrete Applied Mathematics, 230:21–33.

[2] Bixby, E, Flint, T, **Miklós, I.**, (2016) Proving the Pressing Game Conjecture on Linear Graphs Involve, 9(1):41–56.

[3] Babai, L. (2015) Graph Isomorphism in Quasipolynomial Time `arXiv: 1512.03547`.

[4] Brightwell, G., Winkler, P. (1991) Counting linear extensions is #P-complete. In: Proceedings of the 23rd Annual ACM Symposium on Theory of Computing, 175–181.

[5] Dyer, M., Goldberg, L.A., Greenhill, C., Jerrum, M. (2004) The Relative Complexity of Approximate Counting Problems. Algorithmica, 38:471–500.

[6] Erdős, E.L., Greenhill, C., Mezei, T. R., **Miklós, I.**, Soltész, D., Soukup, L. (2022) The mixing time of the switch Markov chains: a unified approach, Eur. J. Comb., 99:103421.

[7] Erdős, E.L., Győri, E., Mezei, T. R., **Miklós, I.**, Soltész, D. (2021) Half-graphs, other non-stable degree sequences, and the switch Markov chain, Electronic Journal of Combinatorics, 28:3 #P3.7.

[8] Erdős, P.L., Kiss, Z.S., **Miklós, I.**, Soukup, L. (2015) Approximate Counting of Graphical Realizations. PLoS ONE, 10(7):e0131300.

[9] Erdős, E.L. Mezei, T., **Miklós, I.**, Soltész, D. (2018) Efficiently sampling the realizations of bounded, irregular degree sequences of bipartite and directed graphs. PLoS ONE, 13(8): e0201995.

[10] Erdős, P.L., **Miklós, I.**, Toroczkai, Z. (2015) A decomposition based proof for fast mixing of a Markov chain over balanced realizations of a joint degree matrix. SIAM Journal on Discrete Mathematics, 29:481–499.

[11] Erdős, P.L., **Miklós, I.**, Toroczkai, Z. (2018) New classes of degree sequences with fast mixing swap Markov chain sampling Combinatorics, Probability and Computing, 27(2):186–207.

[12] Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57:97–109.

[13] Hong, L., **Miklós, I.** (2022) A Markov chain on the solution space of edge-colorings of bipartite graphs. Discrete Applied Mathematics, accepted.

[14] Jerrum, M., Sinclair, A. (1996) The Markov chain Monte Carlo method: an approach to approximate counting and integration. In Dorit S. Hochbaum, editor, Approximation Algorithms for NP-hard Problems, pages 482–520. PWS, 1996.

[15] Jerrum, M., Sinclair, A., Vigoda, E. (2004) A Polynomial-Time Approximation Algorithm for the Permanent of a Matrix with Nonnegative Entries. Journal of the ACM, 51(4):671–697.

[16] Jerrum, M., Snir, M. (1982) Some Exact Complexity Results for Straight-Line Computations over Semirings. Journal of the ACM, 29(3):874–897.

[17] Jerrum, M., Valiant, L., Vazirani, V. (1986) Random generation of combinatorial structures from a uniform distribution. Theoretical Computer Science, 43:169–188.

[18] Karzanov, A., Kachiyan, L. (1991) On the conductance of order Markov chains. Order, 8:7–15.

[19] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953) Equations of state calculations by fast computing machines. J. Chem. Phys., 21(6):1087–1091.

[20] **Miklós, I.** (2019) Computational complexity of counting and sampling, CRC Press.

[21] **Miklós, I.**, Mélykúti, B., Swenson, K. (2010) The Metropolized Partial Importance Sampling MCMC mixes slowly on minimum reversal rearrangement paths ACM/IEEE Transactions on Computational Biology and Bioinformatics, 4(7):763–767.

[22] **Miklós, I.**, Meyer, I.M. (2005) A linear memory algorithm for Baum-Welch training. BMC Bioinformatics 6:231

[23] **Miklós, I.**, Meyer, I.M., Nagy, B. (2005) Moments of the Boltzmann distribution for RNA secondary structures Bul. Math. Biol., 67(5):1031-1047

[24] **Miklós, I.**, Smith, H. (2015) Sampling and counting genome rearrangement scenarios, BMC Bioinformatics, 16(Suppl 14): S6.

[25] **Miklós, I.**, Smith, H. (2019) The computational complexity of calculating partition functions of optimal medians with Hamming distance. Advances in Applied Mathematics, 102:18–82.

[26] **Miklós, I.**, Tannier, E. (2010) Bayesian sampling of genome rearrangement scenarios via Double Cut and Join. *Bioinformatics*, 26: 3012–3019.

[27] **Miklós, I.**, Tannier, E. (2012) Approximating the number of Double Cut-and-Join scenarios, Theoretical Computer Science 439:30–40.

[28] **Miklós, I.**, Tannier, E., Kiss, Z.S. (2014) On sampling SCJ rearrangement scenarios. Theoretical Computer Science, 552:93–98.

[29] **Miklós, I.**, Podani, J. (2004) Randomization of presence/absence matrices: comments and new algorithms, Ecology, 85:86–92.

[30] Pólya, G. (1913) Aufgabe 424.Arch. Math. Phys., 20:271.

[31] Szegő, G. (1913) Lösung zu 424 .Arch. Math. Phys., 21:291–292.

[32] Valiant, L.G. (1979) The complexity of computing the permanent.Theoretical ComputerScience, 8:189–201.

[33] Valiant, L.G. (1979) The complexity of enumeration and reliability problems. SIAM Journal on Computing, 8(3):410–421.

[34] Valiant, L.G. (2006) Accidental Algorthims. Foundations of Computer Science, IEEE Annual Symposium on. IEEE Computer Society. pp. 509–517.