Bioinformatikai eredetű kombinatorikai problémák

Erdős Péter

Bevezetés

A disszertáció 1990-óta keletkezett, alapvetően bioinformatikai eredményeket ismertet: a problémák döntő többsége a molekuláris biológia jelenlegi forradalmában felmerült kombinatorikai kérdésekből ered.

A dolozatban három fő rész található, összesen kilenc szakaszból áll, továbbá nyolc cikk szerepel mellékletként. A első két részben un. *evolúciós fákat* vizsgálok. Ezek (gyakran gyökeres) bináris fák, melyek levelei egy-egy értelműen címkézettek, míg belső (elágazó) csúcsaik nem. A biológusok ezeket használják a fajok közötti leszármazási kapcsolatok ábrázolására (és megtalálására). A biológiai adatokat kevés (tipikusan 2, 4 vagy 20) szín felhasználásával alkotott színvektorok hordozzák, továbbá a fával ábrázolt történések valamilyen biológusok által feltételezett modell szerint történnek. (Nem-biológusok ezeket az objektumokat gyakran *X-fáknak* nevezik, ahol az X halmaz a címkék összessége.)

Az első részben ez a modell a statisztikából ismerős parsimonia elv. A kérdések általában NP-nehezek, ezért a lehetséges modellfák közül gyakran statisztikai alapon választanak. Ebben a részben ilyen statisztikákkal kapcsolatos kombinatórikai problémákat vizsgálunk. Közülük az első egy leszámlálási kérdés, amely megoldása a Menger tételeken alapuló dekompozíciót használ. A módszerek kettőnél több színre történő alkalmazásához a *multiway cut* probléma jobb megértése lehet szükséges, amely az első rész másik témája.

A dolgozat második része evolúciós fák néhány sztochasztikus modelljével foglalkozik. Részben mutatószámokat illetve eszközöket fejleszt ki a modellek illetve módszerek összehasonlítására, részben pedig gyors algoritmusokat ad egy modellosztályban a helyes evolúciós fák 1 valószínűségű megtalálásához.

A disszertáció harmadik része véges ábécé feletti korlátos hosszúságú szavak rész-szavakból történő rekonstrukcióját vizsgálja, amely microarray kisérletek illetve úgynevezett *DNS kódok* tervezéséhez nyújthat segítséget.

1. A multiway cut probléma

A modern kombinatorikus optimalizálás egy sokat vizsgált területe a *multiway cut* (**MC**) probléma: adott a G gráf élein egy w súlyfüggvény. Adott továbbá *terminál pontok* egy k elemű halmaza. Keressünk minimális összsúlyú élvágást, ami a terminál pontokat *páronként szeparálja*: az élek elhagyásával keletkezett gráfban különféle színű pontok között nincsenek utak. A $k = 2$ eset a klasszikus él-Menger probléma. Az MC probléma általában NP-nehéz még a legegyszerűbb esetben is, de síkgráfokon a probléma kezelhető polinomiális időben, ha a színek száma korlátos.

Székely Lászlóval közös cikkeinkben ([1, 2, 7, 10, 13]) bevezettük az eredeti multiway cut probléma egy általánosítását: legyen $G = (V, E)$ egy egyszerű

gráf, $C = \{1, 2, \ldots, r\}$ pedig egy színhalmaz. Ha $N \subseteq V(G)$ a terminál pontok halmaza, akkor egy $\chi : N \to C$ leképezést *parciális színezés*-nek hívunk. Ekkor egy $\bar{\chi} : V(G) \to C$ leképezést akkor mondunk *színezésnek*, ha a két leképezés megegyezik a terminál pontokon. Az *általánosított* (avagy *színezett*) multiway cut (**szMC**) probléma egy olyan legkisebb súlyú élrendszer megtalálása, amely bármely két, eltérő színű terminál pontot szeparál.

Fenti definíció azért igazi általánosítás, mert bár az szMC tetszőleges gráfokon megegyezik az eredeti multiway cut problémával, speciális gráfosztályokon azonban (mint síkgráfokon vagy acyclikus gráfokon) eltérőek. Például síkgráfokon az szMC már három szín mellett és egységsúlyú élekkel is NP-teljes.

Az idézett cikkekben bevezettünk egy új típusú alsó korlátot a multiway cut súlyára, továbbá egy új típusú pakolási feladat felhasználásával illetve egy minimax tétel bebizonyításával teljesen megoldottuk a fák multiway cut problémáját. Ennek egyrészt elméleti következményei vannak, másrészt az eredmények maguk felhasználásra kerültek az evolúciós fák elméletében is. A színezett multiway cut-nak párhuzamos SQL-lekérdesések tervezése témakörében, vagy kommunikációs hálózatok elméletében is vannak alkalmazásai. Ez utóbbi esetben a kommunikációs költségeket minimalizálják szétosztott processzor hálózatok esetén.

Minimális súlyú színezések

A (számunkra fontos) biológiai alkalmazásokban a konstans élsúlyoknál bonyolultabb súlyfüggvényekre van szükség . Ehhez jelölje $E(G)\times 2$ a gráf irányított éleit (azaz mindegyik él mindkét irányítással jelen van). Egy $W : E(G) \times 2 \to \mathbb{N}^{r\times r}$ leképezés egy (színfüggő) *súlyfüggvény,* ha a $W(p, q)$ és $W(q, p)$ mátrixok megegyeznek, továbbá a főátlókban csupa nulla van. A ${}_iW(p, q)_j = w(p, q; i, j)$ elem azt mondja meg, hogy a (p, q) élnek mennyi a súlya egy $\bar{\chi}$ színezésben, ha $\bar{\chi}(p) = i, \bar{\chi}(q) = j$ (avagy $\bar{\chi}(p) = j, \bar{\chi}(q) = i$, ami ugyan azt az értéket adja). A W színfüggetlen, ha minden főátlón kívüli elem azonos. A súlyfüggvény értelemszerűen lesz élfüggetlen. Végül W *konstans*, ha egyszerre szín- és élfüggetlen. Bármely χ parciális színezés partícionálja a terminál pontokat: az azonos színű pontok kerülnek azonos osztályba. Ebben a gráfban élek egy halmaza, amelyek együtt bármely két, eltérő színű terminál pontot elválasztanak, egy (színezett) *multiway cut*-ot alkot. Világos, hogy egy $\bar{\chi}$ színezés színváltó élei mindig multiway cut-ot alkotnak. Egy $\bar{\chi}$ színezés súlya a színváltó élek összsúlya. Az adott gráfon egy χ parciális színezés $\ell(G, \chi)$ *hossza* (avagy a *súlyozott* MC nagysága) az összes lehetséges színezés súlyának a minimuma.

A $\ell(G, \chi)$ mennyiség meghatározásának komplexitása függ a súlyfüggvény és a gráf szerkezetétől. Biológiai alkalmazásokban a gráfok általában címkézett levelekkel és nem-címkézett belső pontokkal rendelkező bináris fák, ahol a parciális színezés a leveleken adott. Ezeket az objektumokat hívják *evolúciós fáknak*. A Székely Lászlóval közös [10] cikk tetszőleges, levél színezett fákra ad unárisan polinomiális algoritmust színfüggő súlyfüggvény esetén a hossz meghatározására. Az algoritmus arra is alkalmas, hogyha minden belső pontban

megadunk egy megendegett színhalmazt, akkor az algoritmus valamelyik megengedett színt rendeli a belső pontokhoz is. A cikk egyébként ennél egy kicsit általánosabb állítást igazol:

1. Tétel. *Legyen a gráf olyan, amelynek minden körét a terminál pontok lefedik. Ekkor létezik unárisan polinomális algoritmus egy optimális színezés meghatározására, feltéve, hogy a súlyfüggvény színfüggetlen*

Lényegesen bonyolultabb kérdést kapunk, ha levelek egy adott L halmazához és a rajtuk adott χ parciális színezéshez meg akarjuk határozni az összes, a levelekre illeszkedő bináris fa közül azt, amelyiknek a legkisebb a hossza a χ-re nézve. Ha a leveleket ma élő fajok alkotják, és a színezés pedig valamilyen biológiai jellemzőjüket jelenti (például morfológiai jegyek, vagy az átörökítő anyag egy jellemző része), akkor a legrövidebb fa megtalálása azt a nézetet testesíti meg, hogy a természet az élet kialakításánál takarékos volt, a lehető legkevesebb változást használta fel az összes létező élőlény kialakításához. Ezt *parsimonia elvnek* (avagy a filozófiában *Occam borotvájának*) hívják, és tipikus feltevés különböző statisztikai vizsgálatoknál.

Az evolúció kutatói ezeket a biológiai jellemzőket *karakter*-eknek hívják. Azaz az i-ik karakter matematikai értelemben a színvektor i-ik koordinátáját jelenti.

A valós helyzetekben, azaz létező biológiai rendszerek vizsgálatakor, persze nem csak egyetlen jellemző ír le egy-egy fajt, ezért minden fajt (azaz a keresett bináris fa leveleit) hosszabb színvektorok jellemeznek. Annak eldöntése, hogy ilyen színvektorok esetén létezik-e pontosan k hosszúságú fa a χ parciális színezésre nézve (ilyenkor az adott fára minden koordinátában külön kiszámoljuk a hosszat, majd összeadjuk) NP-nehéz feladat, ezért az érdekes gyakorlati esetekben ezt lehetetlen eldönteni. Ezen vizsgálatok egyik első lépése az adott levélszínezéshez tartozó, éppen k hosszúságú fák leszámlálása.

A legegyszerűbb eset megtárgyalásához rögzítsünk egy adott egy-karakteres, azaz egy hosszú színvektorokból álló 2-színezést az L levél halmazon. Legyen a és b a két színosztály mérete, és legyen $f_k(a,b)$ azon evolúciós fák száma, amelyek hossza az adott levélszínezés mellett éppen k. Már 1990 óta ismertes, hogy:

$$f_k(a,b) = (k-1)!(2n-3k)N(a,k)N(b,k)\frac{b(n)}{b(n-k+2)} \qquad (1)$$

ahol $a+b=n$, $a>0$, $b>0$, és ahol $N(x,k)$ jelöli az összesen x levéllel rendelkező és k darab evolúciós fából álló erdők számát. (A [9] cikkem, egyebek között, egy bijektív bizonyítást adott az $N(x,k)$ mennyiségekre.) Az (1) formulára adott eredeti bizonyítás többváltozós Lagrange inverziót és computer algebrát alkalmazott. M.A. Steel talált egy jobb, bijektív megközelítést, amire Székely Lászlóval közös [7] cikkünkben adtunk viszonylag rövid és transzparens bizonyítást. A módszer legfőbb érdekessége, hogy a leszámlálás előtt bebizonyítja a k hosszú evolúciós fák egy struktúra tételét, amely eredmény az él-Menger és a pont-Menger tételek felváltott alkalmazásain alapul.

A kettőnél több színnel színezett evolúciós fák leszámlálásához szükség lenne az evolúciós fákra vonatkozó analóg tételek bebizonyítására. A több színű pont-Menger tétel fákra változtatás nélkül teljesül, de ugyanez az él-Menger (azaz a multiway cut) problémára nem igaz.

Egy minimax eredmény fák szMC problémájára

Mivel az általánosított multiway cut probléma már $k = 3$ esetben is NP-nehéz, természetesen nem lehet elvárni általánosan érvényes, a Menger tételhez hasonló minimax eredményt vele kapcsolatban. Valóban, mint az közismet, már a $k = 3$ esetben sem igaz az él-Menger tétel analógja: egyszerű ellenpélda rá az egység élsúlyokkal ellátott, a leveleket terminál pontokként tartalmazó $K_{1,3}$ csillag. Azonban a [1, 2, 10] cikksorozatban Székely Lászlóval közösen sikerült fákra egy hasonló minimax tételt kimunkálnunk. Megjegyzendő, hogy ennek felhasználásával új-zélandi kutatók tovább léptek a leszámlálási feladat tárgyalásában.

A [1] cikkben a súlyozatlan esettel foglalkoztunk (pontosabban szólva itt minden él súlya 1), míg a [2, 10] dolgozatokban színfüggetlen súlyfüggvények esetére dolgoztuk ki a megfelelő minimax eredményt. A továbbiakban irányítatlan gráfokban két-két terminál pont közé, *irányított (oriented)* utakat pakolunk. Irányított út úgy keletkezik egy iranyítatlan P útból, hogy megmondjuk, hogy a határoló terminál pontok közül melyik az $s(P)$ kezdő pont, és melyik a $t(P)$ végpont, továbbá feltesszük, hogy az utak nem érintenek más terminál pontot. Egy út akkor *színváltó*, ha χ szerint eltérő színű terminál pontok között fut.

A Székely Lászlóval közös [10] cikkben hurokél mentes gráfok tetszőleges, azaz él- és színfüggő, súlyozása mellett tanulmányoztuk egy lehetséges alsó becslést a (súlyozott) multiway cut értékére, és találtunk egy minimax eredményt erre a problémájára.

Legyen G hurokél mentes gráf terminál pontok egy N halmazával, ahol a parciális színezés megint k színt használ . Legyen $\mathcal{P}$ színváltó irányított N utak multihalmaza (egyetlen út sem tartalmaz N-beli belső pontot, de valamely út több példányban is jelen lehet). Legyen továbbá $e = (p, q) \in E(G)$ egy rögzített él. Ekkor legyen

$$n_i(e, \mathcal{P}) = \#\{P \in \mathcal{P} \ : \ (p, q) \in P \text{ és } \chi(t(P)) = i\},$$

ahol a $t(P)$ újra az illető út végpontját jelöli, a $(p, q) \in P$ jelölés pedig azt jelenti, hogy az út a p pontban lép be az élbe, és a q pontban hagyja el az élt. Ezután színváltó utak egy rendszerét *útpakolás*nak mondjuk, ha minden $i \neq j$ színpárra és minden (p, q) élre teljesül:

$$n_i\big((p, q), \mathcal{P}\big) + n_j\big((q, p), \mathcal{P}\big) \leq w(p, q; j, i).$$

Ekkor

2. Tétel. *Legyen G hurokél mentes gráf az N terminál halmazzal és a χ parciális színezéssel. Legyen W egy (színfüggő) súlyfüggvény a gráfon és $\mathcal{P}$ egy útpakolás.*

Ekkor teljesül:

$$\ell(G, \chi) \geq |\mathcal{P}|.$$

Teljesül továbbá a következő minimax tétel is (a súlyfüggvény itt kevésbé általános):

3. Tétel. *Tetszőleges T fára és tetszőleges színfüggetlen $w : E(T) \to \mathbb{N}$ súlyfüggvényre minden $\chi : L(T) \to C$ levélszínezés esetén van olyan $\mathcal{P}$ útpakolás, amire teljesül*

$$\ell(G, \chi) = |\mathcal{P}|.$$

Vegyük észre, hogy azonosan 1 élsúly mellett az utak a fa felhasznált élein egyértelműen meghatároznak egy irányítást. Van-e mód ennek az irányításnak a meghatározására az útrendszer rögzítése nélkül?

A kérdésfeltevés mögött az a gondolat, hogyha sikerül megtalálni az említett irányítást, akkor már a szokásos él-Menger tétel k-szoros alkalmazásával meg lehet határozni az útrendszert. Nevezetesen egy színt elkülönítünk az összes többitől, és az irányított gráf ezen 2-színezésében keresünk irányított utakat. A vázolt gondalatmenetet a Frank Andrással és Székely Lászlóval közös [13] cikkben sikerült bizonyítássá érlelni. A cikkben tanulmányoztunk még néhány, mások által bevezetett szMC alsó becslést, és megállapítottuk ezek egymáshoz viszontított méretét. Azt is kimutattuk, hogy a fastruktúra igen hangsúlyos szerepet játszik a minimax tétel érvényességében.

2. Az evolúciós fák sztochasztikus elmélete

Ebben a fejezetben olyan problémákat tárgyalok, amelyek ugyan tisztán matematikai jellegűek, és amelyek nagy apparátust mozgatnak meg, azonban eredetük egyértelműen a biológiához köthető. A problémák háttere egy széles körben elfogadott biológiai modell, amely szerint az élővilág fejlődése, az új fajok kialakulása véletlen eseményeken alapul. A un. Kimura modell számba veszi ezen véletlen mutációk törvényszerűségeit, de nem foglalkozik azzal a kérdéssel, hogy a keletkezett egyedet mi tesz képessé a túlélésre, azaz mikor válhat egy új faj ősévé.

A fejezet előbb az evolúciós fák rekonstrukciójának sok lehetséges módszere közül két, alapvetően különböző megközelítést tárgyal. Az egyik egy un. karakter alapú módszer, amely minden rendelkezésre álló információt párhuzamosan használ, ezért nagy biztonsággal tudja a keresett evolúciós fát felépíteni, de eléggé lassú. A második megközelítés un. quartet alapú: ilyenkor egy evolúciós fa ismert levél-négyeseiből történik az evolúciós folyamat rekonstrukciója. Ezt a módszercsaládot általában a távolság alapú eljárások közé helyezik (bár ez nem törvényszerű).

Végezetül a fejezet utolsó szakasza az evolúciós fák egy nem-klasszikus értelemben vett rekonstrukciós eljárását tárgyalja, amelynek itt a helye, mert egy, a supertree módszerek közé (is) besorolható eljárást ismertetek fák rekonstrukciójáról.

Hadamard konjugáció

Az 1980-as évek elején M. Kimura japán biológus egy 3-paraméteres, véletlenen alapuló mutációs modellt dolgozott ki a fajok változékonyságának megmagyarázására. Mára ez vált a biológusok által legelfogadottabb modellé. Az az alapfelvetése, hogy az élőlények átörökítő anyagában a változások teljesen véletlenszerűen, egymástól nem befolyásolva zajlanak le.

A Kimura modell szerint a fajok fejlődését egy bináris fa szemlélteti, ahol a gyökér jelképezi a közös őst, míg a (címkézett) levelek a vizsgálandó fajokat. Ezek után az élek mentén lejátszódó betű-változások egymástól függetlenül, véletlenszerűen történnek. Mivel a fejlődés a közös őstől a ma élő fajok irányában történik, ezért a változásoknak egyértelmű iránya van, azonban a Kimura modell szerint egy változásnak és az ellentett változásnak ugyanannyi a valószínűsége. Továbbá az egyes élek mentén a változások eltérő valószínűséggel következ(het)nek be, de az ezeket leíró mátrixok szerkezete állandó.

Ezek alapján vezethette be Evans és Speed azt a modellt, ahol az egyes éleken történő változások a négy elemű Klein csoport hatásaként értelmezhetők. (Érdekes megjegyezni, hogy a Klein csoport definiálta változásoknak biológiai leírását is meg lehet adni.) Ebben a modellben a véletlen változások generálta "fejlődés" úgy jelentkezik, hogy a fa gyökerében található fajból rekurzívan határozhatók meg a folyamat közben létrejövő leszármazott fajok: az eleddig meghatározott fajokból kiinduló éleken meghatározzuk, milyen véletlen változások fognak lezajlani, majd a Klein csoport hatásaként meghatározzuk, milyen fajok jönnek létre.

Ilyenkor az éleken illetve a leveleken található valószínűségi elosztások között – bizonyos ésszerű megszorítások mellett – egy Fourier inverz párkapcsolat van, amely miatt valamelyik elosztásból pontosan meghatározható a másik eloszlás. Ezen a gondolatmeneten alapul az evolúciós fák un. *spektrál elmélete.* A módszer ősét (két színre), M. Hendy és D. Penny dolgozta ki, amelyet az Hadamard konjugáltak módszerének neveznek.

A módszer négy színre történő általánosítása a Székely László, Mike Steel és David Penny hármassal közös [5] cikkben kezdtük meg, illetve a Mike Steellel, Székely Lászlóval és Mike Hendyvel közös [3] cikkben fejeztük be. Szintén ebben a cikkben foglalkoztunk avval a kérdéssel, hogy a gyakorlati életben, ahol a leveleken megfigyelhető eloszlások csak bizonyos hibákkal észlelhetők, hogyan lehet egy megfelelő approximációs eljárást kifejleszteni. A kapott módszert *closest tree method*-nak nevezik.

A spectrál módszert a Klein csoport helyett tetszőleges véges Abel csoportra a Székely Lászlóval és Mike Steellel közös [6] cikkben általánosítottuk. Ennek közvetlen haszna ott lehet, ha a fajokat például nem DNS-kkel, hanem protein savaikkal (amiból az emberben például 20 van) azonosítjuk. A módszernek egyébként filozófiai értelemben nagy előnye, hogy képes bizonyos esetekben kimutatni, ha az adatokra teljesen "rossz" modellt kívánunk ráhúzni, azaz popperi értelemben falszifikálható.

A Short Quartet módszerek

Jelölje $B(n)$ az n címkézett levéllel ámde címkézetlen elágazási pontokkal bíró, gyökértelen fák halmazát. (Ezeket *X-fák*-nak is nevezik, ahol az X a levélcímkék halmaza. Azért nem használom itt az evolúciós fa kifejezést, hogy érzékeltessem a szélesebb kontexust.)

Legyen T egy $B(n)$-beli X-fa és legyen S a levelek egy részhalmaza. Ekkor jelölje $T_{|S}$ az S által generált részfát, míg jelölje $T^*_{|S}$ a generált bináris (topológikus) részfát. Ha adott az S levélhalmazon egy T-vel jelölt X-fa, akkor a fa egy élének a törlése egy 2-partíciót hoz létre a leveleken, amit a továbbiakban *split*-nek nevezünk. Ha mindkét osztály legalább két levelet tartalmaz, akkor a split *nem-triviális*. Buneman régi tétele, hogy bármely X-fát egyértelműen meghatároznak nem-triviális splitjei.

Legyen $q = \{a, b, c, d\}$ egy T-beli levél-négyes. Azt mondjuk, hogy a $t_q = ab|cd$ egy *érvényes* (angolul *valid*) *quartet* split, ha ez a generált $T^*_{|q}$ bináris részfának a valódi, a fában szereplő splitje. Jelölje $Q(T) = \left\{t_q : q \in \binom{[n]}{4}\right\}$ a T X-fa összes érvényes quartet splitjét. A jól ismert klaszszikus eredmény szerint bármely T fára a $Q(T)$ halmaz egyértelműen meghatározza a T-t.

Erre a tényre igen sokféle evolúciós fa rekonstrukciós módszert alapoztak, amelyek sajnos gyakran vezetnek ellentmondáshoz, mivel szinte sohasem sikerül minden quartetre meghatározni az érvényes splitet, az eredmények általában ellentmondóak. Mint az könnyen kiszámítható, ennek oka a "hosszú" quartetek léte. Ennek a problémának a megoldására vezette be kutatócsoportunk (Mike Steel, Székely László, Tandy Warnow és jómagam) a "short quartet" módszereket.

Quartet alapú rekonstrukciós módszereknél alapvetően két problémát kell megoldani. Egyfelöl tudni kell, hogy quartetek milyen (rész)rendszere alkalmas a fa (determinisztikus) meghatározására, másfelöl pedig azt kell eldönteni, hogy quartetek "zajos" rendszeréből hogyan kell kiválasztani azokat, amelyek alkalmasak a fa előbb említett determinisztikus rekonstruálására.

Erre az elvi eljárásra többféle módszer is ismeretes. Egy lehetséges mód az, hogy a rendelkezésre álló érvényes quartet splitekből, az eredeti adatok további vizsgálata nélkül, következtetési szabályok felhasználásával határozzuk meg a többi splitet. Ha példáaul két érvényes splitből gyártunk egy harmadikat, akkor egy *diadikus* szabályt alkalmaztunk.

Azt mondjuk, hogy érvényes quartet splitek egy rendszere *szemi-diadikusan* meghatározza a T fát, ha a legegyszerűbb következtetési szabályok rekurzív alkalmazásával előállítható a fa minden érvényes quartet splitje (és persze csak azok). *Diadikus* előállításról akkor beszélünk, ha még egy, valamivel bonyolultabb szabályt is alkalmazunk. Maga az eljárás, amikor rekurzívan kiszámítjuk az új quartet spliteket az eredeti quartet halmaz *(szemi-)diadikus lezárása.*

A [12] preprint egyik fő eredménye a következő: jelölje $L_T(q)$ a q nevű quartet generálta $T_{|q}$ (nem feltétlenül bináris) részfában a leghosszabb, a $T^*_{|S}$ fában egy élbe összehúzódó út élszámát. Ekkor teljesül:

4. Tétel ([12]). *Legyen $T \in B(n)$ legalább négy levéllel. Jelölje $D(T)$ az összes olyan quartet halmazát, amelyekre $L_T(q) \leq 18 \log n$. Ekkor $D(T)$ szemi-diadikus lezárása a levélszám függvényében polinomiális időben előállítja a fát.*

A tétel lehetővé tette az irodalomban megtalálható első olyan evolúciós fa rekonstrukciós algoritmus megszerkesztését, amelynek teljes valószínűségi analízise elvégzésre került. Az analízis lényeges pontja annak meghatározása, milyen hosszú sorozatok elégségesek a levelek jellemzésére, hogy a rekonstrukciós eljárás lényegében 1 valószínűséggel határozza meg a keresett fát.

Az algoritmus elméleti jelentőségét az adja, hogy - véletlenül - ez az elégséges karakter szám nagyon közel van a szintén ebben a cikkben meghatározott információelméletileg szükséges minimális hosszhoz, ami nagy n estén durván $\log n$. Az is fontos, hogy a futásidő is polinomiális (bár nem túl jó paraméterekkel).

Az 1997-es [14] cikk a 4. Tételre talált jelentős élesítést. Egy T evolúciós fában egy él *mélysége* (*depth*) az éltől a lehető legközelebbi levélhez vezető út élszáma. A fának magának a $d(T)$ *mélysége* pedig a benne található legnagyobb él mélység.

5. Tétel ([14]). *Legyen T egy X-fa n levéllel és legyen*

$$D(T) = \left\{ q \in \binom{[n]}{4} : L_T(q) \leq 2d(T) + 1 \right\}$$

ahol csak olyan 4-levelű részfákat veszünk figyelembe, amelyek középső útja egyetlen élből áll. Ekkor T meghatározható a $D(T)$ szemi-diadikus lezártjából.

A ([15, 16, 17, 18]) cikksorozat részleteiben dolgozta ki a *Short Quartet Módszerek*-t (avagy röviden *SQM*-t). Érdemes itt megemlíteni, hogy a szerzők, Karl Popper szellemében, a séma erősségének tekintették a falszifikálás képességét: a módszer felismerte, ha az input elégtelen vagy ellentmondó.

A [17] cikk teljes általánosságban bebizonyítja az információelméleti alsó korlátot egy X-fa determinisztikus vagy véletlen módszeren alapuló rekonstrukciójához szükséges minimális sorozat-hosszra, majd bebizonyítja a 5. Tétel egy még erősebb változatát.

A cikk ezután leírja az SQM egyik megvalósítását, a Dyadic Closure Tree Construction algoritmust (rövidítve DCTC algoritmust). Az algoritmus eredményeit a következő módon lehet összegezni:

6. Tétel. *Legyen a Q quartet splitek egy rendszere. Ekkor:*

(i) *Ha a DCTC meghatároz egy fát Q-ra, és egy másikat quartet splitek egy bővebb rendszerére is, akkor a két fa megegyezik.*

(ii) *Ha a DCTC eredménye inkonzisztens, azaz ellentmondó quartet splitek is keletkeznek, akkor hasonló történik minden bővebb quartet rendszerre is.*

(iii) *Ha a DCTC nem képes Q-ból kiszámolni a fát, akkor hasonló a helyzet bármely szűkebb quartet rendszerre is.*

(iv) *Végül ha Q ellentmondás mentes és eleme minden reprezentatív quartet, akkor a DCTC előállítja a fát.*

Megjegyzendő, hogy a cikk a DCTC algoritmusra egy $O(n^5)$ implementációt mutat be.

A DCTC algoritmus-magra sokféle faépítő algoritmust lehet alapítani. Ezek mindegyikének quartetek egy-egy Q halmazát kell meghatározni, amely eléggé bő ahhoz, hogy tartalmazza az összes reprezentatív quartetet, de eléggé szűk ahhoz, hogy ne legyen ellentmondó. Az Short Quartet Módszer séma alapfeltevése az, hogyha sikerül a Q meghatározásakor csupa rövid quartet felhasználni, akkor az ellentmodásmentesség automatikusan teljesül.

Egy lehetséges stratégiát a *Diadic Closure Módszer* (DCM) ír le: a DCM egy távolság-becslés alapú eljárással dönti el, hogy mely quarteteket kívánja rekonstruálni, magát a rekonstrukciót pedig a még Buneman által bevezetett un. *four point* módszerrel hajtja végre. Ekkor:

7. Tétel ([17]). *Tegyük fel, hogy a Cavender-Farris modell alatt k karakter fejlődik a T evolúciós fa mentén, ahol minden e élen a változás valószínűségére teljesül $p(e) \in [f, g]$, ahol f és g az n függvényei. Ekkor a DCM módszer $1-o(1)$ valószínűséggel rekonstruálja a T fát, amennyiben a karakterek számára teljesül a*

$$k > \frac{c \cdot \log n}{(1-\sqrt{1-2f})^2(1-2g)^{4depth(T)+6}} \tag{2}$$

összefüggés (ahol c valamilyen rögzített konstans).

Mint a tételből látható, a szükséges sorozat-hossz a fa mélységétől függ, amíg más ismert módszerek hatékonysága általában a fa átmérőjének a függvénye. Ezért a [17] dolgozat ezután két gyakran tekintett valószínűségi eloszlás mellett elemzi a fák mélységét és átmérőjét. A két eloszlás: az egyenletes, ahol minden fa egyformán valószínű, és a Yule-Harding féle, amelynél a "lombosabb" (ezért időben hamarabb kifejlődő) fák valószínűsége nagyobb.

A cikksorozat utolsó cikke ([18]) először különféle távolság alapú fa-rekonstrukciós algoritmusok hatékonyságának összehasonlítására fejleszt ki egy módszert. Az ilyen módszerek általában szólva nem a levelekben lévő karaktersorozatokkal magukkal foglalkoznak, hanem először meghatározzák az egyes levelek egymástól való "távolságát", amely a sorozatok "nem hasonlóságán" (*dissimilarity*) alapulnak: minél kevésbé hasonló két sorozat, annál nagyobb a távolságuk.

A cikk fő hozzájárulása a quartet módszerek témájához egy újonnan fejlesztett algoritmus a *Witness-Antiwitness Method* (WAM). Az algoritmus valószínűségi elemzése azt mutatja, hogy a WAM sikeresen képes rekonstruálni a fát a DCM eljárásával lényegében megegyező paraméter tartományban, méghozzá lényegesen gyorsabban, mint a DCM. Az is lényeges, hogy eközben a szükséges sorozat-hossz csak kicsit múlja felül a DCM-nél szükségeset.

Az SQM módszerek eddig jelentős hatást mutattak az evolúciós fák rekonstrukciójának kutatásában. Az egyik legelső példa erre a Disk Covering Method, amely módszer az SQM alapján egyéb ismert módszerek heurisztikus fel-

gyorsitását igéri. Az E. Mossel vezette Berkeley-beli kutatócsoport egy sorozat cikkben jelentősen kiterjesztette az SQM-ben kifejlesztett elveket.

Összességében úgy gondolom, hogy az ebben a szakaszban kifejett eredmények a legfontosabbak a disszertációban.

X-fák és súlyozott quartetek

A fejezet utolsó szakaszában egy Andreas Dress-szel közös eredményt ismertetek ([20]). Legyen X egy véges halmaz és jelölje $\mathcal{S}_{2|2}(X)$ az X össszes négyeseiből megalkotható 2-2 splitet, azaz

$$\mathcal{S}_{2|2}(X) := \left\{ \left\{ \{a,b\},\{c,d\} \right\} \middle| \{a,b\},\{c,d\} \in \binom{X}{2};\ \{a,b\} \cap \{c,d\} = \emptyset \right\},$$

Jelölje $E_1 = E_1(T)$ a T fa összes belső élét, legyen továbbá $\ell : E_1 \to \mathbb{R}_{>0}$ egy tetszőleges, de szigorúan pozitív, valós *hossz-függvény*. Minket az a $W = W_{T,\ell}$ függvény érdekel, amelyet a következő módon definiálunk $\mathcal{S}_{2|2}(X)$-en:

$$W : \mathcal{S}_{2|2}(X) \to \mathbb{R}_{\geq 0} : ab|cd \mapsto \sum_{e \in E(ab|cd)} \ell(e) \tag{3}$$

ahol az összegzés a $E(ab|cd)$ halmazra történik, amely az összes olyan $e \in E$ élt tartalmazza, amely a T fában szeparálja az a, b leveleket a c, d levelektől. A W függvény nyilván a $T|_{\{abcd\}}$ részfa "középső részének" hosszát méri, amennyiben a $ab|cd$ egy érvényes split, egyébként pedig nulla az értéke.

A cikk fő megfigyelése, hogy a hossz-függvény axiomatizálható: van néhány olyan, könnyen látható tulajdonsága, amely biztosítja, hogy az ezeket kielégítő nem-negatív valós függvények ilyen hossz-függvényként állíthatók elő.

3. Szavak rekonstrukciója - DNS kódok

A szavak kombinatorikája (combinatorics on words) széles körben vizsgált, jól megalapozott területe a matematikának. A vizsgált objektum általában egy véges $\Gamma = \{1, 2, \ldots, k\}$ ábécén értelmezett összes véges *szó* (avagy *sorozat*) Γ^* összessége alkotta végtelen poset, amelyet a *részsorozatnak lenni* reláció rendez el.

Ugyanezen objektumok fontos szerepet játszanak a molekuláris biológia alapvető problémáiban is. Ilyenkor a vizsgálandó rendszert leíró biológiai sorozatok a négy nukleotidát (A, C, G, T) tartalmazhatják. Ha DNS helyett RNS sorozatokat vizsgálunk, akkor a T (azaz tymine) helyett U (azaz uracyl) szerepel a sorozatokban. A sorozatok (vagy szavak) vehetik betűiket az aminosavakból is (az emberi szervezetben ebből húsz féle létezik, de az összes élőlényben sem ismeretes 26-nál több). Továbbá tekinthetjük a kromoszómákon előforduló géneket is, ahol a valódi biológiai sorozatokban az egyes gének egynél nagyobb multiplicitással és kétféle irányítással is szerepelhetnek. Ezeknél a sorozatoknál

különféle véges optimalizálási számításokat kell elvégezni. Ezekkel a feladatokkal a *string* (*fűzér*) algoritmusok tudománya foglalkozik.

Hibákat is megengedő paraméteres párositások

Ebben a szakaszban a string elmélet egyik alapvető problémájának egy általánosítását tárgyalom a [24] cikk alapján. A különféle string keresések a számítógépes eljárások egyfajta alapvető "primitívjei": olyan építőelemek, amelyeket a legkülönfélébb eljárásokban használnak. A szokásos megfogalmazásánál adott egy (általában hosszú) *szöveg (text)*, és egy (általában sokkal rövidebb) *minta (pattern)*, ahol a minta összes szövegbeli előfordulását kell megtalálni. Ezt hívják a minta *párosítás*ának. Az alapprobléma sokféle változata ismert: megengedhetünk például korlátos számú hibát a minta előfordulásában, vagy törléseket illetve beszúrásokat is. A paraméteres változatban a szöveg és a minta ábécéje különbözhet egymástól, és akkor gondoljuk, hogy egy adott pozicióban a minta megjelenik a szövegben, hogyha létezik a két ábécé között olyan injektív leképezés, ami teljes aznosságot garantál. A probléma a software engeneeringben, programok tömörítésénél merült fel.

A *közelítő* (*hibákat megengedő*) paraméteres párosítás a következő feladatot jelenti: legyen $t = t_1t_2...t_n$ egy (hosszú) szöveg és legyen $p = p_1p_2...p_m$ egy (rövidebb) minta, amelyek az (esetleg) eltérő Σ_t és Σ_p ábécé fölöttiek. Ezután mindegyik i szöveg-pozicióhoz keressük azt a $\pi_i : \Sigma_p \to \Sigma_t$ injekciót, amely maximalizálja a megegyezések számát a $\pi_i(p)$ leképzett minta és a $t_it_{i+1}...t_{i+m-1}$ szövegdarab között ($i = 1, 2, ...n - m + 1$).

A probléma általános esete könnyen megoldható $O(nm(\sqrt{m}+\log n))$ lépésben, ha a kérdést a szöveg minden poziciójában visszavezetjük páros gráfok maximális súlyú párosításaira (ez már 1974-ben is ismert volt).

A [24] cikk azt az esetet vizsgálja, amikor mind a szöveg, mind a minta futamokkal van kódolva: megadjuk az első pozicióban levő betű megszakítás nélküli, (maximális számú) egymást követő előfordulásainak számát, majd megadjuk a rákövetkező betűt, és annak a multiplicitását, stb. Jelölje r_t és r_p a szövegben illetve a mintában jelenlevő futamok számát.

A dolgozat egy $O(r_p \times r_t)$ idő komplexitású algoritmust fejleszt ki arra az esetre, amikor legalább az egyik ábécé bináris. A futásidőt terheli még egy (szöveghosszban) lineáris előkészítő fázis, továbbá egy logaritmikus szervezési overhead.

Szavak rekonstrukciója - klasszikus eset

A Sziklai Péterrel és David Torney-val közös [19] cikk a véges Γ ábécéből vett szavak alkotta véges posetekkel foglalkozik: legyen $\mathcal{P}^{(n)}$ az ábécé betűiből vett összes, legfeljebb n hosszú sorozat részben rendezett halmaza. A kapott posetben a szavak hossza egy alkalmas rang függvényt határoz meg, ezért a $\mathcal{P}^{(n)}$ poset szintezett. Jelölje $\mathcal{P}_i^{(n)}$ az i-edik szintet, amely az összes i hosszú részsorozatból áll ($0 \leq i \leq n$).

Míg a végtelen változat napjainkban rengeteget vizsgált objektum, addig a véges változat szinte semmilyen figyelmet sem kapott. Jelentőségét többek között az adja, hogy a DNS vizsgálatokban használt *törlés - beszúrás* (*delition-insertion*) metrikán (avagy Levenshtein távolságon) alapuló hibajavító kódok tanulmányozásának természetes közege lehet.

A dolgozat először is meghatározta a $\mathcal{P}^{(n)}$ poset automorphismus csoportját, közben új, egyszerű bizonyítást adott Burosch és kollégáinak régebbi eredményeire kételemű ábécék felett. A módszer továbbfejleszthető az általános ábécé esetére is, ezt Ligeti Péter és Sziklai Péter végezte el.

Ezután a poset klasszikus kombinatorikai tulajdonságait vizsgáltuk meg. Könnyen látható, hogy az azonos hosszú szavak eltérő méretű (alsó) árnyékokkal rendelkezhetnek. Ugyanakkor teljesül, hogy:

8. Tétel. *Legyen ξ egy rögzített sorozat és legyen j olyan egész, hogy $|\xi| \leq j \leq n$. Ekkor azon j-sorozatok száma, amelyek ξ-t részsorozatként tartalmazzák a következő:*

$$N(j, \xi; k) = \sum_{i=0}^{j-|\xi|} \binom{j}{i} (k-1)^i.$$

Ennek köbetkezményeként azt is meg lehetett mutatni, hogy a poset rendelkezik a normalizált matching tulajdonsággal, ezért BLYM tulajdonságú is.

Szavak rekonstrukciója lineáris időben

Ebben a részben az Andreas Dressel közös [22] cikk alapján a véges Γ ábécé feletti n-hosszú szavak részszavaiból lineáris időben történő rekonstrukcióját tárgyalom.

Simon Imre 1975-ben megmutatta, hogy a véges Γ ábécé felett minden $2m+1$ hosszú szót egyértelműen meghatároz legfeljebb $m+1$ hosszú részszavainak halmaza. Érdemes megjegyezni, ha a részszavak halmazán kívül minden egyes részszó multiplicitását is ismerjük, akkor minden szót egyértelműen meghatároz a legfeljebb $\sim 7\sqrt{n}$ hosszú részszavainak kollekciója.

Az ismert megközelítések csupán egzisztencia bizonyítást adtak Simon tételére, azonban nem vizsgálták a rekonstrukciót ténylegesen végrehajtó algoritmust. A jelzett cikkben megmutattuk, hogy ha

9. Tétel. *Adott a legalább kételemű Γ ábécé, továbbá az n és m természetes számok, ahol $2m > n$, akkor bármely $w \in \Gamma^{[n]}$ szó rekonstruálható $|\Gamma| + 2n$ kérdéssel legeljebb m hosszú részszavainak halmazából.*

Szavak rekonstrukciója - fordított komplemens eset

Ebben a szakaszban a [25] cikk eredményeit ismertetem. Legyen $\Gamma = \{a, \bar{a}; b, \bar{b}\}$ ahol a betűk un. *komplemens párokban* vannak. Definiáljuk a következő műveleteket: $\bar{\bar{a}} = a$, $\bar{\bar{b}} = b$ továbbá valamely $w = w_1 w_2 ... w_t$ szóra legyen $\widetilde{w} = \overline{w_t}\,\overline{w_{t-1}}$ ... $\overline{w_1}$, amelyet az eredeti szó *fordított* (*reverse*) *komplemens*ének nevezünk. Könnyen látható, hogy $\widetilde{(\widetilde{w})} = w$. Ezután minden szót azonosítunk

a fordított komplemensével. Ezek után a fordított komplemens rendezésben $w \prec v$ (azaz az első *megelőzi* a másodikat) akkor és csakis akkor teljesül, ha w részszava v-nek vagy részszava $\widetilde{v}$-nek. Jelölje most $S(m, w)$ mindazon legfeljebb m hosszú v szavakat, amelyek megelőzik w-t (azaz vagy w vagy $\widetilde{w}$ szavak részszavai). A Simon Imre tételének megfelelő kérdés az, hogy milyen hosszű w szavakat lehet biztosan rekonstruálni az $S(m, w)$ halmazból. A cikk egyik fő eredménye a következő állítás:

10. Tétel. (i) *Minden legfeljebb* $3m-1$ *hosszú* $w \in \{a, \bar{a}\}^*$ *szót egyértelműen meghatároz a hossza, továbbá részszavainak* $S(2m, w)$ *halmaza.*
(ii) *Minden legfeljebb* $3m+1$ *hosszú* ($m > 1$) *szót, amely tartalmaz betűt mind az* (a *vagy* $\bar{a}$) *mind a* (b *vagy* $\bar{b}$) *párból, egyértelműen meghatároz a hossza, továbbá részszavainak* $S(2m, w)$ *halmaza.*

Az utóbbi állítás akkor is igaz marad, ha a w szó $k \geq 2$ különféle komplemens párból tartalmaz betűket. Érdemes megjegyezni, hogy a bizonyításokban a nehézséget mindenütt az jelenti, hogy bár sok (megelőző) részszó van jelen, nem tudjuk róluk, hogy a szónak, vagy annak fordított komplemensének a részszavaie. Ez ad magyarázatot arra is, miért kell ennyivel hosszabb részszavakat ismernünk a fordított komplemens esetben. Azt is érdemes hozzátenni, hogy ebben az esetben még nem ismeretes a rekonstrukció komplexitása.

DNS kódok

Az előző szakaszban leírt részbenrendezés a szokásos Levenshtein (vagy delition - insertition) metrikához hasonló távolság fogalmat eredményez. Itt is lehet ennek megfelelően hibajavító kódokat keresni. Ezeknek már a Human Genome program idején nagy gyakorlati hasznunk volt, és megkonstruálásuk kézzel, heurisztikus alapon történt. A sokszerzős [21] cikk ennek a problémának próbált elméleti megalapozása lenni. Fő célja a fogalmak és feladatok rögzítése volt. A téma meglepően népszerű, a cikk megjelenése óta eltelt szűk egy évben már jónéhány hivatkozás történt rá.

Hivatkozások

[1] P.L. Erdős - L. A. Székely: Evolutionary trees: an integer multicommodity max-flow – min-cut theorem, *Advances in Appl. Math* **13** (1992) 375-389.

[2] P.L. Erdős - L.A. Székely: Algorithms and min-max theorems for certain multiway cuts, *Integer Programming and Combinatorial Optimization* (Proc. of a Conf. held at Carnegie Mellon University, May 25-27, 1992, by the Math. Programming Society, ed. by E. Balas, G. Cornuèjols, R. Kannan) 334-345.

[3] M.A. Steel - M.D. Hendy - L.A. Székely - P.L. Erdős : Spectral analysis and a closest tree method for genetic sequences, *Appl. Math. Letters* **5** (1992), 63-67.

[4] L.A. Székely - P.L. Erdős - M.A. Steel: The combinatorics of evolutionary trees–a survey, *Séminaire Lotharingien de Combinatoire, (Saint-Nabor, 1992)*, D. Foata, éd, Publ. Inst. Rech. Math. Av. **498** (1992), 129–143.

[5] L.A. Székely - P.L. Erdős - M.A. Steel - D. Penny: A Fourier inversion formula for evolutionary trees, *Appl. Math. Letters* **6** (1993), 13-17.

[6] L.A. Székely - M. Steel - P.L. Erdős: Fourier calculus on evolutionary trees, *Advances in Appl. Math* **14** (1993), 200-216.

[7] P.L. Erdős - L. A. Székely: Counting bichromatic evolutionary trees, *Discrete Applied Mathematics* **47** (1993), 1-8.

[8] M.A. Steel - L.A. Székely - P.L. Erdős - P. Waddell: A complete family of phylogenetic invariants for any number of taxa, *NZ Journal of Botany*, **31** (1993), 289-296.

[9] P.L. Erdős : A new bijection on rooted forests, *Discrete Mathematics* **111** (1993), 179-188.

[10] P.L. Erdős - L. A. Székely: On weighted multiway cuts in trees, *Mathematical Programming* **65** (1994), 93-105.

[11] L.A. Székely - P.L. Erdős - M.A. Steel: The combinatorics of reconstructing evolutionary trees, *J. Comb. Math. Comb. Computing* **15** (1994), 241-254.

[12] M.A. Steel - L.A. Székely - P.L. Erdős: The number of nucleotide sites needed to accurately reconstruct large evolutionary trees, *DIMACS, Rutgers University, New Brunswick, New Jersey, USA* 1996.DIMACS Technical Reports 96-19

[13] P.L. Erdős - A. Frank - L.A. Székely: Minimum multiway cuts in trees, *Discrete Appl. Math.* **87** (1998), 67–75.

[14] P.L. Erdős - M.A. Steel - L.A. Székely - T.J. Warnow: Local quartet splits of a binary tree infer all quartet splits via one dyadic inference rule, *Computers and Artificial Intelligence* **16** (1997), 217-227.

[15] P.L. Erdős - K. Rice - M.A. Steel - L.A. Székely - T.J. Warnow: The Short Quartet Method, to appear in Math. Modelling and Sci. Computing Special Issue of the papers presented at the Computational Biology sessions at the 11th ICMCM, March 31 - April 2, 1997, Georgetown University Conference Center, Washington, D.C., USA.

[16] P.L. Erdős - M.A. Steel - L.A. Székely - T.J. Warnow: Constructing big trees from short sequences, *Automata, Languages and Programming* 24th International Colloquium, ICALP'97, Bologna, Italy, July 7 - 11, 1997, (P. Degano,; R. Gorrieri, A. Marchetti-Spaccamela, Eds.) Proceedings (Lecture Notes in Computer Science. Vol. 1256) (1997), 827-837.

[17] P.L. Erdős - M.A. Steel - L.A. Székely - T.J. Warnow: A few logs suffice to build (almost) all trees (I), *Random Structures and Algorithms* **14** (1999), 153-184.

[18] P.L. Erdős - M.A. Steel - L.A. Székely - T.J. Warnow: A few logs suffice to build (almost) all trees (II), *Theoretical Computer Science*, **221** (1-2) (1999), 77–118.

[19] P.L. Erdős - P. Sziklai - D. C. Torney: A finite word poset, *Electr. J. Combinatorics*, **8** No 2. (2001), R# 8.

[20] A.W.M. Dress - P.L. Erdős: X-trees and Weighted Quartet Systems, *Ann. Combin.* **7** (2003), 155-169

[21] A.G. D'yachkov - P.L. Erdős - A.J. Macula - V.V. Rykov - D.C. Torney - C-S. Tung - P.A. Vilenkin - P. Scott White: Exordium for DNA Codes, *J. Comb. Opt.* **7** (4) (2003), 369–379.

[22] A.W.M. Dress - P.L. Erdős: Reconstructing Words from Subwords in Linear Time, *Annals of Combinatorics*, **8** (4) (2004), 457–462.

[23] P.L. Erdős - P. Ligeti - P. Sziklai - D.C. Torney: Subwords in reverse complement order - extended abstract, invited paper to *Proc. Conf. on "Combinatorial and Algorithmic Foundations of Pattern and Association Discovery" - Schloss Dagstuhl, International Conference And Research Center For Computer Science, Germany* May 14-19. 2006, 1–7.

[24] A. Apostolico - P.L. Erdős - M. Lewenstein: Parameterized Matching with Mismatches, *J. of Discrete Algorithms* **5** (2007), 135–140.

[25] P.L. Erdős - P. Ligeti - P. Sziklai - D.C. Torney: Subwords in reverse complement order, *Annals of Combinatorics* **10** (2006) 415–430.