

EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR

Filogenetikai és ultrametrikus fák

Szakdolgozat
Matematika BSc

Mohay Lili Veronika

Témavezető:
Tóthmérész Lilla

Operációkutatási Tanszék



Budapest, 2025

Köszönetnyilvánítás

Elsősorban szeretném megköszönni témavezetőmnek, Tóthmérés Lillának a szakmai és formai tanácsokat, a rendszeres megbeszéléseket, ahol elmondhattam ötleteimet, meghallgatott és hasznos javaslatokkal, ötletekkel látott el. Köszönöm továbbá szüleimnek támogatásukat és megértésüket.

Tartalomjegyzék

1. Faépítés	4
1.1. Filogenetikai és ultrametrikus fák	4
1.2. Ultrametrikus fa	11
2. Filogenetikai fák tere	19
2.1. A $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ tér és a trópusi metrika	19
2.2. Trópusi geometria, konvexitás és szakasz	21
2.3. Trópusi szakasz tulajdonságai	22
2.4. Trópusi szakasz és ultrametrikus fák kapcsolata	27
2.5. Trópusi konvex burok és az ultrametrikus fák kapcsolata	32
2.6. BHV tér, fa, metrika és a trópusi metrikával a kapcsolata	38
3. Összefoglalás	43

Bevezetés

A filogenetikai fákat főleg biológiai és kémiai folyamatok (pl. evolúciós kapcsolatok, molekulák, vírusok kialakulása, fejlődése) vizsgálatánál és leírásánál használják. A filogenetikai fák a gráfelméletből ismert, többnyire gyökeres, élsúlyozott fákat jelentik, amiknek gyökeréből lefele ágaznak ki az élek, címkézett leveleit pedig utódoknak, fajoknak nevezzük. Többféle folyamatot is ábrázolhatnak, mint például közös tulajdonság öröklődése a gyökértől a levelekig tartó utakon, fajok szétválása és a köztük eltelt idő szemléltetése, stb. A dolgozat utóbbival foglalkozik. A témakör két alapvető kérdése a következő: ha van egy adathalmazunk, ami azt írja le, hogy az egyes fajok mikor vagy mennyi ideje váltak el egymástól, akkor abból visszafejthető-e a fa, ami lerajzolja a folyamatot és ráírhatóak-e pozitív élsúlyok az élekre úgy, hogy az adatoknak megfelelő "evolúciós" úthosszakot adjunk vissza. Látni fogjuk, hogy ez ultrametrikus távolságnál lehetséges, sőt hatékony algoritmus létezik rá (1. fejezetben definiáljuk a fogalmat és bizonyítjuk). Azonban nem minden esetben tartozik hozzá fa, mert például ha pontatlanok a mérések, akkor kaphatunk olyan adathalmazt is, amihez nem létezhet fa (pl. negatív élsúly keletkezne). Ilyenkor közelítő fákat kell keresni, de erre nem térünk ki részletesebben. Előfordulhat, hogy létezik az adatokhoz fa, csak az a valóságtól eltérő fát adunk vissza, ezért ilyenkor is egy "közelítő" fát keresünk. A dolgozat egy speciális tulajdonságú adathalmazzal foglalkozni, az ultrametrikus távolsággal és fakkal.

A másik jelentős probléma, hogy hasonló szerkezetű fák között szeretnénk valami kapcsolatot találni aszerint, hogy milyen "messze" vannak egymástól. Több cikk is foglalkozott ezzel a problémával, mivel az euklideszi tér és az ott ismert metrikák nem bizonyultak eredményesnek, mert a kis távolságból nem feltétlenül következett a fák alakjának hasonlósága. A nem-euklideszi térhez tartozó metrikák mellett a másik nehézséget az okozza, hogy már a gyökeres, bináris, címkézett, n levelű fák száma is $(2n - 3)!! = 1 \cdot 3 \cdot \dots \cdot (2n - 3)$ ([14]). Ezért ha egy mérési hibából más fát kapunk, akkor nagy n esetén nagyon sok fa jöhetne szóba, mint "igazi" fa.

A dolgozat egyik legfőbb célja, hogy a sokféle, de ugyanarról értekező szakirodalmat összekösse és lényeges eredményeket foglaljon össze. Először az ultrametrikus fák konstruálását mutatjuk be, majd a *trópusi geometria* segítségével hatékony "legrövidebb útkeresést" adunk a *filogenetikai fák* (egy speciális *terében*, aminek segítségével vizsgálhatjuk a megmaradó tulajdonságokat két közeli fa között, valamint ha több mérésünk is lenne, bizonyítjuk, hogy létezik olyan ultrametrikus fa, ami mindegyikhez "közel" van.

1. Faépítés

1.1. Filogenetikai és ultrametrikus fák

1.1.1. Definíció. ([9]) Legyen $n \in \mathbb{N}$ és $2 \leq n$. n levelű filogenetikai fáknak nevezzük azokat az n db címkézett levelből álló fákat, amiknek éleire nemnegatív súlyokat írunk. Jelölésük: T_n

A fák nem levél csúcsait *belső csúcsoknak* nevezzük, azokat az éleket, amiknek mindkét végpontja belső csúcs *belső éleknek* hívjuk, a többi *külső éleknek*. A szakirodalomban néha máshogy definiálják a filogenetikai fákat, mivel a belső csúcsokhoz is szoktak számokat rendelni, ami összefüggésben van az élekre írt súlyokkal. Az eltérő definiálást az okozza, hogy van ahol előbb az élekre írunk nemnegatív súlyokat, aztán az alapján rendelünk számokat a belső csúcsokhoz, de van ahol fordítva. Erről a filogenetikai fák egy részcsoportjánál, az ultrametrikus fáknál az 1.2.1. Tételben lesz még szó. Mi a továbbiakban azt feltételezzük, hogy az éleken vannak megadva nemnegatív súlyok (amik azt jelentik, mennyi idő telt el két belső csúcs/esemény között). Ha valahol 0 lenne, ott gondolhatunk úgy rá, hogy a két belső csúcs egyben van.

A továbbiakban fontos lesz a következő mátrix.

1.1.2. Definíció. ([9]) Ha adva van egy T filogenetikai fa, $n \geq 2$ levéllel, akkor csináljunk egy olyan $n \times n$ D mátrixot, aminek i . sorában a j . oszlopban az i és j levelek közti (egyértelmű) út hossza lesz (jelöljük $\omega_{i,j}$ -vel). Ez szimmetrikus és a főátlóban végig nulla szerepel. Valamint a nemnegatív élsúlyozás miatt bármely indexhármásra teljesülni fog a háromszög egyenlőtlenség. Ezért a filogenetikai fákat szokás metrikus fáknak is nevezni, hiszen D mátrix $\omega_{i,j}$ értékeire teljesülnek a metrika tulajdonságai. Ezt a mátrixot szokás T -hez tartozó távolság mátrixnak is nevezni.

1.1.3. Megjegyzés. A metrika tulajdonságai, hogy $\omega_{i,j} = 0$ pontosan, akkor ha $i = j$, de ez teljesül, hiszen a pozitív súlyozás miatt egy levélnek csak önmagától lehet nulla a távolsága. Továbbá a szimmetria is teljesül, hiszen az i, j levelek közti út hossza független attól, hogy melyik irányból nézzük. A háromszög egyenlőtlenség azért igaz, mert ha van egy filogenetikai fánk, benne i, j, k levelek, akkor azokra megszorítva a fát egy háromágú csillagot kapunk (3. ábra a, részén látható fát "kiegyenesítjük" a gyökérsínél). Ekkor ha $d(v, v')$ -vel jelöljük a v, v' csúcsok közti út hosszát a fában, akkor

$$\omega_{i,j} = d(i, v) + d(j, v) \leq d(i, v) + d(v, k) + d(v, k) + d(j, v) = \omega_{i,k} + \omega_{j,k}.$$

Így valóban teljesülnek a metrika tulajdonságai a D mátrixra.

Eddig ha egy fa távolságmátrixának egy értékére hivatkoztunk a $\omega_{i,j}$ jelet használtuk, de a következőkben, ha egy mátrix értékére akarunk csak utalni (amiről nem feltétlenül tudjuk, hogy távolság mátrix-e), akkor a $D(i, j)$ jelölést fogjuk használni.

1.1.4. Definíció. ([5, Definition 5.]) Adva van egy D $n \times n$ valós mátrix, aminek i . sorában és j . oszlopában a számot $D(i, j)$ jelöli. A **4-pont feltétel**

(**Four-point condition**) azt jelenti, hogy ha veszem a tetszőleges, de különböző $1 \leq i, j, k, l \leq n$ indexeket, akkor

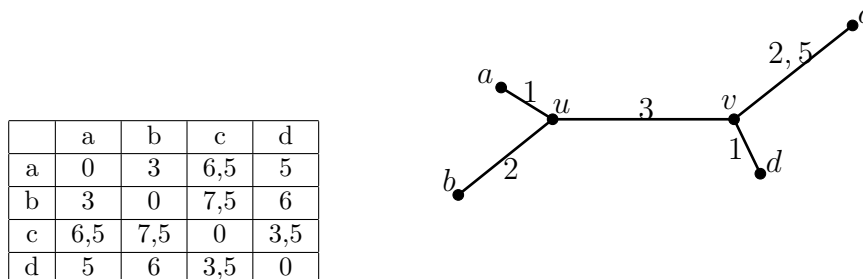
$$D(i, j) + D(k, l) \leq \max(D(i, k) + D(j, l), D(i, l) + D(j, k)),$$

vagy ekvivalensen a maximum legalább kétszer lesz felvéve a $D(i, j) + D(k, l)$, $D(i, k) + D(j, l)$, $D(i, l) + D(j, k)$ értékek között.

1.1.5. Állítás. ([9, Proposition 1.]) Egy D szimmetrikus, valós $n \times n$ mátrix egy filogenetikai fát reprezentál, ha teljesíti a metrika feltételeit és a 4-pont feltételt.

Ennek bizonyításába nem megyünk bele ([2] forrásban olvasható a bizonyítás), mert minket egy speciálisabb fa és mátrix érdekel, aminél bizonyítani fogunk egy hasonló állítást (1.1.8. Tétel).

Az 1. ábrán látható egy filogenetikai fa négy levéllel, ami például nem gyökeres fa, mert nem jelöltünk ki egy csúcsot sem erre a szerepre. Hogy ne keverjük össze a távolság mátrix értékekkel, a leveleket az abc betűivel jelölöm most, de a későbbiekben többnyire számokkal fogom.



1. ábra. ([9])Távolság mátrix és filogenetikai fa. A levelek: a, b, c, d és a belső csúcsok: u, v. Itt csak a 3 súlyú él belső él.

A filogenetikai fák általánosságban olyan fákat jelölnek, amelyek valamilyen evolúciós változást vagy kapcsolatot írnak le. A következőkben definiáljuk az ultrametrikus fákat, amik a filogenetikai fáknak egy részcsoportja (1.1.11. Állítás) és majd látni fogjuk, hogy ezek valóban egy evolúciós folyamatot írnak le, a kezdetektől a végpontig (1.2.3. és 1.2.4. Definícióknál és az előttük lévő bekezdésekben lesz róla bővebben szó). Most megadjuk az ultrametrikus mátrix és fa definícióját, majd belátjuk, hogy a két fogalom ekvivalens (mint a távolság mátrix és a filogenetikai fa). Azután megmutatjuk, hogy az ultrametrikus fánál teljesülő *3-pont feltétel* erősebb, mint a *4-pont feltétel*.

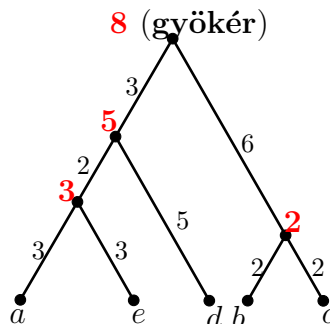
1.1.6. Definíció. ([4, 451.oldal]) Legyen D egy szimmetrikus $n \times n$ nemnegatív, valós mátrix. Azt mondjuk, hogy ultrametrikus távolságot definiál, ha bármely három i, j, k indexre a $D(i, j)$, $D(i, k)$, $D(j, k)$ értékekből a maximum legalább kétszer fordul elő. A szakirodalomban ezt **3-pont feltételnek** (**Three-point condition**) nevezik. Ilyenkor D -t ultrametrikus mátrixnak hívjuk.

A következő definíciónál nem egyértelmű, hogy miért kell ilyen tulajdonságúnak lennie az ultrametrikus fának, de ugyancsak a 1.2.4 Definíciónál fogjuk látni ennek az okát.

1.1.7. Definíció. ([4, 449. oldal]) Legyen D egy szimmetrikus $n \times n$ nemnegatív, valós mátrix. T gyökeres fát D ultrametrikus fájának nevezzük, ha a következők teljesülnek rá:

- 1. T -nek pontosan n levele van és mind D 1-1 sorával van címkézve.
 - 2. Minden belső csúcsnak legalább két gyereke van és D egy mezőjében szereplő számmal vannak címkézve.
 - 3. A gyökérből indulva bármely levélíg az útmenti belső csúcsokon a címkézések szigorúan csökkennek.
 - 4. Bármely i, j levélpárra a legközelebbi közös ősök címkéje $D(i, j)$.
- Az n levelű ultrametrikus fákat U_n -nel szoktuk jelölni.

	a	b	c	d	e
a	0	8	8	5	3
b	8	0	2	8	8
c	8	2	0	8	8
d	5	8	8	0	5
e	3	8	8	5	0



2. ábra. Ultrametrikus mátrix és a hozzá tartozó fa, a piros számok a belső csúcsok címkézése. Az élek súlyozása nem szerepelt a definícióban, de a későbbiek lesz róla szó. Viszont azt már most megjegyezzük, hogy nem véletlen, hogy a piros számok a belső csúcsnak a levél utódjaitól a távolsága.

Az előbbi definícióból következik, hogy ha T egy ultrametrikus fa és bizonyos leveleknek a páronkénti legközelebbi közös őse megegyezik, akkor a hozzájuk tartozó $D(i, j)$ értékek szükségszerűen meg kell, hogy egyezzenek. A 2. ábrán az a, d és e, d levélpároknak a legközelebbi közös őse megegyezik ($D(a, d) = D(e, d)$), de a, e őse nem azonos vele, és $D(a, e)$ értéke kisebb az övéknél. Az ultrametrikus fákat a szakirodalomban általában binárisnak veszik (minden belső csúcsnak, ami nem a gyökér 3 a foka, a gyökérnek 2). Bár a definíciókat általánosan mondom ki és a tételek, állítások is igazak többnyire többgyerekes belső csúcsok esetén is, mi is inkább a bináris fákat használjuk a bizonyításoknál. A szakirodalomban pozitív számok szoktak a mátrixban lenni, kivéve a főátlót, ahol nullák vannak. Nyugodtan feltehetjük, hogy (mindkét definícióbeli) D mátrix főátlójában nulla szerepel. Az elsónél azért, mert ha i, j

indexeket nézem, akkor $D(i, j) = D(j, i) \geq 0$, így $D(i, i) = 0$ nem mond ellent a 3-pont feltételnek. Valamint az i, i levélindexeknek nem igazán értelmezhető a legközelebbi közös őse, hiszen az az előbbi belső csúcs a gyökérhez vezető úton, de annak lesz egy másik gyereke ($j \neq i$) és ezért arra a belső csúcsra a $D(i, j) > 0$ érték fog kerülni. Így a $D(i, i)$ értéket nem szoktuk felírni a fára és az értéke éppen ezért nem is fontos, definiálhatjuk nullának. A nem egyenlő i, j indexekre pedig innentől $D(i, j) > 0$ feltehető, mert a szakirodalomban is ezt fel szokták tenni.

1.1.8. Tétel. ([4, Theorem 17.1.1.]) Egy D $n \times n$ szimmetrikus (nemnegatív) valós mátrixnak pontosan akkor van ultrametrikus fája, ha D ultrametrikus mátrix.

Bizonyítás. Legyen T ultrametrikus fája D -nek a 1.1.7 Definíció szerint és lássuk be, hogy D ultrametrikus a 1.1.6 Definíció szerint. Ekkor ha veszek három indexet (i, j, k) , azok T három levelének felelnek meg. T -t megszorítom a három levélre és a gyökérből hozzájuk vezető (egyértelmű) utakra úgy, hogy a gyökértől az első szétválásig tartó utat levágom, hiszen addig egyik belső csúcs sem volt legközelebbi közös ősök, vagyis nincs rajta $D(i, j)$, $D(i, k)$ vagy $D(j, k)$ érték. Ekkor szimmetria miatt feltehető, hogy a 3. ábra a , részén látható fát kapom (csak a legközelebbi közös ősök és a levelek láthatóak rajta). A legfelső belső csúcsra a $D(i, k)$ és $D(j, k)$ értékek kerülnek (amik egyenlőek, mert csak egy szám kerülhet a belső csúcsokra), a v csúcsra a $D(i, j)$. Mivel a fa ultrametrikus, így $D(i, j) < D(i, k) = D(j, k)$, tehát a maximum kétszer lesz felvéve, D ultrametrikus mátrix.

Visszafele egy algoritmust is adunk, hogy ha adott egy D (1.1.6 Definíció szerint) ultrametrikus mátrix, akkor abból hogyan kapható hozzá tartozó T (1.1.7 Definíció szerinti) ultrametrikus fa.

Először egy konkrét i indexet vizsgálunk. Az előbbiekből láthattuk, hogy $D(i, i) = 0$ feltehető, ekkor minden $j \neq i$ indexre $D(i, j) > D(i, i)$. Tegyük fel, hogy m db különböző érték szerepel D i . sorában. Ha létezne D -hez egy T ultrametrikus fa, akkor a gyökértől az i levélhez vezető úton szerepelhetnek csak ezek a $D(i, j)$ számok, hiszen a nem az úton szereplő belső csúcsoknak nem lesz ármazottja az i , így nem kerülhet rájuk $D(i, \cdot)$ szám (1.1.7 Definíció). Valamint az út minden belső csúcsánál az egyik irányba haladva i levélbe jutunk előbb-utóbb, egy másik irányba egy $j \neq i$ levélbe. Ezért arra a belső csúcsra kerül a $D(i, j)$ érték, hiszen ő a legközelebbi közös őse i -nek és j -nek. Valamint minden levélbe el lehet jutni ha i -ből felmegyünk a legközelebbi közös ősökig (legfeljebb a gyökérig) és onnan lemegyünk a j -be, így tényleg minden $j \neq i$ -re $D(i, j)$ szerepelni fog az út egyik belső csúcsán. Tehát a gyökértől az i levélhez tartó úton szereplő belső csúcsokon lesz az összes i . sorban szereplő szám ($D(i, i) = 0$ nem szerepel a föntiek szerint). Ez azt is jelenti, hogy mivel m különböző számunk van, és az úton szereplő számoknak szigorúan csökkenniük kell, így pontosan m db csúcs van ezen az úton és mindhez fentről lefelé a következő legnagyobb $D(i, j)$ érték tartozik (az m . csúcs maga az i levél, amihez a " $D(i, i)$ " tartozik, csak nem szoktuk felírni rá). Ekkor a maradék $n - 1$ indexet szét lehet osztani $m - 1$ db csoportba, aszerint, hogy $D(i, j)$ érték melyik útbeli

belső csúcshoz tartozik (hanyadik legkisebb érték D i . sorában). Tehát olyan algoritmus kellene, ami megkonstruálja azt a fát, aminél ha $D(i, j) = D(i, k)$, akkor j és k útja ugyanannál a belső csúcsnál szakad el i útjától és az így létrejött fa D -értékeire teljesülnek a 1.1.7 Definíció 3. feltételei (a többi feltétel az előző "index-csoportra bontás" miatt látszik, hogy teljesülni fog).

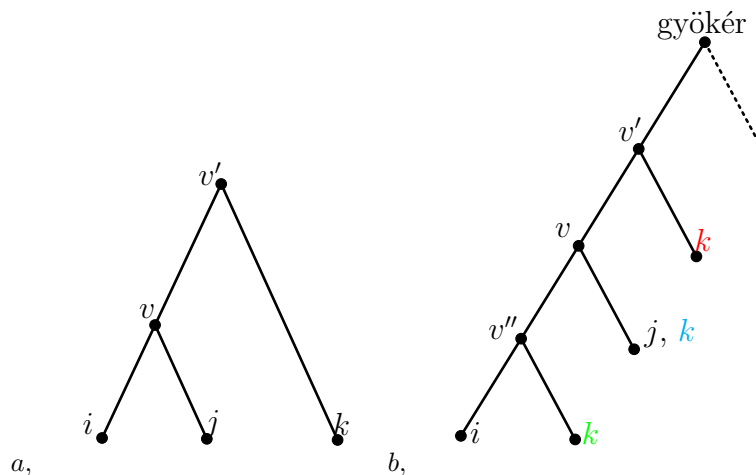
Először is D i . sora alapján elvégzem a $1, 2, \dots, i-1, i+1, \dots, n$ indexek szétválogatását az előbbi módon és "megrajzolom" a gyökérből az i -be vezető utat, a belső csúcsokra pedig ráírom a $D(i, j)$ értékeket. Ebből az állapotból akarok tovább lépni. Legyen egy belső csúcs az úton v, j egy onnan leágazó utód-index, k pedig egy tetszőleges másik index (3. ábra b.). Három esetet vizsgálunk meg. 1. eset: $D(i, j) = D(i, k)$ vagyis j és k is ugyanabba az indexhalmazba van besorolva. Itt úgy folytatjuk az algoritmust, hogy a $D(j, \cdot)$ értékeket is sorba rendezem. Azok, akik nagyobbak $D(i, j)$ -nél azok föntebb lesznek (következő eset), tehát csak a többi fontos. Ez alapján v -ből j -be rajzolunk egy belső csúcsokkal és $D(j, \cdot)$ címkézéssel ellátott utat. Ha $D(j, k) = D(i, j) = D(i, k)$ lenne, akkor a föntiek miatt a k index a v belső csúcshoz lenne a v, j úton rendelkezve, így v lenne a legközelebbi közös őse j és k leveleknek a konstruált fában. Ekkor a v -ből i, j és k felé is szétválík az út, tehát ez nem bináris fák esetén lehetséges csak (de azokra a fákra is igaz az állítás). Mivel D ultrametrikus, ezért $D(j, k) \leq D(i, j)$, ami ha nem egyenlőséggel teljesül, akkor azt jelenti, hogy a k egy v alatti belső csúcshoz lesz rendelve, az lesz a legközelebbi közös őse j -nek és k -nak. Vagyis a v gyökerű részfa egy nem-gyöker belső csúcsa, így az útmenti szigorú csökkenés teljesülni fog rájuk (1.1.7 Def. 3. pontja). Ugyanígy belátható, hogy a v -nél lévő D érték nagyobb lesz az alatta lévő belső csúcsok értékeinél.

2. eset: $D(i, k) > D(i, j)$, ekkor $D(j, k) = D(i, k)$ kell hogy legyen a 3-pont feltétel miatt. Ekkor az index szétválasztás miatt k egy v fölötti v' belső csúcs leszármazottja lesz. Ez azt is jelenti, hogy j és k útjaik is elválnak v' -ben, hiszen k nem v leszármazottja. Így valóban v' belső pontba kell $D(j, k)$ értéknek kerülnie, mert az a legközelebbi közös ősiük, és teljesülni fog a $D(j, k) > D(i, j)$ szigorú csökkenés is.

3. eset: $D(i, k) < D(i, j)$, így az eddigiekhez hasonlóan $D(j, k) = D(i, j)$. Ekkor az indexbesorolás miatt k egy v alatti v'' -ből fog leágazni. v -hez lesz rendelve $D(j, k)$ érték, ami valóban helyes, mert egyrészt v lesz a legközelebbi közös őse j -nek és k -nak, másrészt $D(i, k) < D(j, k)$ is teljesül, tehát a gyökértől lefelé a szigorú csökkenés sem romlik el.

Azt kaptuk tehát, hogy ha egy adott indexre megcsináljuk a gyökérből bele vezető utat és ráírjuk szigorúan csökkenve a D értékeket, akkor szét tudom bontani utódcsoportokra a maradék indexeket. Ha egy v belső csúcshoz tartozó indexeket veszem (akiknek a $D(i, \cdot)$ értéke egyenlő), akkor az 1. esetben leírtak miatt az ő v gyökerű részfájuk a többi részfa leveleit és a belső csúcsok címkézését nem rontják el. Vagyis elegendő a részfákat külön-külön jól megkonstruálni, akkor a nagy fa is jó lesz. Tehát vehetek a v -hez tartozó indexekből egy j indexet, kiválasztom D j . sorából azokat a $D(j, \cdot)$ értékeket, amik $D(i, j)$ -nél nem nagyobbak. $D(j, k) > D(i, j)$ esetén $D(j, k) = D(i, k)$, ami a 2. eset volt, vagyis a hozzá tartozó másik (k) levél másik részfában lesz. A nem nagyobbak az 1. esetet

adják, amiről láttuk, hogy tényleg jó részfat ad. Ezeket a nem nagyobb $D(j, \cdot)$ értékeket szigorú csökkenő sorrendbe állítom, majd elvégzem a részindexhalmaz felosztását az előbbiek szerint. Ezzel egy újabb utódindex kerül jól be a konstruálandó T fába. Rekurzióval befejezem a faépítést és a 1.1.7 feltételei mind teljesülnek a kapott T fára, hiszen n levele lesz, a belső csúcsoknak legalább két gyereke lesz, és rajtuk a címkézések is megfelelőek lesznek. $\square$



3. ábra.

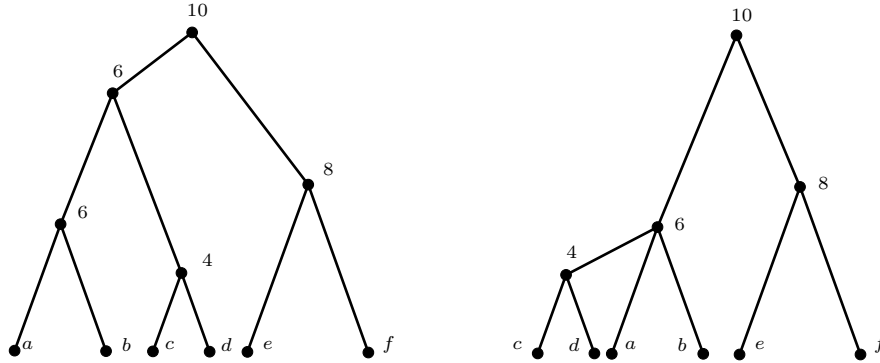
1.1.9. Állítás. ([4, Theorem 17.1.2.]) Ha D ultrametrikus mátrix, akkor az ultrametrikus fája egyértelmű.

Bizonyítás. A 1.1.8 Bizonyítás lépéseiből látszódik, hogy a "szigorúan csökken" feltétel miatt a fat egyértelműen meg lehet konstruálni a mátrixhoz. $\square$

Ezzel beláttuk, hogy az ultrametrikus mátrix és fa fogalma ekvivalens, mert a 1.1.7. Definíció szerinti fa csak az lehet az egyértelműség miatt, amit az előbbi algoritmussal konstruáltunk a D mátrixhoz.

Itt érdemes megjegyezni, hogy ha a "szigorúan" feltétel nem lenne és bizonyos belső csúcsoknak több mint két gyereke lenne, akkor már nem egyértelmű az ultrametrikus fa (ekkor természetesen az egyenlő belső csúcs címkézés miatt az élsúlyozást is kissé át kellene fogalmazni). Ekkor meg lehet az azonos D értékű belső csúcsok részfat cserélni, vagy össze lehet vonni, mint például a 4. ábrán. Ott mivel a két belső csúcsnak 6 az értéke, így például nem egyértelmű, hogy a két csúcs egybe van-e, vagy külön. Valamint, ha lenne egy harmadik 6 súlyú csúcs a 6 és a 10 súlyú gyökér között, azzal felcserélhetnénk a 6 súlyú algyökeres részfatokat.

Az előbb elmondott algoritmus jó és polinomiális. Ez azért igaz, mert egy lépésben egy indexre rendezzük a mátrixbeli elemeket (vagy egy részüket) és ez



4. ábra. Ultrametrikus fa nem egyértelmű, ha megengedünk egyenlőséget a szomszédos belső csúcsoknál.

alapján a részfában lévő indexeket. A rendezés az polinomiális idő alatt megy. Ha esetleg minden érték egyforma egy részfa indexeinél, akkor az azért van, mert minden részfabeli indexnek a távolsága ugyanannyi egymástól, tehát egy kis csillagfát alkotnak. De azt is egy polinomiális végigolvasással ellenőrizhetjük. Így egy lépés valóban polinomális idejű. Minden egyes lépésnél legalább egy új belső csúcsot hozzáveszünk a nagy fához, vagy egy előbb létrejött belső csúcs alatti levelekről megállapítjuk, hogy csillagfát alkotnak, de így is egy belső csúcs legfeljebb két lépésben szerepelhet. Ezek száma azonban n levelű fa esetén legfeljebb $n - 1$ (1.2.5. Állítás). Így tehát legfeljebb $2 \cdot (n - 1)$ lépés lesz és mind polinomiális, tehát az algoritmus is polinomális. Létezik ennél gyorsabb algoritmus is, de ennek bizonyításába nem megyünk bele, mert nem is szükséges a továbbiakban.

1.1.10. Állítás. ([4, Theorem 17.1.3.]) Ha D ultrametrikus mátrix, akkor $O(n^2)$ időben megkonstruálható a ultrametrikus fája.

Most lássuk be, hogy a 3-pont feltétel erősebb a 4-pont feltétel-nél. A következő fejezetben részletezzük a helyességét, de most fogadjuk el a szakirodalomban használt élsúlyozást. Ha adva van egy D ultrametrikus mátrix és hozzá elkészítjük a T ultrametrikus fát, akkor az élekre a két végpont különbségének a felét írjuk. Így ha a fa távolságmátrixát veszem, akkor az pont a D mátrixot adja vissza. Előfordul, hogy az élekre maga a különbség kerül, ekkor a távolságmátrix a D mátrix értékeinek feléből áll. Erre utaltunk a 2. ábránál és azon ellenőrizhetjük is. Így definiáltunk egy nemnegatív élsúlyozást is a fán.

1.1.11. Állítás. Ha van egy ultrametrikus T fánk, akkor az filogenetikai fa is, más szóval ha teljesül egy D mátrix elemeire a 3-pont feltétel, akkor a 4-pont feltétel is. Fordítva azonban nem igaz, tehát $U_n \subset T_n$, de nem egyenlő a két halmaz.

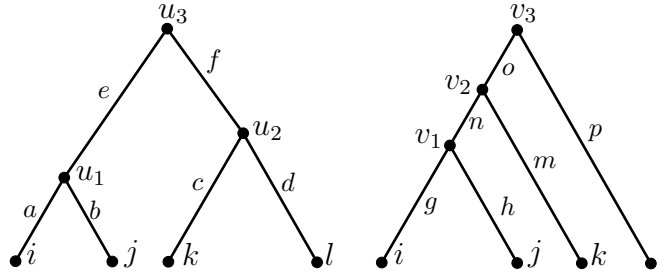
Bizonyítás. (saját bizonyítás) Tegyük fel, hogy van egy ultrametrikus T fánk és D távolságmátrixában ω jelöli az értékeket (az előbbiek miatt ez a D az ultrametrikus mátrixa a fának). Vegyünk négy indexet/levelet: i, j, k, l . Ha ezekre megszorítom a fát, akkor kétféle alakú ábrát kaphatok (eltekintünk az indexek sorrendjétől, csak az alak számít és itt a belső éleken szereplő számok lehetnek nullák is, ha nem bináris a fa). Ezek az 5. ábrán láthatóak. Ekkor az első ábrán:

$$\omega(i, j) + \omega(k, l) = a + b + c + d \leq \omega(i, k) + \omega(j, l) = (a + e + f + c) + (b + e + f + d) = (a + e + f + d) + (b + e + f + c) = \omega(i, l) + \omega(j, k),$$

mert $e, f \geq 0$. Hasonlóan a második ábránál:

$$\begin{aligned} \omega(i, j) + \omega(k, l) &= g + h + m + o + p \leq \omega(i, k) + \omega(j, l) = \\ &= (g + n + m) + (h + n + o + p) = (g + n + o + p) + (h + n + m) = \omega(i, l) + \omega(j, k). \end{aligned}$$

Így tehát teljesül a *4-pont feltétel*, a T filogenetikai fa.



5. ábra. 4 levelű ultrametrikus fa lehetséges alakjai

Egyenlőség azért nem lehet, mert például a 1. ábrán a fa nem ultrametrikus, mert a mátrix sem az. Vegyük például a, b, c indexeket. Ekkor $\omega(a, b) = 3 < \omega(a, c) = 6, 5 < \omega(b, c) = 7, 5$, tehát nem teljesül a *3-pont feltétel*, a *4-pont feltétel* viszont igen, mert $\omega(a, b) + \omega(c, d) < \omega(a, c) + \omega(b, d) = \omega(a, d) + \omega(b, c)$. Ha zavaró a gondolat, hogy nincs ott gyökér kijelölve, míg az ultrametrikus fáknál kikötöttük, akkor vegyük úgy, hogy u a gyökér a filogenetikai fában. Igazából nem fontos, mert a 1.1.5 Állítás és a 1.1.8 Tétel miatt elég a mátrixot figyelni és annak értékei függetlenek attól, hogy van-e gyökér kijelölve (természetesen addig, amíg nem az egyik levelet jelöljük ki gyökérnek). $\square$

1.2. Ultrametrikus fa

A továbbiakban csak az ultrametrikus fákkal fogunk foglalkozni. Az előzőekben láttuk, hogyan lehet fát konstruálni egy mátrixból. Azonban több kérdést is megválaszolatlanul hagytunk. Először is, hogy miért igaz az, hogy az élsúlyozás és a belső csúcs súlyozás felcserélhető művelet, aztán azt, hogy mit is ábrázolunk egy ilyen fánál, miért pont csökkennek a belső csúcsokra írt értékek. Valamint pár hasznos megfigyelést is meg kell állapítanunk.

Először is rögzítsünk egy szemléletet. A gyökeret a legközelebbi közös őznek tekintjük, a *belső* csúcsok azt a pillanatot jelentik, amikor két utód-csoport szétvált, a gyökeret is ilyennek tekintjük. Bár a 1.2.3. és 1.2.4. Definíciókban részletesebben is lesz szó róla, megfogalmazzuk újra a 1.1.11. Állítás előtti élsúlyozást ultrametrikus fákra. Ha adva vannak a belső csúcsokra írandó mátrix értékek a fában (a leveleken 0 szerepel), akkor az élekre a két végpont különbségének a felét írjuk (felső–alsó), ami a szigorú csökkenés miatt pozitív érték lesz minden élen. Ha pedig az élsúlyozás lenne adott, akkor a szakirodalom szerint veszek egy belső csúcsot és két különböző utódját (legalább két utódja van) és kiszámolom a köztük menő út hosszát, majd az az érték lesz a belső csúcs címkéje. Ennél mindegy melyik alatta lévő levél utódpárt választom, az érték ugyanannyi lesz, ezt igazolom a 1.2.1 Tételben. Ezzel a definiálással látszik, hogy valóban mindegy, hogy a belső csúcsok vagy az élsúlyozás van-e adva, és az is, hogy ilyenkor a D ultrametrikus mátrix megegyezik a fa távolságmátrixával (következő tételben ezt is bebizonyítjuk).

Most belátjuk, hogy ha ultrametrikus D mátrixunk van, akkor annak belső csúcsaitól valóban egyenlő távolságra vannak lefelé a levelek.

1.2.1. Tétel. *D nemnegatív értékekből álló, szimmetrikus, $n \times n$ mátrix, ha ultrametrikus, akkor a hozzá tartozó T fára igaz, hogy a gyökértől és minden más belső csúcstól is az alatta lévő levelek azonos távolságra vannak, ha a fenti élsúlyozást alkalmazzuk (az ilyen fa a szakirodalomban az **equidistant tree**). Visszafelé, ha egy T fa élsúlyozására igaz az előbbi tulajdonság, akkor legyen D mátrix i . sorában, j . oszlopában az i és j levelek (egyértelmű) T fabeli útjának hossza (tehát a távolságmátrix), és ekkor a D mátrix ultrametrikus (és T tartozik hozzá, mint ultrametrikus fa).*

Bizonyítás. (saját bizonyítás)

Tegyük fel, hogy D ultrametrikus, T a hozzá tartozó ultrametrikus fa, aminek éleire a fentiekben leírtak szerint lehet élsúlyozást készíteni. Tehát az élekre a két végpont különbségének felét írunk (pozitív számot). Ekkor ha veszek egy tetszőleges v belső csúcsot, akkor tőle lefelé az éleken ha összeadom az értékeket a levelekig, akkor a v -hez tartozó súly fele fog kijönni. Ez a teleszkópikus összeg könnyen látszik: jelöljük $D(u)$ -val az u belső csúcshoz tartozó D -beli értéket, ha v -ből az i levélig a $v_1, v_2, \dots, v_k, i$ csúcsokon lehet eljutni, akkor az általuk meghatározott úthossz

$$d(v, v_1) + d(v_1, v_2) + \dots + d(v_{k-1}, v_k) + d(v_k, i) = \frac{D(v) - D(v_1)}{2} + \frac{D(v_1) - D(v_2)}{2} + \dots + \frac{D(v_{k-1}) - D(v_k)}{2} + \frac{D(v_k) - D(i, i)}{2} = \frac{D(v) - D(i, i)}{2} = \frac{D(v) - 0}{2} = \frac{D(v)}{2}$$

amivel az előbbi állítást beláttuk. Mivel azonban ez tetszőleges levélutódra igaz, így valóban mind ugyanakkora távolságra van az adott v belső csúcstól, a tétel egyik irányát beláttuk. Ekkor nyilván az is teljesül, hogy ha i, j és k, l levélpárokra is a legközelebbi közös ősök v , akkor a $d(i, j)$ és $d(k, l)$ távolságok egyenlőek, és pont $D(v)$ -vel egyenlőek, tehát D a fa távolságmátrixa. Ebből következik a fent megadott *élsúlyozás* $\Rightarrow$ *belső csúcs súlyozás* lépés helyessége is.

Visszafelé, ha tudjuk, hogy T -ben a gyökértől és minden belső csúcstól egyenlő

távolságra vannak a levelek, akkor ugyancsak 3 tetszőleges levelet véve, az 3. ábra a , részét kapjuk (ennek a betűzését használva), i, k és j, k távolsága is megegyezik, hiszen a v, k útszakasz közös, utána pedig a levelek azonos távolságra vannak v -től. Tehát $D(i, k) = D(j, k)$, a pozitív élsúlyozás miatt pedig $D(i, j)$ kisebb távolságot ad (esetleg egyenlőt, ha v és v' egybe esik), tehát teljesül tetszőleges ponthármasra a *3-pont feltétel*, ezért D mátrix ultrametrikus. Itt fontos megjegyezni, hogy ha i, j és i, k levélpárok legközelebbi közös őse (v) megegyezik, akkor a feltétel miatt v -től megegyezik a levelek távolsága (d), ezért $D(i, j) = 2 \cdot d = D(i, k)$, így ha az élekre írt súlyok adottak, akkor a legközelebbi közös ősről tényleg ráírhatjuk a tételben definiált D távolságmátrix értékét, nem ütközhetünk olyan ellentmondásba, hogy két eltérő értéknek kellene oda kerülnie. Ekkor pedig teljesülnek T -re a 1.1.7 Definíció feltételei, tehát T a D ultrametrikus mátrixhoz tartozó ultrametrikus fa. $\square$

A következő algoritmus egy újabb bizonyítás arra, hogy ultrametrikus fára igaz, hogy a belső csúcsoktól a levelek egyenlő távolságra vannak. Ezt azért adjuk meg, mert gyakorlatban sokszor egyszerűbb ilyen módon megkapni az ultrametrikus fát, mint a 1.1.8 Tétel bizonyítása szerint. A bizonyítás úgy következik az algoritmusból, hogy egy olyan fát fogunk kapni, amire egyrészt teljesül az *equidistant tree* feltétele, másrészt igazak a 1.1.6 Definíció feltételei is, de mivel egyértelmű az ultrametrikus fa, így csak a megkonstruált fa lehet D ultrametrikus fája és arra igaz lesz a belső csúcsokra vonatkozó feltétel. Valamint a *élsúlyozás* $\Rightarrow$ *belső csúcs súlyozás* lépés helyessége is következik belőle ugyancsak az egyértelműség miatt, mert egy ultrametrikus távolságmátrixhoz csak egy ultrametrikus fa tartozhat. Az algoritmushoz jó nézni a 6. ábrát is, ami egy példát ad rá. A végleges fa a jobb oldali fa.

Új algoritmus az ultrametrikus fa konstruálására (saját bizonyítással)

Tegyük fel, hogy D ultrametrikus és T a hozzá tartozó fa. Konstruálunk egy T' fát a következő módon. Vesszük azokat az i indexeket, amikre D i . sorában előfordul a legnagyobb D -beli szám (d_1). Ezt a távolságot megfelezzük és a gyökértől akkora távolságra felvesszük az előbbi indexekhez tartozó leveleket (6. ábra első fája). A legnagyobb levéltávolság T -ben azokhoz a levélpárokhoz tartozik, amik között a fabeli út átmegy a gyökéren. Minden levélhez tartozik T -ben olyan, aki a gyökérnek másik gyerekéhez tartozik (a gyökérnek legalább két gyereke van és a két gyerek részfájában van legalább 1 – 1 levél), ezért ez a maximális érték D minden sorában szerepelni fog, és így mindet felvettük most a gyökérből. T' -ben a gyökérre ráírjuk ezt a legnagyobb D -beli számot. Utána vesszük a következő legnagyobbat (d_2). Sorban megnézzük, hogy az 1-es levélnek a sorában szerepel-e d_2 , ha nem akkor tovább lépünk a 2-re, de ha igen, akkor az 1-t és azokat a leveleket, akiknek legfeljebb d_2 a távolsága 1-től (vagyis azokat az oszlopindexeket, amik alatt az első sorban legfeljebb d_2 áll), azokat berakjuk egy új belső csúcs alá. Utána vesszük a következő olyan sort a D mátrixban, amit még nem vizsgáltunk, megint kiválasztunk d_2 szerint egy levélhalmazt az 1-es sorhoz hasonlóan, stb. Így a gyökér alá új belső csúcsokat veszünk fel (legalább egyet), amikhez ugyanaz a d_2 érték tartozik.

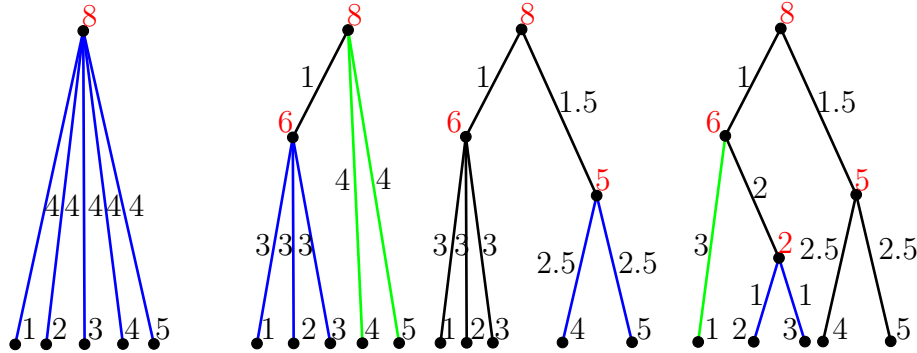
Természetesen kimaradhat levél, aminek nem volt d_2 a sorában (csak ha van d_2 a sorban, akkor nézzük tovább, hogy van-e kisebb érték is az adott sorban). Ezeket úgy vesszük, hogy fennragadtak a gyökérben (6. ábrán a 2. fában zölddel jelöljük ezeket a fennragadt leveleket). Ennél fontos megjegyezni, hogy ha most pl. az i levél miatt a j ki lett választva egy belső csúcs alá, a k levél pedig egy másik alá (akár fennragadt levél), akkor a j és k leveleknek a $D(j, k)$ értéke nem lehet kisebb és egyenlő sem d_2 -vel. Ez a *3-pont feltétel* miatt igaz, hiszen k értékére az i . sorban $D(i, k) > d_2 \geq D(i, j)$ teljesül, ezért $D(j, k) = D(i, k) > d_2$. Így tehát az új belső csúcsok alá olyanok kerültek, amiknek a D -beli értékük (T -ben a távolságuk) nagyobb az aktuális legnagyobb számnál. Ha egy adott belső csúcsot nézünk, alatta olyanok vannak, akik az egyik levélről legfeljebb d_2 távolságra voltak. De ekkor egymástól is legfeljebb ekkora távolságra vannak, hiszen ha lenne egy j, k pár amit i szerint választottunk ki és amire $D(j, k) > d_2$, akkor $d_2 \geq D(i, j)$ és $d_2 \geq D(i, k)$ miatt ellentmondanánk a *3-pont feltételnek*. Így tehát van egy gyökerünk és néhány belső csúcsunk, amikre d_1 , illetve d_2 van írva. Ezek mindegyikéből $\frac{d_2}{2}$ hosszú éleket veszünk fel, amik a belső csúcsokhoz kiválasztott levelekbe mennek. A gyökér és az új belső csúcs távolsága a két "fél D érték" különbsége legyen, tehát $\frac{d_1 - d_2}{2}$. Ezt az eljárást folytatjuk, hogy vesszük a következő legnagyobb D értéket, kiválasztjuk hozzá a leveleket a mátrix sorai alapján és csinálunk legalább egy újabb belső csúcsot egy addigi alá, amiből ezek a kiválasztott levelek ágaznak ki és megcsináljuk az előbbi él és belső csúcs súlyozást. Ilyenkor mindig egy már meglévő belső csúcsnak a leveleit "bontjuk" tovább, méghozzá azokat, akik egy kis csillagfát alkotnak (pl. a 6. ábrán a 2. fánál a 3 címkéjű részfa ilyen alakú és a 4. fánál azt bontjuk). Addig folytatjuk, amíg el nem érjük a nullát a mátrixban (a legkisebb számot), akkor leállunk.

Fontos, az algoritmus során nem akadhatunk el. Azt láttuk a d_2 értéknél, hogy egy adott belső csúcs alá csak olyanok kerülnek, amiknek a D -beli értéke legfeljebb d_2 , a többi belső csúcs alá bekerült levelektől pedig nagyobbnak kell lennie a *3-pont feltétel* miatt a távolságuknak. Ugyanígy lehetne tetszőlegesre is, mert nem használtuk ki a bizonyításban, hogy ez a 2. legnagyobb érték D -ben.

Egyetlen rész van csak, amit nem néztünk, hogy amikor veszünk egy levélhalmazt, akkor annál a későbbiekben nem fordulhat-e elő, hogy egy másik részfából is kellene leveleket kiválasztanunk (tehát ha egyszer szétvált két fa, akkor sose lehet összeérés). Ez bizonyítaná azt, hogy mindig csak egy csillagfa leveleiből választjuk ki az új leveleket. Különben pont az lenne, hogy két részfából veszünk leveleket, amiről most fogjuk belátni, hogy nem lehet.

Tegyük fel, hogy valameddig megkonstruáltuk a T' fát és nem akadtunk el (a 2. legnagyobb D érték kiválasztásáig láttuk, hogy jó az algoritmus). Ekkor kiválasztjuk a következő legnagyobb számot és hozzá a leveleket is a fentiek szerint. Ezek egy eddig meglévő csillag-részfának lesznek a levelei (mint például a 6. ábrán a 4. fán a 2-értékű belső csúcs a 3-értékű belső csúcs alatti csillag-részfából vált ki). Indirekten tegyük föl, hogy van egy i és k indexpár, amiket most kiválasztottunk, de más-más (eddig megkonstruált) részfának a levelei, vagyis nem egy csillag-részfa levelei (segítségnek nézhetjük a 3. ábra a , részét). A kiválasztott levelekből mindig van kettő, hiszen egy adott sorban ha szerepel

egy D érték, akkor ahhoz tartozik egy tőle különböző levél is (akitől T -ben akkora volt a távolsága, és önmagától 0 a távolsága, aminél az aktuális D érték nagyobb, különben leállunk). Ez azt jelenti, hogy volt egy v' csúcs (esetleg a gyökér), aminek két eltérő gyerekébe kerültek az i és k levelek. Ez nem akkor történt, amikor v' -t felvettük, hiszen akkor csillagfát készítettünk hozzá (kék élek az ábrákon), vagyis minden gyereke egy levél volt, így azonos részfában voltak a levelei. Ezért kell lennie egy v belső csúcsnak, ami egy kis részfát hoz létre a v' részfájában és emiatt lesz i és k eltérő részfában. Ekkor a v csúcshoz is tartozik már D érték, ami nagyobb az aktuálisnál (hiszen korábban felvettük már a v belső csúcsot). Mivel belső csúcs létrehozásnál mindig legalább két levél lesz a fában (hiszen a D érték legalább két levélhez tartozik), így v -nek van egy j levele is és a konstrukció miatt a v -hez tartozó D érték ($d(v)$) nagyobb vagy egyenlő, mint a $D(i, j)$ (főnt indokoltuk). Viszont azt is láttuk, hogy ekkor a v alá nem bekerült leveleknek a $D(i, \cdot)$ és $D(j, \cdot)$ értéke nagyobb, mint $d(v)$, így azonban a $D(i, k)$ érték is nagyobb a már vizsgált $d(v)$ értéknél, tehát az aktuális D -értéktől is nagyobb, vagyis nem választhattuk ki most ugyanaz alá a belső csúcs alá az i és k leveleket, mert akkor a $D(i, k)$ legfeljebb az aktuális D érték. Ellentmondásra jutottunk, vagyis nem folyhat össze két, már szétválasztott részfa. Ezért nem is akadhatunk el az algoritmus során.



6. ábra. Faalak konstruálás $n = 5$ levél esetén.

Leállunk, amikor már csak a nulla értéket nem vizsgáltuk a D mátrixban. Ekkor egy ultrametrikus fát kapunk, mert teljesülnek rá a 1.1.7 Definíció tulajdonságai: n levele van; minden belső csúcsnak legalább két gyereke van (mert a D érték mindig legalább két levélhez tartozik) és a gyökértől lefelé haladva csökkenek az értékek az úton, hiszen így konstruáltuk meg. Valamint az i, j levelek legközelebbi közös őse van megcímkezve a $D(i, j)$ értékkel. Ez azért igaz, mert láttuk, hogy az i, j leveleket addig viszem lefele egy fában, amíg az aktuális D -értéknél nem nagyobb távolságuk. Így ha az egyik fennakad egy belső csúcson, vagyis inntől levélként, vagy részfaaként különválnak az i -től, akkor ez a fennakasztó belső csúcs lesz az i és j legközelebbi közös őse a T' fában. Valamint pont erre fog felkerülni a $D(i, j)$ érték, hiszen ennél még éppen ki vol-

tak választva a levelek, vagyis az aktuális D értéknél nem volt nagyobb $D(i, j)$. A belső csúcs alatt valahol szerepelnie kell előbb-utóbb minden D értéknek, ami ezekhez a levelekhez tartozik (így konstruáltuk), ezért mivel a későbbi D értékeknél már nagyobb $D(i, j)$, így most kell ennél a fennakasztó csúcsonál szerepelnie, mint aktuális D érték.

Ezzel beláttuk, hogy a konstruált T' fa a D mátrixhoz tartozó ultrametrikus fa, de akkor a 1.1.9. Állítás miatt csak ez a fa tartozhat D -hez (így T és T' azonos), ez a T' fa viszont teljesíti azt a tulajdonságot, hogy a belső csúcsoktól a levél-utódjaik egyenlő távolságra vannak, hiszen így konstruáltuk a fát.

Az érdekessége ennek az algoritmusnak, hogy ez egy szélességi faépítés, mert szinteket építünk nagyság szerint és úgy haladunk lefele a gyökérből. A 1.1.8. Tétel bizonyításában megadott algoritmus pedig mélységi faépítés, mert haladunk egy i levél irányába, majd visszalépünk egy belső csúcsba és egy másik levél irányába építjük tovább a fát.

Most térjünk vissza a 1.1.7 és 1.1.6 Definíciókhoz.

1.2.2. Definíció. ([4, 449., 451. oldal]) Legyen D egy szimmetrikus $n \times n$ nemnegatív valós mátrix. T gyökeres fát D minimális ultrametrikus (röviden: min-ultrametrikus) fájának nevezzük, ha teljesülnek rá a 1.1.7. Definícióban szereplő feltételek, kivéve a 3.-at, mert min-ultrametrikusban szigorúan nőnek a D értékek a gyökértől a levelek felé.

Hasonlóan D szimmetrikus $n \times n$ nemnegatív valós mátrix értékei min-ultrametrikus távolságot definiálnak, ha bármely i, j, k indexre a minimum legalább kétszer fordul elő a $D(i, j)$, $D(i, k)$, $D(j, k)$ értékeknél. Ekkor D -t min-ultrametrikus mátrixnak hívjuk.

A min-ultrametrikus mátrixra is igaz a 1.1.8. Tétel és 1.1.9. Állítás, mert hasonlóan bizonyíthatóak.

Ha egy élre azt írjuk, hogy a két utódcsoporthoz tartozó szétválás között mennyi idő telt el, akkor a fa és az élsúlyok pontosan leírják az evolúciós folyamatot. Feltehetjük, hogy a súlyok nemnegatívak (igazából pozitívak), hiszen a nulla súlyt vehetjük úgy, hogy a két belső csúcs egyben van, mert nem telt el idő köztük, negatív időnek meg nincs értelme. Ez alapján két címkézést csinálhatunk, ami válaszol arra a kérdésre, hogy mit szemléltetnek az ultrametrikus és min-ultrametrikus fák. Itt az is látszódnia fog, hogy miért az eddigiek szerint adtuk meg az élsúlyozást. Mivel ezek fontosak a fák megértésében, definícióként adjuk meg őket.

1.2.3. Definíció. Az egyik címkézés, hogy adva vannak az élsúlyozások és fentről lefelé haladva ráírjuk a belső csúcsokra, hogy mennyi oda az út hossza a gyökérből. Ez azt adja meg, hogy melyik időpillanatban történt egy utódcsoporthoz tartozó bomlás és egy min-ultrametrikus fát kapunk, hiszen a következő szétválásig pozitív idő telik el, így lefele nőnek az értékek. A $D(i, j)$ értékek, amik a belső csúcsokra vannak írva, a szétválás időpontját adják. Ugyanígy fordítva, ha a min-ultrametrikus mátrix értékei adottak, akkor ezzel a logikával fentről lefele lehet súlyozni az éleket, hogy lefelé az úthosszak azonosak legyenek a $D(i, j)$ értékekkel. A külső élekre az utolsó szétválástól az aktuális mérésig eltelt időt írhatjuk.

1.2.4. Definíció. *Ha nem fentről, hanem lentől felfelé kezdem el összeadni az élhosszakokat és azt írom a belső csúcsokra, akkor azt kapom, hogy az adott utódcsoporthoz tartozó szétválástól mennyi idő telt el mostanáig. Ez ultrametrikus fát ad, hiszen pozitívak az élsúlyok. A $D(i, j)$ érték ekkor az adott belső csúcsnak és az utód-leveleinek a távolsága a fában. (1.2.1 Tétel szerint a szétválási ponttól egyenlő távolságra van a két levél). Azonban a fában az i és j csúcsok távolsága ekkor $2 \cdot D(i, j)$ lesz, tehát a D -ben a távolságmátrix értékeinek a fele szerepel. Erre már régebben is utaltunk. Ez nem egységes a szakirodalomban, hogy a D mátrix értékei a távolságmátrixbeliekkel megegyeznek vagy fele akkorák. Mint azt láthatjuk, ha azt szeretnénk szemléltetni, hogy mennyi ideje történt egy szétválás, akkor előnyösebb a "fél" távolságmátrixot használni D -nek, ugyanúgy ultrametrikus lesz, hiszen csak osztunk kettővel. Ekkor természetesen az élek súlya a végpontjaik különbségével egyenlő. Mi a továbbiakban azonban az elterjedtebb változatot használjuk, amikor D azonos a távolságmátrixszal és ilyenkor az élekre írt súlyok a végpontok különbségének fele (akár az élsúlyozást, akár a belső csúcs címkézést csináljuk előbb).*

A valóságban többnyire az adatok az utóbbinak felelnek meg (mennyi idő telt el a szétválás óta és nem, hogy mikor történt), ezért foglalkozunk inkább az ultrametrikus fakkal.

A 2. ábránál azt is ellenőrizhetjük, hogy ha adva van a fa élsúlyozással, akkor abból hogyan kapjuk vissza a mátrixot. Például az a, e levélpárnak $3 + 3$ a távolsága, és most a "felezős" mátrixkészítés miatt a táblázat hozzájuk tartozó részeinél 3 szerepel (D mátrix a táblázat számmal teli része). b -nek c -t leszámítva mindenkitől $8 + 8$ a távolsága, így csak c -hez van 2 írva.

Nem feltétlenül létezik minden mátrixhoz (adathalmazhoz) ultrametrikus, vagy min-ultrametrikus fa. Egyik féle fa sem létezik, ha egy D mátrixban több mint $n - 1$ különböző pozitív érték szerepel. Ez azért igaz, mert egy n levelű fának legfeljebb $n - 1$ belső csúcsa lehet és minden $D(i, j)$ érték valahova kerülni fog, hiszen minden levélpárnak van (legkésőbb a gyökérben) legközelebbi közös őse. A 1. ábránál látott példában is a mátrixban 6 különböző pozitív szám szerepel, de csak két belső csúcs van.

1.2.5. Állítás. *Egy n levelű gyökeres fának legfeljebb $n - 1$ darab belső csúcsa van (gyökérrel együtt).*

Bizonyítás. (saját bizonyítás)

n szerinti indukcióval haladunk. $n = 2$ esetén két levél van, így csak a gyökér lehet a fában belső csúcs, hiszen egy belső csúcsnak mindig legalább két gyereke, és így legalább két levél utódja van. Ha $n = 3$, akkor vagy mind a három levél a gyökér gyereke, vagy a gyökérnek két gyereke van és pontosan az egyik belső csúcs, és annak két levél utódja van, a másik gyereke a gyökérnek a 3. levél. Tehát legfeljebb $2 = 3 - 1$ belső csúcs lehet. Tegyük fel, hogy $n = m - 1$ -ig igaz az állítás. Ekkor legyen $n = m$ és tegyük fel, hogy a gyökérnek $k \geq 2$ gyereke van (nem feltétlenül belső csúcsok a gyerekei). Lépünk át a k db gyerekre, így k db részfat vizsgálunk $n_1, n_2, \dots, n_k$ levéllel (mind legalább 1), ahol $n_1 + n_2 + \dots + n_k = m$, tehát mind legfeljebb $m - 1$. Ezekre a fákra

az igaz az indukció miatt, hogy legfeljebb $n_i - 1$ belső csúcsuk van (ebben az eredeti gyökér gyerekei is benne vannak), ha $n_i = 1$, vagyis levél volt az egyik gyerek, akkor is igaz az állítás, hiszen akkor abban a "részfában" nincs belső csúcs. Így tehát a belső csúcsok száma (a gyökeret is házzávéve) legfeljebb $(n_1 - 1) + \dots + (n_k - 1) + 1 = n_1 + \dots + n_k - k + 1 = m - k + 1 \leq m - 2 + 1 = m - 1 = n - 1$. Ezzel az állítást beláttuk. $\square$

1.2.6. Megjegyzés. *Az is igaz, hogy pontosan a bináris fának van $n - 1$ belső csúcsa. Ez egyik irányba triviális, hiszen az előző bizonyítást használva $n = 2, 3$ esetén látszik hogy a bináris fának 1 illetve 2 ($n - 1$) belső csúcsa van. Nagyobb n esetén pedig $k = 2$, tehát az egyenlőtlenség egyenlőséggel teljesül. A másik irányhoz is indukciót használhatunk, tehát a belső csúcsok száma a fában $n - 1$ és kell, hogy bináris. Az előbbi bizonyításban egyenlőségnek kell lennie, tehát $k = 2$. Lejebb haladva a kis fákban ott is csak akkor teljesülhet az $n_i - 1$ egyenlőtlenség egyenlőséggel, ha ott is a gyökér két gyerekű. Ilyen módon lefele haladva minden belső csúcs előbb-utóbb gyökér lesz egy részfában, tehát két gyerekes lesz. Végül a fa minden nem levél csúcsának két gyereke lesz, a fa bináris.*

1.2.7. Megjegyzés. *A 1.2.5 Állításból az is következik, hogy egy n levelű fának legfeljebb $n - 2$ belső éle lehet. Hiszen ha levágjuk a külső éleket (vagyis a leveleket), akkor legfeljebb $n - 1$ csúcsból álló fát kapunk, aminek csak belső élei vannak. Így tehát legfeljebb $n - 1 - 1 = n - 2$ belső éle lehet az eredeti fának, egyenlőség csak bináris fa esetén lehetséges.*

2. Filogenetikai fák tere

Ahogy a *Bevezetésben* is jeleztük, nem csak a fák alakja fontos az adatok alapján, hanem az, hogy ha elmérnénk valamit, akkor mennyire kapnánk más formájú fát. Más szóval súlyozott fák távolságát szeretnénk definiálni és vizsgálni. Ezt úgy fogjuk megtenni, hogy a fákhoz valamilyen módon egy vektort rendelünk és ezek között fogunk valamilyen metrika szerint távolságokat vizsgálni. Ebben a fejezetben nemnegatív élsúlyokkal ellátott, gyökeres fákat vizsgálunk és a D mátrix értékei az i, j levelek távolsága a fában (ha létezik az adatokhoz fa). A szakirodalomban sokszor bináris fákat tekintenek, de mint azt látni is fogjuk, előfordulhatnak több mint kétgyerekes fák is az ultrametrikus térben, de az előző szakaszban általánosságban beszéltünk az ultrametrikus fákról, így ez nem fog akadályt jelenteni. Egyelőre általános filogenetikai fákról lesz szó, nem ultrametrikusokról. Euklídeszi geometriát és metrikát (pl. skaláris szorzást használt, *Robinson-Foulds metrika* [13]) nem szoktak használni a gyakorlatban, mert nem tükrözi a pontok távolsága a fa-alakok kapcsolatát. Két nevezetes metrika ismert, az egyik az úgynevezett *BHV*-metrika (2.6. fejezet), a másik, a matematikában jobban ismert, trópusi geometria és metrika. Ez utóbbival foglalkozunk bővebben.

2.1. A $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ tér és a trópusi metrika

Először is, hogyan lesz a fából pont a térben: Ha veszünk egy n levelű címkézett filogenetikai fát, akkor elkészítjük hozzá a távolság mátrixát (az i . sor j . oszlopában az i, j levelek közötti út pozitív hossza lesz, a főátlóban nulla). A mátrix nyilván szimmetrikus. Ebből egy $\omega \in \mathbb{R}^N$ vektort gyártunk, ahol $N := \binom{n}{2}$. Egyszerűen legyen ω első $n-1$ koordinátája D első sora, kivéve az első oszlop értéke (1. levél távolsága a többi levélről), a következő $n-2$ koordináta D második sora, kivéve az első és második levélhez tartozó értékek, hiszen az 1, 2 levelek távolsága már szerepel a vektorban, a főátlóban meg nullák vannak. Ilyen módon D főátló fölötti része kerül be ω -ba, és a konstruálás miatt a mátrixot és a fát is triviálisan vissza lehet fejteni a vektorból. Valamint N definiálása is helyes, mert $\binom{n}{2} = (n-1) + (n-2) + \dots + 1$. Bár $\omega \in \mathbb{R}^N$, mégis nekünk elég $\omega \in \mathbb{R}^N/\mathbb{R}\mathbf{1}$ vektorokat nézni, vagyis olyan teret, ahol x és y $\mathbb{R}^N$ -beli vektorok ugyanabba az ekvivalenciaosztályba tartoznak, ha $x - y$ minden koordinátája ugyanaz a valós szám ($\mathbb{R}^N/\mathbb{R}\mathbf{1}$ teret trópusi projektív tórusznak nevezik). Az indoklás a következő: Veszünk egy T n levelű filogenetikai fát, és a hozzá tartozó ω_T vektort, majd a $\lambda \cdot \mathbf{1} = \lambda \cdot (1, 1, \dots, 1)^T$ vektort hozzáadjuk és az ehhez tartozó fát szeretnénk megnézni. T fának a külső éleit mind megnöveljük $\lambda \cdot 0,5$ -tel, ekkor a hozzá tartozó levél távolságok mind nőnek (az út két végén) $2 \cdot \lambda \cdot 0,5 = \lambda$ -val és ezzel megkaptuk az $\omega_T + \lambda \cdot \mathbf{1}$ vektort és egy hozzá tartozó n levelű filogenetikai fát, ami pontosan úgy néz ki, mint T , csak a külső élei hosszabbak (filogenetikai fákra nem látjuk be, de igaz lesz későbbi $T_n \subset \mathbb{R}^N/\mathbb{R}\mathbf{1}$ állítás). Ultrametrikus fák esetén tudjuk, hogy a fa egyértelműen létezik D -hez és ω -hoz, vagyis tényleg a hozzá tartozó fát kapjuk a külső él nyújtással. Így ultrametrikus fáknál valóban elegendő $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -ben maradni, hiszen a külső éleken már

nincs evolúciós változás, így azoknak a hossza igazából nem fontos, csak annyiban, hogy mindegyiknek pozitív legyen a hossza. Ezekre az élekre gondolhatunk úgy, hogy az utolsó szétválástól a jelen pillanatig az eltelt időt szemléltetik, így csak az a fontos, hogy az utolsó jelentős pillanat után "állítsuk le a stoppert". Ezek alapján megfogalmazhatjuk a következő definíciót, amit igazából az 1. fejezetben is írtunk már, de most átfogalmazzuk kissé.

2.1.1. Definíció. ([5]) T_n az n levelű filogenetikai fák terét, U_n az n levelű ultrametrikus filogenetikai fák terét jelenti (amikről az 1. fejezetben volt szó). Így a T_n és az U_n terek pontjai olyan $\omega \in \mathbb{R}^N/\mathbb{R}\mathbf{1}$ vektoroknak felelnek meg, amikhez létezik T filogenetikai, illetve ultrametrikus fa.

Az eddig leírtak és a 1.1.11 Állítás alapján $U_n \subset T_n \subset \mathbb{R}^N/\mathbb{R}\mathbf{1}$ teljesül. Most definiáljuk a trópusi metrikát $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -ben, amelynek a pontjait $[x]$ -el jelöljük, hiszen itt ekvivalenciaosztályok vannak ($x \in \mathbb{R}^N$ -nel ekvivalens $\mathbb{R}^N$ -beli vektorok halmaza). A későbbiekben a rövidség kedvéért visszatérünk a sima x jelöléshez.

2.1.2. Definíció. ([9, Definition 7.]) $[x], [y] \in \mathbb{R}^N/\mathbb{R}\mathbf{1}$ pontokra a trópusi távolságuk:

$$d_{tr}([x], [y]) = \max_{1 \leq i < j \leq n} |(x_i - y_i) - (x_j - y_j)| = \max_{1 \leq i \leq n} (x_i - y_i) - \min_{1 \leq j \leq n} (x_j - y_j)$$

d_{tr} függvényt trópusi metrikának nevezzük.

T_n teret ezzel a metrikával szokták trópusi fa térnek is nevezni. Belátjuk, hogy ez valóban egy jól definiált metrika.

Bizonyítás. Ahhoz, hogy metrikát kapjunk 3 tulajdonságnak kell teljesülnie.

1. $d_{tr}([x], [y]) \geq 0$ és egyenlőség pontosan akkor teljesül, ha $[x] = [y]$. Az első teljesül, hiszen a maximális koordinátakülönbségből levonva a minimálisat, nyilván nemnegatív számot kapunk. Az is világos, hogy egyenlőség csakis akkor lehetséges, ha $\max_{1 \leq i \leq n} (x_i - y_i) = \min_{1 \leq j \leq n} (x_j - y_j)$, tehát bármely $x_i - y_i$ különbség ugyanannyi, mindre $x_i = c + y_i$ valamilyen konstansra. Ez azonban pont azt jelenti, hogy x és y ekvivalenciaosztálya megegyezik.

2. $d_{tr}([x], [y]) = d_{tr}([y], [x])$, ami ugyancsak teljesül, mert

$$\begin{aligned} \max_{1 \leq i \leq n} (x_i - y_i) - \min_{1 \leq j \leq n} (x_j - y_j) &= \\ -\min_{1 \leq i \leq n} (y_i - x_i) - (-\max_{1 \leq j \leq n} (y_j - x_j)) &= \\ \max_{1 \leq j \leq n} (y_j - x_j) - \max_{1 \leq i \leq n} (y_i - x_i) \end{aligned}$$

3. háromszögegyenlőtlenség: $d_{tr}([x], [y]) \leq d_{tr}([x], [z]) + d_{tr}([z], [y])$ bármely $[x], [y], [z] \in \mathbb{R}^N/\mathbb{R}\mathbf{1}$ esetén.

Tegyük fel, hogy x, y esetén i' és j' -re veszi fel a szélsőértékeket.

$$\begin{aligned} d_{tr}([x], [y]) &= |x_{i'} - y_{i'} - x_{j'} + y_{j'}| = \\ |(x_{i'} - x_{j'} - z_{i'} + z_{j'}) + (z_{i'} - z_{j'} - y_{i'} + y_{j'})| &\leq \\ |x_{i'} - x_{j'} - z_{i'} + z_{j'}| + |z_{i'} - z_{j'} - y_{i'} + y_{j'}| \end{aligned}$$

felhasználva, hogy az abszolútértéknél igaz a háromszög egyenlőtlenség.

Nyilván $d_{tr}([x], [z]) = \max_{1 \leq i < j \leq n} |x_i - x_j - z_i + z_j| \geq |x_{i'} - x_{j'} - z_{i'} + z_{j'}|$

és $d_{tr}([z], [y]) \geq |z_{i'} - z_{j'} - y_{i'} + y_{j'}|$.

Az előzőek alapján $d_{tr}([x], [y]) \leq d_{tr}([x], [z]) + d_{tr}([z], [y])$ teljesül. $\square$

2.2. Trópusi geometria, konvexitás és szakasz

2.2.1. Definíció. ([5, Definition 1.], [15]) Trópusi félgűrű (gyűrű, amelyben nincs megkövetelve az additív inverz létezése) egy $(\mathbb{R} \cup \{-\infty\}, \boxplus, \odot)$ kifejezés, ahol az összeadás és a szorzás a következőket jelenti $(a, b \in \mathbb{R})$:

$$\begin{aligned} a \boxplus b &:= \max(a, b) \\ a \odot b &:= a + b \end{aligned}$$

A nullelem $a - \infty$ és az egységelem $a 0$.

$\mathbb{R}^N$ -ben és $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -ben a műveletek koordinátánként értendők.

A továbbiakban $\mathbf{a} = (a, a, \dots, a)^T$, ahol $a \in \mathbb{R}$. Valamint az ekvivalenciaosztályokat egy reprezentáló elemükkel jelöljük.

2.2.2. Definíció. ([5, Definition 14.], [15]) Legyen $S \subset \mathbb{R}^N/\mathbb{R}\mathbf{1}$.

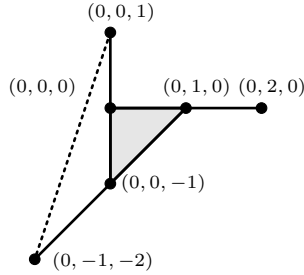
Ha $\mathbf{a} \odot x \boxplus \mathbf{b} \odot y \in S$ minden $x, y \in S$ és $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N$ esetén, akkor azt mondjuk S trópusian konvex.

Egy adott, véges $V \subset \mathbb{R}^N/\mathbb{R}\mathbf{1}$ trópusi konvex burka (vagy trópusi politóppja) a legkisebb olyan trópusian konvex halmaz, amely tartalmazza V -t. Jele (angol szakirodalomnak megfelelően): $tconv(V)$

Az euklideszi geometriához hasonlóan:

$$tconv(V) = \{\mathbf{a}_1 \odot v_1 \boxplus \mathbf{a}_2 \odot v_2 \boxplus \dots \boxplus \mathbf{a}_m \odot v_m \mid \{v_1, \dots, v_m\} = V, \mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^N\}$$

Itt kicsit furcsának tűnhet, hogy a csupaegy vektor egy konstansszorzását adjuk a $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -beli vektorokhoz, de a definíció $\mathbb{R}^N$ -ben is értelmes, így gondolhatunk mi is úgy a műveletekre, hogy elvégezzük az $\mathbb{R}^N$ -beli vektorokra az összeadást, majd a maximum-választást, és a végén "egyszerűsítünk". Az $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ azért lesz fontos a későbbiekben, mert ha két vektor ekvivalens, akkor azonosnak számítanak ebben a térben és a konvex burokba is csak egyszer fog "beszámítani".



7. ábra. ([8, Figure 5.1.]) $(0, 0, 1), (0, 2, 0), (0, -1, -2)$ pontok trópusi konvex burka. Az ábrából látszik, hogy például $(0, 0, 1)$ és $(0, -1, -2)$ között nem a szaggatott szakasz megy (tehát euklideszi értelemben nem konvex a halmaz), hanem a $(0, 0, 1), (0, 0, -1), (0, -1, -2)$ töröttvonal. Ez utóbbira érdemes később visszatérni, a 2.3.3 Állítás bizonyítása után, mert az ott megadott algoritmusból lesz érthető, miért ez a trópusi szakasz köztük. Ott a 2.3.4 Példában le is vezetünk egy trópusi szakasz "készítést".

2.2.3. Definíció. ([5, Definition 13.], [15]) $x, y \in \mathbb{R}^N/\mathbb{R}\mathbf{1}$, a köztük menő trópusi szakasz (tropical line segment) a következő halmazzt jelenti: $\{\mathbf{a} \odot x \boxplus \mathbf{b} \odot y \mid \mathbf{a}, \mathbf{b} \in \mathbb{R}^N\}$

2.2.4. Megjegyzés. ([12]) A trópusi szakasz "hossza" a két pont trópusi metrika szerinti távolsága.

Látni fogjuk a 2.3.3. Állításnál, hogy a trópusi szakasz egy euklideszi töröttvonal. Így a trópusi szakasz "hossza" alatt a töröttvonal trópusi metrika szerinti hosszát, vagyis az egyenes szakaszok trópusi hosszainak összegét értjük. Az előbbi megjegyzést a 2.3.6. Tételnél belátjuk. Világos, hogy két pont közti trópusi szakasz a két pont konvex burka.

2.2.5. Állítás. Egy adott $V \subset \mathbb{R}^N/\mathbb{R}\mathbf{1}$ trópusi konvex burka pontosan a $tconv(V)$ pontjainak trópusi szakaszaiból áll.

Bizonyítás. (saját bizonyítás) Ha veszek két pontot $tconv(V)$ -ből, azoknak a trópusi szakasza benne lesz, hiszen ez volt a trópusian konvex halmaz definíciója. Visszafele, ha veszek egy pontot $tconv(V)$ -ből, akkor az felírható $\mathbf{a}_1 \odot v_1 \boxplus \mathbf{a}_2 \odot v_2 \boxplus \dots \boxplus \mathbf{a}_m \odot v_m$ alakban. Ugyanakkor mivel $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ félgűrű, így teljesül az asszociativitás, tehát $\mathbf{a}_1 \odot v_1 \boxplus \mathbf{a}_2 \odot v_2 \boxplus \dots \boxplus \mathbf{a}_m \odot v_m = ((\dots(\mathbf{a}_1 \odot v_1 \boxplus \mathbf{a}_2 \odot v_2) \boxplus \mathbf{a}_3 \odot v_3) \dots) \boxplus \mathbf{a}_m \odot v_m$, és ha zárójelenként nézzük a műveleteket, akkor mindig két $tconv(V)$ -beli pont trópusi szakaszának egy pontját kapom, ami az előbbieket szerint $tconv(V)$ -ben van. Tehát az összes $tconv(V)$ pont két másiknak a trópusi szakaszán van. $\square$

2.3. Trópusi szakasz tulajdonságai

Innentől U_n -ben vesszük a pontokat és a 2.3.1. Tételben be fogjuk látni, hogy U_n trópusian konvex, tehát nem vezetnek ki a fönti trópusian konvex halmazok és szakaszok belőle. Utána megvizsgáljuk a trópusi szakaszt. Bár a hivatalos jelölésben inkább ω szokott lenni, a későbbiekben mi inkább x, y -t fogunk használni, hogy egyszerűbb legyen a koordinátákra hivatkozni.

2.3.1. Tétel. ([5, Proposition 15.]) Ha ω_1 és $\omega_2 \in U_n$, akkor $\{\mathbf{a} \odot \omega_1 \boxplus \mathbf{b} \odot \omega_2 \mid \mathbf{a}, \mathbf{b} \in \mathbb{R}^N\}$ minden eleme U_n -ben is szerepel, más szavakkal U_n trópusian konvex.

Bizonyítás. Mivel $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ részhalmazában vagyunk, ezért tudjuk, hogy $\mathbf{a} \odot \omega_1$ és $\mathbf{b} \odot \omega_2$ is U_n -ben vannak az ekvivalens párjaik miatt, tehát helyettük vehetjük az ω'_1 és ω'_2 U_n -beli pontokat. Tudjuk $z = \omega'_1 \boxplus \omega'_2$ ultrametrikus pontosan akkor, ha teljesül rá a 3-pont feltétel (vagyis tetszőleges három koordinátát választva, azok között a maximális érték legalább kétszer szerepel). Vegyünk három koordinátát z -ből. Szimmetria miatt feltehetjük, hogy a maximumot $z_{i,j}, z_{i,k}, z_{j,k}$ koordinátákból az elsőn veszi fel és az $\omega'_{1,i,j}$ -vel egyenlő. Ekkor mivel $\omega'_1 \in U_n$, ezért $\omega'_{1,i,k}, \omega'_{1,j,k}$ közül az egyik egyenlő ezzel a maximummal, hiszen az nem lehet, hogy $\omega'_{1,i,k}$ és $\omega'_{1,j,k}$ a nagyobbak, mert akkor $\omega'_{1,j,k} = \omega'_{1,i,k} > \omega'_{1,i,j} = z_{i,j} \geq z_{i,k} \geq \omega'_{1,i,k}$ ellentmondást adna. Legyen $\omega'_{1,i,k} = \omega'_{1,i,j} \geq \omega'_{1,j,k}$. Ekkor

$\omega'_{1_{i,k}} = z_{i,j} \geq z_{i,k} \geq \omega'_{1_{i,k}}$, tehát $z_{i,k} = \omega'_{1_{i,k}} = \omega'_{1_{i,j}} = z_{i,j}$ és ez a maximum a három z koordináta között, tehát valóban legalább kétszer lesz felvéve. Ezzel beláttuk, hogy ultrametrikus a kapott vektor, tehát létezik hozzá ultrametrikus fa is, $z \in U_n$. $\square$

2.3.2. Állítás. $x, y \in U_n$, ekkor $\{a \odot x \boxplus b \odot y \mid a, b \in \mathbb{R}^N\} = \{\lambda \odot x \boxplus y \mid \lambda \in \mathbb{R}^N\}$.

Bizonyítás. (saját bizonyítás) Nyilván a bal oldali tartalmazza a jobb oldali halmazt. A másik irányú tartalmazáshoz, vegyünk egy tetszőleges pontot és alakítsuk át, felhasználva, hogy ha $a + x_i > b + y_i$, akkor $a - b + x_i > y_i$, tehát az i . koordinátára ugyanabból a vektorból fogjuk a maximumot választani (fordított reláció esetén ugyanígy). Valamint nyilván $\mathbf{0} \odot y = y$. Továbbá $\mathbb{R}^N / \mathbb{R}1$ -ben vagyunk, így nem változik a pont, ha a vektor minden koordinátáját b -vel csökkentjük: $\mathbf{a} \odot x \boxplus \mathbf{b} \odot y = (\mathbf{a} - \mathbf{b}) \odot x \boxplus \mathbf{0} \odot y = (\mathbf{a} - \mathbf{b}) \odot x \boxplus y$. Így $\lambda = a - b$ választással a másik irányú tartalmazást is beláttuk. $\square$

Ez az állítás azért fontos, mert a következőkben elegendő az x koordinátáit egy konstansvektorhoz hozzáadni és azt hasonlítani a változatlan y -hoz.

2.3.3. Állítás. *A trópusi szakasz egy euklideszi töröttvonalnak felel meg $\mathbb{R}^N$ -ben ábrázolva. A töröttvonal ($\mathbb{R}^N$ -beli) pontjainak az ekvivalenciaosztályai szerepelnek a 2.2.3. Definíció szerinti trópusi szakaszban.*

Bizonyítás. (saját bizonyítás) Egyszerűbb jelölés miatt legyen $x := \omega_1$ és $y := \omega_2$ és a hozzájuk tartozó ultrametrikus fák T_x és T_y . Vegyük $y - x$ vektor koordinátáit és rendezzük őket nem csökkenő sorrendbe (az egyszerűség kedvéért tegyük fel, hogy az indexek sorrendjében nőnek): $y_1 - x_1 \leq y_2 - x_2 \leq \dots \leq y_N - x_N$

1. $\lambda < y_1 - x_1$, ekkor $\lambda + x_i < y_1 - x_1 + x_i \leq y_i - x_i + x_i = y_i$, tehát $(\lambda \odot x) \boxplus y$ -ből minden koordináta az y -ből lesz kiválasztva, tehát "nem mozdultunk el" az y pontból.

2. $y_1 - x_1 \leq \lambda < y_2 - x_2$ (ha a két különbség egyenlő lenne, akkor természetesen a következő velük már nem egyenlő különbség koordinátát venném felső határnak). Ugyanúgy mint az előbb látszik, hogy az első koordinátánál pont egyenlő lesz $\lambda + x_1$ és y_1 , a többinél azonban még mindig nem érik el az y koordinátáit. Így amíg nem éri el λ a következő legkisebb különbséget, addig az első koordináta mindig $\lambda + x_1$ lesz, hiszen a λ értéke nő és már átlépte az y_1 -t az összeg. A többinél azonban továbbra is az y koordináta lesz a nagyobb, bármekkora is a λ . Ez azt jelenti, hogy olyan pontok lesznek a szakasz elemei, amik az y vektortól az első koordinátában térnek el, még hozzá mind ugyanúgy $\lambda' - (y_1 - x_1) \geq 0$ -gyel nagyobb lesz, ahol λ' az aktuális λ érték (ha az első különbségek közt vannak egyformák, akkor természetesen azoknál is eltér az y -től $\lambda' - (y_1 - x_1)$ -vel, mert $(\lambda' + x_i) - y_i = \lambda' - (y_1 - x_1)$). Ilyen módon az y pontból indulva az $(1, 0, 0, \dots, 0)^T$ vektor irányába haladva jutunk el a következő töréspontig (amikor λ eléri a következő legkisebb különbséget). Hiszen akkor legalább egy újabb koordináta fog "átváltani" y_i -ről $\lambda + x_i$ -re. Pontosabban ott éppen eléri

$\lambda + x_i$ az y_i -t és λ -t tovább növelve jönnek be új " $\lambda + x$ " koordináták, vagyis $(1, 1, 0, \dots, 0)^T$ irányába fogunk onnantól továbbhaladni. Ilyen módon mindig egy $\max(\lambda' + x, y)$ pontból indulunk, ahol λ' éppen egyenlő a k . legkisebb koordinátájával az $y - x$ különbségvektornak. Majd egy $(1, \dots, 1, 0, \dots, 0)^T$ vektor irányába haladok, egészen addig amíg λ el nem éri a $k + 1$. legkisebb különbséget, akkor töréspontunk lesz, onnantól legalább eggyel több egyes lesz az $(1, \dots, 1, 0, \dots, 0)^T$ vektorban. Az eddigiekből látszik, hogy ha eléri valahol az y_i -t a $\lambda' + x_i$, akkor onnantól növelve λ -t az a koordináta is mindig nő, a többi " $\lambda + x$ " koordináta is nő ugyanannyival, de a többi marad " y ".

3. $\lambda \geq y_N - x_N$, ekkor egyenlőség esetén éppen beléptünk egy töréspontba: $y_N - x_N + x_i \geq y_i - x_i + x_i = y_i$, tehát ekkor minden koordinátát $\lambda + x$ határoz meg. Ez azonban azt jelenti, hogy ha nő tovább a λ , akkor is $\lambda + x$ vektor lesz a nagyobb minden koordinátában. Viszont ezek a pontok $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -ben mind x vektorral egyeznek meg, tehát a töröttvonal másik végpontja x lesz, így valóban egy x és y közötti töröttvonalnak felel meg $\mathbb{R}^N$ -ben a trópusi szakasz (ha az ekvivalenciaosztályok megfelelő $\mathbb{R}^N$ -beli megfelelőjét veszem). Mellesleg az is kiderül, hogy legfeljebb N töréspont lehet a trópusi szakaszon (a végpontokkal együtt), hiszen a töréspontokban mindig legalább egy koordináta átvált y -ből $\lambda + x$ -re és vissza nem válthat. Így a töréspontok közötti (euklídeszi geometria szerint) egyenes szakaszok száma is legfeljebb $N - 1$ lehet. $\square$

Ez a bizonyítás egyben egyszerű leírást ad arra, hogyan is lehet egy ilyen trópusi szakaszt elkészíteni. Létezik erre (n levelű fák esetén) $O(n^2 \log n)$ futásidű algoritmus, de ennek bizonyításába nem megyünk bele ([15] Proposition 6).

2.3.4. Példa. ([5, Figure 6.]) $x = (4, 8, 20, 8, 20, 20)$, $y = (20, 20, 20, 8, 8, 4)$ és a 1.2.1. Állítás utáni algoritmust használva (abból kézzel gyorsan vissza lehet fejteni a fát) láthatjuk, hogy a 8. ábra két szélén lévő fát kapjuk (x a jobb oldali). Az előző bizonyítást használjuk, tehát állítsuk sorba az $y - x = (16, 12, 0, 0, -12, -16)$ koordinátákat $\Rightarrow -16 < -12 < 0 < 12 < 16$, tehát 5 töréspont lesz a végpontokkal együtt.

$\lambda = -16$:

$v_1 = \max((4 - 16, 8 - 16, 20 - 16, 8 - 16, 20 - 16, 20 - 16), (20, 20, 20, 8, 8, 4)) = \max((-12, -8, 4, -8, 4, 4), (20, 20, 20, 8, 8, 4)) = (20, 20, 20, 8, 8, 4) = y$ (bal oldali fa). Az is látszik, hogy kisebb λ -ra is az y koordinátái a nagyobbak, valamint az is, hogy csak az utolsó koordinátánál egyenlő a két hasonlítandó vektor, mert ott éri el λ a különbségüket. Azokat a koordinátákat, amiknél a maximum a $\lambda + x$ -ből kerül ki pirossal jelöljük.

$\lambda = -12$:

$v_2 = \max((-8, -4, 8, -4, 8, 8), (20, 20, 20, 8, 8, 4)) = (20, 20, 20, 8, 8, 8)$ vagyis a 2, 3, 4 levelek a közös őszükkel egy kis csillagfát alkotnak és leellenőrizhető, hogy ez pont a 2. ábra balról.

$\lambda = 0$:

$v_3 = \max((4, 8, 20, 8, 20, 20), (20, 20, 20, 8, 8, 4)) = (20, 20, 20, 8, 8, 20)$. Itt az is látszik, hogy ha több koordinátát is érint az aktuális λ , akkor azok egyszerre lesznek pirosak, és innentől pirosak is maradnak a bizonyítás szerint.

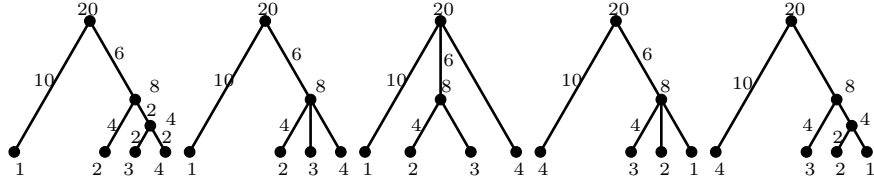
$\lambda = 12 :$

$v_4 = \max((16, 20, 32, 20, 32, 32), (20, 20, 20, 8, 8, 4)) = (20, 20, 32, 20, 32, 32) = (8, 8, 20, 8, 20, 20)$ mivel $\mathbb{R}/\mathbb{R}1$ -ben vagyunk, ezért "visszanormálhattunk" 12-vel.

$\lambda = 16 :$

$v_5 = \max((20, 24, 36, 24, 36, 36), (20, 20, 20, 8, 8, 4)) = (20, 24, 36, 24, 36, 36) = (4, 8, 20, 8, 20, 20) = x.$

Ha vesszük az első töréspontot (v_1), akkor ahogy λ 16-ból 12-be átvált, közben a piros koordináta változik csak, hiszen egyedül őt befolyásolja λ a következő töréspontig. Így a v_1 és v_2 közötti fának az alakja olyan lesz, mint v_1 -é, hiszen csak az utolsó koordináta nő egy kicsit, vagyis a 4 címkéjű belső csúcs (3, 4 levél legközelebbi őse) csúszik felfelé, a többinek maradnia kell, hiszen a többi távolság nem változik. Mivel λ nő, így felfelé csúszik a 4 címkéjű, és amikor eléri a következő töréspontot (amikor a 2, 3, 2, 4 és 3, 4 távolságok megegyeznek), akkor egybecsúszik a két belső csúcs. Mivel ekkor már λ a 2, 4 és 3, 4 párokat is befolyásolja, de a 2, 3-t nem (az még nem piros), ezért 2, 3 távolsága megmarad, vagyis nekik lesz egy külön belső csúcsok, amiből ők 4–4 hosszú éllel válnak ki, de az utolsó két piros koordináta miatt a 4 indexű levéltől távolodniuk kell, tehát egy új belső csúcs felfelé halad, a gyökér felé. Ilyen alakú fa nincsen az ábrán, de könnyen elképzelhető, hiszen a harmadik ábra módusul annyiban, hogy a középső kis részfának a gyökérre eső belső csúcsát mozgatom a jobb "fa száron", de legfeljebb $10 - 4 = 6$ távolságra. Hasonlóan a későbbiekben is bizonyos belső csúcsok "csúszkálnak" a faszárakon.



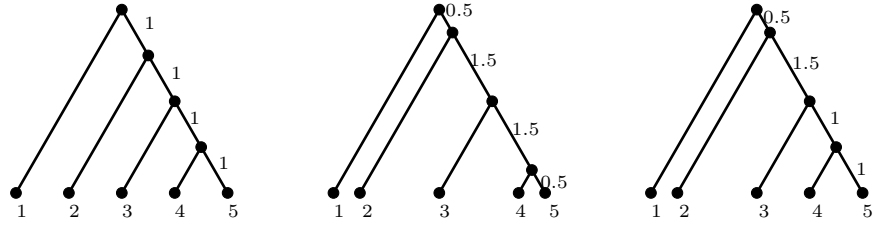
8. ábra. A fák sorban $y = v_1, v_2, v_3, v_4, v_5 = x$ vektorokhoz tartoznak.

A 8. ábrából sejthető, hogy két töréspont között az egyenes szakaszon nem változik a fák alakja, azonban az nem feltétlenül igaz, hogy az egyik végponttal megegyeznek. Ez valóban így van, de csak a 2.5.5. Tétel után a 2.5.6. Állításnál fogjuk belátni. Egyelőre bizonyítás nélkül elfogadjuk ezt. Az azonban nem igaz, hogy ha két fa alakja megegyezik, akkor a köztük menő trópusi szakasz $\mathbb{R}^N$ -ben egy egyenes szakasz.

2.3.5. Példa. $x = (8, 8, 8, 8, 6, 6, 6, 4, 4, 2)$ és $y = (8, 8, 8, 8, 7, 7, 7, 4, 4, 1)$, amiből látszik, hogy a 9. ábrán látható első két fa felel meg nekik.

$y - x = (0, 0, 0, 0, 1, 1, 1, 0, 0, -1)$, tehát lesz egy töréspont a trópusi szakaszukon, mert három eltérő érték van $y - x$ -ben. Az előző példához hasonlóan $\lambda \leq -1$ esetén az y -ban maradunk, ami a második fa. Majd $\lambda = 0$ esetén az utolsó fa alakját kapjuk, aminek a vektora:

$\max((8, 8, 8, 8, 6, 6, 6, 4, 4, 2), (8, 8, 8, 8, 7, 7, 7, 4, 4, 1)) = ((8, 8, 8, 8, 7, 7, 7, 4, 4, 2))$.
Eddig a töréspontig egy egyenes szakaszon haladunk, aminek az irányvektora $(0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$. Ebből a töréspontból azonban az $x = (8, 8, 8, 8, 6, 6, 6, 4, 4, 2)$ -be a $(0, 0, 0, 0, -1, -1, -1, 0, 0, 0)$ irányú egyenes szakaszon jutunk $\mathbb{R}^N$ -ben. Ez nincs ellentmondásban az algoritmussal, mert aszerint $(1, 1, 1, 1, 0, 0, 0, 1, 1, 1)$ irányú vektoron haladunk és a $(9, 9, 9, 9, 7, 7, 7, 5, 5, 3)$ pontba jutunk, de ezt egyszerűsítjük 1-gyel. Tehát az x és y között vezető trópusi szakasz nem egy egyenes szakasz $\mathbb{R}^N$ -ben. Igazából az is látszik ebből a példából, hogy a töréspontnál tényleg irányváltás van, hiszen onnantól újabb koordináták lesznek $\lambda + x$ -ből, amik az irányt megváltoztatják.



9. ábra. Amikor a fák alakja megegyezik, de a trópusi szakasz mégse egy egyenes szakasz.

2.3.6. Tétel. ([15, Proposition 2.]) *A trópusi szakasz hossza az x és az $y \in U_n$ trópusi metrika szerinti távolságával egyenlő. Továbbá a trópusi szakasz egyértelmű két adott pontra.*

Bizonyítás. Az egyenes szakaszok hosszait határozzuk meg először, legyen két töréspont $v_{k'}$ és $v_{k'+1}$, a λ értéke $y_k - x_k$ a $v_{k'}$ pontban, a másikon $y_{k+1} - x_{k+1}$. Az egyszerűség kedvéért továbbra is feltesszük, hogy a 2.3.3 Állítás bizonyításban látott $(1, \dots, 1, 0, \dots, 0)^T$ vektor irányában haladunk a két töréspont között, ahol k db 1-es szerepel az előbbi vektorban. Ekkor 2.1.2 Definíció alapján és a szimmetria miatt $d_{tr}(v_{k'}, v_{k'+1}) = d_{tr}(v_{k'+1}, v_{k'}) =$

$$\max_{1 \leq i \leq n} (v_{k'+1,i} - v_{k',i}) - \min_{1 \leq j \leq n} (v_{k'+1,j} - v_{k',j}) = ((y_{k+1} - x_{k+1}) - (y_k - x_k)) - 0.$$

Ez azért igaz, mert két koordináta vagy azonos a töréspontoknál (ha y -ből van) és akkor 0 a különbség, vagy pedig az egyik koordináta $(\mathbf{y}_{k+1} - \mathbf{x}_{k+1}) \odot x$ -ből, a másik $(\mathbf{y}_k - \mathbf{x}_k) \odot x$ -ből van, amiknél a koordináták különbsége

$$(y_{k+1} - x_{k+1}) - (y_k - x_k) \geq 0.$$

Annál a koordinátánál, ahol $v_{k'+1}$ -ben lett változás, eddig y_{k+1} volt, most pedig az $(\mathbf{y}_{k+1} - \mathbf{x}_{k+1}) \odot x$ miatt éppen egyenlő lesz az ottani koordináta y_{k+1} -gyel, tehát ez is 0 különbségű lesz. Így mivel csak két különbség van és az egyik nemnegatív, nyilván a fenti egyenlőséget kapjuk.

A töröttvonal hosszát úgy kapjuk, hogy összeadjuk a részutakat, vagyis az egyenes szakaszokat. Ezek összege: $(y_2 - x_2) - (y_1 - x_1) + (y_3 - x_3) - (y_2 - x_2) + \dots + (y_n - x_n) - (y_{n-1} - x_{n-1}) = (y_n - x_n) - (y_1 - x_1)$. Ez azonban pont az $y - x$ koordináták közül a maximális mínusz a minimális, tehát a 2.1.2. Definíció szerinti távolsága x és y pontoknak. Ha nem csak egy koordináta változna a

töréspontokban, akkor is ugyanezt a teleszkópikus összeget kapnánk.

A második állítás a 2.3.3. Állítás bizonyításából látszik, hiszen minden lépés egyértelműen következett a definíciókból, így csak az ott leírt töröttvonalat kaphatjuk. $\square$

2.4. Trópusi szakasz és ultrametrikus fák kapcsolata

Ebben a fejezetben további fontos tulajdonságait vizsgáljuk a trópusi szakaszoknak, valamint az alkalmazásukat is bemutatjuk.

2.4.1. Definíció. ([3, Definition 2.20]) Legyen T egy ultrametrikus fa, X pedig a leveleinek a halmaza. Azt mondjuk, hogy az $X_0 \subset X$ levélhalmaz a T fa egy kládja (angol: **clade**, biológiában: élőlények egy csoportját jelenti), ha létezik egy olyan v belső csúcsa T -nek, hogy annak pontosan az X_0 elemei a levél-utódjai. Ha létezik ilyen v csúcs, akkor az X_0 kládja alatt a v gyökerű részfat szoktuk érteni.

2.4.2. Példa. Ha visszanézzük a 8. ábrát, ott például az $X_0 = \{2, 3, 4\}$ levelek egy kládot adnak a bal oldali fánál, mert a 8 címkéjű belső csúcs leszármazottjai. De nem alkot kládot ugyanez a 3 levél a jobb oldaliban, hiszen ott a gyökér lenne az a közös ő, akinek leszármazottja a 4-es levél, viszont akkor az 1-nek is a kládban kellene lennie.

2.4.3. Tétel. ([15, Theorem 8.]) Legyen T_1 és T_2 két ultrametrikus fa, a hozzájuk tartozó vektorok pedig x és y . Továbbá tegyük fel, hogy X levél halmazuk megegyezik ($|X| = n$), és az $X_0 \subset X$ levelekhez tartozó klád ugyanolyan alakú a két fában (tehát ha egy T_1 -beli v csúcsból i, j levelek együtt jobbra ágaznak, hogy később szétváljanak, k pedig balra, akkor a T_2 fához tartozó kládban is ugyanígy ágaznak ki az i, j, k utódok v -ből). Vegyük x és y trópusi szakaszát. Ha ezen egy ω pont szerepel, akkor a hozzá tartozó T ultrametrikus fához is ugyanolyan alakú X_0 levelű klád tartozik, mint a végponti fákhhoz. (2.3.1., 1.1.8. Tételek és 1.1.9. Állítás miatt létezik és egyértelmű T .)

Bizonyítás. A bizonyítás 3 részből áll.

1. Az X_0 -hoz tartozó klád (T_1 fánál) is egy ultrametrikus fa. Ez azért igaz, mert hiszen ez a részfa olyan, hogy a legfelső belső csúcstól lefele csak az ő utódjai vannak. Ha ezekre a belső csúcsokra ráírom az x értékeket, amik az eredeti T_1 fában is rá voltak írva, akkor mivel a levelek legközelebbi ősei és a faalak nem változott, így továbbra is minden feltétel megfelel a 1.1.7. Definíciónak, a klád részfa ultrametrikus. Avagy csak egyszerűen igaz marad a 3-pont feltétel az itteni indexekre, mert a távolságmátrixbeli értékeiket a részfa határozta meg és az nem változott.

2. A közös X_0 klád a T fában is klád lesz. Ehhez azt kell belátni, hogy ha $i, j \in X_0$ és $k \notin X_0$, akkor a T -ben az i, j legközelebbi közös őse lentebb van, mint a k -val közös ősiük. Mivel i, j közös őse lentebb lenne, így a gyökérbe vezető útjuk közös a legközelebbi közös ősiüktől, tehát ha k elérhető a gyökérbe vezető útjuk egy csúcsából, akkor az i, k és j, k legközelebbi közös őse is ugyanaz a

pont. Ha ez teljesül bármely fönti indexhármásra, akkor valóban egy kis részfat alkotnak az X_0 levelek, a többi levél pedig föntebb kapcsolódik a fájukhoz. Így lenne egy v belső csúcs T -ben, aminek pontosan X_0 levelek az utódai. Ehhez azt kell belátni, hogy $\omega_{ij} < \omega_{ik} = \omega_{jk}$, hiszen a gyökértől lefelé szigorúan csökkenniük kell az ω értékeinek.

$\omega_{ij} = \max(\lambda + x_{ij}, y_{ij})$ és $\omega_{ik} = \max(\lambda + x_{ik}, y_{ik})$ az eddigiek szerint. Viszont az X_0 levelű részfa klád volt T_1 -ben és T_2 -ben, ezért ott teljesül az indexekre a fenti feltétel, amiből: $\lambda + x_{ij} < \lambda + x_{ik}, y_{ij} < y_{ik} \Rightarrow \omega_{ij} < \omega_{ik}$, valamint az U_n trópusian konvex, ezért T ultrametrikus, tehát $\omega_{ik} = \omega_{jk}$ is teljesül.

3. Most még azt kell belátni, hogy a T -beli X_0 levelekhez tartozó klád ugyanolyan alakú, mint a két végpontban.

Ehhez először vegyük észre, hogy ha egy ω_T vektornak a koordinátái pontosan ugyanúgy rendezhetőek sorba, mint egy $\omega_{T'}$ vektoré (vagyis ha pl. 1, 2 levelek távolsága kisebb 3, 4 levelekénél, akkor a másikon is ugyanúgy viszonyolnak egymáshoz), akkor T és T' fa alakra pontosan ugyanolyan, más szóval minden belső csúcsból ugyanúgy és ugyanazok az utódok ágaznak ki a két fában. Ez a 1.1.8. Tétel bizonyításából következik, hiszen minden i indexre ugyanúgy rendezzük sorba a hozzá tartozó távolságokat, majd azokhoz ugyanúgy viszonyulnak a többiek távolságai.

Így tehát azt kell belátni, hogy ha veszek 3 indexet X_0 -ból, akkor ha T_1 -ben $x_{ij} \leq x_{ik} = x_{jk}$, akkor $\omega_{ij} \leq \omega_{ik} = \omega_{jk}$ is teljesüljön (az 1. pontban láttuk, hogy a klád részfája ultrametrikus ha az eredeti fa ultrametrikus). Mivel a T_1 és T_2 fában az X_0 klád részfa megegyezik, ezért ott is $y_{ij} \leq y_{ik} = y_{jk}$ teljesül. Ha egyenlőség van az x -es tagnál, akkor y -nál is egyenlőség lesz és azt kell bizonyítani még, hogy pontosan ekkor lesz ω -nál is.

$$\omega_{ij} = \max(\lambda + x_{ij}, y_{ij})$$

$$\omega_{ik} = \max(\lambda + x_{ik}, y_{ik})$$

$$\omega_{jk} = \max(\lambda + x_{jk}, y_{jk})$$

Legyen $\alpha_1 = y_{ij} - x_{ij}$ és $\alpha_2 = y_{ik} - x_{ik} = y_{jk} - x_{jk}$, ezek alapján 4 esetet különböztetünk meg.

$a, \alpha_1, \alpha_2 > \lambda$, ekkor a y koordináták lesznek a nagyobbak ($y_{ij} - x_{ij} > \lambda \Leftrightarrow y_{ij} > \lambda + x_{ij}$, ugyanígy a másik két esetben), tehát ω megfelelő koordinátái megegyeznek y koordinátáival, így nyilván teljesülnek a feltételek.

$b, \alpha_1 \geq \lambda \geq \alpha_2$, ekkor $y_{ij} \geq \lambda + x_{ij}, \lambda + x_{ik} \geq y_{ik}, \lambda + x_{jk} \geq y_{jk}$, tehát $\omega_{ij} = y_{ij}, \omega_{ik} = \lambda + x_{ik}, \omega_{jk} = \lambda + x_{jk}$. Ebből kapjuk: $\omega_{ij} = y_{ij} \leq y_{ik} \leq \lambda + x_{ik} = \omega_{ik}$, hasonlóan $\omega_{ij} \leq \omega_{jk}$ és az ultrametrikusság miatt $\omega_{ik} = \omega_{jk}$. Ebből a levezetésből az is látszik, hogy pontosan akkor lesz ω -nál egyenlőség, ha $y_{ij} = y_{ik} = y_{jk}$ is teljesül (és akkor $x_{ij} = x_{ik} = x_{jk}$ is igaz).

$c, \alpha_2 \geq \lambda \geq \alpha_1$, akkor hasonlóan az előzőhöz: $\omega_{ij} = \lambda + x_{ij}, \omega_{ik} = y_{ik}, \omega_{jk} = y_{jk}$, amiből $\omega_{ij} = \lambda + x_{ij} \leq \lambda + x_{ik} \leq y_{ik} = \omega_{ik} = \omega_{jk}$. Egyenlőség ismét pontosan akkor, ha x és y esetén is egyenlőség volt.

$d, \lambda > \alpha_1, \alpha_2$, ekkor az a , esethez hasonlóan ω megfelelő koordinátái egyenlőek $\lambda + x$ megfelelő koordinátáival, viszont mivel $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -ben vagyunk, így x koordinátáival egyeznek meg.

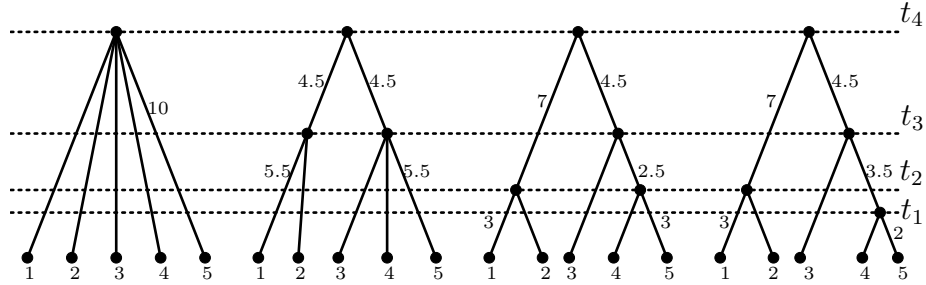
Ezzel beláttuk, hogy X_0 kládot alkot T -ben is, a kládja ultrametrikus részfa és a fa szerkezete olyan mint az eredeti T_1 és T_2 fában. $\square$

Már korábban is említettük, hogy bár többnyire a bináris fákat szeretik a szakirodalmakban használni, vannak többgyerekes fák is. Az $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ origója a $(0, 0, \dots, 0)^T$ pont, ami N koordinátából áll és ennek a csillag ultrametrikus fa felel meg, vagyis amikor minden levél egyből az origóból indul ki. Ezt jelöljük O -val. Igazából ekkor valami pozitív súly lenne az éleken (amennyi idő eltelt a szétválás óta), de a tér miatt ez a $(0, 0, \dots, 0)^T$ vektorral ekvivalens vektort adja, hiszen minden koordináta 2-szer az élekre írt súly.

2.4.4. Definíció. ([15, Definition 7.]) Egy adott n levelű T fához tartozó $\{t_1, t_2, \dots, t_k\}$ halmaz elemeit specializációs időknek nevezzük, ahol t_l az l . legkisebb $\omega_{ij} > 0$ érték fele.

Tehát ezek az időpontok a T vektorában szereplő értékeknek a fele növekvő sorrendben. Vagyis, ha pl. ω_T koordinátái balról jobbra nőnek, akkor $y_{1,2} = t_1$, $y_{1,3} = y_{1,4} = t_2, \dots$ $y_{n-1,n} = t_k$.

2.4.5. Példa. Az alábbi ábrán a 4. fához a vektor $(6, 20, 20, 20, 20, 20, 20, 20, 11, 11, 4)^T$, tehát $t_1 = 2$, $t_2 = 3$, $t_3 = 5.5$ és $t_4 = 10$. Ezt az ábra sorozatot érdemes nézni a 2.4.6 Tétel bizonyításához is, hogy az origóból (bal oldali csillagfa) indulva hogyan alakul ki a jobb oldali fa. Ezek a t specializációs idők egyébként a 1.1.7. és 1.2.4. Definícióknál megadott belső csúcs címkézésekkel egyeznek meg.



10. ábra. Az origótól a $(6, 20, 20, 20, 20, 20, 20, 20, 11, 11, 4)^T$ vektorhoz tartozó fáig az ultrametrikus szakasz töréspontjai.

2.4.6. Tétel. ([15, Theorem 7.]) Tegyük fel, hogy T egy n levelű ultrametrikus fa és a hozzá tartozó specializációs idők halmaza $\{t_1, t_2, \dots, t_k\}$ (ahol $k \leq n - 1$, az 1. fejezetben leírtak szerint). Ekkor az origó (csillagfa) és a T pont között vezető trópusi szakaszon a töréspontok $T = T^1, T^2, \dots, T^k = O$ ultrametrikus fák lesznek, ahol T^i fa úgy néz ki, hogy a hozzá tartozó specializációs idők halmaza $\{t_i, t_2, \dots, t_k\}$ (sőt az időkhöz tartozozó belső csúcsokig alakra megegyezik T -vel a T^i fa, de még lehetnek később (lentebb) szétválások a T fában).

Bizonyítás. Érdemes a bizonyításhoz a 10. ábrát nézni. A régi jelölésekkel legyen $x = (0, \dots, 0)^T = (20, \dots, 20)^T$ és y a T fához tartozó

vektor. Azért használunk itt az origó vektoránál 20-t, mert egyrészt $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ -ben vagyunk, másrészt az ábráknál is ennyi szokott lenni a maximális távolság és így a trópusi szakasz mentén előforduló vektorokat 20-ra fogjuk "visszanormálni". λ az $x - y$ különbségvektor koordinátáin megy végig úgy, hogy az értéke nő, tehát a 0-ról indul (hiszen a maximális koordináta (20) x -ben és y -ban is szerepel, és x 20-as koordinátáiból vonjuk le az y legfeljebb 20, nemnegatív koordinátáit, tehát lesz nulla koordináta és a többi pozitív lesz $x - y$ -ban). A trópusi szakasz töréspontjait, beleértve a két végpontot úgy kapjuk, hogy a $\max(\lambda + y, x)$ vektort kiszámoljuk, ahol λ mindig a következő különbségvektorbeli koordináta. Amíg $\lambda \leq 0$, addig $\max(\mathbf{0} + y, x) = x$, hiszen a legnagyobb koordináta lesz mindenhol a nagyobb. Így az origó csillagfájából indulunk, és bizonyos koordinátáknál a $\omega_0 = \max(\mathbf{0} + y, x)$ vektorban egyenlőség lesz (példában az 1, 3; 1, 4; 1, 5; 2, 3; 2, 4; 2, 5 koordinátáknál, hiszen ott szerepel a 20). A második (igazi) töréspontot akkor kapjuk, amikor λ egyenlő a következő legkisebb koordinátával $x - y$ -ből (a vektor legyen ω_1). Viszont vegyük észre, hogy ez a T fa vektorának legnagyobb és második legnagyobb koordinátájának a különbsége, hiszen az x vektort a maximális koordinátára "állítottuk". Ekkor a töréspont vektorában azoknál a koordinátáknál, akik az előbb éppen egyenlőek voltak x -szel (20-szal), azoknál most $\lambda + y$ adja a koordinátát, mert már nagyobb lesz az x megfelelő koordinátájánál (volt a 2.3.3 Tétel bizonyításában), a többiek közül lesznek, akiknél éppen most lesz $\lambda + y = x$, a többieknél pedig továbbra is $\lambda + y < x$ lesz, tehát x koordinátája marad ott. Viszont mondtuk, hogy visszanormálunk 20-ra. Most bizonyosak éppen 20 értékűek, a többi $20 + \text{aktuális } \lambda$, így a ω_1 minden koordinátáját csökkentem λ -val. (Ez a példában a $\lambda = 20 - 11 = 9$, tehát az $\omega_1 = (20, 29, 29, 29, 29, 29, 29, 20, 20, 20)^T = (11, 20, 20, 20, 20, 20, 11, 11, 11)^T$.) Ekkor igaz lesz, hogy a gyökér fönt marad, az ő "szintjéhez" odaírhatjuk a t_k -t, és a most létrejött új belső csúcsok szintjéhez odaírhatjuk t_{k-1} -t. Ez azért igaz, mert egyrészt létrejött új belső csúcs, hiszen volt 20 és $20 + \lambda$ koordinátánk is, így nem lehet csillagfánk és csak ez a két érték volt, ezért csak két specializációs idő lehet. Valamint azoknál a levélpároknál, akiknek T -ben maximális a távolságuk, már az előbb ω_0 -nál a T -beli távolságuk szerepelt a vektorban és most ω_1 -ben is, hiszen ők voltak a $20 + \lambda$ koordinátájú vektorok. Azok, akikhez T -ben a második legnagyobb távolság tartozott (példában 3, 4 és 3, 5), azoknál most éppen egyenlő lett $\lambda + y$ és x , tehát a λ levonása után visszakaptuk az y -beli távolságukat és innentől végig ez is marad, mert a következő töréspontoknál az ő értékük is $\lambda + y$ -ből lesz, így amikor visszanormálunk majd, akkor y koordinátáját kapjuk. Ekkor a belső csúcstól lefelé az élen valóban a második legnagyobb y koordináta fele lesz, tehát a t_{k-1} . Ezért fölfelé a belső élen $t_k - t_{k-1}$ lesz, ami a T fában is a belső él súlya volt. Ezt az eljárást folytatva, hasonló gondolatmenet szerint, mindig azt fogjuk kapni, hogy a töréspontnál a fa olyan, hogy az aktuális t_i fölött a fa alakja már olyan, mint a T fában (hiszen azokhoz a belső csúcsokhoz már az y -beli távolság-koordináta fele tartozik, mert a $\lambda + y$ -ből van a koordináta). Az aktuális t_i szintjén pedig lesz egy olyan belső csúcs, amihez tartozik legalább két levél, akiknek a távolsága most érte el az T -belit. A többi levél legközelebbi közös őse még lejjebb csúszhat, de a felsők már rögzültek. Így valóban a T fa

alakul ki föntről lefelé és minden t_i specializációs idő szerepelni fog lefelé, hiszen a trópusi szakasz bejárásakor mind szerepelni fog (mert λ a $t_k - t_i$ értékeken halad végig). $\square$

Felmerül a kérdés, hogy a trópusi szakasza két pontnak mikor megy át az origón.

2.4.7. Állítás. ([15, Lemma 1.]) *Ha T_1 és T_2 két ultrametrikus, $n = 3$ levelű, a csillagfától különböző fa, amiknek az alakja eltér, akkor a trópusi szakaszukon a töréspont az origó.*

Bizonyítás. Először is megjegyezzük, hogy ha azonos lenne az alakjuk, akkor egy egyenes szakasz lenne a trópusi szakaszuk. Ez azért igaz, mert feltehetjük, hogy a gyökértől a levelek ugyanakkora távolságra vannak a két fában, hiszen $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -ben vagyunk, úgy csökkentjük az egyik vektort egy konstans vektorral, hogy a maximális koordinátája egyenlő legyen a másikéval. Ekkor feltehetjük, hogy az 11. ábrán látható T_1 fánk van. Akkor a két fa vektora csak abban az egy koordinátában térhet el, ami a 2, 3 levelek távolságát adja a fában. Így a különbségvektorban azok különbsége és a 0 lesz, tehát nem lesz töréspont a trópusi szakaszon. Vagyis a trópusi szakasz egy egyenes szakasz, vagy egy pont, ha a két vektor megegyezne. Emiatt az origón sem mehet át, hiszen akkor változna a fa alakja, az viszont az egyenes szakaszon nem változhat (2.5.6. Állítás) és ha a két végpont nem azonos, akkor nem egy pontból áll az egyenes szakasz.

Feltehetjük, hogy a két eltérő alakú fa az 11. ábra szerinti. Ekkor legyen T_1 vektora $x = (x_{12}, x_{13}, x_{23})$, T_2 -é $y = (y_{12}, y_{13}, y_{23})$. Az ábrából, az ultrametrikusságból és abból, hogy egyik se az origó csillagfája következik, hogy $x_{23} < x_{12} = x_{13}$ és $y_{13} < y_{12} = y_{23}$. Ebből $y - x$ három koordinátája: $y_{13} - x_{13} < y_{12} - x_{12} < y_{23} - x_{23}$, tehát látszik, hogy egyetlen töréspont lesz és annál $\lambda = y_{12} - x_{12}$ (2.3.3 Állítás bizonyításából). Tehát a töréspont három koordinátája:

$$\omega_{12} = \max(y_{12} - x_{12} + x_{12}, y_{12}) = y_{12},$$

$$\omega_{13} = \max(y_{12} - x_{12} + x_{13}, y_{13}) = y_{12} - x_{12} + x_{13} = y_{12} - x_{12} + x_{12} = y_{12} (> y_{13}),$$

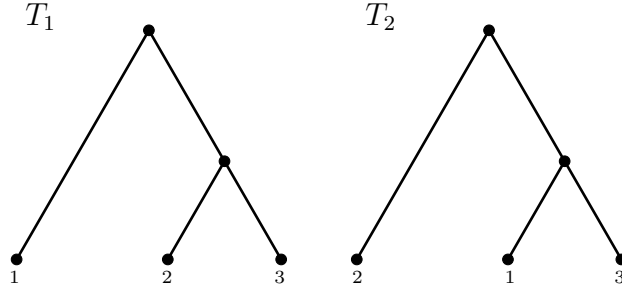
Mivel $-x_{12} + x_{23} < 0$, ezért $y_{12} - x_{12} + x_{23} < y_{12} = y_{23}$

$$\omega_{23} = \max(y_{12} - x_{12} + x_{23}, y_{23}) = y_{23} = y_{12}$$

Mivel azonban $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -ben vagyunk, így ω megegyezik az origó vektorával, tehát a töréspont valóban az origó. $\square$

Most ki fogunk mondani egy tételt, aminek a bizonyításához azonban vissza kell térnünk általánosságban a trópusi konvex burokhöz. Ennek a tételnek és annak az állításnak, hogy a trópusi szakasz egyenes szakaszain a faalak megmarad, a bizonyítása a következő fejezetben lesz.

2.4.8. Tétel. ([15, Theorem 5.]) *Ha $n \geq 5$, akkor 0 annak a valószínűsége, hogy két véletlenül választott n levelű, (nem feltétlenül bináris, de megköveteljük, hogy a gyökérnek két gyereke legyen) ultrametrikus fának a trópusi szakasza átmeny az origón. Más szavakkal, ha egyik kiválasztott végpont sem az origó, akkor a trópusi szakaszuk 0 valószínűséggel megy át az origón.*



11. ábra. $n = 3$ levelű, eltérő alakú fák.

2.5. Trópusi konvex burok és az ultrametrikus fák kapcsolata

A következő definíció $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -ben van kimondva, de nekünk igazából csak az ultrametrikus térben lesz fontos. Ott is csak azért, hogy néhány érdekességet megállapítsunk a konvex burokról és a 2.5.3. Definíciót egy kicsit előkészítsük.

2.5.1. Definíció. ([11, Definition 2.6.]) $P := tconv(V) = tconv(v_1, \dots, v_m)$, ahol $v_1, \dots, v_m \in U_n$. Legyen $\pi_P : \mathbb{R}^N/\mathbb{R}\mathbf{1} \rightarrow P$ egy olyan leképezés, hogy

$$\pi_P(\omega) = \lambda_1 \odot v_1 \boxplus \dots \boxplus \lambda_m \odot v_m,$$

ahol $\lambda_i = \min_j((\omega - v_i)_j)$.

2.5.2. Megjegyzés. Vegyük észre, hogy ezzel a P -be képezünk minden ω pontot, hiszen $\lambda_i \in \mathbb{R}^N$ és így $tconv(V)$ egy elemét kapjuk 2.2.2. Definíció szerint. Ha $\omega \in P$ volt, akkor önmagába képezi π_P . Ekkor ugyanis

$$\omega = \max((\mathbf{a}_1 + v_1), (\mathbf{a}_2 + v_2), \dots, (\mathbf{a}_m + v_m))$$

valamilyen $\mathbf{a}_i$ valósakra. Tehát minden $1 \leq j \leq N$ -re létezik $1 \leq i \leq m$ index, hogy $\omega_j = (\mathbf{a}_i + v_i)_j = a_i + v_{i,j} \Rightarrow \omega_j - v_{i,j} = a_i$ és a többi koordinátára $\omega_{k \neq j} \geq (\mathbf{a}_i + v_i)_k = a_i + v_{i,k}$. Átrendezve kapjuk: $\omega_j - v_{i,j} = a_i \leq \omega_k - v_{i,k}$ minden k -ra, amiből: $a_i = \min_l((\omega - v_i)_l)$, vagyis $\lambda_i = a_i$. Ebből pontosan az következik, hogy ω felírása $tconv(V)$ elemeként ugyanaz, mint a π_P leképezésnél a képe, vagyis önmaga.

2.5.3. Definíció. ([11, Definition 2.7.]) $P = tconv(V) = tconv(v_1, \dots, v_m)$, $\omega \in U_n$ P szerinti típusának (**type**) nevezzük a $Q = (Q_1, \dots, Q_N)$ halmazt, ahol Q_j is egy v -index halmazt jelent. Az $1 \leq i \leq m$ index pontosan akkor kerül a Q_j -be, ha $v_{i,j} - \omega_j = \max(v_{i_1} - \omega_1, \dots, v_{i_N} - \omega_N)$, más szóval, ha az i . indexre nézve a zárójelben lévő különbségek közül a j -re veszi fel a maximumot.

A P trópusi konvex burok pontjai pontosan azok a ω -ák, akiknél egyik Q_j sem üres. Az azonos típusú pontok halmazát cellának hívjuk (tehát amikor a Q_j halmazok páronként megegyeznek).

2.5.4. Megjegyzés. Az utolsó rész nem triviális az előbbi definícióban. Mint azt a 2.5.2 Megjegyzésben is láttuk, ha $\omega \in P$, akkor

$\omega = \lambda_1 \odot v_1 \boxplus \dots \boxplus \lambda_m \odot v_m$, ahol $\lambda_i = \min_j((\omega - v_i)_j) = -\max_j((v_i - \omega)_j)$, vagyis az i index pontosan akkor kerül Q_j -be, ha λ_i a j . koordinátánál veszi fel a minimumot/maximumot. Továbbá, ha $\omega \in P$, akkor minden $1 \leq j \leq N$ indexre $\omega_j = \lambda_i + v_{i,j}$ valamilyen $1 \leq i \leq m$ indexre a 2.2.2 Definíció alapján. Ez azonban tovább alakítva $\lambda_i = \omega_j - v_{i,j} = (\omega - v_i)_j$, vagyis a j . koordinátára veszi fel a minimumot λ_i . Így azt kaptuk, hogy teszőleges $1 \leq j \leq N$ -re Q_j -be fog kerülni egy i index, vagyis nem lesznek üresek.

Visszafele, ha igaz az, hogy egyik Q_j sem üres, akkor legyen $i \in Q_j$ (rögzítsük ezt a két indexet). Ekkor az előbbi $\min - \max$ egyenlőség szerint a π_P leképezésnél definiált λ_i $(\omega - v_i)_j$ lesz, vagyis $\omega_j = \lambda_i + v_{i,j}$. Ha $k \in Q_j$, akkor ugyanígy $\omega_j = \lambda_k + v_{k,j}$, ellenben ha $l \notin Q_j$, akkor $\lambda_l = \omega_{j'} - v_{l,j'} < \omega_j - v_{l,j}$, amiből $\omega_j > \lambda_l + v_{l,j}$. Ezekből viszont az adódik, hogy $\omega_j = \max((\lambda_1 + v_{1,j}), \dots, (\lambda_i + v_{i,j}), \dots, (\lambda_m + v_{m,j}))$, ahol i -re (és minden Q_j -beli indexre) veszi fel a maximumot. Mivel ez minden $1 \leq j \leq N$ indexre ugyanígy igaz lesz, így $\omega = \lambda_1 \odot v_1 \boxplus \dots \boxplus \lambda_m \odot v_m$ -t kapjuk, tehát $\omega \in P$.

Érdemes itt visszanézni a 7. ábrát, ahol az egyenes szakaszok cellákat alkotnak (ezt a következő tételben be is látjuk).

2.5.5. Tétel. ([11, Theorem 3.2.]) $P := tconv(V) = tconv(v_1, \dots, v_m)$, ahol $v_1, \dots, v_m \in U_n$. Ekkor ha $x, y \in P$ és ugyanabban a cellában vannak, akkor a faalakjuk megegyezik.

Bizonyítás. Mivel azonos cellában vannak, így a Q_j halmazaik páronként megegyeznek. Már láttuk a 2.4.3. Tétel bizonyításában (3. pont), hogy ha x és y koordinátái ugyanúgy rendezhetőek sorba (ugyanúgy viszonyulnak egymáshoz a koordinátáik), akkor a faalakjuk megegyezik. Ismét a 3-pont feltételt használjuk, legyen három levélindex i, j, k (innenntől a v_i helyett v_a -t fogunk írni, hogy ne keveredjenek a levélindexekkel).

1. eset. $x_{i,j} = x_{i,k} = x_{j,k}$ kell, hogy ez pontosan akkor igaz, ha $y_{i,j} = y_{i,k} = y_{j,k}$ is teljesül. Nyilván szimmetria miatt elég belátni, hogy ha az x -es rész igaz, abból következik az y -os rész. Mivel $x \in P$, így valamilyen $1 \leq a \leq m$ indexre $v_{a,i,j} - x_{i,j} = \max_{l_1, l_2}(v_a - x)_{l_1, l_2} = -\lambda_a^x$, mert $Q_{i,j}$ nem üres halmaz (tehát $a \in Q_{i,j}$). Tehát $v_{a,i,j} - x_{i,j} \geq v_{a,i,k} - x_{i,k}$ és $v_{a,i,j} - x_{i,j} \geq v_{a,j,k} - x_{j,k}$, amiből az x -ek egyenlősége miatt $v_{a,i,j} \geq v_{a,i,k}$ és $v_{a,i,j} \geq v_{a,j,k}$ következik. Viszont v_a is ultrametrikus, így a maximum kétszer vétetik fel az előző három értéknél, így szimmetria miatt feltehetjük, hogy $v_{a,i,j} = v_{a,i,k}$. Ekkor $-\lambda_a^x = v_{a,i,j} - x_{i,j} = v_{a,i,k} - x_{i,k}$, tehát az a index bekerül a $Q_{i,k}$ halmazba is. De x és y Q halmazaik megegyeznek, ezért $v_{a,i,j} - y_{i,j} = -\lambda_a^y = v_{a,i,k} - y_{i,k} \geq v_{a,j,k} - y_{j,k}$. Mivel $v_{a,i,j} = v_{a,i,k}$, így $y_{i,j} = y_{i,k}$ következik belőle. Mi a föntieket az $v_{a,i,j} = v_{a,i,k}$ egyenlőségből vontuk le úgy, hogy $Q_{i,j}$ halmazból az a pontot vettük és utána $v_a - x$ koordinátáját vizsgáltuk. Ha $Q_{j,k}$ -ból veszek egy b indexet, akkor $v_b - x$ -nek a j, k koordinátája lesz maximális és pontosan egy ugyanilyen gondolatmenet után kijön, hogy $y_{j,k} = y_{i,j}$ vagy $y_{j,k} = y_{i,k}$ attól függően, hogy b index $Q_{i,j}$ -ben vagy $Q_{i,k}$ -ban van-e még benne. Akármelyik is jön ki a $y_{i,j} = y_{i,k} = y_{j,k}$ fog következni belőle. Ezzel az 1. esetet beláttuk.

2. eset. Tegyük fel, hogy $x_{i,j} = x_{i,k} > x_{j,k}$, azt kell belátni, hogy $y_{i,j} =$

$y_{i,k} > y_{j,k}$ következik belőle. Mivel x és y is ultrametrikus, így azt kell belátni csak, hogy a minimum ugyanarra az indexre lesz felvéve, mert akkor a másik kettő koordináta egyenlőek lesznek (mivel nagyobb egyenlőek). Ha egyenlőség lenne a "nagyobbak" és a "kisebbik" között, akkor az első eset szerint következne a másik vektor koordinátáinak egyenlősége is. Tehát tegyük fel, hogy $x_{i,j} = x_{i,k} > x_{j,k}$, de $y_{i,j} = y_{j,k} > y_{i,k}$. Ekkor mivel mindkettő pont P -ben van, így $Q_{i,j}$ nem üres, legyen a egy eleme (ez x és y esetén is teljesül). Definíció szerint: $v_{a_{i,j}} - x_{i,j} \geq v_{a_{j,k}} - x_{j,k}$ és $v_{a_{i,j}} - y_{i,j} \geq v_{a_{i,k}} - y_{i,k}$. Átrendezve az elsőt kapjuk: $v_{a_{i,j}} - v_{a_{j,k}} \geq x_{i,j} - x_{j,k}$ és $v_{a_{i,j}} - v_{a_{i,k}} \geq y_{i,j} - y_{i,k}$, de mivel $x_{i,j} > x_{j,k}$ és $y_{i,j} > y_{i,k}$, így pozitívak a v_a különbségek, tehát $v_{a_{i,j}} > v_{a_{j,k}}$ és $v_{a_{i,j}} > v_{a_{i,k}}$, tehát az ultrametrikus v_a -ban a maximum csak egyszer vétetik fel az előbbi három indexnél, ami ellentmondás. Az ellentmondás abból jött, hogy feltettük, eltérő indexekre veszik fel a minimumot. Ezzel beláttuk, hogy ha csak kétszer veszi fel a maximumot három indexre, akkor a 3. index, amihez a legkisebb koordináta tartozik, megegyezik x -ben és y -ban, és akkor a másik két koordináta nagyobb nála (szigorúan) és egyenlőek. Ezzel a tételt beláttuk, mert ultrametrikus adatok esetén x három koordinátája csak ilyen kapcsolatban lehet egymással. $\square$

Mint azt korábban jeleztük az előző fejezetben, ebből a tételből következik a következő állítás.

2.5.6. Állítás. *Ha egy $\omega_1, \omega_2 \in U_n$ és a trópusi szakasz köztük egy egyenes szakasz, akkor a szakasz bármely két (a végpontokkal nem azonos) x, y pontjához tartozó ultrametrikus fák alakja megegyezik.*

Bizonyítás. (saját bizonyítás) Itt $tconv(V) = tconv(\omega_1, \omega_2)$ és a 2.3.3 Állítás bizonyításából tudjuk, hogy csak két λ érték lehet, amikre az igaz, hogy $\lambda = \lambda_1$ esetén ω_1 -ben maradunk, utána $\lambda_1 < \lambda < \lambda_2$ esetén a szakasz belső pontjait kapjuk, akiknek a vektorában bizonyos koordináták megegyeznek $\lambda + \omega_1$ -val, a többi ω_2 -vel, $\lambda = \lambda_2$ esetén pedig ω_2 -ben vagyunk. Ez viszont két belső pontra pont azt jelenti, hogy ha vesszük trópusian konvex felírását $(\lambda \odot \omega_1 \boxplus \omega_2)$, akkor ugyanabból a (végponti) vektorból veszik fel a maximumot. Láttuk a 2.5.4. Megjegyzésben, hogy ha $x = \lambda \odot \omega_1 \boxplus \omega_2$, akkor az 1-es pontosan akkor kerül be Q_{ij} -be, ha x_{ij} a $\lambda + \omega_1$ -ből van, és a 2 pontosan akkor kerül Q_{ij} -be, ha x_{ij} az ω_2 -ből van. Mivel azonban ez x és y esetén is megegyezik, így minden ij -re Q_{ij} ugyanazokból az indexekből fog állni, tehát x és y ugyanabban a cellában van. Akkor pedig az előző tétel miatt megegyezik a faalakjuk.

Ha egy töröttvonal trópusi szakaszon veszek két szomszédos töréspontot, akkor azokra igaz, hogy a két vektor különbsége egy $0 - 1$ vektor megszorozva egy skalárral, vagyis két különböző értékből áll a különbségvektoruk és a trópusi szakaszuk egyenes, méghozzá az, amit "kiszámolunk" az algoritmus során. Ez mondjuk logikus, de azért jó kimondani. Így tehát az állításból következik a 2.3.4. Példa utáni megjegyzés, hogy a töröttvonal egyenes szakaszain megegyezik a fák alakja. $\square$

Azt is jeleztük, hogy olyan mintha a belső csúcsok csúszkálnának a fa szárazakon, amikor egy egyenes szakaszon mozgunk. Ez azért igaz, mert hiszen feltehetjük,

hogy a legnagyobb koordináta egy x és y vektor esetén ugyanaz, mert $U_n \subset \mathbb{R}^N/\mathbb{R}\mathbf{1}$ térben vagyunk, tudunk úgy csökkenteni a vektorokon, hogy a legnagyobb érték megegyezzen. Így ugyanolyan "mélységű" fákat kapunk. Ezeknek az alakja megegyezik, így nyilván belső csúcsok nem csúszhatnak egymásba vagy egymáson át. Amikor pl. x vektor y vektorrá alakul, akkor bizonyos koordináták változatlanok, a többi meg nő (hiszen λ is nő). Mivel közben a gyökér és a levelek távolsága nem változik (ezt feltehetjük az előbbi szerint), a belső csúcsoktól a levélutódjaik ugyanolyan távol vannak, így valóban bizonyos csúcsok kénytelenek elmozdulni valamilyen irányba. Egymásba csúszás nem lehet, hiszen pozitív távolságra vannak egymástól x fájánál, és λ -t lassan növelve, kicsit változnak csak a fa éleire írt értékek. Pont emiatt a finom változtatás miatt igaz az is, hogy a két végponttól (előző állításban ω_1 és ω_2) a szakaszon lévő fák nem sokban térnek el, csak annyiban, hogy bizonyos belső csúcsok összecúszhatnak és szétválhatnak, de körülbelül megegyeznek vele.

2.5.7. Állítás. ([11, Lemma 3.3.]) $P := tconv(v_1, \dots, v_m)$, ahol $v_b \in U_n$ minden $b = 1, \dots, m$ -re. P pontosan akkor tartalmazza az origót, ha bármely $i \neq j \in \{1, \dots, n\}$ levélpárra létezik olyan v_b , hogy a hozzá tartozó fában az i, j levelek közti út átmegy a gyökéren.

Bizonyítás. Ha minden i, j párhoz létezik a fent megadott tulajdonságot teljesítő v_b , akkor T_{v_b} fában i, j levelek távolsága a legnagyobb. Akiknek az útja nem megy át a gyökéren, azoknak a távolsága kisebb (v_b azon koordinátája kisebb). Legyen $a_b := -(i, j \text{ levelek távolsága } T_{v_b} \text{ fában})$, mert ekkor $\mathbf{a}_b + v_b$ olyan vektor, hogy azok a koordináták, amikhez tartozó levéltávolságok a legnagyobbak (átmegy az útjuk a gyökéren), azoknál 0 lesz, a többinél pedig egy negatív szám. Veszem az összes i, j levélpárt és elkészítem hozzájuk az előbbi a_b értéket. Ha két különböző levélpárhoz is ugyanaz a T_{v_b} fa tartozik, akkor is ugyanaz az a_b szám lesz hozzá rendelve, mert a gyökéren átmenő utak hossza fix egy fában. Ha marad olyan v_c vektor, amihez nem gyártottunk a_c értéket, akkor annál legyen egy nagyon negatív szám az a_c . Ekkor $z = \max((\mathbf{a}_1 + v_1), (\mathbf{a}_2 + v_2), \dots, (\mathbf{a}_m + v_m)) \in P$ -ben minden $z_{i,j}$ koordinátára 0-t kapunk. Hiszen z koordinátáinak i, j levélpárok felelnek meg és mindhez tartozik egy olyan a_b valós szám, aminél a hozzá tartozó $\mathbf{a}_b + v_b$ érték nulla, a többi pedig az a_c választások miatt negatív, így a maximum művelet miatt $z_{i,j} = 0$. Ekkor tehát $z = \mathbf{0}$, vagyis az origó tényleg P -ben van.

Visszafelé tegyük fel, hogy $\mathbf{0} \in P$, vagyis

$\mathbf{0} = \max((\mathbf{a}_1 + v_1), (\mathbf{a}_2 + v_2), \dots, (\mathbf{a}_m + v_m))$ valamilyen a_b valóságokra. Tehát minden koordinátára, vagyis minden i, j levélpárra $0 = a_b + v_{b_{i,j}}$ valamilyen b indexre, a többi indexre $a_c + v_{c_{i,j}} \leq 0$. Így ha másik levélpár koordinátáját nézem, ott hasonlóan $a_b + v_{b_{i,j}} \leq 0$ fog teljesülni. Tudjuk a 2.5.2 Megjegyzésből, hogy $a_b = \min_{k,l}((\mathbf{0} - v_b)_{k,l}) = -\max_{k,l}((v_b - \mathbf{0})_{k,l})$, így az előzőek szerint i, j levélpárra veszi fel a minimumot (átrendezés). Ebből $v_{b_{i,j}} = -a_b = \max_{k,l}((v_b)_{k,l})$, tehát az i, j levelek távolsága a legnagyobb v_b fájában. Ez pont azt jelenti, hogy i, j levelek útja az egyik v vektor fájában átmegy a gyökéren. De mivel ez bármely koordinátára, vagyis levélpárra hasonlóan elmondható, így az állítást fordított irányba is beláttuk. $\square$

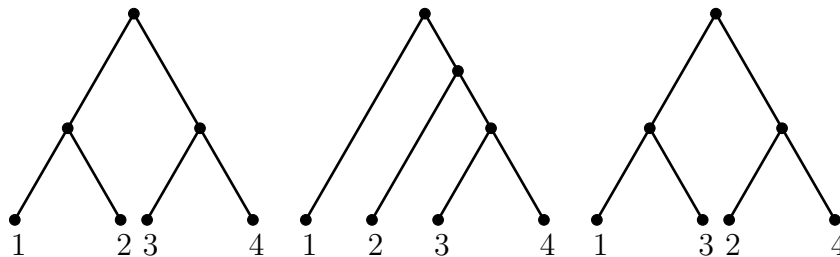
Az előbbi állításból következik, hogy x, y trópusi szakasza pontosan akkor tartalmazza az origót, ha a levélpárookra igaz, hogy az egyik végpont fájában a köztük menő út átmegy a gyökéren. Ez azt jelenti, hogy (mivel feltehetjük, hogy x és y fájában is a gyökér és a levelek távolsága ugyanaz a d szám, hiszen $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -ben vagyunk), így $x \boxplus y = \mathbf{d} = \mathbf{0}$. Így átfogalmazhatjuk trópusi szakaszra az előző állítást, hogy pontosan akkor tartalmazza x, y trópusi szakasza az origót, ha $x \boxplus y = \mathbf{0}$.

2.4.8. Tétel bizonyítása.

Bizonyítás. Legyen a trópusi szakasz két végpontja x és y és nézzük x fáját. $n \geq 5$ esetén a gyökér egyik gyerekének (feltettük, hogy két gyereke van) minimum 3 levél utódja lesz: i, j, k . Ezek között a levelek között nem vezet olyan út, ami a gyökéren átvezetne. Ekkor ha az origót tartalmazza a trópusi szakaszuk (ami a két pont trópusi konvex burka), akkor y -ban eközött a 3 levél között minden útnak át kell mennie a gyökéren. Skatulyaelv szerint azonban legalább kettőnek a gyökér azonos gyereke alatt kell lennie, viszont akkor nem teljesül a gyökéren áthaladás feltétel. Így valóban ha választunk két fát, amiknek a gyökere két gyerekes, akkor azoknak a trópusi szakasza nem mehet át az origón (0 valószínűségű). $\square$

2.5.8. Megjegyzés. $n = 3$ -ra már láttuk, hogy ha a két fa alakja eltér, akkor átmegy az origón, de az előző tételből is következik, hiszen ha eltér a fák alakja, akkor bármely levélpáros útja az egyik fában átmegy a gyökéren.

$n = 4$ esetén előfordulhat, hogy átmegy az origón, de az is hogy nem. Erre példa a 12. ábrán látható.



12. ábra. $n = 4$ esetén az első két fa trópusi szakasza nem megy át az origón, mert a 3, 4 levelekhez tartozó út egyik fában sem megy át a gyökéren. Ugyanígy a második és a harmadik trópusi szakasza se megy át a 2, 4 levélpár miatt. Viszont az első és a harmadik fa trópusi szakasza átmegy, mert látszik, hogy minden levélpárra a köztük menő út az egyik fában átmegy a gyökéren.

Az eddigiek alapján láttuk, hogy ha van két adatunk, akkor azoknak milyen a fájuk és hogyan változnak a trópusi szakaszukon a fák alakjai. Illetve azt is láttuk, hogy meg tudjuk kapni az evolúciós folyamatot is, ha az origóval vesszük

a trópusi szakaszát egy pontnak. Most még azt mutatjuk meg, hogy ha több mérési adatunk lenne, akkor létezik olyan pont, ami a konvex burkukban van és közel van a mérési adatokhoz, vagyis jó közelítése a valóságnak a mérések alapján.

2.5.9. Definíció. ([11, Definition 3.4.]) *Legyenek $v_1, \dots, v_m \in \mathbb{R}^N/\mathbb{R}\mathbf{1}$. x^* -t a $\{v_1, \dots, v_m\}$ halmaz Fermat-Weber pontjának nevezzük, ha minimalizálja a $\sum_{i=1}^m d_{tr}(x, v_i)$ összeget.*

Egy fontos megjegyzés, amit nem bizonyítunk, az előbbi összegnek tényleg létezik minimuma, vagyis valóban létezik olyan $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ -beli pont, amire az összeg minimális ([7]). Infimuma létezik, mert a metrikák és az összegük is nemnegatív, tehát alulról korlátos az összeg.

2.5.10. Állítás. ([11, Lemma 3.5.]) *Tegyük fel, hogy $m \geq 3$ és $\{v_1, \dots, v_m\} \subset U_n$. Ekkor létezik olyan x^* Fermat-Weber pontja a halmaznak, ami $tconv(v_1, \dots, v_m)$ -ben van. Mivel U_n trópusian konvex, így ez az x^* pont ultrametrikus fát ad.*

Bizonyítás. Tegyük föl, hogy van egy x_1^* Fermat-Weber pont, ami nem a konvex burokból van, különben készen vagyunk. Mivel nincs benne, ezért a 2.5.3. Definíció szerint van a Q halmazok között olyan, ami üres, legyen egy ilyen Q_j . Tehát bármely $1 \leq i \leq m$ -re $\{v_{i_1} - x_{1_1}^*, \dots, v_{i_N} - x_{1_N}^*\}$ halmazból nem a $v_{i_j} - x_{1_j}^*$ elem a maximális (vagyis nem $x_{1_j}^* - v_{i_j}$ a minimális, ha a negatívszorosaikat vesszük). Ez azt jelenti, hogy x_1^* j . koordinátáját tudjuk úgy csökkenteni, hogy közben $\sum_{i=1}^m d_{tr}(x, v_i)$ függvény értéke nem változik. Ez azért igaz, mert

$$\sum_{i=1}^m d_{tr}(x_1^*, v_i) = \sum_{i=1}^m (\max_k (x_{1_k}^* - v_{i_k}) - \min_l (x_{1_l}^* - v_{i_l})),$$

ahol a minimumos tagokban a minimumot nem a j indexnél veszi fel. A maximumot sem veheti fel, mert ha igen, akkor most csökkentve a j . indexet a pozitív előjelű maxos tagok csökkenének, a többi nem változna, tehát eleve ez a módosított vektor kisebb összeget adott volna, mint x_1^* . Így valóban nem befolyásolja a j index a távolságösszeget. Ha csökkentjük a j . indexet, akkor $v_{i_j} - x_{1_j}^*$ értéke nő minden i -re. Egészen addig csökkentjük a j . koordinátát, amíg el nem érjük valamelyik i -re a $\{v_{i_1} - x_{1_1}^*, \dots, v_{i_N} - x_{1_N}^*\}$ halmaz maximumát (ezek véges különbségek minden i -re, és véges sok halmaz van, létezik legkisebb halmaz-maximum, akit először elérünk). Ekkor az az i index bekerül a Q_j -be. Ilyen módon csökkenthetjük az üres Q halmazok számát és közben továbbra is Fermat-Weber pontok maradnak. Így amikor elérjük, hogy minden Q_j nemüres, akkor a 2.5.3. Definíció szerint az új x^* pont már a trópusi konvex burokból lesz, de az előzőek szerint Fermat-Weber pont is marad. Ezzel az állítást beláttuk. $\square$

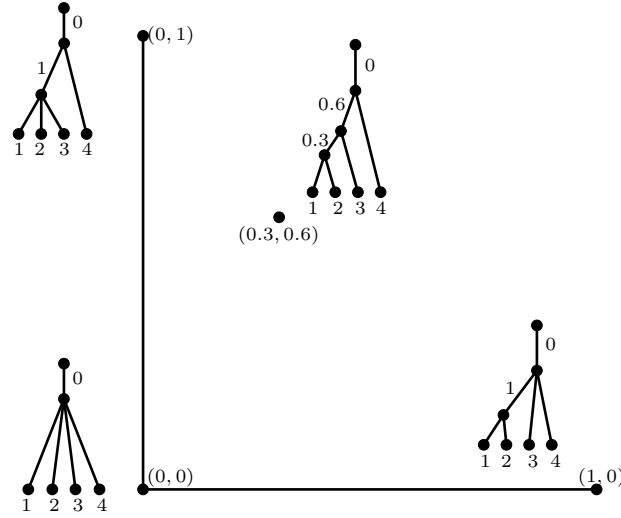
2.6. BHV tér, fa, metrika és a trópusi metrikával a kapcsolata

Az előzőekben láttuk hogy használható a trópusi geometria ultrametrikus adatok esetén a fák alakjának vizsgálatához. Ebben a fejezetben egy másik geometriát mutatunk be, aminél több probléma is fel szokott lépni, ezért elterjedtebb a trópusi. Ugyanakkor az egyik leghíresebb geometria és fa fajta a filogenetikai fák témakörében. Ez a szerzők nevére utaló *BHV* tér (*Louis J. Billera, Susan P. Holmes, Karen Vogtmann*, [1]). Ezt csak röviden tárgyaljuk: bemutattuk a filogenetikai fa szerkezetét, a metrikát, a geometriát és megállapítottuk egy kis kapcsolatot a trópusi metrikával.

2.6.1. Definíció. ([1]) *Filogenetikai fa alatt itt is egy gyökeres, címkézett n levelű, (többszörösen bináris) fát értünk, aminek élére pozitív súlyok vannak írva. A fát gyakran "fellógatjuk a gyökerénél fogva", ami azt jelenti, hogy egy nulla súlyú él a gyökér fölé teszünk (pl. 13. ábrán látható fák). Belső csúcs, él és külső él alatt ugyanazt értjük itt is, mint eddig. Az n levelű filogenetikai fák terét továbbra is T_n -nel jelöljük.*

A *BHV* tér egy pontját úgy kapjuk, hogy a belső élekre írt értékeket adjuk meg mint koordinátákat. Tehát rögzítünk egy elnevezést a belső éleken (egyes él, kettes él, stb.) és az alapján az $\mathbb{R}^E$ tér egy vektorának első koordinátája az egyes élre írt súly, stb. E a belső élek száma, ami 1.2.7. Megjegyzés miatt legfeljebb $n - 2$. Azt is láttuk, hogy bináris fáknál lesz $n - 2$ él, és ugyan itt nem feltétlenül binárisak a fák (belső csúcsok összeecsúsíthatnak), ha olyankor nullának vesszük a belső élre írt súlyt, akkor ugyanúgy egy $n - 2$ koordinátájú vektort kapunk. Egy ilyen $(e_1, e_2, \dots, e_{n-2})^T$ vektor $[0, \infty)^{n-2}$ egy pontja, ahol e_i az i . élre írt pozitív súly vagy azt jelzi, hogy a két végpontja az élnek egyben van. A külső élekkel azért nem kell foglalkozni, mert hasonlóan az $\mathbb{R}^N/\mathbb{R}\mathbf{1}$ térhez, a külső élek hossza nem befolyásolja a fa alakját, és csak nőne a tér dimenziója. A tér a következő képpen néz ki. Minden bináris faalakhoz rendelünk egy $[0, \infty)^{n-2}$ térsíkmetszetet, tehát egy ilyen térsíkmetszetben olyan fák vektorai vannak, amelyeknek az alakja megegyezik. Minden térsíkmetszeti ponthoz tartozik egy *BHV* fa, hiszen itt tudjuk, hogy milyen alakú fa tartozik a térsíkmetszethez, az élekre pedig rá tudjuk írni a súlyokat a vektor alapján. A térsíkmetszeten olyan fák vannak, amelyeknél bizonyos belső csúcsok összeecsúsíthatók, más szóval a vektorukban van nulla koordináta is (13. ábrán látható példa). Tehát ha n levelű bináris fákról van szó, akkor mivel $(2n - 3)!!$ féle faalak lehetséges, így $(2n - 3)!!$ db térsíkmetszet van "összeragasztva" a határaiknál (14. ábrán látható $n = 3$ -ra példa). Ezzel a *BHV* tér legelőnyösebb tulajdonságát el is mondhatjuk: míg a trópusi geometriánál több tétel is azzal foglalkozik, hogy mikor azonos két fa alakja, itt biztosan tudjuk, hogy egy térsíkmetszetben milyen a fák alakja. Itt is legrövidebb utat szeretnénk keresni két fa vektora között. Ha a két pont ugyanabban a térsíkmetszetben van, akkor mivel ezekre az $\mathbb{R}^{n-2}$ -beli térsíkmetszetre euklídeszi térsíkmetszengként tekintünk, a legrövidebb út köztük az euklídeszi egyenes szakasz és a két végpont (sőt az úton lévő pontok) fájának az alakja megegyezik. A nehézség akkor adódik, amikor két eltérő térsíkmetszetben

vannak a pontok. Nyilván egy megoldás, hogy elmegyünk az egyik pontból az origóba a közös térnegyedükben (az origó mindben benne van, hiszen az a csillagfa, aminek minden belső éle nulla), majd onnan a másik pont térnegyedében megyünk tovább. Ez nyilván egy út, könnyen számolható a hossza is, mivel csak két euklideszi szakaszt kell összeadni. Azonban nem biztos, hogy ez a legrövidebb út köztük, lehet, hogy a térnegyedek határán lévő fák pontjain kell átmenni egy másik térnegyedbe, majd több térnegyedet érintve juthatunk el a másik végpontba. A szakirodalomban a legrövidebb utat **geodesic**-nek nevezik (részletek: [1]). Létezik $O(n^4)$ idejű legrövidebb út keresés ([10]). A tér másik nehézsége, hogy véges darab pont konvex burka a *BHV* térben nem tudjuk mekkora, nincs a dimenziójára korlát ([6]).

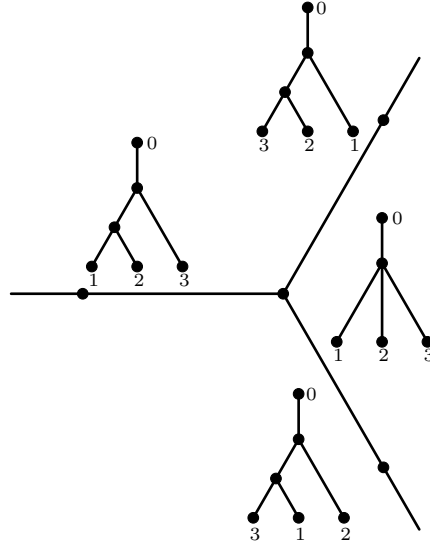


13. ábra. $n = 4$ esetén egy faalak térnegyede a *BHV* térben. A $(0.3, 0.6)$ -hoz tartozó fa alakja "állandó" a térnegyedben. Látszik, hogy a két szélén a 3-as levél fölötti belső csúcs csúszik az alsóba $((0, 1)$ pont), illetve a gyökérbe $((1, 0)$ pont). Itt a térnegyed $n - 2 = 2$ dimenziójú.

Az előzőek alapján megfogalmazhatjuk a *BHV* metrikát.

2.6.2. Definíció. ([1], [9]) Két $\mathbb{R}^{n-2}$ -beli pont között a *BHV* metrika szerinti legrövidebb út egy euklideszi töröttvonal, aminek egyenes szakaszai egy-egy faalakhoz tartozó $[0, \infty)^{n-2}$ térben vannak, a töréspontok pedig a térnegyedek határain. Két pontnak a *BHV* távolsága a szakaszok euklideszi hosszának az összege (ezt d_{BHV} -val jelöljük a d_{tr} trópusi metrikához hasonlóan). A geodesic az az út, amire a hossz minimális.

Ezek után megmutatjuk, hogy a trópusi és a *BHV* metrika között van kapcsolat, de nem ekvivalensek. Ugyan a *BHV* metrikában csak a belső élek hossza



14. ábra. $n = 3$ esetén a bináris faalakok száma $(2 \cdot 3 - 3)!! = 3!! = 3$, és valóban 3 db $n - 2 = 1$ dimenziós terünk van, 3 db $[0, \infty)$ félegyenes. Az ábrán láthatóak az egyes térszögekhez tartozó faalakok és itt határnak csak az origó számít. A "középső" belső csúcs olyankor a gyökérbe csúszik, mert lecsökken 0-ra a belső él értéke.

érdekes, a következő állításban úgy vesszük, hogy azok is beszámítanak a vektorba és a hosszba ([9] cikk szerint).

2.6.3. Állítás. ([9, Theorem 11.]) (**Stability**) Legyen T_1 és T_2 két n levelű filogenetikai fa és feltesszük, hogy a hozzájuk tartozó trópusi pontok ω_1 és ω_2 (nem feltétlenül ultrametrikusak, de $\mathbb{R}^n / \mathbb{R} \mathbf{1}$ -beliek), a BHV térbeli pontjuk pedig v_1 és v_2 . Ekkor $d_{tr}(\omega_1, \omega_2) \leq \sqrt{n + 2} \cdot d_{BHV}(v_1, v_2)$.

2.6.4. Megjegyzés. Igazából a Stability Tételben a gyök alatt $n + 1$ van, de nekünk elegendő itt az $n + 2$ is.

Bizonyítás. Először tegyük fel, hogy a két fa alakja megegyezik (most ha csak annyiban különböznek, hogy esetleg két-két szomszédos belső csúcs összecsatolt-szétvált, azt is azonosnak tekintjük). Ekkor a BHV térben egy térszögben vannak v_1 és v_2 pontok (megengedve, hogy esetleg a térszög határain vannak). Ekkor mivel a BHV és a trópusi metrikánál is koordináta különbségekből számolunk, így vonjuk ki a T_1 faéleinek súlyából a T_2 fát. Ha a belső élekre írt értékek különbsége $e_1, e_2, \dots, e_{n-2}$, a külső éleké meg $p_1, p_2, \dots, p_n$, akkor

$$d_{BHV}(v_1, v_2) = \sqrt{\sum_{i=1}^{n-2} e_i^2 + \sum_{j=1}^n p_j^2}.$$

Mivel a két fa alakja megegyezik, így vehetjük úgy, hogy egy olyan segédfát

készítettünk, aminek az alakja megegyezik T_1 és T_2 -ével, élein pedig a különbségek szerepelnek (esetleg negatívak). Ekkor $\omega_1 - \omega_2$ koordinátái a segédfabeli k, l levélpárok közti út hossza lesz (az út hossza a T_1 fában minusz T_2 -ben, hiszen a trópusi geometriában a távolságmátrixból kapjuk a vektorokat). Viszont ha vesszük $\omega_1 - \omega_2$ -ben a maximális koordináta különbséget és kivonjuk belőle a minimálisat (trópusi metrika), akkor az annak felel meg, hogy a segédfában az egyik levelek közti út hosszából kivonunk egy másikat, tehát bizonyos élek átfedésben lehetnek és ekkor az ellentétes előjelek miatt kiesik az él különbségsúlya. A többi élsúly-különbség legfeljebb egyszer szerepelhet, vagy a maximális k, l levélúton, vagy a minimálison, vagy egyiken sem. Így azonban egy jó felső becslést kapunk, ha az összes e_i különbségsúly abszolútértékét vesszük, és ha a maximumot és a minimumot a k, l és k', l' levélpároknál veszi fel, akkor még hozzáadjuk az előbbi felső becsléshez a $|p_k| + |p_l| + |p_{k'}| + |p_{l'}|$ összeget (esetleg átfedés lehet az indexekben, de akkor is jó a felső becslés). A kapott felsőbecslés tehát: $d_{tr}(\omega_1, \omega_2) \leq \sum_{i=1}^{n-2} |e_i| + |p_k| + |p_l| + |p_{k'}| + |p_{l'}|$

Most használjuk a *Cauchy-Schwartz-Bunyakovszkij egyenlőtlenséget*.

$$\left(\sum_{i=1}^{n-2} |e_i| + |p_k| + |p_l| + |p_{k'}| + |p_{l'}| \right)^2 \leq ((n-2+4) \cdot 1^2) \cdot \left(\sum_{i=1}^{n-2} |e_i|^2 + |p_k|^2 + |p_l|^2 + |p_{k'}|^2 + |p_{l'}|^2 \right),$$

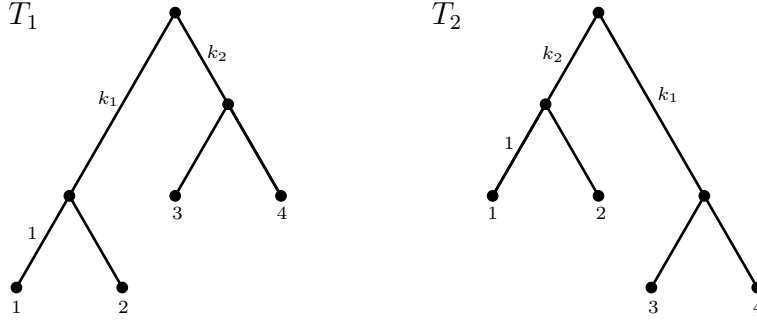
vagyis

$$\begin{aligned} d_{tr}(\omega_1, \omega_2) &\leq \sum_{i=1}^{n-2} |e_i| + |p_k| + |p_l| + |p_{k'}| + |p_{l'}| \leq \\ &\sqrt{n+2} \cdot \sqrt{\sum_{i=1}^{n-2} |e_i|^2 + |p_k|^2 + |p_l|^2 + |p_{k'}|^2 + |p_{l'}|^2} \leq \\ &\sqrt{n+2} \cdot \sqrt{\sum_{i=1}^{n-2} |e_i|^2 + \sum_{j=1}^n |p_j|^2} = \sqrt{n+2} \cdot d_{BHV}(v_1, v_2). \end{aligned}$$

Tegyük föl, hogy a két fa alakja nem egyforma. Ekkor v_1 és v_2 eltérő ténegyedben vannak és a legrövidebb út közöttük egy töröttvonal a BHV térben, ennek töréppontjai legyenek $u_0 = v_1, u_1, \dots, u_k = v_2$. Két szomszédos töréppont azonos ténegyedben van, így a falalakjuk megegyezik (nem szó szerint, hiszen eltérő határvonalon lehetnek, de a fenti számolásban megengedtük az ilyen "megegyezést" is). Mivel a d_{tr} is egy metrika, így teljesül a háromszög egyenlőtlenség. Így ha az u_i -hez tartozó fának a trópusi térbeli vektora ω'_i , akkor nyilván a $\omega'_0 = \omega_1, \omega'_1$, közti trópusi szakaszt, a ω'_1, ω'_2 trópusi szakaszt, ... és a $\omega'_{k-1}, \omega'_k = \omega_2$ trópusi szakaszt összeillesztve egy utat kapunk a trópusi térben ω_1 és ω_2 között, tehát $d_{tr}(\omega_1, \omega_2) \leq \sum_{i=0}^{k-1} d_{tr}(\omega'_i, \omega'_{i+1})$. A fentiek alapján azonban $d_{tr}(\omega_1, \omega_2) \leq \sum_{i=0}^{k-1} d_{tr}(\omega'_i, \omega'_{i+1}) \leq \sum_{i=0}^{k-1} \sqrt{n+2} \cdot d_{BHV}(u_i, u_{i+1}) = \sqrt{n+2} \cdot d_{BHV}(v_1, v_2)$, amivel az állítást beláttuk. $\square$

Mivel filogenetikai fák vizsgálatánál feltesszük, hogy a levelek száma (n) megegyezik, így $n+2$ konstansnak számít. Visszafele azonban nem igaz, más szóval nem létezik olyan $c > 0$ valós szám, hogy bármely n levelű T_1 és T_2 fára (az előző jelöléseket használva) $d_{BHV}(v_1, v_2) \leq c \cdot d_{tr}(\omega_1, \omega_2)$ teljesüljön. Erre egy példa.

2.6.5. Példa. Legyen $n = 4$ és a két fa alakja az alábbi. A T_1 -ben az 1, 2 levelek fölötti belső élnek a súlya k_1 , a 3, 4 levelek fölöttinek k_2 , a T_2 fában pont fordítva legyenek az élsúlyozások. A külső éleken mindenhol legyen 1. Ekkor $\omega_1 = (2, 2 + k_1 + k_2, 2 + k_1 + k_2, 2 + k_1 + k_2, 2 + k_1 + k_2, 2)^T = \omega_2$,



15. ábra.

míg $v_1 = (k_1, k_2)^T, v_2 = (k_2, k_1)^T$, tehát $d_{tr}(\omega_1, \omega_2) = 0$ és $d_{BHV}(v_1, v_2) = \sqrt{(k_1 - k_2)^2 + (k_2 - k_1)^2} = \sqrt{2} \cdot |k_1 - k_2|$, mert az előző állításban szereplő p_i értékek itt nullák. Mivel azonban k_1 -t tetszőlegesen nagyra választhatjuk, k_2 -t pedig akár nagyon kicsinek is, így nyilván nem létezik olyan $c > 0$, amivel a nullát megszorozva felülbecsülhetnénk a BHV távolságot.

Így a két metrika valóban nem ekvivalens. Viszont az látszik, hogy ha két fa a BHV metrika szerint elég közel van, akkor trópusi szerint is. Így ha két pontot vizsgálunk és BHV szerint közel vannak, akkor áttérhetünk a trópusira, mert ott is közel vannak, de jobban tudjuk vizsgálni a köztük vezető utat, valamint azt is tudjuk, hogy ha legalább 5 levelű fáink vannak, akkor nem megy át a trópusi szakaszuk az origón, csak speciális esetben.

3. Összefoglalás

Definiáltuk a filogenetikai, az ultrametrikus és a min-ultrametrikus fákat is. Láthattuk, hogy az ultrametrikus fák egy valódi részhalmazát alkotják a filogenetikai fáknak, valamint azt is megfigyelhettük, hogy miképpen tudnak evolúciós folyamatokat szemléltetni az ultrametrikus és min-ultrametrikus fák. Az 1. fejezetben két algoritmust is megadtunk ultrametrikus fa konstruálására, ha adva volt egy adathalmaz.

A 2. fejezetben már nem csak egy adott adathalmazzal és a hozzá tartozó ultrametrikus fával foglalkoztunk, hanem általánosságban vizsgáltuk a kapcsolatokat közöttük. Láttuk hogyan lehet trópusi szakaszt készíteni két fához, ami bizonyos szempontból egy legrövidebb út, mert a töröttvonal hossza éppen a két végpont trópusi metrika szerinti távolsága. Megfigyeltük, hogy ha azonos részfájuk (kládjuk) van a végpontoknak, akkor az végig megmarad a trópusi szakaszon, ami gyakorlatban hasznos lehet, mert ha több lehetséges fánk van és mind tartalmazza ugyanazt a részfát, akkor az a valóságnak megfelelő fában is benne lesz valószínűleg. Valamint az egyenes szakaszokon megmarad a faalak, de az nem feltétlenül igaz, hogy azonos alakú fának egyenes szakasz a trópusi szakasza. Az origó segítségével megkapható az evolúciós folyamat, vagyis hogy milyen sorrendben váltak szét az egyes utódok. Továbbá az origó azért is különleges, mert csak speciális esetekben lesz a trópusi szakaszon, amit könnyen ellenőrizhetünk is, míg a *BHV* térben ez nem ilyen egyszerű. Bár algoritmust nem láttunk a *Fermat-Weber* pont megtalálására, ha adva lenne egy pont, tudjuk hogyan lehet bevinni adott ultrametrikus fák trópusi konvex burkába. Így azokhoz a fákhoz tartozó vektoroknak egy trópusi konvex kombinációjaként megkapható az a fa, aminek a távolság összege a lehető legkisebb tőlük, így viszonylag közel van mindhez. Megemlítettük az eredeti irányt, a *BHV* teret is, és annak legszebb tulajdonságát, hogy ha adva van több adat és hozzájuk a fák, akkor ha azonos térnegyedbe esnek, akkor a faalakjuk megegyezik.

A dolgozat tehát választ adott a speciális tulajdonságú filogenetikai fák, az ultrametrikus fák konstruálására, illetve arra, hogyan és milyen tulajdonságokat lehet vizsgálni a trópusi geometria segítségével, valamint bemutatta a *BHV* tér alapjait is.

Hivatkozások

- [1] Louis J Billera, Susan P Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.
- [2] Peter Buneman. A note on the metric properties of trees. *J. Combin. Theory Ser. B*, 17(1):48–50, 1974.
- [3] Shelby Cox. Classifying tree topology changes along tropical line segments. *Algebraic Statistics*, 14(1):71–90, 2023.
- [4] Dan Gusfield. Algorithms on stings, trees, and sequences: Computer science and computational biology. *Acm Sigact News*, 28(4):41–60, 1997.
- [5] Bo Lin, Anthea Monod, and Ruriko Yoshida. Tropical geometric variation of phylogenetic tree shapes.
- [6] Bo Lin, Bernd Sturmfels, Xiaoxian Tang, and Ruriko Yoshida. Convexity in tree spaces. *SIAM Journal on Discrete Mathematics*, 31(3):2015–2038, 2017.
- [7] Bo Lin and Ruriko Yoshida. Tropical fermat–weber points. *SIAM Journal on Discrete Mathematics*, 32(2):1229–1245, 2018.
- [8] Diane Maclagan and Bernd Sturmfels. *Introduction to tropical geometry*, volume 161. American Mathematical Soc., 2015.
- [9] Anthea Monod, Bo Lin, Ruriko Yoshida, and Qiwen Kang. Tropical geometry of phylogenetic tree space: a statistical perspective. *arXiv preprint arXiv:1805.12400*, 2018.
- [10] Megan Owen and J Scott Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1):2–13, 2010.
- [11] Robert Page, Ruriko Yoshida, and Leon Zhang. Tropical principal component analysis on the space of phylogenetic trees. *Bioinformatics*, 36(17):4590–4598, 2020.
- [12] María Jesús de la Puente Muñoz. Distances on the tropical line determined by two points. 2014.
- [13] David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147, 1981.
- [14] Ernst Schröder. Vier combinatorische probleme. *Z. Math. Phys*, 15:361–376, 1870.
- [15] Ruriko Yoshida and Shelby Cox. Tree topologies along a tropical line segment. *Vietnam Journal of Mathematics*, 50(2):395–419, 2022.

**Mesterséges intelligencia alapú eszközök használatáról
szóló nyilatkozat**

Alulírott *Mohay Lili Veronika* nyilatkozom, hogy szakdolgozatom elkészítése során semmilyen MI alapú eszközt nem használtam.