



Eötvös Loránd Tudományegyetem
Informatikai Kar

Evolúciós fák és szekvenciaillesztések
Bayes-statisztikai Markov lánc
Monte Carlo mintavételezése

Diplomamunka

Készítette:
Novák Ádám
prog.terv. matematikus
szakos hallgató

Témavezető:
Dr. Miklós István
posztdoktori
ösztöndíjas, PhD.

Budapest
2006.

Tartalomjegyzék

Elfogadott témabejelentő	3
Bevezetés	5
1. Célkitűzés	8
1.1. Előzmények	8
1.1.1. Súlyozásos szekvenciaillesztő eljárások	9
1.1.2. Statisztikus szekvenciaillesztési módszerek	11
1.1.3. Markov-lánc Monte Carlo módszerek	14
1.2. Célok	15
2. Az eljárás felépítése	18
2.1. Markov-lánc Monte Carlo technika	18
2.1.1. Szubsztitúciós modell	20
2.1.2. Beszúrás-törlés modell	20
2.2. A tér komponensei	20
2.3. MCMC lépések	22
2.3.1. Illesztések újramintavételezése	22
2.3.2. Topológia változtatása	24
2.3.3. Evolúciós paraméterek változtatása	26
3. A módszer részletei	28
3.1. A lánc kezdőpontja	28
3.2. Részfa kiválasztása	28
3.2.1. Gyökércsúcs választása	29
3.2.2. Részfa építése	31
3.3. Ablakkivágás	36
4. Eredmények	38
Tartalmi összefoglaló	42
Köszönetnyilvánítás	43
Irodalomjegyzék	47

Elfogadott témabejelentő

x

Bevezetés

A természettudományok történetének egy új korszakát nyitotta meg a biológiai makromolekulák felfedezése. Az DNS szerkezetének 1953-as feltárása (Watson & Crick) óta óriási fejlődés indult meg a biokémiai módszerek területén. Ennek az előrehaladásnak ma már elvülhetetlen eredményei sok faj (köztük az ember) teljes genomjának feltérképezettsége, egyre hatalmasabb mennyiségű fehérje-térszerkezeti és funkcionális ismereteink. Az emberi mértékkel feldolgozhatatlan információhalmaznak jelentős része különböző nyilvános adatbázisokban hozzáférhető a tudományos kutatások számára.

Minden ma ismert élőlény felépítésének struktúráját és folyamatát nukleinsavak írják le. Ugyanakkor a változatos enzimatis tulajdonságokkal rendelkező fehérjéknek köszönhető a sejtek irányítása, fenntartása, osztódása. A biomakromolekulák alapvető szerepe ma már megkérdőjelezhetetlen. Nem meglepő ezért, hogy funkciójuk, működésük, evolúciójuk megértésére óriási az erőfeszítés. A nagy adatmennyiséggel megbirkózni azonban nem kis kihívást jelent: ennek leküzdésére irányuló mozgalom hozta létre a bioinformatika tudományát, amely mára önálló diszciplínává fejlődött.

A bioinformatika számos tudományterülettel szoros kapcsolatban áll. Biológiai kérdésekre keres választ, saját nyelvezete alakult ki. Matematikai, statisztikai, algoritmikai eszköztára egyre szélesebb. Néhány ága mélyebb kémiai ismeretekre is alapoz. Egyre növekvő számú tudományos folyóirat specializálódik e terület eredményeinek publikálására (Bioinformatics, Journal of Computational Biology, stb.).

A bioinformatika tehát elsősorban a biológiai makromolekulákkal foglalkozik. Hogy miért igényel ezeknek a vizsgálata speciális, új módszereket, miért adhatók pontos mate-

matikai modellek az evolúciójuk leírására, azt a nukleinsavak és fehérjék szekvenciális, hasonló ismétlődő egységekből felépülő szerkezete magyarázza.

A DNS-t globális szerkezetét tekintve két, egymáshoz kapcsoló szál kettős spirálja alkotja. A szálak egymáshoz hasonló építőegységekből, nukleotidokból épülnek fel. Egy nukleotid egység egy foszfátot, egy cukormolekulát (dezoxiribózt) és egy bázist tartalmaz, ez utóbbi négyféle lehet. A szomszédos nukleotidok egymáshoz kovalensen kötődő foszfát-cukor egységei adják a szálak vázát, a cukorhoz kötött bázisok sorozata információt kódol. A két szálon egymással szemben lévő bázisok komplementer párokat alakítanak ki: az adenin (A) a timinnel (T), a citozin (C) a guaninnal (G), egymáshoz H-kötések rögzítik őket. Az egyik szál tehát egyértelműen meghatározza a másikat is, ezen alapszik a DNS replikációja. A DNS információtartalma így leírható az egyik szál felépítő nukleotidokat szimbolizáló négy betű egy sorozatával, ezt nevezzük DNS-szekvenciának.

A nukleinsavak másik csoportjának tagjai, az RNS-ek, hasonló felépítésűek. Szinte minden esetben egy szálból állnak, amely nukleotidegységek kovalensen kötődő sorozata. A RNS nukleotidjaiban a cukormolekula a DNS-től eltérően ribóz, és a bázisok közül a timint (T) uracil (U) helyettesíti, amely az adeninnel a timinhez hasonlóan komplementer párt alkot. Globális térszerkezetüket tekintve az RNS-ek igen változatosak lehetnek, legtöbbször a lánc visszahajolva egy-egy szakaszon önmagával bázispárosodik. A másodlagos térszerkezetüket – néhány közel egyformán stabil konformációváltozattól eltekintve – egyértelműen határozza meg a szekvenciájuk a minimális energiájú állapot felé törekvés útján.

Az élőlények szervezetének „dolgozóit”, a fehérjék, szintén szekvenciális, egymáshoz hasonló – de mégis eltérő – építőegységek, aminosavak peptidkötéssel kialakított láncából álló makromolekulák. A természetben leggyakrabban előforduló húsz aminosav mindössze az α -szénatomhoz csatlakozó oldalláncban tér el egymástól. Jelölésükre az angol ábécé betűiből alakítottak ki szabványt. A fehérjék térszerkezete a lehető legváltozatosabb, elsősorban ennek köszönhető, hogy katalitikus aktivitásuk olyan szerteágazó, hogy lényegében bármilyen (energetikailag kedvező) szerves reakció végbemenetelét segítő fe-

hérje létezik vagy felépíthető. Igaz – akárcsak az RNS-ek esetében –, hogy a szekvenciájuk és térszerkezetük között közel egyértelmű megfeleltetés van, annak ellenére, hogy a fehérjék „hajtogatódását” chaperonok (dajkafehérjék) is segítik.

Mindezek ellenére a mai napig nem sikerült olyan hatékony eljárást adni, amely a szekvenciából elfogadható pontossággal megjósolja a fehérje térszerkezetét. A laboratóriumi meghatározás mellett leggyakrabban egyrészt homológiamodellezési módszerekkel dolgoznak, amely egy szekvencia már ismert térszerkezetét vetíti a vele rokon szekvencia egyes szakaszaira, a kapcsolatuk leíró szekvenciaillesztés alapján. Ez különösen akkor hasznos eszköz, ha rendelkezésre állnak az illesztések megbízhatóságát jellemző információk: az erősen konzerválódott, pontosan illeszthető régiók térszerkezete nagy valószínűséggel hasonló. Másrészt, biztató eredményeket érnek el óriási számítási kapacitást igénylő, pontos fizikai-kémiai mechanizmust szimuláló eljárásokkal, amelyek a világhálóra kötött és a projekthez csatlakozó számítógépek szabad processzoridejét hasznosítva hatalmas méretű osztott rendszerként elfogadható időn belül oldják meg a problémát.

Ez is mutatja, hogy a bioinformatika részterületeinek legtöbbször a nukleinsavak és fehérjék speciális, szekvenciális szerkezetén alapszik. A következő fejezetben a korábbi eredmények, fejlődő módszerek közül ismertetem a legfontosabbakat, amelyek témánkhöz kapcsolódódnak. Majd vázolom azokat a problémákat, amelyek miatt szükség van újabb, pontosabb eljárásokra, és meghatározom célkitűzéseinket. A további fejezetekben a kidolgozott módszerünk részleteit tárgyalom, végül annak implementálásával szerzett tapasztalatainkat, eddig elért eredményeinket írom le.

1. fejezet

Célkitűzés

1.1. Előzmények

A biológiai szekvenciák evolúciós kapcsolatának leírására hosszú ideje alkalmaznak szekvenciaillesztéseket. A szekvenciaillesztések alapja az a feltevés, hogy a szekvenciák megváltozásait okozó evolúciós történések három csoportba sorolhatók:

- szubsztitúció
- beszúrás
- törlés

A szubsztitúció a szekvenciát felépítő egységek (nukleotidok illetve aminosavak, ezek után összefoglaló néven karakterek) kicserélődését jelenti (pl. az ősi szekvencia egy nukleotidja más nukleotiddá alakult), a beszúródás új egységek megjelenése a fiatalabb szekvenciában, a törlődés az ősi szekvencia egy szakaszának eltűnése. A modell biokémiai háttérben a DNS molekulák replikációja során megkövetkező hibák állnak. Valójában ez meglehetősen leegyszerűsített képe a valóságnak, hiszen például a transzpozonok a génszakaszok ennél jóval bonyolultabb átrendeződését is eredményezhetik.

Két szekvencia illesztését a szekvenciák egymás alá írásával jelölhetjük úgy, hogy a beszúródott és törlődött karakterek helyén a másik szekvenciába egy „gap” jel (–) kerül. Általában az ősi szekvenciát írjuk fölülre.

$$\begin{array}{ccccccc} A & C & T & - & C & - & - & T \\ - & - & A & G & C & G & C & T \end{array}$$

1.1. ábra. Két szekvencia egy lehetséges illesztése.

Az 1.1. ábrán az ACTCT és AGCGCT szekvenciák egy illesztése látható. Az ősi szekvencia elején történt egy 2 karakter hosszúságú törlődés, majd egy pontmutáció (szubsztitúció), valamint két független beszúródás. A bioinformatika egyik központi problémája, hogy adott szekvenciák esetén keressük meg azt az illesztést, amely az egyik szekvencia másikba alakulásának optimális mutációsorozatának felel meg.

1.1.1. Súlyozásos szekvenciaillesztő eljárások

Hogy mit tekintünk optimálisnak, az természetesen nem magától értetődő. A legkorábbi szekvenciaillesztő algoritmusok az evolúciós események súlyozásán alapuló dinamikus programozási eljárások. Közös tulajdonságuk, hogy valamilyen előre rögzített súlyozás esetén a legkisebb összsúlyú mutációsorozatot keresik, a minimális evolúció elméletének megfelelően.

Sellers algoritmusában [26] a beszúráshoz és a törléshez azonos súlyt rendelt, az egyes szubsztitúciókhoz különbözőt (azok megfigyelt gyakoriságával fordítottan arányosan). Ezzel az egyszerűsítéssel $O(nm)$ idő alatt határozható meg a minimális súlyú illesztés, ahol n illetve m a két szekvencia hossza. Természetesen ez a módszer nem veszi figyelembe, hogy a természetben gyakran bekövetkező hosszú beszúródások sokkal gyakoribbak, mint az ugyanolyan összhosszúságú rövid beszúródások, hiszen az előbbi egyetlen evolúciós esemény. Ennek a problémának a kiküszöbölésére vezette be Waterman [33] a g_k gap-függvényt, amely egy k hosszúságú gap súlyát írja le. A gap a beszúrást és a törlést közös elnevezése (szokás terminológia az indel is), amelyek ebben a modellben azonos súlyt kapnak. Általános gap-függvény esetén $O(nm(n+m))$ időbe kerül az optimális illesztés kiszámítása. Gotoh [7] affin gap-függvények esetén $O(nm)$ -re javította a

szükséges futási időt. A g_k -t akkor nevezzük affinnak, ha

$$g_k = uk + v \quad (\forall k \geq 0), u \geq 0, v \geq 0 \quad (1.1)$$

A v a gap megnyitásának, az u a gap meghosszabbításának súlyát jelenti. Sajnos az affin gap-függvény biológiailag nem alátámasztható, csupán a gyors algoritmusnak köszönheti gyakori használatát. Miller és Myers [20] olyan algoritmust adott viszont konkáv gap-függvények esetén, amely $O(nm(\log n + \log m))$ idő alatt megadja az optimális illesztést. A g_k gap-függvény akkor konkáv, ha

$$g_{k+1} - g_k \leq g_k - g_{k-1} \quad (\forall k \geq 1) \quad (1.2)$$

vagyis ha a gap-ek bővítését egyre kisebb súllyal büntetjük. A gap-függvények e családjában található olyanok is, amelyek biológiailag helytállóak.

Természetes általánosítása a páronkénti szekvenciaillesztésnek a többszörös illesztés, amely kettőnél több szekvencia evolúciós kapcsolatát írja le. D. Sankoff [24] vetette fel először, azóta a bioinformatika egyik legközpontibb feladatává vált, ugyanakkor az egyik legnehezebb is. Wang és Jiang [32] óta tudjuk ugyanis, hogy az optimális többszörös szekvenciaillesztés megkeresése NP-teljes probléma (a szekvenciák számát tekintve).

```

A G A - - T C A
C - A - C T - A
- G A G C G C T

```

1.2. ábra. Többszörös illesztés 3 szekvenciával.

Az illesztést a szekvenciák egymás alá írásával jelöljük, a beszúródások és törlődések helyére gap jelet téve (1.2 ábra). Minden oszlop a homológ karakterek egy csoportját határozza meg. Nem tartalmaz viszont információt arra nézve, hogy a szekvenciák hogyan helyezkednek el egy evolúciós törzsfa mentén. Ez egy lényeges probléma: a többszörös illesztést kiszámító dinamikus programozási algoritmus (amely a páronkénti illesztés algoritmusának egyszerű általánosítása, n dimenziós dinamikus programozási táblázattal)

közvetlenül nem használható arra, hogy fajok leszármazási viszonyaira következtessünk, ráadásul (amennyiben $P \neq NP$) a vizsgált fajok számának növelésével exponenciálisan nő a szükséges számítási idő.

Ezeknek a problémáknak az elkerülése érdekében ma leggyakrabban *iteratív szekvenciaillesztést* [5, 28] alkalmaznak. A módszer lényege az, hogy páronkénti illesztésekből kiindulva egy adott törzsfá mentén azokból egy többszörös illesztés készíthető, a fa leveleitől a gyökér felé. Az eljárás egy lépése során a szekvenciák egy-egy részhalmazát tartalmazó többszörös illesztéseket úgy egyesítjük, hogy a két kiinduló illesztést nem bontjuk szét, azaz amely karakterek egy oszlopban szerepeltek, azok az egyesített illesztésben is egy oszlopban maradnak. Csupa gap-et tartalmazó oszlop viszont beszúrható bármelyik illesztés oszlopai közé. Ezzel a módszerrel előáll tehát egy többszörös szekvenciaillesztés, amely viszont felhasználható egy újabb, pontosabb evolúciós törzsfá készítésére. Ez lehetőséget ad egy jobb többszörös illesztés konstruálására, és így tovább. Szükség van azonban egy kezdő törzsfára. Ezt a célt szolgálja a *vezérfa*, amelyet általában a páronkénti távolságokból klaszterező algoritmussal állítják elő.

1.1.2. Statisztikus szekvenciaillesztési módszerek

Minden eddig bemutatott módszer közös hibája, hogy a páronkénti illesztések meghatározásához az evolúciós események egy szubjektív súlyozását használja fel. Nem adható meg egyértelműen olyan súlyok, amelyek biológiailag a legrelevánsabbak lennének, és a különböző súlyozások igen eltérő optimális illesztéseket adnak.

Ezeknek a gondoknak a kiküszöbölésére kezdtek el kidolgozni olyan módszereket, amelyek egy koherens statisztikai keretben adnak meg optimális illesztéseket. Az evolúciós események súlyozása helyett biológiai ismeretekből származó *a priori* eloszlásokat használnak fel, és az evolúciós paramétereket a megfigyelt szekvenciákból becsülik.

A statisztikus illesztésekhez azonban a szekvenciák evolúciójának statisztikus modelljeire van szükség. A terület fejlődésével fokozatosan elváltak egymástól a szubsztitúciós és beszúrás-törlés modellek, és azokat egymástól függetlenül javították. A statisztikus

szekvenciavizsgálatokhoz egy megfelelő szubsztitúciós és egy beszúrás–törlés modellt kell választani.

A szubsztitúciókat általában folytonos idejű Markov modellekkel írják le. Ez azt jelenti, hogy egy a karakter b -vel való szubsztitúciójának valószínűsége csak az a és b karaktertől valamint az eltelt időtől függ, az adott pozíción a előtt megjelenő karaktertől nem. Így a Markov modell a következő (vektoralakban megadott) lineáris differenciálegyenlettel jellemezhető:

$$\frac{\partial \mathbf{x}}{\partial t} = Q \mathbf{x} \quad (1.3)$$

ahol most $Q \in \mathbb{R}^{k \times k}$ a modellt leíró átmeneti valószínűségi mátrix, $\mathbf{x} \in \mathbb{R}^k$ az egyes karakterek valószínűségét tartalmazó vektor ($k = |\Omega|$ a különböző karakterek száma: nukleinsavszekvenciák esetén 4, fehérjeszekvenciáknál 20). Az egyenlet megoldása:

$$\mathbf{x}(t) = e^{Qt} \mathbf{x}(0) = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!} \mathbf{x}(0) \quad (1.4)$$

amely az $\mathbf{x}(0)$ által meghatározott karakter átalakulási valószínűségeit tartalmazza t idő alatt. Azaz, annak valószínűsége, hogy az i . karakter t idő alatt a j . karakterre cserélődik, $\mathbf{x}_j(t) = (e^{Qt} \mathbf{e}_i)_j$, ahol $\mathbf{e}_1 = (1, 0, \dots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \dots, 0)$, \dots , $\mathbf{e}_k = (0, \dots, 0, 1)$ a k . egységvektorok. A Q mátrix minden oszlopösszege 0, ez garantálja, hogy $\forall t : \sum_{i=1}^k \mathbf{x}_i(t) = 1$, azaz az \mathbf{x} minden t időpontban valóban valószínűségek vektora. Ezenkívül Q minden főátlón kívüli eleme nemnegatív. Q tulajdonságaiból következik, hogy legalább egy 0 sajátértéke van, és minden sajátértéke nemnegatív. Ha csak egy 0 sajátértéke van, akkor létezik egy globálisan stabil egyensúlyi állapot. Az egyensúlyi állapotban az i . karakter egyensúlyi valószínűségét π_i jelöli. A gyakorlatban különböző Q mátrixokat használnak. DNS szekvenciák esetében elterjedt Cantor modell, a SYM modell, Kimura 3 és 2 paraméteres modellje, a Jukes–Cantor modell és a Tamura–Nei modell. Fehérjeszekvenciáknál elterjedt a Poisson modell, a Hasegawa–Fujiwara modell és a Dayhoff mátrixok [1].

Felsenstein algoritmus [4] egy adott fa szubsztitúciós likelihoodját határozza meg

egy szubsztitúciós modell szerint. Az algoritmus leírása megtalálható például [10]-ban is. Eredetileg maximum likelihood keretben (numerikus optimalizálással) használták optimális törzsfák keresésére, ma inkább a beszúrás–törlés modellbe beágyazva, a fa belső csúcaiban nem megfigyelt szekvenciákban az egyes pozíciók, „Felsenstein wildcard”-ok Felsenstein likelihood-jának kiszámítására alkalmazzák. A belső csúcs egy nem megfigyelt pozíciójának az i . karakterhez tartozó Felsenstein likelihoodja a leveleken megfigyelt szekvenciák valószínűsége, feltéve, hogy az adott pozícióban az i . karakter áll. Ezek a likelihood-ok Felsenstein algoritmusával analóg módon határozhatók meg, a levelektől a gyökér felé haladva.

A beszúrás–törlés modellek is jelentős fejlődésen mentek át. Az első statisztikai alapú módszer Thorne, Kishino és Felsenstein nevéhez fűződik [30]. Az irodalomban nagy elterjedtsége miatt csak TKF91 néven hivatkoznak rá. A TKF91 modell reverzibilis, ami egyszeres beszúrásokat és törléseket enged meg. A szekvencia karakterei között *linkek* létét feltételezi, amelyek újabbak karaktereket és linkeket szülhetnek, és a szekvencia elején elhelyezkedő egyetlen halhatatlan link kivételével bármelyik link a hozzá tartozó karakterrel együtt meghalhat. A születési ráta a λ , a halálozási ráta a μ . A szekvenciahosszakra biológiailag nem teljesen megalapozható geometriai egyensúlyi eloszlást jósol.

Az egy évvel később ugyanazon szerzők által publikált javított modell már megengedi hosszabb fragmentumok beszúródását [31]. A számítások megkönnyítése érdekében viszont azzal a feltételezéssel él, hogy a beszúrt fragmentumok csakis egy lépésben törölődhetnek, azaz egy egységként kezelődnek. Az irodalomban erre a modellre a TKF92 rövidítéssel hivatkoznak. Egy plusz paramétert is tartalmaz a TKF91-hez képest, a fragmentumok geometriai eloszlását jellemző r paramétert. Mind ez a modell, mind a TKF91 esetében létezik olyan $O(nm)$ futási idejű dinamikus programozási algoritmus, ami két szekvencia modell szerinti beszúrás–törlés likelihoodját kiszámítja. Explicite megoldható ugyanis a linkek utódszámára felírt differenciálegyenlet.

Megadható olyan modell is, amely hosszú beszúrásokat és törléseket is enged, és a beszúrt darabok tetszőleges hosszúságú fragmentumok formájában törölődhetnek [19]. Saj-

nos erre a modellre a differenciálegyenletek nem oldhatók meg analitikusan, ezért csak numerikus módszerekkel számíthatók ki a szekvenciákra a beszúrás–törlés likelihoodok.

A TKF modellek felírhatók rejtett Markov folyamatként (Hidden Markov Model, HMM) [2, 9]. A rejtett Markov folyamat a nevét onnan kapta, hogy a folyamat maga nem megfigyelhető, csupán az állapotok emissziója. Szekvenciák illesztéséhez páros rejtett Markov-modelleket (pair-HMM) használhatunk. A HMM állapotainak tranzíciós valószínűségeit használva egy dinamikus programozási, ún. „forward” algoritmussal határozható meg az illesztett szekvenciák beszúrás–törlés likelihoodja. Több szekvencia evolúció is leírható többszörös HMM-mel (multiple HMM) [9], és ebben az esetben is kiszámítható a likelihood dinamikus programozással [16], de a számítási igény a szekvenciák számával exponenciálisan nő.

1.1.3. Markov-lánc Monte Carlo módszerek

Ilyen problémák hatására bukkant fel először a Markov-lánc Monte Carlo (Markov Chain Monte Carlo, MCMC) módszer [6], amelynek segítségével a szekvenciaillesztésekből mintát tudunk venni olyan módon, hogy minden illesztést a saját valószínűségének megfelelő arányban kapjuk, azaz az illesztések eloszlásából mintavételezünk. A Markov lánc valószínűségi változók sorozata, amelyben minden tag valószínűsége csakis a sorozat előző tagjától függ. Diszkrét pontokból álló tér esetén megadható a $P(y | x)$ valószínűség, amely azt a valószínűséget fejezi ki, hogy a sorozat valamelyik tagja y , feltéve, hogy az előző tagja x . Egy Markov lánc konvergál egy egyértelműen meghatározott $\pi(x)$ egyensúlyi eloszláshoz, ha ergodikus. A Metropolis-Hastings algoritmus célja egy előre adott $\pi(x)$ eloszláshoz való konvergencia. Ennek biztosításához mindössze arra van szükség, hogy úgy generáljuk a Markov-láncot, hogy egy x állapotból tetszőleges y felajánlott állapotba (proposal) a Metropolis-Hastings hányados által meghatározott valószínűséggel megyünk át (ha nem megyünk át y -ba, akkor a lánc következő eleme is x lesz). A

Metropolis-Hastings hányados:

$$\min \left\{ 1, \frac{T(x | y)\pi(y)}{T(y | x)\pi(x)} \right\} \quad (1.5)$$

ahol $T(x | y)$ annak valószínűsége, hogy y állapotból x -et ajánljuk fel (backproposal), $T(y | x)$ pedig fordítva (proposal). Ekkor ugyanis teljesül a *detailed balance* feltétel, és ha a Markov-lánc ergodikus, akkor a $\pi(x)$ eloszlásba konvergál. Jól látható, hogy a $\pi(x)$ egyensúlyi eloszlást elegendő egy konstans szorzó erejéig ismerni.

A Metropolis-Hastings hányados helyett olyan esetekben, amikor az új állapot egy véletlen ablak kivágásával kezdődik, majd csak az ablakon belül változtatjuk meg az állapotot, módosított proposal és backproposal valószínűséget kell használnunk, és az így kapott módosított Hastings hányados határozza meg az elfogadási valószínűséget. Konkrétan,

$$\min \left\{ 1, \frac{T(x, w | y)\pi(y)}{T(y, w | x)\pi(x)} \right\} \quad (1.6)$$

elfogadási valószínűség esetén a Markov-lánc a $\pi(x)$ eloszlásba konvergál, ha ergodikus, ahol pl. $T(y, w | x)$ a w ablak és az y állapot felajánlási valószínűsége. Ez a mintavételezés, amit „partial importance sampling” néven is illetnek, igen hatékony eszköz olyan esetekben, amikor a teljes állapot megváltoztatása túlságosan számításigényes a proposal és backproposal szempontjából, hiszen nem szükséges az összes w ablakra kiszámolni a proposal valószínűségeket. Ennek a mintavételezésnek a helyességi bizonyítása megtalálható [14]-ben.

1.2. Célok

Az előzmények részletes tárgyalása során fény derült többek között arra, hogy evolúciós törzsfák készítésére ma leggyakrabban alkalmazott algoritmusok és eljárások mindegyike egyetlen, vagy legfeljebb néhány helyesnek ítélt (esetleg többszörös) szekvencia-illesztésből állítja elő a megfelelő törzsfát. Ez több hátrányos következménnyel jár: egy-

részt a szuboptimális illesztések - amelyek lényeges, biológiailag releváns információkat hordoznak - figyelmen kívül maradnak [15]. A törzsfák konstruálásakor nincs lehetőség figyelembe venni, hogy egyes változékony régiókban az illesztés megbízhatatlan. A vezérfán alapuló iteratív algoritmusok által előállított törzsfák nem függetlenek a kiindulópontként használt vezérfától, így gyakran a pontatlan, távolságalapú fakészítő eljárásokkal épített vezérfa topológiája felé tendálnak [5, 29, 27]. Másrészt, a szekvenciaillesztéseket és a törzsfákat meghatározó eljárások tulajdonságaikból adódóan elkerülhetetlenül egymásra épülnek. Régóta ismert, hogy ideális esetben ezért a kettőt egyszerre érdemes becsülni [25].

A diplomamunka célkitűzése ezért egy olyan módszer kidolgozása, amelynek segítségével többszörös szekvenciaillesztések és evolúciós törzsfák együttesen vizsgálhatók egy Bayes statisztikai keretmunkában. Ez a megközelítés egyúttal lehetővé teszi a szekvenciaillesztésekre és az evolúciós törzsfákra adott predikciók megbízhatóságának mérését. A korábbi, hasonló elvekre épülő munkák [14, 22] egyszerűbb, biológiailag nem elég megbízható evolúciós modelleken alapulnak. Ezért célul tűztük ki a TKF92 modell egy továbbfejlesztésének beépítését a módszerünkbe. Az eljárás Markov-lánc Monte Carlo (MCMC) típusú, az illesztések és fák kívánt eloszlásához konvergál, és az eloszlásból autokorrelált mintákat vesz. A random séta evolúciós fák és szekvenciaillesztések közös, magas dimenziójú, bonyolult, nem-euklideszi térben történik. A térben definiált Bayes eloszlást egy szekvenciák evolúcióját leíró időfolytonos Markov modell mint likelihood függvény és biológiai ismeretekből származó prior eloszlások adják meg. Az eljárás kimenetét a mintavételezett evolúciós fák és szekvenciaillesztések alkotják. Ezen minták poszterior analízisére már meglévő eljárások és programok használhatók.

A módszerünket protein szekvenciák másodlagos térszerkezet-predikcióján teszteltük. Egy ismert szekvencia másodlagos térszerkezeti elemeit vetítettük az elemzésben részt vevő többi szekvenciára homológiamodellezéssel. A kapott térszerkezet és az adott szekvencia (ismert) valódi térszerkezetének összevetésével a módszerünk helyességét mérni tudjuk.

A bemenő adatok a szekvenciák, az egyik szekvencia térszerkezete, az MCMC paraméterei (milyen hosszú legyen a bemelegítés (burn-in) fázis, hány mintát veszünk a Markov láncból, mekkora a lépésszám két minta között), illetve opcionálisan a Markov lánc kezdő pontja.

A kidolgozott eljárásunk teszteléséhez egyik elsődleges célunk volt a módszert programként való implementálása. Az elkészült program C++ nyelvű. Linux/Unix valamint Windows 2000/XP operációs rendszerek alatt teszteltük, egyelőre kizárólag parancssori utasításokkal vezérelhető. Távolabbi célként szerepel egy grafikus felülettel való kiegészítése, amelyen könnyen megadhatóak a bemeneti adatok, valamint nyomon lehet követni az eljárás aktuális állapotát olyan, a random sétát jellemző statisztikákkal, mint például a log-likelihood nyom és az autokorreláció.

2. fejezet

Az eljárás felépítése

2.1. Markov-lánc Monte Carlo technika

Többszörös szekvenciaillesztés és evolúciós törzsfakészítés egy közös statisztikai keretben kizárólag Markov-lánc Monte Carlo technikával végezhető hatékonyan. Maximum likelihood keretben szükség lenne vagy az illesztések, vagy a topológiák mint paraméterek becslésére, amely összességében túl sok változó a (nemlineáris) numerikus optimalizálás számára.

Ezzel szemben, az MCMC módszer lehetővé teszi, hogy a magas dimenziójú terünk tartalmazza a szekvenciák illesztését, a törzsfa topológiáját és az evolúciós paramétereket is, és a random séta során e komponensek mindegyikét változtassuk. Így lehetőség nyílik a törzsfák és illesztések közös poszterior eloszlásának mintavételezésére.

2.1. Jelölések. Legyen D a bemenő, megfigyelt adatok halmaza („*Data*”), amely n homológ szekvenciából áll. Célunk a szekvenciák evolúciós kapcsolatát feltárni valamint az őket generáló evolúciós események paramétereit becsülni. Az előbbi két komponensből áll: a szekvenciaillesztések A halmazából („*Alignment*”) és az evolúciós törzsfák lehetséges topológiáinak T halmazából („*Topology*”). Az evolúciós paraméterek tere a Θ .

Ahogy a célkitűzéseink között is kifejtettük, eljárásunkban beszúrás–törlés modellként a TKF92 modell egy olyan javítását használjuk, amely a hosszú beszúrás modell [19]

viselkedését közelíti, és amelyhez mégis explicit átmeneti valószínűségek adhatók meg. Ezzel reményeink szerint biológiailag sokkal relevánsabb eredményeket kapunk, mint ami a korábbi, törzsfákat és illesztéseket egyidejűleg mintavételező MCMC-eljárások egyszerűsített evolúciós modelljeivel [14, 22] elérhető.

Ugyanakkor a pontosabb modellnek a számításigénybeli hátrányai észrevehetőek. A TKF92 modell esetén nem adható ugyanis olyan gyors eljárás, amely kiszámítaná a megfigyelt szekvenciák és azok homológiastruktúrájának likelihoodját adott törzsfa és evolúciós paraméterek esetén (a törzsfa belső csúcaiban a hiányzó, meg nem figyelt adatokat, elsősorban illesztéseket marginalizálva), mint ahogy az a TKF91 modell esetén megtehető [14]. Ezért nem kerülhető el, hogy módszerünkben a törzsfa belső csúcsain a hiányzó adatokat tároljuk.

Nincs azonban szükség minden belső csúcsban az egyes szekvenciák konkrét karaktereinek tárolására, mert a belső csúcsok összes lehetséges karakterére a likelihood összegezhető Felsenstein algoritmusával, a szubsztitúciós modellünk alapján. Így a belső csúcsokban elég az illesztéseket tárolni.

2.2. Jelölés. A törzsfa összes élén megadott illesztések halmazát jelölje A^* .

Az evolúciós modellünk figyelembevételével az evolúciós paraméterek halmaza a következőkből áll:

$$\Theta = \{\lambda, \mu, r\} \cup \tau \quad (2.1)$$

ahol λ , μ és r a TKF92 modellben a születési, halálozási ráta ill. a fragmentumhosszak geometriai eloszlásának paramétere; τ pedig a törzsfa összes élének hosszát tartalmazza.

Jelöléseinkkel most már precízen leírhatjuk, milyen Bayes-eloszlással adható meg a poszterior eloszlásunk, amelyből mintavételezünk, és abban az egyes likelihoodok hogyan számolhatók. Amire elsősorban kíváncsiak vagyunk, az a $P(A, T, \Theta \mid D)$ eloszlás. A poszterior eloszlás, amelyből az eljárás során mintavételezni tudunk, a hiányzó adatokkal (illesztésekkel) kibővített $P(A^*, T, \Theta \mid D)$ eloszlás, amelyet egy normalizációs konstans erejéig a következőképp tudunk kiszámítani:

$$P(A^*, T, \Theta | D) = \frac{P(A^*, D | T, \Theta) \times P(T) \times P(\Theta)}{z} \quad (2.2)$$

A $P(A^*, D | T, \Theta)$ az evolúciós modellből származó likelihood. A $P(T)$ és a $P(\Theta)$ prior valószínűségek, amelyeket biológiai előismeretek alapján adunk meg. Módszerünkben minden evolúciós paraméterre neminformatív exponenciális eloszlást feltételezünk, valamint a fatopológiákat egyenletes eloszlás jellemzi *a priori*.

2.1.1. Szubsztitúciós modell

Eljárásunk szubsztitúciós modellje Dayhoff mátrixon alapul [1]. Az átmeneti valószínűségek gyors kiszámítását a mátrix diagonizált alakja teszi lehetővé.

$$D = V \Lambda W \quad (VW = I) \quad (2.3)$$

és így

$$e^{Dt} = V e^{\Lambda t} W \quad (2.4)$$

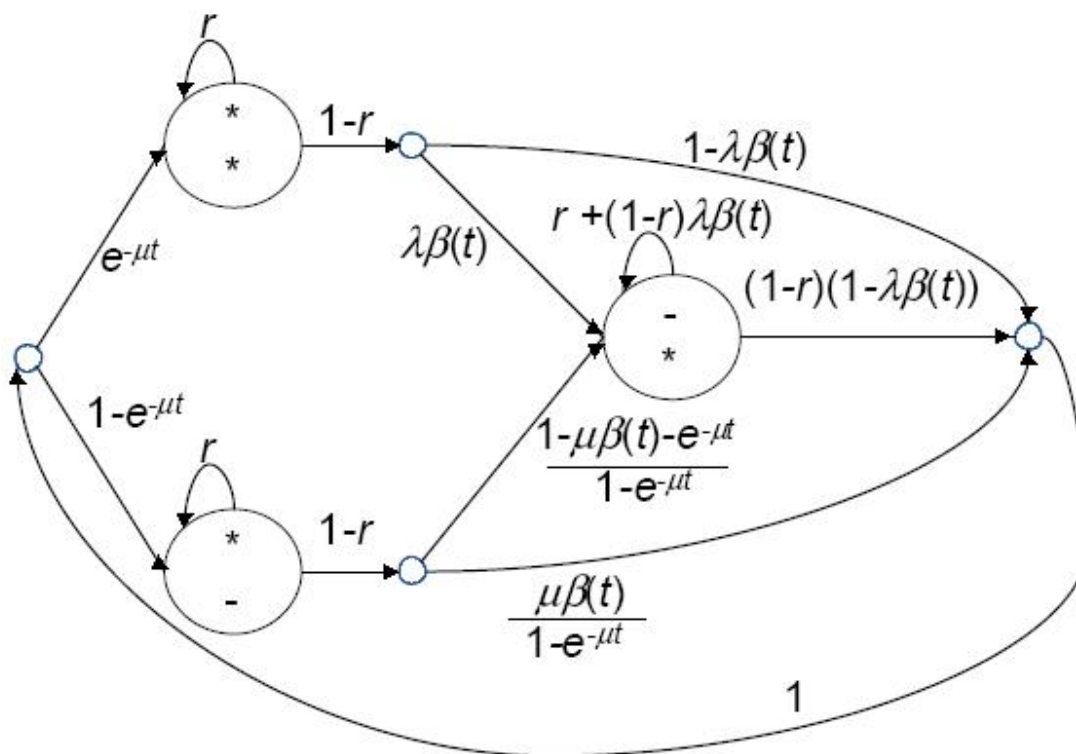
2.1.2. Beszúrás–törlés modell

Módszerünkben a TKF92 modell egy javításának HMM-ként való felírását használjuk a beszúrások és törlések modellezésére. A Markov-modell az átmeneti valószínűségekkel a 2.1 ábrán látható.

2.2. A tér komponensei

A random séta nem-euklideszi tere a következő komponensekből épül fel:

- i.) Az evolúciós törzsfá mint gyökereztetett bináris fa topológiája
- ii.) A törzsfá minden élén egy szekvenciaillesztés valamint a belső csúcsokban a nem megfigyelt szekvencia hossza (ami nem állandó)
- iii.) Valós értékű, nemnegatív evolúciós paraméterek



2.1. ábra. A beszúrás–törlés likelihood számításához használt, a TKF92 modellen alapuló HMM.

A fa topológiáját egyszerű adatszerkezet írja le. Az egyes csúcsokat egy-egy struktúra reprezentálja, amely tartalmaz mutatót a csúcs szülőjére és mindkét gyerekére. A fa élein definiált, szülő és gyerek közötti szekvenciaillesztést a gyerekekhez tartozó struktúrában egy tömb tartalmazza, amelynek hossza a gyerek szekvencia hosszával egyezik meg. A modern szekvencia minden egyes karakterére a tömb értéke megmutatja, hogy

- a karakter homológ-e az ősi szekvencia valamelyik karakterével (pozitív, ha igen, negatív, ha nem)
- az ősi szekvencia hányadik karakterével homológ, vagy
- az ősi szekvencia melyik karaktere előtt történt a modern szekvenciában a beszúrás, a tömbérték ennek az indexnek a negáltja

Ez az adatszerkezet helyesen és egyértelműen írja le a két szekvencia illesztését (2.2 ábra). Az egyértelműség rögtön adódik a definícióból. A szubsztitúciókat és a beszúrásokat biztosan helyesen írja le. Az, hogy az ősi szekvenciából való törléseket is megadja, abból a könnyen bizonyítható megállapításból következik, hogy a modern szekvencia i .

és $i+1$. pozíciója közti törlések száma $|I_{i+1}| - (|I_i| + \text{step}(I_i))$, ahol I_i az illesztési tömb i . értéke, $\text{step}(x) = \chi_{\{x \geq 0\}}(x)$.

ősi szekv. indexei	1	2	3	4	5
ősi szekvencia	A	C	T	-	C - - T
modern szekvencia	-	-	A	G	- G C T
illesztési tömb			3	-4	-5 -5 5

2.2. ábra. Két szekvencia illesztését leíró adatszerkezet.

Nem tartozik a térhez, de hatékonysági okokból érdemes a belső csúcsokban a szekvencia minden egyes karakterére a Felsenstein likelihoodot eltárolni.

2.3. MCMC lépések

Az MCMC eljárás egy iterációja során a tér egyik komponensét változtatjuk meg, a komponens véletlen kiválasztásával. Precízebben, adott valószínűséggel megváltoztatjuk

- az illesztések
- a topológia
- az evolúciós paraméterek valamelyikét.

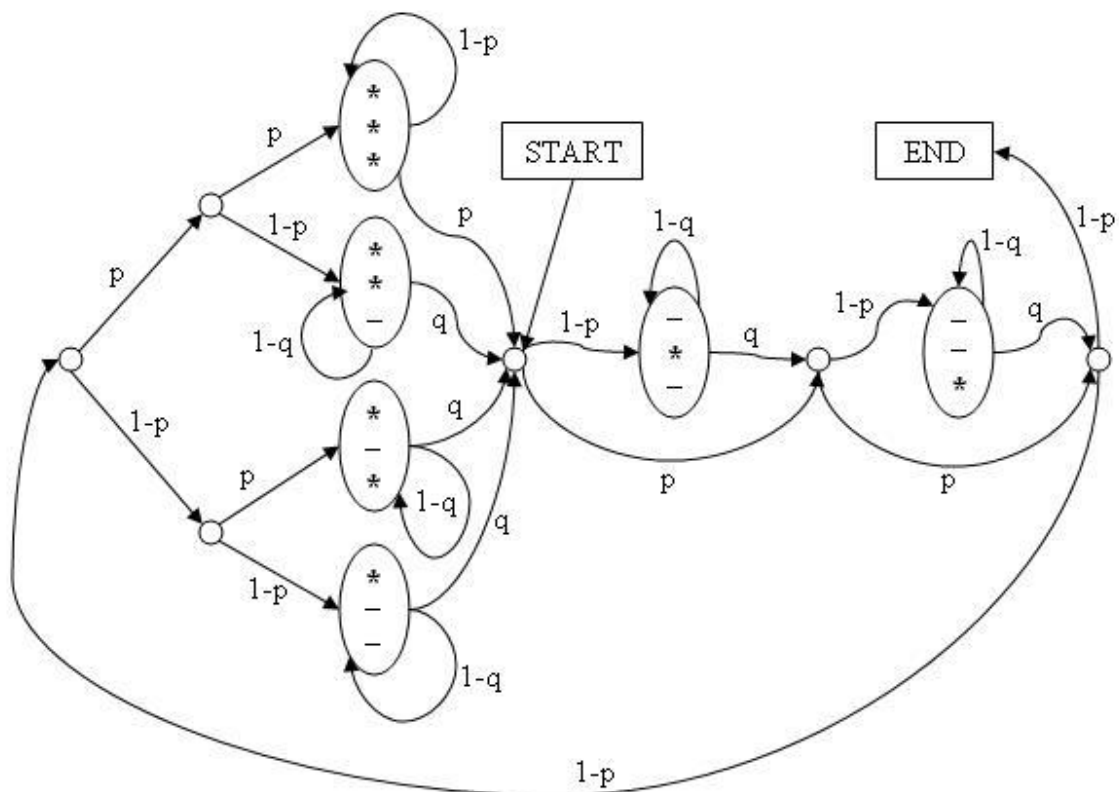
A topológiák lassúbb keveredése miatt a topológiaváltoztatást nagyobb valószínűséggel hajtunk végre. Az értékek megfelelő megválasztását próbálgatással lehet elérni, természetesen a rossz arányok a poszterior eloszlást nem befolyásolják, csupán a keveredésre, a konvergencia sebességére vannak hatással.

2.3.1. Illesztések újramintavételezése

Az összes illesztés egyidejű újramintavételezése nem a legcélravezetőbb megoldás. Ennek oka az, hogy az állapot térben túl nagy lépéseket téve az elfogadási valószínűség óhatatlanul is túl kicsi lesz, ami a keveredés romlásához és az autokorreláció növekedéséhez vezet. Érdemesebb ezért az illesztéseknek csak egy kis részét megváltoztatni. Ezt kétféleképpen is el lehet érni: (1) az illesztéseket csak egy kiválasztott részében található

szekvenciák között mintavételezzük újra, ill. (2) a szekvenciaillesztéseket csak az illesztés egy véletlenül kiválasztott „ablakába” eső részén változtatjuk meg.

Módszerünkben mindkét lehetőséggel élünk. Első lépésként a törzsfa egy részfáját rögzítjük, majd a részfa gyökerében található szekvencián egy „ablakot” jelölünk ki (az ablak első és utolsó karakterének indexével). Ezt az ablakot a fában lefelé haladva kiterjesztjük a részfa összes szekvenciájára, a szekvenciaillesztések segítségével (a gyerek szekvencia egy adott karaktere akkor esik az ablakba, ha az illesztésben ahhoz a karakterhez a szülői szekvenciában tartozó karakter vagy gap jel a szülői ablakba esik).



2.3. ábra. 3-szekvenciás rejtett Markov modell, p és q rögzített, 1-hez közeli valószínűsűségek.

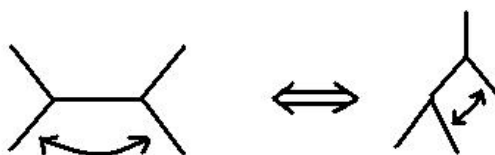
Az illesztések újramintavételezésére az ablakon belül ezek után kerül sor. A részfa belső élein lévő illesztések változtatásakor 3-szekvenciás rejtett Markov-modellt (3-HMM) használunk. Meghatározzuk a két megfigyelt modern szekvencia minden prefixpárjára annak a valószínűségét („Forward” algoritmussal), hogy a HMM azt a két szekvenciaprefixet bocsátotta ki. A két modern szekvencia illesztéseit ezután sztochasztikus vissza-

lépéssel mintavételezzük, az ősi szekvenciát és annak a modern szekvenciákhoz illesztését a HMM kibocsátása adja. Így megváltozhat az ősi szekvenciának az ablakon belülről eső karaktereinek a száma, viszont a HMM kibocsátási valószínűségeinek kiszámításához elegendő a négyzetes dinamikus programozási táblázat (pontosabban, a táblázat mérete a két modern szekvencia hosszának szorzata). A 3-HMM lehet olyan háromszekvenciás „transducer”, amivel a TKF modelleket pontosságát közelítő illesztés kaphatók [8]. Eljárásunkban mégis inkább a 2.3 ábrán látható, evolúciós paramétereiktől független, konstans átmeneti valószínűségeket használó 3-HMM-t használtuk, amelynek előnye a gyors kiszámíthatósága. A keveredést valamelyest rontja, de eredményeink arra mutatnak, hogy a futási időn nyerve az effektív mintavételezési sebességet növeli.

A részfa gyökércsúcsának szekvenciája a 3-HMM használata miatt megváltozik (a karakterek száma változik). Emiatt a részfagyökér és annak szülőcsúcsa közti illesztést is újra el kell készíteni, amennyiben a részfa gyökércsúcsa nem maga a gyökércsúcs. Ehhez az ablakot ki kell terjeszteni a szülőcsúcsra, ami pontosan úgy tehető meg, mint a levelek felé haladó kiterjesztés. Az illesztés ablakon belüli részének újramintavételezését ezek után kétszekvenciás, azaz páros rejtett Markov-moddal (pair-HMM) végezzük.

2.3.2. Topológia változtatása

A topológia változtatása „legközelebbi szomszédok cseréje” (Nearest Neighbour Interchange = NNI) lépések sorozatán keresztül valósul meg. Bebizonyítható, hogy egy gyökerezetlen fában két szomszédos csúcs egy-egy másik szomszédjának kicserélésével mint alaplépéssel bármelyik topológia átalakítható bármelyik másikká, az alaplépések egy sorozatával.



2.4. ábra. Egy NNI lépés gyökerezetlen és gyökereztetett fán.

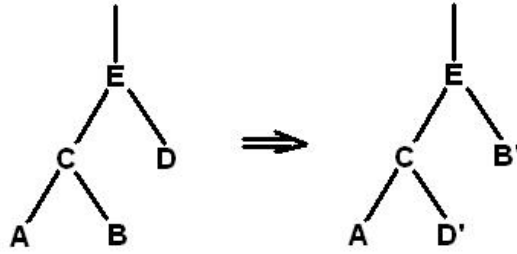
Ahogy a 2.4 ábrán is látható, az NNI lépés gyökereztetett fában egy csúcson és annak „nagybátyjának” (azaz a csúcs szülője testvérének) megcserélését jelenti. A leggyökereztetés helyétől eltekintve ezzel a lépéssorozattal bármilyen topológia megkapható. Nemreverzibilis modell esetén a fa gyökerének áthelyezése („átbillentés”) is szükséges lépés lehet, amivel tetszőleges gyökereztetett fa topológia is elérhető. A módszerünkben használt modellek mindegyikének reverzibilitása miatt ennek most nincs jelentősége.

Eljárásunkban tehát véletlenszerűen választunk egy gyökértől és annak közvetlen lezármazottaitól különböző csúcst, és azt a nagybátyjával megcseréljük. Van egy lényeges probléma viszont ezzel a lépéssel: mivel a csúcs szülője megváltozik, a kettő közötti szekvenciaillesztés érvénytelenné válik. Ugyanez a helyzet a nagybátya csúccsal és annak szülőjével is. Szükséges ezért a szekvenciaillesztések újrafelépítése. Négy lehetséges módszer jön szóba (a 2.5 ábra jelölései alapján):

- 1.) C és E szekvenciák változatlanok maradnak, C–D’ és E–B’ szekvenciaillesztéseket újramintavételezzük (páros HMM)
- 2.) C szekvenciát nem hagyjuk változatlanul: az C–A–D’ illesztést és C szekvenciáját 3-HMM segítségével mintavételezzük újra (C’-t kapunk); de E-t fixen hagyjuk: E–C’ és E–B’ páros illesztéseket készítjük el
- 3.) C és E szekvenciáját is megváltoztatjuk a C–A–D’ és E–C’–B’ háromszekvenciás újramintavételezésekkel; végül ha E nem gyökércsúcs, akkor E’-t hozzáillesztjük páros HMM-mel a szülőjéhez
- 4.) Az A, B, C, D, E szekvenciákban megkeressük azokat a homológ pozíciókat, amelyeknél egyik szekvenciában sincs gap. Ezek között a blokkok között teljes szekvencia- és illesztés-újramintavételezést végzünk.

A Felsenstein likelihoodokat minden esetben újra kell számítani a C csúcstól egészen a fa gyökeréig.

A különböző lehetőségek közül nehéz objektíven választani. Ugyanis egyrészt érdemes minél kevesebbet változtatni az eredeti állapoton a topológián kívül, hogy ezzel kis



2.5. ábra. Egy NNI lépés eredménye gyökereztetett fán.

lépéseket tegyünk az állapottérben, másrészt viszont vigyázni kell arra, hogy a topológia megváltoztatása mellett az eredeti illesztésekhez közeli állapot likelihoodja lehet nagyon kicsi, ami az elfogadási valószínűséget és ezáltal a keveredést csökkenti.

Implementációnkban a 3. változat szerint végeztük a topológiaváltoztatásokat.

2.3.3. Evolúciós paraméterek változtatása

Az evolúciós paraméterek mindegyike nemnegatív valós szám. Ugyanez a fa élhosszaira is igaz, ezért annak változtatását szintén e részben leírtak szerint végezzük.

2.3. Jelölések. Legyen x az evolúciós paraméter eredeti értéke. Rögzítsünk egy s konstans, amely a változtatás maximális mértékét jelenti.

Válasszunk egy r véletlen számot egyenletesen a következő intervallumból:

$$r \in [0, s + \min\{s, x\}] \quad (2.5)$$

és ezzel az r -rel válasszuk meg az evolúciós paraméter új értékét:

$$x' := x + r - \min\{s, x\} \quad (2.6)$$

2.4. Állítás. Az x' ilyen választásával $x' \geq 0$ és $|x - x'| \leq s$.

Bizonyítás. (2.5) és (2.6) alapján:

$$x' \in [-\min\{s, x\}, s] = [\max\{x - s, 0\}, x + s]$$

amiből mindkét állítás azonnal adódik. \square

Ez tehát egy megfelelő módja a paraméterek változtatásának, és lehetővé teszi tetszőlegesen nagy értékek felvételét is (kis valószínűséggel). A jó keveredés érdekében s -et az evolúciós paraméterek átlagos értékénél legalább egy nagyságrenddel kisebbre kell választani.

A Metropolis-Hastings hányados számításához tudnunk kell még a proposal és backproposal valószínűségeket. Könnyen belátható módon

$$T(x' | x) = \frac{1}{s + \min\{s, x\}} \quad (2.7)$$

és így a backproposal:

$$T(x | x') = \frac{1}{s + \min\{s, x'\}}. \quad (2.8)$$

3. fejezet

A módszer részletei

3.1. A lánc kezdőpontja

Az elfogadási valószínűség használata magában is garantálja az ergodikus Markov-lánc konvergenciáját az általunk vizsgálni kívánt poszterior eloszlásba. A lánc kezdőpontjának megválasztása ezért csak a burn-in fázis hossza szempontjából lehet érdekes.

Teljesen megfelelő az, ha az első szekvenciaillesztéseket súlyozás alapú illesztőalgoritmussal készítjük el, a kezdő törzsfát pedig a szekvenciák távolságát felhasználó klaszterező algoritmussal.

Implementációnk Gotoh dinamikus programozási algoritmusát [7] alkalmazza a szekvenciatávolságok meghatározására. A távolságokból a Szomszédok egyesítése algoritmussal készítjük el a törzsfát, amelynek élein a szekvenciaillesztéseket a 2.3 ábrán látható 3-HMM-mel mintavételezzük („Forward” algoritmus, majd sztochasztikus visszalépés). Így a belső csúcsokon lévő szekvenciák is elkészülnek.

Az evolúciós paraméterek kezdőértéke tetszőleges választható.

3.2. Részfa kiválasztása

A szekvenciaillesztések újramintavételezésének első lépése annak a részfának a véletlen kiválasztása, amely mentén új illesztéseket (és új belső szekvenciákat) készítünk.

A részfa választása eljárásunkban két fő lépésből áll:

- 1.) részfa gyökércsúcsának rögzítése
- 2.) részfa véletlen építése

Az 1. lépésben minden csúcshoz valamilyen súlyt rendelve a súlyokkal arányos valószínűséggel választjuk a részfa leendő gyökércsúcsát. A 2. lépés során az így lefixált gyökérből elindulva rekurzívan minden csúcsban adott valószínűséggel a csúcs gyerkeinek részfához vétele történik meg (a levélcúcsokhoz érve az rekurzív eljárás véletlen választás nélkül visszatér). Mivel az illesztések újramintavételezésénél a jobb keveredés érdekében mindig a szülő szekvenciáját is mintavételezzük (3-szekvenciás rejtett Markov-modellel), ezért a részfacúcsokban azoknak vagy mindkét gyereket egyszerre vesszük a részfához, vagy egyiket sem.

3.2.1. Gyökércsúcs választása

Jogos elvárásunk lehet a súlyozástól, hogy a csúcsokat az általuk meghatározott részfa valamilyen nagysági jellemzőjének függvényével arányos valószínűséggel válasszuk. Emellett szükséges, hogy az alsóbb szintek kiválasztási valószínűsége (azaz az adott szinten lévő csúcsok választási valószínűsége összege) egyre kisebb legyen. Ez a két tulajdonság biztosítja, hogy a választott gyökércsúcs csak ritkán legyen a levelek közelében, és így ne okozza a levélcúcsokban végződő, kis mélységű fák túlzott reprezentáltságát.

Minden c csúcs esetében jelöljük l_c -vel az ő részfájában lévő levelek számát! Első ránézésre a c csúcshoz l_c súly rendelése minden szempontból megfelelőnek tűnhet. Ez azonban nincs így.

Vizsgálatainkhoz tekintsünk egy m mélységű teljes bináris fát (legyen most a mélység a szintek számánál eggyel kisebb érték). A szinteket számozzuk 0-tól (a gyökércsúcs szintje) m -ig (a levelek szintje). Ekkor:

$$\forall c \in L(k) : l_c = 2^{m-k} \quad (3.1)$$

ahol $L(k)$ a k . szinten lévő csúcsok halmaza.

Gyorsan látható, hogy a k . szint összsúlya egységesen $2^k 2^{m-k} = 2^m$ (2^k darab csúcs 2^{m-k} súllyal), ami nem teljesíti a második elvárásunkat. Egy ilyen súlyozás esetén a levelek szintje is azonos valószínűséggel választódik, mint például a gyökér szintje. A kiválasztott szint várható értéke így

$$\frac{\sum_{k=0}^m k 2^m}{\sum_{k=0}^m 2^m} = \frac{m(m+1)}{m+1} = \frac{m}{2}. \quad (3.2)$$

Ésszerűnek látszik az l_c egy 1-nél nagyobb hatványát használni súlyfüggvényként. Tegyük fel tehát, hogy a c csúcs súlya $(l_c)^a$ ($a > 1$). Tekintsük ekkor az m mélységű fa esetében a választott szint várható értékét:

$$E_m^{(a)} = \frac{\sum_{k=0}^m k 2^k (2^{m-k})^a}{\sum_{k=0}^m 2^k (2^{m-k})^a} = \frac{2^{ma} \sum_{k=0}^m k 2^{k-ka}}{2^{ma} \sum_{k=0}^m 2^{k-ka}} = \frac{\sum_{k=0}^m k q^k}{\sum_{k=0}^m q^k}, \quad (3.3)$$

ahol $q = 2^{1-a}$ ($q < 1$). Bár $E_m^{(a)}$ értékét explicite meg lehet határozni, mégis érdemes inkább a határértékét vizsgálni, ha a fa mérete végtelenhez tart:

$$E^{(a)} = \lim_{m \rightarrow \infty} E_m^{(a)} = \frac{\frac{q}{(1-q)^2}}{\frac{1}{1-q}} = \frac{q}{1-q} \quad (3.4)$$

Itt a nevező a geometriai sor összegképletéből, a számláló a sor tagonkénti deriváltjából származtatható. Ez azt jelenti tehát, hogy például $a = 2$ esetén $E^{(a)} = \frac{2^{1-2}}{1-2^{1-2}} = 1$. Vagyis, ha a csúcsokhoz az l_c^2 súlyt rendeljük, akkor a fa mélységének tetszőlegesen nagyra növelésével is az 1. szint (a gyökér gyerekeinek szintje) környékén lesz nagy valószínűséggel a kiválasztott csúcs.

Még inkább célravezető lehet ezért az a kitevőt az $(1,2)$ intervallumból venni. Például $a = 1.5$ esetén $E^{(a)} = \frac{1}{\sqrt{2}-1} = \sqrt{2} + 1 \approx 2.414$, így csökken a gyökércsúcs túlzott kiválasztási valószínűsége.

3.2.2. Részfa építése

A részfa gyökércsúcsának rögzítése után egy rekurzív eljárás minden, a részfába addig bekerült csúcsnál dönt, hogy a csúcs gyerekei szintén bekerüljenek-e a részfába. Nem triviális kérdés azonban, hogy az egyes szinteken milyen valószínűséggel csatoljuk a csúcsok gyerekeit az épülő részfához, ha biztosítani szeretnénk, hogy a választott részfa várható mélysége véges legyen.

Természetes módon vetődik fel a lehetőség, hogy a csúcsokban azok szintjétől függetlenül $\frac{1}{2}$ valószínűséggel vesszük fel azok gyerekeit. Vizsgáljuk meg ezt az esetet részletesebben!

Tekintsük most úgy, mintha a részfa gyökércsúcsaként lefixált csúcs a teljes fa gyökércsúcsa lenne. Megállapításainkat így a részfa gyökércsúcsához viszonyítva lesznek érvényesek.

3.1. Definíció. Minden $n \geq 0$ -ra legyen P_n annak a valószínűsége, hogy a részfa kiválasztása során elértük az n . szintet, azaz az eredmény részfa legalább n mélységű.

Nyilván $P_0 = 1$, $P_1 = \frac{1}{2}$, és P_n szigorúan monoton csökken. Hogyan határozható meg azonban P_n általánosan?

Érezhető, hogy valamilyen rekurzív összefüggés adható a kiszámítására. A levelek felőli megközelítés nem működik. A gyökércsúcs felőli viszont annál inkább.

3.2. Állítás.

$$P_n = P_{n-1} - \frac{1}{2}P_{n-1}^2 \quad (n \geq 1)$$

Bizonyítás. Pontosan akkor választunk a gyökérből legalább n mélységű fát, ha

- 1.) a gyökér gyerekeit beválasztjuk, és
- 2.) legalább az egyik gyerekétől indulva legalább $n - 1$ mélységű fát választunk

Az 1. és 2. esemény független egymástól. Az 1. esemény valószínűsége $\frac{1}{2}$, a 2.-at érdemes a komplementer esemény irányából közelíteni: mindkét gyerektől indulva $n - 1$ -nél kisebb mélységű fát választunk. Annak valószínűségű, hogy ez a bal gyerekre fennáll,

$1 - P_{n-1}$, a jobb gyerekre ugyanennyi, és a két esemény független. Azt kaptuk, hogy a 2. esemény $1 - (1 - P_{n-1})^2$ valószínűséggel következik be.

A kettőt összevetve, P_n -re a következő rekurzív egyenlet adódik:

$$P_n = \frac{1}{2} (1 - (1 - P_{n-1})^2) = P_{n-1} - \frac{1}{2} P_{n-1}^2 \quad (n \geq 1)$$

□

A 3.2 állításból nem látszik P_n csökkenési üteme. Vezessünk azonban be a következő sorozatot:

3.3. Definíció. $C_n := \frac{2}{P_{n-1}} - 1 \quad (n \geq 1)$

Ekkor $C_1 = \frac{2}{P_0} - 1 = 1$, $C_n \geq 1 \quad (n \geq 1)$. 3.3 átrendezésével P_n kifejezhető, ha tehát C_n aszimptotikus viselkedését meghatározzuk, akkor P_n -ét is megtudjuk:

$$P_n = \frac{2}{C_{n+1} + 1} \quad (n \geq 0) \tag{3.5}$$

3.4. Lemma. $n \leq C_n \leq 2n \quad (n \geq 1)$

Bizonyítás. Felhasználva a 3.2 állítást és (3.5)-öt, a C_n -re is adható egy rekurzív összefüggés:

$$\begin{aligned} \frac{2}{C_{n+1} + 1} &= \frac{2}{C_n + 1} - \frac{1}{2} \left(\frac{2}{C_n + 1} \right)^2 \implies \\ C_{n+1} + 1 &= \frac{(C_n + 1)^2}{(C_n + 1) - 1} \implies \\ C_{n+1} &= C_n + \frac{1}{C_n} + 1 \end{aligned} \tag{3.6}$$

Ebből az alakból teljes indukcióval azonnal belátható, hogy $C_n \geq n \quad (n \geq 1)$, hiszen $C_1 = 1 \geq 1$ és az indukciós feltételt kihasználva és (3.6)-ból $C_{n+1} \geq n + \frac{1}{C_n} + 1 \geq n + 1$. Másrészt, szintén indukcióval, $C_n \leq 2n \quad (n \geq 1)$, ugyanis $C_1 = 1 \leq 2$, és $C_{n+1} \leq 2n + \frac{1}{C_n} + 1 \leq 2n + 1 + 1 = 2(n + 1)$. □

3.5. Következmény. $C_n = \Theta(n)$, tehát P_n csökkenési sebessége lineáris:

$$P_n \geq \frac{2}{2(n+1)+1} = \frac{2}{2n+3}$$

A kiválasztott részfa várható mélységének becsléséhez vezessük be M_n -et.

3.6. Definíció. M_n annak a valószínűsége, hogy a gyökérből **pontosan** n mélységű fát választunk.

Ezzel a jelöléssel a várható mélység a következőképp írható:

$$E^{(1/2)} = \sum_{n=0}^{\infty} nM_n \quad (3.7)$$

3.7. Tétel. $E^{(1/2)} = +\infty$

Bizonyítás. Könnyen kapcsolat teremthető M_n és P_n között, hiszen P_n a **legalább** n mélységű fa kiválasztási valószínűségét jelenti:

$$P_n = \sum_{k=n}^{\infty} M_k \quad (3.8)$$

Ebből néhány átalakítással:

$$P_n = M_n + \sum_{k=n+1}^{\infty} M_k = M_n + P_{n+1} \implies M_n = P_n - P_{n+1} \quad (3.9)$$

Vizsgáljuk tehát meg a várható mélységet:

$$\begin{aligned} E^{(1/2)} &= \sum_{n=0}^{\infty} n(P_n - P_{n+1}) = \\ &= 0P_0 - 0P_1 + 1P_1 - 1P_2 + 2P_2 - 2P_3 + 3P_3 - \\ &\quad - \dots - (n-1)P_n + nP_n - \dots = \sum_{n=1}^{\infty} P_n \end{aligned} \quad (3.10)$$

Felhasználva 3.5-öt is:

$$E^{(1/2)} \geq \sum_{n=1}^{\infty} \frac{2}{2n+3} \geq \sum_{n=1}^{\infty} \frac{2}{2n+3n} = \frac{2}{5} \sum_{n=1}^{\infty} \frac{1}{n} = +\infty \quad (3.11)$$

□

Összegezve, nem jó megoldás tehát $\frac{1}{2}$ valószínűséggel csatolni a csúcsok gyerekeit a részfához, mert ezzel az eljárással várhatólag végtelen nagy fát fogunk választani. Nyilván annál nagyobb valószínűséget használva is erre az eredményre jutnánk.

Vizsgáljuk a kérdést általánosabban!

3.8. Definíció. Jelölje p a csúcsok gyerekeinek kiválasztási valószínűségét, az adott csúcs szintjétől függetlenül.

3.9. Definíció. $P_n^{(p)}$, $C_{n+1}^{(p)}$ és $M_n^{(p)}$ ($p < \frac{1}{2}$, $n \geq 0$) jelentse a 3.1, 3.3, 3.6 definíciókban szereplő megfelelőik analógját, $\frac{1}{2}$ helyett p kiválasztási valószínűséggel.

Általános esetben is igazak a következők:

- i.) $P_0^{(p)} = 1$, $P^{(p)} \downarrow$
- ii.) $C_1^{(p)} = 1$, $C_n^{(p)} \geq 1$ ($n \geq 1$), $C^{(p)} \uparrow$
- iii.) $M_n^{(p)} = P_n^{(p)} - P_{n+1}^{(p)}$

A 3.2 állítás általánosítása a következő:

3.10. Állítás. $P_n^{(p)} = 2pP_{n-1}^{(p)} - p(P_{n-1}^{(p)})^2$

Bizonyítás. Lényegében egyezik a 3.2 állítás bizonyításával, de a csúcsok gyerekeit p valószínűséggel választjuk. □

$P_n^{(p)}$ csökkenési ütemét most is $C_n^{(p)}$ segítségével fogjuk becsülni.

3.11. Lemma. $C_n^{(p)} \geq (s^{(p)})^{n-1}$ ($n \geq 1$), ahol $s^{(p)} = \frac{1}{2p}$

Bizonyítás. Adjunk (3.6) képlettel analóg rekurzív összefüggést $C_n^{(p)}$ -re (3.5) megfelelőjét valamint a 3.10 állítást felhasználva:

$$\begin{aligned} \frac{2}{C_{n+1}^{(p)} + 1} &= p \left(\frac{2}{C_n^{(p)} + 1} - \frac{2}{(C_n^{(p)} + 1)^2} \right) \implies \\ C_{n+1}^{(p)} &= \frac{1}{2p} \frac{(C_n^{(p)} + 1)^2}{C_n^{(p)}} - 1 = \frac{1}{2p} \left(C_n^{(p)} + \frac{1}{C_n^{(p)}} \right) + \frac{1}{p} - 1 \implies \\ C_{n+1}^{(p)} &= s^{(p)} \left(C_n^{(p)} + \frac{1}{C_n^{(p)}} \right) + t^{(p)} \end{aligned} \quad (3.12)$$

ahol $s^{(p)} = \frac{1}{2p}$, $t^{(p)} = \frac{1}{p} - 1$, és $p < \frac{1}{2}$ miatt $s^{(p)}, t^{(p)} > 1$. Sejthető (3.12) alapján, hogy $C_n^{(p)} = \Omega(2^n)$. Konkrétan, a lemma állítását kapjuk teljes indukcióval, ugyanis $C_1^{(p)} = 1 \geq (s^{(p)})^{1-1}$, és az indukciós feltételből és (3.12)-ből $C_{n+1}^{(p)} = s^{(p)} \left(C_n^{(p)} + \frac{1}{C_n^{(p)}} \right) + t^{(p)} \geq s^{(p)} C_n^{(p)} \geq (s^{(p)})^n$. \square

3.12. Következmény. $P_n^{(p)}$ exponenciálisan csökken, egész pontosan (felhasználva (3.6) analógját is):

$$P_n^{(p)} \leq \frac{2}{(s^{(p)})^n + 1} = \frac{2}{\left(\frac{1}{2p}\right)^n + 1} \leq \frac{2}{\left(\frac{1}{2p}\right)^n} = 2(2p)^n$$

Most már minden rendelkezésünkre áll, hogy megállapítsuk:

3.13. Tétel. $E^{(p)} \leq \frac{2}{1-2p} - 2 < +\infty$ ($p < \frac{1}{2}$)

Bizonyítás. A 3.7 tétel bizonyításához hasonlóan, felhasználva az $M_n^{(p)}$ és a $P_n^{(p)}$ közti összefüggést:

$$E^{(p)} = \sum_{n=0}^{\infty} n M_n^{(p)} = \sum_{n=1}^{\infty} P_n^{(p)} \leq \sum_{n=1}^{\infty} 2(2p)^n = \frac{2}{1-2p} - 2$$

\square

Beláttuk, hogy bármilyen $\frac{1}{2}$ -nél kisebb valószínűséget használva a csúcsok gyerekeinek részfába választásához, a kiválasztott fa várható mélysége véges lesz, és erre egy

felső korlátot is kaptunk. Például $p = 0.4$ esetén a várható mélység kisebb, mint $\frac{2}{1-0.8} - 2 = 8$. A várható érték pontos, analitikus meghatározása nem könnyű. A 3.10 rekurzív összefüggés alapján viszont numerikusan jól számítható, a konvergencia $\frac{1}{2}$ -hez nem túl közeli p -kre nagyon gyors.

p	0.4	0.45	0.47	0.49	0.499	0.4999
$E^{(p)}$	1.40	2.34	3.13	5.02	9.42	13.99

3.1. táblázat. Várható mélységek p függvényében

A 3.1 táblázatban a várható értékek numerikusan meghatározott értékei láthatók, különböző p -k mellett. Olyan p érték választása célszerű, amely esetén a várható mélység (esetleg jóval) kisebb, mint a törzsfa átlagos mélysége.

3.3. Ablakkivágás

Az illesztések újramintavételezése előtti utolsó lépés a már rögzített részfa gyökérszekvenciájában egy ablak kiválasztása (az első és utolsó karakter indexének meghatározása), amelyen belül az illesztéseket meg kell változtatni.

Az ablakkivágással akkor van probléma, ha előállhat olyan helyzet, hogy egy x állapotból egy olyan y állapotot javasolunk, amelyből ugyanolyan ablakkivágási mechanizmus esetén az x állapot nem érhető el. Más szavakkal, ha a backproposal valószínűség nulla is lehet.

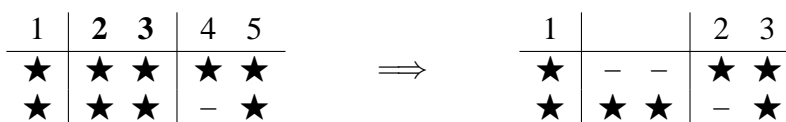
Olyan kivágási módszert kell tehát találni, ahol a backproposal mindig pozitív. Kétféle problémával szembesülhetünk:

- 1.) Előfordulhat, hogy a kivágott ablakban a szekvenciaillesztéskor az ősi szekvencia hossza megnő, ezért a kapott állapotból csak nagyobb ablak kivágásával juthatunk vissza az eredeti állapotba.
- 2.) Ahhoz, hogy az ablakot a kiterjesztési folyamat során az ősi szekvenciáról a modern szekvenciára „adhassuk át”, az ablak elejét és végét az ősi és a modern szekvencia

egy-egy kezdő- és végpozíciójához kell rendelni. Ha a kezdő- vagy végpozíció az ablak belsejében van, akkor az ősi szekvencia megváltozása miatt lehetséges, hogy nem tudjuk ugyanazt az ablakot kiválasztani.

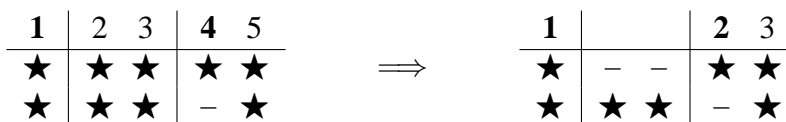
Az első probléma könnyen elkerülhető úgy, hogy a kivágható ablakméret tetszőlegesen nagy lehet, a méretet nem korlátozzuk felülről. Hasonlóan, alulról sem korlátozhatjuk, tehát a 0 hosszú szekvenciárészletnek ugyanúgy, mint a teljes szekvenciának, kivághatónak kell lennie.

A második problémát szemlélteti a 3.1 ábra. Ha az ősi szekvencia 2. és 3. indexű karaktere jelzi az ablak elejét és végét, akkor az ábrán látható új szekvencia esetén nem fogjuk tudni kivágni ugyanazt az ablakot, legfeljebb az 1. és 2. karakterrel jelzett nagyobb ablakot, aminek viszont van egy megfelelője az eredeti szekvenciák esetében is (az 1.-4. ablak).



3.1. ábra. Ablakkivágás és az újraillesztés hatása.

A megoldást az jelenti, ha a kivágásnál az ablak bal szélét az a karakter jelzi, amelyik még éppen *nincs benne* az ablakban, a jobb szélét pedig az a karakter, amelyik már nincs benne, a 3.2 ábrának megfelelően.



3.2. ábra. Ablakkivágás és az újraillesztés hatása 2.

Természetesen így nem állhat elő az a helyzet, hogy ne tudnánk ugyanazt az ablakot kiválasztani, hiszen az ablakot jelző karakterek nem veszhetnek el (hiszen nincsenek az ablakon belül), legfeljebb a sorszámuk változhat meg.

Mindezt szem előtt tartva az ablak kiterjesztése a levélsekvenciák irányába problémamentes, és a backproposal valószínűség mindig pozitív.

4. fejezet

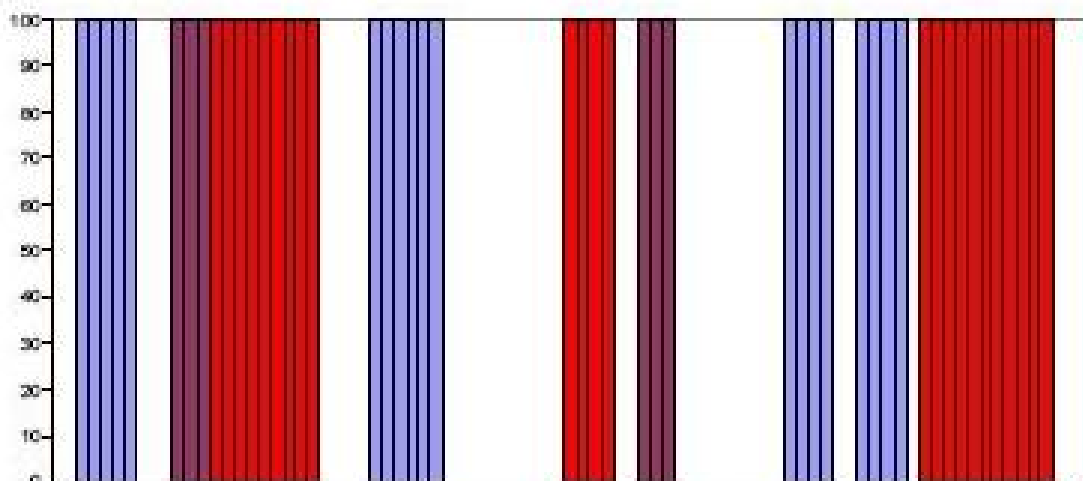
Eredmények

A módszerünket protein szekvenciák másodlagos térszerkezet-predikcióján teszteltük, úgy, hogy az egyik szekvencia másodlagos térszerkezetét bemenetként megadtuk, és megvizsgáltuk, hogy milyen predikciókat kapunk a többi szekvencia térszerkezeteire. Ezek a térszerkezetek ismertek, ami alapján a módszerünk helyességét mérni tudjuk.

Három különböző fehérjecsaládból származó proteinek szekvenciára teszteltük az eljárásunkat. A családok között lényeges eltérés van a szekvenciák konzerváltságát illetően.

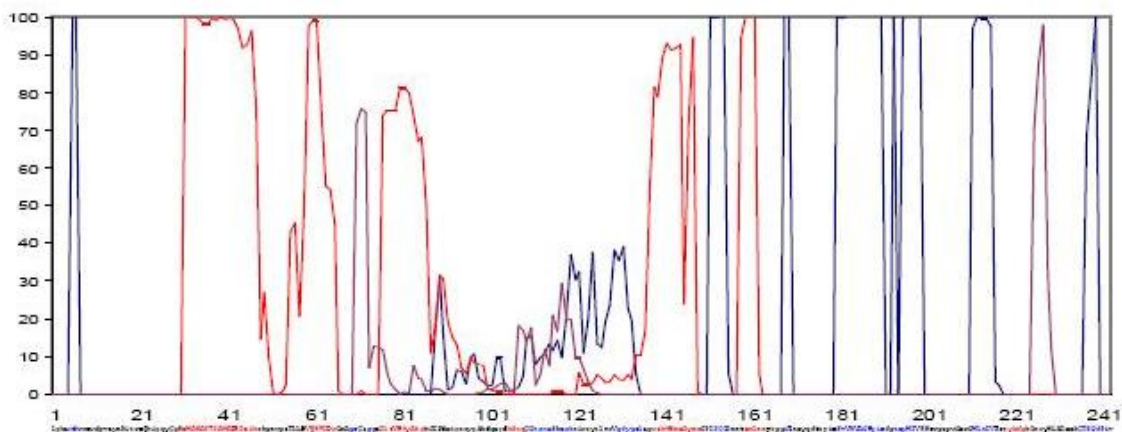
1. A *Glutation S-transzferáz* családból származó fehérjék között igen nagy a homológia. A háromdimenziós szerkezet tanulmányozása azt mutatta, hogy gyakorlatilag a fehérjék néhány régiótól eltekintve teljes mértékben fedésbe hozhatók.
2. A *Papain cisztein-proteináz* család fehérjéire a szekvencia jelentős részén nagy homológia, míg egy régióban nagy variabilitás jellemző.
3. Az *Alfa-béta hidroláz* fehérjecsalád a szekvenciák több részén igen nagy evolúciós eltérést mutat.

Az alábbi ábrákon az egyes fehérjecsaládok vizsgálata során az ismeretlen térszerkezetűnek megjelölt fehérje szekvenciájának különböző pontjain a jóslott másodlagos térszerkezeti elemek és azok megbízhatósága látható.



4.1. ábra. Glutathion S-transzferáz fehérjecsalád.

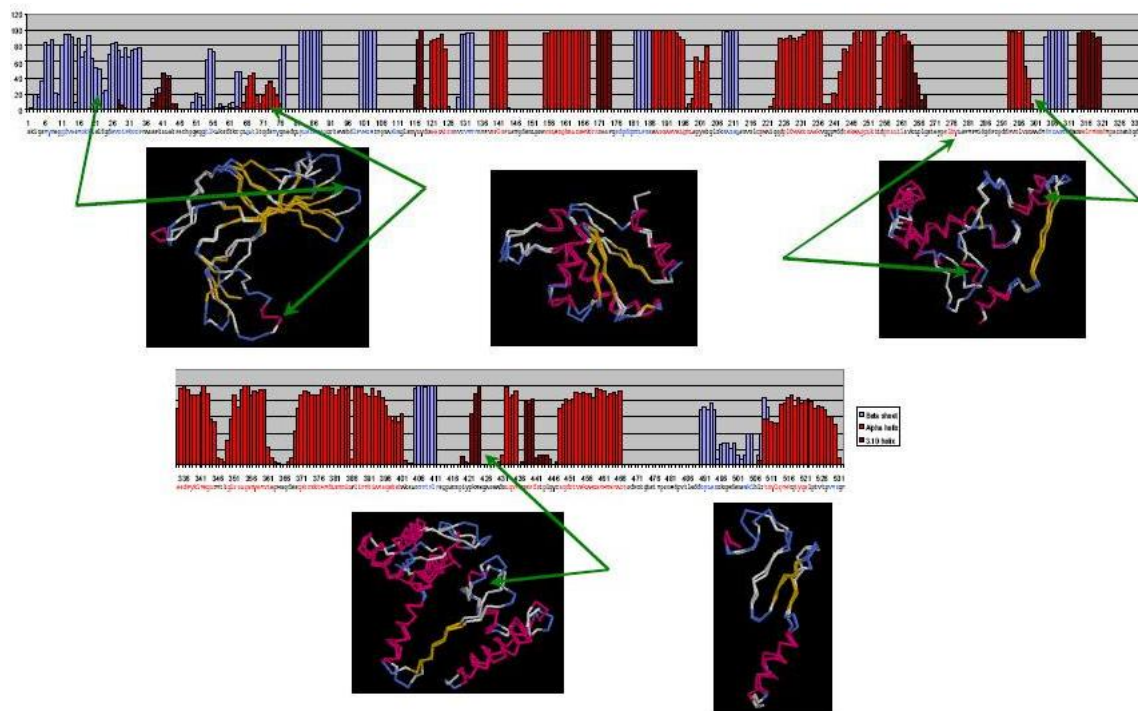
A 4.1 ábrán a nagyon konzervatív Glutathion S-transzferáz család egyik tagjának prediktált másodlagos térszerkezeti elemei vannak feltüntetve. Jól érzékelhető az igen magas, közel 100%-os prediktálási megbízhatóság. Az eljárás ebben az esetben valóban hibátlanul, minden pozíción helyesen jósolta a térszerkezetet.



4.2. ábra. Papain cisztein-proteináz fehérjecsalád.

A 4.1 ábrán megfigyelhető az alacsony megbízhatósági indexszel jelzett középső, va-

riálabilis régió. Ezen a területen kívül az eljárás szinte mindenhol a megfelelő másodlagos térszerkezetet prediktálta.



4.3. ábra. Alfa-béta hidroláz fehérjecsalád.

A 4.3 ábrán egy olyan család esetén tüntettük fel a jóslt szerkezeteket azok megbízhatóságával, ahol az evolúciós folyamatok jelentős térszerkezetbeli eltéréseket okoztak az egyes proteinek között. Az elég erős homológia hiányában nem várható a másodlagos szerkezetek pontos jóslása. Valóban, a szekvencia több pontján hibás a módszer predikciója, de ami fontosabb: figyelemre méltó korreláció mutatható ki az alacsony megbízhatóságú régiók és a háromdimenziós szerkezet alapján nagy változatossággal rendelkező területek között.

Összefoglalva elmondható, hogy ez a módszer alapot ad egy olyan *in silico* térszerkezetmeghatározásra, amely során egy ismert fehérje térszerkezetét a vele homológ, ismeretlen másodlagos szerkezetű fehérjére vetítjük, és megvizsgáljuk a kapott megbízhatóságokat. Az alacsony indexű, variábilis régiók pontos meghatározását laboratóriumi körülmények között lehet folytatni, a magas indexszel rendelkező területekre adott tér-

szerkezeti becslések viszont nagy valószínűséggel helyesek, és a költséges és időigényes kémiai meghatározást kiválthatják.

Hasonlóan eredményesen alkalmazható a módszerünk olyan evolúcióelméleti kérdések eldöntésére, amelyek törzsfák korai elágazásainak sorrendjét kutatják [23]. Ezeknek a kérdéseknek a megválaszolása szükségszerűen igen pontos evolúciós modelleket igényel [21], és a hosszú számítási idő is megengedhető.

Tartalmi összefoglaló

A (többszörös) szekvenciaillesztéseket és a törzsfákat meghatározó eljárások tulajdonságaikból adódóan elkerülhetetlenül egymásra épülnek. Régóta ismert, hogy a legpontosabb eredmények érdekében a kettőt egyszerre érdemes becsülni. Evolúciós törzsfák készítésére ma leggyakrabban alkalmazott algoritmusok viszont legfeljebb néhány helyesen ítélt szekvenciaillesztésből állítja elő a törzsfát. Ezáltal a szuboptimális illesztések - amelyek lényeges, biológiaiailag releváns információkat hordoznak - figyelmen kívül maradnak.

Célkitűzésünk ezért egy olyan módszer kidolgozása volt, amelynek segítségével többszörös szekvenciaillesztések és evolúciós törzsfák együttesen vizsgálhatók egy Bayes statisztikai keretmunkában. A korábbi, hasonló elvekre épülő munkák egyszerűbb, pontatlanabb evolúciós modelleken alapulnak. Ezért célul tűztük ki a TKF92 modell egy továbbfejlesztésének beépítését a módszerünkbe. Kidolgoztuk a Markov-lánc Monte Carlo (MCMC) típusú eljárás részleteit.

A módszert gyakorlatba is átültetve elkészítettünk egy szoftvert, amely lehetővé teszi a szekvenciaillesztésekre és az evolúciós törzsfákra adott predikciók megbízhatóságának mérését, valamint proteinszekvenciák másodlagos térszerkezetének predikcióját egy vele homológ, ismert térszerkezetű fehérje segítségével.

Eredményeink azt mutatták, hogy a módszer jól használható, elsősorban a konzervatív régiók térszerkezetpredikciójára.

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani **dr. Miklós Istvánnak**, a témavezetőmnek a sok-sok óráért, amit arra szánt, hogy segítsen elmélyedni a bioinformatika mai problémáinak és eszköztárának világában.

Köszönettel tartozom az MCMC módszerünk implementálása során nyújtott állandó támogatásáért, magyarázataiért, az elkészült program átvizsgálásáért, javításáért.

Irodalomjegyzék

- [1] M. O. Dayhoff, R.M. Schwartz, B. C. Orcutt: A model of evolutionary change in proteins. *Atles of Protein Sequence and Structure*, **5**:345–352, 1978.
- [2] R. Durbin, S. Eddy, A. Krogh, G. Mitchinson: *Biological sequence analysis*. University Press, 1998.
- [3] S. R. Eddy, R. Durbin: RNA sequence analysis using covariance models. *Nucleic Acids Research*, **22**(11):2079–2088, 1994.
- [4] J. Felsenstein: Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**:368–376, 1981.
- [5] D. Feng, R. F. Doolittle: Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, **25**:350–360, 1987.
- [6] D. Gamerman: *Markov Chain Monte Carlo*. Chapman and Hall, 1997.
- [7] O. Gotoh: An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, **162**:705–708, 1982.
- [8] I. Holmes: Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics*, **19**(1):i147–i157, 2003.
- [9] I. Holmes, W. J. Bruno: Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**(9):803–820, 2001.

- [10] Iványi A. (szerkesztő): *Informatikai algoritmusok*, 1. kötet. ELTE Eötvös Kiadó, 2004.
- [11] N. C. Jones, P. A. Pevzner: *An Introduction to Bioinformatics Algorithms*. The MIT Press, 2004.
- [12] J. A. Lake: The order of sequence alignment can bias the selection of tree topology. *Molecular Biology and Evolution*, **8**:378–385, 1991.
- [13] Larget et al.: Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *Journal of the Royal Statistical Society Series B – Statistical Methodology*, **64**:681–693, 2002.
- [14] G. Lunter, I. Miklós, A. J. Drummond, J. L. Jensen, J. Hein: Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **83**(6):1471–2105, 2005.
- [15] F. Lutzoni, P. Wagner, V. Reeb, S. Zoller: Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Systematic Biology*, **49**:628–651, 2000.
- [16] I. Miklós: An improved model for statistical alignment of sequences evolved by a star tree. *Phylogenetics Combinatorics, Bielefeld, Germany*, 2001.
- [17] Miklós I.: *Statisztikus szekvencia illesztés*. PhD tézis, Eötvös Loránd Tudományegyetem, 2001.
- [18] I. Miklós, M. Csűrös: *Gene content evolution by gene gain, loss and duplication*. Manuscript, 2005.
- [19] I. Miklós, G. A. Lunter, I. Holmes: A long indel model for evolutionary sequence alignment. *Molecular Biology and Evolution*, **21**(3):529–540, 2004.
- [20] W. Miller, E. W. Myers: Sequence comparison with concave weighting functions. *Bulletin of Mathematical Biology*, **50**:97–120, 1988.

-
- [21] H. Phillipe, P. Forterre: The rooting of the universal tree of life is not reliable. *Journal of Molecular Evolution*, **49**:509–523, 1999.
- [22] B. D. Redelings, M. A. Suchard: Joint bayesian estimation of alignment and phylogeny. *Systematic Biology*, **54**(3):401–418, 2005.
- [23] M. C. Rivera, J. A. Lake: Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science*, **257**:74–76, 1992.
- [24] D. Sankoff: Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, **28**:35–42, 1975.
- [25] D. Sankoff, C. Morel, R. J. Cedergren: Evolution of 5S RNA and the non-randomness of base replacement. *Nature New Biology*, **245**:232–234, 1973.
- [26] P. H. Sellers: On the theory and computation of evolutionary distances. *SIAM Journal of Applied Mathematics*, **26**:787–793, 1974.
- [27] J. S. Sinsheimer: *Extensions to Evolutionary Parsimony*. PhD t zis, University of California, 1994.
- [28] J. D. Thompson, D. G. Higgins, T. J. Gibson: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific penalties and weight matrix choice. *Nucleic Acids Research*, **22**:4673–4680, 1994.
- [29] J. L. Thorne, H. Kishino: Freeing phylogenies from artifacts of alignment. *Molecular Biology and Evolution*, **9**:1148–1162, 1992.
- [30] J. L. Thorne, H. Kishino, J. Felsenstein: An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, **33**:114–124, 1991.

- [31] J. L. Thorne, H. Kishino, J. Felsenstein: Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, **34**:3–16, 1992.
- [32] L. Wang, T. Jiang: On the complexity of multiple sequence alignment. *Journal of Computational Biology*, **1**:337–348, 1994.
- [33] M. S. Waterman, T. F. Smith, W. A. Beyer: Some biological sequence metrics. *Advances in Mathematics*, **20**:367–387, 1976.