



EFOP-3.4.3-16-2016-00014



Pénzes Zsolt, Markó Bálint
Filogenetikai rekonstrukció gyakorlat

Jelen tananyag a Szegedi Tudományegyetemen
készült az Európai Unió támogatásával.

Projekt azonosító: EFOP-3.4.3-16-2016-00014

Szegedi Tudományegyetem
Cím: 6720 Szeged, Dugonics tér 13.
www.u-szeged.hu
www.szechenyi2020.hu



Európai Unió
Európai Szociális
Alap



BEFEKTETÉS A JÖVŐBE

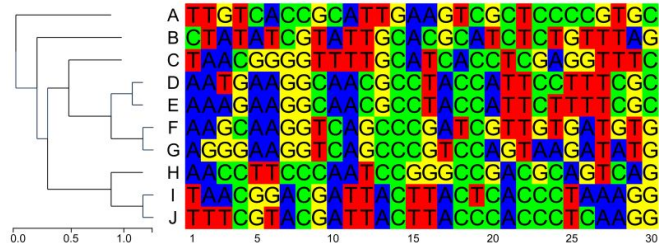
A filogenetika az evolúciós leszármazási kapcsolatok vizsgálatával foglalkozik. Célja az evolúciós történet feltárása (indirekt becslése), ilyen módon a biológia szemléletmódjának alapvető eleme. A leszármazási kapcsolatokat a filogenetikai fával (törzsfával) szemléltetjük. Azonban alkalmazhatóságának feltételei vannak, a fának különböző típusai léteznek, a leszármazáson túl további információt is hordozhatnak. Célunk a történeti rekonstrukció logikájának, az eredmények értelmezésének bemutatása.

Az előadáson a DNS szekvenciákon alapuló rekonstrukciót tárgyaljuk, elsődleges cél a gyakorlatok előkészítése, az alapfogalmak tisztázása. Feltételezett előismeretek: Filogenetika: koncepció, Filogenetika és evolúció, Molekuláris ökológia: populációk történetének rekonstrukciója.

Az olvasólecke átnézése kb. 20 percet vesz igénybe.

Molekuláris filogenetika

- Evolúció: leszármazási sorok szétválása (közös ős), karakterek divergenciája
- Karakterek: DNS szekvenciák – szekvencia pozíciók
 - sok karakter – függetlenség feltétele
 - ortológ szekvenciák
 - karakter állapot (tulajdonság): A, G, T vagy C
 - evolúciós esemény: divergencia egy ősi szekvenciából → szubsztitúció a pozícióban – modellezhető
- Filogenetikai rekonstrukció – indirekt következtetés recens mintázatból (tulajdonságok, mint filogenetikai markerek)
 - molekuláris filogenetika
 - A – J taxonok (távolság)
 - karakterek: szekvencia pozíciók – szubsztitúció
 - előnyök és hátrányok



Filogenetikai rekonstrukció

1/8

A **molekuláris filogenetika** célja a leszármazási kapcsolatok feltárása molekuláris jellemzők, ma elsősorban DNS szekvenciák alapján. DNS szekvenciák meghatározására standard protokollok állnak rendelkezésre, a lépések jelentős része ma már automatizált. A módszerek gyors fejlődésének következtében hatalmas mennyiségű információ áll rendelkezésre a legkülönbözőbb fajokra a szabadon elérhető **szekvencia adatbázisokban**, ilyen például a GenBank.

A szekvencia pozíciókat egyedek jellemzőinek tekintjük (karakterek), minden egyedhez a négy lehetséges **karakter állapot** (A, C, G, T nukleotidok) egyikét rendeljük. A karakter állapot változás (mint evolúciós változás) mutációk, génáramlás, genetikai sodródás, szelekció folyamatok következményei. Így az **evolúció elmélete** alapján a változást értelmezni tudjuk, a közös őst jellemző szekvenciából kiindulva divergencia események sorozatán keresztül, feltételezve, hogy a vizsgált csoportunk monofiletikus. Az így kialakult **változatosság** kiindulva indirekt módon következtetünk (becslést végzünk) a kérdéses csoport leszármazási viszonyaira. Ezt az eljárást **filogenetikai rekonstrukciónak** nevezzük, célunk ennek áttekintése, inkább gyakorlati oldalról megközelítve. Az egyszerűség kedvéért fajok leszármazási kapcsolatairól beszélünk, azonban a történeti rekonstrukció logikáját számos egyéb céllal is használják.

Szekvenciák illesztése

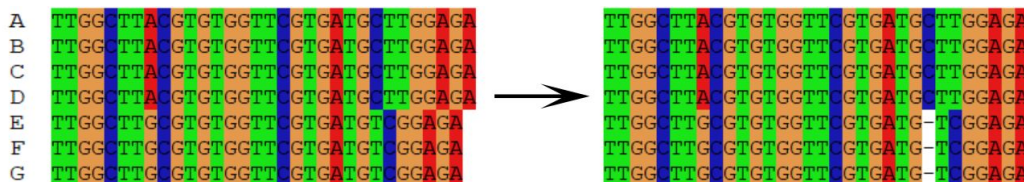
- Rekonstrukció logikai alapja:
 - ortológ szekvenciák hasonlósága evolúciójuk következménye
 - hasonlóság a leszármazásra és az evolúciós mechanizmusra utalhat – homológia
- Szekvenciák illesztése: a szekvencia hasonlóság maximalizálása
 - jelentős divergencia → kis mértékű hasonlóság
 - gyakran nem triviális
 - sok adat: statisztikai szempontok, különböző algoritmusok
 - különböző illesztések – eredmények eltérhetnek
 - döntés közöttük: mérőszám kell – a hasonlóságot tükrözi
 - változás különböző típusainak kezelése (szubsztitúció, indel)
 - szempontok:
 - többszörös illesztés kell (vs. páronként)
 - lokális vagy globális illesztések
- De: illesztés vs. valódi evolúciós történet. . .
 - jó illesztés (= maximális pontszám) $\neq$ valódi evolúciós történet
 - valódi leszármazás $\neq$ maximális hasonlóság

A történeti rekonstrukció első lépése a megfelelő, leszármazási kapcsolatokra utaló karakterek kiszűrése. Ez szekvenciák esetén egyrészt az egy közös ősből származtatható genom régiók kiválasztását jelenti (**ortológok**, történhet pl. PCR reakció során megfelelően megválasztott PCR primer párral). Másrészt a kiválasztott régió belül a szekvencia pozíciókat (a tényleges karaktereket) is rendezni kell **indel** (inzerció, deléció) események bekövetkezése esetén, hiszen így a szekvencia hossza is változik. Az eljárást **szekvenciák többszörös illesztésének** nevezzük – többszörös, mert a rendezést az összes szekvenciára egyidejűleg hajtjuk végre.

Az illesztése során első lépésben a **szekvenciák hasonlóságából** tudunk csak kiindulni (segédinformációk is lehetnek), ezt próbáljuk maximalizálni. Lehet triviális, de gyakran nem az, ha a szekvenciák jelentősen eltérnek egymástól. Különböző módszerek, értékelési szempontok léteznek a hasonlóság kifejezésére, melyek eltérő eredményt adhatnak – súlyozhatunk az azonosságtól a módosulás különböző típusainak eltérő kezeléséig. Létezik optimális módszer (Needleman–Wunsch algoritmus), azonban kivitelezése időigényes. A legjobb illesztést a valódi szekvencia evolúció következményének tekintjük. Ez azonban több kérdést is felvet, amire később még visszatérünk.

Evolúciós távolság

- Divergencia mértéke: evolúciós (genetikai) távolság
- Szekvenciák: pl. szubsztitúciók száma
 - illesztett szekvenciák – hasonlóságból távolság
 - pl. hasonlóság = azonosság → távolság: eltérő pozíciók aránya
 - indel események független kezelése
- Tapasztalt és tényleges evolúciós események száma eltérhet
 - a végtelen pozíció modell
 - egy pozíció többször is módosul – becslés modellekkel
 - szekvencia evolúció modelljei – paraméterek becslése (pl. lehetséges átmenetek szubsztitúciós rátái) a szekvencia adatokból, modell szelekció

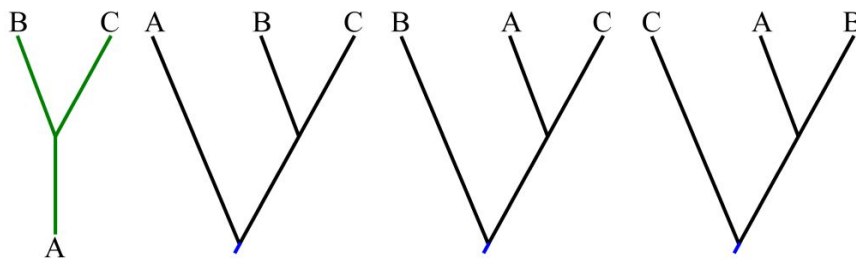


Az ábra szekvenciák illesztésére szemléltet egy példát, feltételezve, hogy 1 indel esemény történt (jele: -). Beszúrásával a szekvenciák közötti hasonlóság nő. Az illesztett szekvenciák hossza 32 bp (bázispár, pozíció) a példában. Amennyiben a hasonlóságot azonos pozíció aránnyal jellemezzük, ez 29/32. Azonban a divergencia mértékére van szükségünk, így informatívabb az eltérő pozíció arány (3/32, ahol 3 a tapasztalt szegregálódó pozíció szám). Példánkban a **tapasztalt változatosságot** jellemeztük, modern fajokra (A-G) jellemzőnek tekintett szekvenciák alapján. De nem zárhatjuk ki annak a lehetőségét, hogy a múltban történt egy szubsztitúció (mondjuk $T \rightarrow C$ az 1. pozícióban) egy közös ősből, majd egy backmutáció ($C \rightarrow T$), amely aztán elterjedt – a tapasztalt változatosság alapján a tényleges szubsztitúció számot így alul becsülhetjük.

A szekvencia evolúció részleteitől eltekintve (lásd neutrális evolúció, koaleszcencia), egyszerű **szekvencia evolúciós modellek** segítségével becsülhető a taxonok közötti, szubsztitúciókra vonatkozó genetikai (evolúciós) távolság az illesztett szekvencia adatokból. A rekonstrukció történhet e modellek (a becsült tényleges szubsztitúció szám) figyelembe vételével. A legjobb modell kiválasztása (a **modell szelekció** általános logikája alapján), azonban a leszármazás ismeretét tételeznék fel.

Fa konstrukció

- Szekvencia adatok → illesztés → evolúciós modell → fa
- Filogenetikai fa: topológia és ághosszak
 - gyökértelen és gyökeres fák
 - dichotómia és politómia
 - lehetséges topológiák száma – taxon számmal gyorsan nő (gyökértelen fák): 5 taxon: 15, 10: 2027025, 20: $2.2 * e^{20}$, 100: $1.7 * e^{182}$
 - additív fák ághossz becslése gyakran egy következő lépés (módszerfüggő)
- Eredmény: a valódi történet becslése
- Fa konstrukciós módszerek: karakter állapotok vs. genetikai távolság



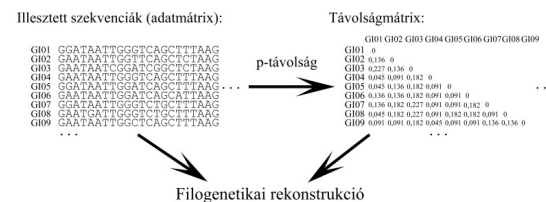
A történeti rekonstrukció következő lépése a történetet szemléltető **filogenetikai fa (törzsfa)** elkészítése. Feltételezzük, hogy egymástól független karakterekből indulunk ki (vagy a nagyszámú karakter miatt a korrelációtól eltekintünk) és az evolúció divergens jellegű, ahol egy közös ősből egyidejűleg két leszármazási sor származhat (lásd filogenetikai fa). Azonban a lehetséges taxon elrendezések (topológiák) száma gyorsan nő a taxon számmal (kiszámolható). 3 taxonnal még egy elrendezés lehetséges (bal oldali ábra, **gyökértelen fa**). Azonban a **gyökér** (a csoport utolsó közös őse, MRCA) rögzítésével ez 3 lehetséges fának felel meg. A rekonstrukció sokszor gyökértelen fát eredményez, a leszármazási sorrendet is tükröző gyökeres fát pl. a külcsoport utólagos kijelölésével kapjuk. Amennyiben a rekonstrukció feltételei teljesülnek (az evolúció valóban a feltételezett divergens jellegű), ezek egyike a valódi történetet szemléltető fa – de melyik?

A fa konstrukciós módszereket különböző szempontok szerint csoportosítjuk. Kiindulhatunk **karakter állapotokból** (nukleotidok, karakter állapot változás modellezhető) vagy **páronkénti genetikai távolságokból** és a fát ezekből készítjük. Néhány eljárás (pl. hibridizációs módszerek) nem szekvenciákat, hanem távolságokat eredményez, ez esetben az utóbbiakat kell használnunk.

Fa konstrukció

- Fa konstrukciós módszerek: algoritmikus és optimalizációs módszerek
- Genetikai távolságon alapú algoritmikus módszerek:
 - gyors, egy additív fát eredményez, de pontatlan
 - UPGMA (ultrametrikus fa), neighbor-joining (NJ)
- Karakter állapotokból kiinduló optimalizáció
 - keresés a fák terében
 - a fa pontszáma → döntés (kritérium)
 - maximális parszimónia: fa hossza, kis divergencia esetén, ált. több fa
 - maximum likelihood (ML): fa likelihood, eredménye egy additív fa, modellek használata
- Minimum evolúció (ME): távolságból kiinduló optimalizációs módszer

- távolságon alapuló módszerek:
genetikai távolság → fa
- karakter állapot változás → fa

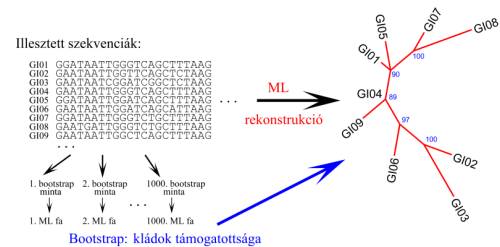


Karakter állapotok alapján a távolság becsülhető (modellek), amelyből fa készülhet. Ez azonban információ veszteséssel jár. A módszerek másik csoportja a karakter állapot változásokat közvetlenül veszi figyelembe a rekonstrukció során. Egy másik osztályozási szempont alapján a módszerek egy része optimalizáción alapul, vagy a fát valamilyen algoritmus alapján „építik”. Ez utóbbira példa a hierarchikus osztályozásra gyakran használt **UPGMA**, genetikai távolságokra alkalmazva. A módszer ultrametrikus fát eredményez, használata csak akkor javasolt, ha a molekuláris óra feltételezhető. Egy másik, általánosan használt szintén agglomeratív, távolságokból kiinduló módszer a **Neighbor-joining (NJ)**.

Az algoritmus módszerek előnye az, hogy gyorsak és egy additív fát eredményeznek. Azonban az **optimalizációs módszereknél** pontatlanabbak. Az optimalizációs módszerek a fákhoz egy pontszámot rendelnek, ez **maximális parszimónia (MP)** esetén a fa hossza (karakter változások száma az összes karakterre), míg a **maximum likelihood (ML)** logikáján alapuló becslések esetén a fa likelihood értéke. A fa kiválasztása pontszám szélsőérték keresést jelent (legrövidebb fa, maximális likelihood). Tekintettel a lehetséges topológiák nagy számára keresést végzünk a fák terében, erre különböző algoritmusok léteznek. Az ML és MP módszer karakter alapú.

Filogenetikai rekonstrukció

- Gyakorlatban főleg az ML preferált – megbízhatósága: bootstrap
- Bayes módszer: poszteriori eloszlás, likelihood
- Konszenzus módszerek
- Eltérő feltételek, nincs legjobb módszer → optimalizációs kell
- Számos egyéb lehetőség (pl. particionált elemzések, szuperfák)
- Evolúció: leszármazási sorok szétválása és karakterek divergenciája
 - nem független problémák
 - illesztés, modell szelekció, fa konstrukció elkülönítése a gyakorlatban – áthidalására egyéb eljárások
 - karakterek filogenetikai markerek – filogenetika ismeretében tesztelhető
- bootstrap szekvencia pozíciókra
- **rekonstrukció** a mintákra
- monofiletikus csoport hány ismétlésben áll elő
- **feltüntetés**e a fán (támogatottság)



Az **ML** a gyakorlatban az egyik leggyakrabban alkalmazott módszer. Előnye, hogy a szekvencia evolúciós modellek kezelése kézenfekvő. Hátránya az, hogy a rekonstrukció időigényes. Mint minden becslésnél, szükség van a becslés megbízhatóságának jellemzésére is. Erre egyik általánosan használt eljárás a **bootstrap**, az ábra ezt szemlélteti. Az egyes kládok karakterek általi támogatottságára ad információt. Egy másik általánosan használt rekonstrukció a **Bayes** logikán alapul. Ebben szintén központi szerepű a likelihood (és így a modellek), eredménye a fák (és ághosszak, paraméterek) poszteriori eloszlása. Előnye, hogy a becslés megbízhatóságára is informatív. Azonban a prior eloszlás definiálása vitatott.

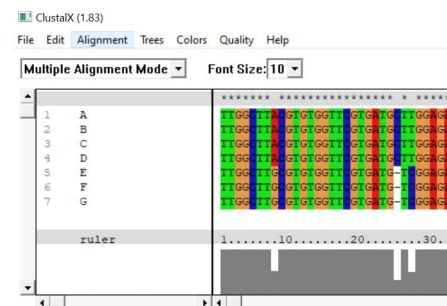
Végezetül szükséges lehet közös konklúzió levonására több fa információ tartalma alapján, melyre különböző, ún. **konszenzus** módszerek állnak rendelkezésre (pl. az MP gyakran több fát eredményez). A módszerek eltérhetnek feltételeikre való érzékenységükben, ezért javasolt több különböző rekonstrukciós logika alkalmazása.

A rekonstrukció célja a valódi történet becslése, azonban van néhány **probléma**, amire figyelemmel kell lennünk. Így például az illesztés, modell szelekciót a fa konstrukciótól függetlenül kezeltük, vagyis az előbbiekből bizonytalansága nem jelenik meg a fa becslésében. Ha rossz az illesztés, akkor a leszármazás becslése is rossz lesz.

Rekonstrukció kivitelezése (gyakorlat)

- GenBank
 - keresés taxonra
 - ortológ szekvenciák, azonosítók, részletes információk
 - keresés szekvencia alapján, BLAST
- Szekvencia file formátumok – .FASTA file, .NEXUS file, formátum konverzió
- Illesztések (Clustal, Muscle)
- Modell szelekció (JModeltest)
- Webszerverek

```
>A
TTGGCTTACGTGTGGTTCGTGATGCTTGGAGA
>B
TTGGCTTACGTGTGGTTCGTGATGCTTGGAGA
>C
TTGGCTTACGTGTGGTTCGTGATGCTTGGAGA
>D
TTGGCTTACGTGTGGTTCGTGATGCTTGGAGA
>E
TTGGCTTACGTGTGGTTCGTGATGCTTGGAGA
>F
TTGGCTTACGTGTGGTTCGTGATGCTTGGAGA
>G
TTGGCTTACGTGTGGTTCGTGATGCTTGGAGA
```

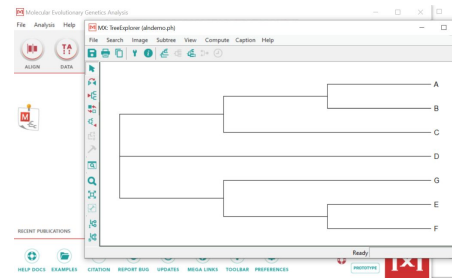


A gyakorlaton a GenBank adatbázisból letöltött **DNS szekvenciák-kal** dolgozunk. Minden szekvenciának egyedi azonosítója van, ezzel tudjuk egyértelműen hivatkozni. A szekvenciákat legegyszerűbben **FASTA** file formátumban tudjuk letölteni. A FASTA egy ASCII text file, ahol a szekvenciákat a '>' karakterrel kezdődő szekvenciát azonosító sorok választják el (néhány további megkötés is van). A nukleotidokat egybetűs kódjukkal adjuk meg, az „indel” jele '-'. A hiányzó adatot 'N' karakterrel jelöljük. A NEXUS szintén egy általánosan használt standard, amely különböző blokkokból áll (taxon, adat, fa stb.) és több információt tárolhat.

Az **illesztésekhez** a ClustalX és a Muscle programokat használjuk. A globális illesztést végző Clustal készít egy segédfat (ez egy UPGMA fa a páronkénti illesztések alapján), ez nem keverendő a filogenetikai fával. Ezek mellett számos egyéb eltérő logikán alapuló szoftver létezik, illetve több online is futtatható webszervereken. Az illesztett szekvencia adatokat FASTA file-ba mentjük. Ez következő lépésben felhasználható szekvencia evolúciós **modell becslésre**, erre is különböző szoftverek léteznek. Egyik ilyen a JModeltest, amely számos különböző modellel számol. A modellek tesztelésének része első lépésként egy fa készítése, a modellek tesztelése ezen a fán történik (szintén csak egy segéd fa).

Rekonstrukció kivitelezése (gyakorlat)

- Fa file formátum – Newick, .NEXUS
- Rekonstrukció Mega szoftverrel
- Bayes rekonstrukció (MrBayes)
- R szoftver
 - szekvenciák, fák, távolságok, modellek
 - APE csomag
 - karakter evolúció R-ben
- Newick (New Hampshire) fa formátum
- monofiletikus csoportok zárójelben
- pl. (((A,B),C),D,((E,F),G));
- megjelenítése MEGA-ban



Következő lépés a fa elkészítése, erre a **MEGA szoftvert** használjuk, melyben több módszer is elérhető. Futtatható belőle a Clustal és Muscle is, illetve a bootstrap is kivitelezhető (de a bootstrap fa nem azonos a legjobb becsléssel!). Emellett számos egyéb céllal is alkalmazható szekvencia elemzésre. Azonban az adott célra készült specifikusabb szoftvereket nem helyettesíti (pl. PhyML, RaxML webszerverek). A MEGA saját file formátummal rendelkezik, de a FASTA importálható.

A fák számítógépes reprezentációjára szolgál a **Newick** fa formátum (pl. text file-ban). A monofiletikus csoportokat zárójelekben adjuk meg, felsorolásuk sorrendje tetszőleges. Így egy adott történet számos különböző módon leírható (és ábrázolható). A ';' karakter a sor végén kötelező elem. Ághosszak, bootstrap értékek is megadhatóak, gyökeres és gyökértelen fa leírására egyaránt alkalmas.

A statisztikai problémák kezelésére megismert **R szoftverben** a megszokott módon kezelhetők a DNS szekvenciák vagy a fák, ehhez egy kézenfekvő megoldást kínál az APE csomag. Ilyen módon könnyedén végezhetünk további elemzéseket is például az ághosszakkal, vagy számolhatunk filogenetikai diverzitást. Karakter evolúció elemzésekre is számos lehetőséget nyújt, erre különböző csomagok állnak rendelkezésre.

Ellenőrző kérdések

- ❶ Mit értünk filogenetikai rekonstrukció alatt?
- ❷ Miért a szekvencia pozíciókat és nem a teljes szekvenciát tekintjük karakternek?
- ❸ Milyen folyamatok magyarázhatják a karakter állapot változást?
- ❹ Mi a szekvencia illesztések célja?
- ❺ Miért lehet félrevezető a szekvencia hasonlóság a leszármazásra vonatkozóan?
- ❻ Mit értünk evolúciós távolság alatt?
- ❼ Miért van szükség szekvencia evolúciós modellekre?
- ❽ Hogyan csoportosítjuk a fa konstrukciós módszereket?
- ❾ Az UPGMA módszer alkalmazása miért jelenti a molekuláris óra feltételezését?
- ❿ Miért lehet szükség bootstrap eljárásra a rekonstrukciót követően?



EFOP-3.4.3-16-2016-00014



Jelen tananyag a Szegedi Tudományegyetemen
készült az Európai Unió támogatásával. Projekt
azonosító: EFOP-3.4.3-16-2016-00014

Szegedi Tudományegyetem
Cím: 6720 Szeged, Dugonics tér 13.
www.u-szeged.hu
www.szechenyi2020.hu



MAGYARORSZÁG
KORMÁNYA

Európai Unió
Európai Szociális
Alap



BEFEKTETÉS A JÖVŐBE