



EFOP-3.4.3-16-2016-00014



Pénzes Zsolt, Markó Bálint

Molekuláris ökológia: populációk történetének rekonstrukciója

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 molekuláris ökológia előadások célja a molekuláris módszerek néhány alkalmazási lehetőségének bemutatása ökológiai és evolúcióböiológiai problémák megfogalmazásában/megválaszolásában. Kérdéseink populációkra, fajokra vonatkoznak – például populációk izolációjának mértéke, egy invazív faj eredete, leszármazási kapcsolatok. A válasz keresése során a molekuláris módszerek eszközökként szolgálnak.

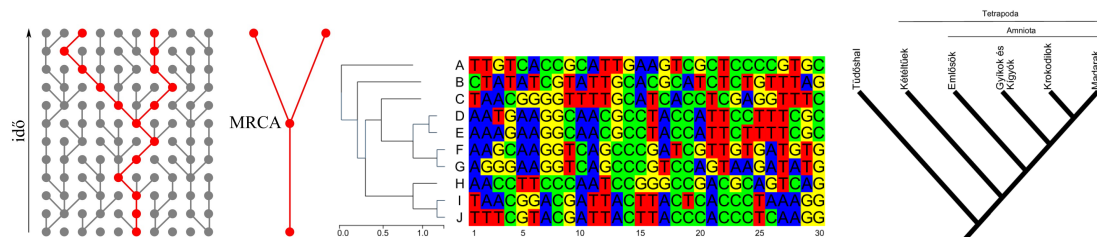
Az előadáson a koaleszcens fák rekonstrukcióját és alkalmazását, a fa alapú populációkra történő következtetés sajátosságait tárgyaljuk. Ajánlott irodalom: Nielsen R, Slatkin M (2013) An Introduction to Population Genetics. Sinauer. (5. fejezet)

Feltételezett előismeretek: molekuláris markerek, a genetikai változatosság mérőszámai, ideális populáció, neutrális változás, koaleszcens modell, populációk szerkezete.

Az olvasólecke átnézése kb. 20 percet vesz igénybe.

Következtetés a történetre

- DNS szekvencia adatok
- Populációk genetikai változatossága → statisztikák
- Információ a leszármazási kapcsolatokra is – történeti rekonstrukció?
 - koaleszcens fa – minta, leszármazás szemléltetése
 - filogenetika rekonstrukciós módszerei – populációkra is?
- A fa értelmezése
- A filogenetikai rekonstrukció módszerei

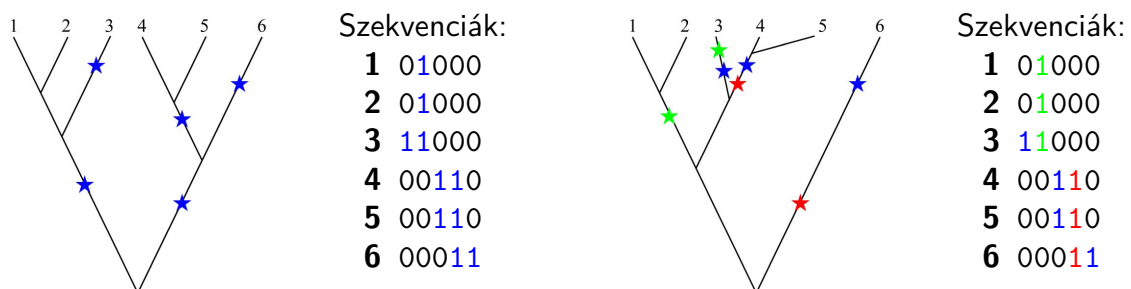


Alapvető információkhoz juthatunk a populációkról, populációk közötti kapcsolatokról különböző, a genetikai változatosságra vonatkozó mérőszámok segítségével (lásd például π vagy F_{ST}). Azonban egy egyszerű mérőszám nem mindig elegendő. A változatosság adatok informatívak lehetnek a populációkat alkotó egyedek kapcsolatára is, de az erre vonatkozó következtetés gyakran más megközelítési módot igényel. Például a leszármazási kapcsolatok feltárása az evolúcióbiológia egy kulcskérdése, melyre vonatkozóan megalapozott metodikával rendelkezik.

Fajok, taxonok leszármazási kapcsolatait **filogenetikai (evolúciós) fákkal** szemléltetjük, melynek becslésére különböző módszerek állnak rendelkezésre. A történeti rekonstrukció gyakran DNS szekvencia adatokon alapul. Az egyedek közötti kapcsolatokat is fákkal, **koaleszcencia fákkal** ábrázoltuk, a genom egy lokuszán tapasztalt változatosságot véve alapul. Felmerül a kérdés, hogyan kapcsolható össze a két logika? Mikor és milyen feltételekkel alkalmazhatjuk a fajok leszármazási kapcsolatainak feltárását célzó filogenetikai rekonstrukció módszereit a populációk jellemzésére? Célunk ennek a kérdéskörnek az áttekintése, függetlenül attól, hogy a rekonstrukció milyen – maximális parszimónia (MP), távolság alapú, maximum likelihood (ML) vagy Bayes – módszerrel történik a gyakorlatban.

Rekonstrukció példa

- Példa: rekonstrukció a maximális parszimónia (MP) módszerrel
 - lehetséges fa topológiák nagy száma – valódi történet?
 - MP: szekvencia adatok magyarázata a lehetséges legkevesebb mutáció (szubsztitúció) feltételezésével
 - optimalizációs módszer – döntés a fa hossza (feltételezett mutáció szám) alapján
 - legrövidebb fa (fák) a valódi történet legjobb becslése a módszerrel
- Ábra: két lehetséges topológia 6 fajra (a lehetséges 945-ből), 5 változatosságot mutató szekvencia pozíció (SNP)



Populáció történet rekonstrukciója

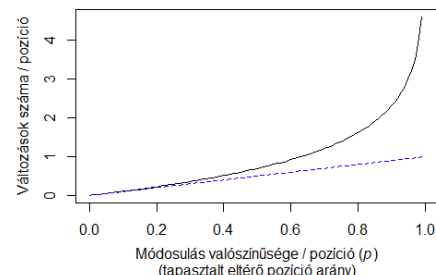
2/9

A **maximális parszimónia (MP)** rekonstrukciós módszerrel a változások számának minimalizálásával következtetünk a leszármazási kapcsolatokra. Az ábrák a bináris szekvencia adatokat (invariáns pozíciók feltüntetése nélkül) és a mutációk eloszlását (★) szemléltetik két lehetséges fa topológián. Most csak a szekvencia pozíció (SNP) változásának ténye érdekes a számunkra, 0 az ősi és 1 a származtatott allél. A baloldali ábra egy a **végtelen pozíció modellel** kompatibilis esetet szemléltet: a fa alapján az adatok a szegregálódnak pozíciók számával azonos számú (5) mutációval értelmezhetők. A jobb oldali ábra, egy másik topológia, már nem kompatibilis a végtelen pozíció modellel. Az adatok legalább 7 mutáció feltételezésével értelmezhetők a fa alapján. A **második** és a **negyedik** pozícióban minimálisan két-két, a **többiben** egy változást kell feltételeznünk. A maximális parszimónia módszerrel a baloldali fát tekintenénk a valódi történet jobb becslésének, mert kevesebb a minimálisan feltételezett teljes változás szám (a fa hossza 5).

A gyakorlatban elvileg az összes lehetséges topológián meg kell határoznunk a minimálisan szükséges változások számát ahhoz, hogy a legkevesebb változást igénylőket megtaláljuk. A maximális parszimónia, mint optimalizációs kritérium azonban számos szempontból vitatható, különösen szekvencia adatokra.

Genetikai távolság

- Genetikai (evolúciós) távolság: a változás mértéke, pl. szubsztitúciók száma (→ fa becslése, távolság alapú rekonstrukciós módszerek)
- Végtelen pozíció modell: tapasztalt = tényleges mutáció (szubsztitúció) szám
- Egy pozíció többször is módosulhat
 - genetikai távolság: tényleges változások száma $\neq$ tapasztalt
 - korrekciók, a szekvencia evolúció modelljei
- változások várható száma a tapasztalt távolság függvényében (egy pozícióra vonatkoztatva), Poisson modell
- eltérés a **végtelen pozíció modell**től nő a módosulás valószínűségével
- a modellek alkalmazásával további feltételek



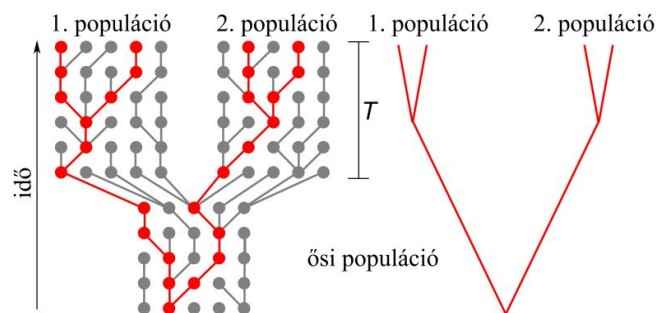
Egy szekvencia pozícióban több mutáció is bekövetkezhet (és terjedhet el, szubsztitúció történik). Ez esetben a végtelen pozíció modell már rossz becslése az evolúciós változás mértékének, mivel a ténylegesen bekövetkezett szubsztitúciók számát alul becsüli. Különböző módszerek léteznek a **többszörös szubsztitúcióra** való korrekcióra egy szekvencia párra vonatkozóan (szekvencia evolúciós modellek). A legegyszerűbb korrekciós modell a **Poisson-folyamaton** alapul (ahol a várakozási idő a következő eseményre exponenciális eloszlású, lásd koaleszcencia). Legyen p annak a valószínűsége, hogy a szekvencia egy adott pozíciója legalább az egyik leszármazási sorban legalább egyszer módosul. p így egyben a szekvencia azon pozícióinak arányát is jelenti, amelyekben változást várunk, amennyiben minden pozícióra azonos a mutációs ráta. Ez a szekvenciák közötti **p -távolság**, ami a tapasztalt változatosság alapján közvetlenül becsülhető. Ha a genetikai távolság (várható módosulás szám) egy szekvencia párra d , annak a valószínűsége, hogy egy pozícióban nincs változás a Poisson modell alapján:

$$1 - p = e^{-d}, \text{ így } d = -\ln(1 - p)$$

Ezt a Poisson-korrigált távolságot szemlélteti az ábra a p -távolság függvényében ($d(p)$ függvény).

Génfa

- Filogenetikai módszerek – fajok leszármazási kapcsolatai
- Ma populációkra is: génfa
 - populáció $\rightarrow$ faj divergencia
 - gén leszármazási sorok szétválása vs. speciáció
 - T divergencia és t koaleszcencia idő
 - ha 1 és 2 faj, T a fajok divergencia ideje – t -től függetlenül
 - szekvenciák $\rightarrow t$ (nem T) becslése $\rightarrow$ génfa
- Ha kizárólagos MRCA a faj vagy populáció minden egyedére (ábra: **mintára**): reciprok monofiletikus $\rightarrow$ génfa és fajfa topológia azonos



Populáció történet rekonstrukciója

4/9

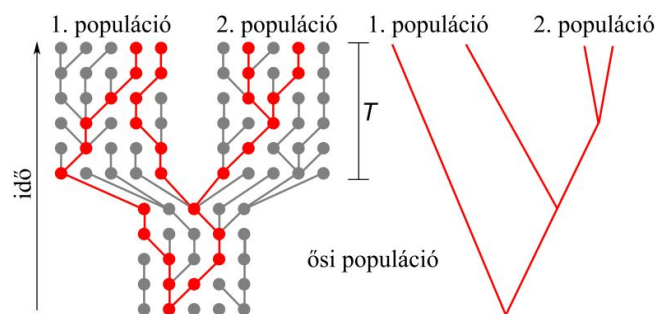
A filogenetikai rekonstrukciós módszerek elsősorban fajok leszármazási kapcsolatainak feltárására születtek, így a leszármazási sorok szétválása a fajképződés folyamatának és fajok divergenciájának felel meg. A leszármazás becslése gyakran DNS szekvencia adatokkal történik (markerek), feltételezve a divergenciát a közös ős megfelelő szekvenciájából (**ortológ** szekvenciák). Az így kapott fa a gén történetét tükrözi, amelyet felhasználunk a populáció vagy faj leszármazási kapcsolatok becslésére – a szekvenciát, szekvencia pozíciókat a populációba, fajba tartozó egyedek tulajdonságának tekintve. Vagyis a szekvenciák változatosságán és a változást leíró modelleken alapuló **génfa** segítségével a populációk, fajok leszármazására is következtethetünk – a **fa topológiájára** és **divergencia időkre**.

Ha kicsi az ősi populáció effektív mérete (N_e) és nagy a divergencia idő (T), azt várjuk, hogy ez utóbbi a koaleszcencia idővel (t) megközelítőleg azonos, így T a szekvenciákkal becsülhető.

A fajfa és génfa topológiája mindig azonos (mintától függetlenül), ha mindkét faj egyedei monofiletikus csoportot alkotnak, vagyis utolsó közös ősök (MRCA) a fajra kizárólagos – ekkor **reciprok monofiletikusak**. Az ábra ezt populációkra, populációként 2-2 elemű **mintára** szemlélteti, $t < T$ külön mindkét populációban a mintákra vonatkozóan.

Génfa

- Becslés szekvenciák alapján $\rightarrow t$
 - minta több fajból: $t > T$ – divergencia idő túlbecslése
 - kis T és az ősi populációra nagy N_e – jelentős az eltérés
 - t az ősi genetikai változatosságának függvénye
- Becslés szekvenciák alapján $\rightarrow$ génfa
 - az ősi populáció változatossága hatással lehet a topológiára is
 - főleg kis T és nagy populáció méret (nagy t) esetén
 - leszármazási sorok nem teljes rendeződése

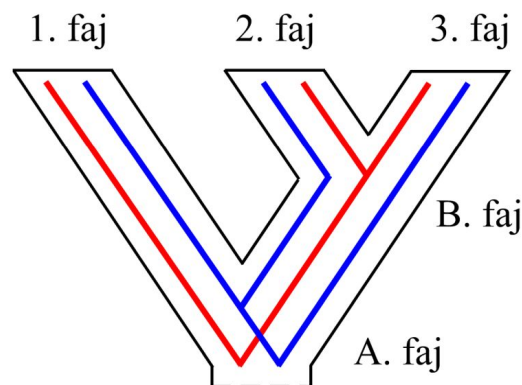


Ha a minta elemei különböző fajokból származnak, a minta utolsó közös őiséig eltelt idő biztosan nagyobb a fajok divergencia idejénél ($t > T$). Szekvenciákból a koaleszcencia időt becsüljük, nem a fajok divergencia idejét. A kettő eltérése mértéke, mennyire becsüljük túl a faj divergencia időt a szekvenciákból számolt koaleszcencia idővel, az ősi faj effektív populáció méretének és a divergencia időnek a függvénye: minél kisebb N_e az ősré és nagyobb T , annál nagyobb az eltérés. t az ősi genetikai változatosságának függvénye. Különösen közeli rokon fajokra a t és T elkülönítése ezért fontos lenne az eredmények interpretálása során.

Az ősi változatosság nem csak a divergencia idő becslésében jelenthet problémát, hanem hatással lehet a fa topológiájára is. Reciprok monofiletikusság esetén a minta koaleszcenciája a divergencia előtt következett be (a jelenből kiindulva). Azonban ez nem szükségszerűen van így, különösen ha nagy a populáció és T kicsi – több leszármazási sor is fennmaradhat a populáció divergenciától függetlenül. Vagyis előfordulhat, hogy egy különböző fajokból vagy populációkból származó **mintára** a koaleszcencia idő kisebb, mint az egy fajból vagy populációból származó mintára. Ezt a **leszármazási sorok nem teljes rendeződésének** nevezzük.

Génfa

- Leszármazási sorok nem teljes rendeződése: valódi génfa és fajfa eltérhet
- Több faj – koaleszcens fa és fajfa
 - divergencia idők és effektív populáció méretek függvénye
 - ha koaleszcencia a 2 és 3 faj közös őseben (*B* faj) – **azonos topológia**
 - ha nincs koaleszcencia *B*-ben (3 leszármazási sor az *A* fajban), de 1 és 2 leszármazási sor koaleszcenciája – **nem kongruens topológia**



Az ősi **leszármazási sorok nem teljes rendeződése** esetén az egy fajba tartozó egyedek egy részének egy másik faj egyes egyedeivel olyan közös őse van, ami nem közös őse a saját faj más egyedeivel, $t > T$. A valódi génfa nem biztos, hogy a fajok valódi leszármazási kapcsolatait tükrözi.

Ha a minta kettőnél több fajból származik, még összetettebb lehet az eredmény, különösen ha kicsik a belső ághosszak és nagyok az effektív populáció méretek. Az ábrán például 2. és 3. jelű faj leszármazási sorának **koaleszcenciája a közös őseben** biztosítja, hogy a koaleszcencia fa topológiája azonos a fajfáéval. Ha ez a **koaleszcencia nem következik be**, viszont 1 és 2 faj leszármazási sorára ez megtörténik (majd ezt követően bekövetkezik ez utóbbiak közös őse és 3 leszármazási sora között is), a koaleszcencia fa és fajfa topológia nem kongruens.

A **nem kongruens eredmény** valószínűsége a koaleszcencia elmélet alapján kiszámítható. Például egy génre, három fajra a standard modell feltételei mellett, ha *B*-ben nem történt koaleszcencia, akkor *A*-ban három lehetőség van: 1 – 2, 1 – 3 és 2 – 3 leszármazási sorok koaleszcenciája, azonos valószínűséggel. Ebből 2 nem kongruens eredményt ad, így ennek valószínűsége $2/3$.

Génfa

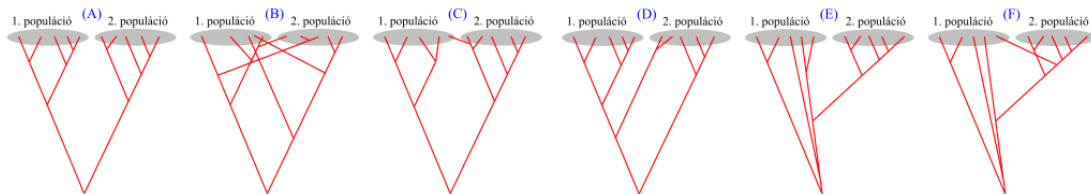
- Nem kongruens eredmény valószínűsége?
 - standard koaleszcencia modell, 1 szekvencia/populáció (faj)
 - nincs koaleszcencia t időn keresztül (időt $2N$ generációban mérve): e^{-t}
 - legyen B populáció mérete $2N$ (konstans), ághossza $2N\tau$
- Ha nincs 2 – 3 koaleszcencia B perzisztencia ideje alatt – valószínűsége: $e^{-\tau}$
 - ez esetben A -ban 3 koaleszcencia történhet elsőként azonos valószínűséggel
 - de csak 2 – 3 esetén azonos kongruencia
- Génfa és fajfa nem kongruens (3 faj, 1 gén) – valószínűsége: $2e^{-\tau}/3$
- Több gén: rekombináció szerepe – egyes génekre kongruens, másokra nem
- Kis τ esetén a mintázás jelentősége...
- Nem kongruens fák egyéb okokból is: becslési bizonytalanságok (kevés adat), horizontális géntranszfer (génáramlás következtében)

A standard koaleszcencia modell alapján annak a valószínűsége, hogy nem történik koaleszcencia B perzisztencia ideje alatt számolható. Ha B populáció mérete nem változik és τ a populáció mérettel átskálázott ághossz, ez $e^{-\tau}$. Így a **nem kongruens eredmény** valószínűsége 3 fajra fajonként 1 lokuszon $2e^{-\tau}/3$. Vagyis kis ághossz ($2N\tau$) esetén a topológia jelentős mértékben függhet attól, hogy milyen, a fajt reprezentáló egyedeket és géneket mintázunk. Ez utóbbi azért lényeges, mert a rekombináció következtében a különböző lokuszok koaleszcencia fája eltérhet – a genom egy része a fajfával kongruens, másik része nem kongruens eredményt adhat.

Például az **ember leszármazási kapcsolataira** irányuló DNS szekvenciákon alapuló molekuláris filogenetikai eredmények kezdeti bizonytalanságának egyik oka a nem teljes leszármazási sor rendeződés lehetett. A gorilla-csimpánz-ember közös ős csimpánz-ember közös ős faj leszármazási sor nagyon rövid. A fentihez hasonló módon a magi genom alapján becsült génfák 2/3 része kongruens a fajfával (csimpánzok az ember testvércsoportja). A magi genom 1/6 része alapján az ember-gorilla, 1/6 része alapján a csimpánz-gorilla együtt alkot monofiletikus csoportot. Feltételezhető továbbá, hogy a gorillák és az ember-csimpánz leszármazási sor között génáramlás is lehetett a kezdeti divergenciát követően (horizontális géntranszfer).

A koaleszcens fa értelmezése

- Különböző történeti és demográfiai modellek → eltérő fák
- Koaleszcencia fák értelmezése (koaleszcencia események) – példák
 - Ⓐ nagy divergencia idő, nincs (csekély) génáramlás
 - Ⓑ recens divergencia, intenzív génáramlás
 - Ⓒ nagy divergencia idő, recens génáramlás
 - Ⓓ nagy divergencia idő és génáramlás vagy kis divergencia idő és ősi leszármazási sorok nem teljes rendeződése
 - Ⓔ nagy divergencia idő és génáramlás, vagy 2. populációra palacknyak, kisebb N_e
 - Ⓕ recens divergencia, génáramlás, ősi leszármazási sorok nem teljes rendeződése, 2. populációra kisebb N_e ...



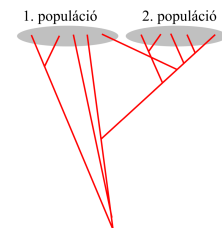
Populáció történet rekonstrukciója

8/9

Következő lépés a **koaleszcencia fák populációgenetikai értelmezése**. A fán az ághosszak a koaleszcencia időt szemléltetik, míg a csomópontok a koaleszcencia eseményekre utalnak. Mivel a koaleszcencia idő az effektív populáció méretek függvénye, a fa alakja függ a populáció demográfiától is. Vagyis a különböző **történeti és demográfiai modellek** alapján eltérő fákat várunk, így a fa alapján a történetre és a demográfiára következtethetünk. Ez néha egyértelműnek tűnik ((A-C) ábra), de sokszor nem triviális ((D-F)). Például, amennyiben a divergencia idő (T) nagy és a divergencia kezdetétől megszűnik a populációk közötti génáramlás, a populációk (reciprok) monofiletikussá válnak, az (A) ábrához hasonló koaleszcens fát várunk. Recens divergencia intenzív génáramlással a (B) ábrához hasonló koaleszcens fát eredményez, míg jóval korábbi divergencia a génáramlás megszűnésével majd később ismét génáramlással a (C)-hez hasonlíthat inkább. (D-F) kompatibilis lehet olyan modellekkel, ahol a génáramlás tartós lehetett és egy (néhány) leszármazási sor fennmaradt mindkét populációban, de hasonló mintázatokat kaphatunk génáramlás nélkül is, például leszármazási sorok nem teljes rendeződésével ((D)), palacknyak hatással ((E)). (E-F) esetén az 1. populációra azt várnánk, hogy N_e értéke nagyobb (t_{MRCA} nagyobb).

A koaleszcens fa értelmezése

- Demográfiai történet hatása a fa alakjára → fa alapján következtetés a demográfiára
- Adott fa valószínűsége egy adott modellel kifejezhető (történet és demográfia) – döntés ez alapján
- Génfák populációkra – nem triviális
 - egy génfa (rekombináldó egység, pl. mtDNS) nem elegendő
 - különböző populáció történeti modellek hasonló koaleszcencia fát eredményezhetnek
 - koaleszcencia sztochasztikus jellege – adott demográfiával különböző fák
- Példa: ember mtDNS génfával kompatibilis koaleszcens fa („out of Africa”)
 - 1. populáció Afrika, 2. Afrikán kívüli
 - MRCA az afrikai változatosságban
 - de egyéb modellekkel is kompatibilis lehet



Például (F) koaleszcens fa kompatibilis lehet a **humán evolúció** „out of Africa” hipotézisével (melyre egyéb bizonyítékok is vannak), azonban a fa a génáramlást feltételező multiregionális hipotézissel is magyarázható (Afrikára nagy N_e feltételezésével).

A következtetés a fa alapján tehát **gyakran nem egyértelmű**, melynek több oka is lehet. Láttuk például, hogy a különböző modellek hasonló fákat eredményezhetnek és egy adott demográfia és eredményezhet különböző fákat. Ezért egy **rekombináldó egység** (mtDNS, Y-kromoszóma, egy genom szakasz) alapján történő következtetés kockázatos. A demográfia alapú következtetéshez koaleszcens keret szükséges.

Ellentétben az egyszerű mérőszámokkal, a **fa alapú megközelítéssel** már lehetővé válik például a divergencia és migrációs modellek kombinálása, illeszkedésének vizsgálata az adatokra. Hátránya azonban többek között a fa becslésének bizonytalansága – a következtetés csak annyira biztos, amennyire a fa. További populációkra alkalmazható megközelítési módok is léteznek, amelyek például az adatok teljes információ tartalmára vonatkozó likelihood függvényen alapulnak.

Ellenőrző kérdések

- ❶ Milyen céllal használunk fa alapú módszereket a populációk genetikai adatainak elemzése során?
- ❷ Melyek a történeti rekonstrukció fő módszerei?
- ❸ Milyen feltételei sérülhetnek a filogenetikai rekonstrukciónak populáció adatokra történő alkalmazása során?
- ❹ Milyen elemei vannak a koaleszcens fának?
- ❺ Milyen okai lehetnek a valódi és tapasztalt genetikai távolság eltérésének?
- ❻ Mi a különbség a koaleszcencia idő és a divergencia idő között?
- ❼ Mikor várjuk, hogy a koaleszcencia idő és a divergencia idő hozzávetőleg azonos?
- ❽ Miért térhet el egymástól a génfa és fajfa topológiája?
- ❾ Mit értünk leszármazási sorok rendeződése alatt?
- ❿ Miért függ a demográfiától a koaleszcens fa alakja (ághossz, topológia)?



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