RÖVID SEGÍTSÉG EGYETEMI HALLGATÓKNAK EVOLÚCIÓBIOLÓGIAI TÉMÁJÚ KUTATÁSAIK ELKEZDÉSÉHEZ, 2.

STATISZTIKAI ELEMZÉSEK, ELTERJEDÉSI TERÜLET REKONSTRUKCIÓ ÉS EGYÉB PROBLÉMÁK

Nagy Jenő
Biológus MSc II. évf.
jenonagy.off@gmail.com
Témavezető: Tökölyi Jácint, egyetemi tanársegéd

Jelen tanulmány egy korábban megjelent munka (Nagy, 2012) folytatásának tekinthető. Az előző tanulmány egy konkrét vizsgálatban felhasznált módszereket ismertetett. Most azonban szeretnék egy általánosabb képet adni, hogy mely statisztikai elemzéseket milyen természetű adatokra lehet használni és mire kell nagyon odafigyelni alkalmazásuk során. A tanulmány második részében pedig az ősi elterjedési területek rekonstrukciójának problémáival, megvalósíthatóságával és lehetséges módjaival foglalkozom a teljesség igénye nélkül *(nem matematikai megközelítésből)*.

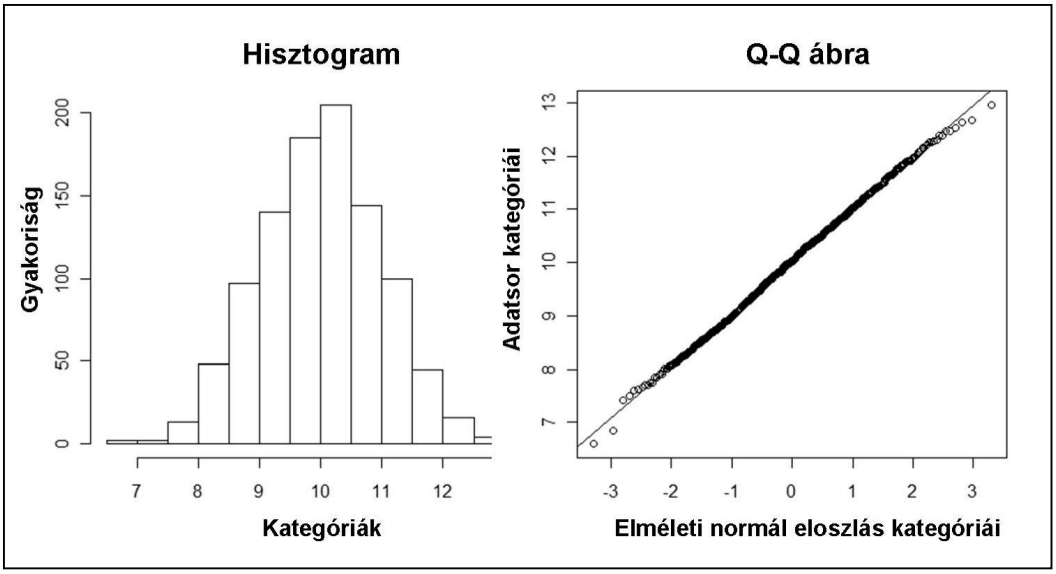
Mielőtt belevágnék a részletekbe, ajánlom az olvasók figyelmébe a következő könyveket, melyekben részletesebben utána tudnak nézni a szóba kerülő módszereknek, programoknak: Biostatisztika nem statisztikusoknak (Reiczigel et al, 2010), The R Book (Crawley, 2007). Az *R* programozási nyelv honlapján minden csomaghoz külön útmutató és számos egyéb segítség található (http://www.R-project.org). A tanulmányban ismertetett általánosan elterjedt metódusok mellett megpróbálom feltüntetni a megfelelő forrásokat, viszont a fentebb említettek mindenképp jó kiindulópontot nyújtanak.

Engedjenek meg egy rövidke emlékeztetőt az *R* programozási nyelvről (R Developement Core Team 2008). Csomagalapú, dinamikusan fejlődő, ingyenes programozási nyelv, mely adatok kezelésére, statisztikai elemzések elvégzésére, egyszerűbb, illetve bonyolultabb modellek készítésére, egyéb számításokra, analízisekre is kiválóan alkalmas. Más statisztikai programokkal szemben előnye, hogy az eredmények kiértékelése egyszerű, minden rendszer alatt futtatható, hátránya lehet, hogy nincs teljes körű grafikus felülete. Az *RStudio* szoftver *(http://www.rstudio.com/)* integráltan tartalmazza az alap *R* programot, különböző grafikus funkciókkal ellátva mind *Mac*, *Linux* és *Windows* környezetben egyaránt.

Most nézzük meg a legegyszerűbb statisztikai elemzéseket, melyeket az *R* alap verziójában, beépítve elérhetünk. Fontos, hogy mielőtt bármit is elvégeznénk, előbb az adataink természetével kell tisztában lennünk. A meg-

felelő próba, modell kiválasztása érdekében érdemes tudnunk, hogy milyen az adatok eloszlása, szórása és varianciája.

Az egyszerűség kedvéért itt most az eloszlás vizsgálatának legegyszerűbb módjára térnék ki, a normál eloszlás megállapítására. A normál eloszlást leíró függvények képileg haranggörbék. A haranggörbe középen csúcsos, vagyis a közepes kategóriákba tartozó elemek gyakorisága nagyobb, mint az alacsonyabb, vagy magasabb kategóriákba tartozó elemeké.



1. ábra: Hisztogram és q-q ábra a normál eloszlás megállapításához. Az ábrák ezer, random módon generált szám eloszlását mutatják.

A normál eloszlás meghatározásának alapvetően kétféle módja lehetséges grafikus és statisztikai próbával tesztelt. A grafikus megközelítésen belül két ábratípust érdemes megkülönböztetni. Az egyik az oszlopdiagram *(hisztogram)*, míg a másik a q-q ábra *(quantile-quantile plot)*. Mindkettőnek meghatározott képe van, mely alapján eldönthető, hogy a vizsgált változónk normál eloszlást követ-e, vagy sem. A hisztogram esetében az oszlopokra illeszthető képzeletbeli haranggörbe mutatja a normál eloszlást. A q-q ábra esetében pedig az egyenesre nagymértékben illeszkedő pontfelhő alapján igazolhatjuk a normál eloszlást (1. ábra). Gyakran az ábrák nem elegendőek az egyértelmű döntéshez, ezért többnyire érdemes statisztikai próbával is alátámasztanunk a választ. Erre szolgál az *R* környezetben a *Shapiro-Wilk-féle* teszt (Royston, 1982). A p-érték alapján eldönthetővé válik, hogy normál eloszlású-e az adatsorunk. A nullhipotézis ennél a próbánál a következő: az adatsorunk megegyezik az elméleti normál eloszlással. Azonban a próba a nagyon kis mintaszám esetében hibás eredményt adhat.

Mivel már tudjuk, hogyan lehet eldönteni, hogy adataink normál eloszlást követnek-e, ezért lássuk röviden a különböző típusú próbákat, modelleket. A statisztikai elemzések egyik nagy csoportjába a parametrikus próbák tartoznak. Az ilyen jellegű próbák, egyik gyakori feltétele az adatok normál eloszlása. Ellenben a nem normál eloszlású adatokat nem parametrikus próbákkal lehet értékelni, ahol nem történik paraméterbecslés. A továbbiakban néhány egyszerű modellel fogunk megismerkedni. Felépítésük közös jellemzője, hogy van egy függő változónk, az a változó, melyet vizsgálni szeretnénk, illetve van egy vagy több magyarázó változónk. A függő és magyarázó változók típusa, eloszlása határozza meg, hogy éppen melyik modellt alkalmazhatjuk.

Az egyszerű lineáris regresszió esetében fontos, hogy a függő változónk folytonos legyen, és a reziduálisok *(adatpontok és a regressziós egyenes közötti függőleges távolság)* normál eloszlást kövessenek. Egyetlen magyarázó változóval vetjük össze, mely szintén folytonos. Gyakorlatilag a két változó közötti lineáris kapcsolat feltérképezésére szolgál. Ebben az esetben két paramétert becslünk: a kapcsolatot jellemző regressziós egyenes *(a pontokra legjobban illeszkedő egyenes)* meredekségét, illetve az Y tengelyen lévő metszéspontját. Ezek az értékek meghatározzák az összefüggés mértékét és irányultságát is. Abban az esetben, ha több magyarázó változónk van, többszörös regressziót hajtunk végre.

Következő, szintén egyszerű típus a varianciaanalízis *(ANOVA)*. Különbség az előző típushoz képest, hogy a magyarázó változónk értékei diszkrét kategóriákat alkotnak. Tulajdonképpen itt csoportokat hasonlítunk egymáshoz, így a becsült paraméterek az első csoport átlaga, illetve a további csoportok ettől az átlagtól való eltérései lesznek. Ha kombinálni szeretnénk a lineáris regressziót a varianciaanalízissel, akkor kapjuk meg a kovariancia analízist.

A fenti elemzéstípusok az általános lineáris modellek (bővebben lásd: Chambers, 1992) körébe tartoznak. Alkalmazhatóságuk feltétele a normál eloszláson kívül még a linearitás, a pszeudoreplikáció és multikollinaritás elkerülése, illetve a varianciák homogenitása.

Ha a függő változónk nem normál eloszlású, akkor az általánosított lineáris modelleket lehet alkalmazni (Hastie–Pregibon, 1992). Az ilyen modellekben a binomiális, Poisson, kvázi-Poisson és egyéb eloszlású adatsorokat is lehet elemezni. A pszeudoreplikációt pedig *(ha a mintavételnél elkerülhetetlen, pl.: egy folyóból több ponton vett minta)*, kevert lineáris modellek *(lme4 csomag R-ben; Bates et al, 2012)* alkalmazásával lehet kezelni. Ez utóbbi esetben a modell két részből áll. A fix változók közé sorolhatóak a függő változó és a magyarázó változók. Random faktorokként pedig be lehet épí-

ni a mintapontok közötti összefüggéseket *(pl.: mintavételi hely sorszáma)* okozó tényezőket.

A több fajjal dolgozó komparatív *(összehasonlító)* vizsgálatokban többnyire olyan modelleket alkalmazunk, melyekben a filogenetikai kapcsolatokra lehet kontrollálni *(pl.: Bayes-i filogenetikai modell, MCMCglmm csomag R-ben; Hadfield – Nakagawa, 2010)*. Erről már volt szó a korábbi tanulmányban, ezért itt csak megemlítem.

A törzsfát azonban nem csak statisztikai elemzésekhez, hanem elterjedési terület rekonstrukciójára is fel lehet használni. A filogeográfia az új informatikai hardverek és szoftverek megjelenésével szintén dinamikusan fejlődő tudományág. Az ősi elterjedési területek rekonstruálásának segítségével a növény és állatfajok elterjedésének, evolúciójának, egymáshoz való viszonyának vizsgálata új szemszögből lehetséges.

Yu és munkatársai nemrég fejlesztettek ki egy új programot, amely minden operációs rendszer alatt hatékonyan képes rekonstruálni az ősi elterjedési területeket, egyidejűleg megbecsülve a különböző események valószínűségeit *(a felhasznált módszertől függően)*. A *RASP (Reconstruct Ancestral State in Phylogenies; http://mnh.scu.edu.cn/soft/blog/RASP;* Yu et al, 2013) három különböző rekonstrukciós módszert foglal magába: a *statistical dispersal-vicariance analysis (S-DIVA; Yu et al, 2010)*, a már korábban szóba került *lagrange* vagy más néven *dispersal-extinction-cladogenesis model (DEC; Ree–Smith, 2008)*, illetve egy harmadik módszer *(Bayesian binary method)* is elérhető.

Számos beállítást lehet alkalmazni az egyes módszerek futtatása előtt (bővebben lásd a használati útmutatóban; Yu et al, 2012). Azonban három alapvető fájl szükséges minden számítás elvégzéséhez. A programba be kell töltenünk a törzsfát. Ez lehet egyetlen törzsfa is, vagy több fából álló adatsor. Ezt követően meg kell adnunk a vizsgálandó fajok jelenkori elterjedési területeit, majd végül vagy magunk adjuk meg, több törzsfa esetén, a konszenzus fát vagy kiszámíttathatjuk a programmal. Fontos, hogy a fajok nevének mind a törzsfán, mind pedig az elterjedési terület adatokat tartalmazó táblázatban azonosnak kell lenniük. Az *S-DIVA* elemzést több fára is elvégezhetjük, míg a *DEC* modell csak a végső törzsfán futtatható. Tehát a jelenkori elterjedési területeket felhasználva lehet következtetnünk az ősi elterjedésekre, illetve a közben lejátszódó eseményekre úgy, mint diszperzió, vikariáns fajkeletkezés és kihalás. A törzsfa egyes csomópontjaiban lehetséges elterjedési területekhez valószínűség értékek kapcsolódnak, melyek alapján el tudjuk dönteni, honnan származhat, milyen módon alakult ki egy adott csoport vagy faj. Ha egyhez közelít a valószínűség érték, akkor szinte 100%-ban biztos, hogy az volt az ősi elterjedési terület.

Nézzük meg egy egyszerű példán keresztül, hogyan kell értelmezni a program által kiadott eredményeket. A számítások alatt minden beállítási lehetőség a program által megadott alapértéken volt. Tételezzük fel, hogy van 11 fajunk *(Species 01-11)*, melyek mindegyike 3 lehetséges területen *(A, B és C)* fordulhat elő (I. táblázat). Jelen esetben a 11 faj lehetséges törzsfáiból véletlenszerűen kiválasztva 100 darabot használtam fel az elemzéshez. Miután a száz törzsfa meg lett nyitva, illetve az elterjedési adatok is be lettek olvasva, lehetőség nyílik ellenőrizni, hogy minden fajnál van-e elterjedési adat, illetve helyesen lettek-e beírva. Ugyanezen táblázatban meg lehet adni azt a taxont, melyet külcsoportként kívánunk használni. Ezt követi a konszenzus törzsfa megnyitása. *(A konszenzus fa több törzsfa eredményeit összesíti egyetlen fán. Az összesítés alapja, hogy az egyes elágazások mekkora százalékban jelennek meg ugyanúgy a különböző fákon.)* Ennél az elemzésnél kész fát alkalmaztam, de ahogy említettem, lehetőség van összesített törzsfa készíttetésére is.

Fajnév	Elterjedés	Külcsoport
Species 01	A	
Species 02	A	
Species 03	A	
Species 04	AB	
Species 05	B	
Species 06	B	
Species 07	B	
Species 08	BC	
Species 09	A	
Species 10	ABC	
Species 11	C	+

1. táblázat: A példában használt kitalált fajok nevei és a hozzájuk tatozó elterjedési adatok. A „Külcsoport" oszlopban levő „+" jel mutatja melyik fajt használtam külcsoportként

A különböző módszerek eltérő eredményeket adhatnak (2. táblázat). A példában vizsgált tizenegy faj összesen 10 belső csomóponttal rendelkezik. A *Species 11* az összes többi faj testvércsoportja, így a legöregebb faj (2. ábra). A *Species 01* a *Species 02-10* fajok külcsoportja. A *Species 02-06* fajokat magában foglaló klád testvércsoportja a *Species 07-10* fajokat összefoglaló kládnak (2. ábra). Jól látszik az is, hogy ez utóbbi csoport fiatalabb.

CSOMÓ-PONT	MÓDSZER	LEGVALÓSZÍNŰBB ŐSI ELTERJEDÉS	LEJÁTSZÓDÓ ESEMÉNYEK*
12	S-DIVA	A	diszperzió
	DEC	AB	diszperzió
13	S-DIVA	A	diszperzió (2) vikariancia
	DEC	AB	diszperzió
14	S-DIVA	AB	viakriancia
	DEC	AB	diszperzió
15	S-DIVA	AB	viakriancia
	DEC	A	diszperzió
16	S-DIVA	B	diszperzió
	DEC	B	diszperzió
17	S-DIVA	B	diszperzió (2) vikariancia
	DEC	AB	vikariancia
18	S-DIVA	AB	diszperzió (2)
	DEC	ABC	diszperzió (2)
19	S-DIVA	B	diszperzió (2)
	DEC	A	diszperzió (2)
20	S-DIVA	AB	vikariancia
	DEC	A	
21	S-DIVA	ABC	vikariancia
	DEC	AC	vikariancia

* ahol szükséges, zárójelben az események száma
nincs meghatározható esemény

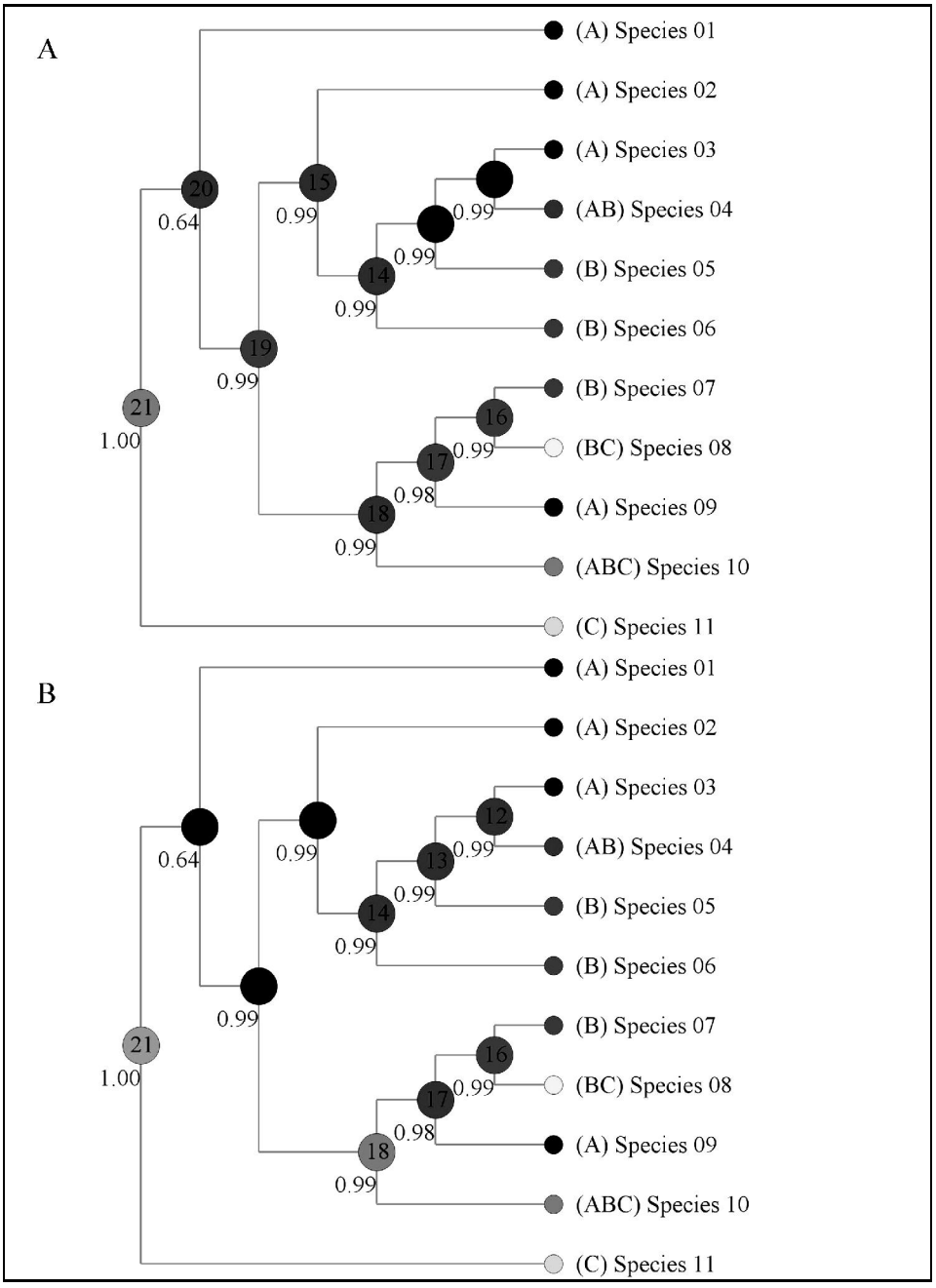
2. táblázat Az elterjedési terület rekonstrukciók eredményei csomópontonként lebontva

A 2. ábrát felhasználva a II. táblázatban összefoglalt eredmények is könnyebben értelmezhetőek. Sajnos a fekete-fehér nyomtatás miatt nem különülnek el megfelelően a színek egymástól. Azt azonban még így is el lehet mondani, hogy valószínűleg a szóban forgó 11 faj közös őse *(21. csomópontban)* nagy valószínűséggel mindhárom jelenlegi elterjedési területen előfordult.

Ha ez egy valós vizsgálat lenne, akkor elképzelhető, hogy olyan földtörténeti korra tehető, mikor a kontinensek még egyben voltak. Így érthetővé válik a fiatalabb csomópontokban megjelenő viakriáns fajkeletkezés *(az egykor egységes elterjedési terület feldarabolódása)*. Majd a *19. csomóponttól* fiatalabb pontokban már túlnyomó részt a diszperziós folyamatok lesznek jelentősebbek, létrehozva azt a 11 különböző fajt az evolúció során, melyek most a fa ágainak végén szerepelnek.

A fenti példához hasonló elemzés kivitelezése száz-kettőszáz faj adatait felhasználva, több lehetséges területtel szinte egészen a *RASP* kifejlesztéséig körülbelül egy hetet vett igénybe még nagyteljesítményű számítógépe-

ken is. Száz darab futtatás esetében, mely csak lineárisan oldható meg, több év telt volna el az eredmények kézhezvételéig. Ezt a problémát az új program már megoldotta, egy-egy számítás csupán néhány percet vesz igénybe. Azonban továbbra sem lehet egyidejűleg több fára alkalmazni a *DEC* modellt. Felmerülhet a kérdés, hogyan lehet megbízhatóan összesíteni a különböző fákra kapott eredményeket. Ugyanis kézzel, csomópontonként „átlagolva" igen hosszadalmas lenne. Egy alternatív megoldás lehet, hogy a több törzsfára elkészített *S-DIVA* elemzést összevethetjük a *DEC* modell eredményeivel, mintegy megerősítve azokat.



2. ábra: (A) S-DIVA eredményei. (B) DEC eredményei

Néhány technikai jellegű probléma is felmerül a programmal kapcsolatban. A *DEC* modellen alapuló számításokat csak akkor tudja elvégezni a program, ha egyszerű elérési utat biztosítunk neki. Ne szerepeljenek ékezetek, különleges karakterek a mappák neveiben *(például ezt az elérési parancsot nem képes helyesen kezelni: D:\Kutatás\Ősi elterjedés\RASP\)*. Ezt úgy lehet megoldani legegyszerűbben, ha a program könyvtárát az operációs rendszer meghajtójának gyökérkönyvtárába tesszük *(pl.: C:\RASP\)*. További érdekesség, hogy habár minden operációs rendszeren lehet futtatni a *RASP*-ot, alapjában véve *Windows*-ra van írva, így *Linux* alatt is emulált *Windows* felületen keresztül lehet használni, ami újabb bonyodalmakat jelenthet. Ezért azt javaslom, használjuk *Windows* rendszer alatt *(Windos XP és Windows 7 rendszereken tökéletesen működik - teszteltem)*.

Természetesen számos előnyös tulajdonsággal rendelkezik a *RASP* program. Először is grafikus felülettel rendelkezik, így nem kell tudnunk parancsokat a kezeléshez. Ebből kifolyólag egyszerűen kezelhető, az adatsorok néhány kattintással betölthetőek, jól átláthatóan ellenőrizhetőek, a beállítási lehetőségek egyértelműek. Az elemzésre fordított idő lényegesen lecsökkent a korábbi verziókhoz képest. Az eredményeket szövegfájlban is kinyerhetjük, de van lehetőség grafikus ábrázolásra, mely nagyban megkönnyíti azok értelmezését.

Mindent egybevetve nagyon körültekintően kell alkalmaznunk ezeket a módszereket. Nagyon sokszor azért nem működnek az elemzések, parancssorok, mert gépelési hiba van benne, vagy nem egyeznek meg a különböző adatsorok fajlistái, változó nevei. Mindig ellenőrizzük le őket. Fontos, hogy az eredmények értékelésénél csak olyan következtetéseket vonjunk le, amelyeket valóban alátámasztanak az adatok is. Mielőtt nagyobb terv kidolgozásához fognánk, érdemes egy kisebb számítást végezni, hogy kiderüljön, jól tettük-e fel a kérdéseinket, megfelelően fogalmaztuk-e meg hipotéziseinket, illetve egyáltalán érdemes-e a témával tovább foglalkoznunk. Amennyiben igen, lehet bővítenünk fajokkal, vagy újabb változókkal vizsgálatunkat.

FELHASZNÁLT IRODALOM

Bates, D.–Maechler, M.–Bolker, B. (2012): *Linear mixed-effects models using S4 classes.* http://cran.r-project.org/web/packages/lme4/ – letöltve: 2013. június 16.

Chambers, J. M. (1992): Linear models. In: J. M. Chambers–T. J. Hastie (szerk.): *Statistical Models in S.* Wadsworth & Brooks, Cole

Crawley, Michael J. (2007): *The R Book.* Wiley, Hoboken.

Hadfield, J. D.–Nakagawa, S. (2010): General quantitative genetic methods for comparative biology: phylogenies, taxonomies, and multi-trait

models for continuous and categorical characters. *Journal of Evolutionary Biology,* 23, 494–508.

Hastie, T. J.–Pregibon, D. (1992): Generalized linear models. In: J. M. Chambers - T. J. Hastie (szerk.): *Statistical Models in S.* Wadsworth & Brooks, Cole.

Nagy, Jenő (2012): Rövid segítség egyetemi hallgatóknak evolúcióbiológiai témájú kutatásaik elkezdéséhez. Néhány módszer rövid áttekintése egy korábbi vizsgálatunk alapján. In: Dávid Ágnes–Jován Katalin (szerk.): *„A mi tendenciáink..." Szakkollégiumi tanulmányok I.* Debrecen.

R Development Core Team (2008): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna. Available: http://www. R-project.org – letöltve: 2013. július 19.

Ree, R. H.–Smith, S. A. (2008): Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis, *Systematic Biology,* 57, 1, 4–14.

Reiczigel, Jenő–Harnos, A.–Solymosi, N. (2010): *Biostatisztika nem statisztikusoknak.* Pars Kft., Nagykovácsi.

Royston, Patrick (1982): An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31, 115–124.

Yu, Yan–Harris, A. J.–He, X. J. (2010): *S-DIVA (Statistical Dispersal-Vicariance Analysis): a tool for inferring biogeographic histories.* Molecular Phylogenetics and Evolution, 56, 848–850.

Yu, Yan–Harris, A. J.–He, X. J. (2012): *A rough guide to RASP.* Available: http://mnh.scu.edu.cn/soft/blog/RASP – letöltve: 2013. július 14.

Yu, Yan–Harris, A. J.–He, X. J. (2013): *RASP (Reconstruct Ancestral State in Phylogenies) 2.1 beta*. http://mnh.scu.edu.cn/soft/blog/RASP – letöltve: 2013. július 15.