RÖVID SEGÍTSÉG EGYETEMI HALLGATÓKNAK EVOLÚCIÓBIOLÓGIAI TÉMÁJÚ KUTATÁSAIK ELKEZDÉSÉHEZ. NÉHÁNY MÓDSZER RÖVID ÁTTEKINTÉSE EGY KORÁBBI VIZSGÁLATUNK ALAPJÁN

Nagy Jenő
Biológus MSC I. évf.
jenonagy.off@gmail.com
Témavezető: Tökölyi Jácint, PHD

Az evolúcióbiológiai kutatások igen szerteágazóak, rengeteg módszert alkalmaznak, melyeknek megvannak a saját előnyeik és természetesen hátrányaik is. (Nem minden esetben használható ugyanúgy egy már korábban jól bevált metódus.)

Ebben a tanulmányban szeretnék egy rövid áttekintést adni azokról a módszerekről, melyeket kutatásaink során alkalmaztunk és alkalmazunk a *Debreceni Egyetem Viselkedésökológiai Kutatócsoport*jában egy-egy viselkedésbeli probléma evolúciós szemszögű feltárásához. Mivel egy dinamikusan fejlődő és napról napra változó tudományágról van szó, a teljesség igénye nélkül készült ez az írás, mindössze néhány példán keresztül szemléltetjük ezeket.

Nem terepi vizsgálatokról lévén szó, az ilyen témájú kutatások erősen teoretikai jellegűek. Megpróbálunk olyan kérdésekre válaszolni, melyeket sok esetben már nem tudunk megfigyelni a természetben. Ugyanis ma már csak az eredményeit látjuk azoknak a folyamatoknak, melyek egykoron, nem egyszer igen hosszú idő alatt játszódtak le. Alapvetően megpróbáljuk a kezdeti állapotokat rekonstruálni, feltérképezve a kialakulás körüli és a jelenlegi tulajdonságok között végbemehetett folyamatokat.

Egy terepbiológushoz hasonlóan az evolúcióbiológus is adatok gyűjtésével kezdi a munkáját. Mivel a kis mintaszám nem megfelelően mutatja meg az általunk vizsgált kérdés mögött meghúzódó problémát, így indokolt nagy adatbázisok felépítése. Nem is lehetne alacsony mintaszámmal dolgoznunk, hiszen nagyrészt összehasonlító *(komparatív)* vizsgálatokra kell gondolnunk, ha az evolúcióról beszélünk. Általában azt az általános vonalat követjük, hogy minél nagyobb a mintaszám, annál biztosabbak, pontosabbak saját eredményeink is.

Minél több információval, minél több fajjal és a hozzájuk tartozó tulajdonságokkal *(változóval)* dolgozunk, annál nagyobb az esélye az ún. *becsúszó* hibáknak. Az ilyen hibák többsége általában adatbeviteli *(gépelési)* hiba. Cél-

szerű a teljes adatbázis összeállítását ugyanannak a személynek elvégezni, ezáltal csökkentve a különböző személyektől származó eltérő hibatípusok mértékét. Ha elkészültünk az adatbázissal, akkor a legjobb, ha munkatársunk, akivel együtt dolgozunk a vizsgálaton, vagy egy kollegánk szúrópróbaszerűen kiválasztott adatok átnézésével ellenőrzi munkánk. Ez azért olyan fontos, mert a hibás adatok megnehezíthetik későbbi munkánk, vagy ami még rosszabb, téves következtetéseket eredményezhetnek.

Következő igen fontos részegysége az evolúcióbiológiai kutatásoknak az összehasonlítás alapjául szolgáló filogenetikai törzsfa elkészítése. A filogenetikai törzsfák jól mutatják a jelenkori fajok rokonsági viszonyait és lehetséges leszármazási vonalaikat. A törzsfa rekonstruálása alapvetően két úton történhet. Megépíthetjük magunk a fát a különböző adatokat *(genetikai, morfológiai tulajdonságokat)* felhasználva, a szükséges számítógépes program és rekonstrukciós módszer felhasználásával. Másik lehetőségként a legfrissebb irodalmi adatokat felhasználva rajzoljuk meg azt.

Egyik kutatásunk során a GenBank adatbázisból (http://www.ncbi.nlm.nih.gov/genbank/) 217 nappali ragadozó madár DNS szekvenciáját töltöttük le. A letöltött szekvenciák közül a citokróm-b, a rag-1 és a NADH-dehidrogenáz 2. alegységének szekvenciáját használtuk törzsfa rekonstruálására. Kiválasztásuk az alapján történt, hogy korábban alkalmasnak bizonyultak a ragadozó madarak filogenetikai kapcsolatainak vizsgálatára (pl. Lerner–Mindell, 2005) és jó a taxonómiai lefedettségük (mindhárom elérhető legalább 150 fajra). Természetesen egy részfolyamat elvégzéséhez időnként több segédprogramot is fel kell használni. A szekvenciák sorba rendezése Clustal X segítségével történt (Thompson et al., 1997). Magát a törzsfa-rekonstrukciót az MrBayes program hajtotta végre (Ronquist–Huelsenbeck, 2003). Az elemzés során tíz millió generációt alkalmaztunk, de az első egymillió generáció adatai nem kerültek felhasználásra *(„burn-in")*, és a mintavételezés 1000 generációnként történt. Az elkészült törzsfát a Mesquite program segítségével öntöttük formába (Maddison–Maddison, 2008).

Ha összegyűjtöttük az adatokat és elkészítettük a törzsfát, következhetnek a tényleges elemzések. A biológiai vizsgálatoknál használatos a modellezés, statisztikai feldolgozás, jól bevált az R programozási környezet. Mi is ezt a nyelvet alkalmazzuk kutatásainkban. Viszonylag könnyen elsajátítható és nagy léptékben fejlődő, képlékeny programozási nyelv, mely a legtöbb operációs rendszeren elérhető. (Legstabilabban talán Linux környezetben fut, saját tapasztalat alapján.)

Az R csomagalapon működik, ami azt jelenti, hogy egyes funkciók eléréséhez telepítenünk kell a használni kívánt függvényt tartalmazó csoma-

got. A törzsfa-rekonstrukcióhoz szükséges molekuláris adatok letöltésének automatizálásához a „seqinr" és „ape" csomagokat használtuk (Paradis–Strimmel, 2004; Charif–Lobry, 2007; R Development Core Team, 2008).

A továbbiakban néhány olyan technika rövid leírása következik, melyekhez felhasználtuk a fentebb említett törzsfát, a filogenetika *(rokonsági)* kapcsolatokra kontrollálva.

A lagrange módszer (Ree et al, 2005; Ree–Smith, 2008) alkalmas arra, hogy a jelenkori fajok elterjedéseiből visszakövetkeztessünk az ősi elterjedésre. A diszperziós és kihalási eseményeket modellezi egy filogenetikai törzsfán „maximum likelihood" eljárással. Ezáltal lehetségessé válik a ma élő fajok őseinek feltételezett elterjedési területének meghatározása, anélkül, hogy egy specifikus fajképződési mechanizmust feltételeznénk (diszperzió és vikariáns fajkeletkezés egyaránt lehetséges). A program minden egyes csomópontban *(ősnél)* meghatározza az összes, modell alapján lehetséges elterjedési területet és ezekhez hozzárendel egy relatív valószínűséget. Minél nagyobb ez az érték, annál biztosabbak lehetünk a rekonstruált elterjedés pontosságában (1-hez közeli értékek esetén majdnem 100%-os a bizonyosság, míg 0 közeli értékek esetében több alternatív lehetőség van, melyek közül nem tudunk egyértelműen egyet kiválasztani).

Ahogy az előbb láttuk, többségében valamilyen mérőszám *(„valószínűség" érték)* alapján feltételezhetjük, hogyan is játszódtak le az evolúciós folyamatok. Megfigyelhető ez a következő néhány metódusnál is. Maradjunk továbbra is a vonulás evolúciójának vizsgálatánál.

A különböző vonulási stratégiák evolúciójának vizsgálatára a BayesTraits program MultiState alegységét használtuk (Pagel et al, 2004). Ebben az esetben az egyes vonulási kategóriák egymásba alakulásának változásaira kell gondolni *(nem vonuló–részlege vonuló–teljes vonuló)*. A MultiState egy több kategóriából álló jelleg ősi állapotait rekonstruálja egy törzsfán és eközben megbecsüli a tranzíciós rátákat, tehát, hogy milyen „gyakorisággal" alakul ki evolúciósan az egyik állapotból egy másik (itt azért meg kell említenem, hogy a gyakoriság kifejezés nem teljes mértékben azonos a tranzíciós rátával, de példaértéke könnyebben érthető). Ezáltal meg lehet határozni, hogy a hosszú távú vonulás *(teljes vonulás)* a nem vonuló vagy inkább a részlegesen vonuló állapotból fejlődött ki. A tranzíciós ráták becslését „maximum likelihood" módszerrel végzi el ez az analízis.

Az evolúciós változások irányának meghatározásához diszkrét teszteket alkalmaztunk (Pagel–Meade, 2006). A diszkrét teszt során úgy lehet becsülni a vonulási viselkedésben megfigyelhető evolúciós átmenetek rátáit, hogy közben figyelembe vesszük egy másik jelleg állapotát is; ezáltal meghatározható, hogy függ-e a vonulás kialakulása vagy eltűnése a másik jelleg

állapotától (vagy fordítva). A diszkrét tesztekhez a BayesTraits program Discrete alegységét használtuk (Pagel–Meade, 2006), a tranzíciós ráták becslése az előzőhöz hasonlóan történt.

Végül az említett vizsgálatban a vonulást meghatározó tényezőket Bayes-i filogenetikai, többváltozós, kevert modellekkel elemeztük (MCMCglmm csomag R-ben; Hadfield–Nakagawa, 2010). Ez a módszer figyelembe veszi, hogy a fajok a közös eredet miatt nem függetlenek egymástól, és a modell felépítésében kontrollálni lehet a törzsfára. Ugyanakkor nemcsak folytonos, hanem pl. bináris függő változókon is alkalmazható. Ugyan úgy működik a modell felépítése, mint az egyszerűbb lineáris, általánosított lineáris modellek esetében. Létrehozunk egy teljes modellt, amely magyarázó változóként az összes lehetséges ökológiai tényezőt tartalmazza, majd ebből a minimálisan megfelelő modellt (Crawley, 2007). A minimális modell a teljes modell redukálásával keletkezik. Több lehetőség is van ilyen modell létrehozására. Mi egy általánosan elfogadott redukálást alkalmaztunk: p-érték *(szignifikancia szint)* alapján, lépésenként hagytuk el a legkisebb értékkel rendelkező változókat, míg el nem jutottunk a megfelelő modellig.

Fentebb használt két fogalom, melyekről talán érdemes még néhány szót szólni. A „maximum likelihood" módszer az evolúciós modell, alapján az adatokból elkészíti az összes lehetséges variációt *(pl. törzsfát)* és valószínűségi értékkel jellemzi, mely alapján el tudjuk dönteni, hogy elfogadjuk-e. Parametrikus becsléseknél széles körben elterjedt statisztikai módszer. A statisztikában továbbá általánosan használt mennyiség a szignifikancia szint *(p-érték)*. Annak a valószínűsége, hogy a nullhipotézisünk fennállása esetén a véletlen következtében a nullhipotézisnek legalább annyira ellentmondó mintát kapunk, mint a ténylegesen megfigyelt minta. Hogyha $p < 0.05$, akkor azt mondjuk, hogy az eredményünk szignifikáns, ellenkező esetben nem szignifikáns. Így eldönthetjük, hogy megtartjuk, vagy elvetjük a hipotézist.

Ezek alapján is látható tehát, hogy az evolúcióbiológiában, igen nagy szerepe van a statisztikának és a számítástechnikának. Eddig nem esett szó róla, de ezeknek az elemzéseknek a hátterében (tekintve a nagy adatmennyiséget) nagy teljesítményű számítógépek állnak. Éppen ezért ilyen méretű vizsgálatokat nem érdemes otthon elkezdeni. Egyrészt a háztartásokban gyakoribb áramszünetek, másrészt a kis teljesítményű számítógépek nagyban megnövelhetik az analízisek lefutásának idejét.

Továbbá még egy dologra szeretném felhívni a figyelmet. Mivel az általunk bemutatott és a fentiekhez hasonló technikák viszonylag könnyen alkalmazhatóak, képlékenyek, sőt ötvöződnek a számítógépek nyújtotta további lehetőségekkel, ezért fontos, hogy kutatásunk során az adatok által valóban meghatározott eredményeket közöljük. Ne használjuk úgy azokat,

hogy az adatokból az *„általunk várt" (előre megfogalmazott)* következtetéseket mutassák meg ellenben a valóban fennálló összefüggésekkel. Ugyanis a szakirodalomban előfordulnak olyan esetek, melyekben *„kiszínezett"* értékeket közöltek (a jobb eladhatóság érdekében), viszont újbóli elemzéssel sokszor teljesen más eredmények születhetnek.

A tanulmányban megjelölt források segítséget nyújthatnak az egyes módszerek teljesebb megértéséhez. Az evolúcióbiológiai kutatások iránt érdeklődők leginkább az evolúcióbiológiában használatos módszerek, programok működése iránt érdeklődnek.

FELHASZNÁLT IRODALOM

Charif, D.–Lobry, J. R. (2007): *SeqinR 1.0–2: a contributed package to the R Project for Statistical Computing devoted to biological sequences retrieval and analysis,* In: Bastolla, U.–Porto, M.–Roman, H. E.–Vendruscolo, M. (eds.): Structural Approaches to Sequence Evolution: Molecules, Networks, Populations, Biological and Medical Physics. Biomedical Engineering. Springer-Verlag, New York

Crawley M. 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.

Lerner, H. R. L.–Mindell, D. P. (2005): Phylogeny of eagles, Old World vultures, and other Accipitridae based on nuclear and mitochondrial DNA. *Molecular Phylogenetics and Evolution,* 1–20.

Maddison, W. P.–Maddison, D. R. (2008): *Mesquite: a modular system for evolutionary analysis.* http://mesquiteproject.org/ (Letöltve: 2012. március 14.)

Pagel, M.–Meade, A.–Barker, D. (2004): Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology,* 53, 673–684.

Pagel, M.–Meade, A. (2006): Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo, *The American Naturalist,* 167, 6, 808–825.

Paradis, J. C.–Strimmer, K. (2004): Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics,* 20, 289–290.

R Development Core Team (2008): R: A language and environment for statistical computing, Vienna. R Foundation for Statistical Computing, Available: http://www. R-project.org (Letöltve: 2012. március 14.)

Ree, R. H.–Moore, B. R.–Webb, C. O.–Donoghue, M. J. (2005): A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution,* 59, 11, 2299–2311.

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.

Ronquist, F.–Huelsenbeck, J. P. (2003): MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19, 1572–1574.

Thompson, J. D.–Gibson, T. J.–Plewniak, F., Jeanmougin, F.–Higgins, D. G. (1997): The CLUSTAL_X windows interface, flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research,* 25, 4876–4882.