

Protracted speciation revitalizes the neutral theory of biodiversity

February 17, 2010

**James Rosindell, Stephen J. Cornell, Stephen P. Hubbell,
Rampal S. Etienne**

1. James Rosindell, (Corresponding author) Institute of Integrative and Comparative Biology, University of Leeds, Leeds, United Kingdom LS2 9JT, James@Rosindell.org , +44 7968 278 436
2. Stephen J. Cornell, Institute of Integrative and Comparative Biology, University of Leeds, Leeds, United Kingdom LS2 9JT. S.J.Cornell@leeds.ac.uk, +44 113 343 2899
3. Stephen P. Hubbell, Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California 900954, USA. shubbell@eeb.ucla.edu , +1 310 206 8165
4. Rampal S. Etienne, Community and Conservation Ecology Group, Centre for Ecological and Evolutionary Studies, University of Groningen, BOX 14, 9750 AA Haren, The Netherlands. R.S.Etienne@rug.nl , +31 50 363 2230

Appendix A: Simulating the model

We simulated the model for the parameters of best fit, computed using the methods in box 1 of the main text conducting 4,000,000 repeats. An excellent estimate of the mean and standard deviation could be obtained from these simulations and could be plotted next to the empirical data in Figures 4 and 5 of the main text. We now proceed to describe the algorithm for these simulations. The simplest case is for a sample from a metacommunity with just point mutation speciation, this is shown in Algorithm 1. The next most complex case introduces protracted speciation and is shown below as Algorithm 2, note that Algorithm 2 calls Algorithm 1 as part of its routine. The most complex case includes both protracted speciation and dispersal limitation, this is shown in Algorithm 3 which in turn calls the earlier algorithms 1 and 2 as part of its routine.

The algorithms are based on the concept of coalescence, instead of simulating the system forwards in time until it reaches equilibrium, we simulate it backwards in time following the ancestry of each individual in the sample back to the speciation event that gives us the species identity of this individual. First let us consider the point mutation case where $\tau = 0$. We take a sample of size J from the metacommunity of size J_M and so are tracing K lineages backwards in time where initially $K = J$. On the previous time-step, one of the J_M individuals will die and be replaced, the probability of choosing one of our K lineages is $\frac{K}{J_M}$. This chosen lineage will then speciate with probability μ . If it speciates, we are no longer concerned with this lineage; if it doesn't speciate, then we might have a coalescence event (where two lineages have a common ancestor) with probability $(1 - \mu) \frac{K-1}{J_M-1}$. In the case speciation or coalescence, the number of lineages K will decrease, otherwise nothing changes. Consequently, we are able to always jump to the next event and then determine whether it was speciation or coalescence. The probability of speciation is $\frac{K\mu}{J_M}$ and the probability of coalescence is $\frac{K(1-\mu)(K-1)}{J_M(J_M-1)}$. The probability that the next event will be speciation is therefore $\frac{\left(\frac{K\mu}{J_M}\right)}{\left(\frac{K\mu}{J_M}\right) + \left(\frac{K(1-\mu)(K-1)}{J_M(J_M-1)}\right)} = \frac{\mu(J_M-1)}{\mu(J_M-1) + (1-\mu)(K-1)}$ and because $J_M \gg J \geq 1$ we can approximate this well with $\frac{\left(\frac{\mu J_M}{(1-\mu)}\right)}{\left(\frac{\mu J_M}{(1-\mu)}\right) + K-1} = \frac{\theta}{\theta + K-1}$. The algorithm attaches a number to each lineage (stored in a vector V) corresponding to the number of individuals

in our sample that are descendants of each lineage. When a lineage speciates, the number attached to it becomes the abundance of the species; when two lineages coalesce to form a single one, the attached numbers are added. The system is initialised with $K = J$ lineages each with the number 1 attached reflecting that at this stage we trace one lineage for each individual in the sample. Pseudo code is as follows.

Algorithm 1: Point mutation metacommunity sample

1. Input θ and J .
2. Define a vector V of length J with $V_i = 1$ for $i = 1, 2, \dots, J$ and define $K = J$, the number of lineages in the vector.
3. Define an empty vector W which will contain a list of species abundances when the algorithm terminates.
4. Pick a random number $0 \leq r \leq 1$ (to choose if coalescence or speciation will happen).
5. Pick a random entry V_i from the vector V (to choose which lineage this will happen to)
6. If $\left(r \leq \frac{\theta}{\theta + K - 1}\right)$, speciation happens: remove V_i from the vector V and append it to vector W (then go to 8.)
7. If $\left(r > \frac{\theta}{\theta + K - 1}\right)$, coalescence happens: pick a random entry V_j from vector V (where $j \neq i$) and set $V_i = V_i + V_j$ then remove V_j from the vector V (then go to 8.)
8. Set $K = K - 1$; now the vector V should have one less element.
9. If $(K > 0)$, we are not finished so go to 4.
10. If $(K \leq 0)$, the algorithm is complete and W contains the abundances of all species.

We wish to extend this basic coalescence algorithm to allow for the case $\tau > 0$. In order to do this we simply require the K sampled lineages to undergo coalescence (but not speciation)

for the most recent τ generations. This means allowing for coalescence without speciation before proceeding with the algorithm described above. We must therefore calculate the number of generations that pass between coalescence events in this model. Because speciation is not of concern here, the probability of coalescence with the next birth in the metacommunity is $\frac{K(K-1)}{J_M(J_M-1)}$. We need to know how many births there will be before the next coalescence event (after which K will be smaller and the probability will change). The distribution of births until the next coalescence event is just the distribution of waiting times between events of a Poisson process with probability $\frac{K(K-1)}{J_M(J_M-1)}$ and this distribution is hence an exponential distribution. The cumulative distribution function is $1 - \exp(-\frac{K(K-1)}{J_M(J_M-1)}t)$ so we can pick a random number $0 \leq r \leq 1$ and obtain $t_{(births)} = \frac{J_M(J_M-1)}{K(K-1)} \ln(r)$ when measured in births. To measure the waiting time to coalescence in generations we note that one generation is the same as J_M births and hence $t_{(generations)} = \frac{J_M(J_M-1)}{K(K-1)} \ln(r)$. Now we take the next step to measure t in units of coalescence time, where one unit of coalescence time is $J_M \gg 1$ generations $t_{(units\ of\ coalescence\ time)} = \frac{-\ln(r)}{K(K-1)}$. We keep drawing from this distribution until τ generations ($\tau' = \frac{\tau}{J_M}$ units of coalescence time) have passed and with each draw a coalescence is performed. The full algorithm is therefore described as follows.

Algorithm 2: Protracted speciation metacommunity sample

1. Input θ , J and τ' .
2. Define a vector V of length J with $V_i = 1$ for $i = 1, 2, \dots, J$ and define $K = J$, the number of lineages in the vector.
3. Define $t = 0$ the total number of units of coalescence time that have passed.
4. Pick a random number $0 \leq r \leq 1$ (to choose how many generations until the next coalescence event).
5. Set $t = t - \frac{\ln(r)}{K(K-1)}$.

6. If $\tau' > t$, pick two random entries V_i and V_j ($i \neq j$) from vector V , set $K = K - 1$ and $V_i = V_i + V_j$. Then remove V_j from vector V (then go to 3.)
7. If $\tau' \leq t$, perform '**Algorithm 1. Point mutation metacommunity sample**' from step 3. onwards which will give the species abundance distribution.

Dispersal-limited local community case

We have a census of a dispersal limited local community of size J . To obtain the species abundances for these J individuals, we again use coalescence. We trace the ancestry of K lineages as before, where initially $K = J$. With each time step there is a probability $\frac{K}{J}$ (initially 1) that we pick one of the lineages, which might then immigrate with probability m or coalesce with probability $(1 - m)\frac{K-1}{J-1}$. The calculations are analogous to the metacommunity case except that instead of a speciation-initiation rate μ we now have an immigration rate m . Lineages that immigrate are then passed to the metacommunity algorithm which accounts for them being drawn from a metacommunity species pool. The final algorithm therefore runs as follows.

Algorithm 3. Local community dispersal limited sample from a metacommunity with protracted speciation

1. Input θ , J , τ' and m .
2. Define a vector S of length J with $S_i = 1$ for $i = 1, 2, \dots, J$ and define $K = J$, the number of lineages in the vector.
3. Define an empty vector V which will contain a list of groups of individuals descending from each immigration from the metacommunity species pool.
4. Pick a random number $0 \leq r \leq 1$ (to choose if coalescence or immigration will happen).
5. Pick a random entry S_i from the vector S (to choose which lineage this will happen to).
6. If $\left(r \leq \frac{m(J-1)}{m(J-1) + (1-m)(K-1)}\right)$. speciation happens: remove S_i from the vector S and append it to vector V (then go to 8.).

7. If $\left(r > \frac{m(J-1)}{m(J-1)+(1-m)(K-1)}\right)$ coalescence happens: pick a random entry S_j from vector S (where $j \neq i$) and set $S_i = S_i + S_j$. Then remove S_j from the vector S (then go to 8.)
8. Set $K = K - 1$; now the vector S should have one less element.
9. If $(K > 0)$, we are not finished so go to 4.
10. If $(K \leq 0)$, perform '**Algorithm 2. Protracted speciation metacommunity sample**' from step 3. onwards which will give the species abundance distribution.

Appendix B: Visibility of the effects of protracted speciation after sampling.

The degree to which this can be detected in empirical data sets will depend on the size of the metacommunity, the value of τ , how many individuals are sampled and the randomness of the sampling. We gather from algorithm 2 above that protracted speciation will be visible in a sample if a significant number of coalescence events occur between sampled lineages during the most recent τ generations. If $\tau = 0$ then we have the point mutation case and of course it is then impossible to observe any effect of protracted speciation. The probability of coalescence in each birth event is given by $\frac{K(K-1)}{J_M(J_M-1)}$ and this yields an approximate rate per generation of $\frac{K(K-1)}{(J_M-1)}$ which continues for τ generations. The effect of protracted speciation can therefore be said to be roughly proportional to τ and proportional to sample size squared but inversely proportional to J_M . If the sample taken from the metacommunity is non random (for example by pooling local communities) then dispersal limitation plays a role in each local community and so the true size of the sample from the metacommunity will be less than the total number of individuals sampled.

Appendix C: Age of common species

Many solutions have been proposed to circumvent the high ages of abundant species in neutral models (Nee, 2005; Ricklefs, 2006), including environmental stochasticity (Allen & Savage, 2007), a historical reduction in metacommunity size (Nee, 2005) and a restatement of the problem. This last explanation argues that the problem has been overstated because only average species lifetimes have been considered (Nee, 2005; Ricklefs, 2006), whereas the model could still be consistent with very abundant species that, by chance, reach these abundances much faster. Indeed, it has been shown that the SAD in a neutral model reaches an equilibrium quickly whereas the average species lifetime does not (Zillio & Condit, 2007), suggesting that some species must reach high abundances in a relatively short time scale.

Here, we show that the age of common species in neutral models with point mutation speciation need not, in fact, be as absurdly high as has been thought. The number of species in the metacommunity with abundance j is

$$\psi_j^{\text{point}} = J_M \mu \frac{(1 - \mu)^{j-1}}{j},$$

while the number of species with abundance j that were brought into being in the interval $(-t, -t - dt)$ is

$$J_M \mu P^1(t) dt = J_M \mu^3 e^{-\mu t} \frac{(1 - \mu)^{j-1} (1 - e^{-\mu t})^{j-1}}{[1 - (1 - \mu)e^{-\mu t}]^{j+1}} dt.$$

Therefore, the probability that a randomly chosen species with abundance j was brought into being in $(-t, -t - dt)$ is

$$\frac{J_M \mu P^1(t) dt}{\psi_j^{\text{point}}} = \mu^2 e^{-\mu t} \frac{j(1 - e^{-\mu t})^{j-1}}{[1 - (1 - \mu)e^{-\mu t}]^{j+1}} dt,$$

so the probability that a species of abundance j is younger than T is

$$q(j, T) = \int_0^T \frac{J_M \mu P^1(t)}{\psi_j^{\text{point}}} dt = \left(\frac{1 - e^{-\mu T}}{1 - (1 - \mu)e^{-\mu T}} \right)^j. \quad (1)$$

The median age $T_{1/2}$ of a species of abundance j is obtained by inverting equation (1) so T is the subject, and setting $q(j, T) = 1/2$:

$$\begin{aligned}
T_{1/2} &= \frac{1}{\mu} \log \left[1 + \frac{\mu}{e^{\frac{\log 2}{j}} - 1} \right] \\
&\rightarrow \frac{1}{\mu} \log \left[1 + \frac{\mu j}{\log 2} \right] \quad \text{assuming } j \gg 1 \\
&\rightarrow \begin{cases} \frac{1}{\mu} \log(j\mu) & \text{when } j \gg 1/\mu \\ \frac{j}{\log 2} & \text{when } 1 \ll j \ll 1/\mu. \end{cases}
\end{aligned}$$

Thus, while other studies that neglect mutation conclude that $T_{1/2} \sim j$, we find that the lifetime of the most abundant species is $\approx \frac{1}{\mu} \log(j\mu)$. The reason for the discrepancy is that, under mutation, species with ages greater than $\sim 1/\mu$ are extremely rare, and hence any species with a very high abundance is overwhelmingly likely to be due a series of unlikely fluctuations rather than because it has been in existence for a long time.

While we have used the point mutation speciation model for illustration purposes, the same result holds for $j \gg \tau$ in the protracted speciation model, for which the dynamics of the most abundant species is the same. Protracted speciation does have an important part to play here by enabling μ to not require such a small value, which is important for $\frac{1}{\mu}$ to be of reasonable size and for $j \ll \frac{1}{\mu}$ to be satisfied.

Appendix D: Incipient species

To approximate the number of incipient individuals, first notice that the expected number of surviving ancestors of each lineage is always one. The total number of speciation-initiation events for each generation is $J_M\mu$ and each such speciation-initiation event from the last τ generations creates on average a single incipient individual at the present day. The total number of incipient individuals is therefore $J_M\mu\tau = J_M\tau\mu$ and the proportion of incipient individuals in any sample is expected to be $\tau\mu$. This remains a good approximation provided that further speciation-initiation events do not occur during the incipient phase of any species; if they did any descendants from these doubly incipient (or incipient squared) species would be counted twice towards the total for $\tau\mu$. We can therefore use $\tau\mu$ as a good approximation for the number of individuals belonging to incipient species and as a definite upper bound.

References

- Allen, A.P. & Savage, V.M. (2007). Setting the absolute tempo of biodiversity dynamics. *Ecol. Lett.*, 10, 637–646.
- Nee, S. (2005). The neutral theory of biodiversity: do the numbers add up? *Funct. Ecol.*, 19, 173–176.
- Ricklefs, R.E. (2006). The unified neutral theory of biodiversity: Do the numbers add up? *Ecology*, 87, 1424–1431.
- Zillio, T. & Condit, R. (2007). The impact of neutrality, niche differentiation and species input on diversity and abundance distributions. *Oikos*, 116, 931–940.