Review: "Identification of the Mode of Evolution in Incomplete Carbonate Successions"

This manuscript uses forward sedimentological modeling to generate synthetic stratigraphic sections, simulates trait evolution in the stratigraphic and time domains, and then fits evolutionary models to assess the degree to which stratigraphic architecture affects evolutionary interpretation. Surprisingly, they find that the geological effects are minor, but that a common way to analyze models of trait evolution seems to perform poorly, regardless of the geological filter.

There is a lot to admire about this manuscript. Explorations of how stratigraphic structure can affect interpretation of trait evolution in time-series has been almost completely ignored since Bjarte Hannisdal's excellent paper almost 20 years ago. I don't have the expertise to evaluate the sedimentological model, but it is well explained, and the exploration of a carbonate system is unique in the literature of evolutionary time-series, as far as I know. The study is well designed, using both idealized and empirical sea level curves, and sampling across different parts of the carbonate platform. Figures are very nice, and it is written clearly and with grace.

The one problem is that the surprisingly bad performance of the evolutionary models is based on a misunderstanding, and, as a result, a lot of the results and discussion will need to be reconsidered. The general issue is that one needs to evaluate not just model support (AICc and Akaike weights), but also model parameters to understand an analysis. In this study, specifically the use of the OU model creates confusion because this model can take on parameter values that cause it to converge to the other three models. The OU model has 4 parameters: the ancestral trait value (anc), the optimal trait value (theta), the strength of attraction to the optimum (alpha), and the stochastic component (vstep).

- When alpha is very strong, the traits are drawn very quickly to the optimum. In the limit (alpha -> Inf), you get a white noise process around theta, with a stationary variance of vstep/(2*alpha), equivalent to how stasis is modeled.
- When alpha is very weak, the optimum has little effect. As alpha -> 0, you get a random walk / Brownian motion.
- When alpha is weak and the optimum is very far from the starting trait value, you get a nearly linear trend from anc to theta.

So, the initially confusing results of evolutionary model fitting can be understood: nearly all model support is received by either the generating model or the OU model with parameter values that cause it to mimic the generating model. For example, when stasis is the generating model, theta and anc will be very similar and alpha values will be quite strong (often easier to judge from the half-life, log(2)/alpha), and furthermore, vstep/(2*alpha) will be nearly exactly equal to the stasis variance. Note that considering parameter values removes all difficulties of

interpretation: the user would find that all the model support goes to two different ways to parameterize white noise, which is the correct, generating model.

The Discussion alludes to this property of the OU model (line 687ff), but doesn't make the connection to interpreting the results in this light. The need to consider parameter values was one of the major points of Grabowski et al. (2023) in their correct criticism of Cooper et al.'s (2016) paper about OU models.

In terms of how to handle this, the paper can be revised to account for this interpretation, but I'd argue it would be better served by simply omitting the OU model, for two reasons. First, this model is not easily to justify biologically. Yes, the OU model can be used to model a population converging to a new adaptive peak. But this dynamic is rapid, and is expected to last a few generations to, at most, a few thousand generations. On the ~2 Myr scale of this study, there really isn't any expectation that this dynamic should be captured. Second: URW, Trend/drift, and stasis are useful because they capture three qualitatively different evolutionary patterns: meandering, directional, and fluctuating, respectively. I don't see any benefit to adding a 4th model that just has the effect of mimicking the best-supported of the other models, essentially splitting the support for the correct dynamic over two nearly equivalent models.

Another contributing factor here is in how Akaike weights work with nested models. If one model is nested within another (e.g., Brownian motion within Brownian motion with drift; the other three models are also nested or nearly nested within OU), it is impossible for the simpler model to decisively beat the more complex one according to Akaike weight. The log-likelihood of the more complex model cannot be lower than that of the simpler model when they are nested. Therefore, the only way for the simpler model to be better is via the parsimony term in AIC. For models that differ by 1 parameter, this leads to a delta AIC of 2, and maximum Akaike weights of 0.73 for the simple model, even when the simple model is correct (see Hunt 2006, p. 596). This is for AIC; with AICc, the exact weight will be initially higher for the simple model and then converge to the AIC value with increasing n. This means that the 0.9 used as a threshold for Akaike weight is inappropriate: it is mathematically impossible for the simpler of nested models to reach this threshold for AIC (and for AICc except when the parsimony penalty is high at low n). This, by the way, also explains the puzzling behavior in Figure 10 in which performance seems to gets worse with increasing n: more complex models will face decreasing parsimony penalties as n increases, which explains the asymptotic increase in support for OU in these plots.

I will say that I am puzzled that the OU model so consistently beats Stasis even with the two extra parameters in the OU model. It doesn't really matter much here because the dynamics will be basically equivalent (as discussed above), but this is something I am curious about.

Below I have added some minor comments, in manuscript order. Despite the problem I have identified above, I want to emphasize how much I like this study. With suitable revision, I think it will be an important contribution to the literature.

Minor comments, in manuscript order

- Line 34: here, and elsewhere in the manuscript, pulsed change is referred to as punctuated equilibrium. I don't think this is quite accurate: the punc eq model has pulsed change but it occurs at lineage splitting. A pulsed change within an unbranched lineage is more evidence against than for punc eq because it involves large changes without speciation. (Gould would sometimes try to cloud this issue.) I'd recommend using terms like pulsed or punctuated change, and not punctuated equilibrium, for unbranched lineages.
- Line 88, before the Fossilized Birth Death model and related approaches, there was a phase in which fossil data was used a lot (sometimes naively) to get constraints for node dating approaches.
- Line 110ff: The presentation of completeness that I am familiar with (e.g., Shanan Peters' work) emphasizes that completeness will depend on the temporal scale of resolution. A section may be mostly complete when considered in 1 Myr bins, but will be much less so if the bins are 10 Kyr.
- Line 169: here and at a few other places, it seems to imply that previous approaches in paleo have required samples to be equally spaced in time. The model fitting approaches used here and in Hunt (2006) cited here have always allowed for arbitrary spacing of samples.
- Line 305: I don't think it needs to be done in this paper, but, as an FYI, it is not difficult to generate realizations of the OU model with unequal sampling. The *sim.OU* function in *paleoTS* does it one way, and there is another approach in which a whole time-series is a single draw from a multivariate normal distribution using the vector of means and covariance matrix from Hansen & Martins (1996).
- Line 322: not quite right as written, as the standard deviation would be sigma *sqrt(t), not sigma. The simulation code is correct, though.
- I would not say scenario 3 is "weakly directional". Both it and scenario 4 are strongly directional, really more so than just about any empirical sequence. This can be seen from the figures – both look almost like straight lines -- and the results are basically the same throughout for both. Calculations from Hunt (2012, Table 1) indicate that directionality accounts for 98% and 99.5% of the evolutionary change in scenario 3 and 4, respectively. I'd recommend just keeping scenario 3 as representing trends and dropping the unrealistic scenario 4.
- Line 348ff. I see the need for the distinction, but it seems odd to call them both time-series. Perhaps instead they can be stratophenetic series and time-series? The former phrase has been used occasionally in this literature.
- Line 516, about stratigraphic completeness not being the driver of outcomes. This is an interesting and important point.
- Line 640ff: this section should be reconsidered based on my comments above. The discussion about Levy flights is interesting, as I agree that kind of dynamic would probably be favored when there are unrecognized hiatuses in a section. That model isn't implemented in paleoTS, but (within-lineage) punctuations are.

- Line 826: this references Hannisdal (2006). The only other example of a similar study I know of appears in one of the chapters of the Patzkowsky and Holland book. I think that should be cited somewhere in here as well. (It is possible that chapter refers to another paper that I am not remembering at the moment, too.)
- Line 829: you would probably see more artefactual support for stasis if the generating parameters didn't have such high rates. With lower rates, sampling noise would be a larger component, and since sampling noise looks like stasis, you should get more spurious cases of stasis. Analyzing more shorter sequences might have a similar effect.
- I love the forward modeling approach to investigate what happens under known conditions. I am curious if you think these sedimentation models might ever be used for the inverse problem, so as to generate more realistic age models for empirical fossil time-series?

Signed,
Gene Hunt

References

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50(4):1404-1417.

Hunt, G. 2012. Measuring rates of phenotypic evolution and the inseparability of tempo and mode. Paleobiology 38(3):351-373.

Patzkowsky, M. E., and S. M. Holland. 2012. Stratigraphic Paleobiology: Understanding the Distribution of Fossil Taxa in Space and Time. University of Chicago Press, Chicago.