1. **The aim**
2. Infer biogeographic histories of the symbionts.
3. Use belief propagation to infer host repertoires on a symbiont tree decorated with those biogeographic histories.

Key questions on the approach to take:

1. Should biogeographic states be reconstructed as a monomorphic character, with a single parameter governing switches between states, or should occupying multiple states be allowed, and switching governed by a gain-rate and a loss-rate?
2. Should any deterministic change in host repertoire be allowed at the point of geographic range change?
3. **Pros and cons for question (i)**
4. Biogeographic reconstruction was initially conceived as being done as a monomorphic character. Points in favour of this are: few parameters needed (just 1 rate of switching between regions); is comparable to standard discrete-character ancestral state reconstructions (and jump dispersal in DEC+J); would be a good match for datasets where species’ ranges are mostly within single regions.
5. Points in favour of a 2-step range change model (so a gain parameter and a loss parameter): comparable to DEC and to host repertoire evolution; applicable to datasets containing species found in multiple regions; potentially less restrictive in terms of host repertoire scenarios allowed at the point of geographic range change (see next section). Could handle branching in the symbiont phylogeny by assuming identical range inheritance by each offspring lineage, as is the case for host repertoire.

Perhaps this comes down to: is it less realistic to restrict symbionts to single-region geographic ranges at any given point in time, or to assume symbiont speciation never features the splitting of the original geographic range between offspring lineages?

1. **Scenarios for question (ii)**
2. No deterministic change in host repertoire when a symbiont changes its geographic range. Points in favour is that all repertoire changes are described by the estimated transition rates, and so the interpretation of the transition rates is the same as in the original host repertoire model. A con is that it would cause many host repertoire permutations to be incompatible with a switch between geographic regions – at least one host must be found in both the original region and the new region, and no hosts not found in the new region can be in the new repertoire at the point of shifting. Switching between 2 regions is ruled out if there are no hosts found in both regions!

* A workaround the incompatible repertoire issue could be forcing all incompatible hosts to be lost when simulating along the branch leading up to the range shift (and ensuring at least one host is shared by the original and new region) by weighting simulations that don’t finish on a full compatible repertoire to 0, and resampling. But it seems biologically weird to impose that kind of restriction on which symbiont lineages are able to disperse to new regions.
* This issue would be reduced by having a 2-step range change model so that lineages could gain a new region, and lose hosts that aren’t in the new region, before losing their original region. But the requirement to have a host found in the new region already in your repertoire would remain.

1. Deterministic loss of hosts when you switch to a region those hosts aren’t found in. This would mean symbionts aren’t penalised (ie prevented from geographically dispersing) by having a broad host repertoire in their original geographic region – any of these hosts not in the new region will simply be dropped when the switch between regions occurs. Would be useful in either scenario for question 1, but particularly important in the monomorphic biogeography scenario.
2. Using state 1 in the repertoire to describe hosts a symbiont can use but doesn’t geographically overlap with, and having deterministic change from 1 to 2 when geographic overlap is achieved, or deterministic change from 2 to 1 when geographic overlap is lost. This seems nice in that creates much more flexibility in the host repertoires that can be had by lineages about to shift between regions. The transition matrix would vary according to whether a symbiont and host overlap:

|  |  |  |
| --- | --- | --- |
| Geographic overlap of host and symbiont | Symbiont changes its geographic range | No geographic overlap of host and symbiont |
| |  |  |  |  | | --- | --- | --- | --- | |  | 0 | 1 | 2 | | 0 | - | 0 | a | | 1 | 0 | - | 0 | | 2 | b | 0 | - | | |  |  |  |  | | --- | --- | --- | --- | |  | 0 | 1 | 2 | | 0 | - | 0 | 0 | | 1 | 0 | - | d | | 2 | 0 | d | - | | |  |  |  |  | | --- | --- | --- | --- | |  | 0 | 1 | 2 | | 0 | - | c | 0 | | 1 | e | - | 0 | | 2 | 0 | 0 | - | |

*d for deterministic change; a, b, c and e are the parametrised rates of change.*

*a/c and b/e could be constrained to be equal (would keep things much simpler!).*

All the permutations in this scenario are:

If you’re in 0 for a host, nothing happens at the point of changing regions

If you’re in 0 and staying within an appropriate region for the host, you can go to 2 or stay on 0

If you’re in 0 and not in the appropriate region you can go to 1 or stay on 0

If you’re in 1 and you gain a region, you deterministically go to 2

You can’t be in 1 if you’re already in the appropriate region

If you’re in 1 and staying in the region you can move back to 0 or stay on 1

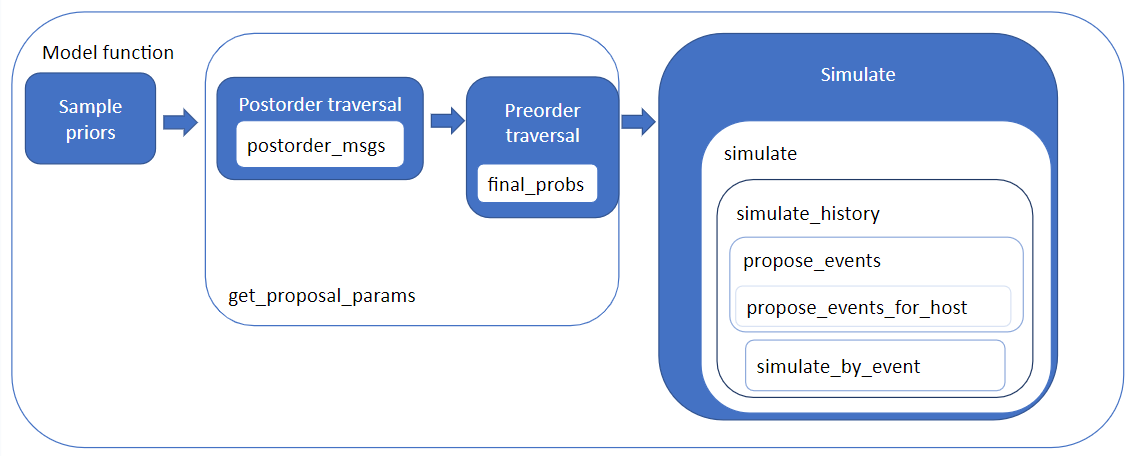
If you’re in 2 and you lose the appropriate region, you go deterministically to 1

If you’re in 2 and you stay in a region, you can go to 0 or stay on 2

You can’t be in 2 if you’re not in the appropriate region.

1. **Parts of the host repertoire code that would need modifying**

*Line references are for host\_repertoire.tppl*

1. Parameter proposal with **postorder\_msgs()** and **final\_probs()** – at this point all hosts are assumed to have the same probability of state changes, hence the sapply1() applies the transition matrix to each host for each tree (L190-191, 242-243).
2. **propose\_events()** – we don’t want to propose gain events for hosts that are in the wrong region (L320).
3. **simulate\_by\_event()** – requires a calculation of the rate of change. The calculation (**totalRate;** L502) assumes all hosts are available.

Schematic of host repertoire code

1. **Ideas and queries about coding the model extensions**
2. Put a matrix containing the regions each host occurs in in the model parameter object so that it can be passed to functions as required (mp.host\_geo\_matrix)
3. Use a 2x2 qMatrix to get the probabilities of moving in & out of the repertoire (code lines in each of 4a-c), then add a step where the host\_geo\_matrix is used to sort the hosts in the repertoire into “1s” (no geographic overlap with the symbiont) and “2s” (geographic overlap with the symbiont).

* A similar process could work in the scenario without deterministic changes between 1s and 2s – calculate probabilities for all hosts, but then a step of setting to 0 the probabilities of being in hosts that are in the wrong regions. Would need to be done each time in the code you use the qMatrix.

1. If no deterministic changes between 1s and 2s – add in an if statement to gainsIf0sOr1s (L526) so that gain rate is 0 if host *i* is geographically unavailable.
2. Have a function that sorts out changes between 1s and 2s at SwitchNodes. And another function that collapses 2s back to 1s when using the qMatrix to calculate gain and loss rates from the repertoire. Both functions would use the host\_geo\_matrix.
3. **Possible extensions that probably aren’t worth the trouble**
4. Model instantaneous loss of hosts at the point of geographic switching. Don’t know if that’s at all possible in continuous time!
5. Model repertoires and geographic ranges in discrete time slices, which would be a fundamentally different approach.
6. Model multiple simultaneous repertoire changes. Potentially particularly useful at switch nodes but could also be allowed at any other time. Could also be an interesting extension to inference of biogeographic histories, as it would give symbionts the option of adding one or more regions to their range (one or more gain events at a given time point), removing one or more regions from their range, or switching their range (simultaneous gain and loss events). Suspect this would make things insanely complicated.
7. Allow repertoire changes at all nodes, ie speciation as well as biogeographic dispersal. Though that would just raise more questions, like whether to allow geographical switching at the point of speciation etc.