Lab 13
Jennifer Lin

Question 1

Geographic range is being modeled here as a discrete character state, where the character state represents the combination of areas that a lineage inhabits at any given point in time. Models of biogeographic range evolution are essentially no different than the other models of discrete character evolution we have looked at over the course of the semester (with the exception of including cladogenetic events).

1. DNA substitution models have 4 discrete states. How many states will our biogeographic model have? Remember, these models allow for a lineage to be in more than one area at a time.

   In our biogeographic data, we have 2 discrete states (0 or 1), which represents whether the species exist in that area (that character) or not. As a result, if the species has character state 0 at that area, this means this species does not exist in this area. On the other hand, if the species has character state 1 at that area, this means this species exists in this area.

   In our biogeographic model, we have 16 area states in total:
   5 none, K, O, M, H,
   6 KO, KM, KH, OM, OH, MH,
   4 KOM, KOH, KMH, OMH,
   1 KOMH

2. Does it make sense to allow the character state 0 0 0 0? What would this represent?

   If there is one species having character state 0 0 0 0, this means this species does not exist in none of the areas we are investigate in. It would make sense only when we try to represent this species does not exist in these areas. However, if it does not exist in these places, we should just remove it.

3. To calculate likelihoods of discrete character evolution models we have to exponentiate the transition matrix, which can be very computationally demanding for large matrices. If we were performing inference on a dataset with 10 areas instead of 4, how many states would our model have? How large would the anagenetic transition rate matrix be?

   If there are 10 areas, there would be 1024 states. Thus, the anagenetic transition rate matrix would be 1024x1024.

4. To model cladogenetic change we need an additional transition probability matrix that represents probabilities for all combinations of the state before cladogenesis and after on each of the two daughter lineages. Given 10 areas, how large would the cladogenetic transition probability matrix be?

   It would be 1024x1024x1024

Question 2

Take a look at the results DEC object. What is the maximum likelihood estimate of the rate of anagenetic “dispersal” (range expansion)? And the rate of anagenetic “extinction” (range contraction)?

> “dispersal” (range expansion) = d = 0.03504546
> “extinction” (range contraction) = e = 0.02835632

Question 3

Compare the estimated ancestral ranges of the lineages leading to *P_hexandra_Oahu* all the way back to the root of the tree. Explain the results in context of biogeographic hypothesis testing. Which hypothesis makes more sense to you given Hawaiian Islands geography?

> In DEC model, the ancestral state of *P_hexandra_Oahu* is O-KO-KO-KO-K(root). We can see that in the transition ancestral states, there are three states KO, which indicates *P_hexandra_Oahu* exists on K and O Islands during these times. In DEC+J model, the ancestral state of *P_hexandra_Oahu* is O-K-K-K-K(root). We can see that the ancestral state jumps from K to O directly.
>
> The Hawaiian Islands were formed by such a hot spot occurring in the middle of the Pacific Plate. While the hot spot itself is fixed, the plate is moving. So, as the plate moved over the hot spot, the string of islands that make up the Hawaiian Island chain were formed. As a result, islands were formed sequentially. We do expect founder speciation events when an island was newly formed. Thus, DEC+J model makes more sense to me because it includes the “jump” parameter for long distance dispersal / founder speciation events.

Question 4

1. Which model does the AIC support?

   > We may always get a closer prediction to the data we have when increasing the number of parameters. However, we may risk overfitting the data. As a result, we use AIC to estimate the models.
   >
   > Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models.AIC checks if increasing the number of the parameters really efficiently increases our prediction. AIC is calculated by "AIC = -2 (log-likelihood) + 2K", where K is the number of model parameters (the number of variables in the model plus the intercept) and log-likelihood is a measure of model fit. Thus, **a lower value of AIC suggests a better model**.
   >
   > stats$AIC1 = **47.89518**
   > stats$AIC2 = 73.08392
   >
   > As a result, AIC supports this model 1 (i.e. **DEC+J**).

2. These models incorporate cladogenetic evolutionary events, where evolutionary change occurs at speciation events. However, in our reconstructed phylogenies we usually only consider the speciation events that led to the extant taxa. How might unobserved speciation events (lineages that went extinct) affect our inferences?

   We may lost some transition species and their biogeography states. For example, in our results, it may looks like species A jumped to another place, went over a speciation event, and become species C. However, we may never notice there was a transition species B, which coexists in both habitats of A and C just because B has already extinct. Consequently, we need to be very careful when dealing with this kind of problems.