How can we use DNA to infer phylogenies?

Evolutionary genetics

February 28, 2018

Abstract

As you are no doubt aware, DNA evidence can be used to identify relatives, to estimate the ancestral origins of human individuals, and to infer the relationships between species. But how do scientists use this data to make inferences about evolutionary history? In this exercise, we will develop models and tools to begin to infer the relationships between species using DNA data.

What is a molecular phylogeny?

A molecular phylogeny is a model that represents the relationships between species. Typically, these are represented as “trees”. An evolutionary tree is then a hypothesis about which species are most closely related to each other, and how far back in time species may have shared common ancestry. We can compare the likelihood of various hypotheses by sequencing the DNA of the species we are interested in learning about.

How distantly are humans and chimpanzees related?

The simplest tree that we can draw shows the relationship between two species. Since there is only one way that two species can be related with a tree, our only task as evolutionary geneticists will be to use DNA to infer the time of the speciation event (*i.e.*, the time at which humans and chimpanzees shared a common ancestor). This time is often referred to as the “divergence time”.

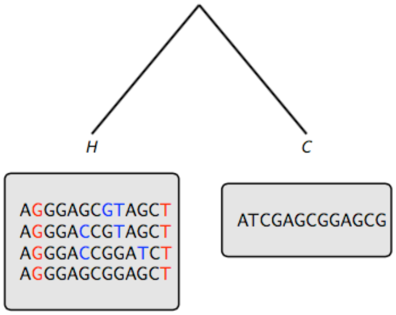

Figure 1: A simple tree relating (H)umans and (C)himpanzees. The sequences at the bottom represent sequence data we may have obtained from each species.

Discuss the following questions:

1. Chimpanzees and humans differ at about 1% of the basepairs in the genome. Given this fact, do you have a guess as to how much two random humans tend to differ at the genomic level?
2. What is the difference between a substitution and a mutation?
3. Using the sequences above, identify which columns are substitutions and which are mutations.
4. Suppose that you sequence a human and a chimpanzee, and you observe substitutions at 1% of the basepairs. Suppose additionally that you know that mutations tend to occur at a rate of 2.5×10^{-8} per base pair with each new birth. Is this enough information to calculate the divergence time? If not, what other information would you need?

Comparing trees

Suppose now that we wish to understand the relationships between three species, humans, chimpanzees, and gorillas. There are three different ways that these three species can be related, each of which is a hypothesis about their evolutionary origins.

(A) (B) (C)

Discuss/answer the following with your group:

1. Draw the trees corresponding to the three different evolutionary hypotheses in the space for A, B, and C above. *(Hint: one hypothesis is that chimpanzees are more closely related to humans. What does this imply about the evolutionary tree?).*
2. Given our understanding of how mutations accumulate and become substitutions, what would you anticipate about the relationship between the DNA sequences of each species under the hypothesis that humans and gorillas are most closely related, while chimpanzees are the out-group? In other words, how much divergence at the DNA level might we expect between the two closely related species as compare to the out-group?
3. You obtain some DNA sequences suggesting that humans and chimpanzees are 1% diverged, while humans and gorillas are 1.5% diverged. Which tree or trees does this data support? What do you additionally suppose you would observe if you compared gorilla and chimpanzee DNA, would the distance be greater, smaller, or the same as the distance between human and chimpanzees?

Complications!

Although it is clear that we can use DNA to test hypotheses about the relationships between species, much of our understanding of how to build trees is rooted in what is called “neutral theory”. Neutral theory supposes that mutations do not confer a fitness benefit or fitness detriment (*i.e.*, there is no natural selection!).

With your group discuss the following.

1. Suppose that we sequenced 100 genes, and used them to infer a phylogeny of five species (the specific number of species isn't really important for this question). Suppose that 10% of the genes are absolutely essential for life and cannot tolerate many changes. How might including these genes in our analysis affect our inference? Would it affect the times that we estimate for divergence between species?
2. Suppose now that an additional 10% of the genes are under positive selection, *i.e.*, new changes in these genes often confer a fitness benefit. Would this affect our estimates?
3. Think about the process of speciation for a moment. Clearly, speciation must occur relatively slowly, since we do not generally capture it in the act in field observations. However, we have been modeling speciation as an instantaneous event. If speciation actually occurs slowly and over millions of years, how might this affect our inferences?