Molecular Phylogenetics

Saul Pierotti

March 11, 2020

Introduction

- Individuals do not evolve, populations do
- Selection can be stabilizing, disruptive, directional, stabilizing
- Genetic drift is stronger in small populations
- The specific name is not univocous, we need also to specify the genous
- The species is the only natural classification, higher classifications are human-made
- An operational taxonomic unit is one of the leaves of the tree (OTU)
  - It is a proxy for the species concept in organisms without clear species boundaries
- Nodes of the tree are hypothetical taxonomic units (HTU)
- The lenght of edges is related to divergence
- Trees can be rooted by using an outgroup
  - The outgrup is by itself an OTU which is for sure more distant to all the other OTUs than the distance among OTUs
- All the OTUs but the outgrup represent the ingroup
- A monophyletic group is a clade that includes the most recent common ancestor of all the leaves and all the descendant of that ancestor
  - A clade is always monophyletic
- A paraphyletic group includes the most recent common ancestor of all the leaves, but not all the leaves of that ancestor
- A poliphyletic group includes leaves from more than 1 taxon
- Evolution is like a branching tree, not like a ladder
  - What is commonly considered ancestor is a sister group, the real ancestor does not exist any more (!)
- The observed genetic distance between 2 species is the sum of the distance between both species and their common ancestor
- The more distant the split, the more the genetic distance
- Frequency of observed mutation is inversely related to the strenght of selective pressure
  - Low mutation rate can be related to higher gene content
  - When selecting a region for phylogenetic analysis, we need to adjust the mutation rate with the distance between the OTUs
    - I cannot use very divergent regions for distantly related organisms or very conserved regions for closely related organisms (!)
  - Differential mutation rate can be observed also inside genes
- The rate of synonimus (S) and non-synonimus (N) mutation is an indication of the selection regime
  - $S > N$ suggests positive selection
  - $S = N$ suggests neutral selection
  - S less than N suggests negative selection
- The neutral theory of molecular evolution (Kimura) states that most molecular divergence is neutral
  - On the contrary, phenotipic evolution is under selection
- For phylogenetic analysis, we aim at using loci under neutral selection
- The molecular clock hypothesis assumes constant mutation rate

  - Implicitely it assumes neutral evolution (!)
  - Double molecular distance means double separation time
- To understand the significance of a phylogenetic hypotesis we can use other information from biogeography
- Relation determined by genes under strong selection can give wrong results (!)
  - Convergent evolution can make me cluster unrelated species, while splitting related species that have adapted to new environments
- When we compare sequences or characters they must be homologous (!)
- On the contrary, analogous characters are a product of convergent evolution
- Homologus genes need to be orthologus in order to be useful for classification
- Paralogous genes cannot be used for classification because they can change or loose function

Molecular markers

- The level of variability is not constant for all organisms and species
  - Citochrome B is really variable in insects but not in mammals
  - Cytochrome C is more variable in mammals
  - Before doing something on a gene look at the literature (!)
- The mtDNA is smaller, aploid and more variable than the nuclear genome
  - It is some orders of magnitude more variable than the nDNA (!)
    - Less efficient proofreading
    - Many more replications per individual
  - mtDNA is useful for analysing shallow divergence
- Gene rearrangments are really unlikely to happen twice in the same way
  - Therefore, they are relly good to establish relationships
- Transcriptome sequencing is better than DNA sequencing in many cases
  - It is easier to assemble and annotate
  - It is easier to handle since it is smaller

Species trees and gene trees

- A species cannot be represented by a single DNA sequence
- When we create a tree we actually reconstruct the phylogeny of the marker, not of the species
- Because of this, we want to use many molecular markers at the same time
- We want to find which gene trees are informative for and overlap with the the species tree