Molecular Phylogenetics

Saul Pierotti

March 23, 2020

Introduction

- Phylogenetic methods apply to both DNA and protein, but we will focus mostly on DNA since it is easier to work with
- We want to compare samples in order to understand their coancestry
- Individuals do not evolve, populations do
- The specific name is not univocus, we need also to specify the genus
- The species is the only natural classification, higher classifications are human-made
- Darwin's postulates of evolution: populations change over time
  - Individuals in a species have a certain variability
  - Some variation is heritable
  - Survival and reproduction are not completely random
- Evolution is more like a brancking tree than a ladder
- Evolution involves first mutation and then selection
- Selection can be of different types
  - Stabilizing, if it tends to increase the frequency of an optimal trait
  - Directional, if a trait becomes more and more extreme
  - Disruptive, if it tends to go away from a certain trait
  - Balancing, if all traits are equally favored
- Other than selection, evolution is promoted by genetic drift
  - The effect is stronger in small populations
- The effective population size is the size of an ideal population with random mating that has the same gene frequency changes as the studied population (the census)
- Two lineages that, going back in time merge in a single ancestor define a coalescent event
  - The mrca of all the individuals in a population almost never dates back to the first generation
  - The colaescent time of a population is the time passed since the mrca of all individuals existed
- Deterministic evolution is possible only in infinite popualtions: real models are stochastic and we can only predict the probabilities of allele frequencies
- 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
- A group of taxa that share the same branch is a monophiletic cluster
- The topology of a tree is its branching pattern
- Nodes of the tree are hypothetical taxonomic units (HTU)
- The lenght of edges is related to divergence time
- 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
  - The root of the tree is the node connecting the outgroup and the ingroup
- If an outgroup is not available, a tree can be rooted by midpoint rooting
  - The root is the node connecting the most distantly related OTUs
- 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
- LUCA: life is thought to be monophiletic
- 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
  - In most populations the effective population size is incredibly small compared to the magnitude of the selective forces
  - Most fixation events are the result of stochastic process on quasi-neutral mutations
  - Adaptive evolution is more predominant when the species is far from the peaks of the fitness landscape
- Speciation is favored by events that reduce the diversity of a population while increasing the diversity with other populations
- The probability of fixation due to genetic drift of a mutation is equal to its frequency
  - If n is the popualtion size, in a diploid population a new mutation has a frequency of 1/2n
  - On average, it takes 4n generation to fix a mutation through drift

Phylogenetic markers and trees

- Initially classification was based on morphological carachters, and taxonomy is still largely based on this
  - Today we tend to use much more molecular data such as DNA and protein sequences, and RFLPs
  - The prefereability of molecular or morphological data is under debeate (Patterson et al, 1993)
  - For estinct species, we often don't have molecular data
- Initially the molecular classification was based on allozymes, then RFLPs became prominent and now microsatellites, SNPs and sequencing data are most used
- The variability of sequences arises from mutations, duplications, recombination, HGT
  - Point mutations, insertions and deletions are the most used data for molecular phylogenetics
- Molecular phylogenetics studies the similarities of 2 sequences assuming that they are homologous
- Most variability in homologous sequences arises from point mutations in the 3rd codon position
- Different genes or portions of gene can have different conservation rates, and distantly related homologs can be identified only in enzymes and structural proteins
  - In introns usually the divergence rate is that derived from neutral evolution, while in exons it is higher or lower
  - When not even sequence of core regions are conserved, homology can be detected at the structural level

  - Closely related species can be compared at the DNA level, families and genera are better compared at the aminoacid level
- 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 (!)
- Paralogous genes derive from duplication, orthologous genes from speciation
  - Studing paralogous sequences is informative for the duplication event
  - Orthologous sequences are informative for speciation events
  - If I want to study speciation I need to be sure that my locus is orthologous (!)
- When we compare sequences or characters they must be homologous (!)
- Homologus genes need to be orthologus in order to be useful for classification
- Multiple substitutions on the same site or equal substitutions in different species can lead to underestimate the genetic distance: homoplasy
- The molecular clock hypothesis assumes constant mutation rate
  - Implicitely it assumes neutral evolution (!)
  - Double molecular distance means double separation time
- 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
  - It tells only about the maternal lineage (!)
- Nucelar DNA is less variable, subject to recombination, polyploid: a mess (!)
- Gene rearrangments are really unlikely to happen twice in the same way
  - Therefore, they are relly good to establish relationships
  - The insertion of transposable sequences is one of these
- 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
- Recombination events can create incoherent trees for the same species
  - In this case It is more adequate to represent the phylogeny with a network, not a tree
- For phylogenetic analysis, we aim at using loci under neutral selection
- To understand the significance of a phylogenetic hypotesis we can use other information from biogeography
- Relations 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
- In some instances tree can be not binary: politomy
  - Hard politomy refers to multiple, almost simultaneous speciation from a single ancestor
    * Its existence is not clear, but it seems to be approximated by explosive radiation events in viruses
  - Soft politomy refers to uncertainty in a given topology
- 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
  - If the genes that I am studying are paralogous, the coalescent event for the gene will be different than for the species (!)
- Higher coalescence time is related to lower probability of wrong trees
- The probability of coalescence for a pair of genes in 1 generation is 1/2N, where N is the size of a diploid population
  - It is the probability that 2 copy of a gene derive from the same parent gene in the previous generation

- It assumes (Kingman's assumptions) a panmittic population, neutral evolution, infinite sites and non-overlapping populations
  - In a panmittic population there is no preferential mating
- The probability that the gene tree and the species tree don't overlap is $\frac{2}{3}e^{\frac{-t}{2N}}$
  - This derives from the probability of coalescence
  - Whith more than 6 genes the probability of a wrong tree is significantly reduced
- Incomplete lineage sorting is the non-overlapping of gene and species tree

Multiple sequence alignments

- A multiple sequence alignment (MSA) is an hypothesis about the homology of multiple sequences
  - We arrange sequences so to have homologous positions in the same column
- In order to find the real alignment of 2 sequences, I need to know the sequence of the mrca (!)
- A simple model for aligning DNA: +1 for matches and -1 for mismatches
- Modelling gaps: we can use different penalties for opening and extending a gap
- Weighted sum of pairs: WSP objective function
  - It is a simple way to score MSAs
  - For each position, I get the pairwise score of each pair and I sum it
  - I can use a weight for each score that balances the over-representation of some sequences
- We could use dynamic programming on a multi-dimensional matrix for maximizing the WSP function, but this requires $O(N^M)$ time
  - N is the sequence lenght and M the number of sequences
  - It is impossible for more than 4 sequences
- Progressive alignment methods are fast but sub-optimal
  - They build a tree and use the tree for guiding the alignment
    - Usually the tree is built with NJ
  - They are by far the most used MSA approaches
  - Once I have the tree, it procedes by pairwise alignment on the most related OTUs and progressively collapses the nodes
  - ClustalX and ClustalW belong to this category
    - ClustalW is textual while ClustalX is GUI
    - ClustalW automatically corrects for over-represented sequences
  - Progressive alignment has a local minimum problem: early errors in the first alignments cannot be corrected later
- Consistency-base MSA: WSP scoring and intermediate sequence information used to improve pairwise alignments
  - T-Coffe is slower than ClustalX, but more accurate
    - It finds the MSA that most agrees with the pairwise alignments
- Iterative approach: the alignment is refined in iteration steps until I reach the maximum possible score
  - It is faster and more effective than the progressive alignment
  - I create a guide tree using a raw distance matrix
  - This is the framework use dby MUSCLE and MAFFT
- Strucutral methods use information about the RNA or protein strucuture
  - A loop can be of variable lenght, but a domain is more constrained
- In many cases (well-behaving datasets) the different alignment appraches give the same result, but there can be subtle differences
- In difficult cases the result can be quite different
- These methods employ a random seed: the same analysis can give slightly different results
- The distance among sequences can be estimated from the number of observed substitutions
  - I cannot observe multiple substitutions, so I tend to underestimate the distance (!)
- The expected distance is calculated by considering the average substitution rate (μ), the relative substitution rate (s) and the nucleotide frequency (π)
  - This is a Markov model and corrects for multiple substitutions (!)

  - It assumes that the substitution rate is constant in time and that the nucleotide frequencies are at equilibrium
  - The relative substitution rate i -> j is assumed to not depend on the status of the position prior to i
- There are many models that use different values for the parameters
  - Jukes Cantor (JK) does not specify any parameter (0 parameters)
  - Kimura 2 parameter (KM) uses equal values for the substitutions, $\pi s = 0.25$ (1 real parameter)
  - HKY uses potentially different substitution rates (4 parameters)
  - TN corrects for transition and transversion (5 parameters)
  - The general time reversible model (GTR) specifies all the parameters (8 parameters)
  - Note that 3 parameters specify 4 values since they have to sum up to 1, so I loose 1 degree of freedom
  - More parameters are not always better, I risk to do overparametrization (!)
- The strenght of a phylogenetic signal decrease with time since it is more probable to have multiple substitutions
  - The plot of observed mutation with respect to distance tends to saturate
- Among-site variation: mutation rate among different position can vary
  - We can model the among-site variation with the gamma distribution
  - The shape parameter of the gamma distribution is called α , while when included in a Markov model it is called γ because of the distribution

Tree reconstruction approaches

- The number of possible trees increases rapidly when increasing the number of nodes: this is the tree-space
- The best tree can be searched with an algorithmic distance-based or characther-based approach
  - An algorithmic approach first obtains the distances, and from them draws the tree
    - UPGMA, WGMA, Neighbour-joining are in this category
    - It is really easy to get wrong trees with them (!)
  - In charachter-based methods I need to know the ancestral sequence (!)
    - They are also called as tree search methods, and they choose a tree in the tree space considering an optimality criterion
    - An exaustive search is almost always impossible
      - The branch and bound approach is apossible solution: I create an optimal tree with a subset of sequences and I add a sequence at a time
      - Heuristic search
    - They can be refined by bayesian inference
    - They determine which tree is more likely, given the sequences
    - They are more reliable than algorithmic methods
    - Maximum likelyhood, maximum parsimony are in theis category
  - There are methods that combine the approach: I create a starting tree with neighbor joning and then refine it with other approaches
    - I can also start from a tree supplied from the user
- WPGMA
  - Start from a star-like tree
  - Create a distance matrix all-against-all
  - Join the closest sequences: I get a new tree and an average sequence instead of the 2 original ones
  - Create a new distance matrix and repeat until I create the whole tree
- UPGMA
- Both WPGMA and UPGMA are really sensitive to differences in rate of mutation among branches (differential branch lenght from a single split)
  - This is defined as rate heterogeneity
  - When I average 2 sequences I am assuming that their rate heterogeneity is equal (!)

- Neighbour-joining tries to overcome the rate heterojeneity problem by trasposing the distance matrix
- Maximum parsimony aims at finding the tree that can be explained with the minimum number of changes
  - For each tree it produces a statistics know as tree lenght, which refers to the number of hypotetical changes (mutations)
  - It chooses the shortest tree according to this optimality criterion
  - Not all variable sites are used: only those for which the ancestral state is known or can be guessed
    - Singlets are excluded (mutation observed only in 1 sequence)
- Maximum likelyhood
  - Likelihood is a posterior probability: it is the probability of the dataset given the model
  - It uses all the variable sites
  - For every site the probability of its state in every sequence is modelled to get a probability for the site
  - The probability for all sites are combined to get a probability for the tree
    - This is a really small number: we use its -log
  - It is used much more than maximum parsimony
- There is a sampling bias in tree reconstruction: we cannot sample the entire population of a species
- The robustness of a tree can be tested in different ways
  - Resampling: modify arbitrarely my dataset and see if the tree changes
    - Bootstrap analysis: removal with replacement
    - Jacknife analysis: removal without replacement
    - The support for a node is the percentage of its appearance in the resampled datasets
  - Charachter based
    - Bremer support: number of minimum steps needed to collapse a node
      - It is used mostly with maximum parsimony
      - If I need 2 steps for collapsing a node the Bremer support for that node is 2
- Bayesian analysis is based on posterior probabilities
  - It is based on the probability of the model being correct given the data
  - I have a probability for each node (!)
  - In general, if M is the model and D the data $P(M|D) = P(D|M)P(M)/P(D)$
  - $P(M)$ is defined as prior, $P(D|M)$ is the likelihood
  - Priors are typically the same for all trees, but we can give some an higher prior
    - This can be for instance because of the taxonomy of the group under investigation
- MCMC (Markov chain Moncte Carlo) is based on bayesian statistics
  - I start from a topology and I test its posterior probability
- ML is probably the most used method now
- In certain conditions the Bayesian analysis consistently overestimates the probability of clades, when compared with ML

---

upload

- This note are just quick and dirty, I will make them better as soon as possible, sorry for any inconvenience
- We will make talks in the last week of April about a paper
  - It is not mandatory
  - We can choose a paper and ask him if it is ok
  - We are expected to do 15-20 minutes presentation
  - There should be intro, methods, result, discussion and have a look also in the supplementary!
  - Have a look also at the main references cited on the paper!
  - The presentation will be done on Teams
- In order to calibrate the molecular clock we need some node that anchors the tree to an absolute timescale
  - I need to know the time of at least one specific node

- This information can be obtained from fossils or biogeographic data
  - I can know that a specific node has a specific age because I can date its fossils
  - I can know when some islands separated, and so I know when 2 population started to evolvbe independently
- Keep in mind that the dating of fossils and biogeographic events is really uncertain!
  - We need to model this uncertainty
- From that node, I can then propagate the absolute dating to the rest of the tree