

# Molecular Phylogenetics

Saul Pierotti

March 30, 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 univocal, 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 branching 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 coalescent time of a population is the time passed since the mrca of all individuals existed
- Deterministic evolution is possible only in infinite populations: 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 monophyletic cluster
- The topology of a tree is its branching pattern
- Nodes of the tree are hypothetical taxonomic units (HTU)
- The length of edges is related to divergence time
- Trees can be rooted by using an outgroup
  - The outgroup 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 outgroup 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 monophyletic
- A paraphyletic group includes the most recent common ancestor of all the leaves, but not all the leaves of that ancestor
- A polyphyletic 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 strength 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 synonymous (S) and non-synonymous (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 population 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 characters, 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 preference of molecular or morphological data is under debate (Patterson et al, 1993)
  - For extinct 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
  - Cytochrome 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 panmictic population, neutral evolution, infinite sites and non-overlapping populations
  - In a panmictic 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
  - With 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
  - Its probability is directly proportional with the ploidy of the species and inversely proportional with the number of generations since the split and with the number of genes under analysis

## 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 length and M the number of sequences
  - It is practically 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 proceeds 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-Coffee 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 used by MUSCLE and MAFFT
- Structural methods use information about the RNA or protein structure
  - A loop can be of variable length, but a domain is more constrained
- In many cases (well-behaving datasets) the different alignment approaches 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

## Distance matrices

- The distance among sequences can be estimated from the number of observed substitutions
  - This is called observed distance or p-distance
  - I cannot observe multiple substitutions, so I tend to underestimate the distance (!)
  - We say that the p-distance saturates with respect to the true distance  $d$  when  $d$  gets high
  - From now, we will refer to true distance with  $d$  and observed distance with  $p$
- The number of mutation expected in a given amount of time can be modelled by a Poisson distribution
- This Poisson process can be described by a Markov chain
  - I can describe the Markov chain with a matrix  $Q$  of transition probabilities
- The transition probability of  $X$  to  $Y$   $a_{X,Y}$  is composed of the product of different terms
  - $p(X \rightarrow Y) = \mu * \pi_X * a_{X,Y}$
  - $\mu$  is the mean substitution rate
  - $\pi_X$  is the relative abundance of the state  $X$
  - $a_{X,Y}$  is the relative mutation rate of  $X$  into  $Y$  compared to the other possible mutations
  - The self-transition probabilities are chosen so to make the sum of outgoing transitions from each state equal to 1
- Note the assumptions we are making
  - Mutations probabilities are only dependent on the immediately preceding state (Markov property)
  - Substitution rates are constant in time (homogeneity)
  - The nucleotide frequencies are at equilibrium (stationarity)
  - These assumptions are not necessarily biologically reasonable, be careful (!)
- It is possible to develop time-reversible and non-time-reversible substitution models
  - In a time reversible model  $p(X \rightarrow Y) = p(Y \rightarrow X)$ , so their matrices are symmetric
  - We will only treat time-reversible models
- Given any  $Q$  matrix, it is possible to compute the probability of change for any evolutionary time  $t$  as exponential of the matrix
  - $p(t) = Q^t$
- The  $Q$  matrix has 8 degrees of freedom
  - I have 6 possible relative mutation rates
    - \* These are the mutation rates  $\mu$ , not the transition probabilities (!)
  - I have 4 possible nucleotide frequencies
  - The 2 groups have to sum up to 1, so I lose 2 degrees of freedom
  - $df = 6 + 4 - 2 = 8$
- There are many models that specify a different number of parameters
  - Jukes Cantor (JK69) does not specify any parameter (0 parameters)
    - \* It assumes equal nucleotide frequencies,  $\pi = 0.25$
    - \* Substitution rates are all equal
  - Kimura 2 parameter (KM) uses equal values for the substitutions,  $\pi_s = 0.25$ , but models transitions and transversions
  - HKY85 is like the KM but it accounts for different nucleotide frequencies
  - TN models purine transition, pyrimidine transition and general transversion (5 parameters), plus different nucleotide frequencies
  - The general time reversible model (GTR) specifies all the parameters (8 parameters)
- More parameters are not always better, I risk to do overparametrization (!)
  - This is true when the exact value for the parameters is unknown
- The strength 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
  - An example: the third codon position mutates faster than the first, that in turn mutates faster than the second
  - In general, different positions are subjected to different evolutionary forces
- We can model the among-site variation with the gamma distribution with expectation 1 and variance

$1/\alpha$

- The modelled variable  $r$  is the relative mutation rate among sites, and its average is of course 1
- $Pdf(r) = \alpha^\alpha r^{\alpha-1} / e^{-\alpha r} \Gamma(\alpha)$
- The shape parameter of the gamma distribution is called  $\alpha$ , while when included in a Markov model it is called  $\gamma$  because of the distribution
- By adjusting the parameter  $\alpha$  I can accomodate different degrees of rate heterogeneity
- When  $\alpha > 1$  the curve is bell-shaped and models weak heterogeneity, with a big peak around 1
- With  $\alpha < 1$  the curve resembles an exponential decay, some position are really variable and others really conserved

— so far so good

## Tree reconstruction approaches

- The number of possible trees increases rapidly when increasing the number of nodes: this is the tree-space
  - With 3 OTUs I have just 1 possible tree
  - With 4 OTUs I have 3 possible trees
  - With  $n$  OTUs I have  $(2n - 5)!2^{n-3}(n - 3)!$  possible trees
- The best tree can be searched with an algorithmic distance-based or character-based approach (tree search)
- Algorithmic approach: first obtains the distances, and from them draw the tree
  - These methods are based on pairwise distances
  - UPGMA, WGMA, Neighbour-joining are in this category
  - It is really easy to get wrong trees with them (!)
  - They were initially developed for phenograms (trees based on phenotypic features)
  - Now they are applied for the construction of ultrametric trees
    - \* A tree is ultrametric when the OTUs are equidistant from the root
  - In general, I start from the most similar sequences and I join them in a new OTU, and I proceed like this until I join all the OTUs
- Tree search: find the tree that maximises an optimality criterion, also called objective function
  - In general these are function for scoring a give tree, not a series of step for obtaining it
  - Maximum likelihood, maximum parsimony are in this category
  - They can be refined by bayesian inference
  - They determine which tree is more likely, given the sequences
  - They are more reliable than algorithmic methods
  - In character-based methods I need to know the ancestral sequence (!)
  - An exhaustive search is almost always impossible
    - \* The branch and bound approach is a possible solution: I create an optimal tree with a subset of sequences and I add a sequence at a time
    - \* I can employ some heuristics
- There are methods that combine the approaches: I create a starting tree with neighbor joining and then refine it with other approaches
  - I can also start from a tree supplied from the user
- UPGMA and WPGMA are also called clustering methods
- WPGMA: the distance from a node  $k$  to another node  $u$  is the average of the distances of the children of  $k$  to  $u$ 
  - Weighted pair group method with arithmetic mean
  - When I join 2 OTUs  $A$  and  $B$ , I place them at the same distance from the parent node
  - Now the distance from the  $(A,B)$  node to any other node is the average of the distances from the node to  $A$  and  $B$
  - When joining the node  $(A,B)$  with the node  $(C,D)$ , their distance is the average among the distance  $C$  to  $(A,B)$  and  $D$  to  $(A,B)$
- UPGMA: like WPGMA but the average is weighted on the numerosity of the OTUs under a node

- Unweighted pair group method with arithmetic mean
- Unweighted refers to the fact that each distance contributes equally to the average, so the actual average is weighted on the numerosity (!)
- In an ultrametric tree it gives the same result as WPGMA
- Both WPGMA and UPGMA are really sensitive to differences in rate of mutation among branches (differential branch length from a single split)
  - This is defined as rate heterogeneity
  - When I average 2 sequences I am assuming that their rate heterogeneity is equal (!)
- To overcome the limitation of clustering methods, algorithms based on additive distances were developed
- Additive distances satisfy the four point metric condition for any 4 taxa A, B, C, D that are joined as (A,B) and (C,D)
  - $d_{ab} + d_{cd} \leq \max(d_{ac} + d_{bd}, d_{ad} + d_{bc})$
  - This is because the branch among the internal nodes is always  $\geq 0$
  - This means that I can estimate distances among taxa by summing intermediate distances
- Additive trees are always superior when the tree is not ultrametric
  - This is when the sequences do not follow a clock-like behaviour
- Real dataset can deviate from the four-point metric because of noise
  - In this case I need to artificially add a systematic error to correct
- Minimum evolution (ME) is a tree scoring function that selects the tree that minimizes overall branch length
  - $S = \sum_{i=1}^{2n-3} v_i$
  - There are  $2n-3$  branches in an unrooted tree of  $n$  OTUs, and I am assuming that distances are additive
  - In this method branch length is inferred from pairwise genetic distances
  - An exhaustive ME search is practically impossible with more than 10 sequences because of the numerosity of the possible trees
- Neighbor-joining (NJ) is an heuristic used for estimating the ME tree
  - It is conceptually related to clustering but it does not assume clock-like behaviour
  - It minimizes the metric  $S$  of ME locally, in pairwise comparisons, but it does not guarantee to find the global minimum of the metric  $S$
  - I always start from a distance matrix
  - I calculate for every OTU the net divergence  $r$  as the sum of the distances from the OTU to all the other OTUs
    - \* It is basically the sum of the column of the matrix corresponding to the OTU
    - \*  $r_a = d_{ab} + d_{ac} + d_{ad}$
  - I create a rate-corrected matrix by subtracting from the pairwise distances the sum of the net divergences of the 2 OTUs considered divided by  $n-2$ 
    - \*  $M_{ab} = d_{ab} - r_a - r_b$
    - \*  $n$  is the total number of OTUs
    - \*  $n-2$  are the degrees of freedom
    - \* Note that in this matrix I have negative values
  - Now I join the closest OTUs (most negative score) in the transposed matrix
  - I calculate the distance from the node to the OTUs
  - I create a new distance matrix with the OTUs fused using the four-point condition
    - \* I know the distance of the C from A from the original matrix
    - \* I know the distance from A to the new node because I just calculated it
    - \* The distance from D to the node is thus the difference among them, since the tree is additive
- Maximum parsimony: the tree or set of trees that can be explained with the minimum number of evolutionary changes
  - This criterion follows from the Okham's Razor
    - \* There is no real statistical justification
    - \* It is still useful as a fallback method when computational power is an issue for maximum likelihood methods
  - Parsimony works better when evolution is slow, but this is NOT an assumption of the method

- It is difficult to state the assumptions of a parsimony method, but we can say when it is good and when it suffers
  - \* Parsimony doesn't work well with long branch attraction
  - \* It fails catastrophically in the Felsenstein zone
    - It converges on the wrong tree with increasing certainty as more data are added
    - The Felsenstein zone is when unrelated taxa share more identity than related taxa by chance
- The objective function of MP is the length  $L$  of the tree  $\tau$ 
  - \*  $L(\tau) = \sum_{i=1}^n l_i$
  - \*  $n$  is the number of characters in the MSA
  - \*  $l$  is the length of that specific character
- For every character  $l$  is the number of changes implied by the tree times the cost of each change
  - \*  $l_i = \sum_{k=1}^{2^n-3} c_{a_k b_k}$
- For each tree it produces a statistics known as tree length, which refers to the number of hypothetical 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 likelihood
  - 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 arbitrarily my dataset and see if the tree changes
    - \* Bootstrap analysis: removal with replacement
    - \* Jackknife analysis: removal without replacement
    - \* The support for a node is the percentage of its appearance in the resampled datasets
  - Character 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 Monte 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 evolve 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

## 25/03

- The supergene approach
- The supertree approach is based on joining different trees
  - In this way I can reconcile trees built with different methods (DNA, phenotype, biogeography)
  - An informal supertree is made by joining subtrees
  - A formal supertree involves also possibly mixing OTUs among trees
  - Joining trees based on a consensus needs the subtrees to be compatible
    - \* The trees should not be in conflict with each other
  - I can represent a tree with a matrix of OTUs vs nodes
    - \* I put one when the OTU is included in the node
  - I can create a matrix containing the nodes of both trees to join
- The supermatrix approach