Molecular Phylogenetics

Saul Pierotti

April 5, 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
  - 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 lenght 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 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 by 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 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
  - $p(X->Y) = \mu * \pi_X * a_{X,Y}$
  - μ is the mean substitution rate
  - π_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 choosen so to make the sum of outgoing transitions from each state equal to 1
- Note the assumptions we are making
  - Mutations probablities are only dependent on the immediately preceding state (Markov property)
  - Substitution rates are constant in time (homogeneity)
  - The nucleotide frequencies are at equilibrium (stationariety)
  - 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 $p(X->Y) = p(Y->X)$, 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 Saul, not the transition probabilities (!)
  - I have 4 possible nucleotide frequencies
  - The 2 groups have to sum up to 1, so I loose 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 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
  - An example: the third codon position mutates faster than the first, that in turn mutates faster than the second
  - In general, different positions are snjected 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 α , while when included in a Markov model it is called γ because of the distribution
  - By adjusting the parameter α I can accomodate different degrees of rate heterogenity
  - 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

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 characther-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 developped 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 likelyhood, maximum parsimony are in theis 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 charachter-based methods I need to know the ancestral sequence (!)
  - An exaustive 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 joning 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 distace 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 avergae 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 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 (!)
- To overcome the limitation of clustering methods, algorithms based on additive distances were developped
- Addittive 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 >= 0
  - This means that I can estimate distances among taxa by summing intermediate distances
- Addittive 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 lenght
  - $S = \sum_{i=1}^{2n-3} v_i$
  - There are 2n-3 branches in an unrooted tree of n OTUs, and I am assuming that distances are addittive
  - In this method branch lenght is inferred from pairwise genetic distances
  - An exaustive 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 colum of the matrix corresponding to the OTU
    - $r_a = d_{ab} + d_{ac} + d_{ad}$
  - I create a rate-corrected matrix by subtarcting 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 trasposed 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 addittive
- 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 maximul 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
        · Long branch attraction is when 2 long branches (distant nodes) are clustered because they both diverged a lot from the original, not because they are similar
        · The similarity arises by chance, is more diversity from all the rest
        · It is a systematic error
    * It fails catastrofically in the Felsentein zone
        · It converges on the wrong tree with increasing certainty as more data are added
        · The Felstenstein zone is when unrelated taxa share more identity than related taxa by chance
  - Not all variable sites are necessarily used: only those for which the ancestral state is known or can be guessed
    * Singlets are excluded (mutation observed only in 1 sequence)
  - The objective function of MP is the lenght L of the tree τ
    * $L(\tau) = \sum_{i=1}^{n} l_i$
    * n is the number of charachters in the MSA
    * l is the lenght of that specific charachter
  - For every charachter l is the number of changes implied by the tree times the cost of each change
    * $l_i = \sum_{k=1}^{2n-3} c_{a_k b_k}$
    * In the simplest model the cost is 0 if the position is conserved, 1 otherwise
  - The costs can be represented by a cost matrix
    * The matrix is symmetrical, so that the lenght of the tree is constant regardless of the position of the root
  - There are dynamic programmim approaches for finding the optimal tree
- Branch and bound: an algorithm for exhaustive search that prunes some of the possibilities
  - It can be applied to MP, but it can use any optimality criterion
  - In practice it is applicable for 15-25 taxa
  - I do an exhaustive search by progressively adding taxa to the tree
  - I keep track of the best solution until now while doing so
  - If I find a subtree with not all the taxa added that is worse than the best tree that I have, I stop evaluating it
    * It will never give a result which is better than the one that I have
  - It is like α/β pruning
- Maximum likelyhood: optimize the likelyhood of observing the data given the model
  - Likelihood is a posterior probability: it is the probability of the dataset given the model
  - For a state of a particular position, its probability is evaluated as the ration among the count for the state and the total count for the position
    * $\theta = h/n$, where h is the count of the state and n the total count
    * This is a probability, not a likelyhood, I am now building the model given the dataset
  - The likelihood of the tree L is the product of the likelyhood of all sites s
    * $L = \prod_j s_j$
  - Since L is usually really small, -Log L is usually used as an optimality criterion
  - It is used much more than maximum parsimony

Nodal support

- Given discording predicitons, I can elaborate a consensus tree
- A strict consensus tree has polytomies for each disagreement
- A consensus based on the majority rule chooses the nodes that appear in most fundamental trees
- We need a measure for node statistical significance: nodal support
  - There is a sampling bias in tree reconstruction: we cannot sample the entire population of a species
  - I never know if my sample is representative of the population
- We distinguish broadly resampling techniques and character-based approaches

- Bootstrap analysis: random resampling with replacement
  - It is useful when I don't know the sampling distribution and I cannot derive it
  - I approximate the real distribution by resampling
  - Bootstrapping means to take a subset of the columns of the MSA with replacement until I get an alignment of the original lenght
    - Since there is replacement I can have the same column twice and not have some columns
  - The new shuffled and resempled alignment is called bootstrap replicate
  - I create in this way a series of bootstrap replicates for my dataset
  - I can get a tree from each replicate
  - The statistical support for a node is the fraction of tree replicates in which the node is present
  - Bootstrap results cannot detect systematic errors in tree reconstruction
    - It is not suitable when LBA (long branch attraction) is likely
  - They also cannot detect a biased sample
- Jacknife analysis: random resampling by independent removal
  - It is similar to bootstrap analysis, but I randomly remove one (or more) of the sites in each resampling
  - My new subset is thus shorter than the original
  - Like in bootstrap, I can get many subsets and I create a tree for each of them
  - The jacknife support for a subtree is the fraction of times it appears in the subsets
- In both bootstrap and jacknife I should be suspicious when the support is under 70%
  - 200 to 2000 resamplings are usually recommended
- Bremer support or deacy index: a method for testing maximum parsimony
  - The decay index is the difference in lenght between the shortest tree and the shortest tree that is incompatible with the node
    - In MP the lenght is measured as number of mutations
  - I start from the best tree in MP and I count the number of steps needed to make a node collapse
  - If I need 2 additional mutations from the best tree to make a node disappear, the decay index is 2 — so far so good

Bayesian analysis

- It is impossible to go from $p(d|M)$ to $p(M|d)$ without knowing $p(M)$
  - $p(M)$ is called prior in bayesian jargon
  - $p(M|d)$ is a posterior
    - It is the probability of the prior update with the available data
- If I don't know anything about the prior I can use a uniform distribution
  - I consider all possible values equally likely
- If I calculate a tree with Bayes, there is no need for validation
  - What I am using for evaluating trees is already the probability of the tree being correct (!)
- In this calculations I can ignore $p(d)$ since it is constant for all models
  - $p(d) = \sum_M p(d|M)p(M)$
- I can therefore assume that $p(M|d) \propto p(d|M)p(M)$
- 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

25/03

- The supergene approach
- The supertree approach is based on joining different trees
    - In this way I can reconsile trees buil 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

DNA sequence databases

- Databases are useful for making sequences freely available and for independent validation
- If I take a sequence from a database, I have a reference for it (!)
- I can find information about the taxonomy related to a sequence
    - NCBI is not autoritative for taxonomy, but still gives an useful indication
- I can find metadata about the sequence

  - In some cases I can also find in which museum the original speciment is conserved (!)
- The BOLD database contains a DNA barcode for many species
- RNAcentral was a database about RNA sequences that now is discontinued
  - There are many alternatives for studying non-conding RNA sequences
- NCBI has many resources
  - Genbank is directly submitted by the user and not validated by NCBI staff
  - Refseq is reviewed and annotated by NCBI staff
  - Pubmed is good for biomedical papers