Genes and their functions

Genes are a relatively recent discovery. Genetics is a field that started without a proper knowledge about what the genes were (studies of Mendel).
The concept of genes was mentioned but there was no clue about which was the physical basis.

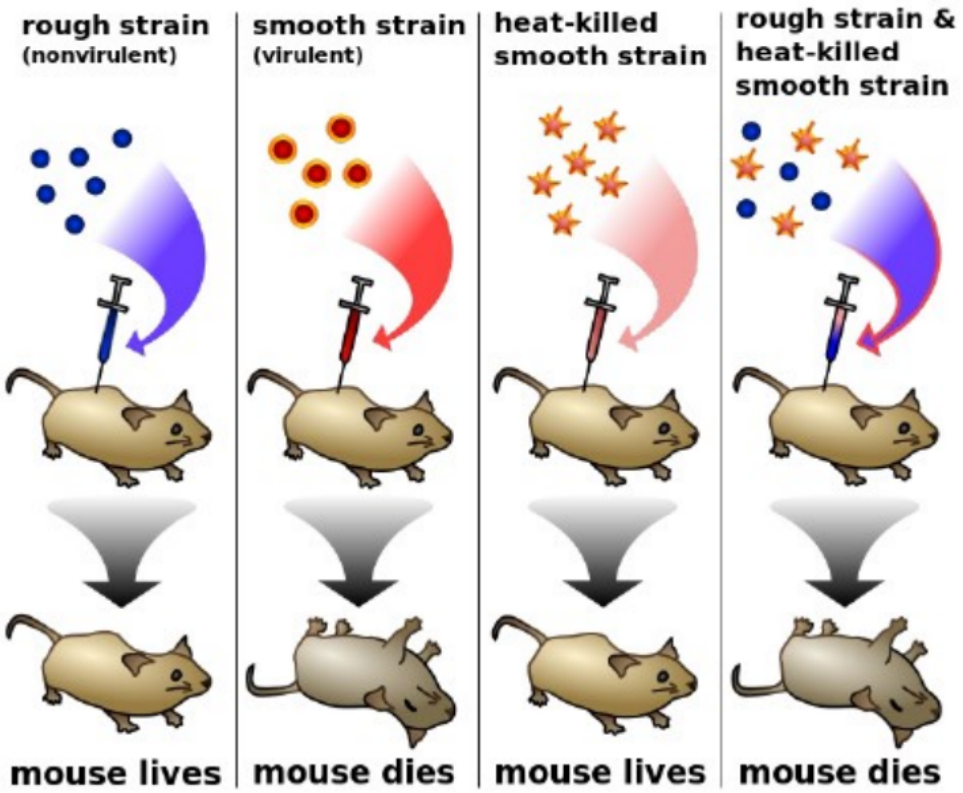
In 1940s there were some key experiments:

- Avery-MacLeod-McCartly experiment with *Streptococcus pneumoniae*
- Hersey-Chase experiments with bacteriophages

Both demonstrated that Nucleic Acids (DNA/RNA) are the substrate of hereditary information (not proteins).

In the first experiment they had 2 strains of *Streptococcus pneumoniae*. One of them was virulent and the other not.

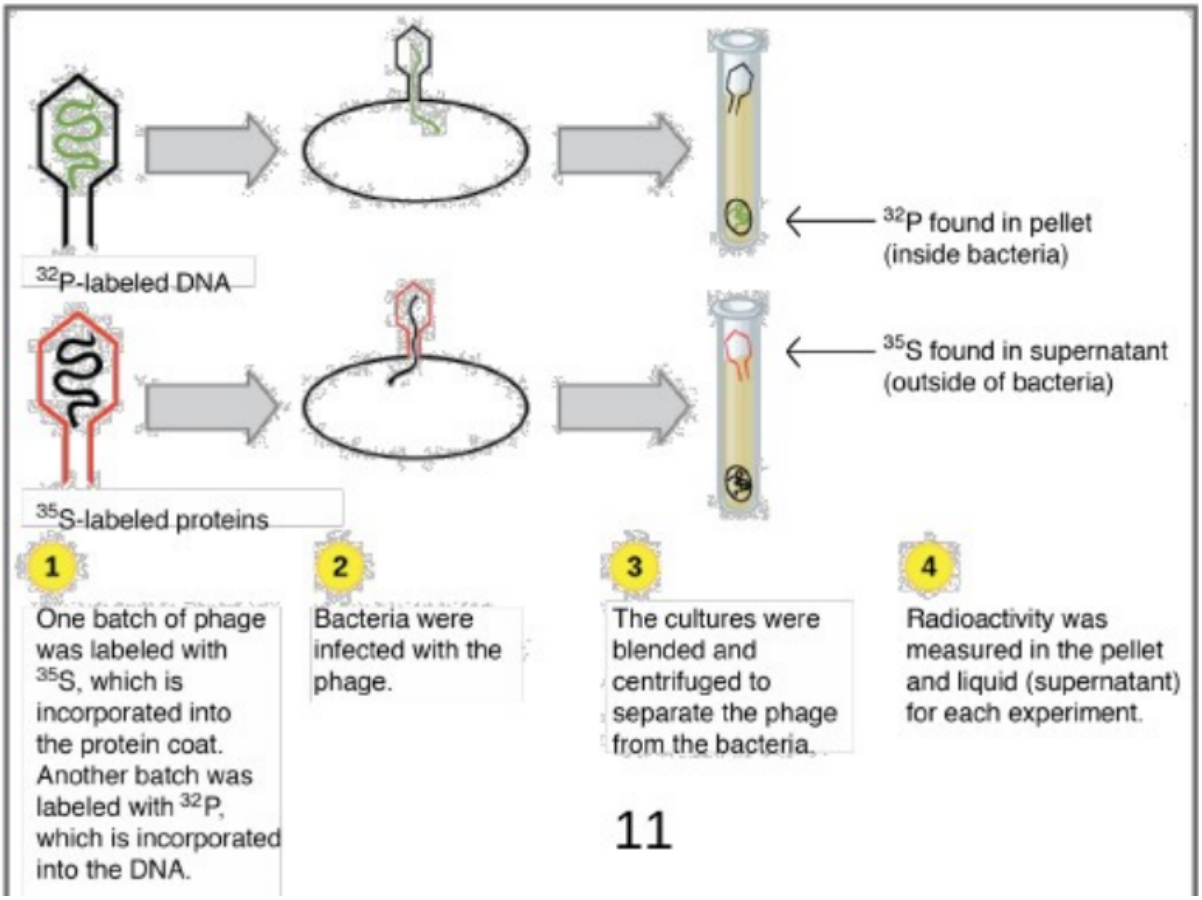
- When co-incubating the virulent (treated with heat) and not virulent strains, the mice dies. The assumption is that the DNA is passing the information because the proteins are less stable and thus they are degraded with heat.



In the second experiment, they used two types of bacteriophages:

- DNA labeled (phosphates)
- Proteins of the capsid labeled (sulfurs)

They infected the cells with the bacteriophages. Later, they saw that the material that remained in the infected cells was the DNA (the proteins were not in the bacteria, but in the supernatant).
Thus, the information carried by the virus was contained in the nucleic acids.



Gene: Any sequence of DNA or RNA that codes for a molecule that has a function

- Promoters, enhancers… are excluded.
- Pseudogenes are excluded because the product has no function.

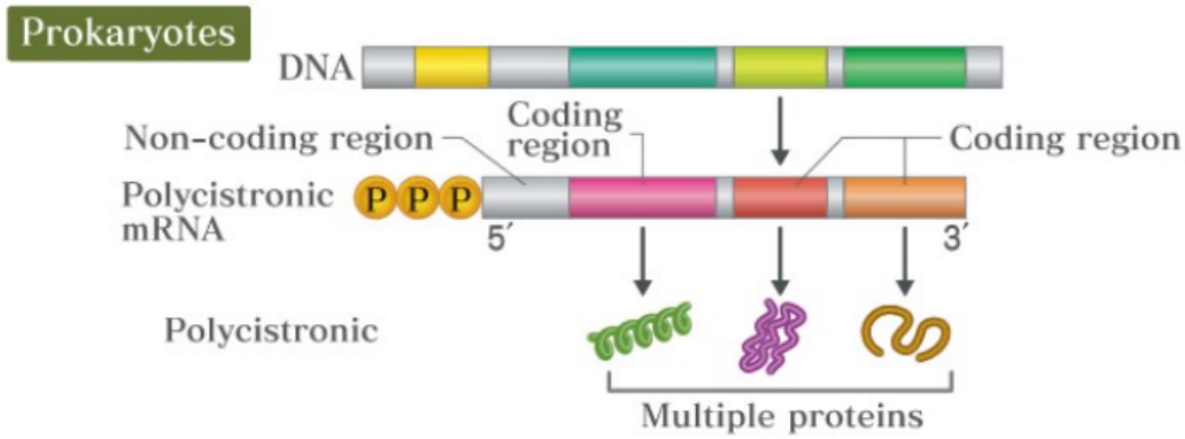
When talking about genes, the DNA/RNA itself does not have a function. It codes for a protein that has a function. In fact, promoters and enhancers have a regulatory function (reason why they are not included).

ncRNA: They already have a function and they are not translated. Like the tRNA, rRNA…

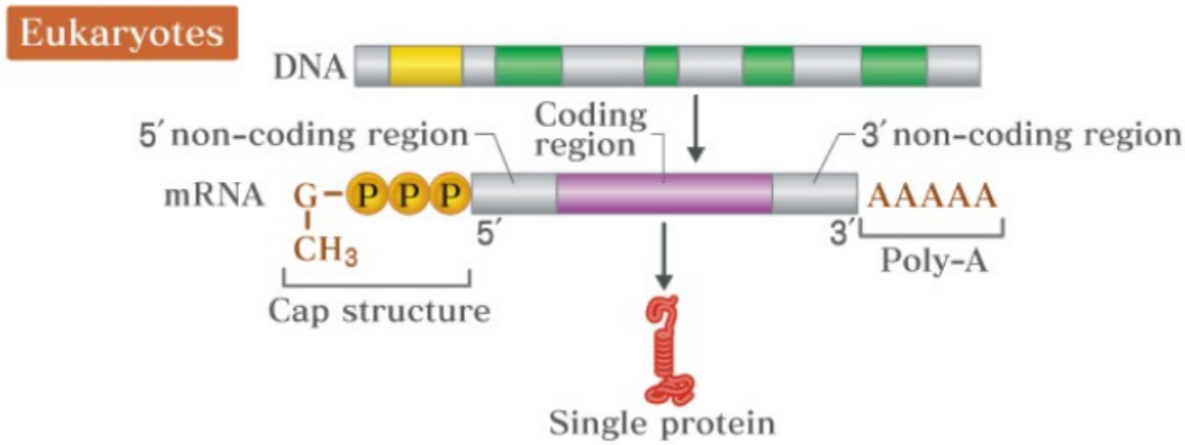
Gene structure and expression

Prokaryotes (cells without a nucleus): Sometimes different genes are transcribed together.

- There is no single mRNA for each gene but a polycistronic mRNA that carries the information for more than one gene.
- This ensures that all the genes of a certain pathway are expressed at the same time (the stoichiometry is 1 to 1).



Eukaryotes: The genes are not in one continuous piece but interrupted by introns (which will be removed during the splicing).
Transcripts are modified by adding a Poly-A tail, a Cap structure (distinguishes mRNA so that they are not degraded).
The RNA goes to the ribosomes (out of the nucleus).



Enhancer and Silencer

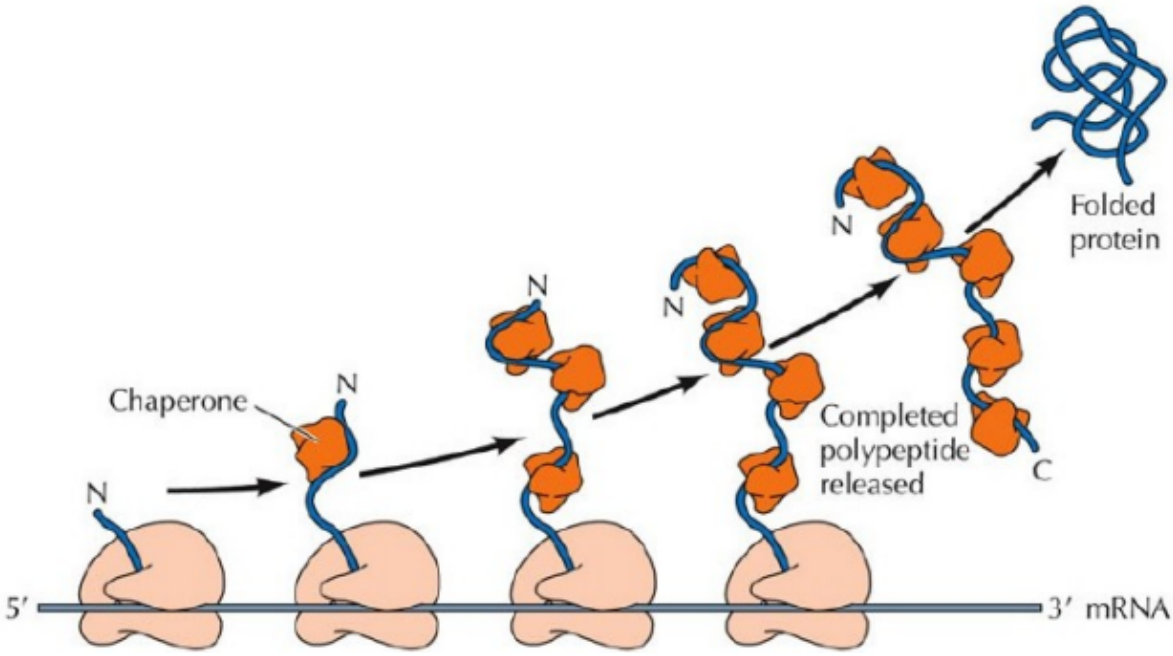
The genes must be expressed in order to synthesize the proteins.
Thus, they are activated when they are needed. This gives information about the function of the gene.
The expression of the genes is tightly controlled by enhancers and silencers which forces the DNA polymerase (transcription factors…) to bind or be removed from the DNA.

Splicing

In eukaryotes, there is splicing.
One gene may code for more than one protein and thus, for more than one function.
Depending on the environment (time, tissue…), the splicing acts in one way or another in order to have the correct final product.

Translation

Finally, the transcript goes out of the nucleus to be translated in the ribosomes.
Many ribosomes go to the mRNA one after another to synthesize the protein.
The chaperones help to fold the protein correctly.



Functional roles of genes

- Structural (Actin)
- Catalytic (Glycogen synthase)
- Regulatory (Transcription factors)
- Etc

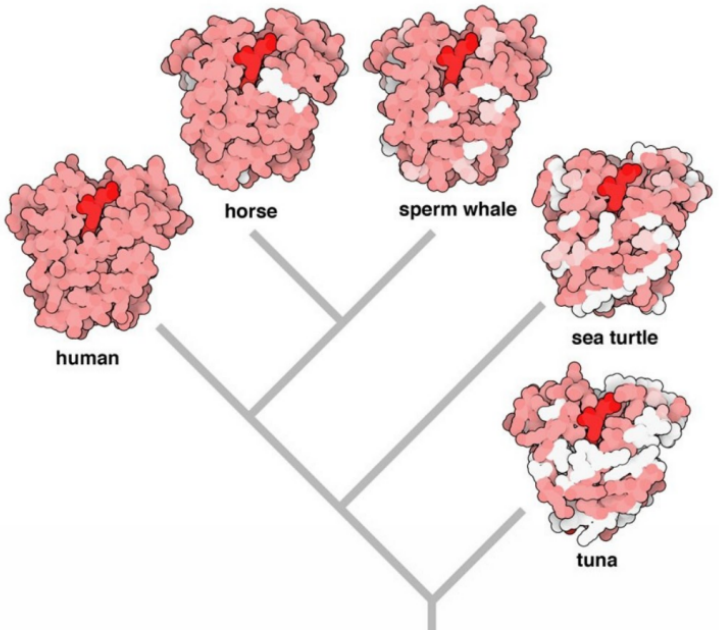
Some genes have essential or non-essential functions (if you remove the gene, the cell will die). Some genes are constitutive (always expressed)...

There are different levels in which you can talk about the function of a gene:

- **Molecular function:** It refers to the actual roles of the molecules.
- **Cellular function:** It refers to the role you have in the cell. You can have the same molecular activity and have different cellular functions.
  For example: 2 different transcription factors can be found in different pathways.
- **Phenotype function:** It refers to the effect on the morphology, physiology… of the organisms.

What determines the function of the gene?

It is determined by the structure of the molecule (protein or ncRNA) that is doing the function. The structure is determined by the sequence.
Since each sequence has a propensity to fold in a given way.



The regions that are essential for the function of the protein are more conservated.
We can just compare the hemoglobine different species. We can see that the heme group is more conserved.

Also notice that the sequence may be different but the key structure remains the same.

Homology based functional inference

It considers that the sequence determines the structure and the structure determines the function. Thus, by just having the sequence we can compare it to the sequence of another protein from which we know the structure (or function) and deduce the structure (or function).

Only 43% of E.coli genes are experimentally characterized. We normally do predictions.

The homology-based prediction is good at predicting molecular function but not at higher levels (cellular and phenotypic functions).
Also be aware that few residue changes can drive changes in substrate affinity…

Protein domains and domain shuffling

Domain: Conserved part of a protein sequence and structure that can evolve, function, and exist independently of the rest of the protein. Each domain forms a compact 3D structure and often can be independently stable and folded.
Some domains are promiscuous meaning they can appear in diverse families in combination with other domains.

Proteins can have a modular structure, in which different parts of the protein do different things. The protein does not act as a whole.
The different domains do different functions.

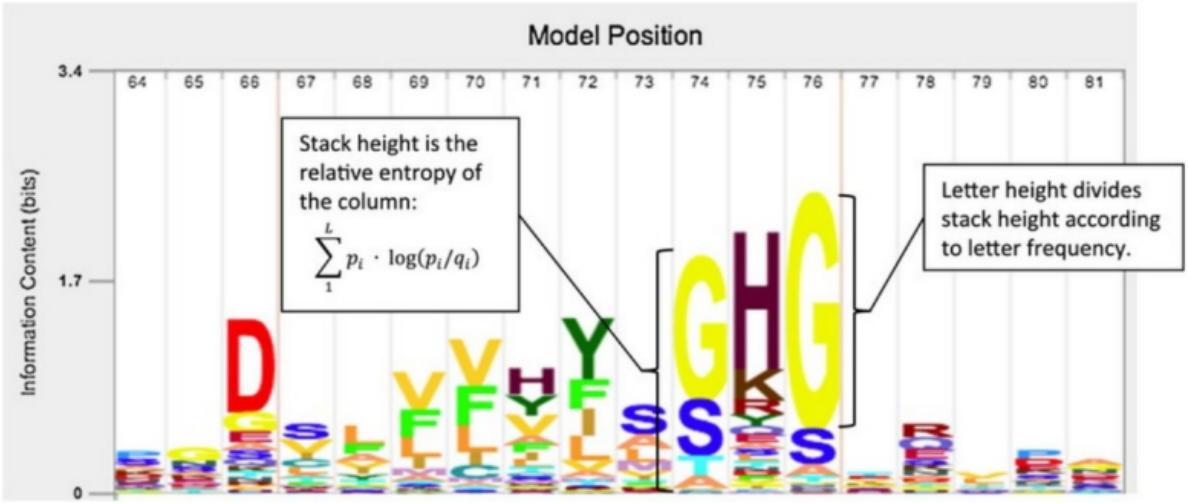
For example: A transcription factor may have a domain that is devoted to the binding to the DNA and another domain that attracts other proteins.

This can lead to confusion when doing a BLAST (homology based predictions). Because maybe the similarities of a protein is only restricted to a single domain.

Resources:

- Pfam
- SMART
- InterPro

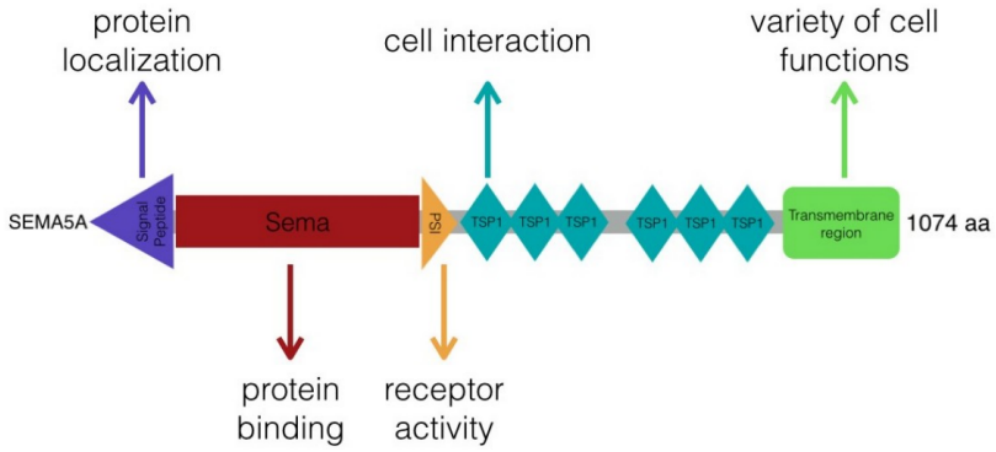
Another way of annotating a protein is to run our sequence against the databases above (store information about the domains).



So, we can compute the likelihood of having a certain sequence.

Prediction of protein subcellular localization

There are domains (motifs) in the N-terminal that indicate the subcellular localization. They are signal peptides.



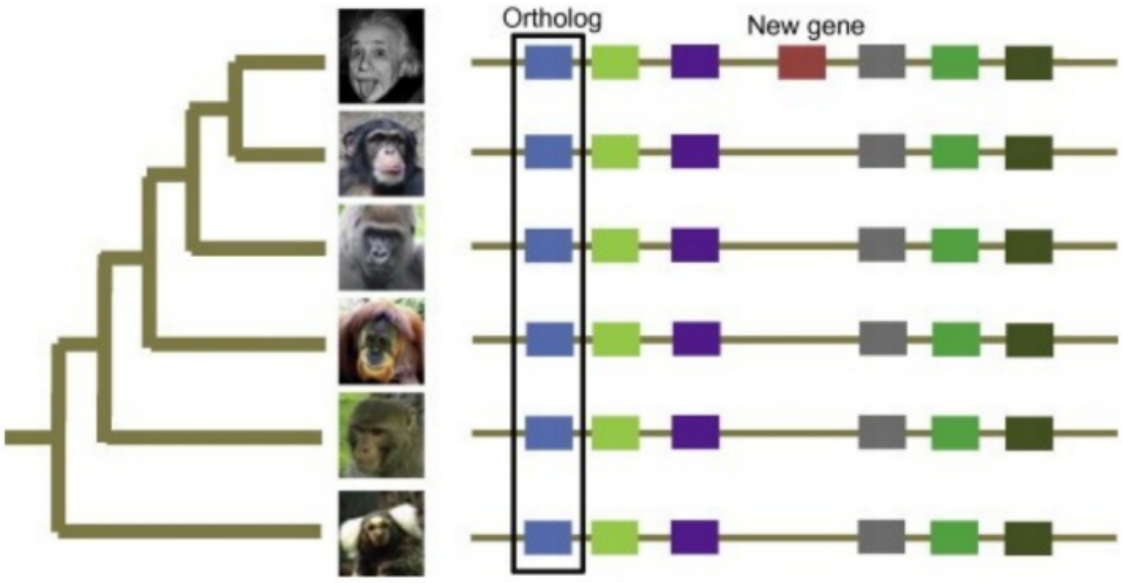
This gives a lot of information about the function of the protein.

De novo origin of genes

Process by which new genes evolve from DNA sequences that previously were not encoding for functional molecules.

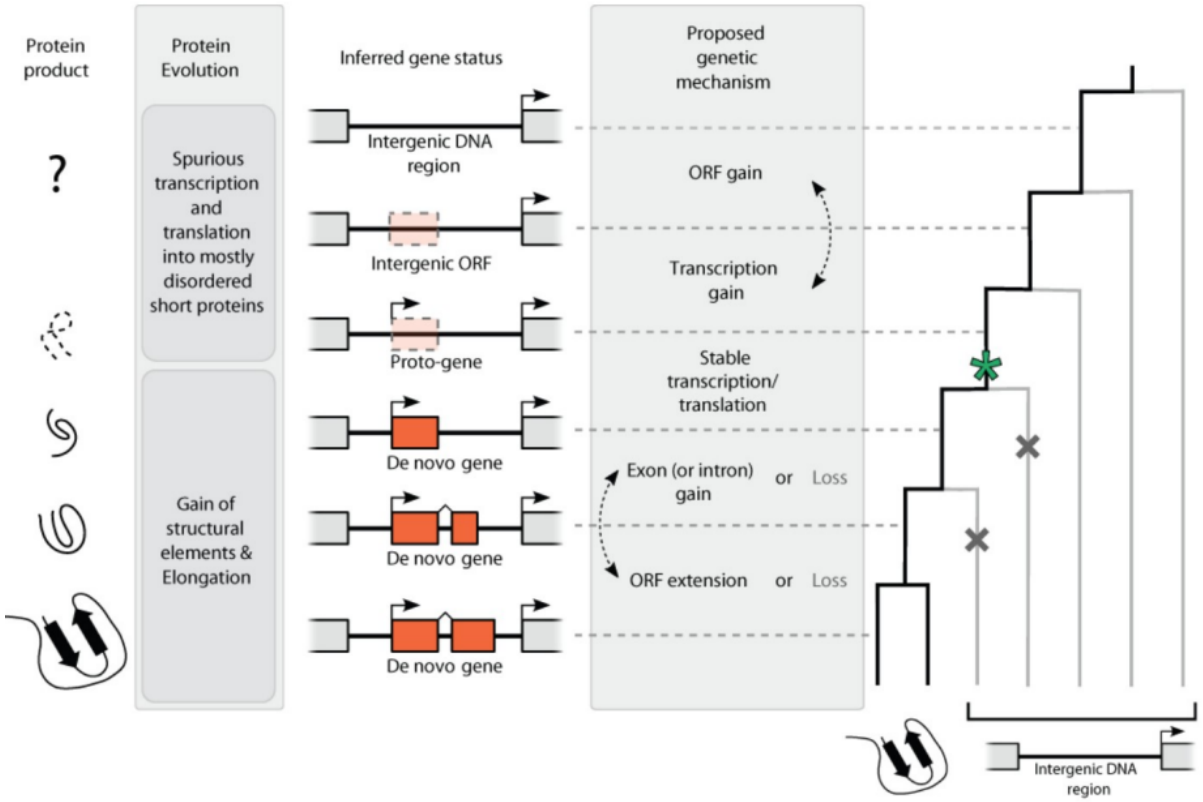
Genes can originate from nothing.

We can see this by just making an alignment of the genome of closely related species.

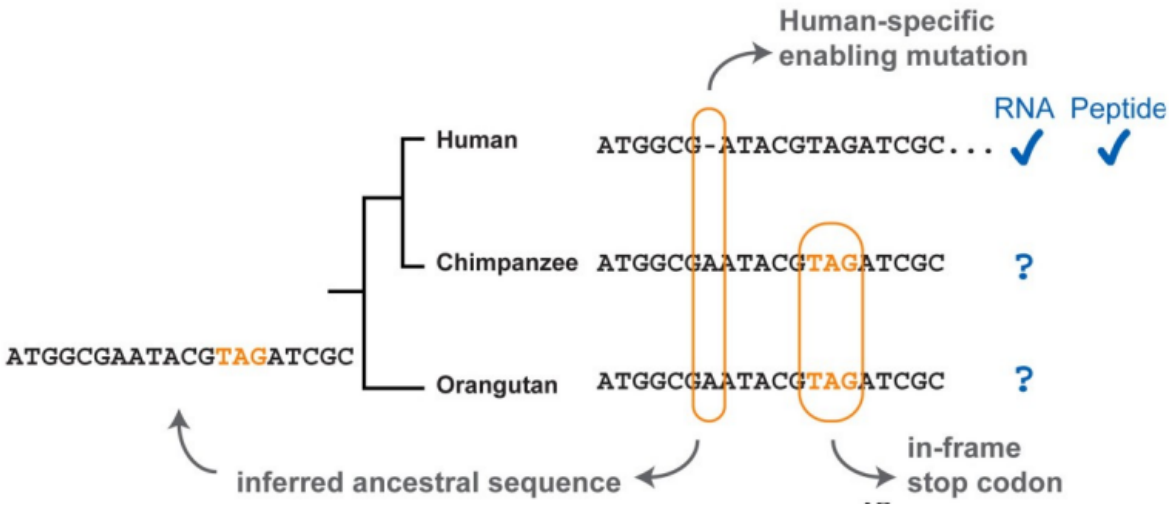


Genes can also be duplicated and accumulate mutations that provoque the formation of a new gene.

Creation of a gene: By chance, an intron can have a mutation that leads to the creation of a gene. If there is an advantage, it is selected.



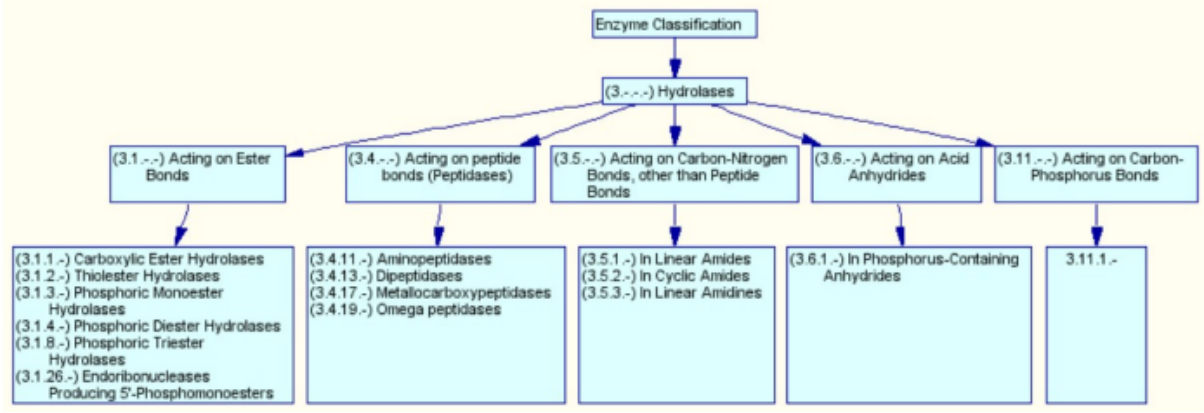
There can also be a deletion, addition of a nucleotide that changes the open reading frame. Thus, the STOP codons are not readed and we obtain a larger or smaller gene.



Enzyme Classification

Enzymatic Commission number (EC): Each number is a hierarchical description of the function of the enzyme.

EC 1.3.1.21



KEGG is the database for pathways.

Gene Ontology

System that describes functions and relates them with others.

- It provides a controlled vocabulary of gene and gene product attributes
- Annotate genes and gene products, and assimilate and disseminate annotation data
- Provide tools for its easy access.

It has 3 different ontologies (the study of ‘being’):

- Molecular function: An elemental activity
- Biological function: A commonly recognized series of events
- Cellular component: Where a gene product is located

Homology, paralogy and orthology

Analogous structures: Similar function but different origin. For example, the wings of bats and birds (they don't have a common ancestor).

Homology: Similarity of the structure, physiology, or development of different species based upon their descent from a common evolutionary ancestor.

The same organ in different animals under every variety of form and function.

Extension of the concept of homology to sequences: Two sequences are homologous if they share a common ancestor.
The problem with sequences is that we do not have a fossil record and thus, we can not get the sequence of the common ancestor of birds and reptiles.
To check if 2 sequences are homologous we must do an alignment . We check if they are really similar and we make an inference. We make an hypothesis that they come from the same ancestor.

But how similar must it be to consider them homologues? Here we define the concept of similarity (degree of likeness between two sequences, usually expressed as a percentage of similar or identical residues over a given length of the alignment). We can not say that the sequences are 50% homologous, but have a similarity of 50%. Because you share a common ancestor or not (there is no middle way).

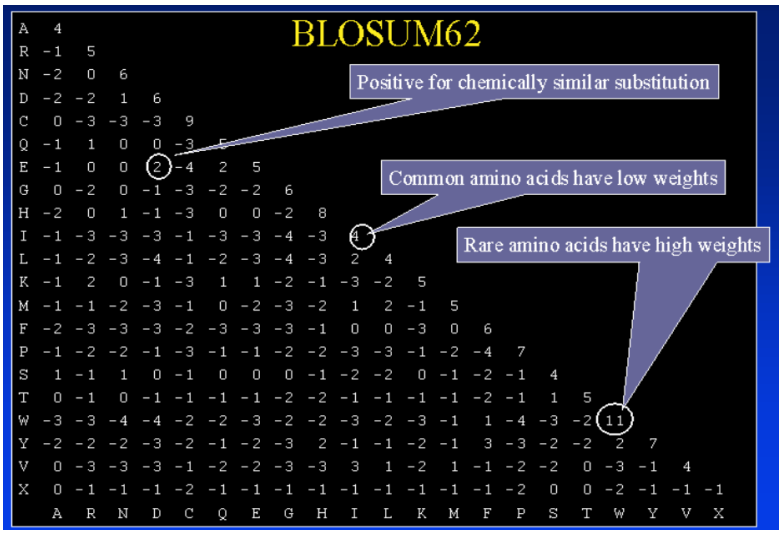
```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Are these 2 sequences significantly similar? To answer this, we must check how likely it is that such alignment is the result of chance. Thus, we do not have to look at the % of similarity but the E-value.

The **score** of a blast is computed using the substitution matrices like the Blosum62.
If in one position there is a different amino acid that has different properties, it will give a high value, for example.

This is much more relevant than the identity. Because we are considering if the aa have similar



properties (not only if they are the same).
The **P-value** is the probability of obtaining the same alignment by chance (it goes from 0-1).
The **E-value** is the expected number of alignments with this score or higher that you would expect by chance when comparing such sequence against this database (it can go from 0 to a positive number).

The process of obtaining the E-value is the following:

- We align different random sequences from a database and we obtain a distribution of scores.
- We compare our score to obtained distribution.
- If the score is in the expected distribution we do not trust it

Thus, the E-value depends on the database you are using (not the score).

When doing a BLAST, you can put "**Low Complexity Filtering**". It refers to the regions in the protein that have low sequence complexity in the terms that they have the same aa or several aa repeated many times.
This is usually removed because they can appear in unrelated sequences because of expansions of some repetitive regions (they have nothing to do regarding the origin).

When introducing the fasta of a single protein, we will obtain multiple alignments. This is due to the fact that BLAST is a local alignment tool. So, it is not trying to align the whole sequence but trying to find optimal alignments (even if they are sub-alignments).

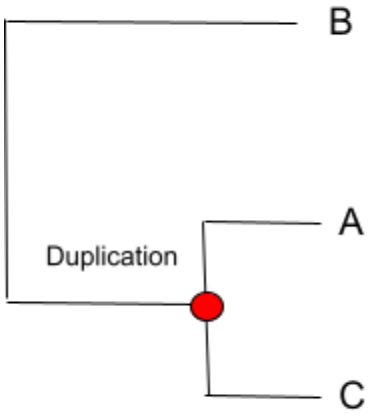
All homologous genes come from a common ancestor. But there are 2 fundamental ways in which 2 genes can evolve from the common ancestor:

- **Orthology:** When the homology is the result of speciation. There is a creation of a new species!
- **Paralogy:** When the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism. In the same species, there is a duplication and they both accumulate different mutations.

All these sentences are wrong

- Orthologs are homologous genes that have the same funcion.
- Orthologs are homologous genes in different species, while paralogs are homologous genes in the same species
- The ortholog is the most similar sequence among the homologs in another species
- Orthologs are genes that do not duplicate and, when they exist, they are always present in single copy
- After a duplication, the orthologous copy is the one that keeps the function of the ancestral gene

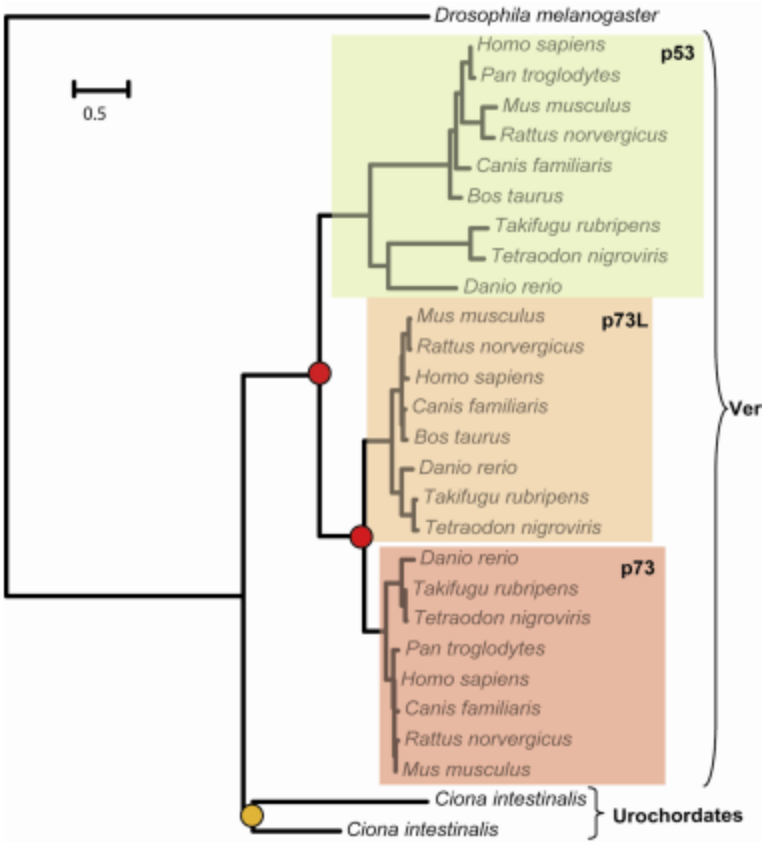
Demonstration that this sentence is wrong: If gene A is orthologous to gene B, and gene B is orthologous to gene C, then A and C are orthologous to each other.



In this case, A is orthologous to B, B is orthologous to C but A is not orthologous to C (they are paralogs)

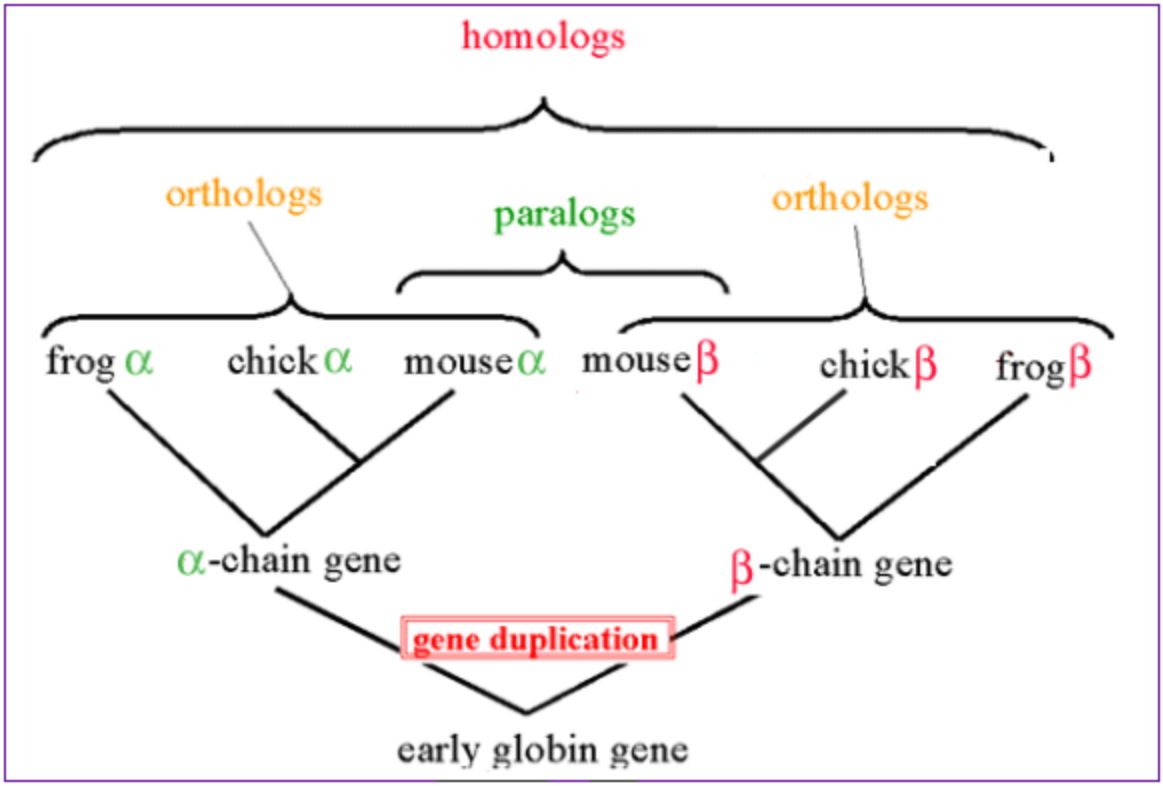
In a real case:

B is Drosophila melanogaster
A is Homo sapiens
C is Canis familiaris



Example

Hemoglobin belongs to a globin family. Initially it was a single gene that duplicated and formed 2 different chains (alpha and beta) in all vertebrates.
Both chains evolved differently and they specialized in different functions.



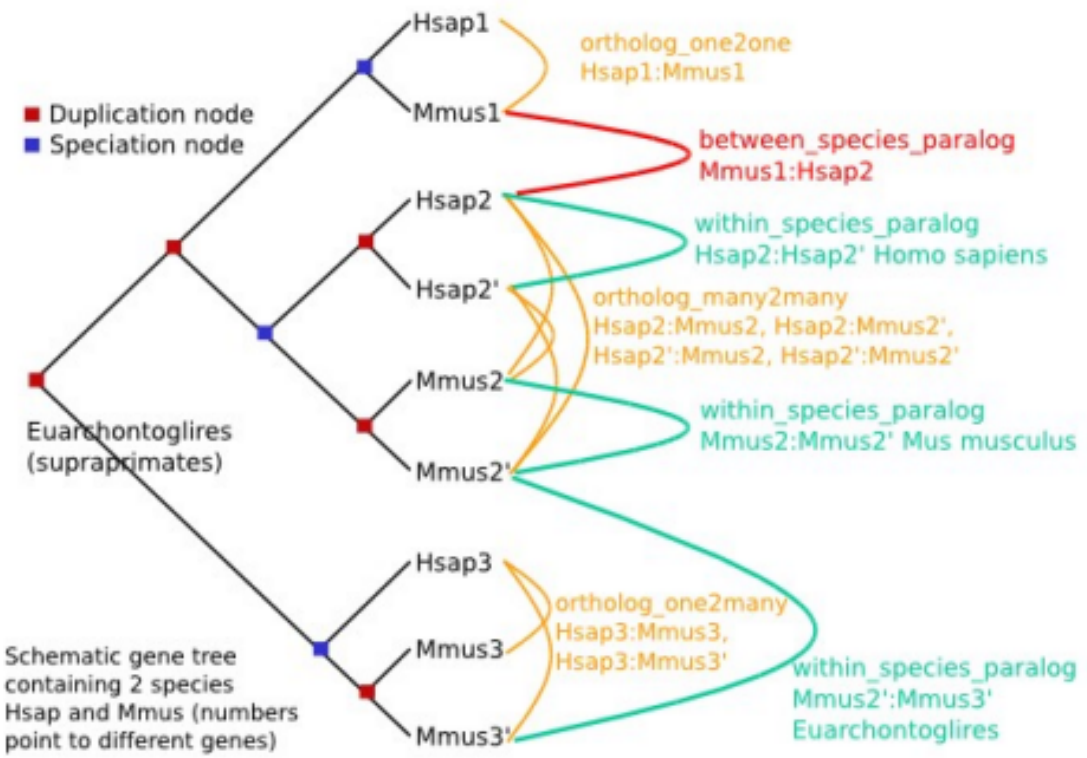
As we can see, all vertebrates have an alpha and a beta chain.
According to the definitions that we mentioned above:

- The alpha or beta chains of different species are orthologs. Because they are a result of speciation.
- If we compare the alpha and beta chains of any species, we can say that they are paralogs. Because they come from a duplication.
    - Mouse alpha and chick beta are paralogs, for example.

Remember

- Orthology definition is purely on evolutionary terms (not functional, not synteny…)
- There is no limit on the number of orthologs or paralogs that a given gene can have
- Many to many orthology relationships do exist (co-orthology)
    - 2 human genes can be orthologous to 2 rat genes
- No limit on how ancient/recent is the ancestral relationship of orthologs and paralogs
- Orthology is non-transitive (as opposed to homology)

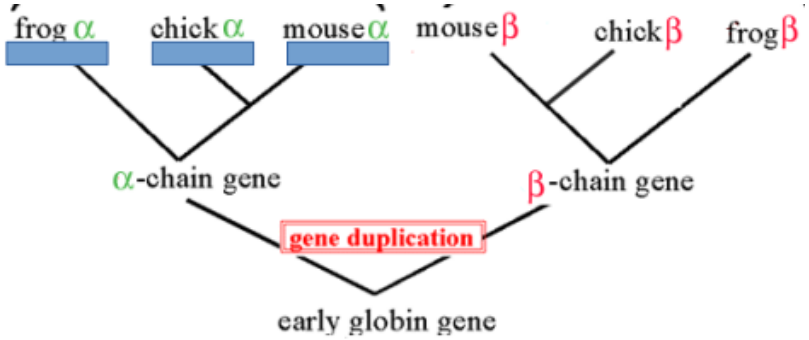
Here we have 2 species and a gene family that has duplicated several times and a single speciation event that produced both species.



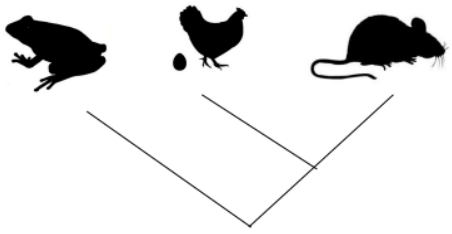
Why is predicting orthology important?

- **Important implications for phylogeny:** Only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions). Because they show the speciation events.

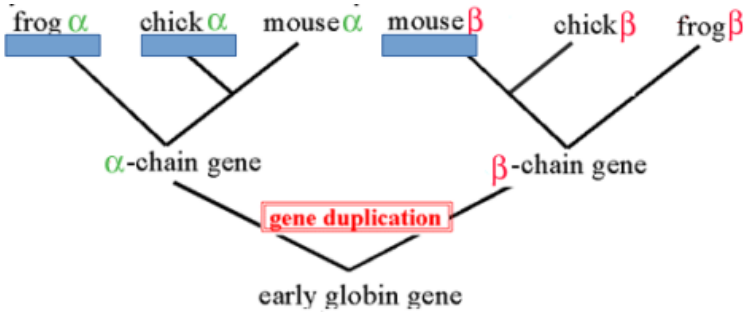
  Example: Imagine that we take one gen from frog, chicken and mouse.
  All 3 genes are orthologous to each other, the alpha chain for example.



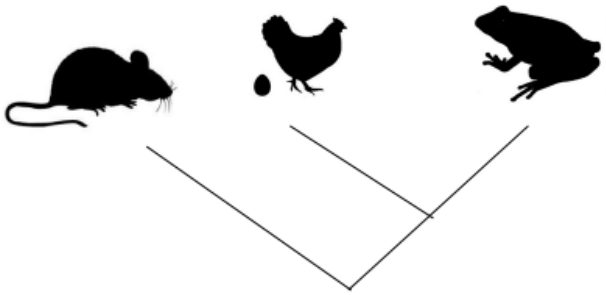
If I make a phylogenetic tree with the 3 sequences I will retrieve this tree.



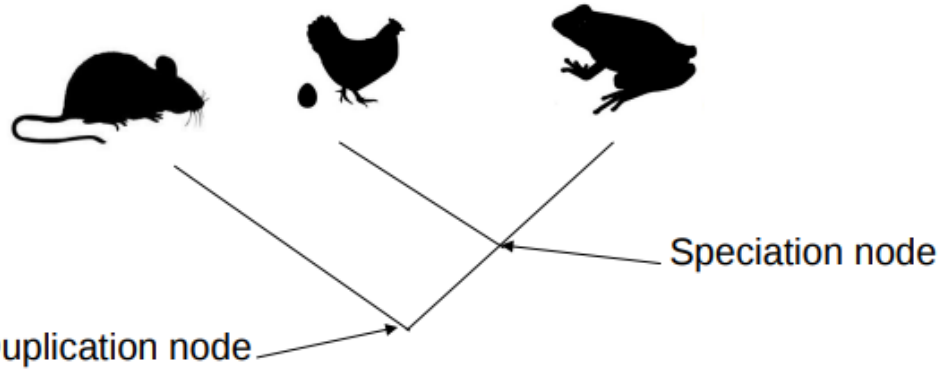
We would not obtain the same result if we use 3 homologous (not orthologous) genes.



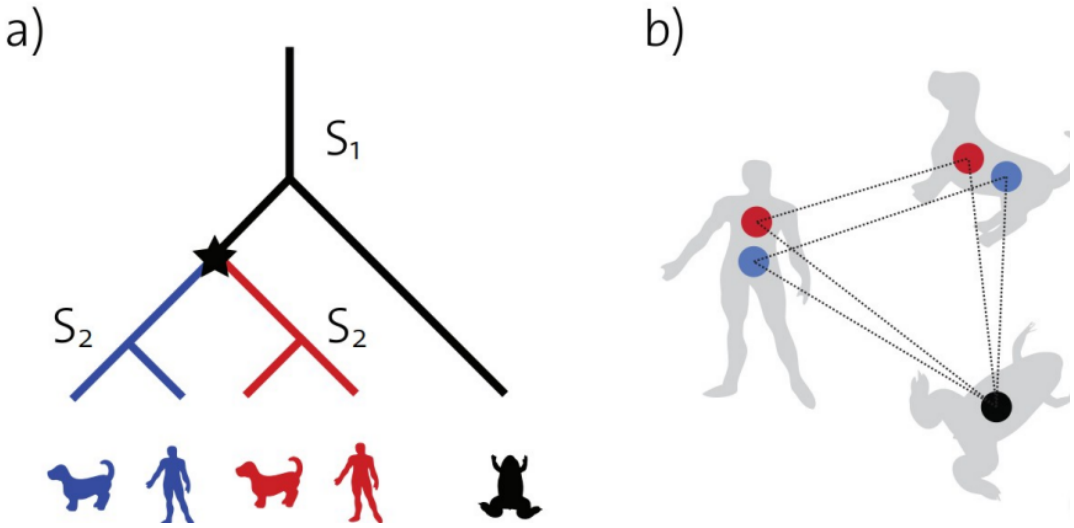
We obtain this wrong species tree:



This is because this is not a species tree.



- The most exact way of **comparing** two or more genomes in terms of their gene content. Imagine that we have this tree (human, dog, frog).



There is a speciation, a duplication and another speciation.
Only by resolving the orthology relationships, we could see that these 2 human genes are equally related to the frog gene. The frog has not missed a gene.

If I work at the level of homology, I would paint all the genes with the same color and I would not know which gene is orthologous to each other. I also would not know if there has been a deletion of a gene in frogs or a duplication in dogs and humans.

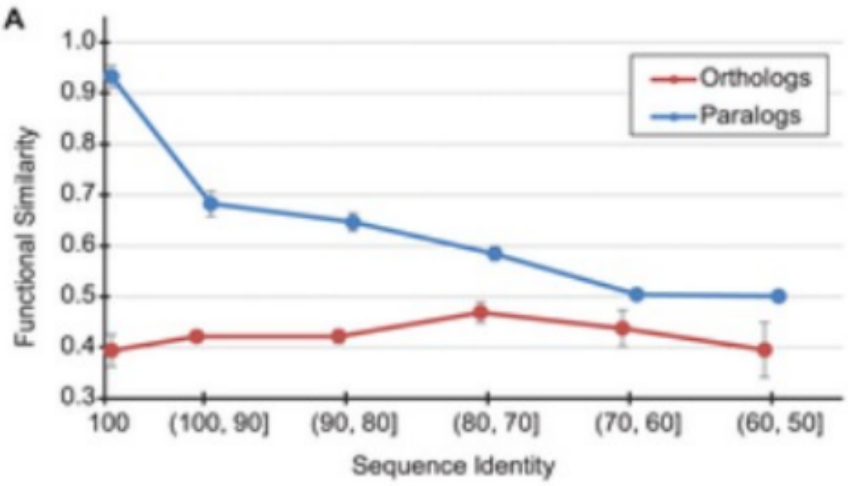
- Implications for **functional inference**: orthologs, as compared to paralogs, are more likely to share the same function.
  Example: We have a gene that has 4 functions and this gene suffers a duplication. Thus, there is a redundancy in the genome (this is not good unless having more quantity of a gene is important). With the time, 3 things can happen:
  - The most common thing is that one copy is lost. The copy incorporates a deleterious mutation that produces a non-functional gene, the gene is turned into a pseudogene and then it is lost.
  - **Neofunctionalization:** One copy adopts a new function
  - **Subfunctionalization:** Both genes coexist and one gene does some of the functions and the other copy makes the rest. Thus, both genes need to be retained.

The process of paralogous retention has something to do with function changes.
Then it is clear that the process of gene mutation and duplication is often associated with processes of changing function. Thus, we must rely more in orthologs than in paralogs to annotate the function.

Note that orthologs can also change the function. An ortholog protein in a fish and in humans is likely to have a different function.

Some scientists did a study regarding this topic, where they searched in GO the function of some paralog and ortholog genes.

They obtained the following graph:

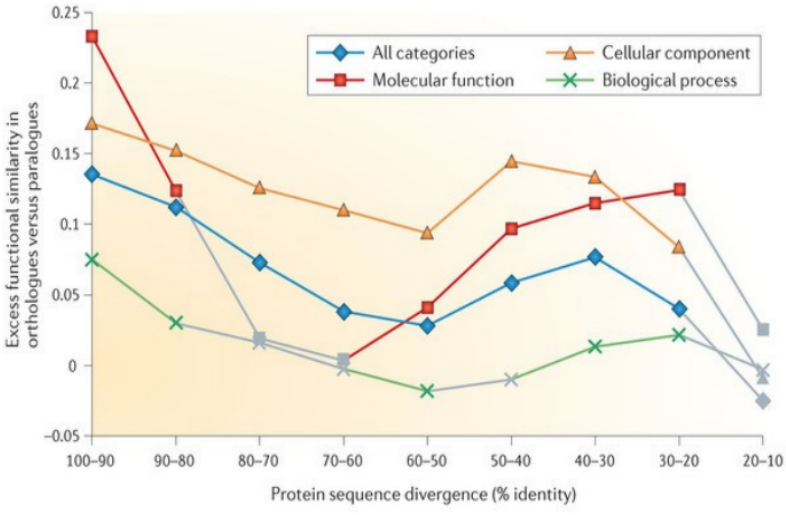


We can see that paralogs are functionally more similar. So, we are obtaining the opposite result.

They only used experimentally determined annotations. The problem with this is that these annotations are done with model species. Also, different groups work with different gene families.
The problem is that there was a bias. If the protein was annotated in the same paper, it had a high GO term similarity.
Each lab is interested in one type of function.

There was another bias. The labs work with a single species, mouse for example. So, they will test the members of the family which will be paralogs to each other (they are in the same species). They won't do the experiment with the human protein because they work with the mouse system.
So, the paralogs are annotated by the same group and thus there is a higher probability of having the same annotation.

When correcting the bias, we obtain the following graph. Here on the Y axis we are doing orthologs minus paralogs. Thus, if it is higher than 0, they are functionally more similar.
But, as we can see, it is not black or white (orthologs can also change function).



Gene families

Group of genes that share a common ancestry (they are homologs).
They have a hierarchical evolutionary relationship (best represented by a tree)
Members of a gene family can be orthologs or paralogs between them
An orthologous group is a (or part of) gene family
Gene families evolve by duplication and loss (birth and death)

Because of loss/duplication dynamics, gene families will vary in size and phylogenetic distribution.
Example: More than 518 protein kinases only in humans (they are involved in different pathways, so they do not have the exact same function, but similar).
We can use homology base function prediction if 2 proteins come from the same gene family, because the function will be similar but not the same.

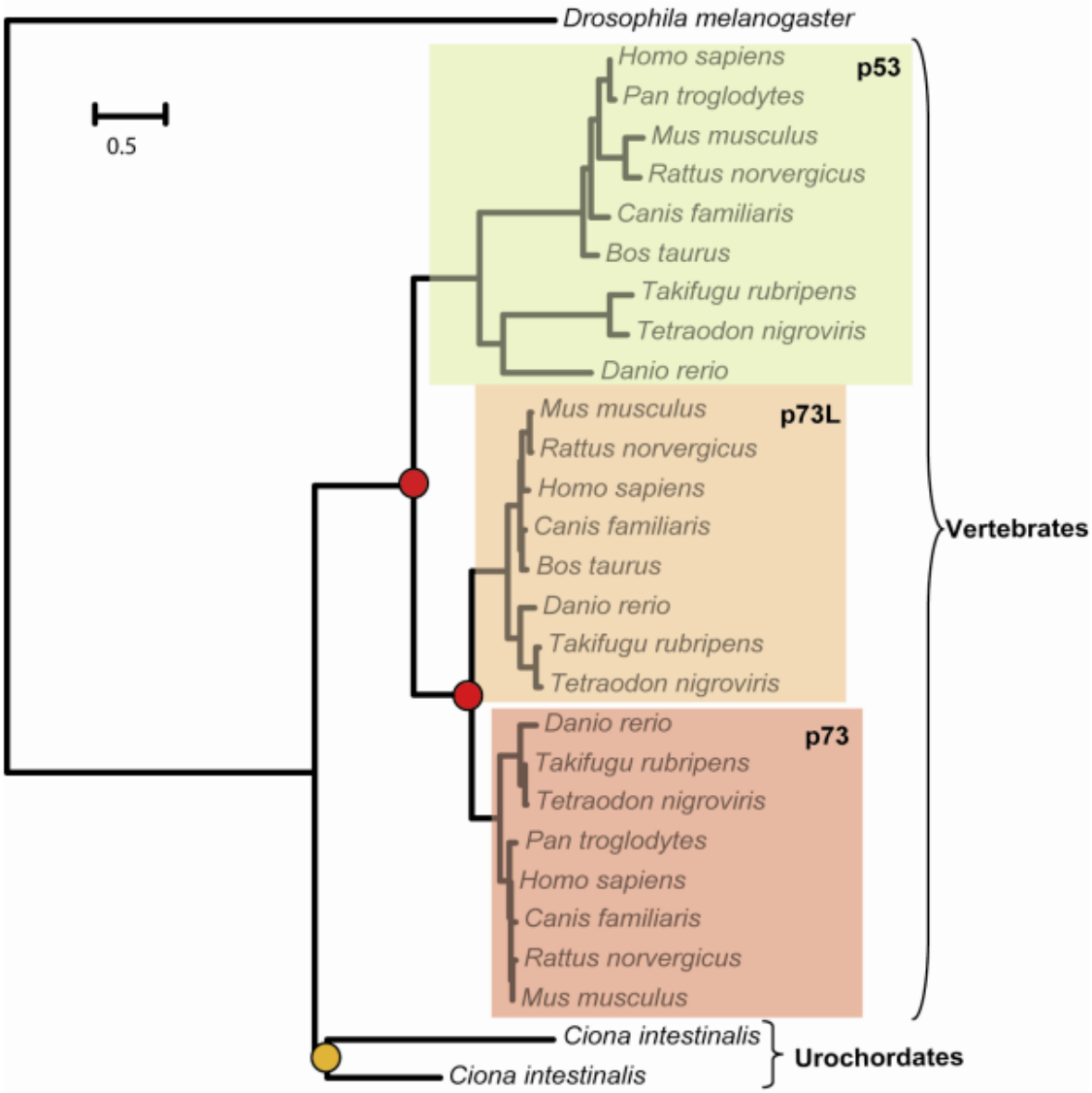
Orthology prediction methods

To predict homologues, we did alignments and we checked if the alignment was significant by looking at the E-value…
For orthology, we use other methods.

Making a phylogenetic inference is the classical approach.

- Build a gene tree
- Compare to the species tree
- Infer duplications and speciation events
- Assign orthology and paralogy relationships accordingly

Question. What is the orthology relationship between homo sapiens and Ciona intestinalis. In other words, how many genes in Homo sapiens are orthologous to how many genes in Ciona intestinalis.



3 Homo sapiens genes are co-orthologous to 2 Ciona intestinalis genes.
If I want to make an experiment in Ciona intestinalis to study p53, I should be aware that there are 3 human genes that are co-orthologous to 2 Ciona intestinalis genes.

So it's a many-to-many relationship

Going genome-wide scale

Now-a-days, we work with a high number of genes, so we can not use the classical approach. Everything must be done automatic and "blind" (not looking at the trees)

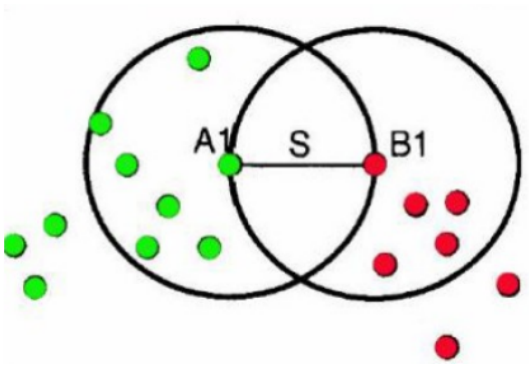
- **Best bidirectional (or reciprocal) hits.** This method is based on BLAST.
  Basically, you have 2 genomes and you make a BLAST from one genome to the other and I get the top hit. Then I do the reciprocal search. If my first hit is the same one, then I have the best reciprocal best hit. Thus, I consider them orthologs.
  The problem that this method has is that it can not predict one to many relationships.
  It is used only for closely related species where there are not many duplications!
  Low rate of false positives but high rates of false negatives

- **COG, MCL-clustering approach**
  We can perform multiple comparisons between different genomes and then build a network of BLAST hits.
  So, we are making a BLAST of all against all and then building a network of these relationships. Where each node is a protein that belongs to different species and the edges are BLAST relationships.
  The most modern methods have weighted relationships, depending on the e-value of the blast.

  In the network we can find orthologous groups (all genes derive from a single gene of a common ancestor). The term orthologous group is confusing, because we can have orthologs and paralogs.
  Also, we need to define how we go from networks to families (we can change the threshold).

- **InParanoid**
  Starts searching for the best bidirectional hits with a protein of interest.
  Then it searches within its own genome for its own hit.
  Any hit that is closer to the initial hit than the initial hit to the protein of interest, is considered a paralog.
  This handles many to many.
  The genes inside the circle will be paralogs that resolve from duplications after the speciation between the genes A and B (called in-paralog. Meaning that it is a paralog more recent than a given speciation).
  The genes outside are called out-paralogs. Meaning that they are paralogs that arrived from duplications more ancient that the especiation.

  Definition of in- and out-paralogues require the specification of a given speciation-node of reference

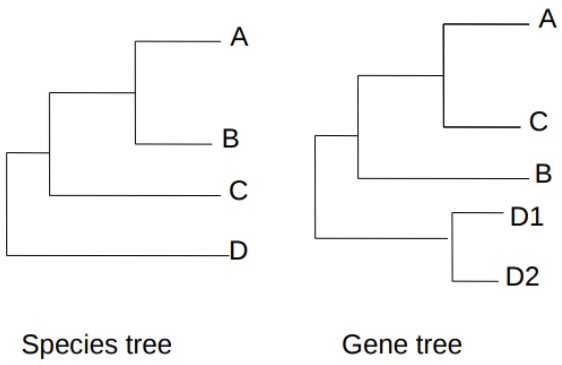


These 3 methods use blast and thus, they are really fast.

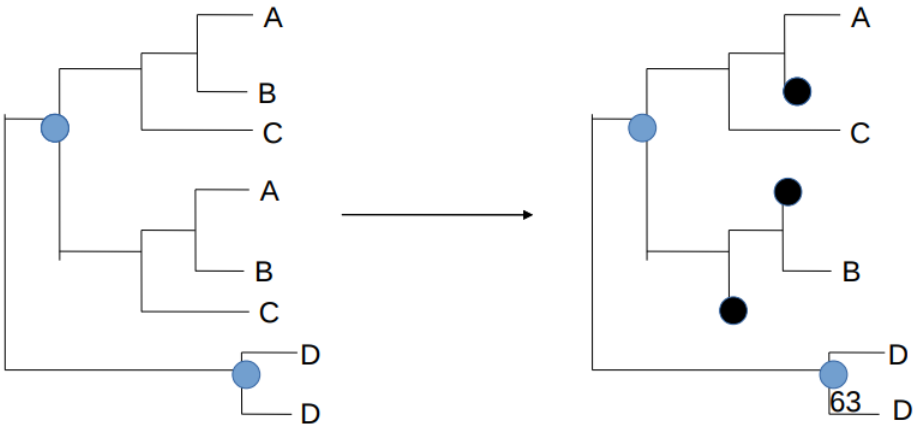
Methods based on phylogeny were not used at a large scale due to limitations in computational power. However these has changed, due to new algorithms.

They reconstruct the evolution of a gene family, detect duplication and speciation nodes and predict orthology and paralogy accordingly.
The 2 main methods for predicting duplication and speciation nodes from a tree are:

- **Species tree reconciliation** (RIO, Ensemble). Hard reconciliation: You want to reconcile every node. It resolves any incongruence between gene tree and species tree by introducing the minimal number of gene duplications and losses.
    - Soft reconciliation: Allows incongruences below a given support value.
        - A gene tree is congruent when It has the same topology as the species tree.



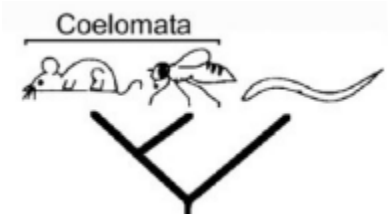
We have a species tree and a gene tree. Are they congruent? No, because in the gene tree A is closer to C than to B.
So, this method will reconcile the species tree and gene tree by adding the minimal number of duplications or losses.



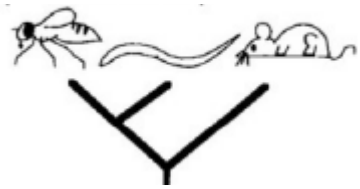
Reconciliation with the species tree readily provides you information on speciation and duplication nodes in a tree. It only works when:

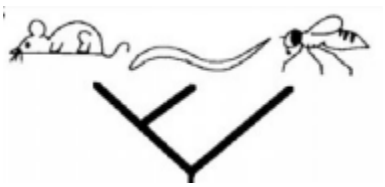
- We know the true species tree.
- The gene tree is correct and reflects the species evolution. Meaning that genes are only inherited vertically.

How often do we have the true species tree and the correct gene tree reflects the species evolution? There is a degree of uncertainty. For example, we would think that this tree is correct.

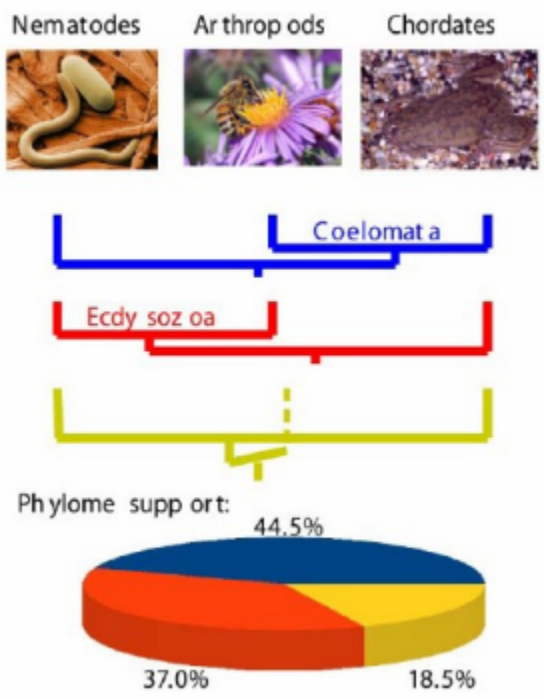


But these 2 other trees also have a certain support.





If we take the human genome and we make a gene tree for every gene in the human genome, let's see which one of these hypotheses is correct.
In other words, what percentage of gene trees from the human phylome support each topology?

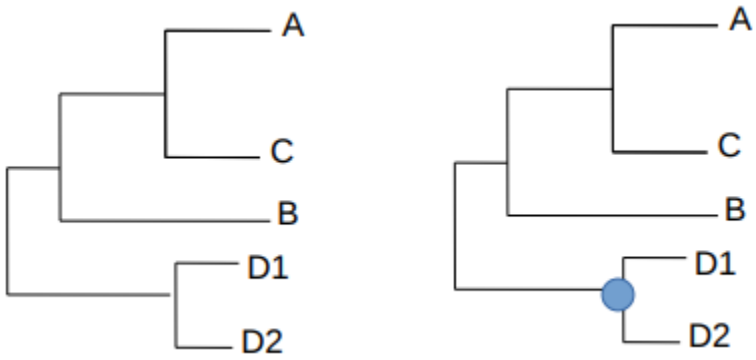


Nowadays, Ecdysozoa is the supported topology even though it is not the winning topology in the gene trees. Thus, you can not reliably use reconciliation because of topological variability.

- **Species-overlap algorithms**
  To deal with topological variability we use this algorithm.
  We simply explore the tree and for every partition in the tree, we ask the question: Is there a species overlap? If there is no species overlap, we put a speciation. Otherwise we put a duplication.

  It does not require a species tree, but needs to know the species to which the genes belong. In essence can be seen as a reconciliation with an unresolved species tree.

  In the last example, we would obtain (note that we do not use a species tree):



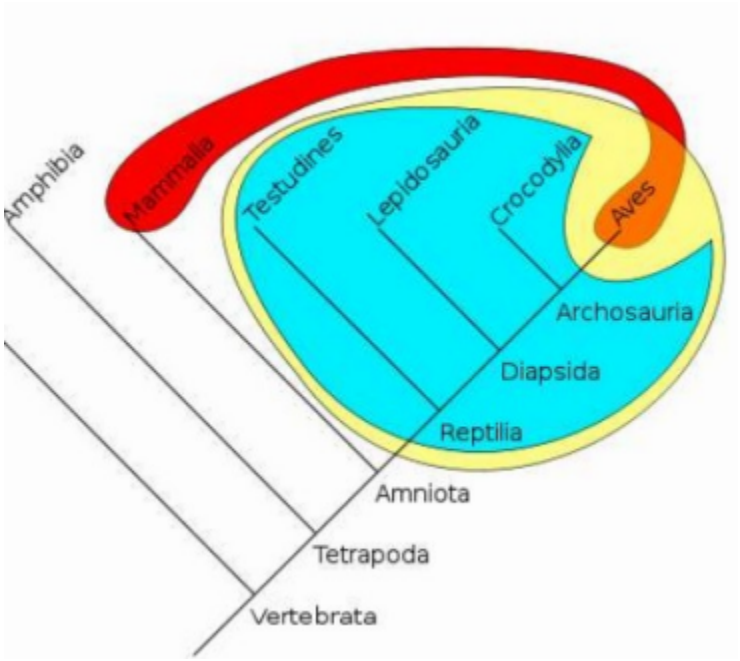
When comparing species overlap with the reconciliation method, we obtain similar prediction values. Meaning that they both say that it's orthologous or not. But the species overlap algorithm has a higher sensitivity.

Phylogenetic tree

A branching diagram (bipartite graph) showing the inferred evolutionary relationships among various biological species or other entities based on similarities and differences in their physical and/or genetic characteristics.

The species tree shows the evolutionary relationships between the different species.
The gene tree represents evolutionary relationships between genes. In this case, nodes can represent duplications or speciations.

Monophyletic group (yellow): All groups that share a common ancestor.
Paraphyletic group (blue): Group's last common ancestor and most of its descendants, excluding a few monophyletic subgroups. Modern reptiles contain not all descendants from a common ancestor.
Polyphyletic group (red): Set of organisms that have been grouped together based on characteristics that do not imply that they share a common ancestor that is not also the common ancestor of many other taxa. Warm-blood animals (mammals and birds) are from 2 different common ancestors.



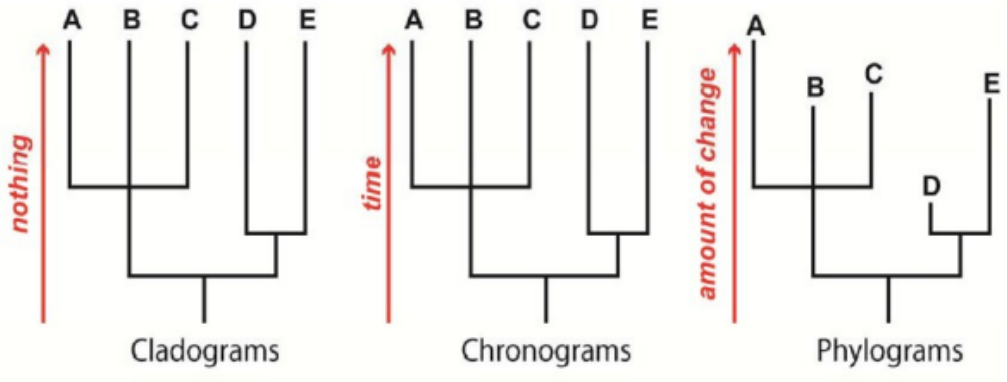
Trees

The topology is the branching structure of the tree.

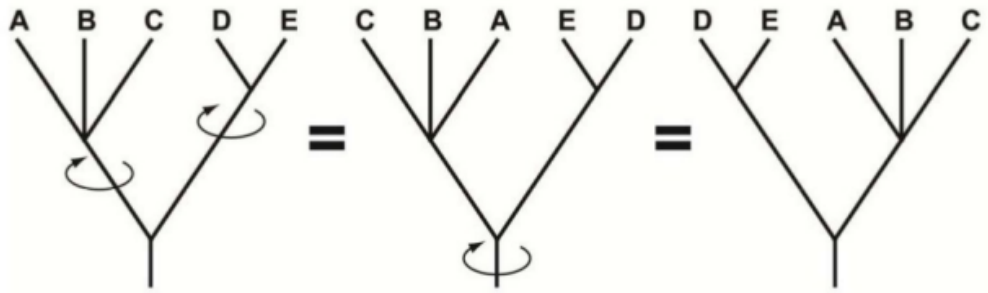
The length of the horizontal axis is not relevant.

The length of the vertical axis can mean different things:

- In cladograms, it does not mean anything. Because there is a constant mutation rate.
- In chronograms it represents the time
- In phylograms it represents the amount of change



We can represent the trees in vertical or horizontal mode or in different shapes. If we rotate the branches, the tree remains the same but with another configuration.



Newick standard: ((A,B,C),(D,E));

To represent a tree, we can use the Newick format (it explains the hierarchical relationships).

Trees can be rooted or unrooted. There are several strategies that can be used to root a tree. For every topology there are always more rooted trees than unrooted trees.

Number of rooted trees with n OTUs: $$(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Number of unrooted trees with n OTUs: $$(2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

How do we reconstruct a phylogenetic tree?

Phylogenetic trees can be based on anything that can tell us similarities and differences. Nowadays, molecular phylogenetics is the most widely used method because we can obtain sequences of different species very easily. These sequences are compared using alignments.

To find the best tree, we can use 2 approaches:

- **Exhaustive search:** Make all trees first and then see which one best fits the data. But this is not possible for a large number of sequences.
- **Heuristic search:** Try to find a way to find an optimal tree (hopefully the best) without testing them all. You also need an optimality criterion and you are not guaranteed to find the best tree, but you save time.

The **distance-based** methods are the fastest but they are not the best methods when obtaining the correct tree. If there are no errors, the correct tree can be obtained in polynomial time. Otherwise, optimization problems are NP-hard

Maximum Parsimony and **probabilistic methods** are NP-hard.

Bootstrapping

The numbers that appear in a branch of a tree are support values that are computed by a bootstrap. It does a sampling with replacement, so that it mixts the order of the sequence to check the robustness of the tree.

So, you will obtain a tree for each alignment and then you can make a consensus tree. Each value of a branch represents the % of times that branch appeared in other trees.

Normally, low supported node correspond to short branches (there has been a short number of changes)

Maximum parsimony method

Given a number of sequences, it builds a tree minimizing the number of mutations for each position of the sequence.
Problem: When comparing sequences that are really far away in evolution, the number of changes is not representative to the number of mutations. This is due to the fact that there is a saturation.
You also have to evaluate many trees to evaluate which is the best one.

For these reasons, it is not used often.

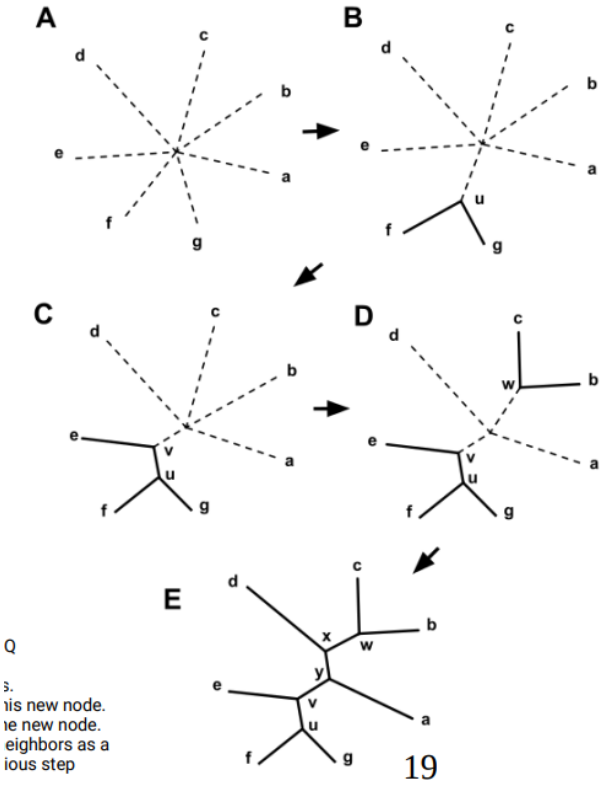
Neighbour Joining

You start with some sequences from which you do not know the relationship between them.
Then you compare all sequences to each other to obtain the pairwise distances.
Then you find which is the closest pair and make a cluster.
Then you repeat for the other remaining sequences.

This is very fast because you only need pairwise comparisons.

The problem is that it is also affected by the saturation of mutations when comparing sequences that are really far from each other.
Also because of convergent evolution

This method is used a lot.



Q
s.
is new node.
e new node.
eighbors as a
ious step

Statistical methods

Methods that try to work with probabilities of observing one tree.
For example the maximum likelihood methods.

The maximum likelihood method computes the probability of observing the data given a hypothesis. In our case, the data is the alignment and the hypothesis is the tree.
So, we are computing the probability that a tree computes a certain alignment.

Example: We toss a coin 10 times and we obtain:

HHTTHTHHTTT

We compute the maximum likelihood of these data as:

$$L = Prob(D/p) = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = p^5(1-p)^6$$

Thus, the maximum likelihood of observing the above sequence of events is p = 0.4545

For a sequence, we will compute the probability of going from one sequence to another.

$$\text{Sequence 1} \quad \mathbf{C\ C\ A\ T}$$

$$\text{Sequence 2} \quad \mathbf{C\ C\ G\ T}$$

To do this, we need the prior information and a transition probability matrix

$$\pi = [0.1, 0.4, 0.2, 0.3]$$

$$\boldsymbol{P} = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}$$

The transition probability matrices are computed using empirical data.

$$L_{(Seq.1 \to Seq.2)} = \pi_C P_{C \to C} \pi_C P_{C \to C} \pi_A P_{A \to G} \pi_T P_{T \to T}$$
$$0.4\text{x}0.983\text{x}0.4\text{x}0.983\text{x}0.1\text{x}0.007\text{x}0.3\text{x}0.979$$
$$= 0.0000300$$

We can do the log transformation and we will obtain a value of -10.414

We can translate this into a real case where we know the sequences (4 for example) and we want to compute the probability of a tree that gives that alignment.
So, we will create n trees for each position of the sequence and we will obtain the probability of having that alignment.

In clustering we created a vector that represents the probability of having each nucleotide and at the end we add all probabilities and multiply them by 0.25.

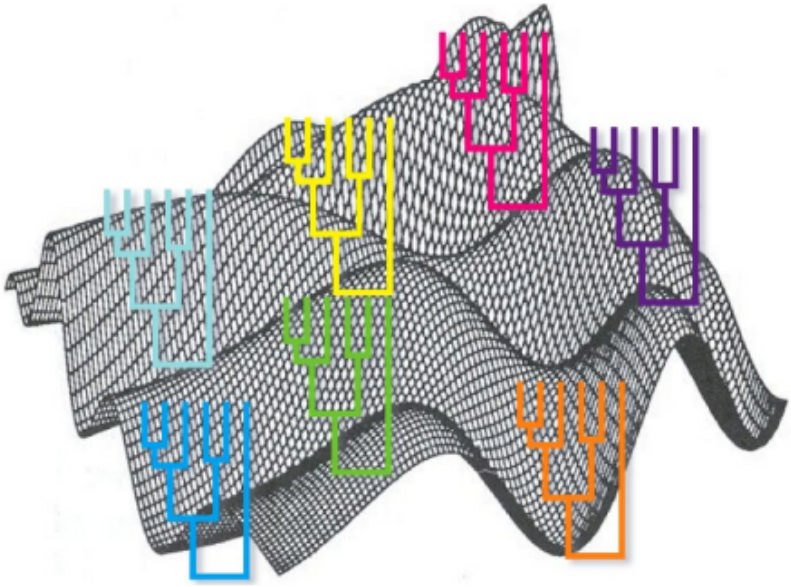
Note that we need to take into consideration the length of the branches by just multiplying the transition probability matrix by the length.

Advantage: In this method we are also computing the probabilities of all internal nodes.

So far we have assumed 1 evolutionary rate for all the sites. But we know that different parts of a protein evolve at different rates. So, to make our model more realistic we can use different probabilities of change in different regions of the alignment.

Heuristic Search

We know that there are millions of possible trees and each of them has a likelihood. Thus, we can imagine a space of possible trees where some of them are more likely than others.



We can explore this space of solutions with a heuristic method.
You can start at a random place or the first operation can be a NJ and then you use an heuristic method.

We will need a way of moving and a way to decide if we want to move or not. All methods have their criterions…

Bayesian inference

In maximum likelihood we compute the probability of observing the alignment (data) given the tree (hypothesis).
The Bayesian approach tries to compute the probability of the tree given the alignment.

So, we will compute the posterior probability and to do this we will need to know the prior information (the probability of knowing the tree before looking at the alignment).

$$P(\theta/D) = \frac{P(\theta)\cdot P(D/\theta)}{P(D)}$$

The problem is that we do not know the probability of knowing the tree before looking at the alignment.

Solution: Prior that are flat distributions. Meaning that all trees have equal probability (thus the result is not influenced).

How it works
We need a method to compute the posterior distribution and then we use the MCMC to look at the space of possibilities.

In the case of maximum likelihood, we were navigating until we were finding the maximum likelihood point.
In the Bayesian approach, the robot has a backpack and in each step he computes the likelihood of that tree and keeps the information in the backpack.
At the end of the day, the Bayesian approach will have a sampling. It will have more trees in the picks of the mountain because the algorithm stays longer in the high parts.

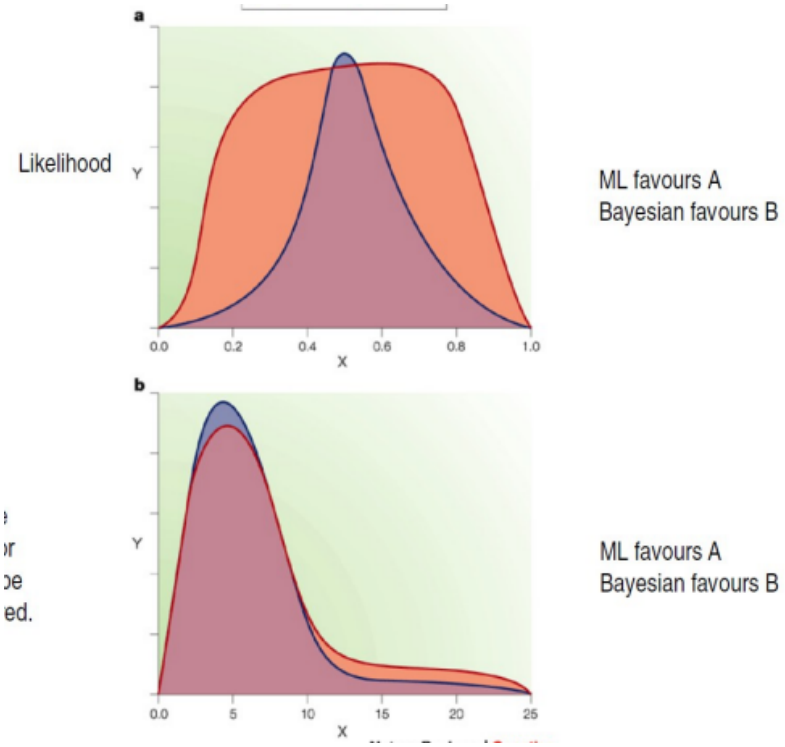
So, ideally, the trees will be sampled in proportion to the posterior probability (trees with higher probability will be sampled more times).

In the Bayesian approach you do not need to do a bootstrap, because you have 100000 trees and you can compute the support of each partition of your tree.

The problem is that you need to run it many times and you don't know when to finish. To solve this, we have many robots searching at random in the space of solutions. At the beginning, the backpacks will be really different, but when the backpacks are similar it means that they have all explored the whole space of solutions and thus you can stop.

You also remove the first trees (burn-in).

Maximum likelihood picks the blue tree because it has the highest peak. But the Bayesian approach chooses the orange one because the orange tree has been sampled more times than the blue one.



In the second case, the blue tree is much better.

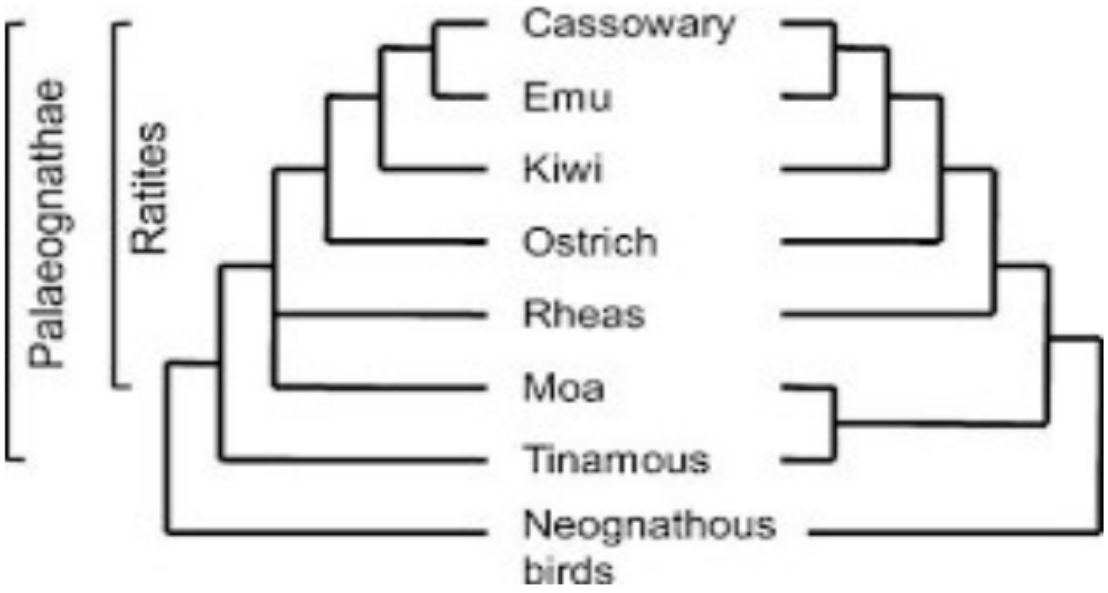
When there are few parameters and plenty of data, then ML and Bayesian inference will agree.

When you have a few data and a lot of parameters, then the BI is more reliable.

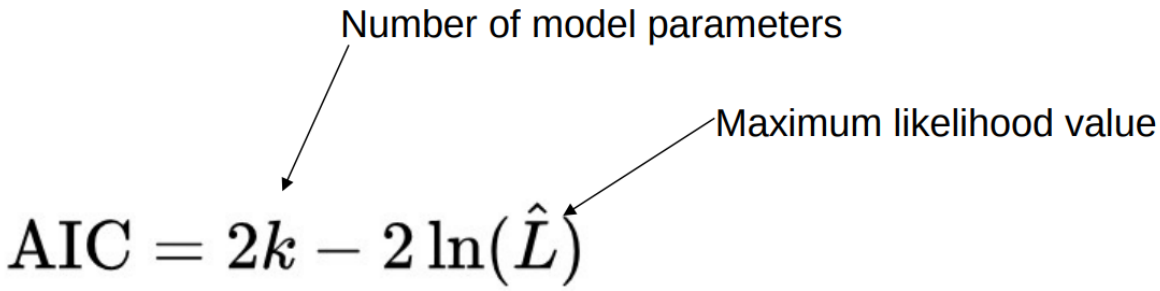
The good thing about this probabilistic based method is that you can use statistical tests to determine how more likely a topology is with respect to another.

Comparing alternative topologies

We have the following topologies and we want to test which is more likely.



We can compute the Akaike information criterion (AIC) to assess which topology is more likely.



AIC is an estimator of the relative quality of statistical models for a given set of data.
Given a set of a model, we can use AIC to choose the one with the minimum value.
As we can see, if we have a large number of parameters we will have a higher AIC (it penalizes a large number of parameters)

Once we have a tree, we are going to use algorithms to infer evolutionary events on our tree:

- Reconciliation algorithm
- Species overlap

The reconciliation algorithm tries to reconcile the incongruences between the species and gene tree using the minimal amount of duplications and gene losses. We saw that there is a hard and soft reconciliation.

The species overlap algorithm does not require a species tree but needs to know the species to which the genes belong. In essence can be seen as a reconciliation with an unresolved species tree.
For every node in the gene tree evaluate whether the daughter partitions share any species. If the overlap (number of species shared over total number of species) is higher than the given threshold. Input a duplication at that node.

You may also want to infer relative timing of speciations and duplications (puting time on the branches of the tree). Because normally the branch length represents the number of changes and not the time.

We can assume a molecular clock.
A molecular clock is said to exist when substitutions accumulate linearly with time. This assumption is violated most of the time, particularly at long evolutionary distances.
Also, different species have different mutation rates.

For a constant mutation rate, neutral substitutions are expected to behave more clock-like than non-neutral substitutions, that is why they are generally used as a proxy for time.

Another way of putting time into a tree is to use other kinds of evidence such as fossil records. But this implies some uncertainties.

dN/dS

We can use non-synonymous mutations to know about the functions of your proteins.
One test we can use is the dN/dS ratio which is the ratio of non-synonymous and synonymous substitutions.
It is useful to measure the strength of natural selection acting on protein-coding genes.

If one gene is evolving neutrally, we expect a value of 1 (there is the same chance of having a synonymous or none-synonymous mutation). This could be the example of a pseudogene where it does not care about the sequence.

If the gene is important, we expect a value close to 0. Because it does not accept non-synonymous mutations (purifying selection).

If there is positive selection, then you expect a lot of non-synonymous mutations and thus a value higher than 1. This is due to the fact that it needed to change of function.

We can also detect radical changes in aa when comparing a group of proteins.
Some regions of the protein accumulate changes that you can consider radical. Meaning that in some species there is a positive aa and in others there is a negative aa.

This can be indicative of functional shifts (radical change in function).

Phylogenomics

Intersection between genomics and evolution. That is, looking at genomes from an evolutionary perspective, often using phylogenetics.

The distinction between phylogenetics and phylogenomics lies in the scale:

- When talking about a single gene or protein in one or few species, we are in the phylogenetics realm
- When talking about a single gene or protein in all species, we are in the phylogenomics realm

Phylogenomics is necessary to provide an evolutionary framework to the deluge of data generated. It is useful to obtain biological knowledge from sequence data. The more data, the more powerful.

But, it is computationally demanding, it needs to be automated and it needs proper scalability.

Phylogenomics to reconstruct species trees

A species tree represents relationships between different species.
In gene trees we represent evolutionary relationships between genes.

Most of the species trees are based on molecular data, so there is a relationship between both trees.

How can we use molecular data to reconstruct a species tree?
At the beginning, the methods only used information from the sequences, without making any alignment. We don't use them any more because they suffer a lot from the effects of convergence (for example, a lot of gene loss is shared by parasitic organisms in the different parts of the tree of life. Thus, they become closer in gene trees but they are distant in evolutionary terms. They are close because they have similar gene content because both went convergently through massive gene loss).

The alternative methods are sequence based methods (they use the information of the sequence of the genes). There are two main methods that both start by aligning homologous genes present in different species:

- **Supermatrix**: Concatenating the alignments of the different genes to make a single tree. Thus, this tree is expected to better represent the evolution of species than a tree made by a single gene.
  By concatenating more and more residues into a longer alignment, then the precision in resolving difficult positions of the tree can improve. This is because noise cancels out.
  The more species you use, the fewer genes you can find shared between all species.

- **Supertree**: From each alignment of a gene, we make a tree. Then, all the trees are converged to a consensus tree or more parsimonious tree (there are multiple methods to converge the trees).

The results obtained in both methods are similar. But the branch lengths change a bit.

Genome-wide phylogenetic analysis (phylome)

Phylome: complete collection of evolutionary histories of all genes encoded in a given genome.
We are now interested in studying gene trees.

Imagine that we have x genes in a genome. For each gene, we interrogate which is the evolution of that gene. And this is best represented by a phylogenetic tree that represents the evolution of that gene and the homologes in the different species.
If we do this for all the genes, we will obtain a collection of all the gene trees derived from all the genes in the genome. This collection is a phylome.

There are 2 approaches:

- **Family-based approach:** First build families (orthologous groups), then reconstruct one tree per family (Ensembl):
  - Build families
  - Make an alignment per family
  - Phylogenetic reconstruction per alignment
- **Gene-based approach:** Sequentially use each gene of interest as a seed to build a gene tree (PhylomeDB):
  - Search for homologs
  - Make an alignment per gene + homologs
  - Phylogenetic reconstruction per alignment.

  In this case we will have to make many more trees.

  We can compare 3 different aligners in forward and reverse mode (we should obtain the same result when forward or reversed, but this does not happen because of the heuristics of the alignment process. Some processes include a gap in the right other to the left so we will accumulate differences)

With the result of the 3 aligners, we make a consensus alignment. We take each pair of aligned residues in each of these alignments and compare them to the others.

We will obtain a consensus alignment and we will also obtain information about which columns were more stable in all the different alignments. We will use this information to trim (remove some parts of the alignment that are not reliable) the alignments.

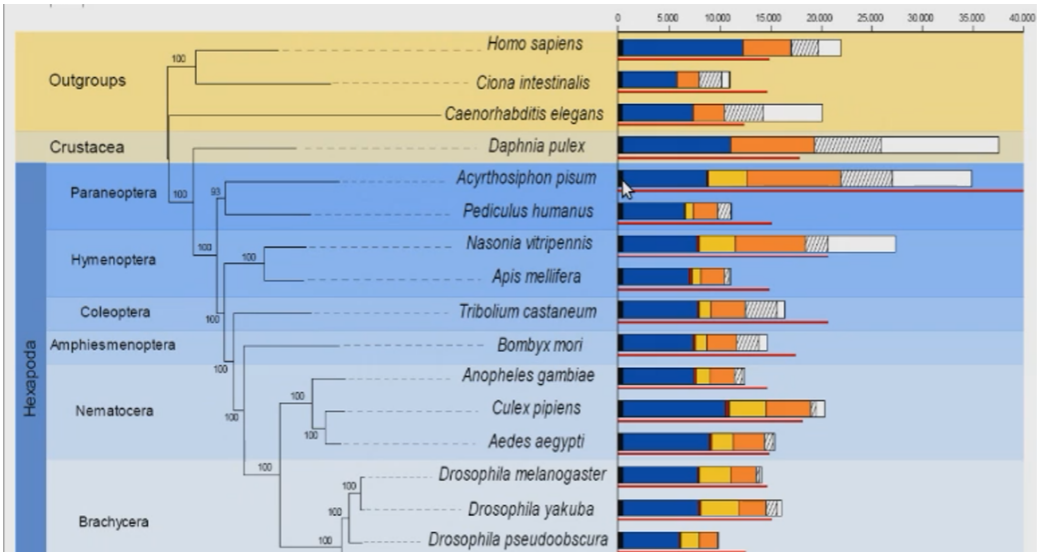
These phylomes provide the following information:

- Families that show a particular topology
- Detect and date duplication events
- Genes that have accelerated evolutionary rates at a particular lineage (because of an adaptation)
- Families expanded at particular lineages
- Footprints of horizontal gene transfer, lineage sorting, gene conversion and other evolutionary processes
- Search for co-evolving genes
- Predict functional properties
- Across-species prediction of orthology and paralogy

Large-scale phylogenetics to assist in the annotation of a newly sequences genomes
We compared its genome with 13 other sequenced arthropods and 3 out-groups. We used the gene-based approach pipeline we mentioned before and we obtained a tree.

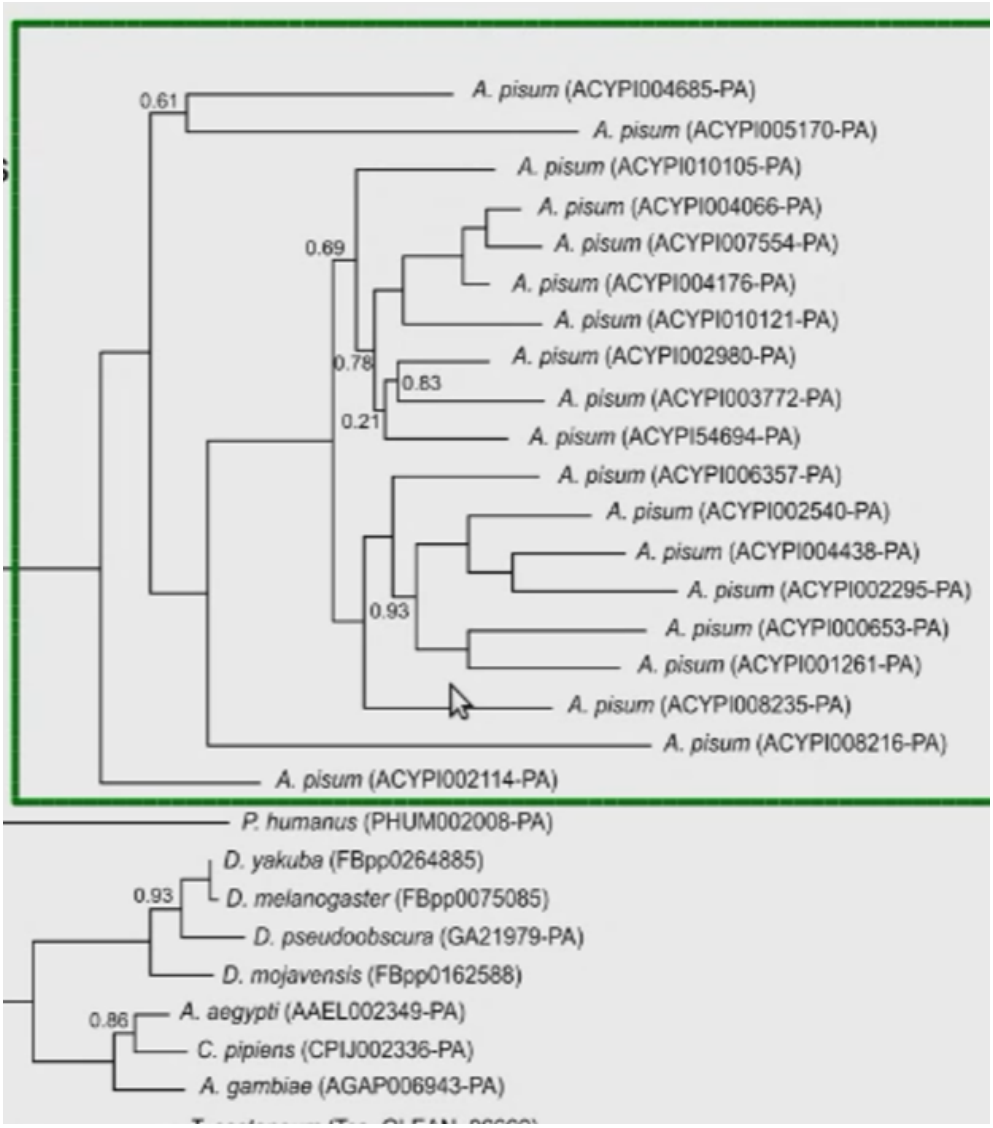
We made a species tree.
Note that Acyrthosiphon pisum is the species of interest.

Yellow are insect specific genes
White are the genes that are duplicated. As we can see, our species of interest has a lot of duplications (reason why it has many more genes than other species). So, we were interested in these gene family expansions.



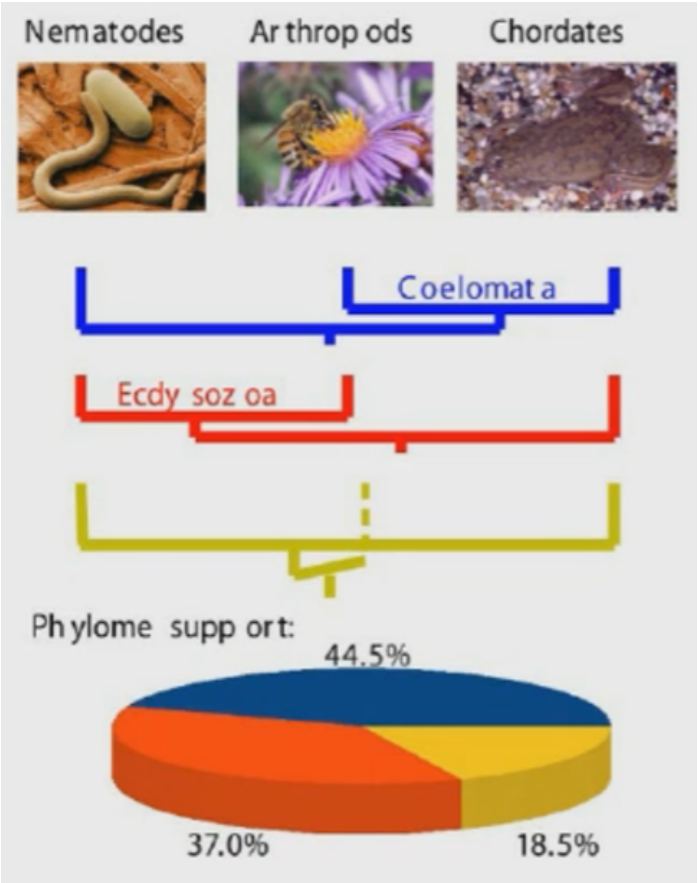
We also used the gene trees to annotate sequence functionalities. We can search the GO terms of the proteins of the homologs and determine the function of the proteins of the species of interest. So, if many species have the same functions it may be that that protein is going to be conserved in all the tree.

As we said, we can also determine the number of duplications that our species has suffered. In our case, it has a lot of duplications in a transporter.



Gene tree vs species tree

There is uncertainty in species trees and topological variability in gene trees.
We have mentioned that we can use gene trees from a human phylome to support a topology. But this was not accurate at all.

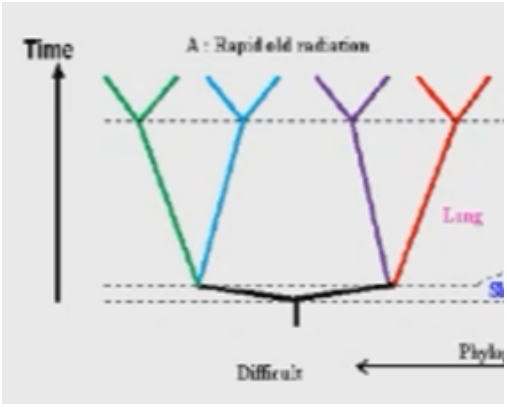


Possible sources of this incongruence between gene tree and species tree:

- **Analytical factors:** They lead to failure in accurately inferring a gene tree (obtain a gene tree that is wrong). These can be either due to stochastic error (insufficient sequence length or taxon samples, noise. They are random) or due to systematic error (observed data far depart from model assumptions. These are the most dangerous ones, because your model explains poorly the data and thus you will always have that error).
  So, there are problems regarding the methodology or data used.

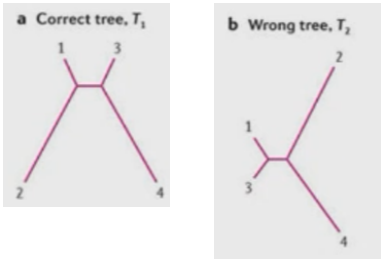
- **Biological factors:** They lead to gene trees that are topologically distinct from each other and from the species tree. Known factors include stochastic lineage sorting, hidden paralogy, horizontal gene transfer, recombination and natural selection.
  So, this factor relates to actual biological processes that may result in true incongruences between the gene and species trees.

Analytical factors



The most common case is when there is “fast radiation”. Meaning that there is a short internode. Therefore, the species that result from that branch will be really similar because there has not been enough time to accumulate enough mutations. Therefore, it will be really difficult to obtain the right topology. Because all species (green, blue, purple and red) can descend from any of the branches.

This will also happen when analyzing short sequences. Because you have even less information.

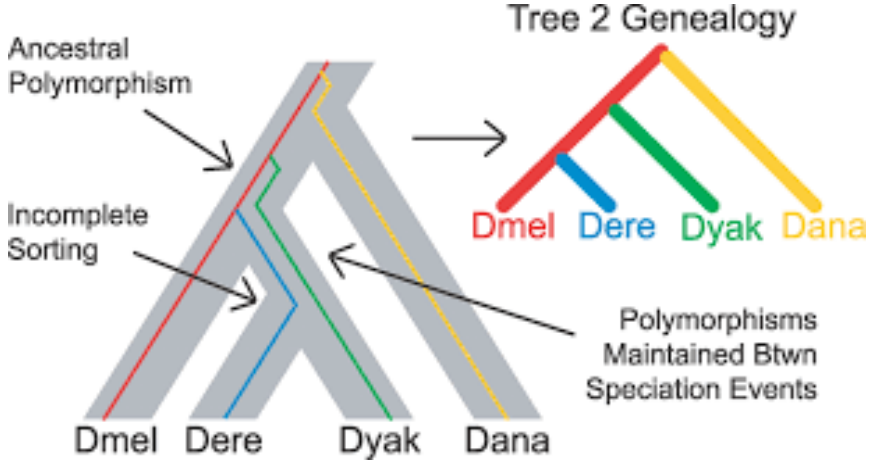


A systematic error could be “long branch attraction artifact”. When you have a dataset where there are sequences that are close to each other and other sequences that are far from all the rest.
You will obtain 2 types of trees and you won't be able to distinguish which one is the correct one.

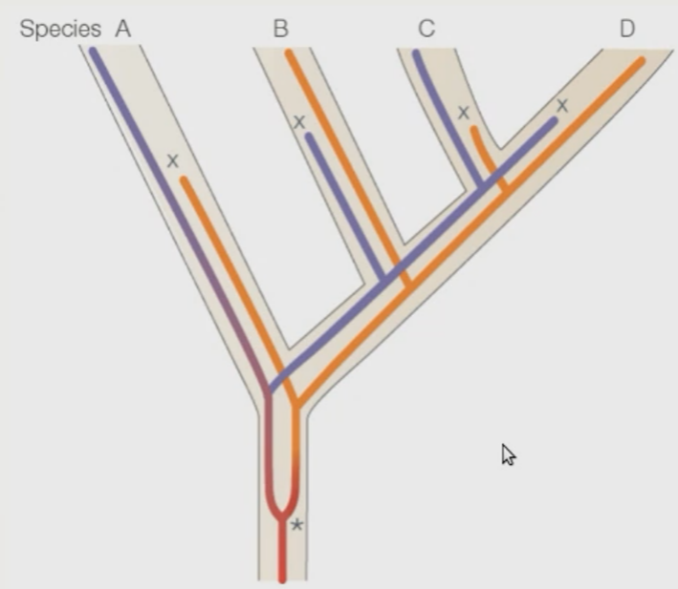
2 species that should be far away are put together to minimize the total length of the gene tree.

Biological factors

- **Incomplete lineage sorting**
  We have to consider that the speciation process happens at the population level. Each population is genetically diverse, meaning that there are different alleles in the population.
  Because of that, if there are 2 consecutive speciation events in a short amount of time, it may happen that some of the alleles distribute in the resulting species in a way that is different to the species tree.



- **Hidden paralogy caused by differential gene loss following duplication**
  In this case, a gene duplicates. So, all the species have 2 paralogs for a long time and suddenly they all lose one of the paralogs (maybe one or the other).
  So, the phylogeny you will reconstruct from these genes will be different

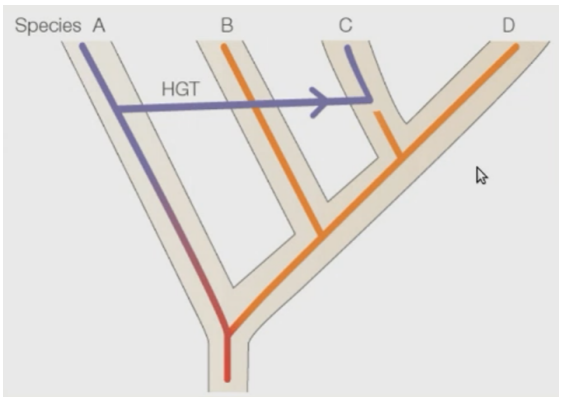


- **Positive selection can cause tree patterns different from the gene tree by convergent mutations.** This is very rare.
  2 genes appear together in the gene tree because they are really similar to each other because of convergent evolution.
  For example the gene responsible for echolocation is present in bats and dolphins.
  So, we will find this gene tree where they are close and in the species tree they are really far.

- **Non-vertical evolution**
  We represent trees vertically, meaning that there are parents that create childs. But what happens when there are genes that go from one species to another or lineages that mix?

  Evolution makes small jumps (Quantum leaps). Meaning that there are gene duplications, symbiosis, hybridization and lateral gene transfer.

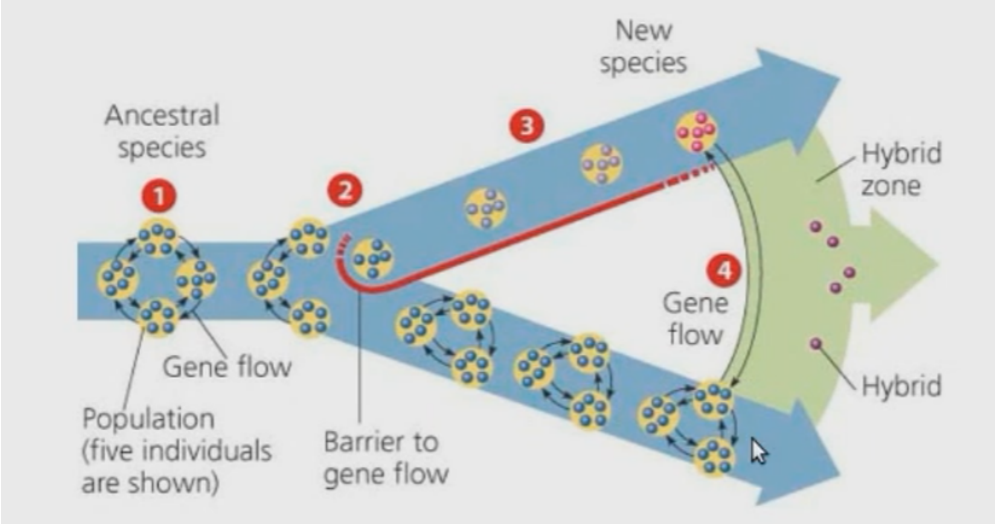
  The typical case is the horizontal gene transfer. Process by which a gene is transferred from an organism of a species to another one from a different species.Where one gene is transferred to another and maybe it substitutes another gene.
  Thus, if you make a tree, you will obtain a topology that is different to the species tree.



In this case A will be next to C and B next to D (which is wrong)

Another process of non-vertical evolution is hybridization (it can be seen as a massive horizontal gene transfer of the whole genome).

In a populations perspective, you can see the hybridization as:
You have one population with a gene flow and at some point there is a barrier that separates part of the population. During the time that there is a barrier (glaciation, for example. A species moved to another island…), they will diverge from each other. At some time, the barrier disappears and there is a gene flow between both populations, forming the hybrids.



Sometimes, the hybrids can still mate with the both of the WT or not. Imagine that the WT yellow and red make an orange hybrid.

- The orange can't mate with the yellow but can mate with the red. If there are successive rounds of mating, we will have introgression (red genome with small pieces of yellow genome).

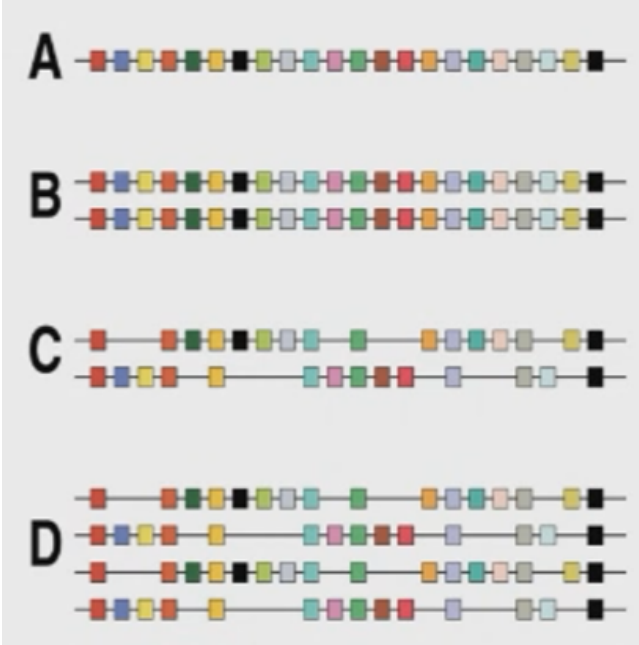
Hybridization can lead to networks rather than trees. Because in hybridization we are joining species.

When 2 species mixt, the 2 genes that were orthologs in different species will become paralogs (we see 2 homologs in the same genome). But they are not paralogous genes, because they especiated. We will call them homeologs:

- Set of genes that were originated by speciation and then joined back together in the same genome due to hybridization

So, hybridization and horizontal gene transfer lead to incongruences when making gene trees.

- **Whole genome duplications**
  The Hox cluster is a group of genes that regulates the development process in vertebrates. These genes are situated together in different chromosomes.
  These blocks sometimes had similar arrangements between genes that are more similar to each other.



  For example, the red genes are always in the first position…

  We proposed that this comes from 2 rounds of **whole** genome duplications.
  So, we have an original Hox cluster (A) that suffers a duplication. Then some of the paralogs are lost and there is another duplication.

Questions

B1) You have a set of 40 complete, annotated genomes. Briefly define what would be your strategy to reconstruct the relationships among these species.

- Make a multiple alignment
- Create a distance matrix
- Run UPGMA, NJ, MCMC algorithms…
- Since we are working with a lot of species, I would use a heuristic algorithm.

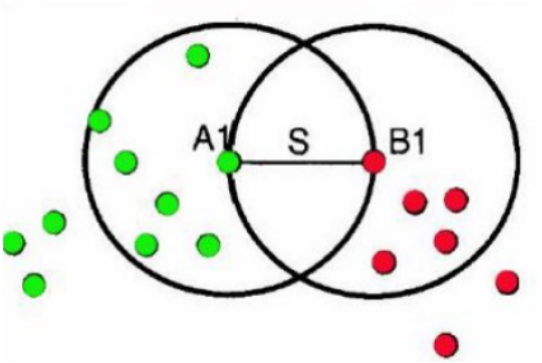
El teacher diu d'explicar el "Supertree approach".

B2) Mention two alternative methods to predict orthology and paralogy relationships, briefly describe their basis and discuss the main advantages and disadvantages for each method in comparison to the other one.

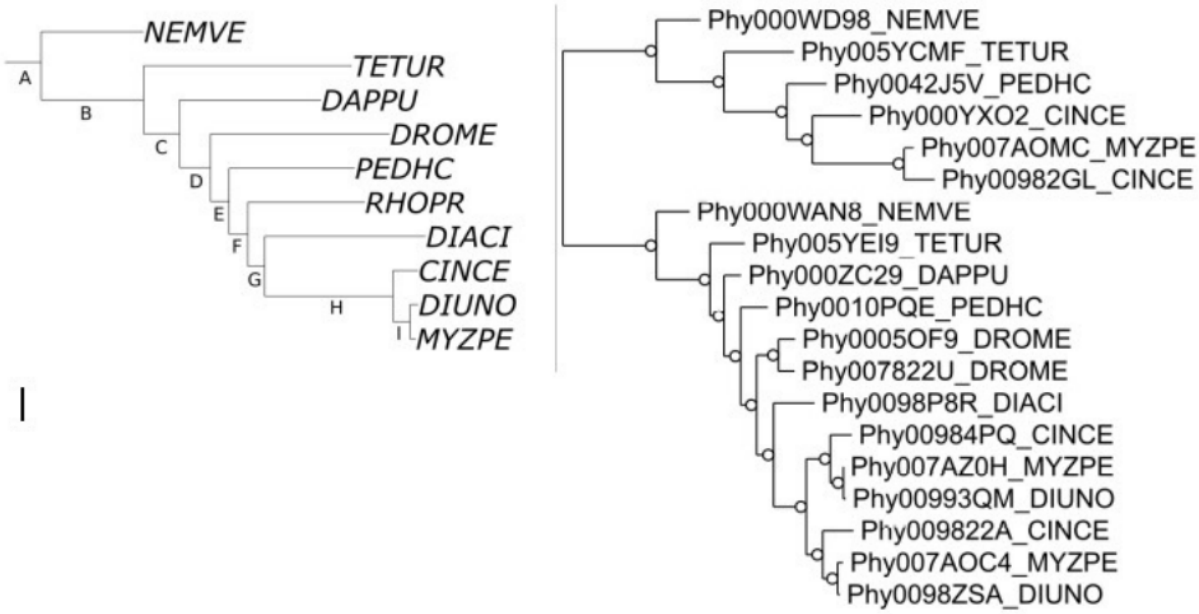
- **Best bidirectional (or reciprocal) hits.** This method is based on BLAST.
  Basically, you have 2 genomes and you make a BLAST from one genome to the other and I get the top hit. Then I do the reciprocal search. If my first hit is the same one, then I have the best reciprocal best hit. Thus, I consider them orthologs.
  The problem that this method has is that it can not predict one to many relationships. It is used only for closely related species where there are not many duplications!

- **InParanoid**
  Starts searching for the best bidirectional hits with a protein of interest.
  Then it searches within its own genome using the obtained initial hit.
  Any hit that is closer to the initial hit than the initial hit to the protein of interest, is considered a paralog.
  This handles many to many.
  The genes inside the circle will be paralogs that resolve from duplications after the speciation between the genes A and B (called in-paralog. Meaning that it is a paralog more recent than a given speciation).
  The genes outside are called out-paralogs. Meaning that they are paralogs that arrived from duplications more ancient that the especiation.

  Definition of in- and out-paralogues require the specification of a given speciation-node of reference



C1) Given this species tree (left) and this gene tree (right), where in the gene tree the species code is found at the end of the gene code, indicate the number of times that each node in the species tree (internal and terminal) has been duplicated in the gene tree using the species overlap algorithm:



There is a duplication in node A, D, H, H.

Genome comparison and evolution of gene order

In **pre-sequencing times**, a way to compare 2 genomes consists of:

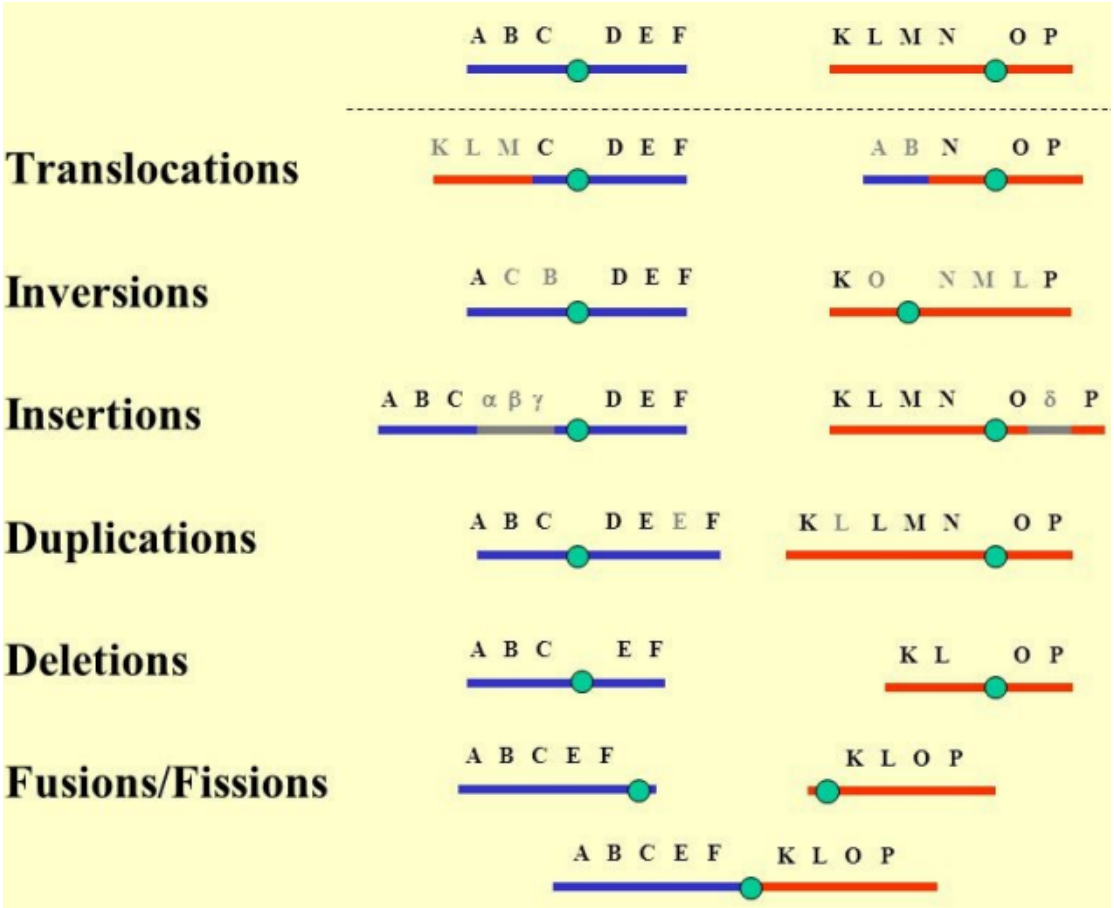
- Obtain both DNA
- Heat to denaturalize
- Combine the single strands of DNA at a lower temperature
- Determine the degree of hybridization

Then you measure the percentage of DNA that is still denatured at different temperatures. If they are really similar, they will remain together longer. A high temperature implies greater genomic similarity.

We can also compare genomes using an electrophoresis that separates entire chromosomes. The smaller chromosomes run faster in the gel.

Using this methodologies, they discovered that the species do not only diverge by accumulating changes like point mutations in the DNA but also by the genomic rearrangements.

Types of chromosomes rearrangements: Mutational changes in the genome that range from a few hundred bp to several Mb that cause structural variations in the genome



Comparing genomes in the **sequencing era**:
We can compare the nucleotide frequencies of two different species:

- Also dinucleotide, trinucleotide, k-nucleotide frequencies (k-mers).
  A k-mer is a fragment of the sequence of a given length.
  So, we can see how many k-mers we have in both genomes…
  If the k-mer content is similar, we will suppose that they belong to close related species. There are methods that do this automatically, like Self Organizing Maps (used in metagenomic analysis). They will try to put in a space closer to each other the sequences that have similar tetranucleotide frequencies and far away sequences that have different tetranucleotide frequencies.

- Compare the GC content (also gives information of the AT content)

The good thing about this method is that you do not need an alignment. We are just counting and therefore this algorithm is really fast.

To compare genomes, we can also do **alignments**:

- Partial alignments: We break the 2 genomes into pieces, we make a BLAST of each of these pieces to the other genome and then make a reciprocal hit. If our piece aligns with an identity over 30% and over 70% of our length, then we count it as a hit. We will store the % of identity and then we will compute the average nucleotide identity (ANI) over the whole genome (we will see that on average 60% of the genome is identical).
  So, as a recap: In the method of ANI we make BLAST searches by aligning small pieces of one DNA against another DNA and doing the reverse.

- Whole genome alignments. Algorithmically aligning two genomes is not simple because genomes do not only evolve by accumulating point mutations but also because rearrangements.

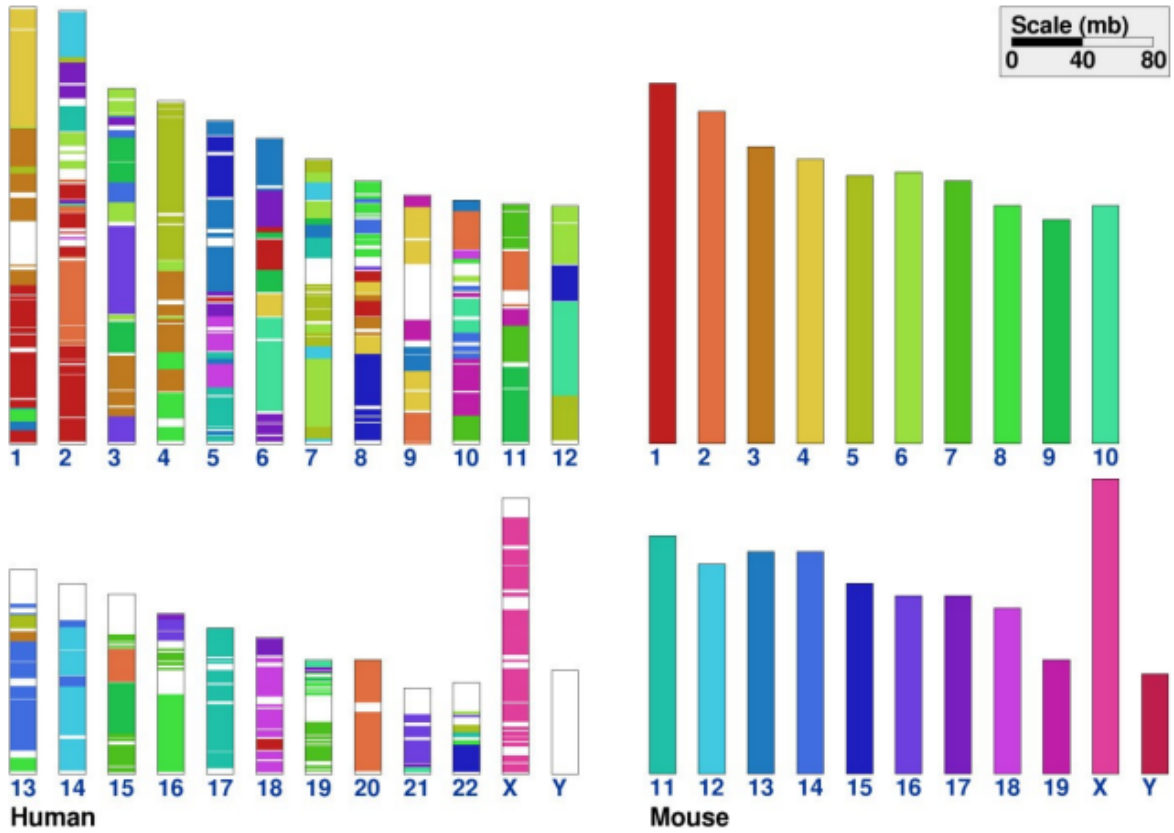
All the previous methods based on kmer frequencies do not capture other types of rearrangements.
Also, standard sequence alignment algorithms such as Needleman-Wunsch or Smith-Waterman do not work as they do not handle re-arrangements (because an alignment expects the things to go in a certain order and if we have a translocation it does not know how to proceed) and computationally they cannot handle such large sequences.

Solution: Align bits and pieces while being aware of the relative coordinates of the pieces.
They try to find anchors (small regions that align well) and then they try to extend them.
Finally they look at how these different anchors are related to each other.

Which genomes should you align? They need to be sufficiently similar (they have diverged recently) so that we can identify homologous regions. But they need to be a bit distant because otherwise there will be no differences.
For reasonable analysis, genomes should:

- Derive from a sufficient recent common ancestor: So that homologous regions can be identified
- Derive from a sufficiently distant common ancestor: So that sufficiently “interesting” changes are likely to have occurred

There are many genome browsers (like Ensembl) that provide pre-computed alignments with closely-related species. We can see the human genome aligned with all the primates or all the other vertebrates…



When doing “**Chromosome painting**” to check if the location of genes is conserved.

We use it to compare genomes, where you pick a color for each mouse chromosome and paint it in the human chromosome, and compare the rearrangements.

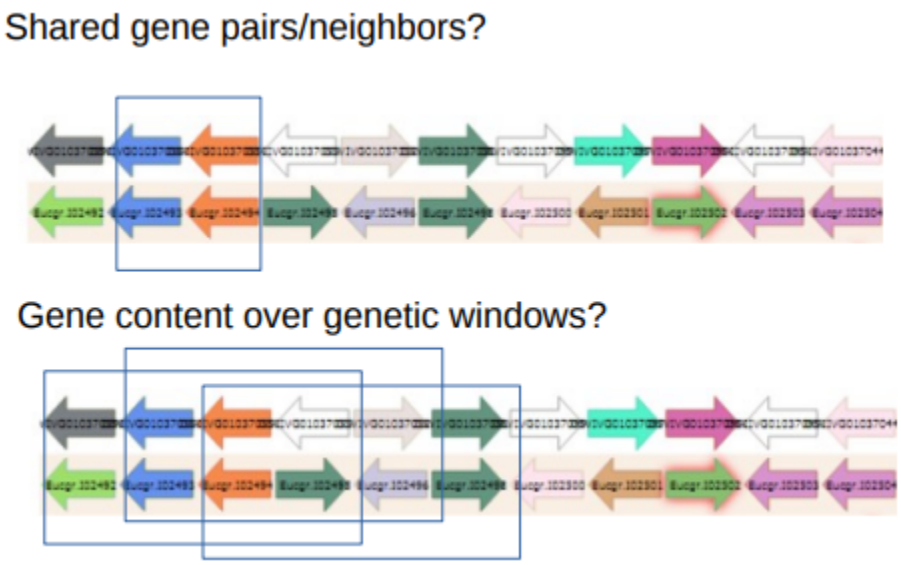
We can see that the chromosome Y does not have homologous regions at the threshold that they use. This is because it evolves much faster than the other chromosomes since it can not be repaired by recombination.
The X chromosome is highly conserved because it never recombines with other chromosomes.

Synteny

In **classical genetics**, synteny describes the physical co-localization of genetic loci on the same chromosome within an individual or species.
In **comparative genomics** synteny refers to the conservation of block order within two sets of chromosomes that are being compared with each other. This concept can also be referred to as shared synteny.
In other words,conservation of DNA blocks which are close in the chromosome and in the same relative order. The gene has stayed in the same relative position in the chromosome, so the neighbors around these gene are the same.

How can we measure gene order conservation? How can we compare the gene order in species A and B? How can we determine how distant they are in terms of synteny?
With nucleotide sequences it was easy, because we only needed to count substitutions, % of identity… But in this case, what should we do?

- We can count shared gene pairs
- We can also create windows of a certain number of genes and see the number of genes that are shared between the 2 windows.



There are other algorithms:

- **Pancake flipping problem**
  - Imagine that you have a pile of pancakes and each pancake has a different size. You want to make a pile like a pyramid.
  - To do this, you can only insert a spatula in any place and flip the pile.
  - They developed an algorithm that obtains the pyramid with the smallest number of movements.
  - We also have the **Burned pancake flipping problem**, where the orientation matters. This is really similar to genome rearrangements.
    So, we can determine the distance between both genomes by counting the number of steps that are needed.

Step 0: π	2	-4	-3	5	-8	-7	-6	1
Step 1:	2	3	4	5	-8	-7	-6	1
Step 2:	2	3	4	5	6	7	8	1
Step 3:	2	3	4	5	6	7	8	-1
Step 4:	-8	-7	-6	-5	-4	-3	-2	-1
Step 5: γ	1	2	3	4	5	6	7	8

Step 0: π	2	-4	-3	5	-8	-7	-6	1
Step 1:	2	3	4	5	-8	-7	-6	1
Step 2:	-5	-4	-3	-2	-8	-7	-6	1
Step 3:	-5	-4	-3	-2	-1	6	7	8
Step 4: γ	1	2	3	4	5	6	7	8

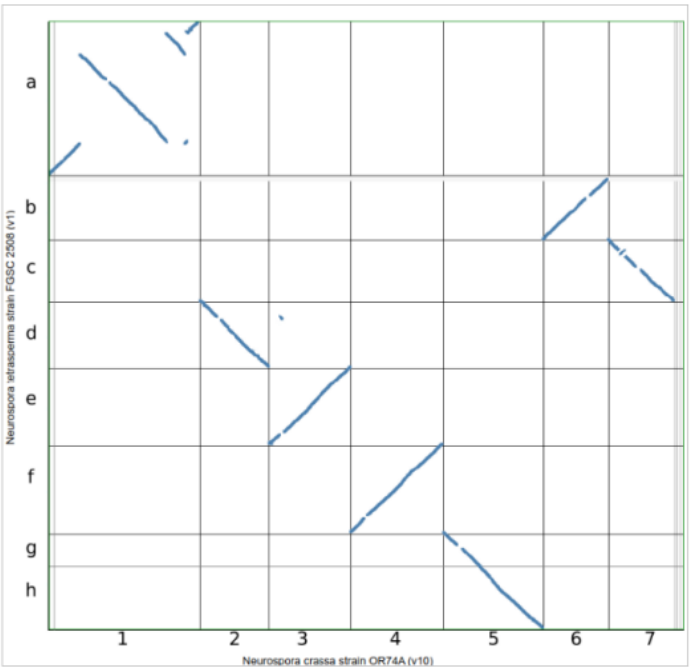
Dot Plot

We compare a genome against another genome (represented in the X and Y axis). Each dot can be a fraction of the genome or a gene.

It's a graphical method for comparing two biological sequences and identifying regions of close similarity after a sequence alignment. Each axis is a sequence.

- **lines:** The line means regions of similarity.

Exam problem: We compare 2 closely related species in terms of gene order conservation and have obtained the following dotplot. Circle the inversions with an "I" and the chromosome fusions with an "F".



There are no whole chromosome inversions, because it would not make any sense.
The reference sequence in one species is in one way and in the other it's in the other way.

Chromosome fusion: G and H are 2 different chromosomes and in the other species they are fused in chromosome 5.

There are genes that tend to be close to other genes.
In biology, when you see something that goes against the trend (in this case the disorder in the genome), is usually the result of selection. There is a selection acting against this trend.

As an example, in bacteria (prokaryotes) we have operons: Genes that are next to each other and therefore transcribed in a polycistronic transcript (a single transcript). They are then translated into different proteins.
The advantage is that all these proteins are regulated under the same operon or promoter region and therefore all these proteins are activated under the same conditions.

In eukaryotes there is no such organization of genes. But there is an influence in where each gene is located. As we know, the chromosomes are organized as:

- Heterochromatin: DNA associated with some proteins that condense the DNA.
- Euchromatin: When the DNA is loose. This space allows transcription factors to bind and activate the genes. The regions of the genome open up or close depending on the regulation. Thus, genes that need to be regulated similarly are located close together.

How can we use this to predict the function of a protein?
So, if we find genes that are kept together through evolution, we can make an hypothesis saying that they are involved in the same pathway or biological process.

Note that this is very different from what we were discussing when talking about homology prediction. Because in this case the genes are not doing the same function but they are cooperating in the same process.

Take into consideration that it is of great help knowing the function of at least one gene. Because then we can make inferences and discover the function of the other genes. Otherwise we only know that they are related but we don't know what they do.

There is one extreme case in which the genes fusionate. For example the tryptophan synthase has 2 subunits but in yeasts there has been a fusion and thus there is a single unit. This is an even stronger indication that both genes are related somehow and moreover, they are really likely to be physically interacting.

Detecting genome rearrangements with Next Generation Sequencing
We want to detect genomic rearrangements not only across species but also within species. For example, cancer genomes have a lot of rearrangements. So, it is interesting to be able to detect them.

How can we do this? We can use sequencing approaches. We just expect a certain distance between the reads and if there is a smaller distance, then there is a deletion…

Phylogenetic profiling and co-evolution

Co-evolution: Occurs when two or more species reciprocally affect each other's evolution through the process of natural selection.

Phylogenetic profile: Describes the presence or absence of a protein in a set of genomes. Similarity between profiles is an indicator of functional coupling between gene products.

Why is it useful to compare the gene content of 2 genomes (what genes has genome A and what genes has genome B)?
If we compare 2 genomes we expect them to share the genes that are responsible for making them share the same phenotype/characteristics.
The genes that are unique for each of the 2 genomes may indicate functions, characteristics, traits… that are specific to each species.

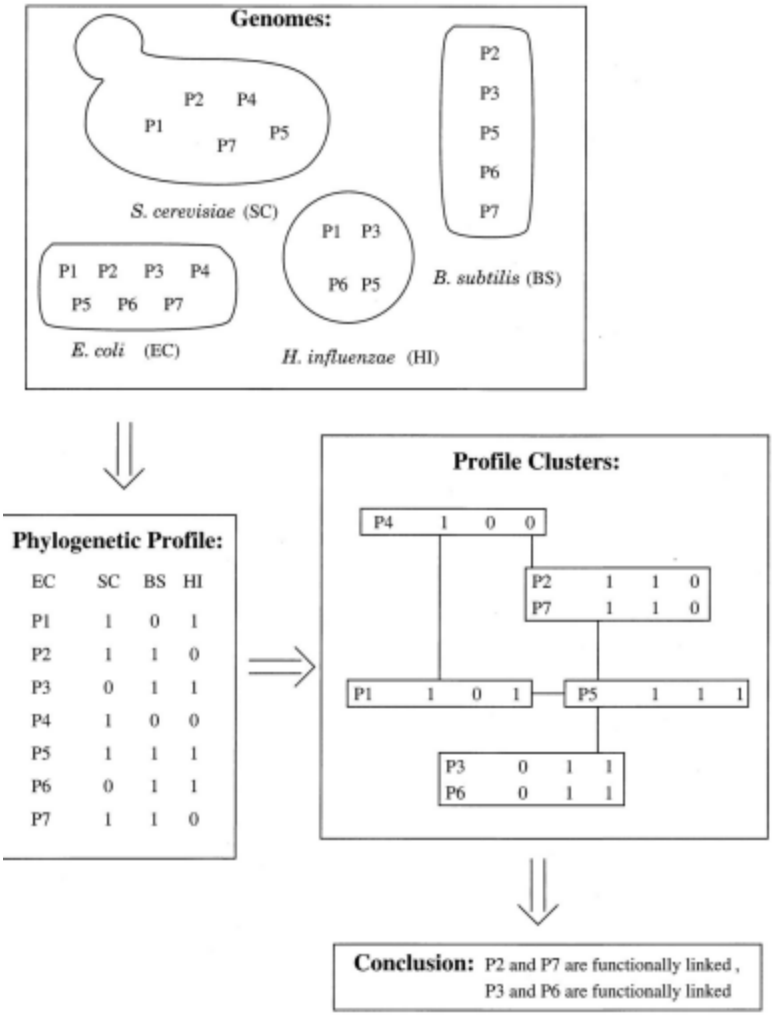
Can we extend this to more than two genomes?
Yes. But we will need to detect homology or, even better, orthology between the genomes.

We have 4 different genomes and we are investigating 7 different proteins that can be shared between the species or not.

We can build a matrix that stores a 1 if the protein is codified in the genome or a 0 if not.

Then we make a clustering. Group proteins or genes that have identical or similar profiles.

The genes that share a similar profile will tend to be working in the same function.
Because if there is a biological process that needs 2 genes, these 2 genes will be present in the genome every time this biological process is required.

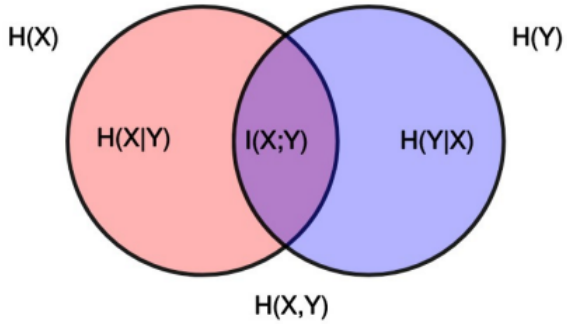


Distance measurements between profiles:

- Hamming distance
- Mutual information
- Jaccard index
- Correlation

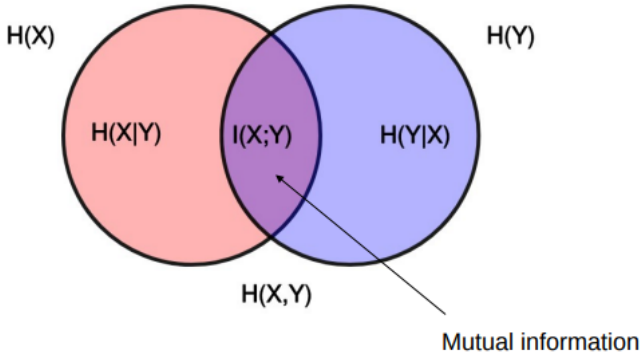
Hamming distances

Counting how many different instances we have between the 2 profiles. In other words, how many times there is a species that has a gene and the other does not.



Mutual information (intersection)

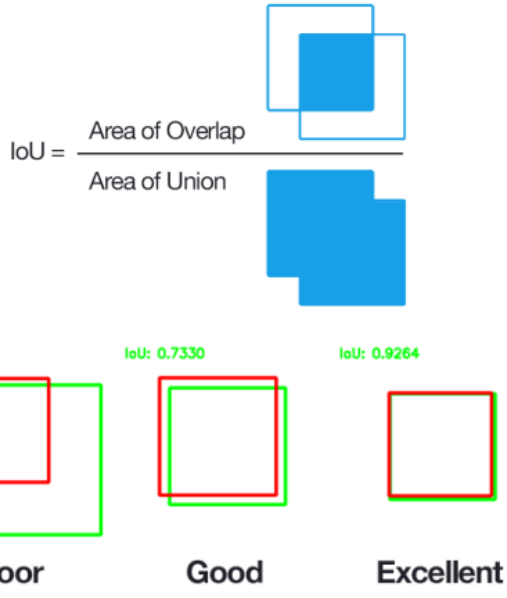
Counting how many instances are shared between the 2 profiles.



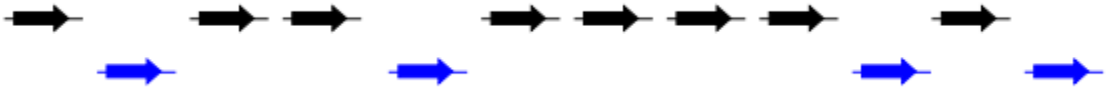
Jaccard index (Intersection/over union)

Counting how many instances are shared between the 2 profiles and dividing it by the total “surface” (all the instances).

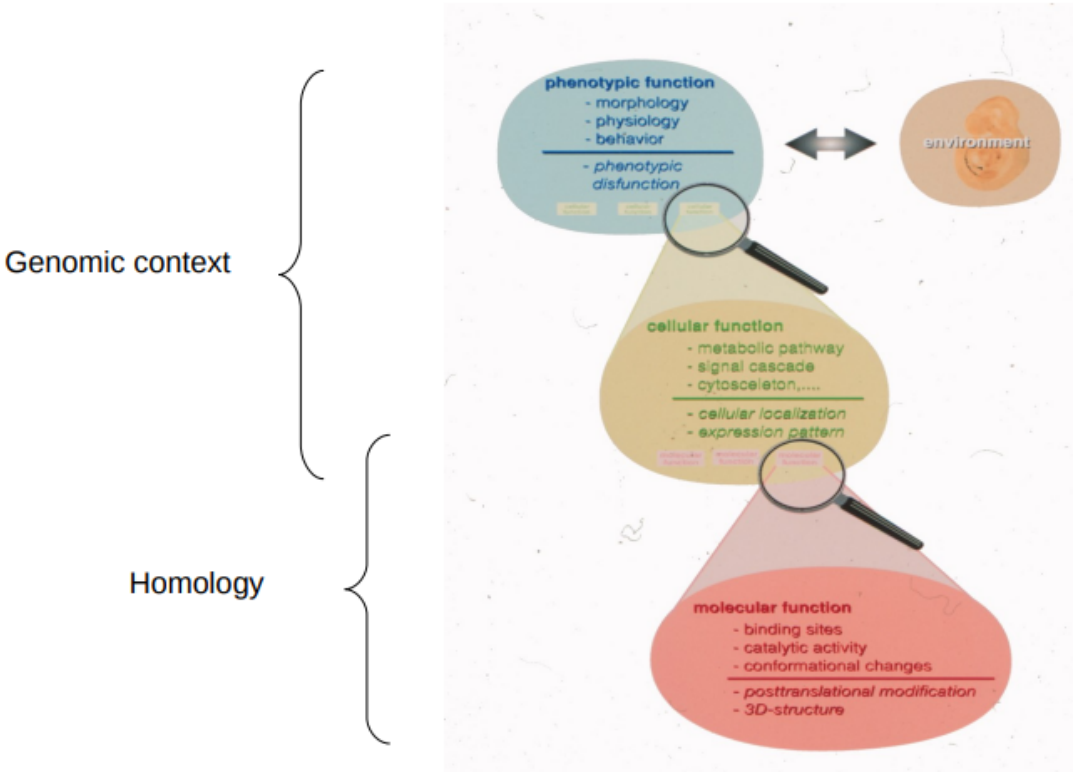
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



Genes with complementary phylogenetic profiles tend to have a similar biochemical function. It makes no sense to have different genes that do the same (redundant).



Types of genomic context



This methodologies that are based on genomic context (like gene fusion/fission, gene order, phylogenetic profiling…) give us information that is totally different and non-overlapping with the information obtained with homology based function prediction (which said: you have a similar sequence, then you have a similar function. If you are similar to a transporter, then you are likely a transporter).

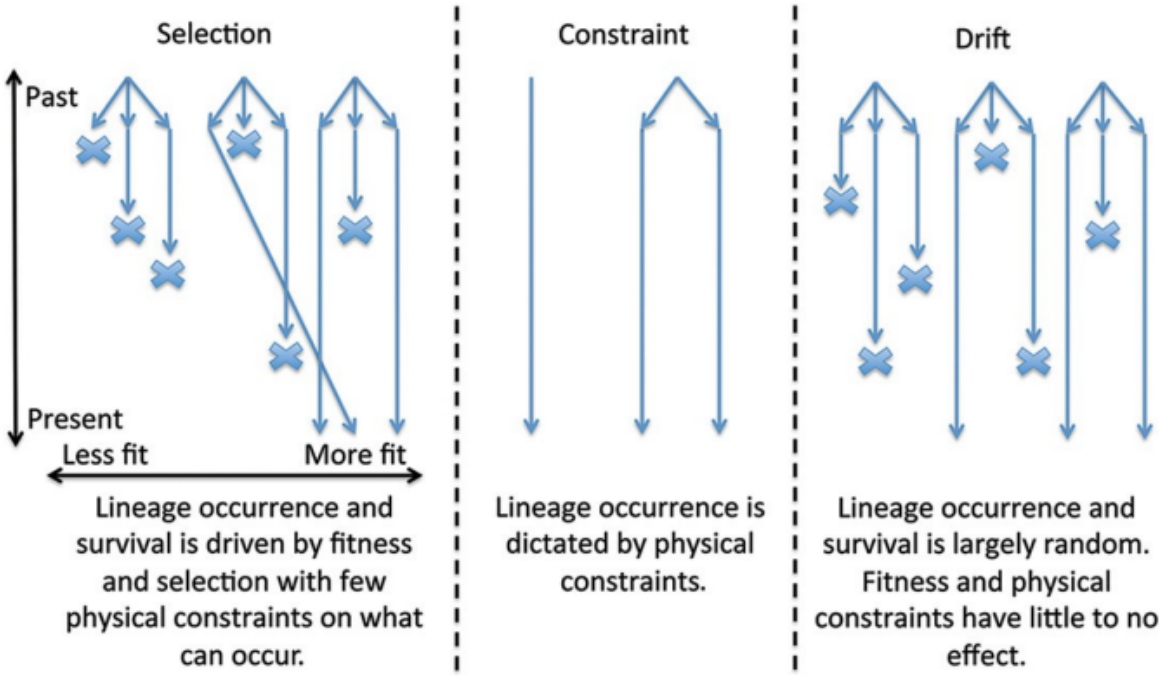
Genomic context: If you are always close to a gene that synthesizes tryptophan, maybe you are involved in the same pathway…

Info for the project: We can check what is similar to the gene (we can infer its function) and go to the STRING DB to check if it's in a specific pathway.

How many genomes must we include in a phylogenetic profiling (predicting the function)? If we add more genomes, we will have a better result. But at a certain number we reach a plato.
Also, there is no difference between using a subset with maximally diverse organisms (2 species for each phylum) or a subset with a randomly selected species.

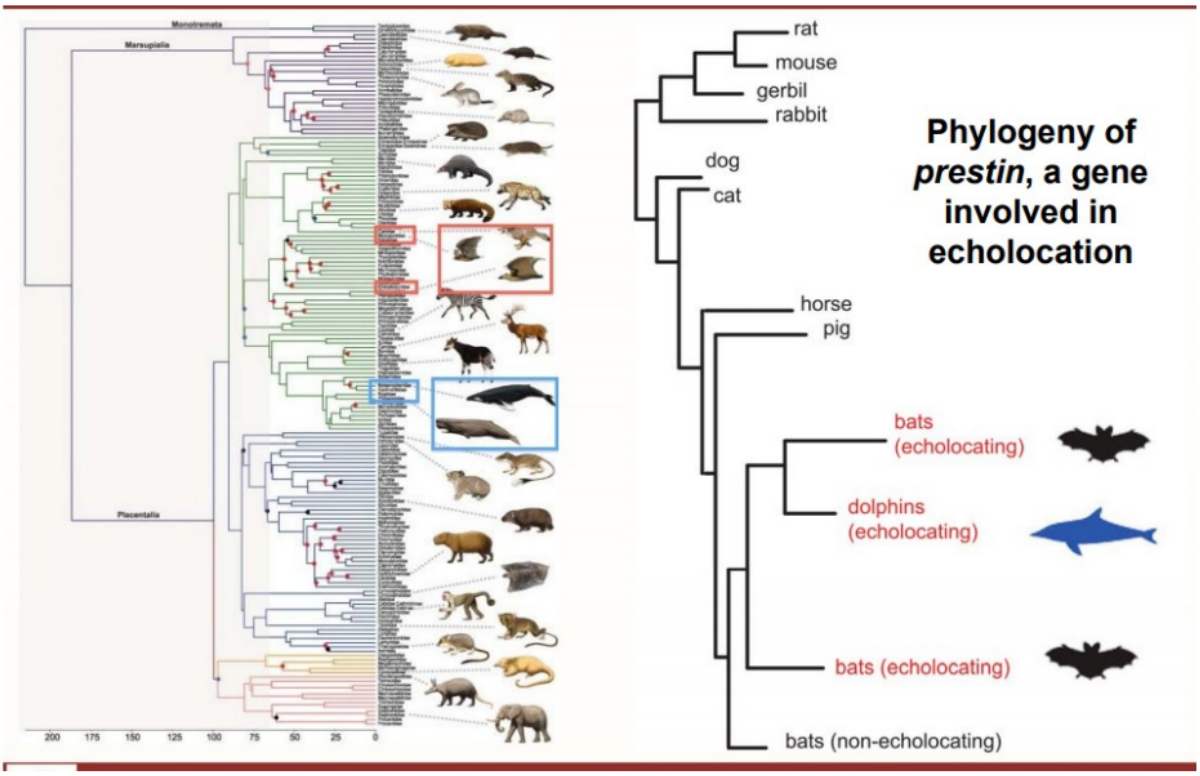
Convergent evolution
They adapt to a similar environment and therefore they acquire very similar traits. It's the independent evolution of similar features.
Convergent evolution creates analogous structures that have similar form or function but were not present in the last common ancestor of those groups.
Example of Cactus and Euphorbia.

There are different ways of explaining convergence:



Constraints: In the case of genome rearrangements, maybe two genes can't be separated because they are at both sites of a centromere (you can not break that part because then the chromosome would be unstable).
If there is a constraint, this implies that this has always been like that.

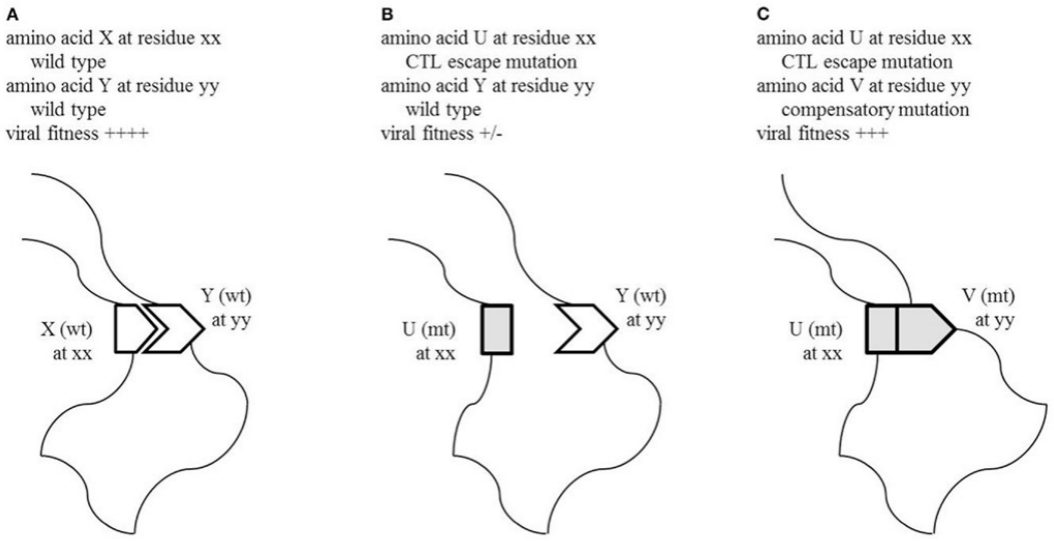
Positive selection can cause tree patterns different from the gene tree by convergent mutations. 2 species that are very distant can have some genes that have converged (gene tree will be really different to the species tree).



Coevolution at the sequence level

The structure of a protein is maintained because of weak interactions (electrostatic interactions) between the aa of the sequence.

If there is a mutation that changes an amino acid that is involved in the structure, then it will tend to have a compensatory mutation (mutations that correct a loss of fitness due to earlier mutations) to return to the initial (or similar) structure.



So, if in an alignment we see that there are some aa that tend to mutate together (in opposite directions. One goes from + to - and the other from - to +), we can make an hypothesis saying that both residues are interacting.

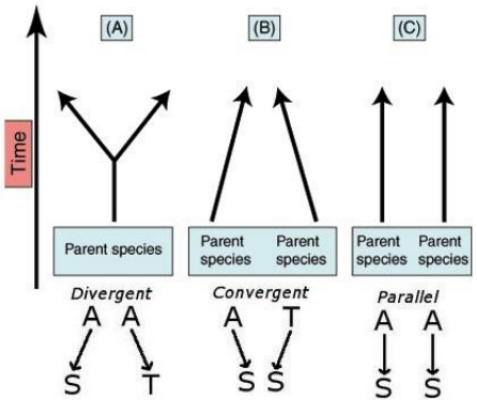
We can do this within a protein or also between different proteins. So, we can predict proteins that interact. So, we will need to check for correlations between aa of different proteins.

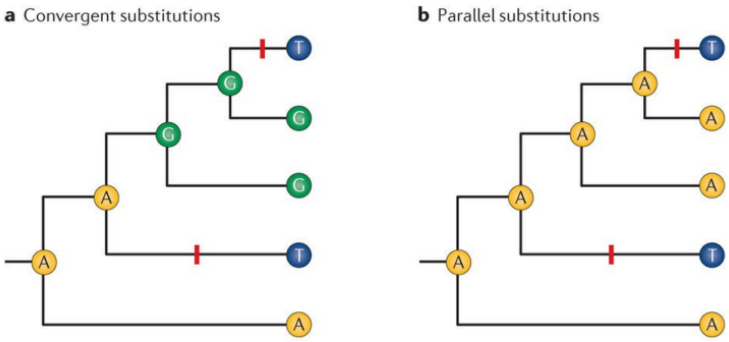
To do this, sometimes we don't even need to look at the alignments.

- Reconstruct the tree from each gene family
- We obtain a distance matrix from each tree
- Measure the correlation between the 2 matrices. If the 2 trees are similar (similar branch lengths…) there will be a high correlation. Then we expect the proteins to interact.
  Because if one protein evolves fast, the other will also evolve fast…

We can also look for coevolution between different organisms that interact (like host parasites, viruses).

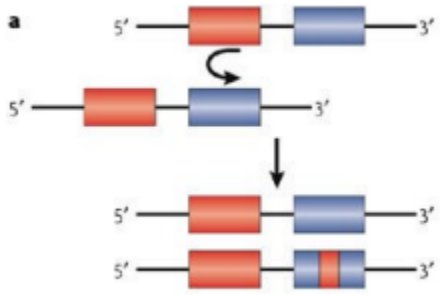
Difference between convergent and parallel evolution
Convergent come from different nucleotides
Parallel come from the same nucleotide: Process in which independent species acquire similar traits while evolving together in the same space and time.





Gene conversion and parallel, concerted evolution

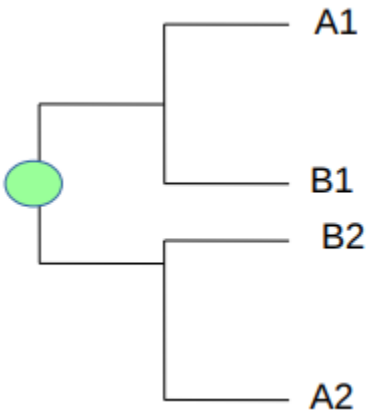
Gene conversion: In the same genome you have 2 similar genes (paralogs, for example) that can recombine and exchange some material. This happens because of DNA repair.



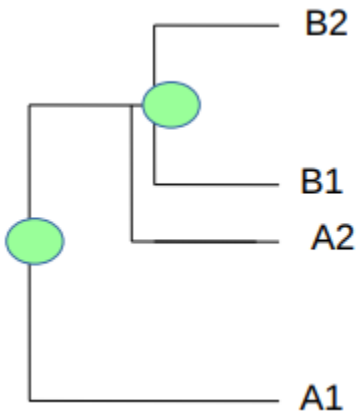
This may lead to confusion when doing analysis.

Imagine that we have a gene duplication and both genes diverge (so they are different but they are still similar because they are paralogs).

Then there is a speciation. So, we have species A and species B that both have the 2 genes. As expected, the same protein of different species is more similar.

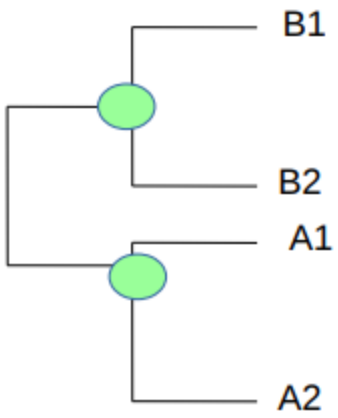


If there is gene conversion. B1 converts to B2. In other words, B2 is copied over B1. Then the gene tree changes, making you think that there has been a duplication in species B.



So, the orthologs are closer than ancient paralogs.

If it happens in species B, it can also happen in species A. Then it looks that there was an especiation at the beginning and then 2 parallel duplications.



Paralogs are closer than orthologs, apparent parallel duplication

Gene Expression Analysis

One way of knowing the function of a gene is knowing when and where this gene is expressed.
So, we can interrogate the transcriptome under different conditions, tissues… to understand the function of the genes.

The transcriptome can be described as the complete collection of transcripts present in a specific cell, tissue, organism… at a given time-point. The transcriptome is very dynamic (reason why every cell has the same DNA information but has a different phenotype).

The transcriptional process is highly regulated. The cell controls very well which genes are going to be expressed in each condition:

- Enhancer: Region in the DNA that attracts some activator proteins that will open the chromatin. That will allow other transcription factors to come and bind to promoter regions, where the transcription will start.
- There are repressors…
- Methilations that also regulate by promoting or inhibiting depending on the case.

The transcripts will be spliced and then they will be modified in the 3’ (addition of poly-A tail) and 5’ ends that will ensure the protection of the transcript.

How do we know when and where a gene is expressed?

The previous technologies that they used to do this analyses:

- RT-PCR: Has a very low throughput
- EST, SAGE… mid-throughput but bad coverage
- DNA microarrays: High-throughput, good information on expression levels, but no direct information on splicing, etc… not suitable to discover new transcripts.

RNA-seq

In most of the cases we convert the RNA into cDNA and then we do the analysis.
Others, like Oxford Nanopore (measures the changes in conductivity that goes through the pore. The conductivity depends on the nucleotide that is passing and the modifications) use the RNA directly.

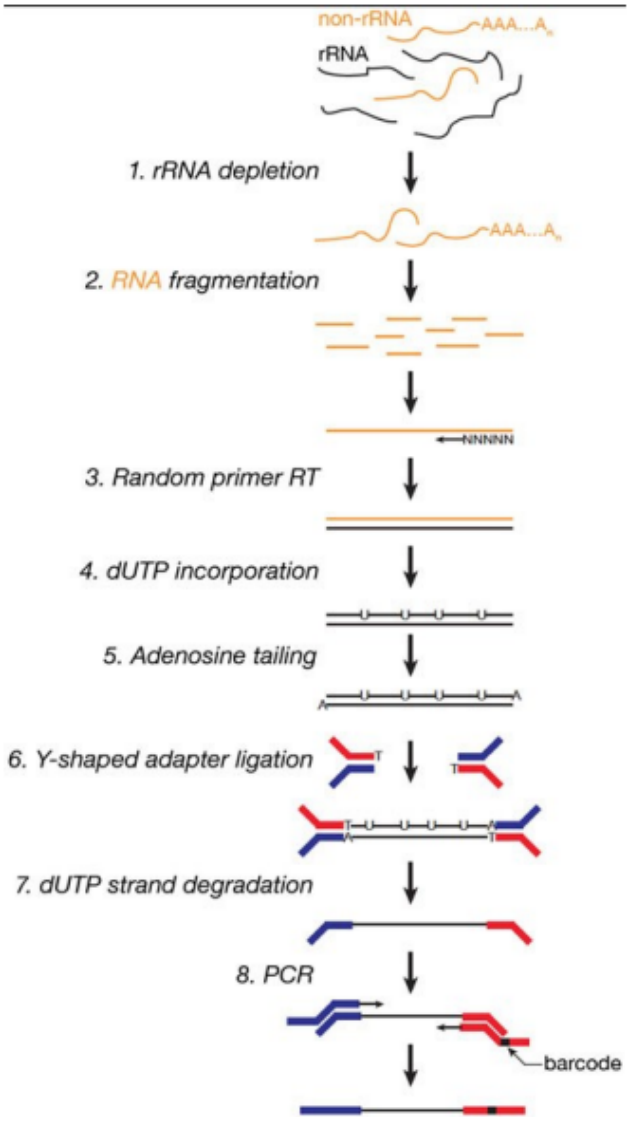
Generally, the term RNA-seq is used to indicate any RNA sequencing method based on a shotgun approach (not a specific fragment).

The advantage of a shotgun, sequence-it-all method, over a tag-based method, is the ability to quantify the expression level of each exon within a transcript, estimate their percent inclusion level and detect (differential) alternative splicing events.

However, it is difficult to identify the exact 3' and 5' ends of transcripts due to various technical biases (such as random hexamer priming or oligo dT priming) leading to underrepresentation of sequences near 5' and 3' ends.

Plenty of different protocols but they have many steps in common:

- rRNA depletion and fragmentation of the RNA
- Conversion of RNA into cDNA (performed by oligo dT or random primers)
- Second strand synthesis
- Ligation of adapter sequences at the 3' and 5' ends
- Final amplification



rRNA depletion: rRNA (90%) constitutes the majority of the RNA.
Thus, we need to get rid of it.

How can we selectively remove the rRNA? Use probes/oligos that bind specifically the rRNA. Then we add an enzyme that degrades the heteroduplex. The problem is that we need to know the sequence of the rRNA.

An alternative that works in eukaryotes is to capture the mRNA using an oligo of T. It will bind to the poly-A tail that is only found in mRNA.

Considerations:

- Bacteria have no poly-A tail.
- Ribosomal depletion kids are based on hybridization to specific sequences so they are optimized for specific species.

Target enrichment
It is also used to enrich some of the transcripts. But it is not based on the poly-A tail.
We have to design probes for the transcripts we want to capture.

Transcript orientation
All tag-based methods are strand specific, meaning that they preserve information about the transcript's orientation, shotgun methods may be strand-specific or not strand specific.

Strand specificity is important to determine the exact gene expression levels in the presence of antisense transcription, or for accurate prediction of certain classes of transcripts (e.g. lncRNAs)

Strand-specific methods can be classified into two categories:

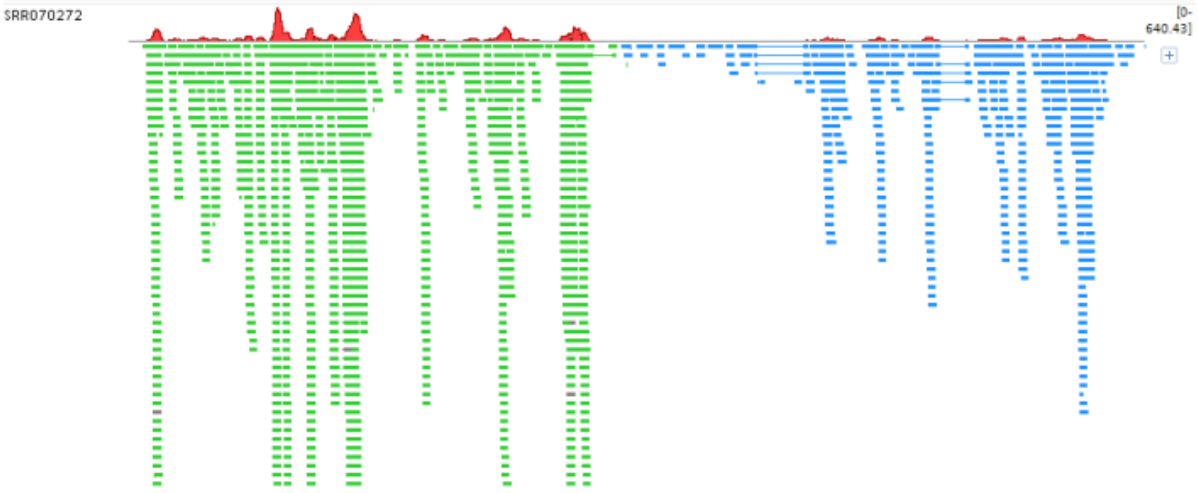
- RNA-seq methods based on ligation of two different adaptors in a known orientation relative to the 5′ and 3′ ends
- RNA-seq methods based on chemical modification of the RNA, either by bisulfite treatment or by the incorporation of dUTPs during the second-strand cDNA synthesis.

In both cases, the non-modified strand is degraded enzymatically

Bioinformatics analyses
Imagine that we have done this enrichment. Now we have all the sequences of RNA and we want to analyze them.
We have to translate this into the expression of different genes.

The easiest way is to map these reads into a reference genome.
We can also do transcriptome assembly (to try to reconstruct the transcripts) or de-novo transcriptome assembly. They do not need a reference genome.

This is an example of the results when mapping the reads into a reference genome.



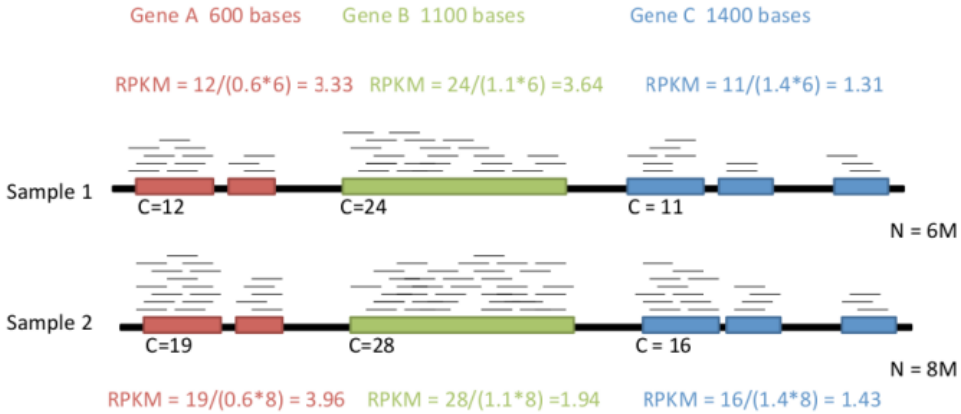
We can see that the green gene is more expressed, because it has more reads.

But can we quantify it?

The expression of genes is assessed by counting how many reads map to the gene, taking into account read length and total number of reads (RPKM or RFKM).

Note that we are normalizing using the length of the gene (because otherwise we could have more reads because the gene is longer).

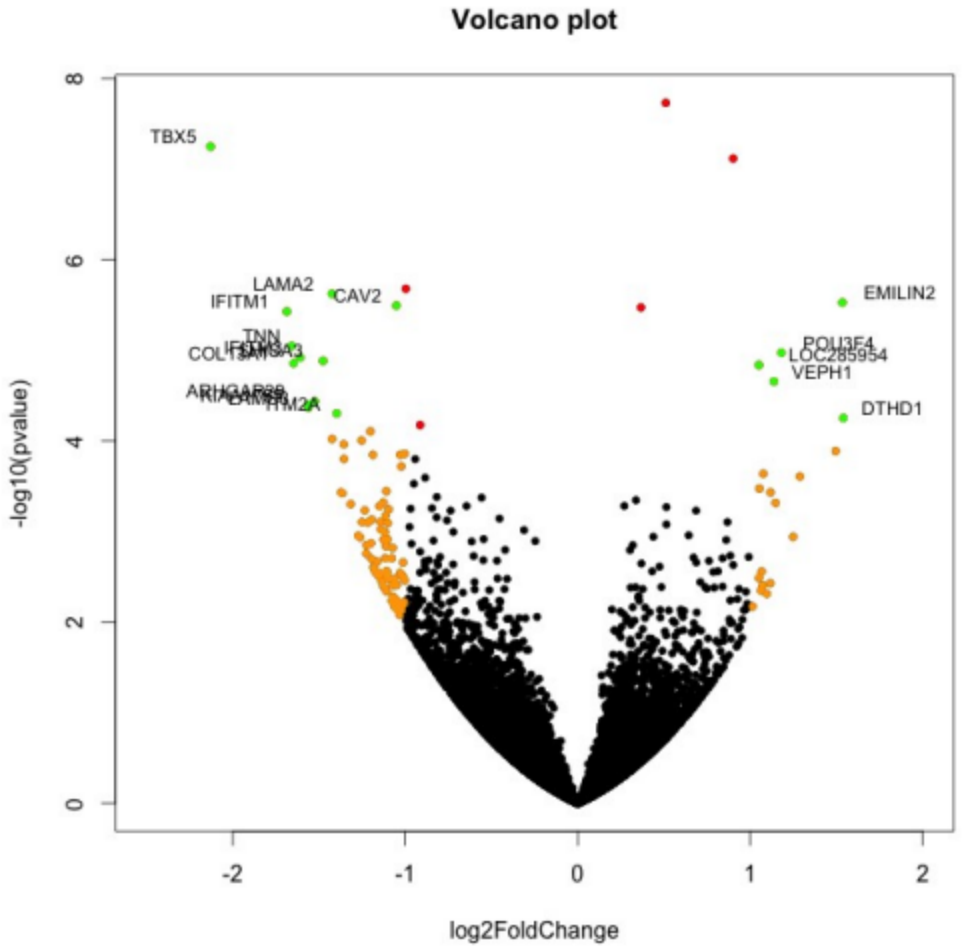
We also take into consideration the number of million reads that we obtained.



Counting rules

- Count reads, not nucleotides
- Count each read at most once.
- Discard a read if
  - it cannot be uniquely mapped
  - its alignment overlaps with several genes
  - the alignment quality score is bad
  - (for paired-end reads) the mates do not map to the same gene

Usually log2 Fold change and p-value thresholds are put to select differentially expressed genes. Each dot is a gene.



We want a small p-value and a big Fold change. Meaning that there is a different expression that is significant.

We can also use the information from the RNA-seq to reconstruct intron-exon structures of the genes and also to recognize alternative splicing. This is because when we map against the reference genome, we can see reads that map to a given exon and other reads that map to half of one exon and half of another exon. Meaning that there is an intron between.

Which of the following entities would necessarily be considered a gene?

- A transcript that is translated into protein that has an enzymatic activity. Wrong because a transcript is not a gene.
- Region of DNA that binds proteins that open the chromatin. This is an enhancer. To be a gene it needs encode for a molecule that has a function
- Region of DNA that is transcribed into RNA molecule that serves as a scaffold for enzymes
- All correct

Which statement is correct regarding orthology and paralogy?

- 2 homologous genes that are in genomes from different species are necessarily orthologs. Look at the incorrect sentences that the teacher explained in class. Homologous genes in the same genome are paralogs but not all homologous genes in different genomes are orthologs.
- Orthologs as compared to paralogs share a higher level of homology
- The best bidirectional hit approach is more prone to error when comparing closely related species.
- Clustering approaches define orthologous groups which contain both paralogs and orthologs.

Maximum likelihood phylogenetic reconstruction methods

- Provide the probability (likelihood) that the reconstructed tree is correct.
- Generally use a method called joined estimation to compute the likelihood of a tree over a set of parameters. Correct because maximum likelihood is the probability of that tree resulting from that alignment and that set of parameters. Not that the tree is correct (that's what bayesian inference tries to do)
- The 2 options above are correct
- Uses exhaustive searches.

Genes that have coevolved in terms of presence/absence (they are present or absent in the same species, so their phylogenetic profiles will be really similar)

- The comparison of their phylogenetic profile will show high jaccard indexes.
- The comparison of the phylogenetic profiles will show high hamming distances.
- 2 are correct
- Their molecular functions are likely to be equivalent. Wrong, because having very similar genetic profiles indicates that they are related in the same process (not the same function).

We want to predict orthology and paralogy relationships using a phylogenetic approach coupled with a reconciliation method.

- A node is a duplication node only when there are common species on either side of the node. This is true if we are using species overlap
- A node is a speciation node only when there are common species on either side of the node.
- A node is a duplication node only when there are no common species on either side of the node. False because a node can be a speciation node when there is no common species on either side of the node.
- A node is a speciation node only when there are no common species on either side of the node.

1. **What statement is correct:**
    1. **genes can be made of DNA or RNA**
    2. **genes are all transcribed regions of a genome**
    3. **genes are all the functional regions of a genome**
    4. **all the statements are correct**
2. **Out-paralogs, as compared to in-paralogs, are more:**
    1. **Appropriate to build species tree**
    2. **similar**
    3. **numerous**
    4. **divergent**
3. **What statement is correct:**
    1. **a species tree comprises speciation and duplication events**
    2. **branch lengths indicate level of confidence of a tree partition**
    3. **maximum likelihood is the fastest tree reconstruction model**
    4. **there are more possible rooted trees than unrooted trees**
4. **Regarding the reconstruction of gene trees:**
    1. **They are always rooted**
    2. **genes that contain duplictions cannot be used**
    3. **only orthologous genes can be used**
    4. **only homologous genes can be used**
5. **Gibbs sampling is a method to:**
    1. **discover conserved motifs in a set of sequences**
    2. **compute the similarity of two genomes in terms of gene order**
    3. **find shared structures between two RNA sequences**
    4. **calculate support of a phylogeny**
6. **Which signature is strongly suggestive of two genes having the same molecular function:**
    1. **the two genes encode the same protein domains**
    2. **all the answers are correct**
    3. **the two genes are co-expressed in the same tissue**
    4. **the two genes appear fused in the genome of another species**

7. **When can we say that a function (function A) is enriched in a gene set?**
    1. **when the function A is the most common among the genes in the set**
    2. **all the statements are correct**
    3. **when there are more genes with function A in the set than expected by chance in random samplings of the genome**
    4. **when the genes with function A are more expressed than the set of the genes**
8. **Which of the following statements about InParanoid is true?**
    1. **provides phylogenetic trees for several gene families**
    2. **is a method to predict protein domains**
    3. **none of the other answers is correct**
    4. **is a method to predict alternative splicing**
9. **RPKM unit of gene expression is obtained by:**
    1. **dividing reads by gene length and adjusting the GC content**
    2. **dividing read counts by total library size and adjusting GC content**
    3. **dividing read counts by total library size**
    4. **dividing read counts by total library size and gene/ transcript length**
10. **When is appropriate to use InterPro?**
    1. **all answers are correct**
    2. **when you have a genomic DNA sequence and are interested in gene annotation**
    3. **when you want to perform structural alignment of protein sequences**
    4. **when you want to know the function of an amino acid sequence or set of sequences.**

1. **If two genes are orthologous to each other, then**
    a) **They must be syntenic**
    b) **All responses are correct**
    c) **They must belong to different species**
    d) **They must have the same function**

2. **What statement is correct?**
    a) **Neighbor joining is more prone to long branch attraction than maximum likelihood**
    b) **Bayesian analysis is faster than neighbor joining methods**
    c) **Maximum parsimony is recommended for distantly related sequences**
    d) **There are more possible unrooted trees than rooted ones**

3. **What is the purpose of a gene concatenation approach?**
    a) **To build a species tree**
    b) **To detect genome rearrangements**
    c) **To predict the function of a gene**
    d) **To find syntenic genes**

4. **Which signature is strongly suggestive of two genes having the same biological function?**
    a) **The two genes encode the same protein domains**
    b) **The two genes are co-expressed across many tissues**
    c) **The two genes appear fused in the genome of another species**
    d) **All the answers are correct**

5. **Sorting by reversals is a method to**
    a) **Predict gene clusters**
    b) **Compute the similarity of two genomes in terms of gene order**
    c) **Discover conserved motifs in a set of sequences**
    d) **Reconcile gene and species tree**

6. **When two phylogenetic profiles are very similar**
    a) **All the statements are correct**
    b) **They have Jaccard index higher than 1**
    c) **They have hamming distance close to 0**
    d) **They have low mutual information**

7. **Which of the following statements about UniProt is wrong?**
    a) **Contains information of manually annotated proteins**
    b) **Allows you to convert other database identifiers to UniProt identifiers but not vice versa**
    c) **Contains information of computationally annotated proteins**
    d) **Is a freely accessible database of protein sequence and functional information**

8. **Which of the following statements about KEGG is true?**
    a) **It provides a genome browser that acts as a single point of access to annotated genomes for mainly vertebrate species**
    b) **Is a database for intensively studies model organisms**
    c) **Is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances**
    d) **Provides complete proteome sets for organisms whose genomes have been completely sequenced**

9. **Which of the following statement is true?**
    a) **Phred quality score of reads is stored in SAM and fastq files**
    b) **GFF file is required for fastq data trimming**
    c) **SAM file is a binary version of BAM file**
    d) **Each gene identifier has 4 lines in fastq file**

1. **What would constitute a gene according to the modern definition**
    1. **a piece of DNA that is replicated and transcribed**
    2. **a piece of DNA that regulates the transcription of another DNA**
    3. **a piece of DNA that is transcribed as an RNA that inhibits the expression of another RNA**
    4. **any DNA that, when deleted confers a phenotype**
2. **The function of a gene…**
    1. **it can be described by GO terms only if there is experimental evidence**
    2. **it remains constant over the course of evolution**
    3. **is best determined by comparing its sequence with a database**
    4. **depends on the structural propertie of the molecule that is encoded by the gene**

3. **Which sentence is wrong?**
    1. **GO is the only controlled vocabulary that can be used to functionally annotate genes**
    2. **gene ontology terms can be used for genes with no experimental information**
    3. **GO terms are useful to detect what functions might be over-represented in a given dataset**
    4. **Gene ontology entries inform on the source of annotation**
4. **Regarding homology:**
    1. **paralogous genes are homologs**
    2. **all statements are correct**
    3. **orthologous genes are homologs**
    4. **homologs are genes that share a common ancestor**
5. **What is a promiscuous domain**
    1. **Domains that have many different functions**
    2. **Domains that can bind many other protein sequences**
    3. **all statements are correct**
    4. **domains that can be present in many different protein families**
6. **Orthologs, as compared to homologs, are more**
    1. **likely to have complementary functions**
    2. **all options are correct**
    3. **appropriate to reconstruct a species tree**
    4. **similar in terms of their sequence**
7. **Select the incorrect statement regarding phylogenetic trees:**
    1. **the newick format is a graphical representation of phylogenetic trees with edges and nodes**
    2. **bootstrap values provide information on how much support a give clade has from the analysis**
    3. **in a phylogenetic tree all edges can be rotated without changing the topology**
    4. **for a given topology, there are more rooted trees than unrooted trees**
8. **Out paralogs are…**
    1. **not necessarily encoded in the same genome**
    2. **gene that result from duplication**
    3. **more distantly divergent as compared to in-paralogs**
    4. **all the responses are correct**
9. **Associate the correct questions and answers:**
    1. **Alpha and beta globin are: paralogous**
    2. **Wings of bats and bird are: non-homologous**
    3. **Mammals are: monophyletic**
    4. **winged animals are: polyphyletic**
10. **Phylogenetic reconstruction methods based on Bayesian analysis:**
    1. **Need a set of prior probabilities for the parameters of the model**
    2. **are generally the fastest among tree reconstruction models**
    3. **need to be run on a bootstrapped set of alignments to assess the support of the tree partitions**
    4. **generally use a method called joint estimation to compute the likelihood of a tree over a set of parameters**

11. **Match the following concepts and algorithms with the correct context:**
    1. **Chromosome painting: is an algorithm to detect large chromosomal rearrangements**
    2. **Sorting by reversals: is an algorithm to find the minimum number of rearrangements between two series of elements**
    3. **Gibbs sampling: is an algorithm to discover enriched motifs that are enriched in a set of sequences**
12. **REgarding the reconstruction of species trees:**
    1. **The more species you compare, the fewer conserved genes you can use in gene concatenation approach**
    2. **all existing methods rely on sequence alignments at some point**
    3. **the super-tree approach is the fastest and most accurate method**
    4. **genes that contain duplications cannot be used**
13. **Which statement is correct regarding protein domains**
    1. **Promiscuous domains mediate binding to other domains**
    2. **they can exert a function independently of the rest of the protein**
    3. **they are longer than 200 amino acids**
    4. **they are separated by introns in eukaryotic genes.**
14. **Genes that co-evolved, being similarly present or absent form genomes:**
    1. **the comparison of their phylogenetic profiles will show high Jaccard Indexes**
    2. **all the answers are correct**
    3. **the comparison of their phylogenetic profiles will show high Hamming distances**
    4. **their molecular functions are likely to be equivalent.**
15. **Maximum likelihood phylogenetic reconstruction methods:**
    1. **all the answers are correct**
    2. **generally use a method called joint estimation to compute the likelihood of a tree over a set of parameters**
    3. **provide the probability (likelihood) that the reconstructed tree is correct**
    4. **uses exhaustive search**
16. **Adjacent genes are:**
    1. **co-regulated**
    2. **part of a gene cluster**
    3. **neighbors to each other**
    4. **all answers are correct**

Which of the following entities would necessarily be considered a “gene” by the modern definition.

a) all statements are correct
b) a DNA region which is transcribed into a tRNA
c) a protein that has enzymatic activity
d) a RNA transcript that inhibits the function of other RNAs

Question 1

Provide the shortest and most inclusive definition of a gene

Any sequence of DNA or RNA that codes for a molecule that has a function

Question 2

What is a phylogenetic profile? How can we use them to know the function of an uncharacterized gene?

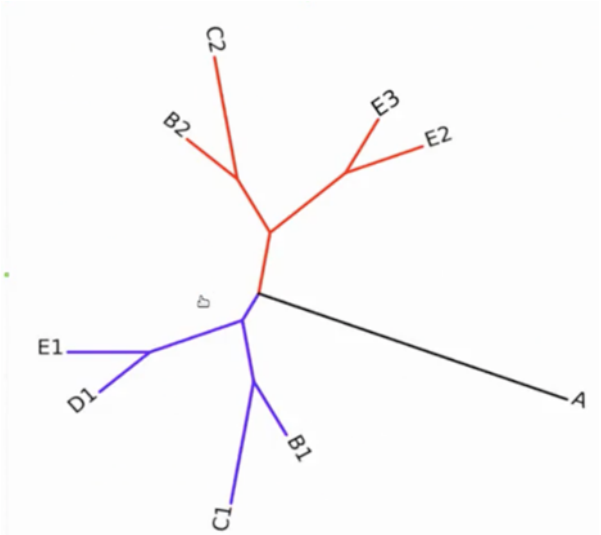
Describes the presence or absence of a protein in a set of genomes. Similarity between profiles is an indicator of functional coupling between gene products.

If our objective is to characterize an uncharacterized gene, we can take a look at the phylogenetic profile and try to find a gene that has a similar phylogenetic profile. Meaning that if there are similarities between both profiles, then both genes are likely to be involved in similar biological processes / does not mean that they have the same function .

Question 3

Given the gene tree below, where letters indicate species and number genes (i.e. B1 and B2 are two genes of species B, and in which A can be used as an out-group). Using the algorithm of your choice

- **List all orthologs of B2: C2, E2, E3, A**
- **List all paralogs of B2: C1, B1, D1, E1**
- **Write the most likely species tree in newick format. (((D, E), (C, B)), A)**



Question 4. The plot below shows a hypothetical comparison between the genomes of two bacterial strains (Strain 1 and Strain 2, of which strain 2 has been evolved from strain 1)

- **Explain what type of plot is this one and its utility**

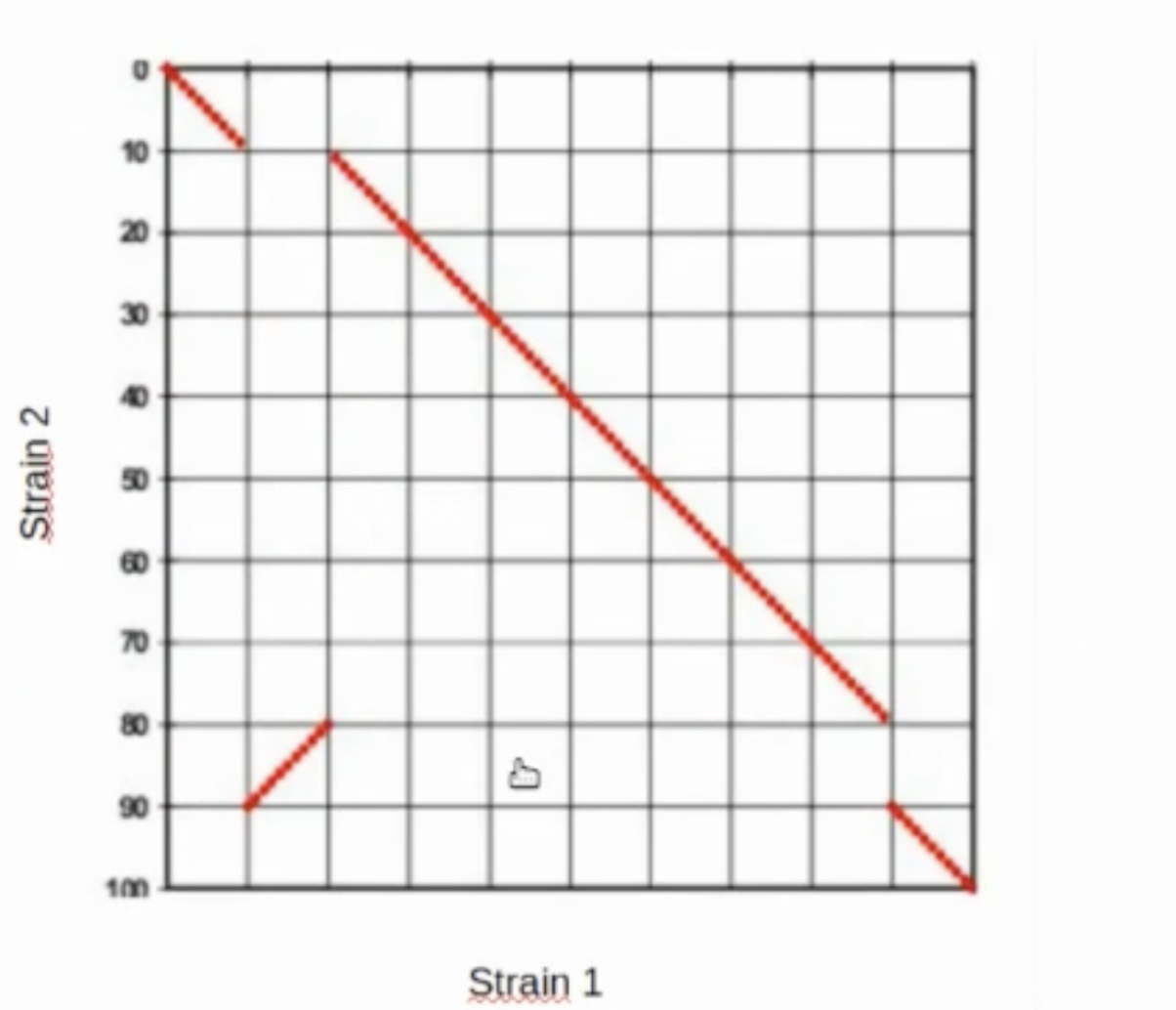
This is a dot plot. It's a graphical method for comparing two biological sequences and identifying regions of close similarity.

- **What could be the unit of the axes?**

Casi segur que es Kilobases (Kb).

- **Can you reconstruct what rearrangement(s) occured during strain 2 evolution from its strain 1 ancestor.**

Translocation in the 10th place, where 10 Kb have been moved to the 80th place. This translocated region has also been inverted.



Question 1

What is the mirror tree approach? What is its purpose and what is it based on? Can you explain how it works?

You have two gene trees from which you derive their distance matrices.
Then we measure correlation between both matrices.
If both matrices are highly correlated, then it is likely that both proteins interact (since they have coevolved for a long time)
You can do this between proteins of different organisms (parasites and hosts)

Question 2

Provide the shortest and most inclusive definition of a protein domain. What is a promiscuous domain?

Conserved part of a given protein sequence and tertiary structure that can evolve, function and exist independently of the rest of the protein. Promiscuous domains are domains that are present in more than one protein family.

Question 3

Given the following gene tree in newick format:

(R1,(((H1,H2),C1),(H3,C2)))

Where R represents Rat genes, H human genes, and C chimpanzee genes. Using the algorithm of your choice (indicate it).

a) **Indicate which genes are orthologous or paralogous to each of the human genes.**

Paralog to H1: H2, H3, C2
Ortholog to H1: C1, R1
Paralog to H2: H1, C2, H3
Ortholog to H2: C1, R1
Paralog to H3: H1, H2, C1
Ortholog to H3: C2, R1

b) **Considering the human lineage as a reference, sort the paralogs of H1 as in- and out- paralogs.**

- In-paralogs: H2
- Out-paralogs: H3, C2

Question 4

This is a typical volcano plot obtained after comparing gene expression of genes in two conditions (A versus B).

a) **Can you explain what is represented in each of the axes, and what units are typically used? -** log-scaled P-value (Y axis) and log-scaled fold change (X axis)

b) **What is represented by each of the dots?** A gene

What is the difference between green, brown, and black dots? Black dots are not significant and don't have a differential expression. Brown dots have a differential expression but are not significant and green dots have a significant differential expression.

And those on the left, and right? If they are up or down regulated.

c) **Can you identify the gene that changed most of its expression?** TBX5