

Molecular evolution and phylogenetics Questions

Lecture 1

1) Pictured below are the water lily (*Nymphaea alba*) and the sacred lotus (*Nelumbo nucifera*).

a) Are these plants related?



water lily



lotus

They belong to different plant families and are not closely related.

b) If they are placed into the same taxonomic group, would you consider this group as monophyletic, paraphyletic, or polyphyletic?

If they did not have a common ancestor I would consider the group polyphyletic as they would be placed in the group based on the homoplasy traits (e.g. floating plants)

However, they do share a common ancestor, therefore I would consider the group either paraphyletic or monophyletic, depending on the classification of different plants species.

c) Is the similarity in appearance due to homology or homoplasy (if homoplasy, then consider the following options: convergent/parallel evolution, secondary loss)?

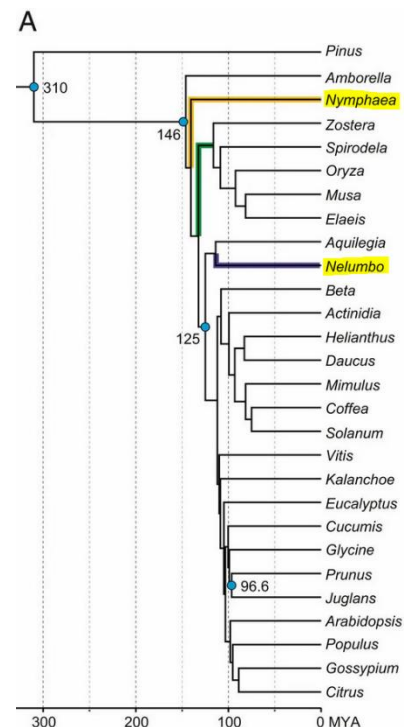
Homoplasy, specifically **convergent evolution**. Homoplasy refers to similarities in traits that arise independently in different evolutionary lineages and are not inherited from a common ancestor.

In the case of water lilies and lotus flowers, they have adapted to similar aquatic environments, so traits such as floating leaves and flowers above the water surface are a result of adaptation to similar ecological niches rather than shared ancestry.

2) Will a taxonomic group “parasitic plants” make any sense?

It would be a **polyphyletic** grouping of taxa without common ancestor, as parasitic behaviour can develop independently across different plant families.

(There are 2 types of parasitic plants: Holoparasites which have lost their chloroplasts, and Hemiparasites: retain functional chloroplasts but still rely on their hosts for certain nutrients.)



3) Would you prefer morphological or molecular markers for generating phylogenies?

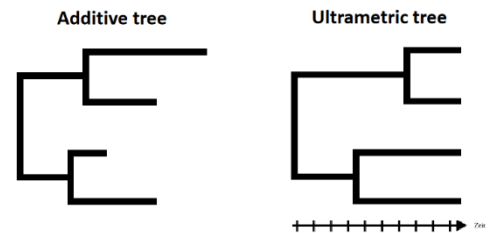
Molecular markers because morphological traits don't always reflect genetic distances. Molecular markers overall provide more objective and precise data and provide insight into genetic diversity that may not be apparent based on morphological characteristics alone (e.g. closest relative of a whale is hippopotamus).

Additionally morphological traits can be influenced by environmental factors and can vary across different life stages of an organism, making them less reliable.

4) What is the difference between an additive tree and an ultrametric tree?

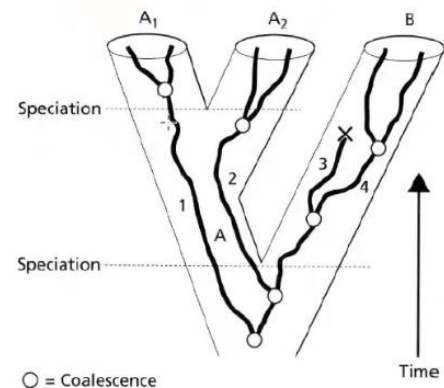
Additive tree - also called metric tree or phylogram. Branch lengths match divergence.

Ultrametric tree - also called dendrogram. Tree plotted against time scale. Branch length indicates time of divergence.



5) Species trees and gene trees may not coincide, even though both could be correct, technically speaking. Explain why.

In the figure on the right the alleles of a certain gene are indicated with numbers 1-4. A and B are different species that developed over time. Allele 2 occurred later in the evolution than allele 1. Both were found in the species A until it diversified into A1 and A2. Later alleles 3 and 4 occurred and are found within species B. However, genetically allele 2 in species A2 can be more closely related to allele(s) 3 or 4 found in species B.



Summarising: Alleles found in a certain species can pre-date the formation of that species (A2) and be therefore more closely related to an allele found in a different species (B) rather than a species that originated from the same ancestor (A1).

6) What information is contained in a distance matrix?

A distance matrix shows the amount of change from one taxon to another one.

7) Can you use chloroplast sequences universally for plant phylogeny, or are there exceptions?

No, there are exceptions. There are parasitic plants that lost their chloroplasts. In that case mitochondrial sequences are used for the phylogeny.

(There are 2 types of parasitic plants: Holoparasites which have lost their chloroplasts, and Hemiparasites: retain functional chloroplasts but still rely on their hosts for certain nutrients.)

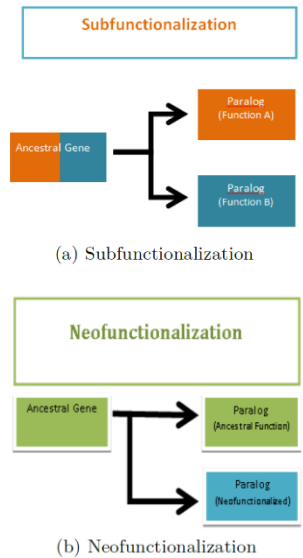
8) How can whole-genome duplication influence the evolution of genomes?

It can lead to tetraploidy. Over time the genes will get lost, become pseudogenes or genome rearrangements take place. In some rare cases some duplicated genes will be retained and evolve independently. Ultimately after a long time, the genome can return to a diploid state.

9) What is subfunctionalization, what is neofunctionalization?

In some rare cases some duplicated genes will be retained and evolve independently

- **Sub-functionalization** – the original gene copies remain unchanged but the duplicate only keeps a part of the function. E.g. if a gene was previously expressed in two different tissues, the original gene can be then expressed in one tissue and the duplicate exclusively in another tissue. Then for example if an enzyme is able to catalyse two different reactions. Each of the new proteins evolve to catalyse one reaction only.
- **Neo-functionalization** – one of the gene copies keeps the original function and one acquires a completely different function (e.g. by mutation)

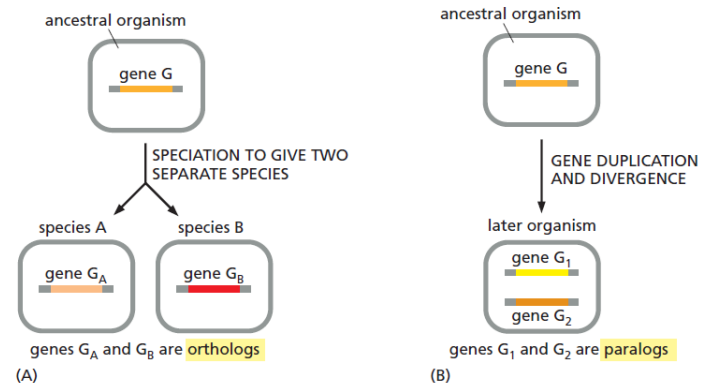


Lecture 2

1) Describe: what is an ortholog, what is a paralog?

Paralogs: Related genes that have resulted from a gene duplication event within a single genome and have diversified in their function.

Orthologs: Genes in two separate species that derive from the same ancestral gene, the genes gradually become different in the course of evolution, but they are likely to continue to have corresponding functions.



2) What is a pseudogene, and how does it differ from a functional gene?

A pseudogene is a DNA sequence closely resembling that of a functional gene, but containing numerous mutations that prevent its proper expression or function. Most pseudogenes arise from the duplication of a functional gene followed by the accumulation of damaging mutations in one copy.

Closely related species can share the same pseudogenes, which makes them useful for phylogenetic analysis.

3) What are four-fold degenerate sites?

Four-fold degenerate sites are specific positions in the DNA/RNA sequence where exchanging a nucleotide results in a silent mutation. Because some amino acids (e.g. glycine, valine, proline) are coded by four different codons, exchanging a nucleotide at the last position of a codon will not result in producing a different amino acid in the corresponding protein.

4) Why do transversions occur less frequently than transitions?

- Transition = exchanging a purine base against another purine base OR pyrimidine against a pyrimidine
- Transversion = exchanging of a purine against a pyrimidine or vice versa.

Transitions occur more frequently due to the chemical and structural similarities between the nucleotide bases involved in transitions. Additionally, transitions maintain the same type and number of hydrogen bonds (two for purines and three for pyrimidines) in the base pair, making the exchange “smoother”.

5) It has been observed that insertion/deletion events within coding regions frequently affect three consecutive bases (rather than 1 or 2 bases). What is the explanation?

The key factor is the **reading frame**. Since the genetic code is read in a frame of 3 nucleotides, a removal/addition of a single nucleotide (or anything that is not a multiple of 3) would lead to a frame shift and a most likely non-functional or very different protein. In contrast by insertion or deletion of 3 nucleotides maintains the reading frame.

6) What mechanisms exist for preventing mutations to occur? Think about point mutations and insertions/deletions?

DNA Polymerase has a proofreading mechanism, therefore errors can be corrected already during replication. After replication the Mismatch-Repair system can remove mispaired bases and correct small insertion/deletion loops. Also during transcription RNA Polymerase and later during translations the ribosomes have a proofreading mechanism.

7) Think about an experimental design how to determine the human nuclear mutation rate with current technology.

Indirectly: The mutation rate can be estimated due to known data such as the error rate of proofreading mechanisms etc.

Directly: by sequencing of the genomes of a “trio”: mother, father, and their child

8) Name the two large groups of transposable elements and mention the key feature that distinguishes them.

Class I: Retrotransposons – move through a “copy-and-paste” mechanism: The retrotransposon is transcribed into RNA, then reverse transcribed into complementary DNA and integrated into the genome.

Class II: DNA transposons – move through a "cut-and-paste" mechanism. The DNA transposon is “cut” from one genomic location and inserted into another site. This is mediated by the enzyme transposase.

9) How can transposable elements interfere with the function of the genes of their hosts?

Depending on where the transposable element (TE) is inserted, will have a different impact:

- Gene disruption: Insertion of a TE into the coding region of a gene can disrupt the functionality of the gene.
- Altered Gene Regulation: insertion into the regulatory regions of a gene can influence its expression either upregulating or downregulating it.
- Epigenetic Changes: TEs can influence epigenetic modifications, such as DNA methylation and histone modifications.
- Remain neutral: Insertion into a non-coding region can remain without any major impact.
- Creation of new genes

10) How can you identify repetitive sequences in the genome?

Various methods can be used, partially depending on the searched element. This also defines where and how I would be looking for the specific element (e.g. LINEs in heterochromatic AT-rich regions, microsatellites have known patterns etc.)

Some methods:

- BLAST - compare a sequence of interest against a database of known repetitive elements.
- FISH (Fluorescent In-situ hybridisation) using a repetitive sequence as probe
- Sequencing

Lecture 3

1) Is gene number a good measure for assessing the complexity of an organism?

No. A higher gene number does not always correspond to the complexity of an organism. E.g. *A. thaliana* has a more genes than humans.

2) Would you expect similar substitution frequencies for genes encoding for proteins of different functions?

No, the substitution frequency of a gene/protein depends strongly on its function. Conserved proteins, like for example ribosomes, show little changes over time. Other proteins, which are not necessarily needed for survival, can have higher substitution rates.

3) What is synteny? Will synteny increase or decrease with evolutionary distance?

Synteny means a **conserved order genes** on chromosomes between different species. When two or more genomes display synteny, it means that they share a similar gene order. The higher the evolutionary distance the more likely it is that some chromosomal rearrangements disrupted the gene conservation resulting in a **decrease** of synteny.

4) Think about organisms which benefit from a high mutation rate. How could they achieve high mutation rates?

Mutation rate increases the less control mechanisms there are. For example, the bacterial Taq polymerase has no proof-reading mechanism which leads to comparatively higher rate of mutations (increase of error rate by about 2 orders of magnitude).

5) What is an allele?

Alleles are different versions of a gene in a population. They differ by nucleotide substitutions or indels. The phenotype can therefore also be affected: different alleles can have different phenotypes. For example, the colour of the eyes can change. Or enzymes with different kinetics.

6) What is meant when we speak about alleles showing qualitative or quantitative differences, respectively?

These are terms used to describe different aspects of the variation in the alleles present at a particular gene:

- **Qualitative** differences in alleles can for example be a different eye colour. Or that a certain metabolic pathway is shut down or compromised.
- **Quantitative** differences in alleles can for example be the expression pattern of proteins. So that the protein itself is not affected, just its regulatory region which in turn influences the amount of protein expressed. An example would be the digestion of lactose in adult humans. In northern Europe an allele is frequent, which enhances the expression of the lactase gene.

7) When is an allele considered to be rare? How can you assess if an allele is rare? What could be the reasons that an allele is rare?

Alleles with frequencies less than $MAF < 5\%$ (minor allele frequency) are generally considered rare. Rare alleles are present in a small proportion of the population.

One possible reason for a rare allele is that it is relatively new. If it is not associated with a fitness disadvantage, it can further increase its frequency in the population. Another reason could be a disadvantageous allele which becomes removed from the gene pool.

8) Explain what is meant by Hardy-Weinberg-equilibrium and under which conditions such an equilibrium will be maintained.

In a Hardy-Weinberg equilibrium, the genotypes and allele frequencies stay constant from one generation to the next. The genetic variation does not get lost in the population. For the Hardy-Weinberg equilibrium, an ideal population is implied:

- Infinitely large population
- No migration
- Random, successful mating
- No selection
- No mutation

- 9) Explain the concept of overdominance. Give an example. Would you expect heterozygosity at a locus upon which overdominant selection is acting to a) persist or b) to disappear.

Overdominance also known as heterozygote advantage, is a phenomenon in population genetics where the heterozygous (Aa) individuals have a higher fitness than both homozygous (AA, aa) individuals.

An example is the Hb^s allele: Homozygotes (aa) become sickle cell anaemia because they have two "bad" alleles. Homozygotes (AA) on the other hand have a high risk of getting infected with malaria. The heterozygotes (Aa) don't get sickle cell anaemia and additionally are better protected from malaria and have therefore an advantage. In regions with a high risk of malaria infection, the Hb^s allele is advantageous and therefore widespread.

Therefore, the over dominant selection will lead to persisting of the heterozygosity unless a factor from the outside (eg. extermination of malaria, migration) causes the balance to change.

- 10) What is disruptive selection? Which biological process may be augmented by disruptive selection?

Disruptive selection is also called underdominance or heterozygote disadvantage. The heterozygotes (Aa) have a disadvantage in contrast to both homozygotes (AA, aa). This means that the hybrids are less fit. Underdominance could be a major driver for speciation.

- 11) Genetic drift and the generation of new mutations are opposing forces. Explain!

Genetic drift change in the frequency of an existing gene variant in a population due to random chance.

In the absence of selective pressures, the fate of neutral mutations is primarily determined by genetic drift. When a new neutral mutation arises, its initial frequency in the population is low, often close to $1/(2N)$, where N is the population size. This is because each individual in a diploid population has two copies of a gene, and the mutation may initially appear in only one copy.

Over time, the frequency of a neutral mutation may increase or decrease due to random sampling effects. Eventually, the mutation may become fixed in the population (present in all individuals) or be lost entirely. By fixation of the new alleles over time, genetic drift pushes the population towards homozygosity, while the generation of new alleles pushes the towards heterozygosity.

- 12) What is "coalescence" and how does it relate to a "most recent common ancestor"?

Coalescence means that all alleles present in a population ultimately trace back to a common ancestral allele of the "most recent common ancestor".

- 13) A population may go through a "population bottleneck". What consequences may arise?

A population bottleneck means a drastic reduction of the size of the population and thereby a decrease in its genetic diversity. Rare variants can be present in the new founder population and increase their frequency by genetic drift, some alleles may be lost entirely from the population, while others may become more prevalent.

14) How can you distinguish a population bottleneck from a selective sweep?

Population bottleneck: Drastic reduction in population size -> Genetic diversity decreases

Selective Sweep: The frequency of a new allele increases rapidly due to positive selection. This leads to a decrease in genetic variation, because other alleles get removed from the gene pool. But a selective sweep is only confined to a small region of the genome.

Lecture 4

1) What is a species?

Two organisms belong to the same species if they can reproduce and if their offsprings have no fitness disadvantage, e.g. they can also reproduce. Members of the same species don't necessarily share the same morphology, but they have a high molecular similarity.

2) Explain what is meant by "negative epistatic interactions".

Most of the time, genes/alleles work in combination. They can have positive or negative effects on each other. If genes/alleles don't work well in combination, it is called negative epistatic interactions. It can lead to reproductive isolation which is a major factor for speciation.

3) Explain why the concept of sympatric speciation is sometimes met with skepticism.

Sympatric speciation is the evolution of a new species from a surviving ancestral species while both continue to inhabit the same geographical area. It has been a subject of debate because the absence of geographic isolation challenges traditional views of speciation.

4) How can one distinguish between incomplete lineage sorting and introgression?

Incomplete Lineage Sorting is when the tree produced by a single gene differs from the population or species level tree. This happens when derived genes/alleles from an ancestor are not distributed evenly to the descendants.

Introgression is the transfer of genes/alleles from one species into the gene pool of another one. This means, distantly related species can share the same gene, even though their common ancestor hadn't had this gene. For example, neanderthal genes in modern homo sapiens.

To distinguish between them one should probably determine the age of the gene of interest and compare it with the age of the last common ancestor. This can be done with the ABBA-BABA test.

5) F_{ST} statistics allows you to identify regions within the genome which are under diversifying selection. Can you point out what data you need to calculate F_{ST} values

The Fixation index (F_{ST}) is a statistical measure of genetic differentiation of subpopulations relative to the total population. For calculating the Fixation index one needs the differences between the taxa. E.g. comparing the genome in a sequence alignment and counting the nucleotide substitutions. The more differences, the more likely it is that there is no gene flow. In other words: In the best case one has the complete sequenced genomes of the species of interest and can align them.

6) When comparing F_{ST} values between closely related species genome-wide, would you expect higher F_{ST} values between pairs of allopatric or sympatric species?

In allopatric species, populations are geographically isolated from each other, leading to limited gene flow. Over time, genetic differences can accumulate due to independent evolutionary processes in each population, resulting in higher F_{ST} values as a measure of genetic differentiation.

In sympatric species, populations coexist in the same geographic area, allowing for potential gene flow and interbreeding. The ongoing gene flow tends to reduce genetic differentiation between populations, resulting in lower F_{ST} values.

7) Are extent of phenotypic difference and reproductive isolation always correlated?

No, **not always**, because there are other factors that play a role as well. E.g. negative epistatic interactions in tomatoes lead to high reproductive isolation but little phenotypic difference (-> the negative interaction between genes may suppress the phenotypic divergence).

Another example: sticklebacks living in the ocean have a very different phenotype than sticklebacks inhabiting sweet waters, but they can be easily bred as there is only a small reproductive isolation.

8) Explain what effects you will see in case of acquisition of genetic material (genes, chromosome segments, etc.) on gene tree topology.

The effects will be different depending on how new genes were acquired:

- **Horizontal gene transfer** (e.g. viral transduction) will result either in the minor tree or dominant tree. One of tree topology will dominate. E.g. minor tree when the transferred gene was derived from different species.
- **Introgression** - depending on the age of introgression also one of the tree formats (dominant or minor) will dominate.
- **Linear fusion** - will result in two codominant, possibly conflicting trees at similar frequencies.

9) What do we mean by a hard polytomy?

In a polytomy, the exact order of relationships among the diverging lineages are not determined, and they are depicted as emerging from a common node without clear hierarchical relationships.

Hard polytomy means that the speciation of three or more lineages took place within very short time interval. Additional data will not resolve the tree as opposed to soft polytomy where unresolved branching is a result of missing data.

Lecture 5

- 1) Explain how the d_N/d_S ratio test can help to study gene evolution. For a randomly chosen pair of orthologs, what d_N/d_S ratio would you expect (<1 , 1 , >1) and why?

d_N/d_S ratio compares the rates of non-synonymous (dN) (=aa change) to synonymous (dS) (=no aa change in the protein) nucleotide substitutions in a pair of orthologous genes. The result of the ratio can have the following meaning:

- **d_N/d_S ratio = 1 neutrally evolving** -> d_N and d_S substitutions are occurring at an equal rate, therefore the gene is evolving without a significant selective pressure.
- **d_N/d_S ratio > 1 positive selection** - d_N substitutions are favoured, suggesting that amino acid changes are being positively selected, e.g. to adapt to new environment.
- **d_N/d_S ratio < 1 negative selection** - means that natural selection is working towards preserving certain amino acid sequences (e.g. functional proteins), and non-synonymous changes are often deleterious.

For a randomly chosen pair of orthologs I would expect **d_N/d_S ratio < 1 negative selection** because an organism needs functional proteins to work, therefore most of protein-coding genes have to be negatively selected.

- 2) Explain the concept behind the McDonald-Kreitman test (MKT). What data are needed? What aspects of gene evolution can be solved with the MKT in a better way compared to a d_N/d_S ratio test?

The McDonald-Kreitman test is used to detect if a specific site within a gene is under positive selection. The test compares polymorphic variations within species to fixed differences between them. The key assumption of the MKT is that the ratio of (P_n/P_s) within species = (D_n/D_s) between species. But under selection this ratio is not equal.

Data needed for the MKT: 1) Sequences of protein-coding genes from a set of closely related species -> derive the divergence (substitutions) from here and 2) Polymorphism Data

The d_N/d_S test looks at the entire gene sequence, and therefore does not allow to make conclusions about specific residues of a protein, while **MKT allows to identify particular residues that are under selection.**

- 3) You would like to study the population structure of 500 wild mice from a particular species which occurs all over Europe. You decide to use RAD-Seq. Explain the concept behind RAD-Seq. Will the data allow you to identify genes under selection?

Rad-Seq selectively sequences genomic regions close to specific enzyme restriction sites, which are shared between individuals. The obtained data represents a subset of a genome. The method is cost-efficient, especially when analysing 500 individuals and does not require fully assembled reference genome of the studied species, which allows to perform analysis of species which do not have a sequenced genome yet.

Will the data allow you to identify genes under selection? Yes, because the studied individuals are all part of the same species. It would also work for subspecies or closely related, but the further the evolutionary distance the less informative the data sets will be -> number of shared restriction sites will decrease with evolutionary distance (cannot for example compare human and mouse)

- 4) Discuss the advantages and disadvantages, respectively, of generating data for phylogenetics/population genetics based on a) whole-genome sequencing (WGS) b) exome sequencing and c) RAD-Seq.

Whole-Genome Sequencing

- Allows to analyse the whole genome, including coding, non-coding, and mitochondrial DNA
- Discover novel genomic variants (structural, single nucleotide, insertion-deletion, copy number)
- Costly, especially when sequencing large number of genomes.

Exome Sequencing

- Datasets biased towards exons (2% of the genome) but can also have off target effects.
- Important and cost-effective for big genomes
- Only cost-effective for large genomes because the selection assay is quite costly and does not pay off for small genomes like *Drosophila*.

RAD-Seq:

- Important for studies of large sets of individuals
- Does not need fully assembled reference genome of the studied species – can perform analysis of species which does not have a sequenced genome yet.
- Low cost
- Works only for closely related species. The further the evolutionary distance the less informative the data sets will be as the number of shared restriction sites will decrease.

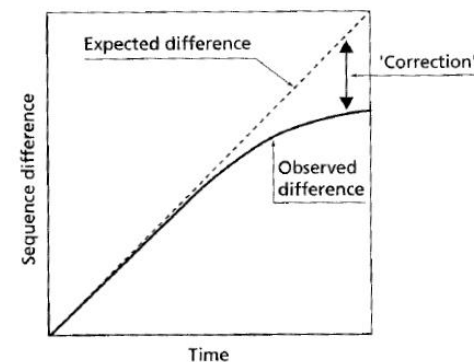
- 5) What is the source behind genome-wide protein data sets used in phylogenetics?

The collaborative contribution of researchers and the advancement of sequencing methods over the years. Currently there are genomic databases (like NCBI), there have been collaborative genomic projects (1000 genomes project), published sequencing results from custom studies.

- 6) Describe the effect of saturation onto the genetic distance observed between two sequences.

Saturation = multiple substitutions occur at the same site in a DNA or protein sequence over evolutionary time, to the extent that the original differences are no longer distinguishable.

The correlation of sequence difference and time is not linear, because despite the accumulation of substitutions, the total number of observed differences between sequences changes only slightly, since (once saturation occurs) additional substitutions at the same site do not contribute to increasing the total difference count.



7) You want to generate a mammalian phylogeny and you have the choice between data from nuclear genes and from mitochondrial genes. Which data sets will be more prone to show saturation?

MtDNA has a higher mutation rate and is therefore more prone to saturation than the nuclear genome.

8) Models for the evolution of nucleotide sequences operate with two key parameters. What are these parameters? Which is the simplest model, which model is the most complex?

- Parameter 1: base frequencies
- Parameter 2: substitution rate.

The simplest model is Jukes-Cantor model. Most complex model is "General reversible".

9) How can you calculate distances between protein sequences?

Calculation of non-identical sites, correction for multiple substitutions

- Jukes-Cantor protein
- Scoredist, based on BLOSUM62 alignment scores, calibrated with evolutionary model

Evolutionary distance is measured based on optimal alignment matrix.

10) Under which condition will mismatches in a sequence alignment directly reflect the genetic distance between the compared sequences?

Assuming that the rate at which substitutions occur is the same for all positions in the sequences and that all lineages evolve at the same rate, mismatches should reflect genetic differences.

11) What is meant by a compensatory change? Give an example.

A compensatory change means that one mutation in a sequence is compensated by another mutation at a different site. The compensatory change typically occurs to preserve the structural or functional integrity of the molecule despite the initial mutation.

RNA: If in a nucleotide pairing C-G a substitution to C-C occurs, a compensatory change would restore the order, by a further substitution resulting in G-C.

12) Four-fold degenerate sites are assumed to evolve neutrally. Explain.

Four-fold degenerate sites are assumed to evolve neutrally because changes at these positions often do not result in amino acid substitutions in the corresponding protein.

13) What is meant by discrete tree building method, what is the key difference that distinguishes them from distance-based methods?

Discrete trees are constructed based on the presence or absence of discrete traits or characters. These trees model the evolutionary changes of discrete traits using substitution models to find evolutionary the most likely development. Discrete methods avoid loss of information that occurs when alignments get converted into distances.

In contrast, distance trees, are constructed based on pairwise dissimilarities calculated from entire sequences or traits and show relationships based on the overall similarity.

14) What is meant by “long branch attraction” and why is it important to take it into account?

It's potential pitfall in phylogenetic tree reconstruction where distantly related lineages with long evolutionary branches are incorrectly grouped together due to a shared, apparent similarity resulting from multiple substitutions over time. For example: one lineage has undergone rapid evolution, accumulating many mutations over time and another lineage, despite being distantly related, may have evolved more slowly and accumulated fewer mutations, resulting in a shorter branch length, but due to the random nature of mutations, some changes may occur convergently in both lineages, leading to similarities and an “apparent” shared ancestry, grouping the long branches together.

Lecture 6

1) Pseudoreplicates are used for bootstrapping. Explain.

Pseudo-replicates are randomly resampled datasets from the original sequencing data. The new datasets are of the same size as the original dataset but with some columns repeated and others left out. They are sequences of same length but different composition.

The pseudo-replicates are then used to calculate phylogenetic trees. The results from the analysis on each pseudo-replicate are then combined to create a distribution of possible outcomes to estimate how well-supported each branch.

2) A well-supported phylogenetic tree may be wrong. Explain why.

In short: due to data properties or analysis strategy, e.g.:

- Sampling error
 - Chosen sequences do not adequately reflect species history.
 - Incomplete lineage sorting, horizontal gene transfer, paralogy/orthology
- Incorrect model of sequence evolution = Chosen model does not fit the data
- Evolutionary history of the samples → Succession of speciation events within short time interval

3) Contemporary sequences can be traced back to an ancestral sequence. Explain why an ancestral sequence is not the consensus sequence calculated from the contemporary sequences.

Because the reconstruction of the ancestral sequence takes more factors into consideration than the consensus sequence. For example, it looks at the evolutionary history that led to the specific diversity and the substitution patterns (syn/nonsyn).

The consensus sequence is based solely on the most frequently observed nucleotide or amino acid at each position in a set of aligned sequences, without explicitly considering their historical context. It is also calculated from a single species or population, while ancestral sequence reconstruction involves a broader phylogenetic perspective, involving multiple species or lineages.

4) The sequences at internal nodes are generally unknown. Are there any exceptions?

Yes. They can sometimes be derived from palaeontological data (e.g. fossils, extracting genetic material from ancient specimen, etc).

5) Explain the concept behind the molecular clock. What is meant by a “stochastic clock”?

Molecular clock makes assumptions about evolution at a molecular level. It suggests a steady accumulation of mutations over time, while the „stochastic clock“ assumes that mutations accumulate randomly but approx. at the same rate in different species.

The so called “strict clock” assumes perfectly constant rate of evolution while the “relaxed” clock uses different evolutionary rates on different branches. However, there are several molecular clock models.

6) Is there a link between generation time and/or metabolic rate of an organism and the observed substitution rate?

Yes, the observed substitution rate in molecular evolution is influenced by various factors, including generation time and metabolic rate. Generally, species with shorter generation times tend to have higher substitution rates. Also, species with higher metabolic rates often exhibit faster rates of molecular evolution (because faster metabolic rate = more cell division = more mutations).

7) Explain when a discrete clock model should be used.

A discrete clock model is used when different branches of the tree are expected to have distinct substitution rates, and these differences are based on biological information. E.g. traits like sexual/asexual reproduction may influence mutation rates or selective pressures, leading to different rates of molecular evolution.

8) Which neutral factors or processes can shape codon usage?

- **Global GC content of the genome** – genomes with higher GC content tend to have a preference for codons that end in G or C.
- **Local variation of GC content** (e.g. isochores in mammals) - Some isochores may have higher GC content than others => variations in the frequency of GC-ending codons within different regions of the genome.
- **Different nucleotide bias of the leading and lagging strands in prokaryotes**
- **Horizontally acquired genes** – may have different codon usage patterns compared to native genes of the organism.

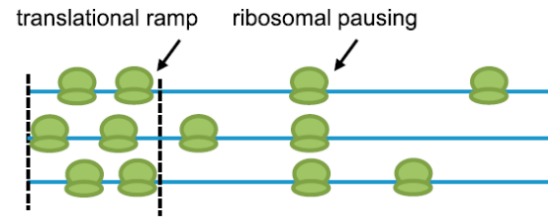
9) What is the strongest predictor of codon usage in mammals.

Local GC content.

Lecture 7

1) How can ribosome profiling tell you whether a particular codon is optimized or suboptimal?

Ribosome profiling measures how many ribosomes along mRNA are occupied. High ribosome density suggests optimized codons, while pauses or fluctuations indicate potential suboptimal codon usage.



2) Explain the connection between codon bias and composition of a cell's tRNA pool.

Codon bias is linked to the composition of the tRNA pool as optimal codons are reflected in the abundance of corresponding tRNA molecules. Cells favour codons for which there are high concentrations of matching tRNAs, optimizing translation efficiency. The genomic repertoire, copy number, wobble base pairing, and post-transcriptional modifications of tRNA collectively influence the connection between codon bias and the cellular tRNA pool.

3) What are optimized codons? Which microbial genes generally contain optimized codons?

Optimised codons are codons that are preferentially used in a particular organism's genome during protein synthesis. These codons correspond to tRNA molecules that are abundant in the cellular tRNA pool, leading to more efficient and accurate translation.

In microbes, **genes of highly expressed and essential proteins** often have optimised codons, increasing translation efficiency and rapid protein synthesis.

4) Are rare codons randomly distributed within an open reading frame?

No, their distribution is not random; instead, they are often found in regions where they are necessary.

5) Explain whether a link exists between codon bias and transcript splicing.

Codon bias influences transcript splicing **through mechanisms such as exonic splicing enhancers (ESEs) and transcription factor binding**. For example, somatic mutations at ESEs, observed in cancer, contribute to the generation of cancer-specific gene isoforms, indicating a functional connection between codon choices, splicing regulation, and disease outcomes. Also, **substitution of optimal codons with unpreferred codons can reduce mRNA half-life**.

6) Synonymous substitutions can be subject to purifying selection. Why?

Because certain synonymous codons may have functional roles beyond encoding amino acid, e.g. building secondary structures or due to their role in splicing regulation and purifying selection acts to preserve the functional aspects associated with specific codon choices.

7) What is meant by domestication and by diversification?

Domestication is the process by which a wild species, typically plants or animals, are selectively bred and controlled by humans for various purposes. This process involves manipulating the genetic, behavioural, and morphological traits of the species to make them more suitable for human needs,

such as agriculture, companionship, or labour. Genomes of domesticated species differ from wild progenitors in specific ways.

Diversification is an evolutionary process where population is splitting into a distinct species. The group diversifies acquiring new characteristics. Diversification can occur through various mechanisms, for example by adapting to different ecological niches, or through gradual accumulation of genetic differences over time.

8) What is referred to as the “domestication syndrome”?

It's a set of common characteristics or traits that are often observed in domesticated plants/animals but not in their wild ancestors. In plants that would include:

- Less, but larger, fruits or grains
- More robust plants, more determinate growth, less branching
- Loss of seed dispersal
- Decrease of bitter tasting substances in edible structures
- Synchronized flowering, germination.

9) Which strategies exist to identify domestication genes?

Selective sweep analysis -> detecting regions of reduced genetic diversity around a favoured allele that has undergone positive selection. Such regions may contain genes critical for domestication.

GWAS (Genome-wide association study) -> examines the entire genome for associations between genetic markers and traits of interest.

10) What is a selective sweep?

Selective Sweep: The frequency of a new allele increases rapidly due to positive selection. This leads to a decrease in genetic variation, because other alleles get removed from the gene pool.

In the context of domestication genes, where strong artificial selection is applied, selective sweeps often result in the complete fixation of the selected allele in the population, especially for traits crucial to domestication. However, fixation can be prevented in certain cases, such as when interference with fertility mechanisms occurs.

10) It has been observed that the genomes of cultivars harbour more non-synonymous mutations than expected. How can this finding be explained?

It means that strong artificial selection during domestication led to the hitchhiking effect where non-synonymous mutations co-occur with beneficial mutations.

11) Explain what is meant by a “domestication bottleneck”?

It's the reduction in the genetic diversity of a population that occurs during the process of domesticating plants or animals. The bottleneck occurs especially when a relatively small number of individuals is chosen for breeding.