Lecture 1 Pairwise Alignment (1)

1.1 Sequence analysis exploits patterns arising during evolution

First, we would like to talk about evolution:

So, what is evolution? We give a brief introduction which copied from chat-gpt: Evolution is the process by which species of living organisms change over time through the gradual accumulation of small genetic variations that are inherited from one generation to the next. Evolution is a conservative process: the vast majority of mutations will not be selected (i.e. will not make it as they lead to worse performance or are even lethal), if such a evolution happens, we say it *negative* (or *purifying*) selection.

There are four r**equirements for evolution: (Know this)**

- **Template**: This refers to the structure provides stability, for example, DNA. This is achieved through the transmission of genetic information from parent to offspring.
- **Copying Mechanism (Meiosis)**: This process is crucial for maintaining the correct chromosome number in sexually reproducing organisms and ensuring genetic diversity.
- **Variation**: This refers to the existence of differences or variations in traits within a population. These variations are often the result of genetic diversity.
- **Selection** (Survival of the Fittest): Some traits lead to greater fitness of one individual relative to another.

Actually, the selection often refers to natural selection. The natural selection often acts on the phenotype, which is the characteristic of the organism that actually interact with the environment. We also have 'genotype' which is the genetic basis of any phenotype. If certain genotype gives phenotype a reproductive advantage, then it would become more common in a population. Throughout the lives of individuals, their genomes interact with the environment, causing variations in traits. Sometimes, intracellular processes (e.g. copying errors) may also lead to DNA iterations. We should emphasis that the environment of a genome includes the molecular environment in the cell, other cells, other individuals, populations, species, and the abiotic environment.

For any evolution, we could specifies it as two parts: Convergent evolution and divergent evolution.

- ***Convergent evolution***: It refers to a independent evolution of similar characteristics in organisms that do not have to be closely related. For example, wings in the different animals that fly, ranging from insects to bats. Structures resulting from convergent evolution are termed **analogous**. The terminology analogous we will talk about later.
- ***Divergent evolution***: It refers to the process by which a single ancestor or ancestral gene is modified over time into two or more descendants that have an increasing degree of dissimilarity as the time since they diverged increases.

1.2 Environment: the Exposome

Paul Wild published a paper in 2005, he described three overlapping components of the exposome of humans. But first, what is 'exposome'? Exposome is a concept that encompasses the totality of environmental exposures throughout an individual's life, inlcuding exposures to physical, chemical, biological and psychosocial factors. These exposures can influence an individual's health and contribute to the development of diseases. Back to our topic, the three overlapping components are:

1. A general external environment including the urban environment, education, climate factors, social capital, population.
2. A specific external environment with specific contaminants, radiation, infections, lifestyle factors (e.g. tobacco, alcohol), diet, physical activity, etc.
3. An internal environment to include internal biological factors such as metabolic factors, hormones, gut and oral microbiota, inflammation, oxidative stress.

1.3 Homology/Orthology/Paralogy

After talking about the evolution and environment, let us talk about homology, orthology, and paralogy. These three topics would help us trace the evolutionary history of genes and understand how they have diversified over time. We first give the definition and some example, then we use a graph to illustrate it.

- ***Homology***: Homology is common ancestry; i.e. homologous genes have a common ancestor. There are two forms of homology: Orthology and paralogy.
- ***Orthology***: Orthologous gene are genes in different species that evolved from a common ancestral gene through speciation. In other words, orthologs are genes that retained the same function in different species due to their divergence during evolution.

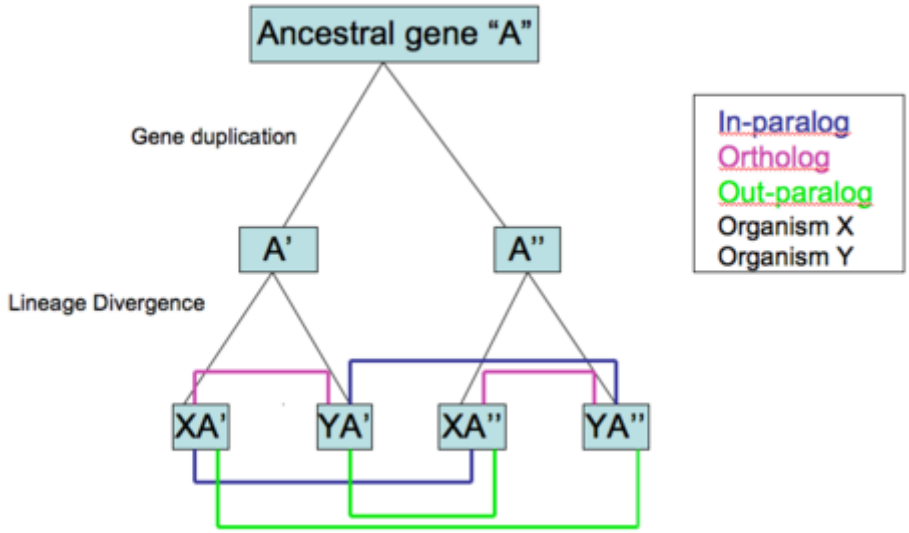
If gene A in species X and gene B in species Y are orthologs, it means that gene A and gene B share a common ancestral gene that existed before the species X and Y diverged.

- ***Paralogy***: Paralogous genes are genes within the same species that arise from gene duplication events. After a duplication event (refers to the process where a segment of DNA is copied or duplicated, resulting in two or more copies of a particular gene or genomic region), the two (or more) copies of the gene can evolve independently, potentially acquiring new functions while retaining some similarities.
    - ***In-paralog***: In-paralogs are genes that result from a duplication event and are present within the same species or genome.

      If gene A is duplicated, and the two resulting genes, A1 and A2, exist in the same genome, they are considered in-paralogs.
    - **Out-paralog**: Out-paralogs are genes that result from a duplication event in one species and are then compared to genes in another species that did not undergo the same duplication event.

      If gene B is duplicated in species X, resulting in genes B1 and B2, and these are compared to a related gene C in species Y that did not undergo a similar duplication event, then B1 and B2 are considered out-paralogs to C.

If gene C and gene D are paralogs within the genome of a single species, it means that they originated from a gene duplication event. Gene C and gene D may have similar functions, but they are not necessarily identical. This graph is vitally important.



Look at the graph, we have two organism X and Y. First, we notice that there are XA' and XA'' should be the same species, so they are paralogs. When we notice that the XA' and YA', they do have the same letter A' but in different species, so they are orthologs. We could notice that the gene A is duplicated to A' and A'', and they are both in X - the same species, so XA' and XA'' are in-paralogs, so do the YA' and YA''. However, when we focus on out-paralogs, we know that the XA' is a different species of YA'', so they are out-paralogs. We have to focus on two things: different species (X and Y) and different gene types (A' and A'').

1.4 Xenology

First we introduce a new terminology: horizontal gene transfer (HGT), also called Lateral gene transfer (LGF), is any process in which an organism incorporates genetic material from another organism without being the offspring of that organism. By contrast, vertical transfer occurs when an organism receives genetic material from its ancestor, e.g. its parent or a species from which it evolved. To be clear, it means the HGT happens from a organism to another organism without being the offspring. The horizontal gene transfer is amongst single-celled organisms a significant form of genetic transfer.

1.5 Changing molecular sequences

The molecular sequences changes would be caused by mutations. Here we introduce two types of point mutation:

- **Synonymous mutation**: mutation that does not lead to an amino acid change in the protein product.
- **Non-synonymous mutation**: mutation does lead to an amino acid change.
    - **Missense mutation**: one amino acid replaced by other amino acid
    - **Nonsense mutation**: amino acid replaced by stop codon

Similarly, there is a term sequence mutation. There are many different evolutionary ways to alter a gene: we do not list them all. We first focus on DNA expression mutation. There are many types of mutations that do not change the protein itself but where and how much of a protein is made. These types of changes in DNA can result in proteins being made at the wrong time or in the wrong cell type. Changes can also occur that result in too much or too little of the protein being made.

1.6 Paralogy and repeating sequences

There is large than 50% of the human genome consists of repeats for DNA, and many proteins consist of sometimes numerous repeats, however, the repeats sometimes lead to gain function but sometimes lead to disease (e.g. single-residue repeats). There are a lot of types of genome repeats, we do not illustrate here.

1.7 Transposons

Transposons are sequences of DNA that can move around to different positions within the genome of a single cell, a process called ***transposition***. Transposons can cause mutations and change the amount of DNA in the genome. It was discovered in the 1940s. There are two classes of transposons:

- **Class I** mobile genetic elements, move in the genome by being transcribed to RNA and then back to DNA by reverse transcriptase.
- **Class II** mobile genetic elements move directly from one position to another within the genome using a **transposase** to 'cut and paste' them within the genome

1.8 Protein Sequence-Structure-Function

We have a sequence, and then we could use it to predict the structure, of course we could use structure to predict the function. Sequence encodes Structure encodes Function. So based on evolution, we can relate biological macromolecules and then 'borrow' annotation of 'neighbouring' proteins or DNA in the databases (DBs). This works for sequence as well as structural information. Now we could stress the problem we will discuss in this course: how do we score the evolutionary relationships. i.e. we need to develop a measure to decide which molecules are probably neighbours and which are not.

Sequence - Structure/function gap: there are far more sequences than solved tertiary structures and functional annotations. This gap is growing so there is an ongoing need to predict structure and function.

1.9 Alignment

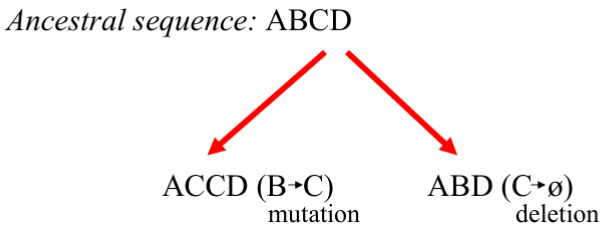
Which alignment is better? We use common sense and call it: simplest, most probable and maximum likely. How do we compare sequences? We use similarity score. Many properties can be used to create such a score: nucleotide or amino acid composition, isoelectric point, molecular weight and morpholohical characters. But!: Molecular evolution understood through sequence alignment.

A main idea is to search for similarities of sequences, because common ancestry is an important notion, which makes it more likely that genes share the same function. As we said before, sharing a common ancestor is homology. When an unknown gene X is homologous to a known gene G, it means that we gain a lot of information on X, what we know about G can be transferred to X as a

good suggestion.

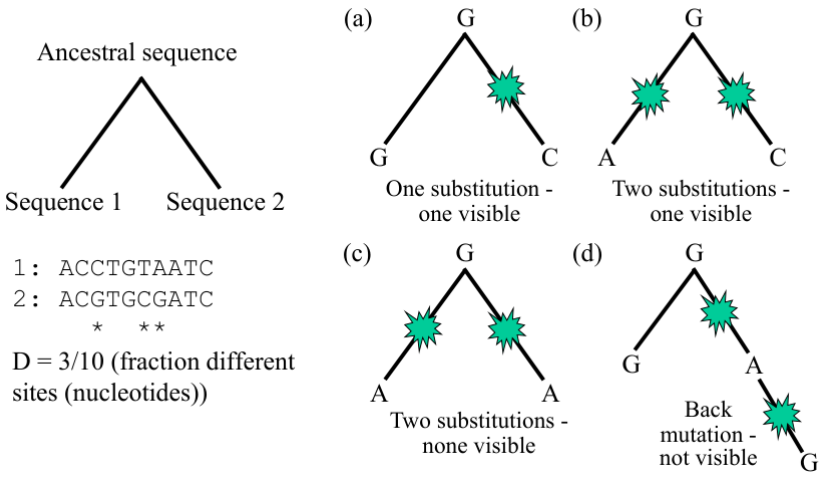
A question is what can and what cannot be aligned: sequences should be related through divergent evolution, that is, have common ancestry. They should be homologous and preferrably orthologous, paralogous sequences can become too distant for correct alignment! Analogous sequences (i.e. convergent evolution) should not be aligned! Sometimes a short functional motif can be detected.

1.10 Divergent evolution

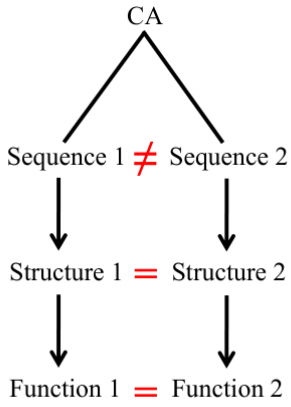


ACCD AB—D	or	ACCD A—BD	*Pairwise Alignment*

Which one is true alignment? We would say the first one. Why? Before we answer this question, we should first know what is 'reconstructing evolution'.



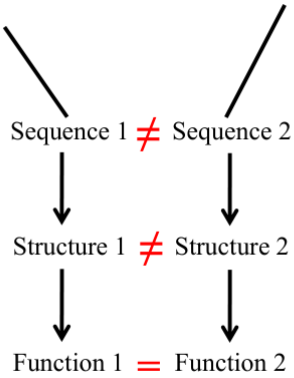
As it shows, the ancestral sequence has two offspring, sequence 1 and sequence 2, the similarity of the two sequence is 0.3. At the right side, we know G to C is a substitution and visible, perhaps the mutation leads to changes of proteins, however, G to A is not visible. So, we could see some none visible substitutions and some back mutations. That is, sometimes, even the two sequences are not identical, the proteins may be the same so that have the same function. Therefore, here we could know that protein structure typically remains the same, and one more thing, **function normally is preserved within orthologous families**. "***Structure is more conserved than sequence***".



Here, we would like to talk a word of caution on divergent evolution. Homology is a term used in molecular evolution that refers to common ancestry. Two homologous sequences are defined to have a common ancestor, just like the computer science, we use a Boolean term: two sequences are homologous then we use 1 to represent and if not, we use 0. As for relative scales, i.e., sequence A and B are more homologous than A and C, are nonsensical. For scalars, you can talk about sequence similarity, or the probability of homology.

1.11 Convergent evolution

A motif is a short, recurring sequence of nucleotides in a DNA molecule or amino acids in a protein. The convergent evolution often involves shorter motifs, e.g. active sites of proteins. The convergent evolution leads to analogous structures, motif (or function) has evolved more than once independently, for example, starting with two very different sequences adopting different protein folds. More than that, sequences and associated structures remain different, but functional motif can become near identical, that is, same functions with different sequences or structures. If original gene is displaced with convergently evolved gene, this is called **non-orthologous displacement**. Convergent evolution is a lot less common than divergent evolution, a classical convergent example is serine proteinase and chymotrypsin, these two enzyme families have similar functions and structures, despite not sharing a recent common ancestor that possessed those traits.



For now, we could say, what is convergent evolution and how can we judge it? First, there is no common ancestry and protein sequence and structures are very different, second, functional motif can arise leading to similar function, lastly, if an analogous protein resulting from divergent evolution takes over function in cell, this is called non-orthologous displacement.