Introduction to bioinformatics

(Database searching, Sequence alignment, and alignment affecting factors)

Course Code –BOTY 4204

Course Title- Techniques in plant sciences , biostatistics and bioinformatics

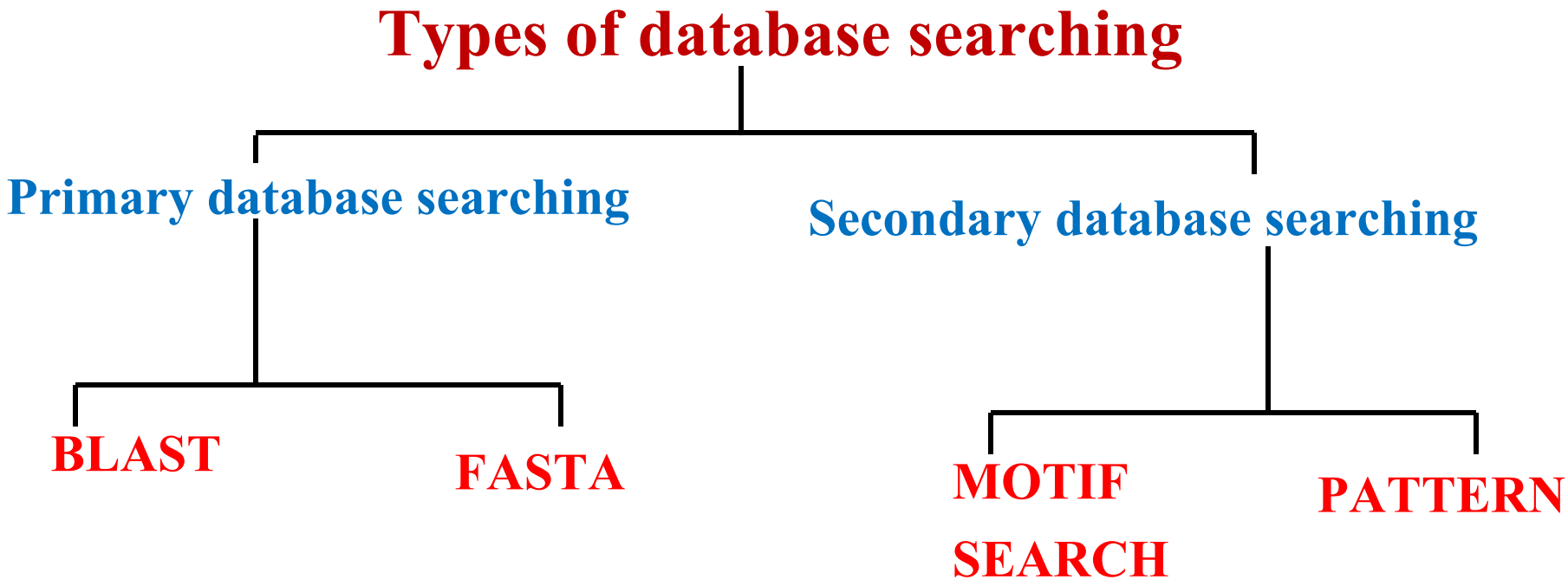
By – Dr. Alok Kumar Shrivastava

Department of Botany

Mahatma Gandhi Central University, Motihari

Database searching ?

- Database searching is the matching of query nucleotide or protein sequences with database sequences. To do this , we align the query sequence with database sequences to find similarity among them. Database searching is the application of knowledge achieved from previous biological experiments to the gene discovery problem.
- A DNA sequence comprises only 4 nucleotides, while a protein sequence is made up of twenty amino acids. Hence it is easier to search for similar patterns in proteins than in DNA. To perform a database search , it is better to translate the DNA sequence for encoding proteins into protein sequence for a more reliable result.


Types of database searching
Primary database searching
Secondary database searching
BLAST
FASTA
MOTIF SEARCH
PATTERN

Primary database searching

- A primary sequence is one that has been experimentally determined, and a primary database is one that contain experimentally derived data.
- Searching a database in an efficient manner is a matter of prime importance.
- Methods that can run on small databases may not be effective with larger databases in terms of time and space efficiency.
- Due to large amount of information in database it is difficult to perform a database search using dynamic programming as it is a computationally intensive programme and is therefore too slow.
- To save time and space heuristic methods (BLAST and FASTA) are used for database search.

- These programs are local similarity search programs that provide short matches.
- These short matches called local alignment (a segment of sequence showing similarity) are more informative than global alignment (complete sequence showing similarity).
- Local alignments can return highly conserved region of the sequence even in two sequences that do not produce any reasonable alignment when aligned globally.
- Dynamic algorithm programs can return local alignments and are guaranteed to find the best alignment, but are very insensitive due to their mathematical rigour.
- BLAST and FASTA are major primary database search method and are very fast though less sensitive used to carry out database searches.

FASTA

- ✓ FASTA developed by Pearson and Lipman in 1905, and was so named because it was faster than other methods used for sequence alignment at that time.
- ✓ FASTA uses the Pearson and Lipman algorithm for similarity search between a query sequence and a database sequence.
- ✓ Given a query sequence , FASTA searches for local alignment with the sequences in the database.
- ✓ Originally, the FASTAP program was designed for protein sequence similarity.
- ✓ It is a rapid alignment program for protein and DNA sequence pairs.
- ✓ No individual residue search is performed, saving time.
- ✓ Input sequence must be in FASTA format for alignment.

Basic FASTA programs

Program name	Query sequence	Database sequence	Algorithm used
Nucleotide BLAST	NUCLEOTIDE	NUCLEOTIDE	DNA/RNA FASTA, FASTM, FASTS
Protein BLAST	PROTEIN	PROTEIN	FASTA, SSEARCH, FASTS, FASTF
FASTX/FASTY	TRANSLATED NUCLEOTIDE	PROTEIN	
TFASTX/TFASTY	PROTEIN	TRRANSLATED DNA	
TFASTs	PEPTIDES	TRANSLATED DNA	

BLAST (Basic local alignment search tool)

✓A local similarity search program, BLAST compares nucleotides or protein sequences to sequence databases and calculate the statistical significance of the matches. The functional and evolutionary relationship between sequences are construed and members of gene families identified by the BLAST search program.

✓It is a simplification of the Smith-waterman algorithm and It is faster than FASTA.

✓This is the algorithm that is most commonly used for database search and sequence alignment. It looks for similar regions in two sequences without allowing a gap, though now there is gapped BLAST (WU-BLAST).

✓It is more selective and less sensitive.

✓It does not allow gaps in the alignment.

✓FASTA is more sensitive than BLAST for nucleotide sequences

✓BLAST(word size 3) is more sensitive for protein sequence as compared to FASTA(Word size 2)

PSI –BLAST(Position specific iterative- BLAST)

- ✓ In PSIBLAST the query sequence is first queried against a protein database using the BLASTP program.
- ✓ This program searches for a certain conserved region by working interactively.
- ✓ In the first round PSIBLAST works like a normal BLAST program producing hits.
- ✓ Significant hits are collected and an MSA(multiple sequence alignment) table is constructed between the query sequence and significant conserved matches.
- ✓ This MSA provides a profile. The database is once again queried using this profile and local matches are investigated and statistically significant matches are taken.
- ✓ The steps are repeated again (from step two) until no significant change occurs in the final alignment.
- ✓ Some types of low complexity sequences may not be detected by the filtering option in BLAST.
- ✓ This program provides better results to identify distant relatives.

Basic BLAST programs

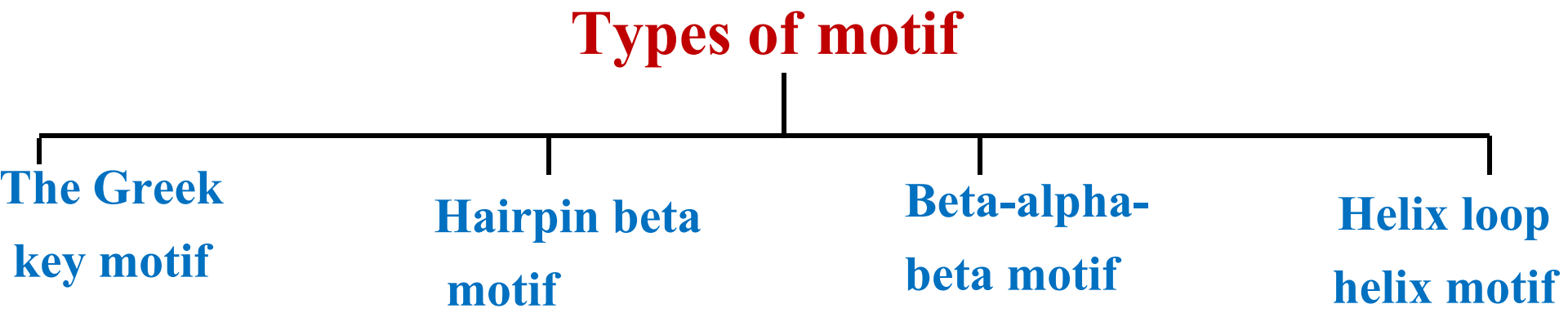
Program name	Query sequence	Database sequence	Algorithm used
Nucleotide BLAST	NUCLEOTIDE	NUCLEOTIDE	BLASTN, MEGABLAST
Protein BLAST	PROTEIN	PROTEIN	BLASTP, PHIBLAST, PSI BLAST
BLASTX	TRANSLATED NUCLEOTIDE	PROTEIN	BLASTP
TBLASTN	PROTEIN	TRRANSLATED NUCLEOTIDE	BLASTP
TBLASTX	TRANSLATED NUCLEOTIDE	TRANSLATED NUCLEOTIDE	BLASTP

Secondary database searching

- ✓ Primary database searching does not always provide a satisfactory answer to the questions of sequence analysis.
- ✓ The presence of highly repetitive and low complexity sequences can result in irrelevant matches and may even complete the interpretation.
- ✓ Secondary databases provides information about the relationship of a given sequence with other sequences within multiple alignment and some more information (family, domain and motif) as well, depending on the method used.
- ✓ These databases contain the results of primary sequence analysis.
- ✓ Some important secondary database searches are motif or pattern search and profile search.
- ✓ PROSITE is a database and a tool consisting of documentation entries describing protein domain, families and functional sites as well as associated pattern and profiles to identify them.

Motif search

- Motifs are specific geometric arrangement of protein secondary structure elements (alpha, beta and loops). Some motif are associated with a particular function and some are part of other structural and functional arrangements. Simple motifs are combined to form complex motifs. These are biologically conserved regions from protein sequences.


Types of motif
The Greek key motif
Hairpin beta motif
Beta-alpha-beta motif
Helix loop helix motif

Simple motif

- In the greek key motif , four adjacent antiparallel beta sheets are arranged in a greek style ornamental pattern.
- Hairpin beta motif is built from two adjacent beta strands, connected by a loop, and occurs repeatedly in protein structure.
- Beta-alpha-beta motif is a commonly used method to connect two beta sheets. It comprises an alpha sheet connected with a loop to a beta helix and again a loop connecting the next beta sheet.
- Helix loop helix motif is specific to a DNA binding region and comprises two helixes connected with a loop.

Pattern search

- The pattern from a conserved residue of a multiple sequence alignment, if used in the right way, can give an enormous amount of information about a protein sequence. Various pattern search methods include regular expressions, rules and block search.
- Regular expression approach is based on pattern recognition, and a single conserved motif found in the multiple sequence alignment gives details about a family. Regular expression gives information about a sequence based only on the conserved region or consensus. To obtain more authentic information, regular expression uses only identical matches and does not allow similarity. The regular expression should not be very short, as very short matches can not give realistic prediction.

Parameters

- ✓ Score value(S) , also called bit score, is derived from the raw alignment score ‘S’. It is numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. Here the statistical properties of the scoring system used have been taken into account to produce ‘S’.
- ✓ Expect value (E) is a parameter which depicts the expected number of hits that happen just by chance when searching a database of a particular size. Thus a match is more significant if E Value is lower or closer to zero. The length of the query sequence is also taken into account for the E value, as shorter sequences have a high probability of occurring in the database purely by chance.

Continue……

✓K value is a statistical parameter that is used to calculate BLAST score. This value is used to convert a raw score (S) to a bit score (S').

✓Probability value (P) is the probability of obtaining a result by chance alone. The P-values closer to zero are highly significant. Like the E value, it is a different way of representing the significance of the alignment.

✓Selectivity is the ability of a method to find the members of a protein family without any false positive matches. or in other words it is a measure of your ability to not include the false matches erroneously.

✓Sensitivity is the ability of a method to find the most similar sequences from the database or a measure of your ability to find all the true matches.

Factor affecting database searching

- Scoring matrices- Scoring matrices are used to determine the relative score made by matching two characters in a sequence alignment.
- Gap- In bioinformatics, gaps are used to account for genetic mutations occurring from insertion or deletions in the sequence, sometime referred to as indels. In sequence alignments, gaps is represented by dashes on a protein/ DNA sequence alignment. Too many gaps can cause an alignment to become meaningless.
- Gap extension penalties- Gap penalties are used to adjust alignment scores based on the number and length of gaps.

- Type of gaps- Constant gap, Affine gap, Convex gap and Arbitrary gap.

- Filtering low complexity regions - Low complexity regions (LCR) is repeat of the same amino acid pattern in protein sequences these sequences produce high scoring alignment artificially and hinder sequence alignment. This region is usually filtered while performing database searches as it hinders the database result.

Similarity and Identity

- These are terms that illustrate the relationship between two proteins with one another.
- The residue position at which both sequences being compared have the same type of residue is called identical residue.
- The residue positions at which both sequences being compared have amino acids with similar properties are called similarity residues.
- Similarity is the likeness (resemblance) between two sequences in comparison while identity is the number of characters that match exactly between two different sequences.

For example

A F N T T (Seq1)

: | | :

L N N T S (Seq2)

AL and TS are similar residues. Similar residues are represented with a colon (:). The residues N and T are identical residues in the given example which represented by solid line (|) .

Sequence alignment

- In bioinformatics , a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

Pairwise sequence alignment

- This alignment used to identify regions of similarity that may indicate functional, structural or evolutionary relationships between two biological sequences.
- EMBOSS, LAGAN, Bl2seq, Dotlet and Dotter are the common tools for pairwise sequence alignment.
- It is of two types ; local alignment and global alignment.

Local alignment

- If the two given sequences are not so similar and it is difficult to align the two sequences across the full length, then local alignment can be used to align the sequences.
- Local alignment provides information about conserved regions or domains. From these conserved regions it is possible to get an idea of the evolutionary history.
- Local alignment is more meaningful than global alignment as it can achieve some alignment even with sequences that are not so similar. It can also be used to align sequences of unequal length or when only a conserved domain is found in two sequences.

Global alignment

- Global alignment is done across the entire length of the sequence, including matches characters, gaps and mismatches.
- Choosing different mismatch and gap penalties may produce different alignments for the same sequences.

Multiple sequence alignment

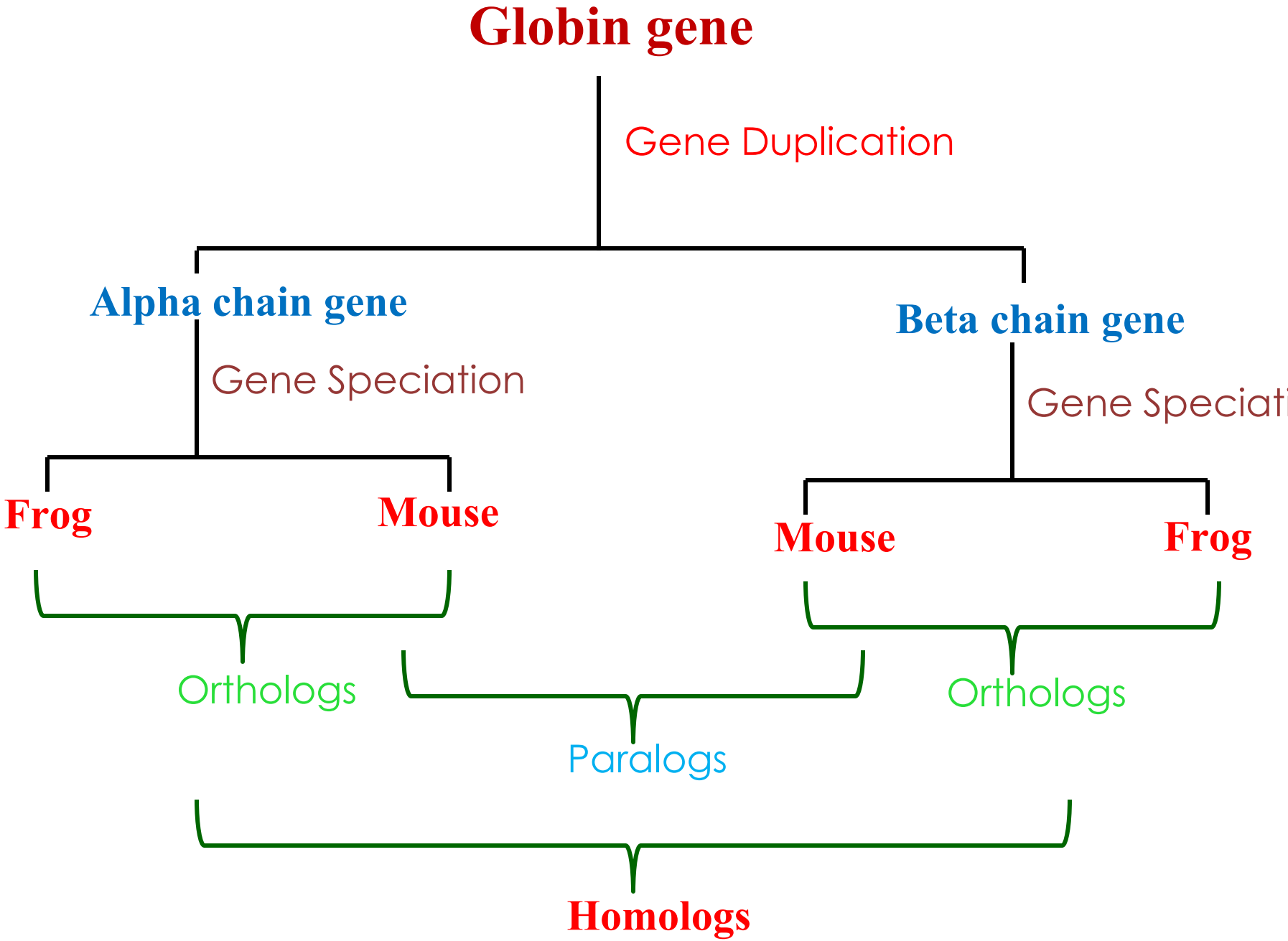
- For multiple sequence alignment more than two sequences are required.
- A database search usually reveals many homologous sequences. The residues of the homologous sequences are aligned together I column for multiple sequence alignment.
- While aligning, wherever a sequence does not possess an amino acid in a particular position, it is denoted by a dash.
- Highly identical sequences are used to give some meaningful results. These multiple sequence alignment can be used to establish phylogenetic relationship.
- ClustalW, T-Coffee, Multalin, DCA, HMMER, DIALIGN are tools for Multiple sequence alignment.

Homologous gene

- Homologous gene is a gene inherited in two species by a common ancestor. While homologous gene can be similar in sequence.

Orthologous and paralogous gene sequences

- ✓ Both orthologs and paralogs are types of homologs.
- ✓ Orthologs are homologous genes where a gene diverges after a speciation event, but the gene and its main function are conserved.
- ✓ If a gene is duplicated in a species, the resulting duplicated genes are paralogs of each other, even though over time they might become different in sequence composition and function.


Globin gene
Gene Duplication
Alpha chain gene
Beta chain gene
Gene Speciation
Gene Speciati
Frog
Mouse
Mouse
Frog
Orthologs
Orthologs
Paralogs
Homologs

Thank you