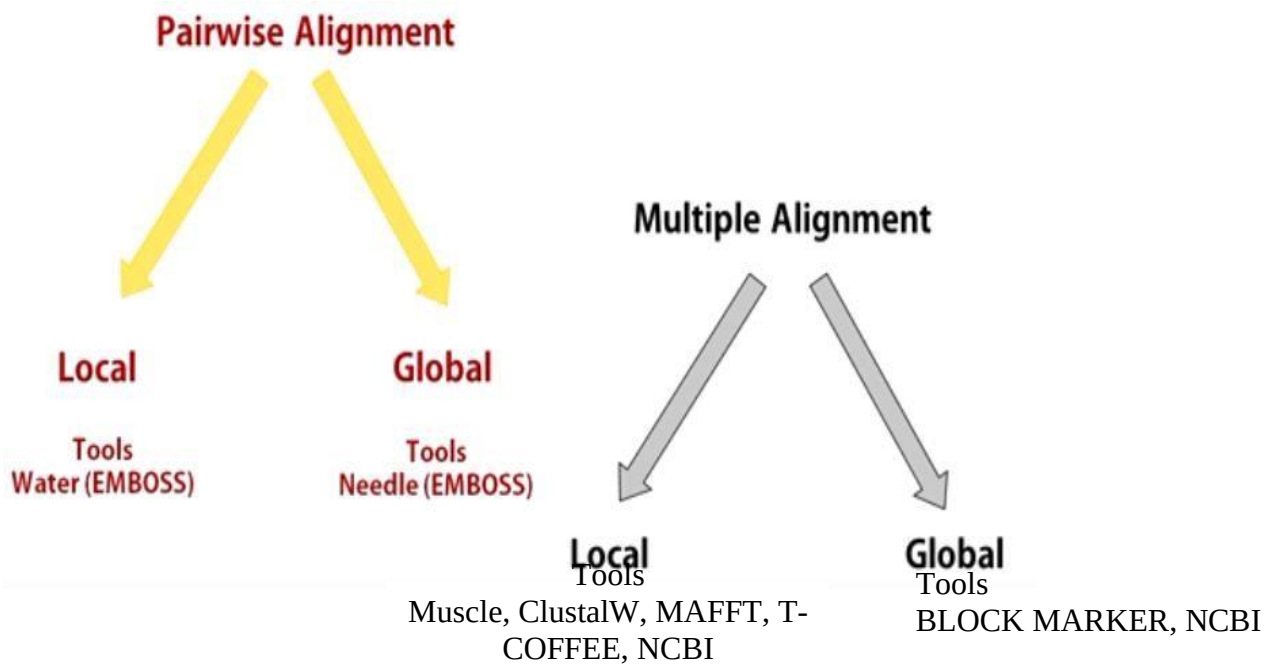
- Multiple sequence alignment (MSA):

1. It is basically an alignment of more than two sequences.
2. A pairwise alignment tell us about the similarity of two sequence, while MSA tell us about similarity among multiple sequence.



- Goals of multiple sequence alignment

❖ **Evolutionary analysis**

1. Identify homology
2. Build phylogenies
3. Test evolutionary models

❖ **Functional analysis**

1. Identify conserved residues
2. Identify protein family

❖ **Structural analysis**

1. Identify sequence co-variation
2. Homology modeling

❖ **Practical application**

1. Identify conserved primer binding site
2. Design of mutagenesis experiments
3. Mutant analysis

- Multiple sequence alignment: evolutionary history

1. In the first figure below, you will see four top protein sequences which are more related and the sequence down (four green sequences) also related sequence.
2. All these sequences have evolutionary history when you look at them as MSA columns (MSA columns = homology).

Figure 1: Evolutionary history of related sequences

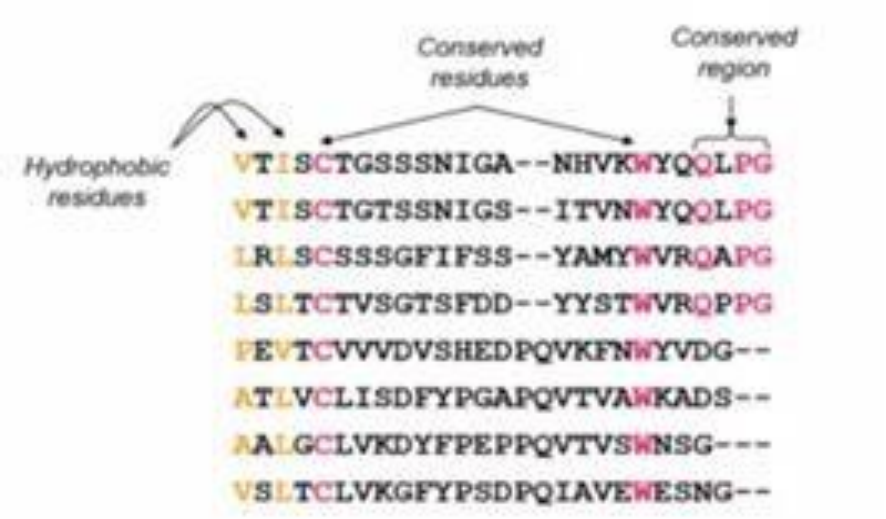


Figure 2: Multiple sequence comparison of related protein sequences

- **Multiple sequence alignment: structure/ function**

1. The second figure shows the multiple comparisons of the sequences in the first figure. Also, MSA columns = homology.
2. Identification of homologous residues greatly facilitated by multiple comparisons.
3. Homologous selectivity maintained due to structural and functional constraints.
4. MSAs identify conserved and structural equivalent residues and regions.

- **Conserved sequences**

1. Conserved sequences refer to identical or similar sequences of DNA or RNA or amino acids (proteins) that occur in different or same species over generations.
2. These sequences show very minimal changes in their composition or sometimes no changes at all over generations.

- **Common examples of conserved sequences include,**

1. Translation and transcription related sequences which are found conserved in the genome at multiple places
2. Certain RNA components in ribosomes are found to be highly conserved over various species
3. tmRNA is found to be conserved in multiple bacteria species
4. Other examples like TATA (repetitive regions) and homeoboxes (involved in regulating embryonic development in a wide range of species).

- **Types of Conserved Sequences**

1. Orthologous (conserved residue) identical sequences are found across species.
2. Paralogous (conserved region) identical sequences are found within the same genome over generations.

- **Importance of Conserved Sequences**

❖ **Biological importance**

1. Conserved sequences found in different genomes can be either coding sequences or non-coding sequences.
2. As coding sequences, amino acids and nucleic acids are often conserved to retain the structure and function of a certain protein.
3. These sequences undergo minimal changes.

4. When changes happen, they usually replace an amino acid or nucleic acid with one which is biochemically similar.
5. Similarly, other mRNA related nucleic acid sequences are often conserved.
6. Non coding sequences, like ribosomes sites, transcriptional factors, binding site, etc., are also conserved sequences.

❖ **Computational importance**

1. Conserved sequences help us find homology (similarity) among different organisms and species.
2. Phylogenetic relationships and trees could be developed and effective ancestry could be found using the data on conserved sequences.
3. A common example is the conserved sequence "16S RNA" which is used to reconstruct phylogenetic relationship among various bacterial phyla.
4. Conserved sequence can also be used to mark the origination of genetic disorders and mutations.
5. By comparing genomes which have a certain conserved sequence common to them we can easily identify anomalies, any exist.

- Hydrophobic residues

1. Hydrophobic Residues such as phenylalanine, isoleucine, leucine, methionine and valine play an important role in protein structure and activity.
2. hydrophobic amino acids pack in the interior of proteins, away from the aqueous environment.
3. The hydrophobic effect is considered to be the major driving force for the folding of globular proteins.

- Methods of multiple sequence alignment

❖ **Dynamic program method**

1. This is an extension of dynamic programming approach for pairwise alignment to multiple sequence.
2. Unfortunately, the cost of computation grows up exponentially with the number of sequence and the sequence lengths.
3. For this reason, this approach is not effective in multiple sequence alignment.

❖ **Progressive multiple sequence alignment**

1. Any two sequence can be aligned accurately and rapidly via dynamic program.
2. Once alignment is made, equivalent to any other sequence.
3. Aligned set of sequences = profile

4. Dynamic program accurately aligns pairs of profiles.
5. Progressive align more distantly related profiles and sequences.

❖ **Iterative multiple sequence alignment**

1. The major problem in progressive method is the propagation errors in the initial alignment throughout the MSA.
2. Iterative method solves this problem by repeatedly aligning subgroups and then realigning these subgroups into the global alignment.
3. Selection of groups can be based upon:
   a) Order of sequence of phylogenetic tree.
   b) Separation of the sequence from the rest.
   c) Random sampling.

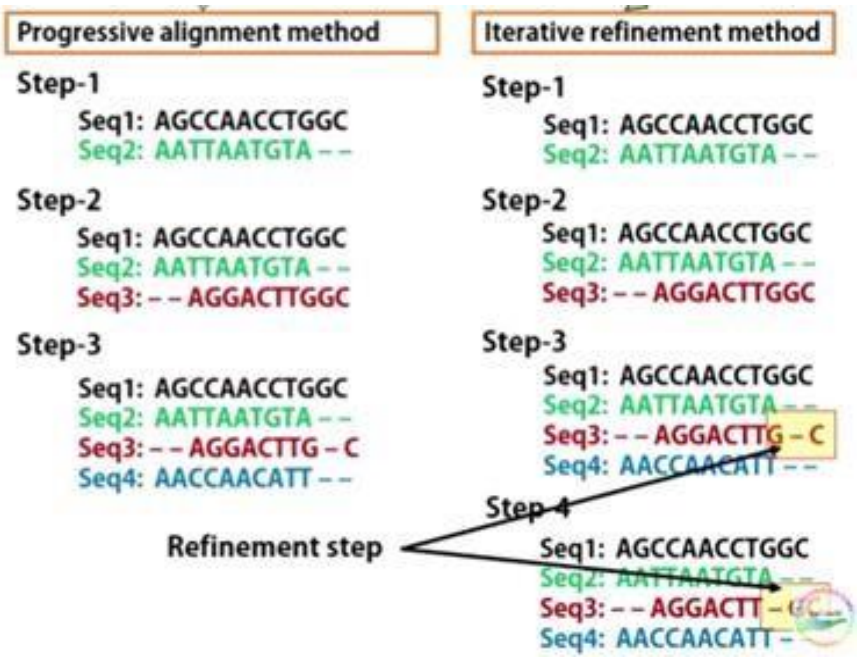


Figure 3: progressive and iterative alignment method