**Similarity and homology**

These terms are often confused. Sequences are **homologous** only if they evolved by divergence from common ancestor. Sequence similarity searches are used very commonly to predict gene or protein function, the underlying theory is that similar sequences are likely to be homologous and therefore to have similar functions.

However, the evolutionary mechanisms of gene duplication allows organisms to acquire reductant copies of genes. These reductant copies are then free to evolve new functions, and become homologous gene with different functions.

**Orthologs-** when two homologous genes in different species have the same function, they are known as **orthologs.**

**Paralogs-** when two genes in the same or different species have different functions they are known as **paralogs.**

**Scoring matrices**

Scoring system is a set of values for qualifying the set of one residue being substituted by another in an alignment.

* It is also known as substitution matrix.
* Scoring matrix of nucleotide is relatively simple.
* A positive value or a high score is given for a match & negative value or a low score is given for a mismatch.
* Scoring matrices for amino acids are more complicated because scoring has to reflect the physicochemical properties of amino acid residue.

**Scoring a sequence alignment.**

Match score: +1

Mismatch score: +0

Gap penalty: –1



Matches: 18 × (+1)

Mismatches: 2 × 0

 Gaps: 7 × (– 1)

 Score = +11

**Using scoring matrices**

Scoring matrices are useful when you want to compare two sequences. Scoring matrices used for comparing alignment scores are based on observation substitution rates derived from the substitution frequencies. Every possible identity and substitution is assigned a score based on the observed frequency of occurrences in alignments of related proteins. The score is calculated from the frequency of occurrence of a match of the two individual amino acids in evolutionary related sequences. This provides a measure of a random alignment of two amino acids. This score also reflects the frequency that a particular amino acid occurs in nature, as some amino acids are more abundant than others.

Higher scores indicate that the probability that those two amino acids aligned by chance is very small, and lower scores indicate a high probability that the two amino acids aligned by chance are evolutionary unrelated. Thus, identities are assigned the most positive sore.

**The two most popular two scoring matrices are PAM and BLOSUM matrices.**

**PAM** means “**Percent accepted mutations”. Margaret Dayhoff and coworkers** originally proposed **PAM model of evolution in the 60s.** Here “**accepted”** means **fixed in populations and is therefore a more complex process than simply mutations.** PAM is usually used for global alignment of closely related proteins.

PAM matrices are derived by counting observed evolutionary changes in closely related protein sequences, and then extrapolating the observed transition probabilities to longer evolutionary, in the distances. It is possible to derive PAM matrices for any evolutionary distance, but in practice, the most commonly used matrices are PAM120 and PAM250. PAM matrices with smaller evolutionary distances represents shorter evolutionary distances. Choosing the matrix for the most appropriate evolutionary distance might result in the best possible alignment of two sequences, but in practice, it is rarely possible to know what the evolutionary distance is, and experience shows that the PAM250 matrix usually produces reasonable alignments.

**PAM** tries to model what happens at long evolutionary distances based on a simple Markov model derived from closely related sequences.

**BLOSUM- Blocks substitution matrix**

BLOSUM matrices are best for detecting local alignments of distantly related proteins. BLOSUM do outperform PAM matrices in practice. The BLOSUM62 matrix is the best for detecting the majority of weak protein similarities and the BLOSUM45 matrices for detecting long and weak alignment.