Pairwise Sequence Alignment

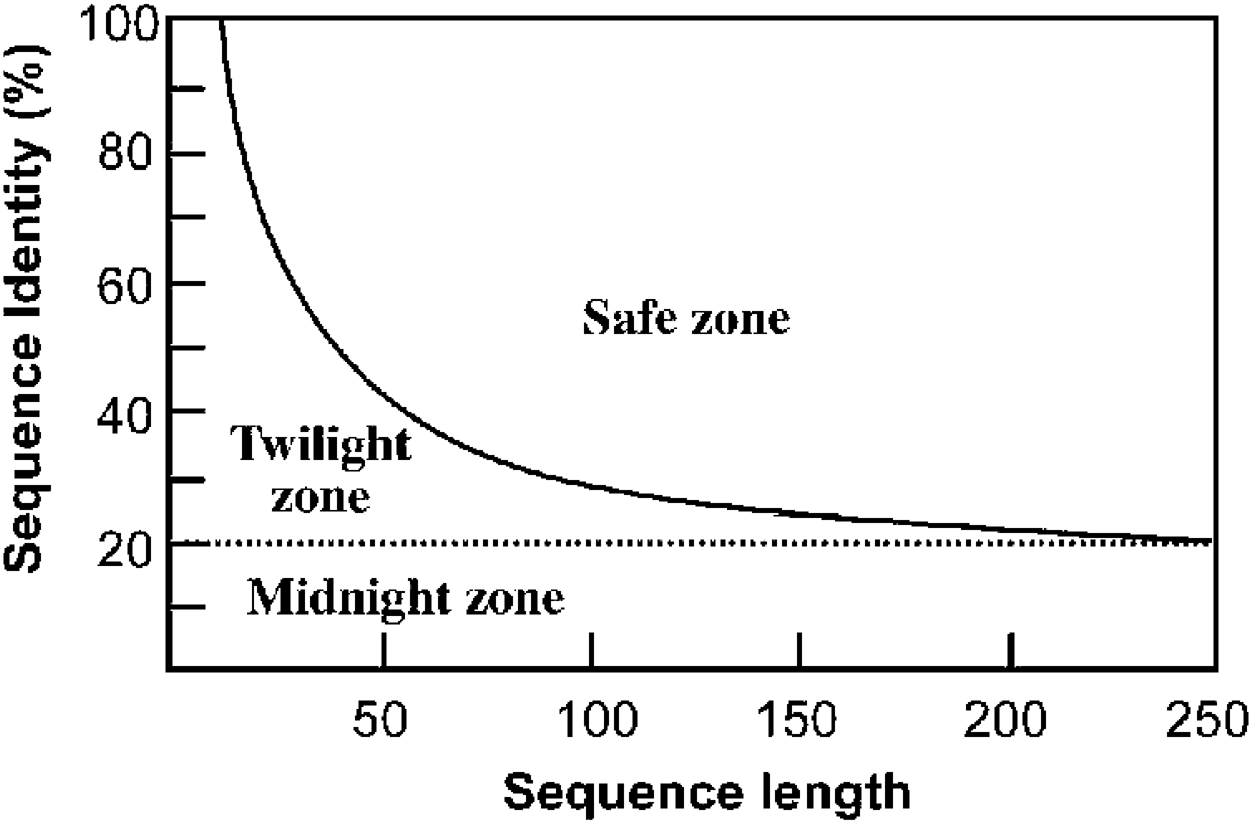
Sequence comparison lies at the heart of bioinformatics analysis. It is an important first step toward structural and functional analysis of newly determined sequences. As new biological sequences are being generated at exponential rates, sequence compare is on is becoming increasingly important to draw functional and evolutionary inference of a new protein with proteins already existing in the database. The most fundamental process in this type of comparison is sequence alignment. This is the process by which sequences are compared by searching for common character patterns and establishing residue–residue correspondence among related sequences. Pairwise sequence alignment is the process of aligning two sequences and is the basis of database similarity searching and multiple sequence alignment.

EVOLUTIONARY BASIS

Identifying the evolutionary relationships between sequences helps to characterize the function of unknown sequences. When a sequence alignment reveals *significant* similarity among a group of sequences, they can be considered as belonging to the same family. If one member within the family has a known structure and function, then that information can be transferred to those that have not yet been experimentally characterized. Therefore, sequence alignment can be used as basis for prediction of structure and function of uncharacterized sequences.

SEQUENCE HOMOLOGY VERSUS SEQUENCE SIMILARITY

An important concept in sequence analysis is sequence homology. When two sequences are descended from a common evolutionary origin, they are said to have a *homologous relationship* or share *homology.* A related but different term is *sequence similarity*, which is the percentage of aligned residues that are similar in physiochemical properties such as size, charge, and hydrophobicity. It is important to distinguish sequence homology from the related term sequence similarity because the two terms are often confused by some researchers who use them interchangeably in scientific literature. To be clear, *sequence homology* is an inference or a conclusion about a common ancestral relationship drawn from sequence similarity comparison when the two sequences share a high enough degree of similarity. On the other hand, *similarity* is a direct result of observation from the sequence alignment. Sequence similarity can be quantified using percentages; homology is a qualitative statement. The shorter the sequence, the higher the chance that some alignment is attributable to random chance. The longer the sequence, the less likely the matching at the same level of similarity is attributable to random chance. This suggests that shorter sequences require higher cutoffs for inferring homologous relationships than longer sequences. For determining a homology relationship of two proteins sequences, for example, if both sequences are aligned at full length, which is 100 residues long, an identity of 30% or higher can be safely regarded as having close homology. They are sometimes referred to as being in the "safe zone". If their identity level falls between 20% and 30%, determination of homologous relationships in this range becomes less certain. This is the area often regarded as the "twilight zone," where remote homologs mix with randomly related sequences. Below 20% identity, where high proportions of nonrelated sequences are present, homologous relationships cannot be reliably determined and thus fall into the "midnight zone."



SEQUENCE SIMILARITY VERSUS SEQUENCE IDENTITY

Another set of related terms for sequence comparison are sequence similarity and sequence identity. Sequence similarity and sequence identity are synonymous for nucleotide sequences. For protein sequences, however, the two concepts are very different. In a protein sequence alignment, *sequence identity* refers to the percentage of matches of the same amino acid residues between two aligned sequences. *Similarity* refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other. There are two ways to calculate the sequence similarity/identity. One involves the use of the overall sequence lengths of both sequences; the other normalizes by the size of the shorter sequence.

The first method uses the following formula:

$$S = [(L\text{s} \times 2) / (L\text{a} + L\text{b})] \times 100$$

Where S is the percentage sequence similarity, L s is the number of aligned residues with similar characteristics, and La and L b are the total lengths of each individual sequence. The sequence identity (I %) can be calculated in a similar fashion:

$$I = [(L\text{i} \times 2) / (L\text{a} + L\text{b})] \times 100$$

where (Li) is the number of aligned identical residues. The second method of calculation is to derive the percentage of identical/similar residues over the full length of the smaller sequence using the formula:

$$\text{I(S)}\ \% = L\text{i(s)}/L\text{a}\%$$

where La is the length of the shorter of the two sequences.