A screen shot of a number

Description automatically generatedOne hot encoding forms vectors of sequences in binary form with length equal to number of possible amino acids(typically 20 for the standard amino acids). Each position in the vector corresponds to a specific amino acid, with “1”s indicating the presence of amino acids at that positions and “0”s else- where. This method treats each amino acids completely independent from others, without any notion of similarity all difference beyond presence or absence.

Although one hot encoding gives a high-dimensional representation, the amino acids being represented as 1’s and 0’s doesn’t not provide sufficient information of the nature of proteins represented, as it fails to capture the biological and chemical properties of amino acids eventually limiting models ability. Upon applying one-hot encoding to extensive sequence data, it may generate highly sparse data matrices, which could result in inefficient utilization of computational resources.

On the other hand, BLOSUM62(Block substitution matrix 62) represent the score for substituting one amino acid for another during protein sequence alignment. These scores are computed by analyzing substitutions observed in protein sequences that lack close relationships. The calculations are derived from aligning blocks of sequences belonging to protein families. The values in the matrix indicate the probability of one amino acid being substituted for another during evolutionary processes, in comparison to random chance.

A diagram of a number of letters

Description automatically generated with medium confidence