Substitution Matrices

Anamaria Hodivoianu

April 9, 2025

1 Introduction

Substitution matrices are used in bioinformatics to measure the similarity between two DNA or amino acid sequences. They are mostly seen in sequence alignment algorithms, and are an application of stochastic matrices. The most common substitution matrices are PAM (Point Accepted Mutation) and BLOSUM (BLOck SUbstitution Matrix).

2 Main Idea

The main idea behind substitution matrices and how they work is that they are based on tha fact that some mutations from one amino acid to another are more likely to occur than others. For example, the mutation from arginine to glutamine, both hydrophillic residues, is more likely to occur than the mutation from arginine to leucine, a hydrophobic residue. A substitution matrix is a matrix in which the position (i, j) represents the score of substituting the amino acid i with the amino acid j. The higher the score, the more likely the substitution is to occur.
An example of a substitution matrix is the identity matrix, which is a square matrix in which all the elements of the principal diagonal are ones and all other elements are zeros. This means that the substitution of an amino acid with itself is very likely to occur, while the substitution of an amino acid with another one is not likely to occur. The identity matrix is not actually used in practice.
A more useful substitution matrix is a log-odds matrix. The formula for a log-odds matrix is:

$$S(i,j) = \log\left(\frac{p(i)\cdot M(i,j)}{p(i)\cdot p(j)}\right) = \log\left(\frac{M(i,j)}{p(j)}\right) \tag{1}$$

where $S(i,j)$ is the score of substituting amino acid i with amino acid j, $p(i)$ and $p(j)$ are the probabilities of amino acids i and j occurring in a sequence, and $M(i,j)$ is the probability of amino acid i being substituted with amino acid j.

3 PAM Matrices

The PAM (Point Accepted Mutation) matrix is one of the first amino acid substitution matrices. It was invented in the 1970s by Margaret Dayhoff. The PAM matrix is calculated with the differences between closely related sequences. A PAM unit represents that 1% of the amino acids in a sequence have changed. A PAM1 substitution matrix is calculated by taking some closely related sequences that have mutations that can be considered PAM1, and then observing the mutations that occur. The PAM1 matrix can be used to calculate other matrices such as PAM2, PAM3, etc. This is based on the assumption that mutations will occur the same as in PAM1, and so to calculate PAM2 the probabilities are sqaured. The most common PAM matrices are PAM30 and PAM70.

4 BLOSUM Matrices

The BLOSUM (BLOck SUbstitution Matrix) matrix is another amino acid substitution matrix. It was invented by Henikoff and Henikoff to solve some of the problems of PAM matrices, such as the fact that they are not very acccurate for evolutionarily divergent sequences. Instead of looking at small changes, BLOSUM matrices look as blocks of sequences that are conserved through evolution. The more conserved the block, the less likely it is to be substituted, since it is probably important for the function of the protein.

5 Conclusion

Substitution matrices are important in bioinformatics, and they are used to measure the similarity between two sequences. They are based on the fact that some mutations are more likely to occur than others. The most common substitution matrices are PAM and BLOSUM.