Substitution Matrix

The substitution matrix represents a fundamental concept in bioinformatics and computational biology, and it is an essential instrument in analyzing biological sequences, especially proteins. This matrix is used to quantify the similarity between different biological elements, like amino acids or nucleotides, and they play a crucial role in sequence alignment, protein structure prediction, and molecular evolution studies.

Definition and purpose

A substitution matrix is, in essence, a table of numerical values that show the probability or cost of substituting one element for another in a biological sequence. In the case of proteins, these matrices reflect the probability of replacing one amino acid with another during evolution. The values in the matrix can be positive or negative: positive means favorable (conservative) substitutions, while negative values mean unfavorable (non-conservative) substitutions.

The main purpose of the substitution matrices is to optimize the biological sequence alignment. When two sequences are compared, the matrix of substitution allows the assignment of a score for each pair of aligned elements, reflecting thus the degree of similarity between them. This score is essential in determining the best sequence alignment.

Types of substitution matrices

PAM Matrices (Point Accepted Mutation)

PAM matrices are among the first substitution matrices used in bioinformatics. They are based on the observation of evolutionary changes in closely related proteins. PAM1 is a matrix calibrated to reflect substitutions that occur with a probability of 1% over a defined evolutionary period. PAM matrices with higher numbers (PAM40, PAM120) are mathematically derived from PAM1 and are used to compare sequences with different degrees of evolutionary divergence.

BLOSUM Matrices (BLOcks SUbstitution Matrix)

BLOSUM matrices are constructed based on the frequency of substitutions observed in blocks of aligned sequences without gaps from the BLOCKS database. Unlike PAM, BLOSUMs are derived directly from multiple alignments of real protein sequences. BLOSUM62, which is calibrated for sequences that share approximately 62% identity, is one of the most widely used matrices in sequence searching and alignment.

Constructing substitution matrices

The construction of a substitution matrix involves several steps:

1. Collecting data: Protein or DNA sequences that are known to be evolutionarily related are gathered.
2. Sequence alignment: The sequences are aligned to identify homologous positions.
3. Counting the substitutions: The frequency with which each amino acid or nucleotide is replace with one another in these alignments is counted.
4. Calculating the probabilities or scores: These frequencies are converted in probabilities or logarithmic scores.
5. Calibration: The matrix is often calibrated to reflect a certain level of evolutionary divergence.

Uses in bioinformatics

Substitution matrices have numerous applications in bioinformatics:

Sequence Alignment

The most common use is in sequence alignment algorithms. These algorithms use the substitution matrix to decide how to best align two sequences, giving positive scores for good alignments and penalties for mismatches and gaps.

Database Search

Programs such BLAST (Basic Local Alignment Search Tool) use substitution matrices to identify similar sequences in a large database. Choosing the right matrix can significantly influence the search results.

Protein Structure Prediction

In protein structure modeling, substitution matrices help identify conserved regions and predict the impact of mutations on protein structure and function.

Limitations and Challenges

The substitution matrices have important limitations including the wrongful assumption that substitutions at different positions are independent (when in reality they are often influenced by sequence context) and the fact that they are constructed based on limited datasets that do not always reflect all the evolutionary pressures and constraints present in nature.

Conclusion

Substitution matrices are fundamental in bioinformatics, offering a quantitative basis for comparing biological sequences. With the continued evolution they remain essential for sequence alignment, database searching, and evolutionary studies, contributing significantly to the advancement of molecular biology research.