**Computational Sequence Alignment**

**(using Needleman-Wunsch and Smith Waterman Algorithms)**

**Objective:**

To perform pairwise sequence alignment on a set of protein sequences using both global (Needleman-Wunsch) and local (Smith-Waterman) alignment algorithms, evaluate similarity metrics, and visualize functional impacts of amino acid substitutions using the BLOSUM62 substitution matrix.

**Methodology:**

1. **Data Collection:**
   * Protein sequences are collected and stored in a CSV file named protein\_sequences.csv, with one column titled seq containing amino acid sequences.
   * The BLOSUM62 substitution matrix is also provided in CSV format as BLOSUM62\_matrix.csv.
2. **Preprocessing:**
   * The protein sequences are read from the CSV file using pandas.
   * The BLOSUM62 matrix is parsed into a dictionary for efficient lookup during alignment scoring.
3. **Sequence Alignment Algorithms:**
   * **Needleman-Wunsch Algorithm (Global Alignment):**
     + This dynamic programming algorithm aligns entire sequences from end to end.
     + A scoring matrix is created where each cell contains the optimal alignment score up to that position.
     + Traceback is used to construct the aligned sequences.
   * **Smith-Waterman Algorithm (Local Alignment):**
     + Focuses on aligning the most similar regions (subsequences) between two sequences.
     + The scoring matrix allows local alignment by including 0 as a valid score.
     + The alignment ends at the cell with the highest score and traceback is performed until a zero is encountered.
4. **Pairwise Comparison:**
   * All unique sequence pairs are aligned using both algorithms.
   * For each pair, the alignment score and aligned sequences are recorded.
   * Similarity is calculated as the ratio of matching characters in the local alignment.
5. **Visualization and Analysis:**
   * **Substitution Matrix Heatmap:**
     + A heatmap of the BLOSUM62 matrix is generated to show the substitution scores between amino acids.
   * **Similarity Heatmap:**
     + A matrix of local similarity scores is visualized to identify closely related sequences.
6. **Output:**
   * The results, including sequence indices, alignment scores, aligned sequences, and similarity metrics, are saved to a CSV file named alignment\_results.csv.
   * Visualizations provide intuitive insights into sequence similarity and substitution impact.

**Tools & Libraries:**

* Python
* Biopython
* pandas, numpy, matplotlib, seaborn

**Conclusion:**

This project enables comprehensive comparison of protein sequences using two fundamental alignment techniques. The integration of substitution matrices and visual analytics provides deep insights into sequence relationships and potential functional impacts of mutations.