Problem Statement and Goals
Substitution-Matrix benchmarking with pairwise sequence alignment

Uriel Garcilazo Cruz

January 18th, 2025

Table 1: Revision History

Date	Developer(s)	Change
18 January	Uriel Garcilazo Cruz	Document's first release
18 January	Uriel Garcilazo Cruz	Title's correction

1 Problem Statement

Substitution matrices make one of the axioms on which to elaborate hypotheses in comparative biology. There are many substitution matrices used in the literature, and their effects in different types of sequences is not always easy to evaluate through benchmarking.

1.1 Problem

Aligning DNA strands from two individuals or species helps revealing past evolutionary events between them. However, finding empirical values for the rate at which one nucleotide or aminoacid changes into another is difficult, because any evidence of substitutions that may have occurred as intermediate stages gets erased by new mutations. A square matrix that encodes the substitution rates among nucleotides or aminoacids is called a substitution matrix. Although multiple substitution matrices have been proposed, and subsequently adopted as a standard for alignment [1, 2], it is of the utmost importance to determine how a given substitution matrix may impact the quality of an alignment. For this it is useful to use methods of pairwise alignment [3] that ensure the best global score between two sequences. This program enables the benchmarking of substitution matrices to determine their effects in a pairwise alignment.

1.2 Inputs and Outputs

Given two sequences of DNA, determine the effects that commonly used substitution matrices have in the quality of their alignment.

1.3 Stakeholders

Evolutionary Biologists and researchers with genomic or protein data, interested in the effects of hyperparameters in the quality of their research.

1.4 Environment

Linux terminal recommended
Windows 10 or higher is recommended.
MacOS Sierra or later is recommended

2 Goals

- Gives a way to measure the effects that a substitution matrix has over the quality of an alignment.
- Yields a comparative tool to evaluate the effects of diverse substitution matrices.
- Provides a benchmarking tool to evaluate the utility of matrices found in the literature.

3 Stretch Goals

- By finding the best possible alignment, the data can be used to train machine learning algorithms.
- framework could be extended to optimize substitution matrices for specific sequence types.
- The analysis could help identify which matrix elements are most important for accurate alignment.

References

[1] S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology*, 219(3):555–565, jun 1991.

[2] D. W. Mount. Comparison of the PAM and BLOSUM amino acid substitution matrices. *Cold Spring Harbor Protocols*, 2008:pdb.ip59, jun 2008.

[3] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.