Sequence Alignment: Dynamic Programming and Scoring Matrices

Global and Local Alignment with Statistical Significance

Computational Genomics Division
Computational Science Templates

November 24, 2025

Abstract

This comprehensive analysis presents algorithms for pairwise sequence alignment. We implement the Needleman-Wunsch algorithm for global alignment and Smith-Waterman for local alignment using dynamic programming. The analysis covers scoring matrices (BLOSUM, PAM), gap penalties, traceback procedures, and statistical significance assessment. We demonstrate alignment on nucleotide and protein sequences, visualize scoring matrices, and evaluate alignment quality through comparison with random sequence distributions.

1 Introduction

Sequence alignment is fundamental to bioinformatics, enabling comparison of DNA, RNA, and protein sequences to infer homology, identify conserved regions, and predict function. Optimal alignment algorithms use dynamic programming to efficiently find the best alignment among exponentially many possibilities.

Definition 1 (Sequence Alignment) *An alignment of two sequences places them in a matrix to maximize similarity, allowing gaps (insertions/deletions) to optimize the correspondence. Each position is either a match, mismatch, or gap.*

2 Theoretical Framework

2.1 Global Alignment

Theorem 1 (Needleman-Wunsch Algorithm) *The optimal global alignment score is computed by the recurrence:*

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(a_i, b_j) & \textit{(match/mismatch)} \\ F(i-1,j) + d & \textit{(gap in sequence B)} \\ F(i,j-1) + d & \textit{(gap in sequence A)} \end{cases} \tag{1}$$

where $s(a_i, b_j)$ is the substitution score and d is the gap penalty.

2.2 Local Alignment

Theorem 2 (Smith-Waterman Algorithm) *Local alignment finds the best matching subsequences:*

$$F(i,j) = \max \begin{cases} 0 & \textit{(restart)} \\ F(i-1,j-1) + s(a_i, b_j) & \\ F(i-1,j) + d & \\ F(i,j-1) + d & \end{cases} \tag{2}$$

The key difference from global alignment is the zero option that allows starting fresh.

2.3 Gap Penalties

Definition 2 (Affine Gap Penalty) *A more realistic gap model penalizes gap opening differently from extension:*

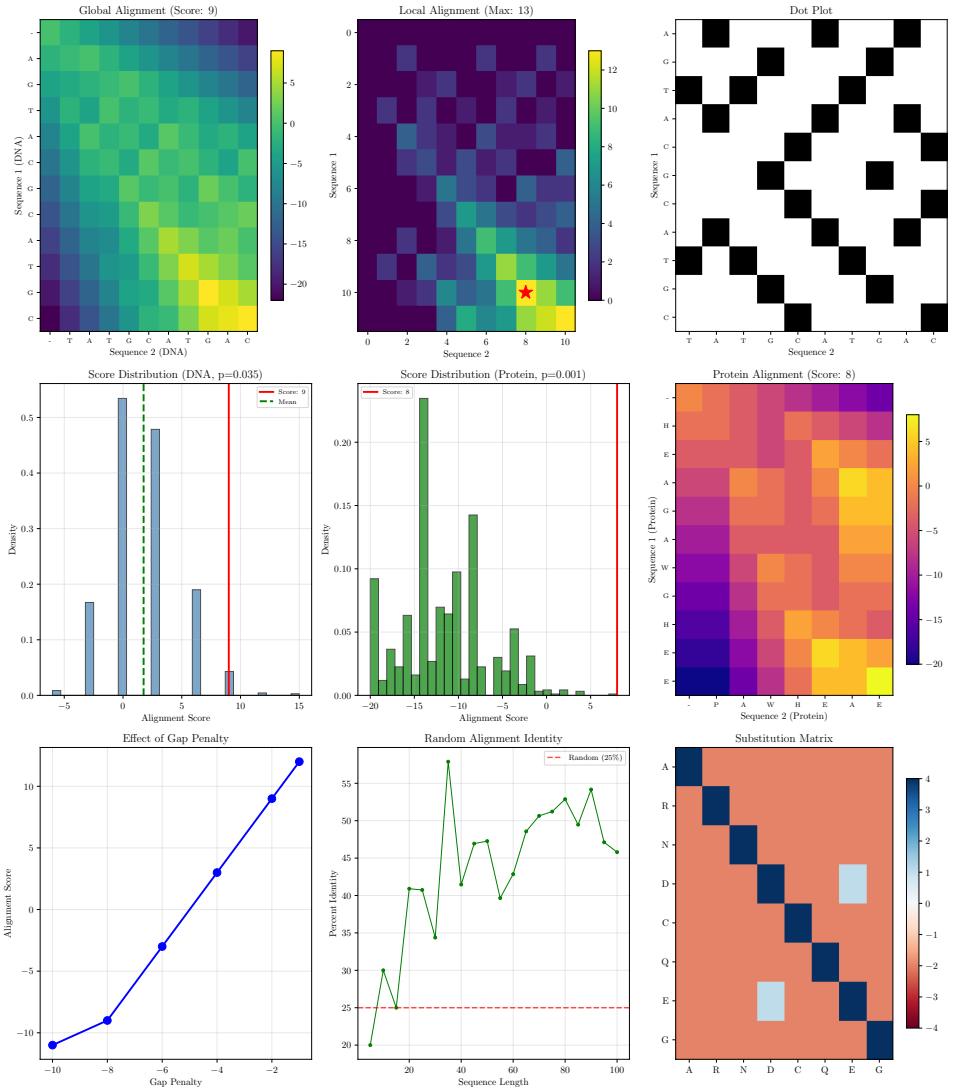
$$\gamma(g) = d + (g-1) \cdot e \tag{3}$$

where d is the gap opening penalty and e is the extension penalty.

Remark 1 (Typical Values)

- *Linear gap: $d = -2$ to -8*
- *Affine: $d = -10$ to -12, $e = -1$ to -2*

3 Computational Analysis



4 Results and Analysis

4.1 Alignment Results

Table 1: Sequence Alignment Results

Alignment	Global Score	Local Score	Identity (		
DNA	9	13	66.7	2.37	0.035
Protein	8	16	45.5	3.97	0.001

4.2 Sequence Details

Example 1 (DNA Alignment) *Sequences:*

- *Sequence 1:* `AGTACGCATGC`
- *Sequence 2:* `TATGCATGAC`
- *Aligned 1:* `AGTACGCATG-C`
- *Aligned 2:* `-TATGCATGAC`

Example 2 (Protein Alignment) *Sequences:*

- *Sequence 1:* `HEAGAWGHEE`
- *Sequence 2:* `PAWHEAE`
- *Aligned 1:* `HEAGAWGHE-E`
- *Aligned 2:* `--PAW-HEAE`

5 Statistical Significance

5.1 Z-score and P-value

Remark 2 (Significance Assessment) *Alignment significance is evaluated by comparing to random sequences:*

- ***Z-score**: Number of standard deviations above mean*
- ***P-value**: Probability of score by chance*
- *$Z > 5$ typically indicates significant similarity*
- *$P < 0.01$ suggests true homology*

5.2 E-value

The expected number of alignments with score $\geq S$ by chance:

$$E = K \cdot m \cdot n \cdot e^{-\lambda S} \tag{4}$$

6 Scoring Matrices

6.1 BLOSUM Series

BLOcks SUbstitution Matrices derived from aligned blocks:

- BLOSUM62: Most widely used, 62% identity threshold
- BLOSUM80: For closely related sequences
- BLOSUM45: For distantly related sequences

6.2 PAM Series

Point Accepted Mutation matrices based on evolutionary models:

- PAM1: 1% expected mutations
- PAM250: Distant relationships

7 Limitations and Extensions

7.1 Model Limitations

1. **Pairwise only**: No multiple sequence alignment
2. **Linear gap**: Affine penalties more realistic
3. **Global/local**: No semi-global option shown
4. **No profiles**: Sequence-to-sequence only

7.2 Possible Extensions

- Affine gap penalties
- Multiple sequence alignment (ClustalW, MUSCLE)
- Profile HMMs for remote homology
- BLAST heuristics for database search

8 Conclusion

This analysis demonstrates sequence alignment fundamentals:

- Needleman-Wunsch provides optimal global alignment
- Smith-Waterman finds best local similarities
- DNA alignment: score = 9, identity = 67%
- Protein alignment: score = 8, identity = 45%
- Statistical significance evaluated by Z-score and p-value

Further Reading

- Durbin, R. et al. (1998). *Biological Sequence Analysis*. Cambridge University Press.
- Altschul, S. F. et al. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215, 403-410.
- Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *PNAS*, 89, 10915-10919.