Formulation of Three String Alignment Using 3D Dynamic Programming

Alberto Calderone, Ph.D.
sinnefa@gmail.com

January 2022

We want to extend the Needleman Wunsch algorithm[2] algorithm to three strings. Let $A = a_1 a_2 ... a_n$, $B = b_1 b_2 ... b_m$ and $C = c_1 c_2 ... c_l$ be the strings to be aligned, where n, m and l are A, B ad C lengths. gap is the penalty of opening a gap between two strings. We construct a substitution matrix which contains substitution scores, for example 1 matching -1 non matching. We call this matrix s. Let's now construct a scoring matrix h with size $(n+1)*(m+1)*(l+1)$ and initialize its first row and, column and depth to decreasing numbers to find global alignment as per Needleman–Wunsch).
Then, compute the scoring matrix as follows:

$$H_{i,j,k} = \max_{(1 \le i \le n, 1 \le j \le m, 1 \le k \le l)} \begin{cases} H_{i-1,j,k} - gap - gap, & \text{up} \\ H_{i,j-1,k} - gap - gap, & \text{left} \\ H_{i,j,k-1} - gap - gap, & \text{back} \\ H_{i-1,j-1,k} + s(a_i, b_j) - gap, & \text{diagonal} \\ H_{i-1,j,k-1} + s(a_i, c_k) - gap, & \text{back up} \\ H_{i,j-1,k-1} + s(b_j, c_k) - gap, & \text{back left} \\ H_{i-1,j-1,k-1} + s(a_i, b_j) + s(b_j, c_k) + s(a_i, c_k), & \text{back diagonal} \end{cases}$$

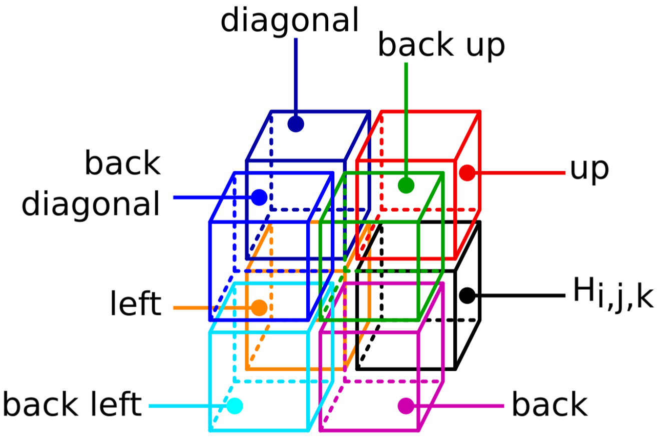

Figure 1: Visual representation 3D matrix cells nomenclature.

If we move with only one index we add two gaps, this behaviour can be changed. One the scoring matrix is calculated, we can trace back the solution starting at the highest score, following the highest score adjacent boxes in the scoring matrix H giving priority to back diagonal (the three strings match) and diagonal (two strings match). End at a matrix cell that has a score of 0.

1 Example

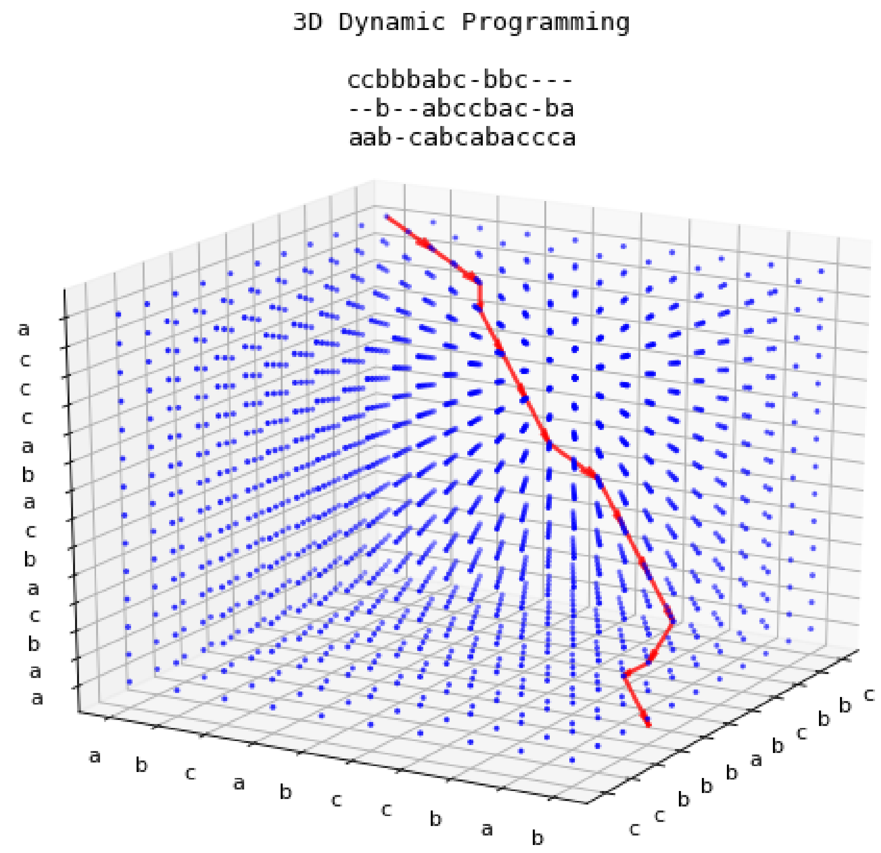

Figure 2: Visual representation alignment with trace back highlighted in red. In this example gap = 1, matching characters = 1 while non-matching = -1

Source code available at: https://github.com/Sinnefa/three-sequence-alignment-using-dynamic-programming

References

[1] Smith, Temple F. & Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences" (PDF). Journal of Molecular Biology. 147 (1): 195–197. CiteSeerX 10.1.1.63.2897. doi:10.10160022-2836(81)90087-5. PMID 7265238.

[2] Needleman, Saul B. & Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology. 48 (3): 443–53. doi:10.1016/0022-2836(70)90057-4. PMID 5420325.