**Course: Advanced Bioinformatics**

**Module title: Local Alignment**

**Module no. : 31**

***Local Alignment: Smith-Waterman Algorithm***

In 1981, Temple Smith and Mike Waterman proposed a modification to the Needleman-Wunsch algorithm in order to obtain a local sequence alignment resulting in the highest-scoring local match between two sequences.

Why choose a local alignment algorithm?

* More meaningful – point out conserved regions between two sequences
* Aligns two sequences of different lengths to be matched
* Aligns two partially overlapping sequences
* Aligns two sequences where one is a subsequence of another

There are only two slight modifications that need to be made to the Needleman-Wunsch Algorithm in order to make it a local alignment algorithm. The first modification requires negative scores for mismatches. The second modification requires that when the dynamic programming scoring matrix value becomes negative, the value is set to zero, which has the effect of terminating any alignment up to that point. This has the effect of changing the matrix score to:

**Si,j = MAXIMUM[**

**Si-1, j-1 + s(ai,bj**) (match/mismatch in the diagonal),

**Si,j-1 + w** (gap in sequence #1),

**Si-1,j + w** (gap in sequence #2),

**0]**

The local alignments are then produced by starting at the highest-scoring positions in the scoring matrix and following a trace path from those positions up to a box that scores zero.

**Initialization Step.** In the initialization step of local alignment, each row Si,0 is set to 0. In addition, each column S0,j is set to 0. Using the scoring scheme described above, the initialization step results in the following:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-** | **G** | **A** | **A** | **T** | **T** | **C** | **A** | **G** | **T** | **T** | **A** |
| **-** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| **G** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **G** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **A** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **T** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **C** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **G** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **A** | **0** |  |  |  |  |  |  |  |  |  |  |  |

**Matrix Fill Step**. One possible solution of the matrix fill step finds the maximum local alignment score by starting in the upper left hand corner in the matrix and finding the maximal score Si,j for each position in the matrix. In order to find Si,j for any i,j it is minimal to know the score for the matrix positions to the left, above and diagonal to i, j. In terms of matrix positions, it is necessary to know Si-1,j, Si,j-1 and Si-1, j-1.

For each position, Si,j is defined to be the maximum score at position i,j; i.e.

**Si,j = MAXIMUM[**

**Si-1, j-1 + s(ai,bj**) (match/mismatch in the diagonal),

**Si,j-1 + w** (gap in sequence #1),

**Si-1,j + w** (gap in sequence #2),

**0]**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-** | **G** | **A** | **A** | **T** | **T** | **C** | **A** | **G** | **T** | **T** | **A** |
| **-** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| **G** | **0** | **5** |  |  |  |  |  |  |  |  |  |  |
| **G** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **A** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **T** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **C** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **G** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **A** | **0** |  |  |  |  |  |  |  |  |  |  |  |

Note that in the example, Si-1,j-1 will be red, Si,j-1 will be green and Si-1,j will be blue.

Using this information, the score at position 1,1 in the matrix can be calculated. Since the first residue in both sequences is a G, s(a1b1)= 5, and by the assumptions stated earlier, w = -4. Thus, S1,1 = MAX[S0,0 + 5, S1,0 - 4, S0,1 – 4,0] = MAX[5, -4, -4, 0].

Now we proceed to S1,2. Since a1 = G and b2 = A, there is a mismatch. Therefore, sa1b2 = -3 and by the assumptions stated earlier, w = -4. Thus, S1,2 = MAX[S0,1 -3, S1,1 - 4, S0,2 – 4, 0] = MAX[0 - 3, 5 – 4, 0 – 4, 0] = MAX[-3, 1, -4, 0] = 1. An arrow is placed back into the cell that resulted in the maximum score, which is the cell S1,1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-** | **G** | **A** | **A** | **T** | **T** | **C** | **A** | **G** | **T** | **T** | **A** |
| **-** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| **G** | **0** | **5** | 1 |  |  |  |  |  |  |  |  |  |
| **G** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **A** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **T** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **C** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **G** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **A** | **0** |  |  |  |  |  |  |  |  |  |  |  |

Now we proceed to S1,3. Since a1 = G and b3 = A, there is a mismatch. Therefore, sa1b2 = -3 and by the assumptions stated earlier, w = -4. Thus, S1,3 = MAX[S0,2 -3, S1,2 - 4, S0,3 – 4, 0] = MAX[0 - 3, 1 – 4, 0 – 4, 0] = MAX[-3, -3, -4, 0] = 0. Since the maximum score is 0 (all other possible scores are negative), no arrow is drawn back from this location.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-** | **G** | **A** | **A** | **T** | **T** | **C** | **A** | **G** | **T** | **T** | **A** |
| **-** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| **G** | **0** | **5** | 1 | 0 |  |  |  |  |  |  |  |  |
| **G** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **A** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **T** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **C** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **G** | **0** |  |  |  |  |  |  |  |  |  |  |  |
| **A** | **0** |  |  |  |  |  |  |  |  |  |  |  |

We can then proceed to fill in the rest of the matrix in a similar fashion. The resulting matrix is as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-** | **G** | **A** | **A** | **T** | **T** | **C** | **A** | **G** | **T** | **T** | **A** |
| **-** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **G** | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 5 | 1 | 0 | 0 |
| **G** | 0 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 5 | 2 | 0 | 0 |
| **A** | 0 | 1 | 10 | 7 | 3 | 0 | 0 | 5 | 1 | 2 | 0 | 5 |
| **T** | 0 | 0 | 6 | 7 | 12 | 8 | 4 | 1 | 2 | 6 | 7 | 3 |
| **C** | 0 | 0 | 2 | 3 | 8 | 9 | 13 | 9 | 5 | 2 | 3 | 4 |
| **G** | 0 | 5 | 1 | 0 | 4 | 5 | 9 | 10 | 14 | 10 | 6 | 2 |
| **A** | 0 | 1 | 10 | 6 | 2 | 1 | 4 | 14 | 10 | 11 | 7 | 11 |

Each cell has one to three arrows indicating from which cell the maximum score was obtained. The matrix fill step is now complete.

Traceback Step. After the matrix fill step, the maximum local alignment score for the two sequences is 14, which can be found by locating the highest values in the score matrix. Note that 14 is found in two separate cells, indicating there are multiple alignments producing the maximal alignment score. The traceback step will find the actual local alignments resulting in the maximum score.

The traceback begins in the position with the highest value. Since pointers have been kept back to all possible predacessors, the traceback is simple. At each cell, we look to see where we move next according to the pointers. When we reach a cell where there is not a pointer to a previous cell, then we have reached the beginning of the local alignment.

First, consider the case where the 14 is in the last row.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-** | **G** | **A** | **A** | **T** | **T** | **C** | **A** | **G** | **T** | **T** | **A** |
| **-** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **G** | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 5 | 1 | 0 | 0 |
| **G** | 0 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 5 | 2 | 0 | 0 |
| **A** | 0 | 1 | 10 | 7 | 3 | 0 | 0 | 5 | 1 | 2 | 0 | 5 |
| **T** | 0 | 0 | 6 | 7 | 12 | 8 | 4 | 1 | 2 | 6 | 7 | 3 |
| **C** | 0 | 0 | 2 | 3 | 8 | 9 | 13 | 9 | 5 | 2 | 3 | 4 |
| **G** | 0 | 5 | 1 | 0 | 4 | 5 | 9 | 10 | 14 | 10 | 6 | 2 |
| **A** | 0 | 1 | 10 | 6 | 2 | 1 | 4 | 14 | 10 | 11 | 7 | 11 |

Note that the blue letters and gold arrows indicate the path leading to the maximum score.

We can continue to follow the path until we get to the following situation:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-** | **G** | **A** | **A** | **T** | **T** | **C** | **A** | **G** | **T** | **T** | **A** |
| **-** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **G** | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 5 | 1 | 0 | 0 |
| **G** | 0 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 5 | 2 | 0 | 0 |
| **A** | 0 | 1 | 10 | 7 | 3 | 0 | 0 | 5 | 1 | 2 | 0 | 5 |
| **T** | 0 | 0 | 6 | 7 | 12 | 8 | 4 | 1 | 2 | 6 | 7 | 3 |
| **C** | 0 | 0 | 2 | 3 | 8 | 9 | 13 | 9 | 5 | 2 | 3 | 4 |
| **G** | 0 | 5 | 1 | 0 | 4 | 5 | 9 | 10 | 14 | 10 | 6 | 2 |
| **A** | 0 | 1 | 10 | 6 | 2 | 1 | 4 | 14 | 10 | 11 | 7 | 11 |

At this point, or alignment (which is built starting at the end of the alignment) is as follows:

**C - A**

**| |**

**C G A**

Now the current cell gets its score either from a match of the T’s or a gap in the second sequence. We’ll consider both as possibilities: Match of the T’s (1) and gap in second (2).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-** | **G** | **A** | **A** | **T** | **T** | **C** | **A** | **G** | **T** | **T** | **A** |
| **-** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **G** | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 5 | 1 | 0 | 0 |
| **G** | 0 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 5 | 2 | 0 | 0 |
| **A** | 0 | 1 | 10 | 7 | 3 | 0 | 0 | 5 | 1 | 2 | 0 | 5 |
| **T** | 0 | 0 | 6 | 7 | 12 | 8 | 4 | 1 | 2 | 6 | 7 | 3 |
| **C** | 0 | 0 | 2 | 3 | 8 | 9 | 13 | 9 | 5 | 2 | 3 | 4 |
| **G** | 0 | 5 | 1 | 0 | 4 | 5 | 9 | 10 | 14 | 10 | 6 | 2 |
| **A** | 0 | 1 | 10 | 6 | 2 | 1 | 4 | 14 | 10 | 11 | 7 | 11 |

Once we reach the node with 0 and there are no pointers from this node, we are finished. The two local alignments resulting in a score of 14 in the final row are:

**G A A T T C - A**

**| | | | |**

**G G A T – C G A**

**+ - + + - + - +**

**5 3 5 5 4 5 4 5**

**G A A T T C - A**

**| | | | |**

**G G A – T C G A**

**+ - + - + + - +**

**5 3 5 4 5 5 4 5**

As you can see, each of these has 5 matches, 1 mismatch, and 2 gaps, so the score is 5(5) – 1(3) – 2(4) = 25 – 3 – 8 = 14. This coincides with the maximum local alignment score calculated in the matrix.