Needleman–Wunsch Algorithm (Global Alignment)

CHAPTER 9

CS Mukhopadhyay and RK Choudhary
School of Animal Biotechnology, GADVASU, Ludhiana

9.1 INTRODUCTION

The Needleman–Wunsch algorithm (NWA; Needleman and Wunsch, 1970) is used for global alignment. We compare homologous molecular sequences character by character to achieve sequence alignment. Global alignment is the end-to-end alignment between two sequences; hence, it introduces gaps that represent insertions/deletions. This is useful for identifying "InDels" (**In**sertion and **Del**etions), and for overall comparison of two or more comparable (i.e., similar) sequences. Phylogenetically close sequences of the same length are the most suited for global alignment.

The best alignment can be identified by quantifying or scoring the possible alternative alignments. Scoring matrices are used to award the match(es) and penalize the mismatch(es) and gap(s), so that the best alignment with the highest score can be identified. The scores in the matrix are integer values (e.g., +1, 0, –1).

Dynamic Programming (DP) means that the scores of the subsequent cells can only be determined if the initial cells (towards the top left) have been scored. The DP methodology requires computation of the initial values (at the left and top side of the matrix) to obtain the later values of the cells (located towards the right and bottom side) in the matrix.

9.2 OBJECTIVE

To align two comparable sequences (nucleotide or amino acid) for obtaining a global alignment using the NWA.

Basic Applied Bioinformatics, First Edition. Chandra Sekhar Mukhopadhyay, Ratan Kumar Choudhary and Mir Asif Iquebal.

9.3 PROCEDURE

Let us start with two short sequences which are to be globally aligned using an NWA:

- Seq1: CTAGTAG
- Seq2: CAGGTAGTG

If the sequences are of length "*m*" and "*n*", respectively, we will obtain one scoring matrix of dimensions $m \times n$.

9.3.1 Step 1: define a scoring scheme

A scoring scheme is first defined, and then the dynamic scoring is done, starting from the top left to the bottom right of the matrix:

Match score (S_{ij}) = 2
Mismatch score (S_{ij}) = –1
Gap penalty (d) = –2

Here, "*i*" (varying from 1, 2, …, *m*) and "*j*" (varying from 1, 2, …, *n*) refer to the indexes for the row and column numbers, respectively, of the cell of the scoring matrix (described in Step 2). Please note that the scoring scheme can differ according to the *evolutionary relatedness* of the input sequences. A gap extension penalty is also introduced in advanced methods of such dynamic algorithms, to direct proper alignment and select the best one, based on the scores.

9.3.2 Step 2: initiation of matrix construction

This step starts with creating a matrix that represents the score for each pair of residues belonging to the pair of sequences.

a. Seq1: written along the vertical axis (or *Y*-axis) of the matrix (i.e., along the rows).
b. Seq2: written along the horizontal axis (or *X*-axis) of the matrix (i.e., along the columns).

TABLE 9.1

← Sequence 1	"*i*" varying from 1, 2, 3 …, *m*		Sequence 2 →								
			"*j*" varying from 1, 2, 3 …, *n*								
			C	A	G	G	T	A	G	T	G
		C									
		T									
		A									
		G									
		T									
		A									
		G									

Finally, we will get a matrix with each of its cells filled up with three scores obtained from three different types of movements, as shown below:

- **Horizontal movement**: represents a gap in the sequence written vertically (along the *Y*-axis, row numbers have been denoted by "*i*")
- **Vertical movement**: represents a gap in the sequence written horizontally (along the *X*-axis, column numbers are indicated by "*j*")
- **Diagonal movement**: representing either match or mismatch between the two sequences (cell positions indexed by "*i*" and "*j*", respectively) (Figure 9.1).

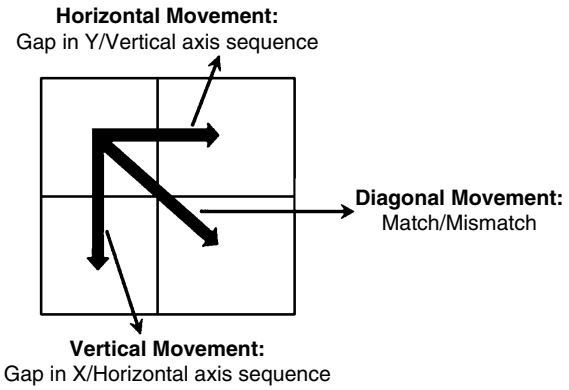

FIGURE 9.1 Three types of movement along the matrix in dynamic programming.

Annex one row at the top and one column at the left of the original $m \times n$ dimensional matrix (to make it an $(m+1) \times (n+1)$ dimensional matrix). These are termed the 0th row and 0th column, respectively. Next, fill the first cell (i.e., the cell located at the top left-most corner of the 0th row and 0th column) with a zero.

9.3.2.1 Scoring method: dynamic programming

The score of each cell can be determined from three different movements towards the cell. The formula used to calculate the scores for each cell is shown in Figure 9.2.

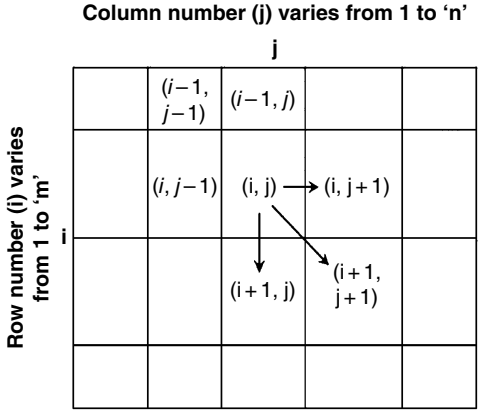

FIGURE 9.2 Increment in the respective indexes of the cells (denoting row and column numbers, respectively) of the matrix, to indicate the movement along the cells.

$$F(i,j) = \max \begin{Bmatrix} Fi-1, j + d\left(\text{Gap in horizontal} - \text{axis sequence; Vertical movement}\right) \\ Fi-1, j-1 + s(i,j)\left(\text{match/mismatch; Diagonal movement}\right) \\ Fi, j-1 + d\left(\text{Gap in vertical} - \text{axis sequence; Horizontal movement}\right) \end{Bmatrix}$$

[9.1]

Select the highest score(s): Now, after obtaining these three scores in the cell being studied, the highest value is selected, which is to be used as the score of that cell. This highest score is used for calculating the respective scores of the cells located just right (hence, pertaining to horizontal movement), just bottom (vertical movement) and just at the bottom-right (diagonal movement) positions of the current cell. If more than one score shows the highest value, then that highest value will be considered for calculating the score of the adjacent cells. However, the backtracking will be through all these cells contributing to the same highest values (see Figures 9.3 and 9.4).

Let us clarify this with the first few cells of the matrix shown in Step 3:

a. Put "0" value in the first cell (utmost top left cell).
b. The value of the next cell (right side) will be 0+(–2)=–2; since it is a horizontal movement, so the gap penalty of –2 will be given. The next cell at the right side will have the score of –2+(–2)=–4, due to the gap penalty for the same horizontal movement. Likewise, the subsequent cells on the right side will be awarded –6, –8, –10, –12, –14, –16, and the last one –18.
c. Now, for the downward movement in the very first column, the gap penalty will be consecutively awarded to each cell. The cells will have the scores of –2, –4, … –14.
d. The diagonal movement starts from 0 to the bottom right cell; we will check whether there is a match or mismatch. In this case, there is a match of "C" to "C". Hence, a score of +2 will be awarded.
e. Please note that, except for the first row and first column (which were annexed to the original matrix), every cell of the matrix can have three values, due to three possible movements towards that cell:
   - vertical movement from the cell just above,
   - horizontal movement from the cell just at left, and
   - diagonal movement from the adjacent cell just at the top-left position).

   In the following matrix, three colors have been used: Blue for vertical movement (Gap), Yellow highlight for diagonal movement (Match or Mismatch), and Red for horizontal movement (Gap).
f. Now, out of these three values, the highest one will be selected as the score for that cell. The chosen score(s) has/have been highlighted in yellow in the matrix below.

9.3.3 Step 3: trace-back step

Finally, the trace-back step starts from the last cell at the right side of the bottom row in a way such that arrow(s) will be drawn to the cell(s) from which the current score (highest of three scores of the present cell) has been obtained. If there are two or three cells which contribute to the highest value, then two (or three, respectively) arrows will be drawn to indicate the previous cells.

The arrows are drawn from the bottom right cell towards the top left cell. If certain cell(s) have more than one arrow, this will lead to more than one path at every such junction.

		C	A	G	G	T	A	G	T	G
	0	–2	–4	–6	–8	–10	–12	–14	–16	–18
C	–2	[–4, 2, –4]	[–6, –3, 0]	[–8, –5, –2]	[–10, –7, –4]	[–12, –9, –6]	[–14, –11, –8]	[–16, –13, –10]	[–18, –15, –12]	[–20, –17, –14]
T	–4	[0, –3, –6]	[–2, 1, –2]	[–4, –1, –1]	[–6, –3, –3]	[–8, –2, –5]	[–10, –7, –4]	[–12, –9, –6]	[–14, –8, –8]	[–16, –13, –10]
A	–6	[–2, –5, –8]	[–1, 2, –4]	[–3, 0, 0]	[–5, –2, –2]	[–4, –4, –4]	[–6, 0, –6]	[–8, –5, –2]	[–10, –7, –4]	[–12, –9, –6]
G	–8	[–4, –7, –10]	[0, –3, –6]	[–2, 4, –2]	[–4, 2, 2]	[–6, –3, 0]	[–2, –5, –2]	[–4, 2, –4]	[–6, –3, 0]	[–8, –2, –2]
T	–10	[–6, –9, –12]	[–2, –5, –8]	[2, –4, –7]	[0, 3, 0]	[–2, 4, 1]	[–4, –1, 2]	[0, –3, 0]	[–2, 4, –2]	[–4, –1, 2]
A	–12	[–8, –11, –14]	[–4, –4, –10]	[0, –3, –6]	[1, 1, –2]	[2, 2, –1]	[0, 6, 0]	[–2, 1, 4]	[2, –1, 2]	[0, 3, 0]
G	–14	[–10, –13, –16]	[–6, –9, –12]	[–2, –2, –8]	[–1, 2, –4]	[0, 0, 0]	[4, 1, –2]	[2, 8, 2]	[0, 3, 6]	[1, 4, 4]

FIGURE 9.3 Each cell is assigned three scores obtained from three possible movements – namely, horizontal, diagonal and vertical. The arrows indicate back-tracing based on the highest score out of the three scores.

Tracing back is done from the bottom right cell towards the top left cell. The path is determined based on the source of the cell which contributes to the highest score in the present cell.

The dynamic programming thus proceeds from one cell to the other (in the defined directions), until the whole matrix obtains the score for each of the cells.

		C	A	G	G	T	A	G	T	G
	0	–2	–4	–6	–8	–10	–12	–14	–16	–18
C	–2	**2**	**0**	**–2**	**–4**	–6	–8	–10	–12	–14
T	–4	**0**	**1**	–1	–3	**–2**	–4	–6	*–8*	–10
A	–6	–2	***2***	**0**	–2	–4	***0***	–2	–4	–6
G	–8	–4	0	***4***	***2***	0	–2	***2***	0	*–2*
T	–10	–6	–2	2	3	***4***	2	0	***4***	2
A	–12	–8	*–4*	0	1	2	***6***	**4**	**2**	3
G	–14	–10	–6	*–2*	*2*	0	4	***8***	**6**	***4***

FIGURE 9.4 Trace-back starts from the bottom right cell towards the top left cell, according to the highest score(s) obtained in the previous step. There could be more than one path at a point (i.e., cell), if that cell has been awarded more than one highest score, due to two or three movements in the previous step.

All possible paths are obtained from the scoring matrix during the process of global alignment.

9.3.4 Step 4: calculating the scores for each alignment

Points to remember during scoring:

When only the column number (but not the row number) is increased by 1 (i.e., "j" is increased to "$j+1$"):

a. Indicates no change in row position, but there is one cell movement to the right.
b. Horizontal sequence (written along the X-axis) has one residue which is missing in the vertically written sequence (i.e., along the Y-axis).
c. A gap is introduced in the vertically written sequence.

When only the row number (but not the column number) is increased by 1 (i.e., "i" is increased to "$i+1$"):

a. Indicates no change in column position, but there is one cell movement to the bottom.
b. The vertical sequence (written along the Y-axis) has a residue which is missing in the horizontally written sequence (i.e., along the X-axis).
c. A gap is introduced in the horizontally written sequence

When both row and column numbers are increased by 1 (i.e., "i" and "j" are increased to "$i+1$" and "$j+1$", respectively):

a. Indicates one-cell diagonal movement towards the bottom left.
b. The vertical sequence (written along the Y-axis) and the horizontally written sequence (i.e., along the X-axis) have one residue each, which may be matching (award match score) or mismatching (penalize with the mismatch score).
c. No gap is introduced in either vertical or horizontally written sequences.

The score for each of these alignments is calculated according to the scoring scheme set at the beginning of scoring. The alignment score with the highest value is considered as the best global alignment. One could get more than one highest score (same value) when there are multiple (more than one) pair-wise alignments. In such a situation, all those alignments with the highest score are equally good, and any one of these can be accepted as the global alignment.

In the example shown above, we get seven possible global alignments, all of which have the highest score (i.e., equal to 4). Here, we can select any one of these alignments as the best alignment (Figure 9.5).

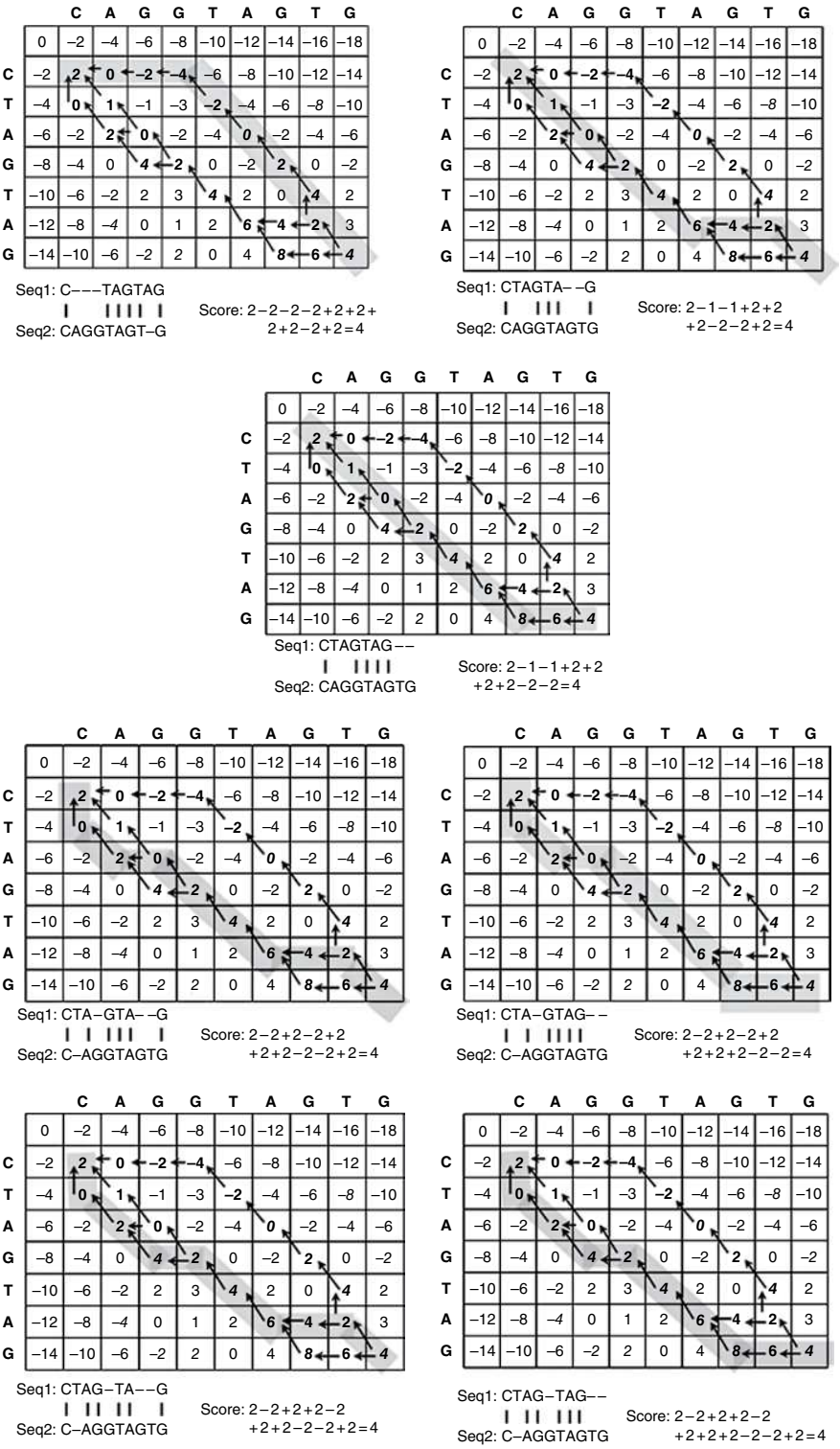

FIGURE 9.5 Global alignment (by NWA) has yielded seven equally good (same alignment score of 4) alignments.