**Describe the Needleman-Wunsch algorithms for pairwise nucleotide sequence alignment and discuss how they can be used for genome assembly.**

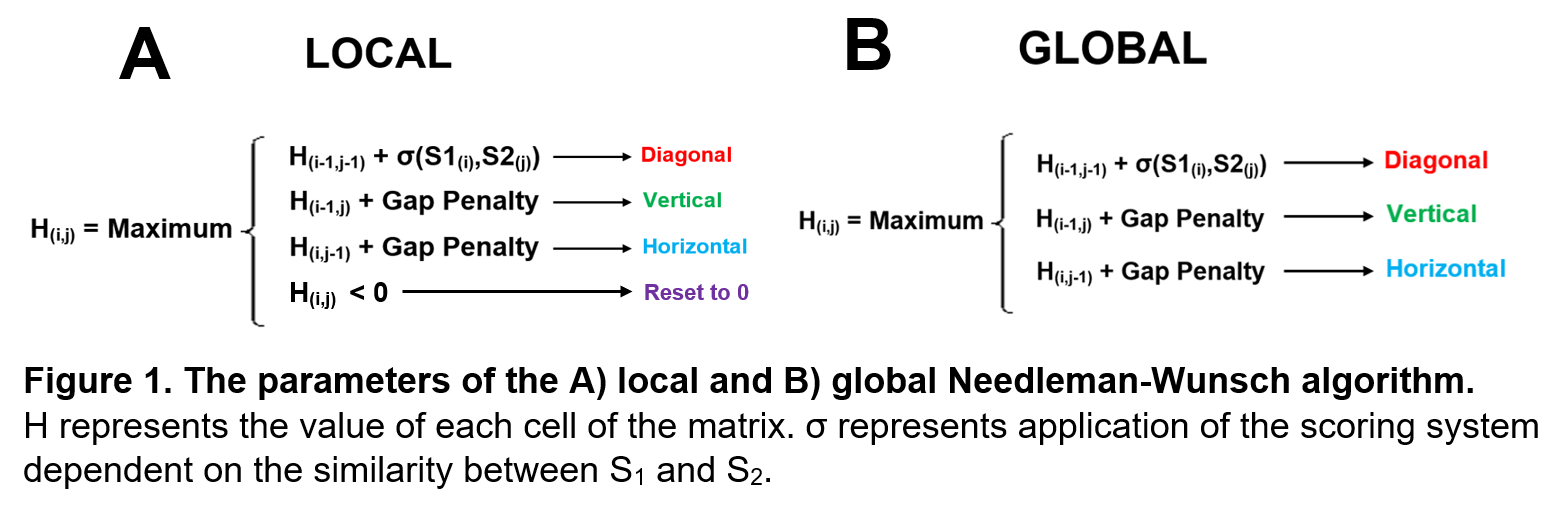
In the modern era of bioinformatics, dynamic programming represents an established computational approach to the rationalisation of biological complexity. However, the expansion of dynamic programming was only facilitated through innovation of foundational algorithms that deconvolute biological information, such as the Needleman-Wunsch algorithm derived in 1970. The Needleman-Wunsch algorithm generates an alignment between query and reference nucleotide sequences by comparing every character in a pairwise-manner. Extensive application of this algorithm thereby enables *de novo* genome assembly, as overlapping sequence reads can be concatenated into contigs and assembled onto a scaffold to constitute a continuous chromosome or region of DNA.

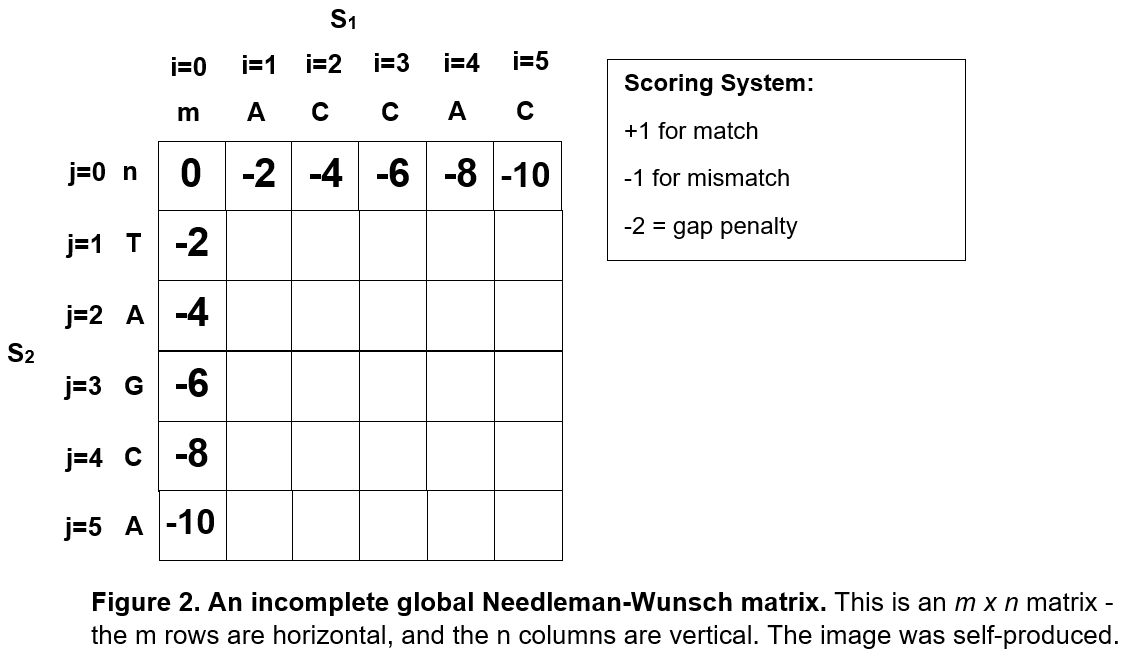
However, the original Needleman-Wunsch algorithm contained several inherent limitations, not least the unrealistic nature of a fixed gap-penalty; the model has been subsequently upgraded to include an enhanced penalty for the first opening in pairwise sequence alignment compared to each consecutive extension of the gap, as described by Smith and Waterman. Nevertheless, the ability to map the location of large numbers of sequence reads relative to a reference sequence proved paradigmatic for genome assembly, as overlapping contigs could be assembled together to produce continuous segments of sequenced genome. The ongoing relevance of the Needleman-Wunsch algorithm should therefore not be underestimated.

Needleman-Wunsch algorithms can be broadly classified into two pairwise approaches for sequence alignment: a global alignment aligns both the query and reference sequences across their entire length, whereas a local alignment aligns a substring of the query sequence with a substring of reference sequence. The biological context of a particular experiment will determine which type of Needleman-Wunsch algorithm is selected. Local alignment is more readily employed for evolutionarily divergent sequences, where any small region of nucleotide alignment is likely to correlate with high conservation in a particular gene. Global alignments are more useful in comparing homologous genes through end-to-end alignment of both sequences, to determine the location and prevalence of genomic variation. Both types of Needleman-Wunsch algorithm therefore contribute to genome assembly in a different way; repeated global alignment enables concatenation of sequence read overlaps into adjacent contig loci to form a scaffold, whereas local alignment is used to distinguish regions with highest similarity between highly-conserved sequence reads, often in conjunction with a reference genome.

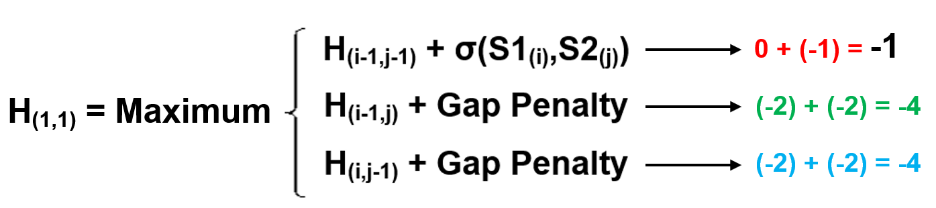
Both categories of Needleman-Wunsch algorithm also satisfy the recursive and heuristic requirements for any dynamic programming system; pairwise sequence alignment is rationalised into multiple less-complex problems that are assigned quantitative values using a variable scoring-system which favours matches over mismatches, with penalties for gaps. Furthermore, both classes of Needleman-Wunsch algorithm are highly conservative with low false-positive rates, exert little computational cost and are relatively quick to run – a query sequence of 1,000 nucleotides requires one second and 32 MB for pairwise alignment to a reference sequence [1].

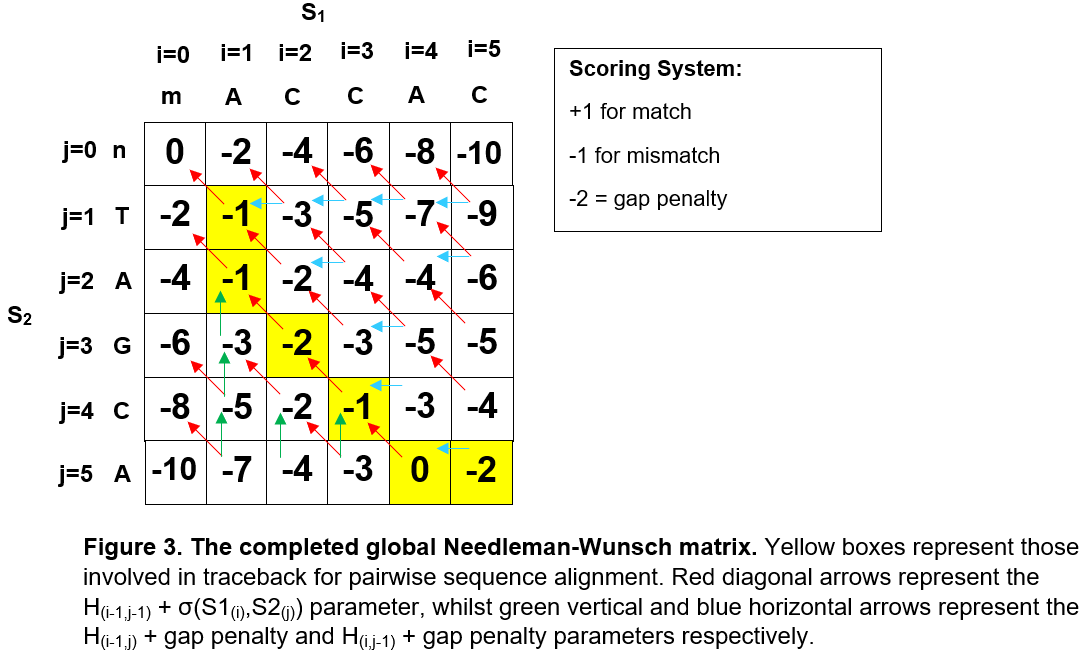
However, the local Needleman-Wunsch alignment introduces a fourth parallel term into the parameters of the Needleman-Wunsch matrix shown in Figure 1a, where negative scores derived from prior mismatches or gap penalties are reset to zero. The differences between the two algorithmic categories are perhaps best-exhibited through comparing examples of the two Needleman-Wunsch algorithms.

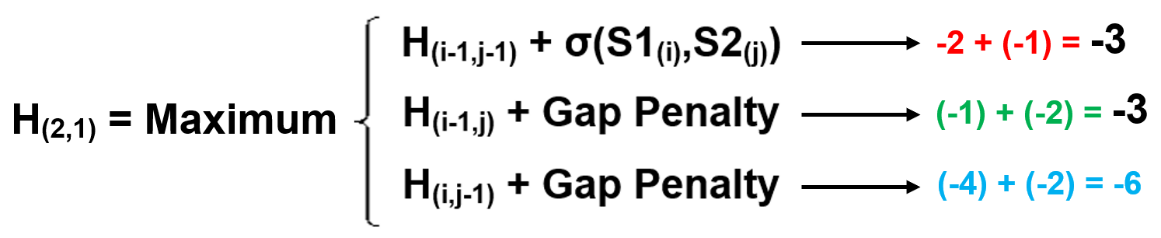


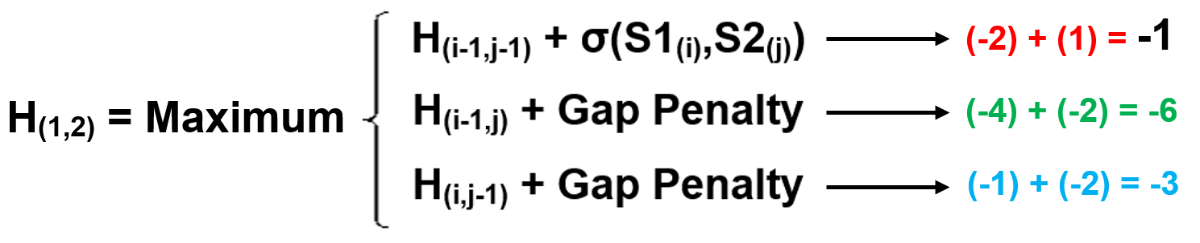
As displayed in Figure 1b, global Needleman-Wunsch parameters contain only three terms, where H represents the value of a matrix cell. Figure 2 displays the original blank global Needleman-Wunsch matrix for exemplar query-sequence ACCAC (S1) and reference- sequence TAGCA (S2), alongside a scoring-system that assigns a value to each of the three options for potential sequence alignment (i.e., match, mismatch, or gap). For both i0j𝑥 and i𝑥j0, a score equal to -2𝑥 is assigned to the matrix cell (the maximum of the three parameters is always located at 𝑥-1, with a further gap penalty of two subtracted for each successive cell).

In order to explain how scores are ascribed to each matrix cell, consider row j=1:

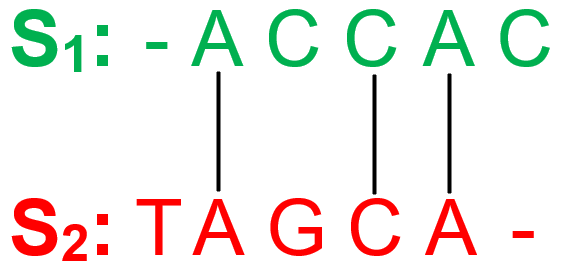
For H(1,1), the maximum value of the three parameters of the global Needleman-Wunsch algorithm is -1, derived from subtracting a mismatch penalty from the H(0,0) value of zero, as shown below. Figure 3 shows the recursively completed Needleman-Wunsch matrix; the value of H(1,1) is obtained from the top parameter, and hence is denoted with a red-diagonal arrow pointing to H(0,0) in accordance with the colour-scheme first introduced in Figure 1.



For H(2,1), the maximum value for the three parameters of the global Needleman-Wunsch algorithm is -3. As shown below, this arises from both the H(1,0) and H(1,1) matrix cells, where we subtract a mismatch penalty of one and a gap penalty of two from initial values of -2 and -1 respectively. Figure 3 therefore represents the H(2,1) cell with a red-diagonal arrow drawn to H(1,0) and a blue-horizontal arrow drawn to H(1,1).

Where j=1, there are no matches between the query and reference nucleotide sequences, so we now consider H(1,2), where an adenine in the query aligns with an adenine in the reference sequence. Within the parameters of the Needleman-Wunsch matrix, the maximum value of H(1,2) is -1, as shown below. One point due to a sequence match is added to the -2 value of the H(0,1) cell.

Once the global Needleman-Wunsch matrix has been completed, the query and reference sequences are aligned by traceback from the bottom-right corner, which here has a value of -2. Traceback is conducted following the direction of the arrows used to generate the value of each matrix cell – the boxes highlighted in yellow represent the path of traceback for our example algorithm. The traceback sequence is then inverted to produce a final query-reference alignment – our example pairwise-alignment is shown below:

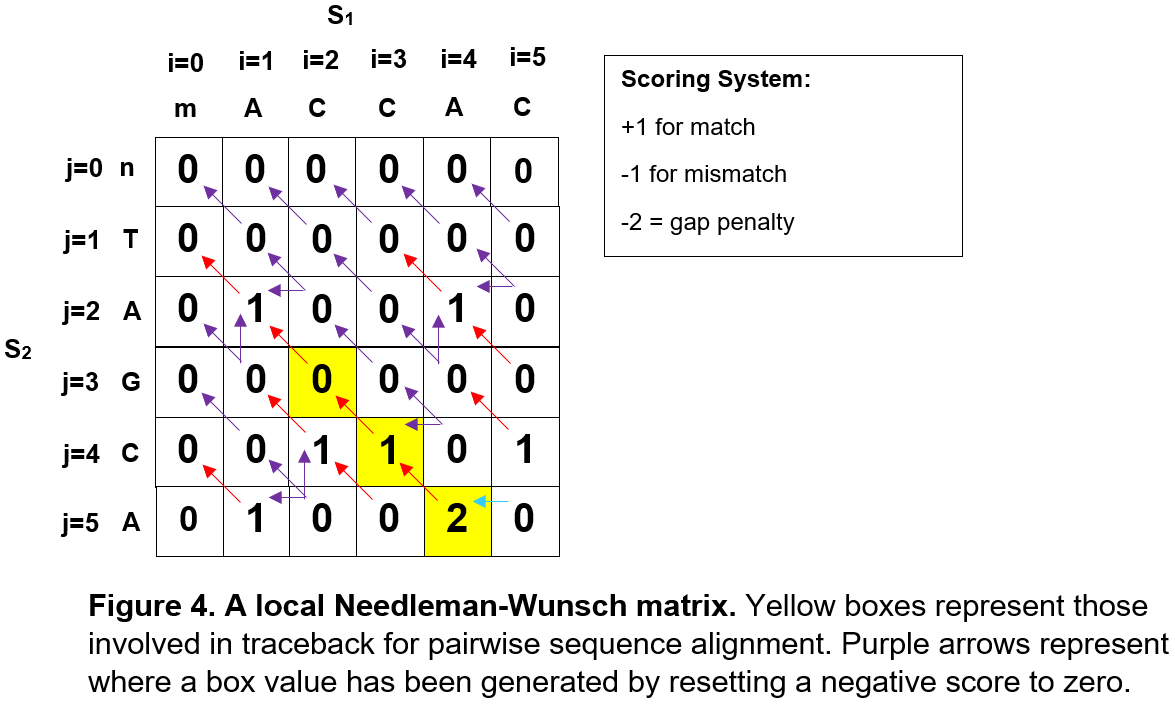


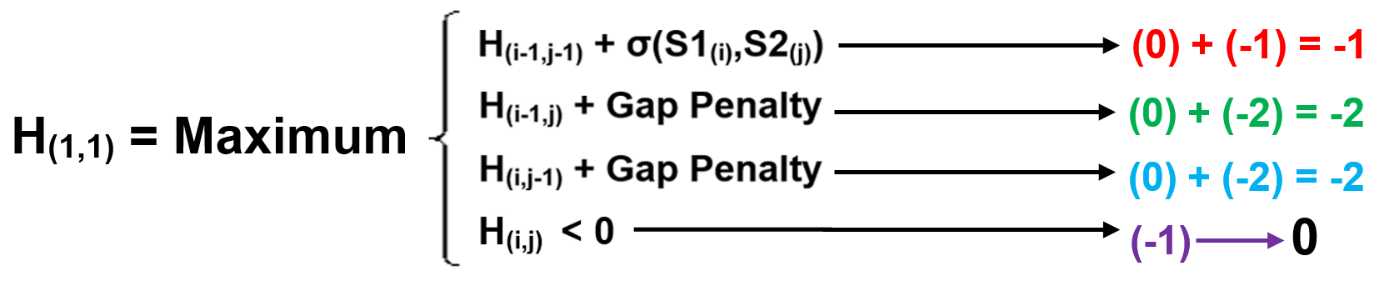
Repeated global sequence alignment can therefore identify regions of overlapping sequence between adjacent contigs to allow their concatenation into an uninterrupted scaffold and ultimately a contiguous genome. In *de novo* assembly, sequence reads are compared to each another instead of a reference sequence. Mauve software aligns reads according to Needleman-Wunsch principles, and concatenates the two fragments with the largest overlap as contigs [2]. These two fragments are then removed from the data-set for this particular estimation so that the two fragments with the next-largest overlap can be merged without re-using the same sequence reads. Mauve provides a ranked list of the estimated optimum arrangements of concatenated contigs in the final scaffold, which often constitutes the entire genome assembly [2].

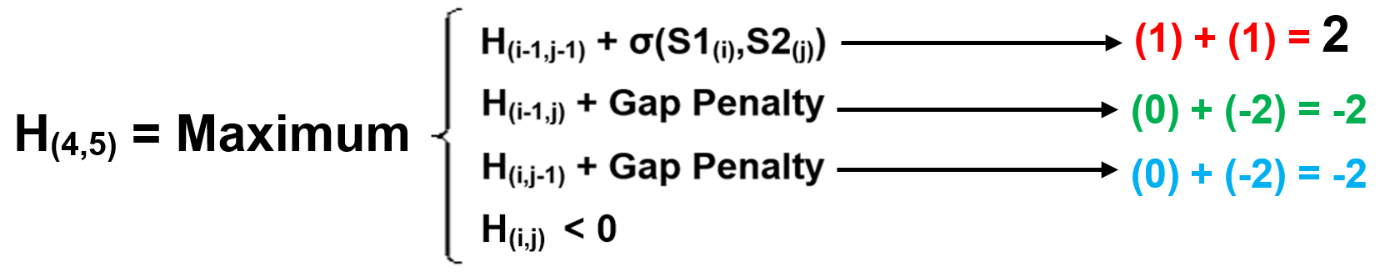
The global Needleman-Wunsch alignment process therefore identifies regions of potentially overlapping sequence termed contigs, which are subsequently concatenated together and assembled onto a single scaffold. However, when large segments of query read sequences are highly repetitive, local alignment is critical to identifying contig overlaps by quantitative determination of the most highly-conserved regions between query sequence reads. A genomic reference map is often used in conjunction with *de novo* assembly to denote the location of contigs within the scaffold of the assembling genome. Local algorithms can use the known distance between each paired query-reference sequence read to map their scaffold position more accurately against reference genomes. A full-dovetail join of perfect local alignment between a query and reference sequence can therefore be used to assemble adjacent contig sequences onto the correct scaffold location for genome assembly.

The quality of the *de novo* assembly can be defined according to the size and continuity of assembled contigs on the scaffold – a common metric of genome assembly quality is N50, equal to the sequence length of the shortest contig that covers fifty-percent of the genome. The greater the number and size of gaps between adjacent contigs, the poorer the genome assembly, which is signified by a low N50 value. However, it is worth note that the N50 metric does not account for the misassembly of merged contigs, therefore an assembly with frequently incorrect contig concatenations can still have a misleadingly high N50 value [3]. Vezzi *et al.* have recently developed a feature-response curve metric which aims to evaluate genome assembly quality more accurately by characterising its quality as a function of the maximum number of potential errors permitted in the contigs [3].

Whilst very-short sequence reads are quicker to align, they are more difficult to reliably merge into contigs that can be assembled onto a scaffold - the statistical significance of very short regions of alignments is weak and may be repeated thousands of times within a genome [4]. The local alignment of short reads against a reference sequence is therefore particularly important to constructing a complete genome assembly; the local Needleman-Wunsch algorithm represents a logical technique to accurately localise short sequence reads onto a genomic scaffold.

To demonstrate the difference between global and local Needleman-Wunsch algorithms, we consider the same query and reference sequences from Figure 2 in a new local matrix (Figure 4); all i=0 and j=0 cells are equal to zero, as the local Needleman-Wunsch algorithm requires introduction of a fourth parameter where negative scores are reset to zero. Local algorithms trace back aligned sequence from the highest scoring cell of the matrix and end when their value becomes zero, unlike the global Needleman-Wunsch algorithm, which proceeds from the bottom-right to the top-left corner of the matrix irrespective of the cell value.

For H(1,1), the maximum value of the three Needleman-Wunsch parameters is -1, derived from subtracting a mismatch penalty of one from the H(0,0) value of zero, as shown below. However, due to the nature of the local Needleman-Wunsch algorithm, this is reset to zero to ensure nearby regions of query-reference alignment are still observed.

H(4,5) has the greatest value of the entire matrix at 2, and therefore acts as the starting point for traceback of an aligned sequence. This value was derived from the match reward score of 1 in addition to the magnitude of H(3,4), in accordance with the top parameter as shown below. Pairwise alignment thus continues from this H(3,4) cell, which was derived from H(2,3). The H(2,3) cell has a value of zero, therefore we stop the traceback. The region of aligned sequence for this simple model is therefore CA – in reality, local Needleman-Wunsch algorithms require more than five nucleotides to display statistically significant regions of similarity.

The Needleman-Wunsch algorithm therefore provides a mechanism for pairwise alignment of query and reference sequences on both a global and a local scale. Global sequence alignment is essential for *de novo* genome assembly, as aligned overlapping contigs can be merged and scaffolded to construct a genome. Local sequence alignment is also relevant for assembling draft genomes, by identifying regions of greatest similarity between highly-conserved sequences. These dynamic Needleman-Wunsch algorithms therefore combine to deconvolute the vast quantity of data produced by DNA sequencing into a rationalised genome.

**References:**

1. de la Torre, L. & Seguel, J. (2010). A Parallel Needleman-Wunsch-Hirschberg Bio-sequence Alignment Algorithm. *CSC,* 269-272.

2. Darling, A. *et al.* (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research,* **14(7):** 1394-403.

3. Vezzi, F., Narzisi, G. & Mishra, B. (2012). Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *PloS one,* **7(12):** e52210.

4. Mitrophanov, A. & Borodovsky, M. (2006). Statistical significance in biological sequence analysis. *Briefings in Bioinformatics,* **7(1):** 2-24.