Julius D. Higiro

Project Proposal

Sequence alignment is a technique in the biological sciences that is used to compare nucleotide sequences in DNA, RNA or amino acid sequences in proteins. The purpose of sequence alignment is to discover ancestral and evolutionary relationships among sequences by the identification of patterns and similarities. The patterns are used to predict and understand the structures and functions of different molecules.

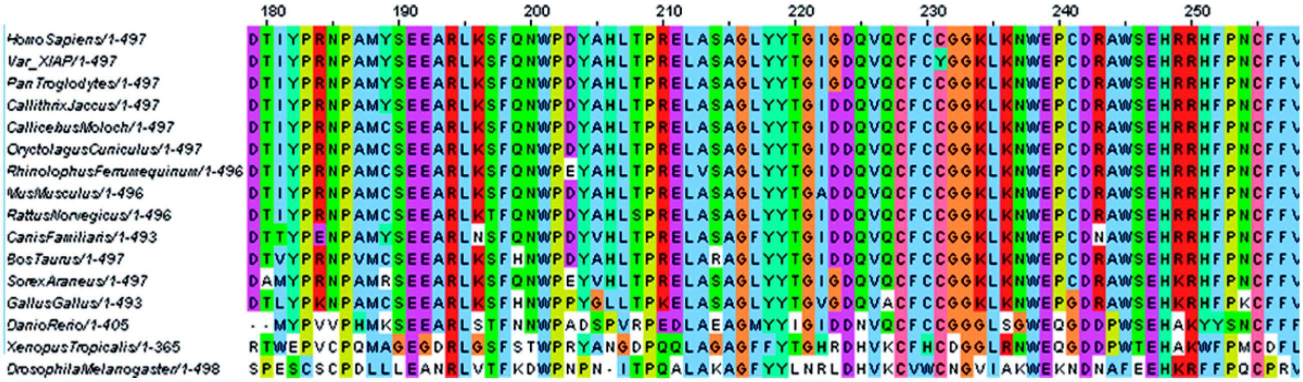

Figure 1. Multiple sequence alignment of the XIAP protein in different species [1]**.**

The above figure displays the sequence alignment of the XIAP protein in several different species. The figure was taken from an article that presented how the application of sequencing alignment was utilized to diagnose and treat a child with intractable inflammatory bowel disease. The first line of the sequence alignment contains the XIAP protein sequence from a human that served as a human reference. The second line of the sequence alignment contains the XIAP protein sequence taken from a child that was later diagnosed with inflammatory bowel disease. The remaining lines are the protein sequences that originated from different species (1).

[1] **Image source:** http://www.nature.com/gim/journal/v13/n3/fig_tab/gim9201146f2.html

In column 231, there is a mutation with the replacement of the amino acid Cysteine represented by C with the amino acid Tyrosine represented by Y in the child's sequence (1). As you can see in the multiple sequence alignment, Cysteine is conserved in the human reference and in the sequences of the other species. The authors blamed this mutation for the child's illness and identified a suitable treatment. This is one of many examples for the importance of sequence alignment in biology.

The Needleman-Wunsch algorithm is a dynamic programming algorithm that is used to generate the optimal sequence alignment for analysis when comparing two sequences (2). The National Center for Biotechnology Information provides a popular tool for scientists that utilize the algorithm to perform nucleotide and amino acid sequence alignments. The problem the algorithm solves is finding the optimum alignment of two nucleotide or amino acid sequences. The algorithm works by building a matrix with an alignment score and using the alignment score to find the optimal path or paths (2).

Let us consider a matrix M[i, j] where i is a row index and j is a column index. A matrix of the form M[6, 6] is presented on page 3 with a comparison of the sequences (A, C, T, T, C and A, C, T, G, C). We utilize the following weights and rules to assign the scores:

Weights:

1. Assign a value of 5 to matching nucleotides or amino acids.

2. Assign a value of -1 to mismatching nucleotides or amino acids.

3. Assign a value of -2 for gaps in the sequences.

Rules:

1. Set equal to zero the left most row/column.

2. Set first row/column in the sub-matrix S[5, 5], a gap value that increases by a factor of 2.

3. The value assigned to the quadrants originates from three directions (top, bottom and diagonal) as depicted in figure 2. The value is the maximum value from computing the following:

Top-down value = M[i-1, j] + gap value.

Right-left value = M[i, j-1] + gap value.

Diagonal value = M[i-1, j-1] + match/mismatch.

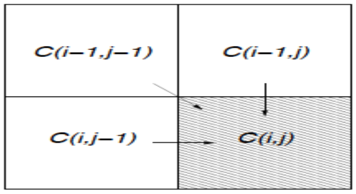

Figure 2. Source of values for computing alignment score [2]**.**

	(col 1)	**A** (col 2)	**C** (col 3)	**T** (col 4)	**G** (col 5)
(row 1)	0	-2	-4	-6	-8
A (row 2)	-2	5 ↘	3 →	1 →	-1 →
C (row 3)	-4	3 ↓	10 ↘	8 →	6 →
T (row 4)	-6	1 ↓	8 ↓	5 ↘	7 ↘
T (row 5)	-8	-1 ↓	6 ↓	13 ↘	11 →

[2] **Image source: Likic V. (2008) The Needleman-Wunsch algorithm for sequence alignment. Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne.**

http://www.science.marshall.edu/murraye/clearer%20matrix%20slide%20show.pdf

Below is an example calculation of the alignment score for M[2, 2]:

Top-down = M[1, 2] + gap = -2 + -2 = -4

Right-left = M[2, 1] + gap = -2 + -2 = -4

Diagonal = M[1, 1] + matching = 0 + 5 = 5 (max)

The maximum value of top-down, right-left and diagonal is 5, so M[2,2] is assigned a score of 5 and it is marked with a directional arrow (diagonal). After assigning the alignment scores, a trace back is performed starting with the last M[i, j] that was assigned a score (i. e. M[5, 5]). A trace back is the process of tracing back from the last assigned score to the preceding value from which it originated (2). As such, the alignment score of 11 in M[5, 5] is traced back to the alignment score of 13 in M[5, 4] from which it originated and so on.

Next, the optimum alignment is generated by aligning the letters of two sequences marked by a diagonal, gapping the nucleotide or amino acid in the left hand sequence of the matrix for quadrants that contain a right-left arrow and gapping the nucleotide or amino acid in the top side sequence of the matrix for quadrants that contain a top-down arrow (2). Utilizing this scheme produces the below alignment without gaps. This particular sequence alignment is too basic of an example to show the various optimal sequence alignments that can be produced.

```
A C - T G
A C T T -
```

Nevertheless, I propose the encoding of the Needleman-Wunsch algorithm using ASP in order to answer the query, what is the optimum sequence alignment when two DNA sequences are provided as input? I believe that ASP is a good choice for representing this problem because the Needleman-Wunsch algorithm is used to solve a combinatorial problem with optimization and ASP is suitable for solving combinatorial search problems.

References:

1. **http://www.nature.com/gim/journal/v13/n3/fig_tab/gim9201146f2.html**

2. Likic V. (2008) The Needleman-Wunsch algorithm for sequence alignment. Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne.