TEACHING THE GENOME GENERATION

Introduction to Sequence Comparison

The Jackson Laboratory
Leading the search for tomorrow's cures

NIH National Institutes of Health
Turning Discovery Into Health

SEPA SCIENCE EDUCATION PARTNERSHIP AWARD
SUPPORTED BY THE NATIONAL INSTITUTES OF HEALTH

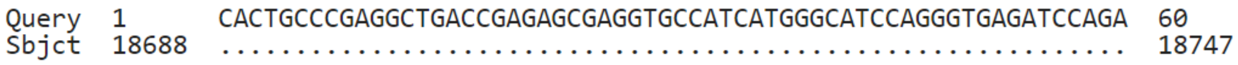
Introduction to Sequence Comparison

Sequence comparison is a technique to determine how similar two or more nucleotide or protein sequences are to each other. As you complete the activity, you will discover different applications of sequence comparison used by scientists.

Sequence Alignment

In sequence comparison, the first step is aligning the sequences. The goal of sequence alignment is to line up the nucleotides or amino acids of each sequence such that there are as many matches as possible. Researchers typically use computer programs, like the online Basic Local Alignment Search Tool (BLAST), to align sequences.

BLAST displays aligned sequences one above the other. The **Query** sequence is typically the sequence of interest and the Subject (**Sbjct**) sequence is typically the sequence you are comparing to.



```
Query  1      CACTGCCCGAGGCTGACCGAGAGCGAGGTGCCATCATGGGCATCCAGGGTGAGATCCAGA  60
Sbjct  18688  ............................................................  18747
```

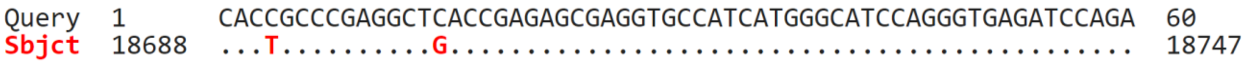
The numbers at the beginning and end of each sequence represent the position of the first and last nucleotide or amino acid within the whole sequence. Dots in the Subject sequence line indicate positions at which the nucleotides or amino acids **match** the Query sequence. In the above example, all the nucleotides in the query and subject sequences match.

Sequence Comparison

After the sequences are aligned, the next step is to identify differences between the sequences. There are two main kinds of differences: mismatches and gaps.

Mismatches

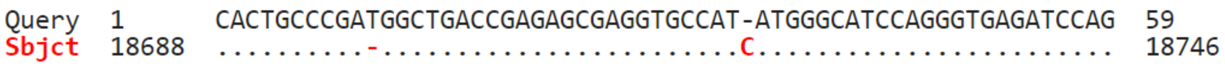
Mismatches are positions where the nucleotides or amino acids in the sequences are not the same, or do not match. Mismatches are represented as letters in the Subject sequence. In the example below, there are two mismatches: the Query sequence has a cytosine (C) at one position where the Subject sequence has a thymine (T), and at another position where the Subject sequence has a guanine (G).



```
Query  1      CACCGCCCGAGGCTCACCGAGAGCGAGGTGCCATCATGGGCATCCAGGGTGAGATCCAGA  60
Sbjct  18688  ...T..........G.............................................  18747
```

Gaps

Gaps are positions where one sequence has one or more nucleotides or amino acids that the other sequence does not have. Gaps are represented as dashes (-). In the example below, there are two gaps: the Query sequence has an extra thymine (T) relative to the Subject sequence, and the Subject sequence has an extra cytosine (C) relative to the Query sequence.



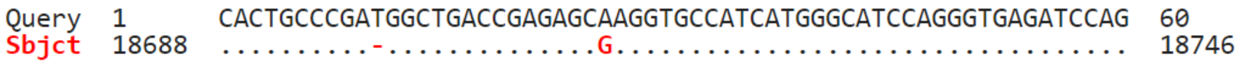
```
Query  1      CACTGCCCGATGGCTGACCGAGAGCGAGGTGCCAT-ATGGGCATCCAGGGTGAGATCCAG  59
Sbjct  18688  ..........-........................C.......................  18746
```

Percent Identity

When comparing sequences, **percent identity** provides a measure of how similar two sequences are. The formula for percent identity uses the total number of nucleotide or amino acid **positions** in the sequence comparison and the number of nucleotide or amino acid positions that are different, or **divergent**, between the sequences:

$$Percent\ Identity = \frac{\#\ positions - \#\ divergent\ positions}{\#\ positions}\ x\ 100\%$$

In the example below, there are two divergent positions when comparing the two sequences.



```
Query  1      CACTGCCCGATGGCTGACCGAGAGCAAGGTGCCATCATGGGCATCCAGGGTGAGATCCAG  60
Sbjct  18688  ..........-..............G..................................  18746
```

In BLAST, different lengths of sequences can be compared. When the alignment is complete, BLAST will display the alignment such that each full line within a sequence comparison has 60 total positions, regardless of the number of nucleotides in each individual sequence. In the sequence comparison above, Query has 60 nucleotides and Subject has 59 nucleotides and one gap. In total there are 60 positions being compared. Two of those positions are divergent (one gap and one mismatch).

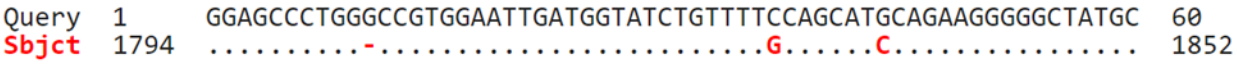
The percent identity for this comparison could be calculated as follows:

$$Percent\ Identity = \frac{60 - 2}{60}\ x\ 100\% = \frac{58}{60}\ x\ 100\% = 96.67\%$$

Knowledge Check

1. What is the difference between a mismatch and a gap?

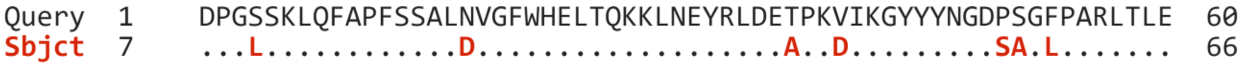
2. What is the percent identity for the following comparison of 60 nucleotide positions?



```
Query  1     GGAGCCCTGGGCCGTGGAATTGATGGTATCTGTTTTCCAGCATGCAGAAGGGGGCTATGC  60
Sbjct  1794  ..........-.........................G......C................  1852
```

3. You have two sequences you want to compare with BLAST. Your Query sequence is 100 nucleotides long, and your Subject sequence is 70 nucleotides long. When you compare the two sequences, you find that all 70 nucleotides from the Subject sequence exactly match the first 70 nucleotides from the Query sequence. Using 100 nucleotides as the total number of positions, what is the percent identity between these two sequences?

4. The sequence comparison below shows two aligned amino acid sequences, with the amino acids abbreviated as single letters. What is the percent identity for this comparison of 60 amino acid positions?



```
Query  1    DPGSSKLQFAPFSSALNVGFWHELTQKKLNEYRLDETPKVIKGYYYNGDPSGFPARLTLE  60
Sbjct  7    ...L............D...................A..D.........SA.L.......  66
```