Sequence alignment

An alignment of DNA or protein sequences X and Y is a new pair X' and Y' such that

1. Both X' and Y' are of equal length, and
2. Removal of "gaps" from X' and Y' yields the original sequences X and Y.

For example if X = ACCG and Y = AGAACCGG then

```
X' = ACCG----
Y' = AGAACCGG
```

is an alignment of X and Y. However,

```
X' = ACCG--           and   X' = ACCG----
Y' = AGAACCGG               Y' = AGCCAAGG
```

are both not alignments of X and Y.

There are many different alignments of a given pair of sequences X and Y. We can evaluate the quality of an alignment by computing its score. The score of an alignment is just the sum of the scores of individually aligned pairs of nucleotides and gaps. For example let the score of a "match" be 8, the score of a mismatch be 2, and the score of a gap be -2. Then the score of the alignment

```
                X' = ACCG----
                Y' = AGAACCGG
is given by     score= 8+2+2+2-2-2-2-2=6
```

However, the alignment below has more matches and thus a better score of 24

```
 X' = ---ACCG-
 Y' = AGAACCGG
score= -2-2-2+8+8+8+8-2=24
```

Instead of match and mismatch scores we can assign specific matches and mismatches different scores. For example suppose the score of A aligned to A is 8, C aligned to C is 6, G aligned to G is 7, T aligned to T is 5, mismatch is 2, and gap is -2. Then the score of the above alignment becomes 19 as shown below.

```
 X' = ---ACCG-
 Y' = AGAACCGG
score= -2-2-2+8+6+6+7-2=19
```

Alignments play an important role in genomics. Among many other applications they are useful in identifying conserved regions of a protein or the genome. These conserved regions usually play an important functional role since they have been preserved across years of evolution.