Further Information on alignments

1. Summary - an overview of steps in modern biology:

Step 1: Cell → extract DNA → Analysis (eg SNP, PCR, RF:P etc) and / or clone (gene library).

Step 2: From gene library undertake sequencing (gene or genome)

Step 3: Contig assembly of sequence fragments to produce a contiguous sequence.

Step 4: Sequence analysis:

(a) Inspecting the sequence and data base searching- six frame translation, G+C content etc
(b) Infer homology – calculate probability that sequences have a common ancestor (search against database) - use BLAST)
(c) Identify ancestrally related sites in a pair of sequences, assuming that they have a common ancestor – changes over evolutionary period - mutations (deletions or insertions) – use eg Pfam, PRINTS, BLOCKS etc

2. Aligning sequences:

You have two sequences:

Seq A: ATC and
Seq B: AGC

We cant get a perfect match but we can insert blank spaces to represent symbols that might once have been in the sequences but were later deleted. These blanks are written as - (aka inde – a deletion or insertion)

Some possible alignments are….:

SeqA: ATC	SeqA:AT-C	SeqA: A-TC	SeqA:A--TC
SeqB: AGC	SeqB:A-GC	SeqB: AG-C	SeqB:AG--C
(a)	(b)	(c)	(d)

But it is not obvious how to choose between them.

There are two types of clashes: a direct clash [such as T against G, in example (a)] and an indel clash in which a letter clashes with a blank [such as – against G, in example (b)].

3. Ways of measuring alignment matches (scoring methods)

There seem to be two ways of scoring matches:

1. To create a score that measures the "distance between" the two sequences. For distance measurement, the ideal score will be 0 with high positive scores representing different sequences. No negative scores will be possible.

2. To create a score that measures the similarity of the two sequences. For this high scores are good and low (even negative scores) will be bad.

Unfortunately, both ways of measuring matches exist in the literature. You will see at the end of the exercise that by changing the scoring scheme you can change the "best" alignment. The word "best" is rather problematic.

4. The very simplest way to score an alignment- distant scores.

The following exercise relates for scoring <u>distances</u>. We set a score for how well each letter in SeqA matches the corresponding letter in SeqB.

The definition we will use is:

s(x,y) = 0 If x and y match <u>and</u>
s(x,y) = 1 if x and y are different

Once we have scored each letter of SeqA with the corresponding letter of SeqB, we get the total score for the possible alignment by adding up these individual scortes.

We can now compare the score for the SeqA and SeqB:

```
SeqA: ATC          SeqA: AT-C         SeqA: A-TC         SeqA: A-TCA
SeqB: AGC          SeqB: A-GC         SeqB: AG-C         SeqB: AG-CA
Score:010=1              0110=2             0110=2             01010=2
    (a)                (b)                (c)                (d)
```

The best score is 1. Note that scoring an indel against an indel is redundant and is usually not considered

5. Making the scoring complicated:

A more complicated score will be to give different penalties to clashes (like A against T) as opposed to a letter against an indel (like A aginst -). Extending this we could have different scores for different types of clashes (eg A vs C might be different from A vs T).

This is also referred to as assigning WEIGHTS. The applications of WEIGHTS for scoring is useful for scoring if we have an idea of the frequency of mutation at particular site(s) for the genes under study. So, WEIGHTS can be heavily weighted for one nucleotide in which the nucleotide mutates at a lower rate than to another, which mutates readily. Can you think to which weights could / should be applied?

Now suppose we use WEIGHTS for the examples of SeqA and SeqB, so that a direct clash scores as 2 and an indel clash scores as 1. In this cas, all 3 alignments have a score of 2.

```
SeqA: ATC          SeqA:AT-C          SeqA: A-TC
SeqB: AGC          SeqB:A-GC          SeqB: AG-C
Score: 020=2             0110=2             0110=2
       (a)                (b)                (c)
```

<u>Exercise:</u>

1. Align the two sequences and compute a distance score:
   SeqA:TCAGACCGATTG
   SeqB: TCGGAGCTG
2. A number of alignment programs allow the user to change WEIGHTS. Can you search on the web for these programs and provide a list?

6. Dynamic programming for sequence alignments- Similarity scoring method

A dynamic programming matrix is a diagrammatic way of representing all possible alignments of two sequences. A table is generated which seeks to find the best alignments as more pieces from the two sequences are added. This will be explained later.