Matlab Group Project Hints

For those who are doing the gene sequencing by producing two vectors of start codon positions and end codon position (i.e. using "regexp" or "strfind"), following are some hints.

You can do the script using three while loops rather than for loops as the exact number of iterations is not known. However by using while loops you will need to have three of your own counters; one counter for the start index, one counter for the end indexing and one for the gene number

Set all the counters to 1 and define two new vectors to store the start gene position and the end gene position (any size as we do not know how big they will be). Set the initial value of gene start vector to first start codon index as well as set the gene end vector initial vale to the first end codon index.

The first while loop is used to make sure your start index counter and you end index counter do not go bigger than the size of the respective vector they are indexing – when that happens you need to end.

Inside this while loop, the next while loop check to find the next end codon index that is larger than the current start codon index. It may look like this:

```
% Make next end index larger than the current start index
while EndCodon_Vector(end_codon_counter) < GeneStart_Vector(gene_counter)
    end_codon_counter = end_codon_counter + 1;
end
GeneEnd(gene_counter) = End_Index(end_codon_counter);
```

Now increase your gene counter as we assume we have got to the end of the first gene:

```
gene_counter = gene_counter + 1;  %we found the end so increase the gene counter
```

With the next while loop make sure the next start codon is after the last end codon. It may look like:

```
% Make sure the next start is the first one after the previous end
while Start_Index(start_codon_counter) < GeneEnd(gene_counter - 1)
    start_codon_counter = start_codon_counter + 1;
end
GeneStart(gene_counter) = Start_Index(start_codon_counter);
```

This will also be the end of the first while loop.

Notice that as I used while loops I had to increase my counters by adding '1' to them each time.

One problem with this method is that the last end codon may have been omitted, however you can add some code at the end to check if there is one last end codon.

With the long *sequence_long.txt* I get 5810 gene pairs using both methods – first 6 being:

74 → 75; 89 → 131, 279 → 282, 399 → 451; 461 → 462; 711 → 724

If you allowed for the fact that a stop must be at least 3 nucleotides after a start then I get 5467 combinations. The first 6 being:

74→ 131, 279→ 282, 399→ 451, 461→ 491, 711→ 724, 797→ 807