GTPB

The Gulbenkian Training Programme in Bioinformatics
(Since 1999)

Pedro Fernandes, Organiser

INSTITUTO
GULBENKIAN
DE CIÊNCIA

ELB18F

Entry Level Bioinformatics

19-23 February 2018

(First 2018 run of this Course)

Basic Bioinformatics Sessions

Practical 2: Pairwise Sequence Alignment

Tuesday 20 February 2018

Sensitive Pairwise Alignment

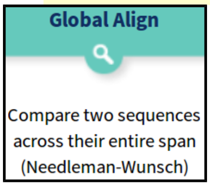
The purpose of this exercise is to look at some aspects of **Pairwise Sequence Alignment** using the most accurate methods available.

As hopefully has been discussed, sequences can be aligned using a **global** strategy, in which the two sequences being aligned are assumed to be homologous from end to end, or using a **local** approach, in which the sequences are assumed to just have homologous region(s).

Global Pairwise Sequence Comparison

First the **global** approach. In a previous exercise, you already have used the **blast** facility at the **NCBI** to perform crude pairwise alignment. **blast** also offers a sensitive option, so maybe that would be a good place to start.

So, once more to the NCBI home page (**http://www.ncbi.nlm.nih.gov/**). From there chose BLAST from the **Popular Resources** list. Scroll down to the **Specialized searches** section and chose the option.



Global Align

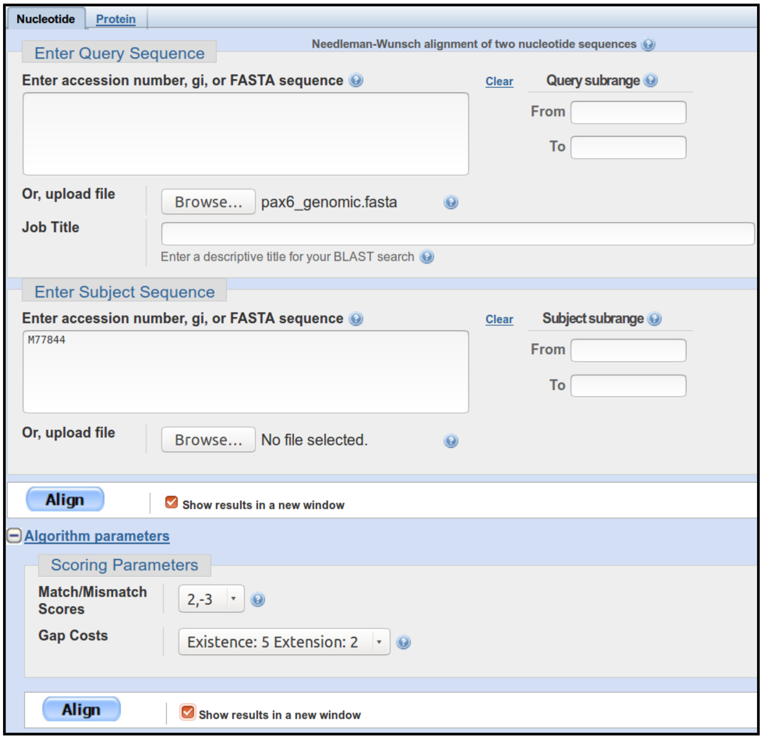
Compare two sequences across their entire span (Needleman-Wunsch)

A choice of settings for **Nucleotide** or **Protein** alignment is offered. As we are going to investigate the alignment of DNA sequences, the default choice is fine. For the first sequence, browse for the file **pax6_genomic.fasta**, which you created when looking at **Ensembl**. It contains the region of **Chromosome 11** containing the entire **PAX6** gene (with a few extra base pairs either end).

To specify the second sequence, you could load the file **pax6_mrna.fasta**, but just typing the corresponding **Accession** code in the appropriate box seems far more sophisticated, so that is what I chose to do.

Open the **Algorithmic Parameters** section, and see that they are as one might expect. The defaults are fine here as the alignment to be computed is trivial (given the way **blast** will go about the task), so anything not outrageous should work.

Ask to **Show results in a new window** and then click on the **Align** button.



Needleman-Wunsch alignment of two nucleotide sequences

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Query subrange

From

To

Or, upload file pax6_genomic.fasta

Job Title

Enter a descriptive title for your BLAST search

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence

M77844

Subject subrange

From

To

Or, upload file No file selected.

☑ Show results in a new window

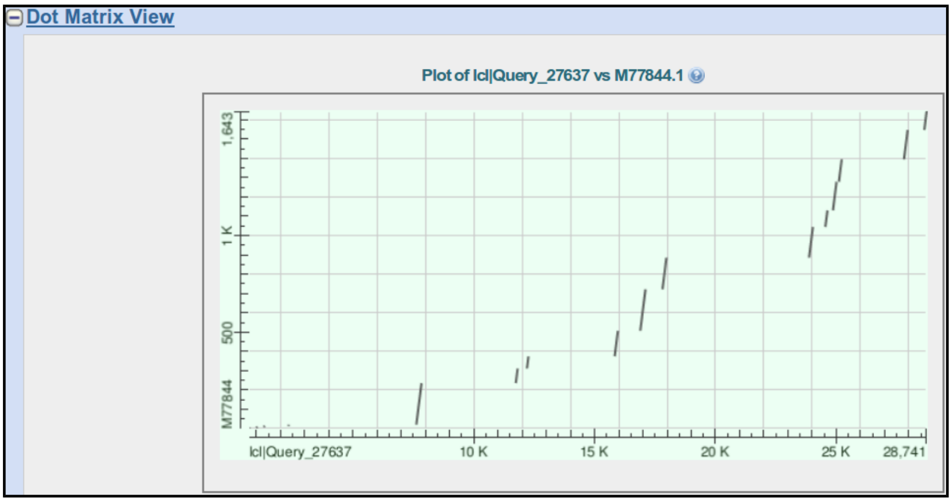
Scoring Parameters

Match/Mismatch Scores: 2,-3

Gap Costs: Existence: 5 Extension: 2

☑ Show results in a new window

After some significant Rollin' and Tumblin' **blast** will proclaim its lyrical conclusions. First examine the **Dot Matrix View**. This sort of representation has rather gone out of fashion in recent years. A shame, I say, this picture represents such a succinct summary of what should be expected of the textual alignment(s) that are the "real" detailed output of this sort of program.



How would you interpret this picture?

What do the diagonal(ish) lines represent?

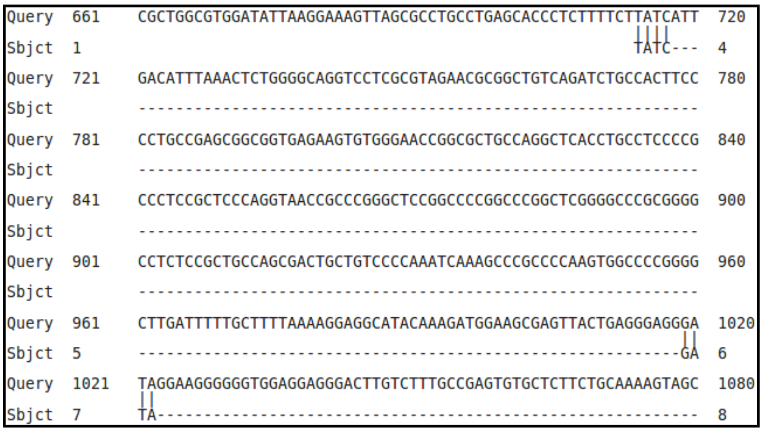
What are the gaps in between the lines?

Which axis represents the genomic sequence and which the mrna?

Move down to the textual alignment. There are some weird little bits and pieces at the front of the alignment which defy logic. I decide not to dwell on these to much, beyond noting that the mRNA has some odd bases at the front.

Also, I have faith that the alignment you look at yields the highest alignment score, but equally. I doubt most **people** would have chosen to throw these odd bases about with quite such abandon! **People** are best!

You can just see evidence of the little patches of whimsy in the **Dot Matrix View**.



```
Query  661    CGCTGGCGTGGATATTAAGGAAAGTTAGCGCCTGCCTGAGCACCCTCTTTTCTTATCATT  720
                                                                   ||||
Sbjct  1                                                           TATC---  4
Query  721    GACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCTGCCACTTCC  780
Sbjct         ------------------------------------------------------------
Query  781    CCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCG  840
Sbjct         ------------------------------------------------------------
Query  841    CCCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGG  900
Sbjct         ------------------------------------------------------------
Query  901    CCTCTCCGCTGCCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGG  960
Sbjct         ------------------------------------------------------------
Query  961    CTTGATTTTTGCTTTTAAAAGGAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGA  1020
                                                                        ||
Sbjct  5      ----------------------------------------------------------GA  6
Query  1021   TAGGAAGGGGGGTGGAGGAGGGACTTGTCTTTGCCGAGTGTGCTCTTCTGCAAAAGTAGC  1080
              ||
Sbjct  7      TA----------------------------------------------------------  8
```

Moving down there are a series of far more convincing near perfect alignments.

You must know what these aligned regions represent by now?

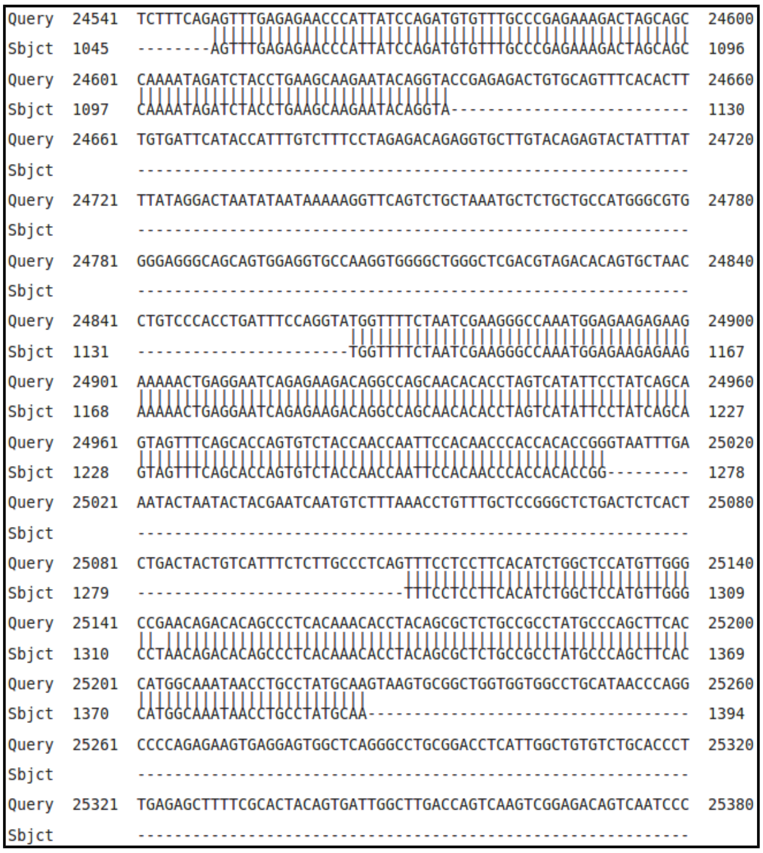
But, just in case:

What do you suppose these regions represent?

How many are there and do they correspond nicely to the lines of the **Dot Matrix View**?

How many exons would you say this mrna has?

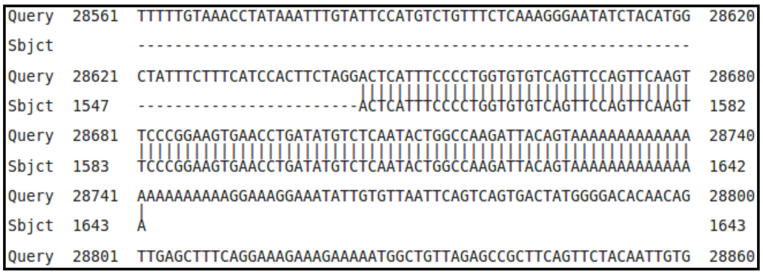
If one was to forgive the strange “bits” at the start, would you say **blast** seems to have done a reasonable job here?



```
Query  24541  TCTTTCAGAGTTTGAGAGAACCCATTATCCAGATGTGTTTGCCCGAGAAAGACTAGCAGC  24600
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1045   --------AGTTTGAGAGAACCCATTATCCAGATGTGTTTGCCCGAGAAAGACTAGCAGC  1096
Query  24601  CAAAATAGATCTACCTGAAGCAAGAATACAGGTACCGAGAGACTGTGCAGTTTCACACTT  24660
              ||||||||||||||||||||||||||||||||||
Sbjct  1097   CAAAATAGATCTACCTGAAGCAAGAATACAGGTA--------------------------  1130
Query  24661  TGTGATTCATACCATTTGTCTTTCCTAGAGACAGAGGTGCTTGTACAGAGTACTATTTAT  24720
Sbjct         ------------------------------------------------------------
Query  24721  TTATAGGACTAATATAATAAAAAGGTTCAGTCTGCTAAATGCTCTGCTGCCATGGGCGTG  24780
Sbjct         ------------------------------------------------------------
Query  24781  GGGAGGGCAGCAGTGGAGGTGCCAAGGTGGGGCTGGGCTCGACGTAGACACAGTGCTAAC  24840
Sbjct         ------------------------------------------------------------
Query  24841  CTGTCCCACCTGATTTCCAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAG  24900
                                     |||||||||||||||||||||||||||||||||||||
Sbjct  1131   -----------------------TGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAG  1167
Query  24901  AAAAACTGAGGAATCAGAGAAGACAGGCCAGCAACACACCTAGTCATATTCCTATCAGCA  24960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1168   AAAAACTGAGGAATCAGAGAAGACAGGCCAGCAACACACCTAGTCATATTCCTATCAGCA  1227
Query  24961  GTAGTTTCAGCACCAGTGTCTACCAACCAATTCCACAACCCACCACACCGGGTAATTTGA  25020
              |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1228   GTAGTTTCAGCACCAGTGTCTACCAACCAATTCCACAACCCACCACACCGG---------  1278
Query  25021  AATACTAATACTACGAATCAATGTCTTTAAACCTGTTTGCTCCGGGCTCTGACTCTCACT  25080
Sbjct         ------------------------------------------------------------
Query  25081  CTGACTACTGTCATTTCTCTTGCCCTCAGTTTCCTCCTTCACATCTGGCTCCATGTTGGG  25140
                                           |||||||||||||||||||||||||||||||
Sbjct  1279   -----------------------------TTTCCTCCTTCACATCTGGCTCCATGTTGGG  1309
Query  25141  CCGAACAGACACAGCCCTCACAAACACCTACAGCGCTCTGCCGCCTATGCCCAGCTTCAC  25200
              || |||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1310   CCTAACAGACACAGCCCTCACAAACACCTACAGCGCTCTGCCGCCTATGCCCAGCTTCAC  1369
Query  25201  CATGGCAAATAACCTGCCTATGCAAGTAAGTGCGGCTGGTGGTGGCCTGCATAACCCAGG  25260
              |||||||||||||||||||||||||
Sbjct  1370   CATGGCAAATAACCTGCCTATGCAA-----------------------------------  1394
Query  25261  CCCCAGAGAAGTGAGGAGTGGCTCAGGGCCTGCGGACCTCATTGGCTGTGTCTGCACCCT  25320
Sbjct         ------------------------------------------------------------
Query  25321  TGAGAGCTTTTCGCACTACAGTGATTGGCTTGACCAGTCAAGTCGGAGACAGTCAATCCC  25380
Sbjct         ------------------------------------------------------------
```

I think I would.

The final alignment section even has a poly A tail!



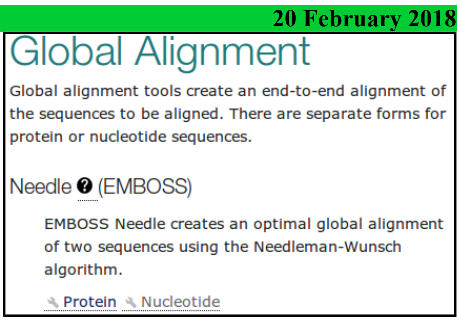
```
Query  28561  TTTTTGTAAACCTATAAATTTGTATTCCATGTCTGTTTCTCAAAGGGAATATCTACATGG  28620
Sbjct         ------------------------------------------------------------
Query  28621  CTATTTCTTTCATCCACTTCTAGGACTCATTTCCCCTGGTGTGTCAGTTCCAGTTCAAGT  28680
                                      ||||||||||||||||||||||||||||||||||||
Sbjct  1547   ------------------------ACTCATTTCCCCTGGTGTGTCAGTTCCAGTTCAAGT  1582
Query  28681  TCCCGGAAGTGAACCTGATATGTCTCAATACTGGCCAAGATTACAGTAAAAAAAAAAAAA  28740
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1583   TCCCGGAAGTGAACCTGATATGTCTCAATACTGGCCAAGATTACAGTAAAAAAAAAAAAA  1642
Query  28741  AAAAAAAAAAGGAAAGGAAATATTGTGTTAATTCAGTCAGTGACTATGGGGACACAACAG  28800
              |
Sbjct  1643   A                                                             1643
Query  28801  TTGAGCTTTCAGGAAAGAAAGAAAAATGGCTGTTAGAGCCGCTTCAGTTCTACAATTGTG  28860
```

Wonderful, but it is not safe to assume that just selecting any service that claims to do a sensitive global pairwise alignment will just work for any pair of sequences. I fact, pretty though it appears, the alignment **blast** has generated is not as entirely logical as it might first seem. For example, consider:

How might the gap around **24,750** in the genomic sequence been positioned more intelligently?

Next, try aligning the same two sequences with another program (implementing the same algorithm) at the **EBI**.

Go to the **Pairwise Sequence Alignment EBI page (http://www.ebi.ac.uk/Tools/psa/)**. From there, select the **Nucleotide** option for the **Global Alignment** program **Needle**. **Needle** implements the best global pairwise algorithm faithfully.

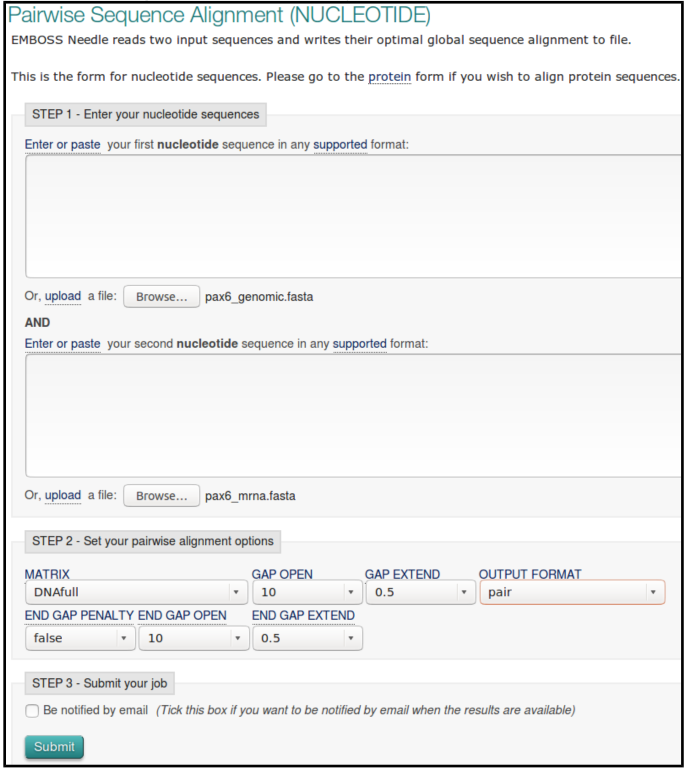


Global Alignment

Global alignment tools create an end-to-end alignment of the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

Needle ❓ (EMBOSS)

EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.



Pairwise Sequence Alignment (NUCLEOTIDE)

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.

This is the form for nucleotide sequences. Please go to the protein form if you wish to align protein sequences.

STEP 1 - Enter your nucleotide sequences

Enter or paste your first **nucleotide** sequence in any supported format:

Or, upload a file: Browse... pax6_genomic.fasta

AND

Enter or paste your second **nucleotide** sequence in any supported format:

Or, upload a file: Browse... pax6_mrna.fasta

STEP 2 - Set your pairwise alignment options

MATRIX	GAP OPEN	GAP EXTEND	OUTPUT FORMAT
DNAfull	10	0.5	pair

END GAP PENALTY	END GAP OPEN	END GAP EXTEND
false	10	0.5

STEP 3 - Submit your job

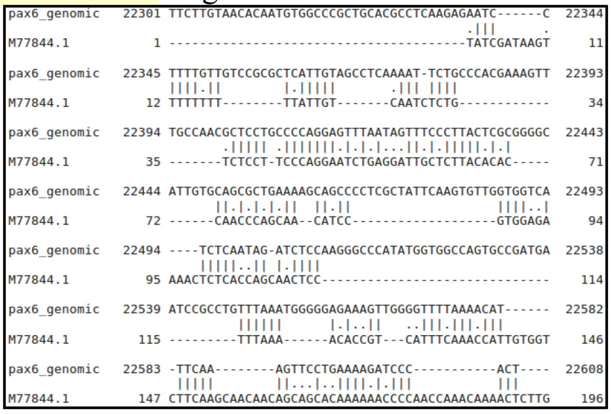
□ Be notified by email (*Tick this box if you want to be notified by email when the results are available*)

Load up the first sequence from **pax6_genomic.fasta**.

Load up the second sequence from **pax6_mrna.fasta**.

Click on the **More options** button to see what parameters you can set. They should be as you might expect. The defaults are fine for the first run.

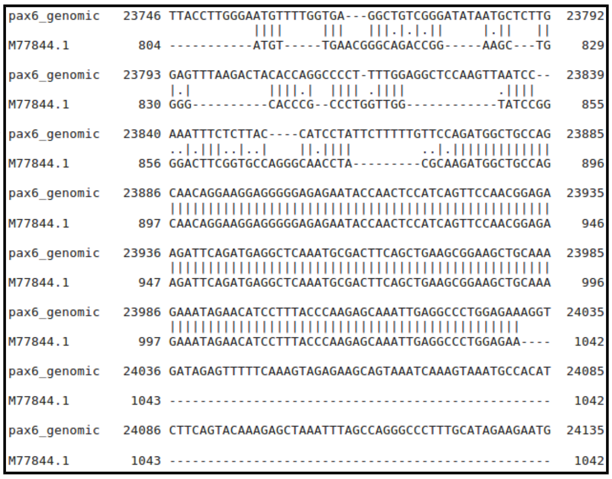
Click on the **Submit** button to get **Needle** into action.



```
pax6_genomic   22301 TTCTTGTAACACAATGTGGCCCGCTGCACGCCTCAAGAGAATC------C  22344
                                                            .|||      .
M77844.1           1 ---------------------------------------TATCGATAAGT     11

pax6_genomic   22345 TTTTGTTGTCCGCGCTCATTGTAGCCTCAAAAT-TCTGCCCACGAAAGTT  22393
                     ||||.||        |.|||||       .||| ||||
M77844.1          12 TTTTTTT--------TTATTGT-------CAATCTCTG------------     34

pax6_genomic   22394 TGCCAACGCTCCTGCCCCAGGAGTTTAATAGTTTCCCTTACTCGCGGGGC  22443
                            .||||| .|||||||.|.|.|...||.|.|||||.|.|
M77844.1          35 -------TCTCCT-TCCCAGGAATCTGAGGATTGCTCTTACACAC-----     71

pax6_genomic   22444 ATTGTGCAGCGCTGAAAAGCAGCCCCTCGCTATTCAAGTGTTGGTGGTCA  22493
                           ||.|.|.|.||  ||.||                   ||||..|
M77844.1          72 ------CAACCCAGCAA--CATCC-------------------GTGGAGA     94

pax6_genomic   22494 ----TCTCAATAG-ATCTCCAAGGGCCCATATGGTGGCCAGTGCCGATGA  22538
                         |||||..|| |.||||
M77844.1          95 AAACTCTCACCAGCAACTCC------------------------------    114

pax6_genomic   22539 ATCCGCCTGTTTAAATGGGGGAGAAAGTTGGGGTTTTAAAACAT------  22582
                              ||||||      |.|..||   ..|||.|||.|||
M77844.1         115 ---------TTTAAA------ACACCGT---CATTTCAAACCATTGTGGT    146

pax6_genomic   22583 -TTCAA--------AGTTCCTGAAAAGATCCC-----------ACT----  22608
                      |||||        ||...|..||||.|.|||           |||
M77844.1         147 CTTCAAGCAACAACAGCAGCACAAAAAACCCCAACCAAACAAAACTCTTG    196
```

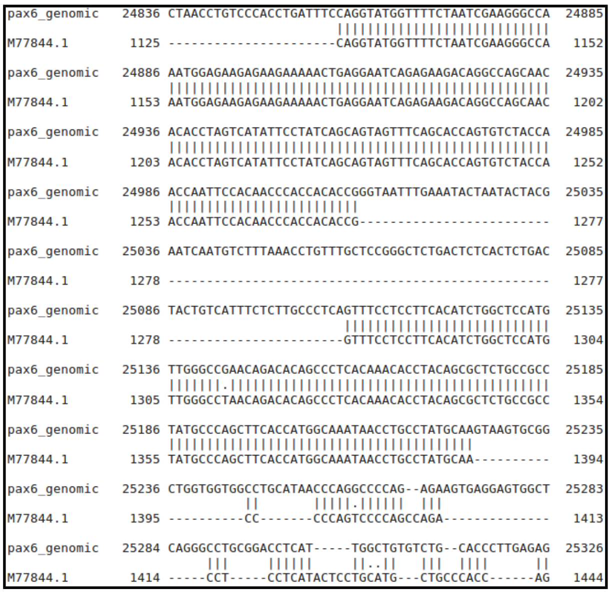
Well! Nothing like as convincing as the alignment **blast** produced!

Alignment does not even begin until over **22,300** base pairs along the genomic sequence. Even then it is not convincing, as in ***wrong***, if we accept the results already obtained from **blast** as a fair approximation of the truth.



```
pax6_genomic   23746 TTACCTTGGGAATGTTTTGGTGA---GGCTGTCGGGATATAATGCTCTTG  23792
                                |||||     |||   |||.|.|.||     |.||   ||
M77844.1         804 -----------ATGT-----TGAACGGGCAGACCGG-----AAGC---TG    829

pax6_genomic   23793 GAGTTTAAGACTACACCAGGCCCCT-TTTGGAGGCTCCAAGTTAATCC--  23839
                     |.|          ||||.|  |||| .||||            .||||
M77844.1         830 GGG----------CACCCG--CCCTGGTTGG------------TATCCGG    855

pax6_genomic   23840 AAATTTCTCTTAC----CATCCTATTCTTTTTGTTCCAGATGGCTGCCAG  23885
                     ..|.|||..|..|    ||.||||         ..|.|||||||||||||
M77844.1         856 GGACTTCGGTGCCAGGGCAACCTA---------CGCAAGATGGCTGCCAG    896

pax6_genomic   23886 CAACAGGAAGGAGGGGGAGAATACCAACTCCATCAGTTCCAACGGAGA  23935
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1         897 CAACAGGAAGGAGGGGGAGAATACCAACTCCATCAGTTCCAACGGAGA    946

pax6_genomic   23936 AGATTCAGATGAGGCTCAAATGCGACTTCAGCTGAAGCGGAAGCTGCAAA  23985
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1         947 AGATTCAGATGAGGCTCAAATGCGACTTCAGCTGAAGCGGAAGCTGCAAA    996

pax6_genomic   23986 GAAATAGAACATCCTTTACCCAAGAGCAAATTGAGGCCCTGGAGAAAGGT  24035
                     ||||||||||||||||||||||||||||||||||||||||||||||
M77844.1         997 GAAATAGAACATCCTTTACCCAAGAGCAAATTGAGGCCCTGGAGAA----   1042

pax6_genomic   24036 GATAGAGTTTTTCAAAGTAGAGAAGCAGTAAATCAAAGTAAATGCCACAT  24085

M77844.1        1043 --------------------------------------------------   1042

pax6_genomic   24086 CTTCAGTACAAAGAGCTAAATTTAGCCAGGGCCCTTTGCATAGAAGAATG  24135

M77844.1        1043 --------------------------------------------------   1042
```

There are some well aligned regions after genomic position **24,500**.



```
pax6_genomic   24836 CTAACCTGTCCCACCTGATTTCCAGGTATGGTTTTCTAATCGAAGGGCCA  24885
                                           ||||||||||||||||||||||||||||
M77844.1        1125 ----------------------CAGGTATGGTTTTCTAATCGAAGGGCCA   1152

pax6_genomic   24886 AATGGAGAAGAGAAGAAAAACTGAGGAATCAGAGAAGACAGGCCAGCAAC  24935
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1        1153 AATGGAGAAGAGAAGAAAAACTGAGGAATCAGAGAAGACAGGCCAGCAAC   1202

pax6_genomic   24936 ACACCTAGTCATATTCCTATCAGCAGTAGTTTCAGCACCAGTGTCTACCA  24985
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1        1203 ACACCTAGTCATATTCCTATCAGCAGTAGTTTCAGCACCAGTGTCTACCA   1252

pax6_genomic   24986 ACCAATTCCACAACCCACCACACCGGGTAATTTGAAATACTAATACTACG  25035
                     |||||||||||||||||||||||||
M77844.1        1253 ACCAATTCCACAACCCACCACACCG-------------------------   1277

pax6_genomic   25036 AATCAATGTCTTTAAACCTGTTTGCTCCGGGCTCTGACTCTCACTCTGAC  25085

M77844.1        1278 --------------------------------------------------   1277

pax6_genomic   25086 TACTGTCATTTCTCTTGCCCTCAGTTTCCTCCTTCACATCTGGCTCCATG  25135
                                            |||||||||||||||||||||||||||
M77844.1        1278 -----------------------GTTTCCTCCTTCACATCTGGCTCCATG   1304

pax6_genomic   25136 TTGGGCCGAACAGACACAGCCCTCACAAACACCTACAGCGCTCTGCCGCC  25185
                     |||||||.||||||||||||||||||||||||||||||||||||||||||
M77844.1        1305 TTGGGCCTAACAGACACAGCCCTCACAAACACCTACAGCGCTCTGCCGCC   1354

pax6_genomic   25186 TATGCCCAGCTTCACCATGGCAAATAACCTGCCTATGCAAGTAAGTGCGG  25235
                     ||||||||||||||||||||||||||||||||||||||||
M77844.1        1355 TATGCCCAGCTTCACCATGGCAAATAACCTGCCTATGCAA----------   1394

pax6_genomic   25236 CTGGTGGTGGCCTGCATAACCCAGGCCCCAG--AGAAGTGAGGAGTGGCT  25283
                               ||       |||||.||||||  |||
M77844.1        1395 ----------CC-------CCCAGTCCCCAGCCAGA--------------   1413

pax6_genomic   25284 CAGGGCCTGCGGACCTCAT-----TGGCTGTGTCTG--CACCCTTGAGAG  25326
                          |||     ||||||     ||..||   |||  ||||      ||
M77844.1        1414 -----CCT-----CCTCATACTCCTGCATG---CTGCCCACC------AG   1444
```

Then a resumption of chaos after **25,230** or so.

How many convincingly aligned regions did you see?

How many did you expect?

Clearly, this alignment is not correct. Can you explain why?

I assume you have all read the lucid answers to the question above? If so, I am confident you will agree that there are **3** ways to get an answer, similar to that generated by **blast**, from the tools offered at the **EBI**. They are:

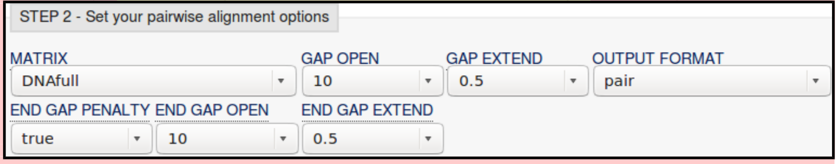
- Make gap penalties so cheap that **Needle** will have no excuse to avoid gaps where they are needed. This works if you use a gap opening penalty of **1.0** (the lowest allowed by the web interface) and a gap extension penalty of **0.0**, allowed by the program ***but not by the EBI web interface!!*** The lowest value the web interface allows is **0.0005**, which really should be sufficiently small, but provably is not. The most important question being "*Why would a web interface restrict a program's capabilities other than to prevent excessive resource use?*". I have no answer for that one, I will just petulantly include some extra low gap alignments (made without a web interface) in your **Backup_Results** directory and retire with self righteous hauteur! Note that making gaps completely free (i.e. both gap **opening** and **extension** equal to **0.0**) will not work at all! **needle** would simply match each base of the mRNA with the next identical base of the genomic sequence until it runs out of letters. You could do this from the command line, but it would clearly not make sense.

  Actually, using gap penalties to suit huge gaps that are really introns, will only work when the exons are so similar (as here) that any gap penalties will work for their alignment. Generally, you need to pick gap penalties to optimise exon alignment. So this is a very horrible way to "fix" the situation anyway.

- Tell **Needle** to penalise the gaps it puts at either end of the alignment in the same way it penalises gaps it puts in the middle. By default, end gaps are free!! Which is not very logical here. This ***is*** possible using the website.

- Use **Stretcher**, which uses essentially the same algorithm as **Needle**, except, it also applies a bit of common sense (**heuristics**, if you like). **Stretcher** takes a look at the sequences before it starts to do any serious computation. It identifies any "*good regions*" (all **12** exon matches in this case) and then says "*OK, I am definitely having those, how best can I deal with the rest?*". In essence, **Stretcher** does a quick **Dot Matrix View** before it starts and so only goes to work when it has a pretty good idea what the answer should look like It works in this case, but not always. **Stretcher** is faster than **Needle** but does not necessarily generate the highest scoring alignment. **Stretcher** works in a fashion far closer to the way a human would work, which has to be good! Well, usually anyway.

So, try the **Needle** with penalised **End Gaps** approach by returning to the **Needle** launch page from your results. You should find the two sequences are still selected, so you should only have to click on **More Options** again and change the **END GAP PENALTY** field from **false** to **true**.



STEP 2 - Set your pairwise alignment options

MATRIX	GAP OPEN	GAP EXTEND	OUTPUT FORMAT
DNAfull	10	0.5	pair

END GAP PENALTY	END GAP OPEN	END GAP EXTEND
true	10	0.5

Click on the **Submit** button and **Needle** will be on the road again.

How many matching regions are there this time?

Is the count **now** roughly as you would expect?

Finally, check that **Stretcher** works as expected.

Go again to the **Pairwise Sequence Alignment EBI** page (**http://www.ebi.ac.uk/Tools/psa/**).

From there, select the **Nucleotide** option for the **Global Alignment** program **Stretcher**.

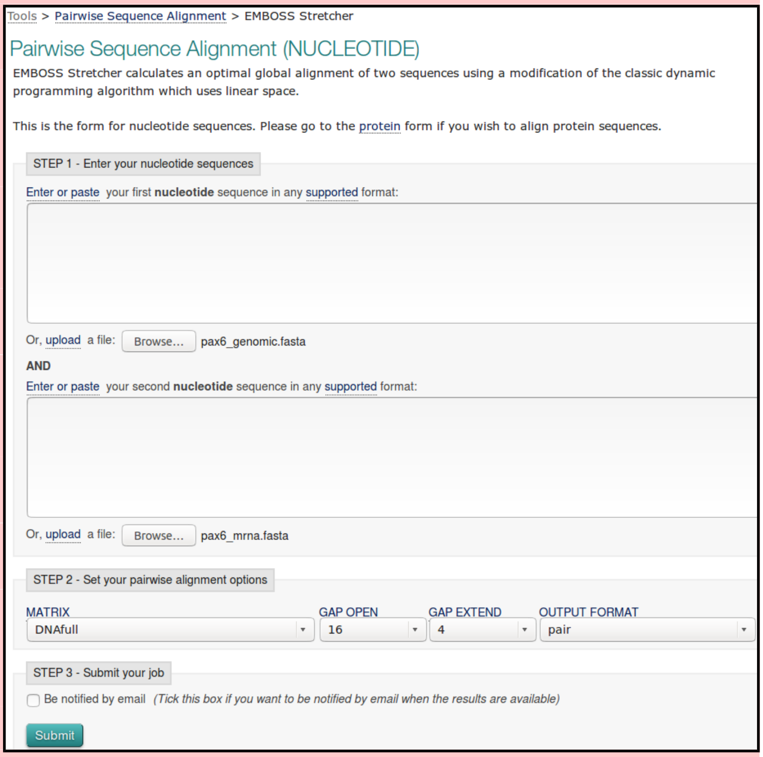
Load up the sequences exactly as for **Needle**.

Take a look at the parameters and see there is nothing unexpected hiding there.

Set **Stretcher** sequence rope stretching.

How do you feel about the results this time?

How do you think **blast** achieve the correct results without any fuss?



Tools > Pairwise Sequence Alignment > EMBOSS Stretcher

Pairwise Sequence Alignment (NUCLEOTIDE)

EMBOSS Stretcher calculates an optimal global alignment of two sequences using a modification of the classic dynamic programming algorithm which uses linear space.

This is the form for nucleotide sequences. Please go to the protein form if you wish to align protein sequences.

STEP 1 - Enter your nucleotide sequences

Enter or paste your first **nucleotide** sequence in any supported format:

Or, upload a file: Browse... pax6_genomic.fasta

AND

Enter or paste your second **nucleotide** sequence in any supported format:

Or, upload a file: Browse... pax6_mrna.fasta

STEP 2 - Set your pairwise alignment options

MATRIX	GAP OPEN	GAP EXTEND	OUTPUT FORMAT
DNAfull	16	4	pair

STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

Pairwise Sequence Comparison using Specialised Software

None of the alignments generated thus far have been entirely correct.

By persuading the general global alignment software to treat huge gaps (i.e. the introns) in some sort of special manner, a reasonable answer was obtained. However, the general software could not know that something more than just **Substitutions** and **Indels** were at issue here. Consequently, it stood no chance of dealing with the intron/exon boundaries sensibly.

The solution is not to fiddle around with the parameters of the general tools. Aligning **mRNA**s with **Genomic** sequence is simply not "*General Alignment*". It is an example of a problem that is sufficiently particular to require specialised software for an optimal solution.

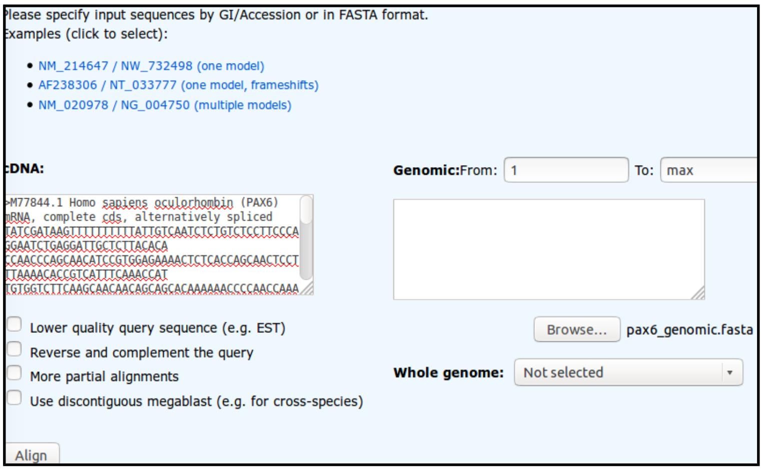
There is a program in the **EMBOSS** package (the same collection of programs as **Needle** and **Stretcher**), called **est2genome**, which is specifically designed for the alignment of cDNA/mRNA and genomic sequences. **est2genome** (and similar programs) may assume much more about the sequences to be aligned than can a general purpose alignment program. Gaps representing introns can be placed far more accurately if they are **known** to represent introns. Programs such as **est2genome** seek the highly conserved bases that occur at intron/exon boundaries, **C/T** rich intronic regions, **polyA** regions and **Stop/Start** codons to assist its detection of exons and gene structures.

est2genome is a fine program, but the option offered at the **NCBI** in America does the same job, I think, somewhat more nicely. The **NCBI** program is called **splign**. To investigate, go to the home of **splign** at:

http://www.ncbi.nlm.nih.gov/sutils/splign

Click on the **Online** button. In the **Genomic** section, **Browse** to upload pax6_genomic.fasta.

In the **cDNA** section, paste the sequence pax6_mrna.fasta. Where **cDNA** and **Genomic** sequences share exons that are nearly identical, **splign** uses the comparison algorithm **megablast** (default). Where exons are less similar (e.g. when the **cDNA** and **Genomic** sequences are from different organisms) the more sensitive option **discontinuous megablast**, is a better choice[1]. Note the option to compare your **cDNA** with a **Whole genome** (including Human). Today, the default options are fine. Click the **Align** button.



Please specify input sequences by GI/Accession or in FASTA format.
Examples (click to select):

- NM_214647 / NW_732498 (one model)
- AF238306 / NT_033777 (one model, frameshifts)
- NM_020978 / NG_004750 (multiple models)

cDNA:

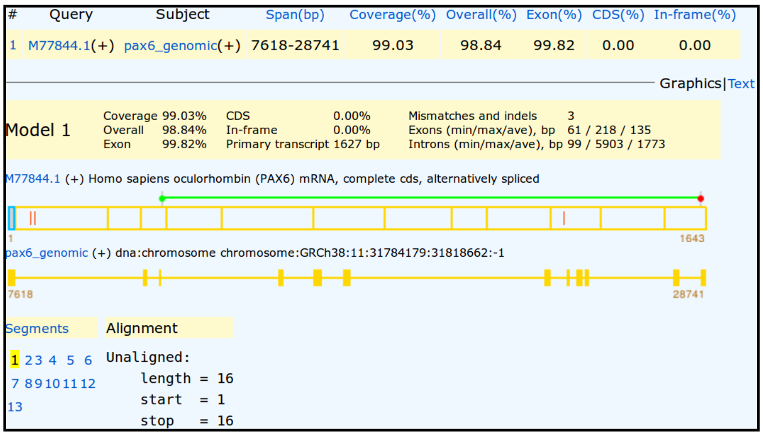
```
>M77844.1 Homo sapiens oculorhombin (PAX6)
mRNA, complete cds, alternatively spliced
TATCGATAAGTTTTTTTTTTATTGTCAATCTCTGTCTCCTTCCCA
GGAATCTGAGGATTGCTCTTACACA
CCAACCCAGCAACATCCGTGGAGAAAACTCTCACCAGCAACTCCT
TTAAAACACCGTCATTTCAAACCAT
TGTGGTCTTCAAGCAACAACAGCAGCACAAAAAACCCCAACCAAA
```

Genomic:From: 1 To: max

Browse... pax6_genomic.fasta

Whole genome: Not selected

□ Lower quality query sequence (e.g. EST)
□ Reverse and complement the query
□ More partial alignments
□ Use discontiguous megablast (e.g. for cross-species)



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	M77844.1(+)	pax6_genomic(+)	7618-28741	99.03	98.84	99.82	0.00	0.00

Model 1

Coverage	99.03%	CDS	0.00%	Mismatches and indels	3
Overall	98.84%	In-frame	0.00%	Exons (min/max/ave), bp	61 / 218 / 135
Exon	99.82%	Primary transcript	1627 bp	Introns (min/max/ave), bp	99 / 5903 / 1773

M77844.1 (+) Homo sapiens oculorhombin (PAX6) mRNA, complete cds, alternatively spliced

pax6_genomic (+) dna:chromosome chromosome:GRCh38:11:31784179:31818662:-1

Alignment

```
Unaligned:
    length = 16
    start  = 1
    stop   = 16
```

Your results will appear showing the cDNA split into **12** sections (the predicted exons) corresponding to **12** regions of the genomic sequence indicated by yellow rectangles. A **13th** region of **16** base pairs is displayed and declared to be **unaligned**. These are the **16** mystery base pairs at the start of this particular mRNA that **Needle** and **Stretcher** had trouble treating sensibly also. I wonder what they are?

Any theories?

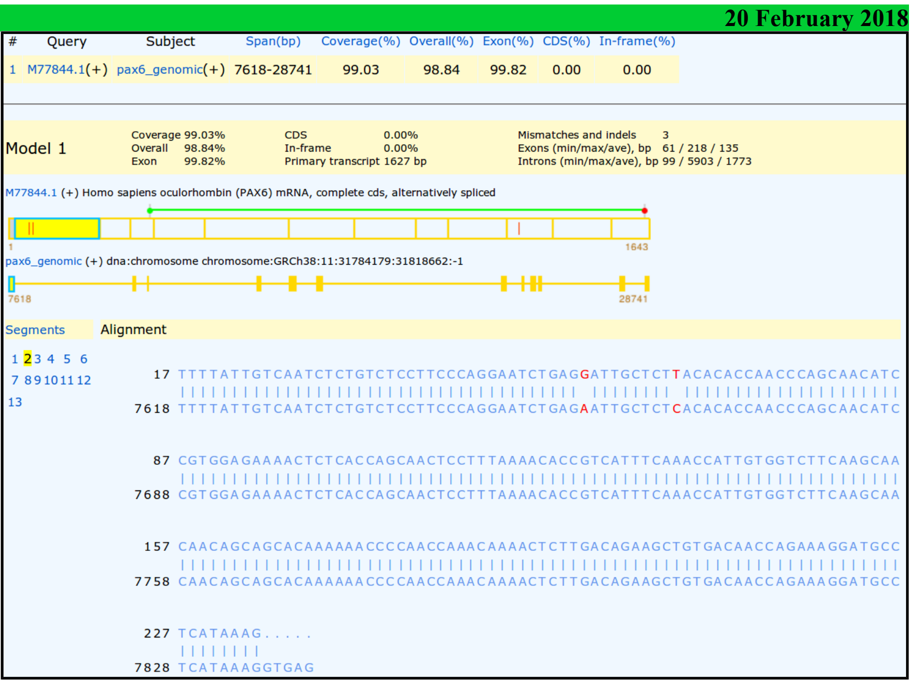
[1] Why this is so will be considered later when we look at the database searching program **blast**.

Click on the first exon section of the cDNA display.

Here there shows two **substitutions**. These were also apparent in the successful **blast**, **Needle** and **Stretcher** alignments. You might have spotted them?

Though these are in a non-coding region, they could easily still be very significant. However, for the purposes of this exercise, let us assume they are not.

The **Start** (green) and **Stop** (red) codons delimiting the **CoDing Sequence** (**CDS**) are illustrated by the bar above the cDNA display.



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	M77844.1(+)	pax6_genomic(+)	7618-28741	99.03	98.84	99.82	0.00	0.00

```
Model 1   Coverage 99.03%   CDS                0.00%     Mismatches and indels          3
          Overall  98.84%   In-frame           0.00%     Exons (min/max/ave), bp    61 / 218 / 135
          Exon     99.82%   Primary transcript 1627 bp   Introns (min/max/ave), bp  99 / 5903 / 1773
```

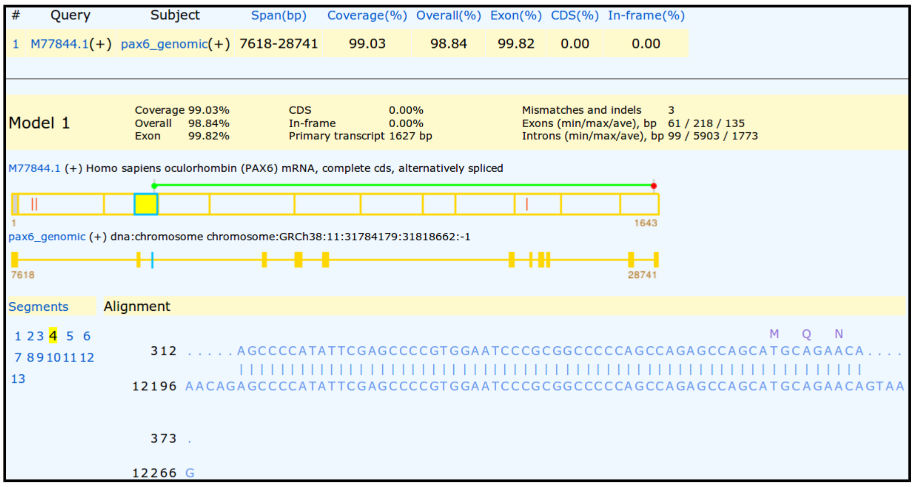
M77844.1 (+) Homo sapiens oculorhombin (PAX6) mRNA, complete cds, alternatively spliced

pax6_genomic (+) dna:chromosome chromosome:GRCh38:11:31784179:31818662:-1

Segments: 1 2 3 4 5 6 7 8 9 10 11 12 13

Alignment

```
  17 TTTTATTGTCAATCTCTGTCTCCTTCCCAGGAATCTGAGGATTGCTCTTACACACCAACCCAGCAACATC
     ||||||||||||||||||||||||||||||||||||||| |||||||| |||||||||||||||||||||
7618 TTTTATTGTCAATCTCTGTCTCCTTCCCAGGAATCTGAGAATTGCTCTCACACACCAACCCAGCAACATC

  87 CGTGGAGAAAACTCTCACCAGCAACTCCTTTAAAACACCGTCATTTCAAACCATTGTGGTCTTCAAGCAA
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
7688 CGTGGAGAAAACTCTCACCAGCAACTCCTTTAAAACACCGTCATTTCAAACCATTGTGGTCTTCAAGCAA

 157 CAACAGCAGCACAAAAAACCCCAACCAAACAAAACTCTTGACAGAAGCTGTGACAACCAGAAAGGATGCC
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
7758 CAACAGCAGCACAAAAAACCCCAACCAAACAAAACTCTTGACAGAAGCTGTGACAACCAGAAAGGATGCC

 227 TCATAAAG.....
     ||||||||
7828 TCATAAAGGTGAG
```



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	M77844.1(+)	pax6_genomic(+)	7618-28741	99.03	98.84	99.82	0.00	0.00

```
Model 1   Coverage 99.03%   CDS                0.00%     Mismatches and indels          3
          Overall  98.84%   In-frame           0.00%     Exons (min/max/ave), bp    61 / 218 / 135
          Exon     99.82%   Primary transcript 1627 bp   Introns (min/max/ave), bp  99 / 5903 / 1773
```

M77844.1 (+) Homo sapiens oculorhombin (PAX6) mRNA, complete cds, alternatively spliced

pax6_genomic (+) dna:chromosome chromosome:GRCh38:11:31784179:31818662:-1

Segments: 1 2 3 4 5 6 7 8 9 10 11 12 13

Alignment

```
                                                                 M  Q  N
  312 .....AGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCCCCAGCCAGAGCCAGCATGCAGAACA....
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
12196 AACAGAGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCCCCAGCCAGAGCCAGCATGCAGAACAGTAA

  373 .

12266 G
```

Click on the exon including the green **Start** codon (the **3rd**).

The first coding exon is now displayed with translation of the mRNA where appropriate.

The statistics at the top of the display include the claim that there are **3** discrepancies (**Mismatches** and **Indels**) between the **cDNA** and **Genomic** sequences.

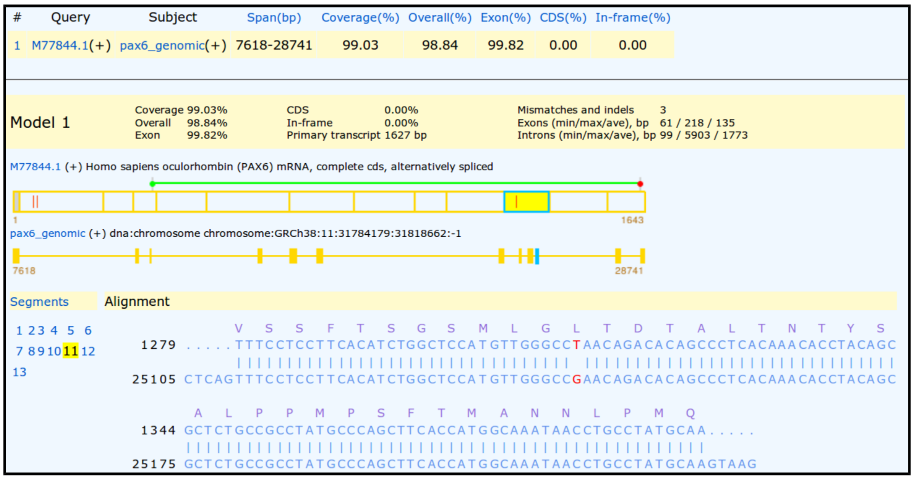
Two of these are the **substitutions** we have already seen in the first exon of the cDNA. The third is indicated by the red bar in the **10th** exon of the **cDNA** display.

Click on the **10th** exon section of the cDNA display.

The third difference, a substitution, should be clear to see. Given it changes the coded protein, this substitution is likely to be the most significant.

Irritatingly, in the extreme! **splign** only translates the mRNA. So one has to work to discover the alternative suggested by the Genomic sequence.

Vital if we were really doing this seriously, but for an exercise, it is fine to relax. I do not intrude on real life much and **it,** largely, leaves **me** untouched in grateful response.



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	M77844.1(+)	pax6_genomic(+)	7618-28741	99.03	98.84	99.82	0.00	0.00

```
Model 1   Coverage 99.03%   CDS                0.00%     Mismatches and indels          3
          Overall  98.84%   In-frame           0.00%     Exons (min/max/ave), bp    61 / 218 / 135
          Exon     99.82%   Primary transcript 1627 bp   Introns (min/max/ave), bp  99 / 5903 / 1773
```

M77844.1 (+) Homo sapiens oculorhombin (PAX6) mRNA, complete cds, alternatively spliced

pax6_genomic (+) dna:chromosome chromosome:GRCh38:11:31784179:31818662:-1

Segments: 1 2 3 4 5 6 7 8 9 10 11 12 13

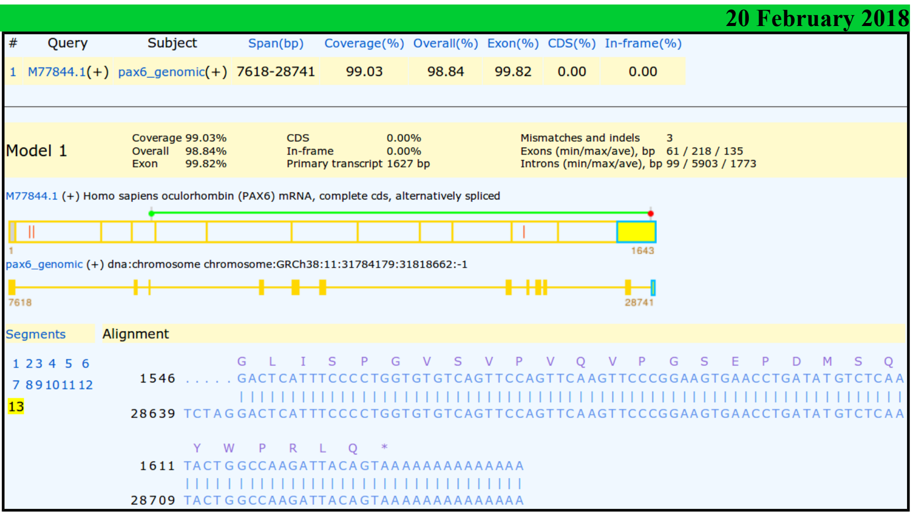
Alignment

```
              V  S  S  F  T  S  G  S  M  L  G  L  T  D  T  A  L  T  N  T  Y  S
 1279 .....TTTCCTCCTTCACATCTGGCTCCATGTTGGGCCTAACAGACACAGCCCTCACAAACACCTACAGC
           ||||||||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||
25105 CTCAGTTTCCTCCTTCACATCTGGCTCCATGTTGGGCCGAACAGACACAGCCCTCACAAACACCTACAGC

       A  L  P  P  M  P  S  F  T  M  A  N  N  L  P  M  Q
 1344 GCTCTGCCGCCTATGCCCAGCTTCACCATGGCAAATAACCTGCCTATGCAA.....
      |||||||||||||||||||||||||||||||||||||||||||||||||||
25175 GCTCTGCCGCCTATGCCCAGCTTCACCATGGCAAATAACCTGCCTATGCAAGTAAG
```

What is the amino acid corresponding to the mutated position in the **Genomic** sequence?

What are the **Genomic** and **mRNA** base positions corresponding to the mutation at amino acid position **33**?

Click on the last exon section in the cDNA display. You should now see the final exon of the cDNA with the **Stop** codon and polyA region.



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	M77844.1(+)	pax6_genomic(+)	7618-28741	99.03	98.84	99.82	0.00	0.00

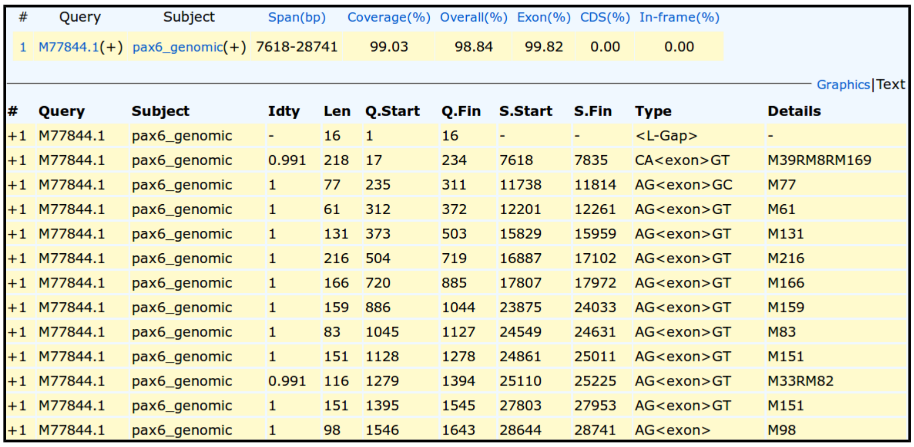
```
Model 1   Coverage 99.03%    CDS                0.00%       Mismatches and indels         3
          Overall  98.84%    In-frame           0.00%       Exons (min/max/ave), bp       61 / 218 / 135
          Exon     99.82%    Primary transcript 1627 bp     Introns (min/max/ave), bp     99 / 5903 / 1773
```

M77844.1 (+) Homo sapiens oculorhombin (PAX6) mRNA, complete cds, alternatively spliced

pax6_genomic (+) dna:chromosome chromosome:GRCh38:11:31784179:31818662:-1

```
Segments       Alignment

1 2 3 4 5 6                G  L  I  S  P  G  V  S  V  P  V  Q  V  P  G  S  E  P  D  M  S  Q
7 8 9 10 11 12   1546 .....GACTCATTTCCCCTGGTGTGTCAGTTCCAGTTCAAGTTCCCGGAAGTGAACCTGATATGTCTCAA
13                         |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
                28639 TCTAGGACTCATTTCCCCTGGTGTGTCAGTTCCAGTTCAAGTTCCCGGAAGTGAACCTGATATGTCTCAA

                       Y  W  P  R  L  Q  *
                 1611 TACTGGCCAAGATTACAGTAAAAAAAAAAAAA
                      ||||||||||||||||||||||||||||||||
                28709 TACTGGCCAAGATTACAGTAAAAAAAAAAAAA
```

Finally, click on the **Text** link to view the textual summary of the **splign** results.



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	M77844.1(+)	pax6_genomic(+)	7618-28741	99.03	98.84	99.82	0.00	0.00

Graphics|Text

#	Query	Subject	Idty	Len	Q.Start	Q.Fin	S.Start	S.Fin	Type	Details
+1	M77844.1	pax6_genomic	-	16	1	16	-	-	<L-Gap>	-
+1	M77844.1	pax6_genomic	0.991	218	17	234	7618	7835	CA<exon>GT	M39RM8RM169
+1	M77844.1	pax6_genomic	1	77	235	311	11738	11814	AG<exon>GC	M77
+1	M77844.1	pax6_genomic	1	61	312	372	12201	12261	AG<exon>GT	M61
+1	M77844.1	pax6_genomic	1	131	373	503	15829	15959	AG<exon>GT	M131
+1	M77844.1	pax6_genomic	1	216	504	719	16887	17102	AG<exon>GT	M216
+1	M77844.1	pax6_genomic	1	166	720	885	17807	17972	AG<exon>GT	M166
+1	M77844.1	pax6_genomic	1	159	886	1044	23875	24033	AG<exon>GT	M159
+1	M77844.1	pax6_genomic	1	83	1045	1127	24549	24631	AG<exon>GT	M83
+1	M77844.1	pax6_genomic	1	151	1128	1278	24861	25011	AG<exon>GT	M151
+1	M77844.1	pax6_genomic	0.991	116	1279	1394	25110	25225	AG<exon>GT	M33RM82
+1	M77844.1	pax6_genomic	1	151	1395	1545	27803	27953	AG<exon>GT	M151
+1	M77844.1	pax6_genomic	1	98	1546	1643	28644	28741	AG<exon>	M98

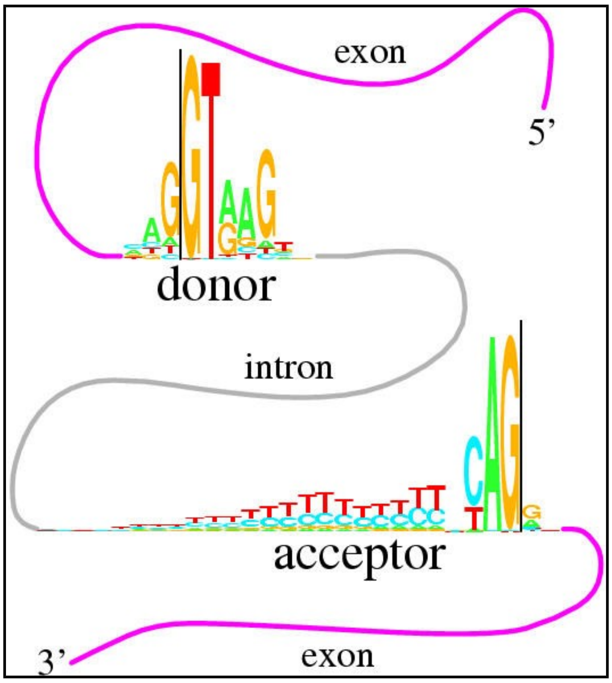
How do you interpret the **Details** column for exons 1 and 10?

Where is the **3rd** substitution in the mRNA?

Where is the **3rd** substitution in the Genomic Sequence?

Compare the predicted **splign** intron/exon boundaries with the conservation suggested by the logo[2]?

What deviation(s) from the model suggested by the logo can you see?



[2] The original label for this very nice graphic is:
This figure shows two ‘‘sequence logos’’ which represent sequence conservation at the 5’ (donor) and 3’ (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern ‘‘**CAG|GT**’’, which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992).

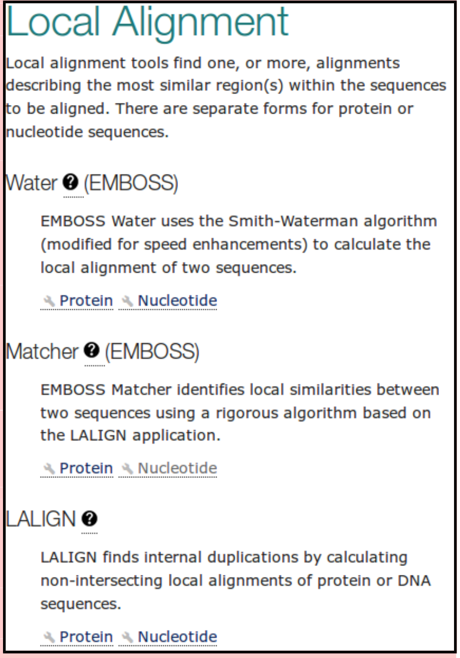
Sensitive Local Pairwise Sequence Comparison

Finally, a swift look at sensitive local pairwise sequence alignment. You have already used **blast** to do a local pairwise alignment in the last Practical, when you aligned the two human genomic sequencing contigs that covered the **PAX6** location in **Chromosome 11**. **blast** did not use a sensitive approach however, nothing subtle was required for that particular alignment.

For a more accurate alignment, return to the **Pairwise Sequence Alignment EBI** page (**http://www.ebi.ac.uk/Tools/psa/**).

From there, select the **Nucleotide** option for the **Local Alignment** program **Matcher**.

Water or **LALIGN** would also be fine options, but I declare the nucleotide option of **Matcher** to be choice of the day.



Local Alignment

Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

Water ❓ (EMBOSS)

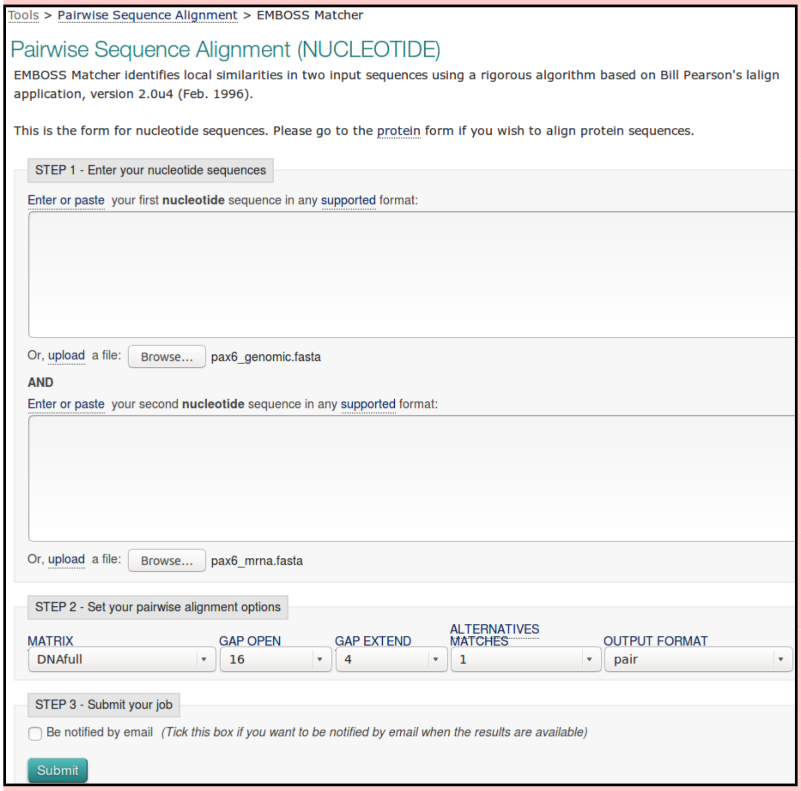
EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.

Matcher ❓ (EMBOSS)

EMBOSS Matcher identifies local similarities between two sequences using a rigorous algorithm based on the LALIGN application.

LALIGN ❓

LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or DNA sequences.



Tools > Pairwise Sequence Alignment > EMBOSS Matcher

Pairwise Sequence Alignment (NUCLEOTIDE)

EMBOSS Matcher identifies local similarities in two input sequences using a rigorous algorithm based on Bill Pearson's lalign application, version 2.0u4 (Feb. 1996).

This is the form for nucleotide sequences. Please go to the protein form if you wish to align protein sequences.

STEP 1 - Enter your nucleotide sequences

Enter or paste your first **nucleotide** sequence in any supported format:

Or, upload a file: Browse... pax6_genomic.fasta

AND

Enter or paste your second **nucleotide** sequence in any supported format:

Or, upload a file: Browse... pax6_mrna.fasta

STEP 2 - Set your pairwise alignment options

MATRIX	GAP OPEN	GAP EXTEND	ALTERNATIVES MATCHES	OUTPUT FORMAT
DNAfull	16	4	1	pair

STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Load up the **Genomic** and **mRNA sequences** as you did for **Needle**.

Click on the **More options** button to see what parameters you can set. They should be as you might expect. The defaults are fine for the first run.

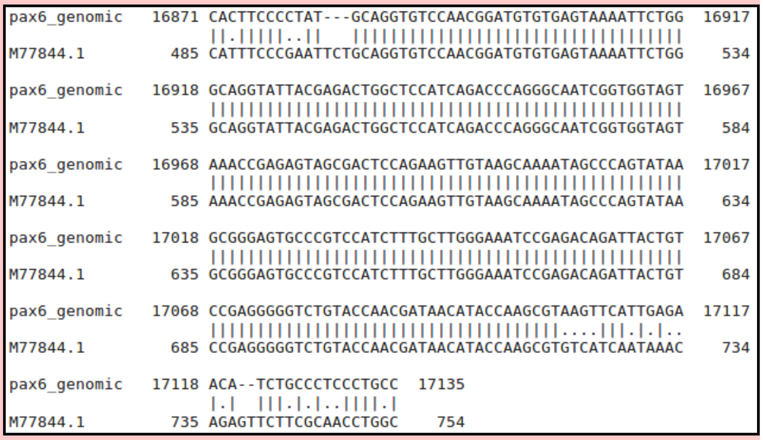
Click on the **Submit** button to get **Matcher** into Matchbox mode.

After due consideration of all the possibilities, **Matcher** will enrich your screen with its conclusions.

But, only one alignment? A good one, covering the highest scoring region of all those considered, but it cannot be the whole story, which must tell the tale of **12** exons! Here is but one.

In common with most local alignment programs, by default **Matcher** will only show you the single best local alignment between two sequences.

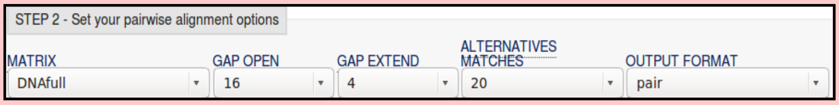
A good reason to have a **Dot Matrix View** to inform one of roughly what to expect, which is not one miserable alignment in this case.



```
pax6_genomic   16871 CACTTCCCCTAT---GCAGGTGTCCAACGGATGTGTGAGTAAAATTCTGG  16917
                     ||.|||||..||   |||||||||||||||||||||||||||||||||||
M77844.1         485 CATTTCCCGAATTCTGCAGGTGTCCAACGGATGTGTGAGTAAAATTCTGG    534

pax6_genomic   16918 GCAGGTATTACGAGACTGGCTCCATCAGACCCAGGGCAATCGGTGGTAGT  16967
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1         535 GCAGGTATTACGAGACTGGCTCCATCAGACCCAGGGCAATCGGTGGTAGT    584

pax6_genomic   16968 AAACCGAGAGTAGCGACTCCAGAAGTTGTAAGCAAAATAGCCCAGTATAA  17017
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1         585 AAACCGAGAGTAGCGACTCCAGAAGTTGTAAGCAAAATAGCCCAGTATAA    634

pax6_genomic   17018 GCGGGAGTGCCCGTCCATCTTTGCTTGGGAAATCCGAGACAGATTACTGT  17067
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1         635 GCGGGAGTGCCCGTCCATCTTTGCTTGGGAAATCCGAGACAGATTACTGT    684

pax6_genomic   17068 CCGAGGGGGTCTGTACCAACGATAACATACCAAGCGTAAGTTCATTGAGA  17117
                     |||||||||||||||||||||||||||||||||||||....|||.|.|..
M77844.1         685 CCGAGGGGGTCTGTACCAACGATAACATACCAAGCGTGTCATCAATAAAC    734

pax6_genomic   17118 ACA--TCTGCCCTCCCTGCC  17135
                     |.|  |||.|.|..||||.|
M77844.1         735 AGAGTTCTTCGCAACCTGGC    754
```

Of course, it is also miserable biologically! **Matcher** fails to align the exons accurately for all the same reasons that the **Needle** failed to represent the *biological* reality.

So, what can one do but try again! By returning to the **Matcher** launch page from your results. You should find the two sequences are still selected, so you should only have to click on **More Options** again and set the **ALTERNATIVE MATCHES** field **20**.



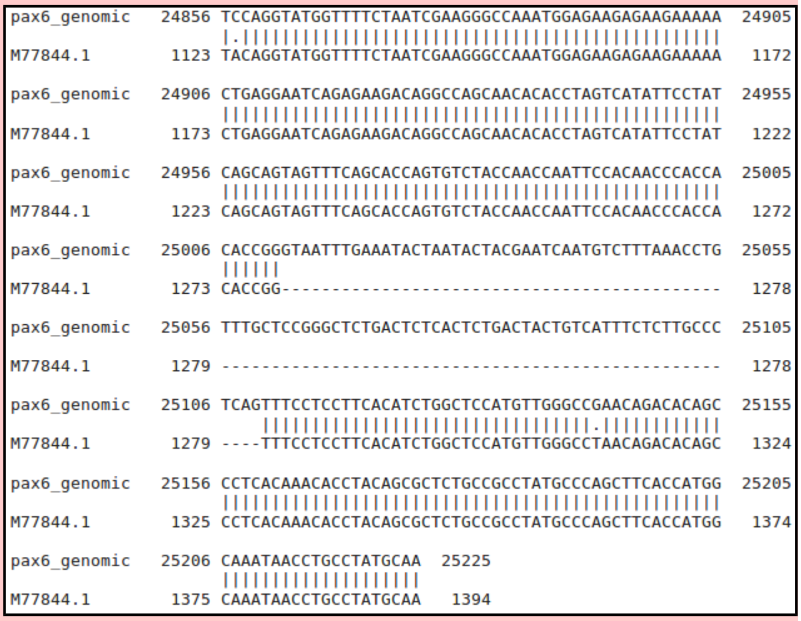
STEP 2 - Set your pairwise alignment options

MATRIX	GAP OPEN	GAP EXTEND	ALTERNATIVES MATCHES	OUTPUT FORMAT
DNAfull	16	4	20	pair

Actually, as you know there are only **12** exons. And that some might well be close enough to be included in the same alignment, you do not need to go as high as **20**. However, the web interface restricts choice (**WHY!?**) such that this is the most sensible cautious choice.

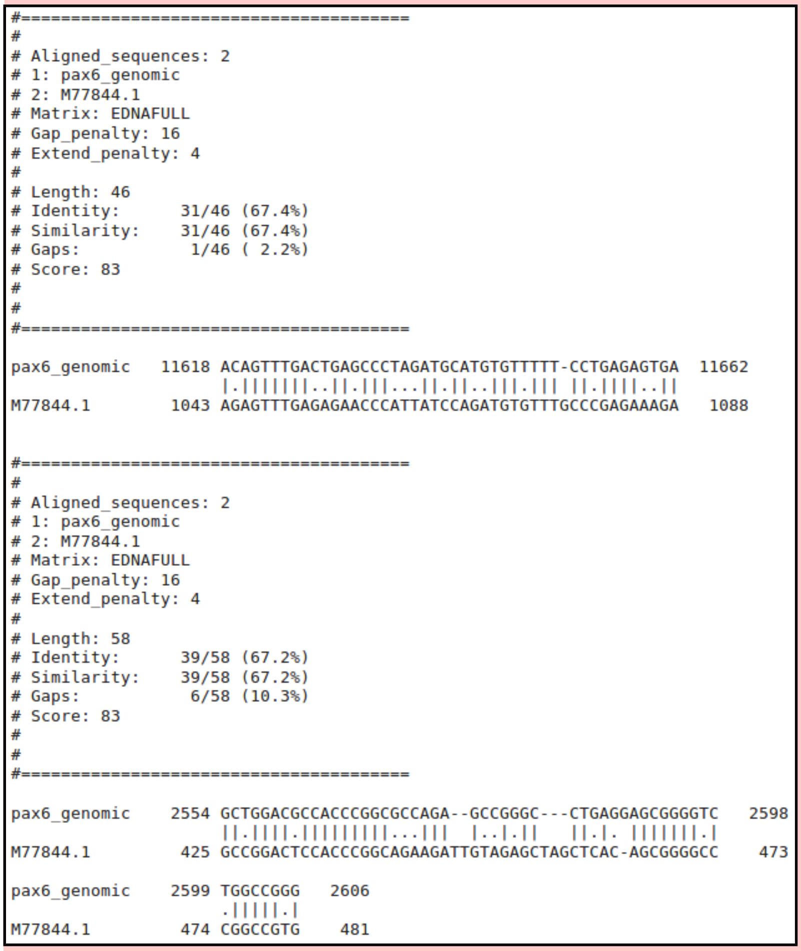
Click on the **Submit** button and **Matcher** will trust and obey.

At the top of your output will be some nice believable local alignments, some involving more than one exon.



```
pax6_genomic   24856 TCCAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAGAAAAA  24905
                     |.||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1        1123 TACAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAGAAAAA   1172

pax6_genomic   24906 CTGAGGAATCAGAGAAGACAGGCCAGCAACACACCTAGTCATATTCCTAT  24955
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1        1173 CTGAGGAATCAGAGAAGACAGGCCAGCAACACACCTAGTCATATTCCTAT   1222

pax6_genomic   24956 CAGCAGTAGTTTCAGCACCAGTGTCTACCAACCAATTCCACAACCCACCA  25005
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1        1223 CAGCAGTAGTTTCAGCACCAGTGTCTACCAACCAATTCCACAACCCACCA   1272

pax6_genomic   25006 CACCGGGTAATTTGAAATACTAATACTACGAATCAATGTCTTTAAACCTG  25055
                     ||||||
M77844.1        1273 CACCGG--------------------------------------------   1278

pax6_genomic   25056 TTTGCTCCGGGCTCTGACTCTCACTCTGACTACTGTCATTTCTCTTGCCC  25105

M77844.1        1279 --------------------------------------------------   1278

pax6_genomic   25106 TCAGTTTCCTCCTTCACATCTGGCTCCATGTTGGGCCGAACAGACACAGC  25155
                         |||||||||||||||||||||||||||||||||.||||||||||||
M77844.1        1279 ----TTTCCTCCTTCACATCTGGCTCCATGTTGGGCCTAACAGACACAGC   1324

pax6_genomic   25156 CCTCACAAACACCTACAGCGCTCTGCCGCCTATGCCCAGCTTCACCATGG  25205
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
M77844.1        1325 CCTCACAAACACCTACAGCGCTCTGCCGCCTATGCCCAGCTTCACCATGG   1374

pax6_genomic   25206 CAAATAACCTGCCTATGCAA  25225
                     ||||||||||||||||||||
M77844.1        1375 CAAATAACCTGCCTATGCAA   1394
```

Matcher tries to make each alignment as long as it can, stopping only when, to stretch the alignment any further would involve the alignment score deceasing due to the necessity for gap penalties.



```
#=======================================
#
# Aligned_sequences: 2
# 1: pax6_genomic
# 2: M77844.1
# Matrix: EDNAFULL
# Gap_penalty: 16
# Extend_penalty: 4
#
# Length: 46
# Identity:      31/46 (67.4%)
# Similarity:    31/46 (67.4%)
# Gaps:           1/46 ( 2.2%)
# Score: 83
#
#
#=======================================

pax6_genomic   11618 ACAGTTTGACTGAGCCCTAGATGCATGTGTTTTT-CCTGAGAGTGA  11662
                     |.|||||||..||.|||...||.||..|||.||| ||.||||..||
M77844.1        1043 AGAGTTTGAGAGAACCCATTATCCAGATGTGTTTGCCCGAGAAAGA   1088


#=======================================
#
# Aligned_sequences: 2
# 1: pax6_genomic
# 2: M77844.1
# Matrix: EDNAFULL
# Gap_penalty: 16
# Extend_penalty: 4
#
# Length: 58
# Identity:      39/58 (67.2%)
# Similarity:    39/58 (67.2%)
# Gaps:           6/58 (10.3%)
# Score: 83
#
#
#=======================================

pax6_genomic    2554 GCTGGACGCCACCCGGCGCCAGA--GCCGGGC---CTGAGGAGCGGGGTC   2598
                     ||.||||.|||||||||...|||  |..|.||   ||.|. |||||||.|
M77844.1         425 GCCGGACTCCACCCGGCAGAAGATTGTAGAGCTAGCTCAC-AGCGGGGCC    473

pax6_genomic    2599 TGGCCGGG   2606
                     .|||||.|
M77844.1         474 CGGCCGTG    481
```

Go to far down the list of alignments and you will realise what a literal interpretation **Matcher** has of its duties.

You asked for **20** alignments?

So here are the best **20** alignments and it is entirely up to you to decide where “silly” begins.

Not too difficult in this case I suggest.

Why do you suppose your aligned exons are not presented in the correct positional order?

THE END

DPJ – 2017.12.23

Model Answers to Questions in the Instructions Text.

Notes:

For the most part, these “**Model Answers**” just provide the reactions/solutions I hoped you would work out for yourselves. However, sometime I have tried to offer a bit more background and material for thought? Occasionally, I have rambled off into some rather self indulgent investigations that even I would not want to try and justify as pertinent to the objective of these exercises. I like to keep these meanders, as they help and entertain me, but I wish to warn you to only take regard of them if you are feeling particularly strong and have time to burn. Certainly not a good idea to indulge here during a time constrained course event!

Where things have got extreme, I am going to make two versions of the answer. One starting:

Summary:

Which has the answer with only a reasonably digestible volume of deep thought. Read this one.

The other will start:

Full Answer:

Beware of entering here! I do not hold back. Nothing complicated, but it will be long and full of pedantry.

This makes the Model answers section very big. **BUT**, it is not intended for printing or for reading serially, so I submit, being long and wordy does not matter. Feel free to disagree.

From your investigations of **Global Alignment**:

What do you suppose these regions represent?

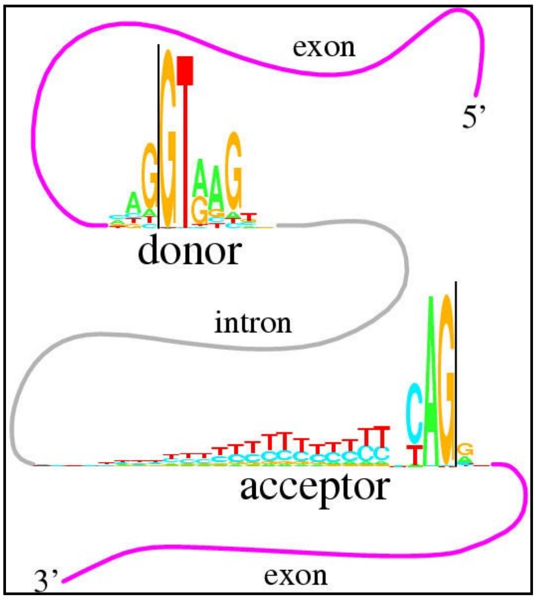
Exons

How might the gap around **24,750** in the genomic sequence been positioned more intelligently?

blast has positioned a gap in this region merely to maximize the overall alignment score. There is more than one way of achieving this simple goal. However, if it were to be recognized that the gap to be positioned was to represent an intron, then one of the arithmetically equivalent options becomes far more attractive than the others. This "best" option is not the one chosen by **blast**, which is forgiveable as **blast** had nor reason to expect an intron and was not written to understand the properties of introns anyway.

The alignment chosen for this region by **blast** was:

```
Genomic  24601  CAAAATAGATCTACCTGAAGCAAGAATACAGGTACCGAGAGACTGTGCAGTTTCACACTT  24660
                ||||||||||||||||||||||||||||||||||
mRNA     1097   CAAAATAGATCTACCTGAAGCAAGAATACAGGTA--------------------------  1130

                                          • • •

Genomic  24781  GGGAGGGCAGCAGTGGAGGTGCCAAGGTGGGGCTGGGCTCGACGTAGACACAGTGCTAAC  24840

mRNA            ------------------------------------------------------------

Genomic  24841  CTGTCCCACCTGATTTCCAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAG  24900
                                       |||||||||||||||||||||||||||||||||||||
mRNA     1131   -----------------------TGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAG  1167
```



Shifting the gap **3** places to the left neither changes the size of the gap nor the perfection of the alignment either side of the gap and so does not affect the alignment score.

However, it does mean the gap begins with an **GT** and ends with a **AG** which is what one might expect if it were known that the gap represented an intron. I include the beautiful **Intron/Exon** logo. As you might gather, I rather like this one.

So, if **blast** was a little better informed, the improved alignment would have been:

```
Genomic  24601  CAAAATAGATCTACCTGAAGCAAGAATACAGGTACCGAGAGACTGTGCAGTTTCACACTT  24660
                |||||||||||||||||||||||||||||||
mRNA     1097   CAAAATAGATCTACCTGAAGCAAGAATACAG-----------------------------  1130

                                          • • •

Genomic 24781   GGGAGGGCAGCAGTGGAGGTGCCAAGGTGGGGCTGGGCTCGACGTAGACACAGTGCTAAC  24840

mRNA            ------------------------------------------------------------

Genomic  24841  CTGTCCCACCTGATTTCCAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAG  24900
                                    ||||||||||||||||||||||||||||||||||||||||
mRNA     1131   --------------------GTATGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAG  1167
```

This is the alignment that one might expect from any program customized to align **mRNA** with **Genomic** sequence, as you will see in the fullness of time.

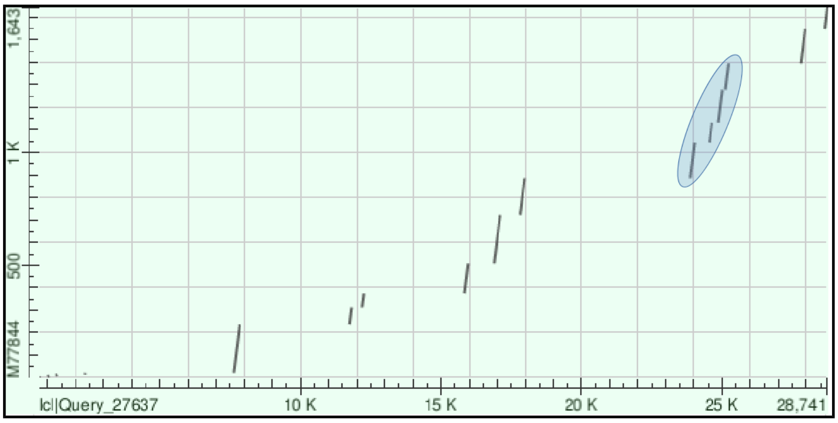
How many convincingly aligned regions did you see?

4

How many did you expect?

12, as that was how many **blast** found, not including the silly ones at the beginning.

The **4** that were found correspond the illustrated **4** diagonal lines grouped together in the **Dot Matrix View** made by **blast**.



Clearly, this alignment is not correct. Can you explain why?

This alignment algorithm only wishes to maximise an alignment score. It sees ***ALL*** the high scoring exon regions, however, as the gaps between many of the exons (introns that is) are so long that the penalties for representing them correctly are greater than the gain achieved by the inclusion the extra exons in the alignment. Arithmetically, it is better to align all the exons either side of the **4** exons that were aligned sensibly, in the biologically improbably fashion shown. Arithmetically the best alignment, biologically ridiculous!

This behaviour is exaggerated because this program regards the enormous gaps in has suggested at the start and end of the alignments as "free". Some global alignment programs (including this one if you ask politely, as you will see) offer the option of penalising the ends gaps in the same way as for internal gaps. Normally, not penalising end gaps is sensible as it allows for the sequences to have slightly different lengths. In this case, penalising end gaps will result in a far better alignment.

Had you used **stretcher** (also offered by the **EBI**) you would have got a much improved answer in this case (but not necessarily in generally). This is because **stretcher** works in a way far closer to the way an informed human might think. **stretcher** does not mindlessly insist of the highest alignment score. Instead, it looks for all the high scoring regions (i.e. all the exons) and then computes the best way to link them together. The result is a far more convincing alignment, but not the arithmetically best scoring answer.

How many matching regions are there this time?

Were you to trawl though your textual output carefully (or simply take my immaculate word for it), you would find **12** perfectly (or nearly so) aligned regions, implying **12** exons.

To be pedantic, the nicely aligned regions do not match the exons exactly (as has been discussed), but well enough to claim definite evidence for the number of exons. **12** is good enough for me.

Is the count **now** roughly as you would expect?

Yes, exactly the same as **blast** predicted in the first place. More exons that **17** might have been a surprise as that is how many the gene record for **PAX6** at the **NCBI** suggested. Any given transcript may have less than **17** exons or exactly **17** exons, but not more than **17** exons if the heroes of the **NCBI** are not mistaken.

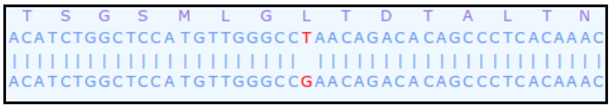
How do you think **blast** achieve the correct results without any fuss?

The only way **blast** could have got the right answer, as it did, would be to use one of the strategies listed previously. **blast** did not use the horrible idea of making gaps super cheap! Not only is that a disgustingly dirty trick, but **blast** actually declares that it is using quite sensible gap penalties.

Leaving **penalising end gaps** and/or using the same sort of heuristics employed by **stretcher**. I would strongly suspect **blast** uses a **stretcher** approach. After all, **blast** has clearly already identified all the "promising regions" in order to construct its **Dot Matrix View**. Also the **stretcher** strategy is similar to that of all **blast** searches (discussed in the next Practical). Finally, **blast** is often used to align very long DNA sequences to detect very strongly similar large regions. This is exactly what the faster (if less pure) **stretcher** approach is all about.

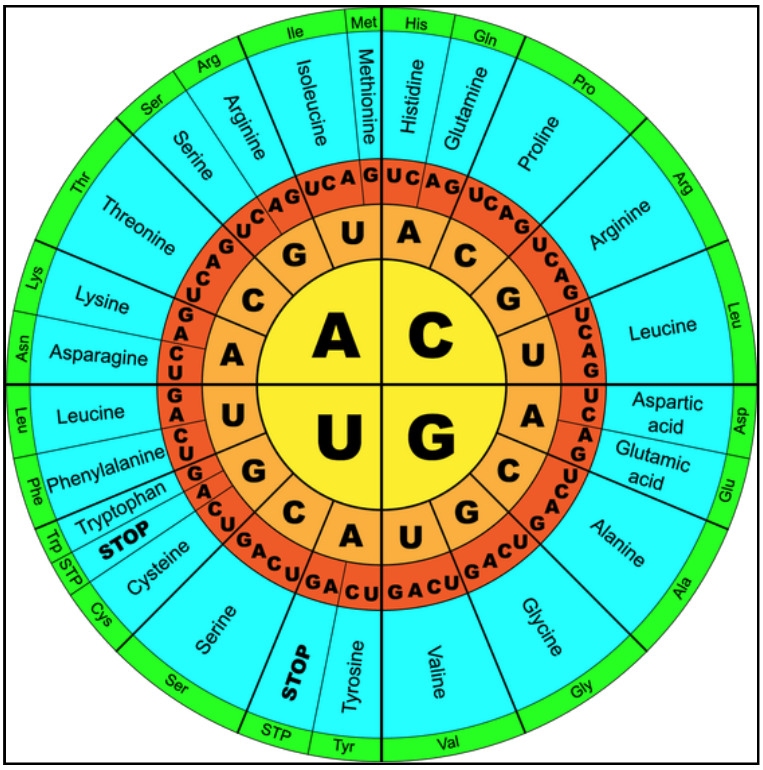
From your investigations comparing mRNA/cDNA with genomic DNA:

What is the amino acid corresponding to the mutated position in the **Genomic** sequence?



```
 T  S  G  S  M  L  G  L  T  D  T  A  L  T  N
ACATCTGGCTCCATGTTGGGCCTAACAGACACAGCCCTCACAAAC
|||||||||||||||||||||| ||||||||||||||||||||||
ACATCTGGCTCCATGTTGGGCCGAACAGACACAGCCCTCACAAAC
```

The top sequence is the mRNA. **splign** is kind enough to explicitly inform us that the “mutated” codon, **CTA**, will be expressed a **L**eucine.



So, why not translate the **Genomic** sequence also **splign**?! Easy enough to look up. But I resent having to do so!

From this rather beautiful representation of the **Genetic Code**, I conclude:

mRNA	**CTA**	→	**Leucine** (**L**)
Genomic	**CGA**	→	**Arginine** (**R**)

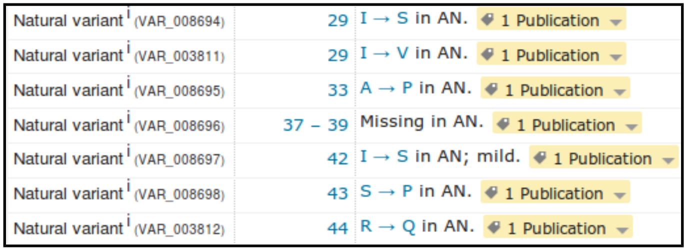
I checked, and this does not appear to be a substitution that is associated with any “interesting” phenotype.

There is no real reason why it should. We did not pause to find out anything about the mRNA downloaded from the **NCBI**, The annotation is particularly unrevealing by itself (it is in **Backup_Files** if you really want to check).

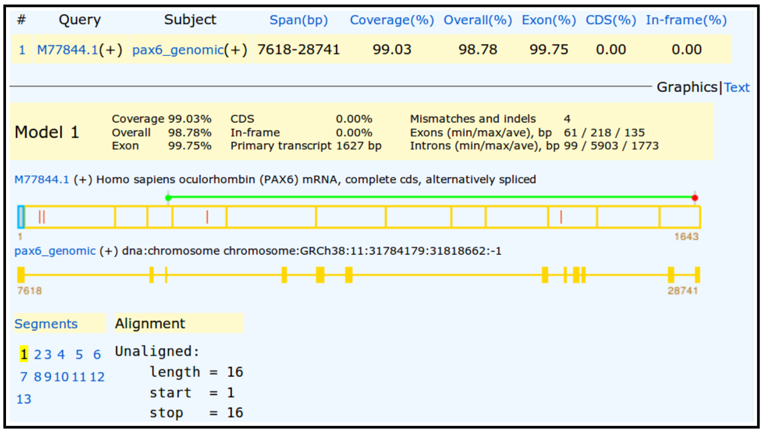
Let us simply assume it is a benign **A**ccepted **P**oint **M**utation (**PAM**). Yes indeed, that feels comfortable. Not so very tricky this Science stuff after all what!

What are the **Genomic** and **mRNA** base positions corresponding to the mutation at amino acid position **33**?

Remember the **Natural variation** at amino acid position **33**? You looked at it in passing during the course of the first exercise. It is a major cause of **Aniridia**. An **Alanine** mutated to a **Proline** at the end of a **Helix** vital to the **DNA Binding** function of the **PAX6** protein.



Natural variant (VAR_008694)	29	I → S in AN.	1 Publication
Natural variant (VAR_003811)	29	I → V in AN.	1 Publication
Natural variant (VAR_008695)	33	A → P in AN.	1 Publication
Natural variant (VAR_008696)	37 – 39	Missing in AN.	1 Publication
Natural variant (VAR_008697)	42	I → S in AN; mild.	1 Publication
Natural variant (VAR_008698)	43	S → P in AN.	1 Publication
Natural variant (VAR_003812)	44	R → Q in AN.	1 Publication



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	M77844.1(+)	pax6_genomic(+)	7618-28741	99.03	98.78	99.75	0.00	0.00

```
Model 1   Coverage 99.03%   CDS                 0.00%    Mismatches and indels        4
          Overall  98.78%   In-frame            0.00%    Exons (min/max/ave), bp      61 / 218 / 135
          Exon     99.75%   Primary transcript 1627 bp   Introns (min/max/ave), bp    99 / 5903 / 1773
```

M77844.1 (+) Homo sapiens oculorhombin (PAX6) mRNA, complete cds, alternatively spliced

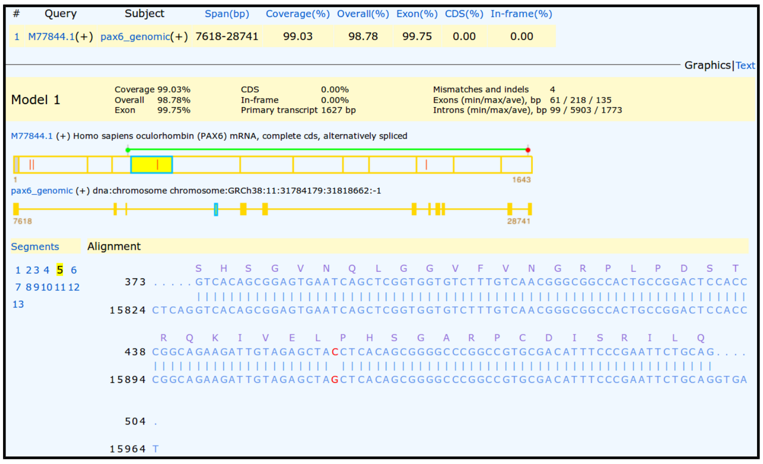
pax6_genomic (+) dna:chromosome chromosome:GRCh38:11:31784179:31818662:-1

```
Alignment
Unaligned:
   length  = 16
   start   = 1
   stop    = 16
```

splign shows alignments for all exons and from those alignments the answer to this question is thus clearly available. To make finding the right spot in the alignment to study easier, I ran **splign** again with an edited version of the **mRNA** (saved as **pax6_mrna_edited.fasta** amongst your cheat files) against the same **Genomic** sequence. Had there been a suitable **mRNA** sequence in the databases, I would have used it for the exercise, but there is not.

You should be able to clearly see the extra mutation is in the **5th** segment.

Focussing on the **5th** segment, the substitution is clear. Using the same methods as were used for the previous question, it is easy to confirm that the variation at amino acid position **33**[3] amounts to:



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	M77844.1(+)	pax6_genomic(+)	7618-28741	99.03	98.78	99.75	0.00	0.00

```
Model 1   Coverage 99.03%   CDS                 0.00%    Mismatches and indels        4
          Overall  98.78%   In-frame            0.00%    Exons (min/max/ave), bp      61 / 218 / 135
          Exon     99.75%   Primary transcript 1627 bp   Introns (min/max/ave), bp    99 / 5903 / 1773
```

M77844.1 (+) Homo sapiens oculorhombin (PAX6) mRNA, complete cds, alternatively spliced

pax6_genomic (+) dna:chromosome chromosome:GRCh38:11:31784179:31818662:-1

```
Alignment
             S  H  S  G  V  N  Q  L  G  G  V  F  V  N  G  R  P  L  P  D  S  T
  373 .....GTCACAGCGGAGTGAATCAGCTCGGTGGTGTCTTTGTCAACGGGCGGCCACTGCCGGACTCCACC
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
15824 CTCAGGTCACAGCGGAGTGAATCAGCTCGGTGGTGTCTTTGTCAACGGGCGGCCACTGCCGGACTCCACC

        R  Q  K  I  V  E  L  P  H  S  G  A  R  P  C  D  I  S  R  I  L  Q
  438 CGGCAGAAGATTGTAGAGCTACCTCACAGCGGGGCCCGGCCGTGCGACATTTCCCGAATTCTGCAG....
      |||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
15894 CGGCAGAAGATTGTAGAGCTAGCTCACAGCGGGGCCCGGCCGTGCGACATTTCCCGAATTCTGCAGGTGA

  504 .

15964 T
```

Affected Patient protein:

CCT → Proline (P)

Canonical protein:

GCT → Alanine (A)

Squinting madly, you can also discover that the variation base positions are:

Affected Patient mRNA:

Base position 459 → C

Wild Type Genomic DNA:

Base position 15915 → G

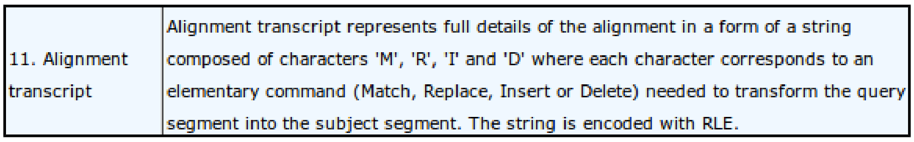
In case you were wondering, chasing these values around is a little more than tragic pedantry. You will need this information later when you investigate **Primer Design**. No need to take notes, I will remind you of what you need when the time comes. Here I just want to show how the values could be determined, if you had to. Not difficult, just tedious!

[3] Proving beyond reasonable doubt that that substitution is exactly at amino acid position **33** requires a little more counting, dividing by **3** and subtracting the number you first thought of. For now, just trust me? I really am more honest than I look.

How do you interpret the **Details** column for exons 1 and 10?

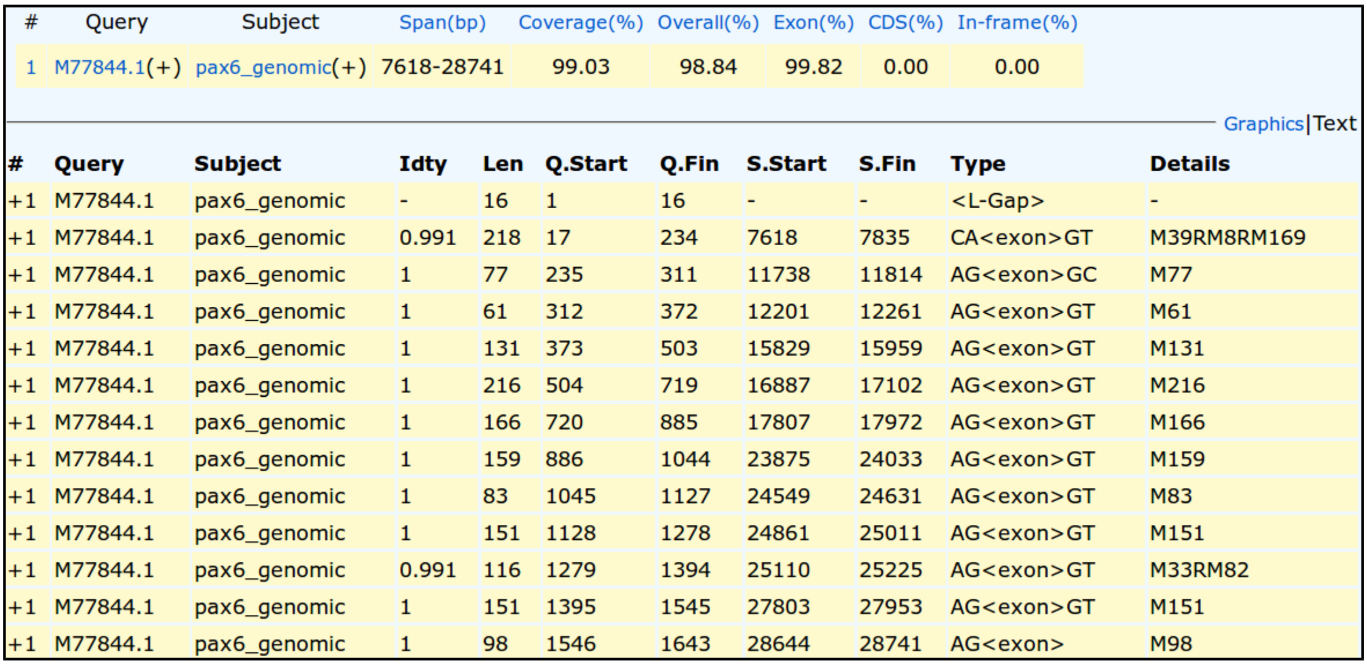
Summary:

The **Details** column shows the alignments of each exon in a compressed format described in the **splign** documentation as illustrated.



11. Alignment transcript	Alignment transcript represents full details of the alignment in a form of a string composed of characters 'M', 'R', 'I' and 'D' where each character corresponds to an elementary command (Match, Replace, Insert or Delete) needed to transform the query segment into the subject segment. The string is encoded with RLE.

The majority of the exon alignments are trivial.



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	M77844.1(+)	pax6_genomic(+)	7618-28741	99.03	98.84	99.82	0.00	0.00

Graphics|Text

#	Query	Subject	Idty	Len	Q.Start	Q.Fin	S.Start	S.Fin	Type	Details
+1	M77844.1	pax6_genomic	-	16	1	16	-	-	<L-Gap>	-
+1	M77844.1	pax6_genomic	0.991	218	17	234	7618	7835	CA<exon>GT	M39RM8RM169
+1	M77844.1	pax6_genomic	1	77	235	311	11738	11814	AG<exon>GC	M77
+1	M77844.1	pax6_genomic	1	61	312	372	12201	12261	AG<exon>GT	M61
+1	M77844.1	pax6_genomic	1	131	373	503	15829	15959	AG<exon>GT	M131
+1	M77844.1	pax6_genomic	1	216	504	719	16887	17102	AG<exon>GT	M216
+1	M77844.1	pax6_genomic	1	166	720	885	17807	17972	AG<exon>GT	M166
+1	M77844.1	pax6_genomic	1	159	886	1044	23875	24033	AG<exon>GT	M159
+1	M77844.1	pax6_genomic	1	83	1045	1127	24549	24631	AG<exon>GT	M83
+1	M77844.1	pax6_genomic	1	151	1128	1278	24861	25011	AG<exon>GT	M151
+1	M77844.1	pax6_genomic	0.991	116	1279	1394	25110	25225	AG<exon>GT	M33RM82
+1	M77844.1	pax6_genomic	1	151	1395	1545	27803	27953	AG<exon>GT	M151
+1	M77844.1	pax6_genomic	1	98	1546	1643	28644	28741	AG<exon>	M98

For example:

For **Exon 2**, **splign** informs us **M77**, meaning “There are **77** bases aligned and they all **M**atch perfectly”.

For **Exon 4**, **splign** informs us **M131**, meaning “There are **131** bases aligned and they all **M**atch perfectly”.

The only **2** interesting entries are those were there are some disagreements. That is, the entries for **Exons 1** and **5**, which, following the documentation, I translate thus:

Exon 1 – M39RM8RM169

An alignment of **218** bases, the first **39** of which **M**atch perfectly (**M39**), there then follows an **R**eplacement (**R**), a further **8** **M**atched bases(**M8**), a second **R**eplacement (**R**) all finished off with **169** **M**atched bases (**M169**).

Exon 10 – M33RM82

An alignment of **116** bases, the first **33** of which **M**atch perfectly (**M33**), there them follows a **R**eplacement (**R**) and a further **82** **M**atched bases(**M82**).

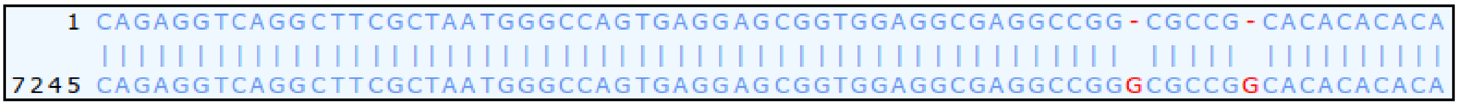
Its a pity there are no **I**nsertions (**I**) and **D**eletions (**D**), but this was the best **mRNA** I could find.

Full Answer:

A point of pedantry to commence. From a different example, which included **InDel**s, I got the display illustrated.

The exon was reported as: **M53IM5IM43**

This implies that the choice of **I**nsertion (**I**) or **D**eletion (**D**) is made to describe the type of variation required to transform the **cDNA** (**Query**) sequence into the **genomic** (**Subject**). Hence the two **InDels** displayed here are considered to be **I**nsertions.



```
   1 CAGAGGTCAGGCTTCGCTAATGGGCCAGTGAGGAGCGGTGGAGGCGAGGCCGG-CGCCG-CACACACACA
     ||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||| ||||||||||
7245 CAGAGGTCAGGCTTCGCTAATGGGCCAGTGAGGAGCGGTGGAGGCGAGGCCGGGCGCCGGCACACACACA
```

Not that it is a vital issue, but I would have thought the other way around was more logical? That is, to consider the **genomic** sequence as the **reference** against which a particular **mRNA** might vary. In other words, what we see here would surely be more relevantly recorded as “This **mRNA/cDNA** has two **D**eletions relative to the **genomic** sequence which, presumably, attempts to represent the norm in the general population”? Just the reflection of an irretrievable pedant, but I am right, nevertheless!!!

In the documentation (see illustration in the **Summary** answer) it enigmatically states “The string is encoded with **RLE**.”. Just in case, **RLE** stands for **Run-length encoding** which is succinctly defined by **Wikipedia**. In a nutshell, it is a very simple form of data compression that recognizes that:

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

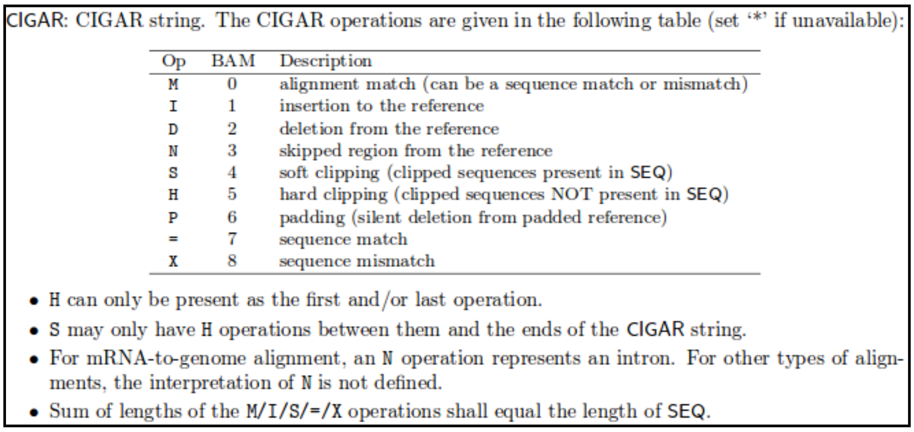
can be compressed to:

60X

which has to be very effective for any data that has runs of identical characters of significant length. This is certainly the case here where one would expect long stretches of **M**s in most alignments. Of course, life would get tricky if the data included numeric characters, but that is not an issue here[4].

I think it worth mentioning, that this way of representing an alignment is a simplification of CIGAR format[5]. This format is used for **SAM** (**S**equence **A**lignment **M**ap) and **BAM** (**B**inary **A**lignment **M**ap, exactly the same as **SAM**, except compressed) files. You will be engulfed in **SAM/BAM** files if you ever do any **N**ext **G**eneration **S**equencing (**NGS**).

So, straight from the **SAM/BAM Format Specification** I copy the table of **CIGAR** enlightenment.



CIGAR: CIGAR string. The CIGAR operations are given in the following table (set ‘*’ if unavailable):

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

Note, in particular, the extended range of **Op**erators and the different meaning associated with the operator '**M**'. The operators '=' and '**X**' are such that any '**M**' is either an '=' or and '**X**' but never both. Which leaves one pondering when one might use '**M**' in preference to either an '=' or an '**X**'?

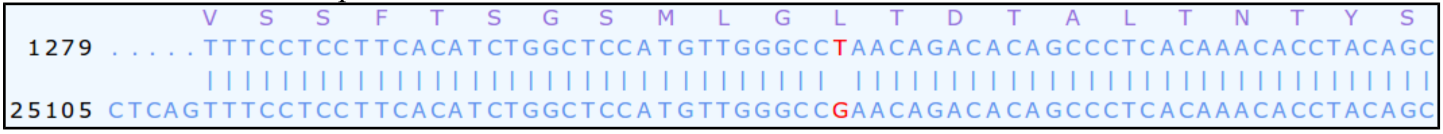
[4] The **Wikipedia** article shows how this complication might be overcome.

[5] There may or may not be some justification for calling the format **CIGAR**, but if there is, I have no idea what it might be.

Where is the 3rd substitution in the mRNA?
Where is the 3rd substitution in the Genomic Sequence?

splign makes one work quite hard to answer this one! Unless I am missing something.

From the alignment of **Exon 10**, the exon including the **3rd R**eplacement, with a bit of squinting, it can be confirmed that the **3rd R**eplacement is at:

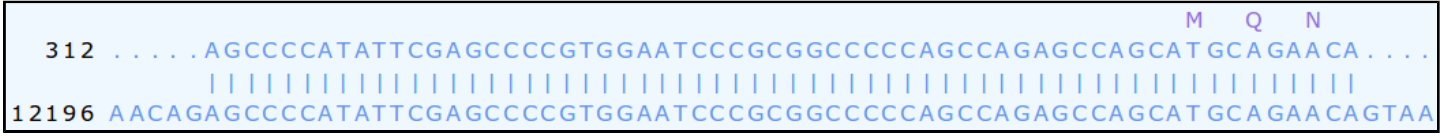


```
           V  S  S  F  T  S  G  S  M  L  G  L  T  D  T  A  L  T  N  T  Y  S
 1279 .....TTTCCTCCTTCACATCTGGCTCCATGTTGGGCCTAACAGACACAGCCCTCACAAACACCTACAGC
           ||||||||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||
25105 CTCAGTTTCCTCCTTCACATCTGGCTCCATGTTGGGCCGAACAGACACAGCCCTCACAAACACCTACAGC
```

Base pair position **1,312** of the **mRNA**

Base pair position **25,143** of the **genomic** sequence

It might also have been relevant to ask which amino acid position corresponded to the **R**eplacement. To discover this one would need to look at the alignment of **Exon 3**, where the coding begins.



```
                                                               M  Q  N
  312 .....AGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCCCCAGCCAGAGCCAGCATGCAGAACA....
           |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
12196 AACAGAGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCCCCAGCCAGAGCCAGCATGCAGAACAGTAA
```

More squinting, and I conclude the **A** of the **ATG** representing the initial **Methionine** of the protein coding region is at position **363**. That is, the **5' UTR** ends at position **362**. So the **R**eplacement is at:

Base position **1312** – **362** = **950** of the protein coding region of the **mRNA**.

As **950** / **3** is **316** remainder **2**, the **R**eplacement is at codon position **2** of the **317th** amino acid of the protein.

Cannot help thinking that **splign** might have helped a bit more here?

I also reflect that I cannot fully recall why I wanted to know where the mutation was, especially given we have decided to reject any chance that it might be a mutation of consequence. Oh well, some things a man must do, just because they are there to be done!!

Time to move on … without checking my arithmetic. Bound to be right, I used to be a mathematics teacher you know! Several lifetimes ago.

Postscript:

After the passage of many months, I now recall why I obsessed as to the position of this amino acid substitution. I wondered if it was in the region of one of the major domains of this protein. If it was, it might increase its chances of being significant?

Well, it is not. In the last exercise, we discovered that:

The **Paired-box** domain is between positions **4** and **128** (**Consensus isoform**) or **4** and **142** (**isoform 5a**).

The **Homeo-box** domain is between **214** and **266** (**Consensus isoform**) or **228** and **280** (**isoform 5a**).

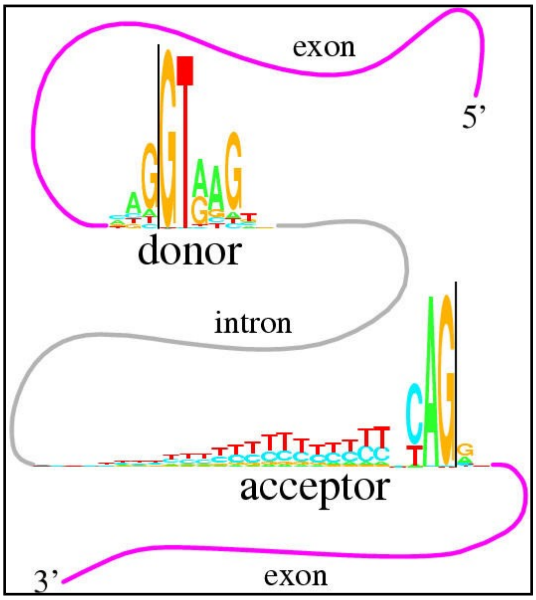
So the **Substitution**, at position **317**, is in a relatively neutral region and so, maybe, less likely to be of great consequence?

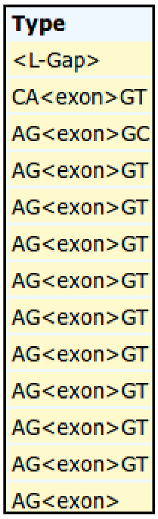
Compare the predicted **splign** intron/exon boundaries with the conservation suggested by the logo?
What deviation(s) from the model suggested by the logo can you see?

You may have gathered, I rather like this logo, although I rather think it is leading me to make the same point a trifle to often?

The logo is in almost **100%** agreement with the predictions of **splign**.

As you will have noted previously, when looking at the **Ensembl** predictions of exons locations of a similar transcript of the **PAX6** human gene (previous Practical), there is a single exception.





Type
<L-Gap>
CA<exon>GT
AG<exon>GC
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>

The easiest way to show this in the **splign** output is to look at the **splign** text output again.

The **Type** column records the type of all the **<exon>** alignments it predicts. It also records **2 flanking intron base pairs**.

It is clear that the only time the **splign** prediction deviates from the model suggested by the logo is at the end of the **2nd** exon. Here there is **GC** rather then **GT**. Well, nothing is perfect!

From your investigations of **Local Alignment**:

Why do you suppose your aligned exons are not presented in the correct positional order?

To **Matcher**, the logical order in which to present the alignments is that governed by quality rather than position. So, the highest scoring alignment, rather than the first exon alignment, will be at the top of the list. I think this is generally logical. Once again, the program **splign**, knowing it was looking for an ordered set of exons, was more specifically logical.

DPJ – 2017.12.23

Discussion Points and Casual Questions arising from the Instructions Text.

Notes:

Work in progress I fear.

The intention is to provide a full consideration of some issues skimmed over in the exercise proper.

If you are attending a "supervised" presentation of the exercise, I would hope to have conducted a live discussion of all these issues to an extent that reflects:

- the depth that seems appropriate
- the time available
- the degree to which the issues seem to match the interests of the class
- how many of you are awake

Here, I hope to write out very full answers were such a response exists. Accordingly, I suggest you will not need to read much of many of these discussions. There will be much detail of interest to rather few of you. Possibly a bit self indulgent, but I wish to make a note of all the background I have discovered while writing these exercises.

In a nutshell, the exercises are trying to make very general points avoiding too much detail. Nevertheless, I record the detail outside the main exercise text, just in case it might be if interest. Some of the answers to the "**Casual Questions**" are exceedingly trivial. Some of the "**Discussion Points**" are exceedingly long and rambling. You have been warned.

How would you interpret this picture?

What do the diagonal(ish) lines represent?

Exons

What are the gaps in between the lines?

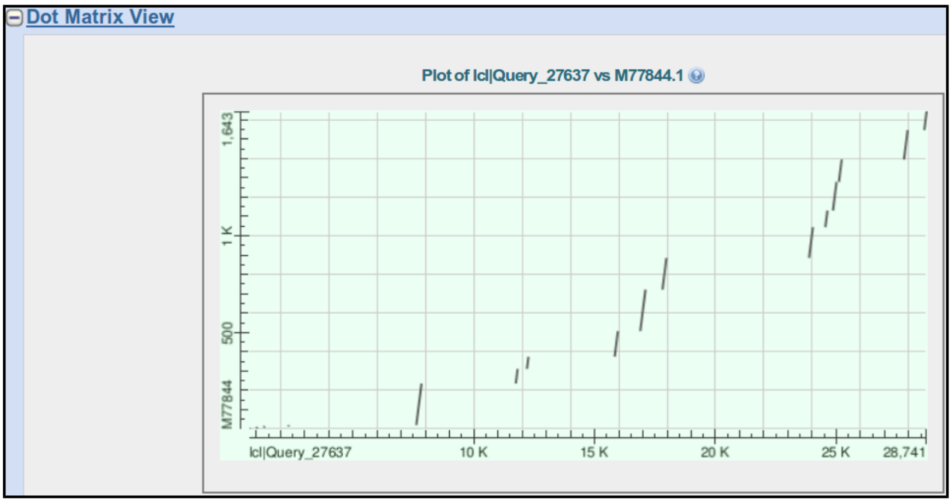
Introns

Which axis represents the genomic sequence and which the mrna?

The **Horizontal** axis is the **genomic** sequence.

The **Vertical** axis is the **mRNA**.

The axes are not in strict proportion, but the **genomic** axis is longer than the **mRNA** axis, which feels and looks intuitively correct.



How many are there and do they correspond nicely to the lines of the **Dot Matrix View**?

How many exons would you say this mrna has?

If one was to forgive the strange “bits” at the start, would you say **blast** seems to have done a reasonable job here?

How do you feel about the results this time?

Any theories?

I cannot help you here? Maybe some sequencing artefact? It is a sequence of some antiquity after all.

DPJ – 2017.12.23