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 3: Database Searching

Tuesday 20 February 2018

Searching for sequence similarities in databases.

The most popular way to investigate a sequence has always been to compare it with one of the sequence databases now accessible from sites all over the world. When sequences databases were more sparsely populated than now, the objective was to search hopefully, not always with success, for any convincingly similar sequence(s). When such a match was discovered, it could be supposed that known properties of the "similar" database sequence might provide insight to the properties of the query sequence. Now, the databases are full of sequences representative of most interesting conditions. Similarity searches are conducted in the expectation of finding many close "hits" for almost any sequence. Fewer database searches are conducted in complete ignorance of what the query sequence might be.

Database Searching to determine gene structure.

Here, take the **PAX6** genomic DNA sequence retrieved from **Ensembl** and conduct two searches analogous to those run in the **Ensembl** pipeline (or the equivalent **NCBI** pipeline for **Map Viewer**). Results should confirm that which has already been discovered using other sources.

blast is not the only sequence database searching program available, but it is the most popular by a very long way. **blast** searches are offered in many forms by many servers all over the world, but the most comprehensive and reliable service has to be that offered by the **NCBI**.

Comparing **Genomic** sequence against **mRNA** sequences to predict exon splicing alternatives.

Go to the **NCBI** homepage at:

```
http://ncbi.nlm.nih.gov
```

Select the BLAST option (from the **Popular Resources** list). In the **Basic BLAST** section, select nucleotide blast. Use the **Enter Query Sequence Browse** (or **Choose File**) button to upload the file:

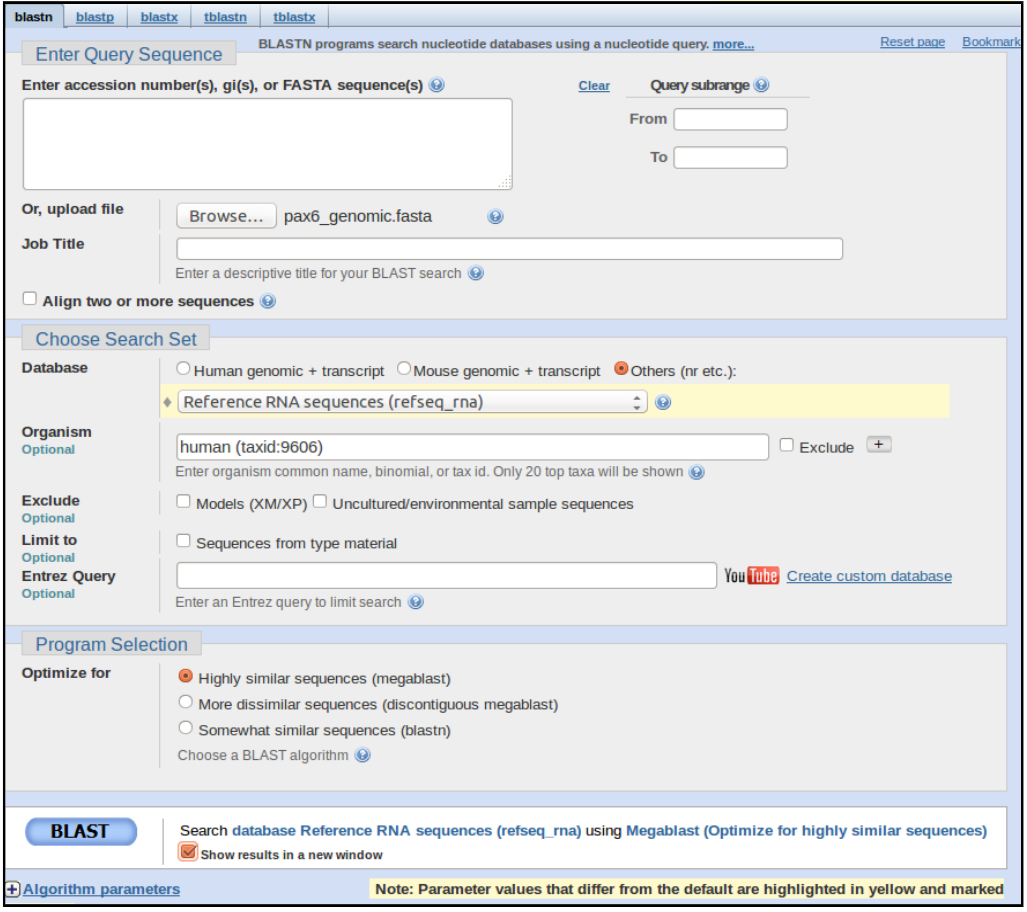
pax6_genomic.fasta.

For results like those used by **Ensembl** to predict **PAX6** transcripts, you must compare your genomic sequence to a reliable set of human mRNA/cDNA (or similar) sequences.

In the **Choose Search Set** section, set the **Database** to **Reference RNA sequences (refeseq_rna)**.

You are now able to specify an **Organism**, choose **human (taxid:9606)**.

blast is now set to compare the **PAX6** genomic region with all **Human** mRNA sequences in **RefSeq**.



BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange From To

Or, upload file: pax6_genomic.fasta

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Reference RNA sequences (refseq_rna)

Organism (Optional): human (taxid:9606) □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to (Optional): □ Sequences from type material

Entrez Query (Optional)

Enter an Entrez query to limit search

Program Selection

Optimize for:
● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database Reference RNA sequences (refseq_rna) using Megablast (Optimize for highly similar sequences)

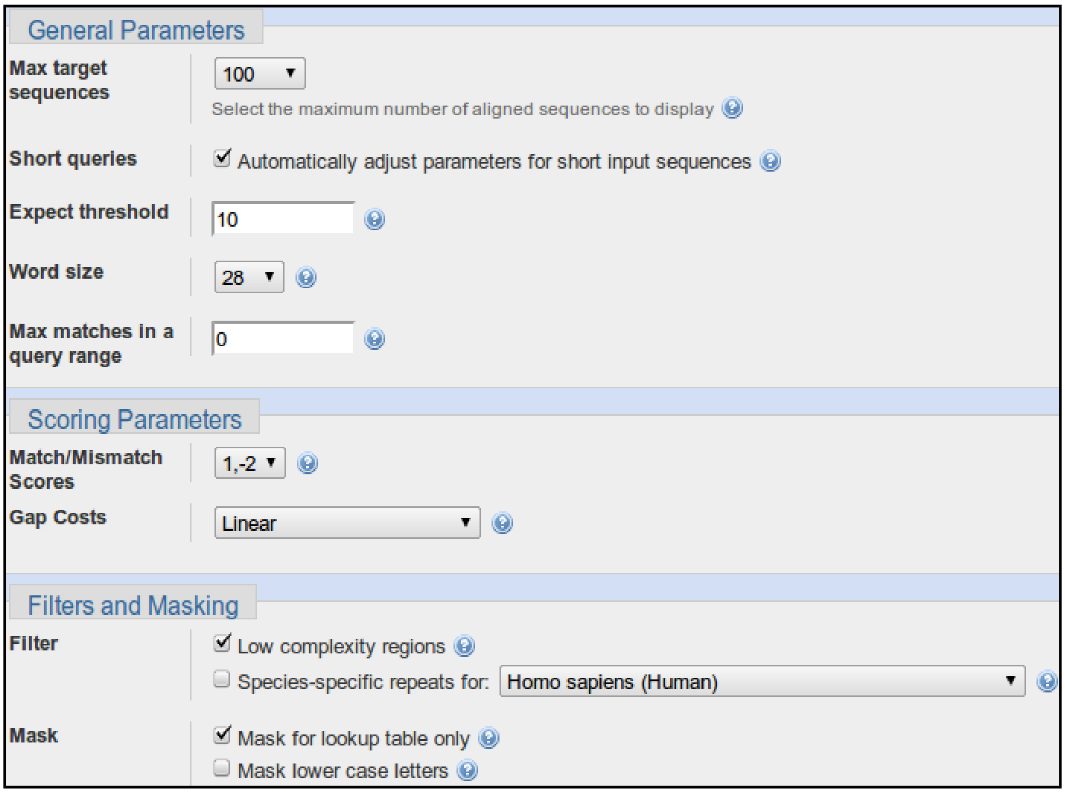
☑ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked

Note that the default **Program Selection** is **Highly similar sequences (megablast[1])**, which seems appropriate here as all the mRNA that correctly match should surely do so almost perfectly.

[1] **megablast** is a less sensitive but even faster version of **blast** only suitable when, as now, almost identical matches are sought.

Click on the **Algorithm Parameters** button. The defaults are fine here, but before starting your search, try changing the **Program Selection** and observing the different **Algorithm Parameters**.



General Parameters

Max target sequences 100 ▾
Select the maximum number of aligned sequences to display

Short queries ☑ Automatically adjust parameters for short input sequences

Expect threshold 10

Word size 28 ▾

Max matches in a query range 0

Scoring Parameters

Match/Mismatch Scores 1,-2 ▾

Gap Costs Linear ▾

Filters and Masking

Filter ☑ Low complexity regions
☐ Species-specific repeats for: Homo sapiens (Human) ▾

Mask ☑ Mask for lookup table only
☐ Mask lower case letters

The default settings of all shared parameters are identical for the two slower more sensitive **Program Selections**.

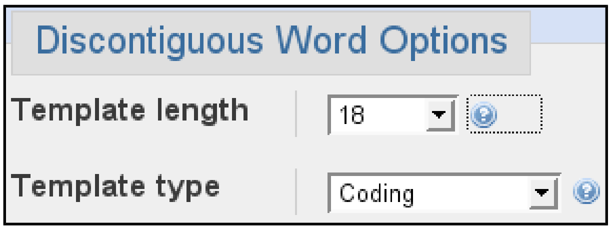
There are differences for **megablast**, where speed is of the essence and sensitivity can be sacrificed.

Smaller **Word sizes** slow searches but increase sensitivity. For **megablast** the default **Word size** is **28** otherwise it is **11**.

Gapped alignment is time consuming and, by default, considered more crudely by **megablast** than the other two algorithms[2].

Filtering and Masking matches with organism specific repeats and/or low complexity regions takes time, and so only avoiding **Low complexity regions**[3] is on by default for all **Program Selections**.

When **discontinuous megablast** is selected, an extra options section appears. Discussing how this flavour of **blast** works is a little beyond the scope of these note, but briefly. Unlike the other **Program Selections**, **discontinuous megablast** does not just look for exactly matching "words" of given size as a first step towards identifying matching regions between sequences. It looks for a pattern of matching bases within a word. For example, the default choice assumes your query is **coding** and looks for **11** matching bases within a word of **18**. Approximately, every third base is allowed not to match. Biologically, this can be justified as allowing for third codon position wobble. For more detail, use the appropriate ? button. Notice there are ? buttons by every parameter selection. Try one or two. In the process, discover:



Discontiguous Word Options

Template length 18 ▾

Template type Coding ▾

When would **Mask lower case letters** be a useful thing to do?

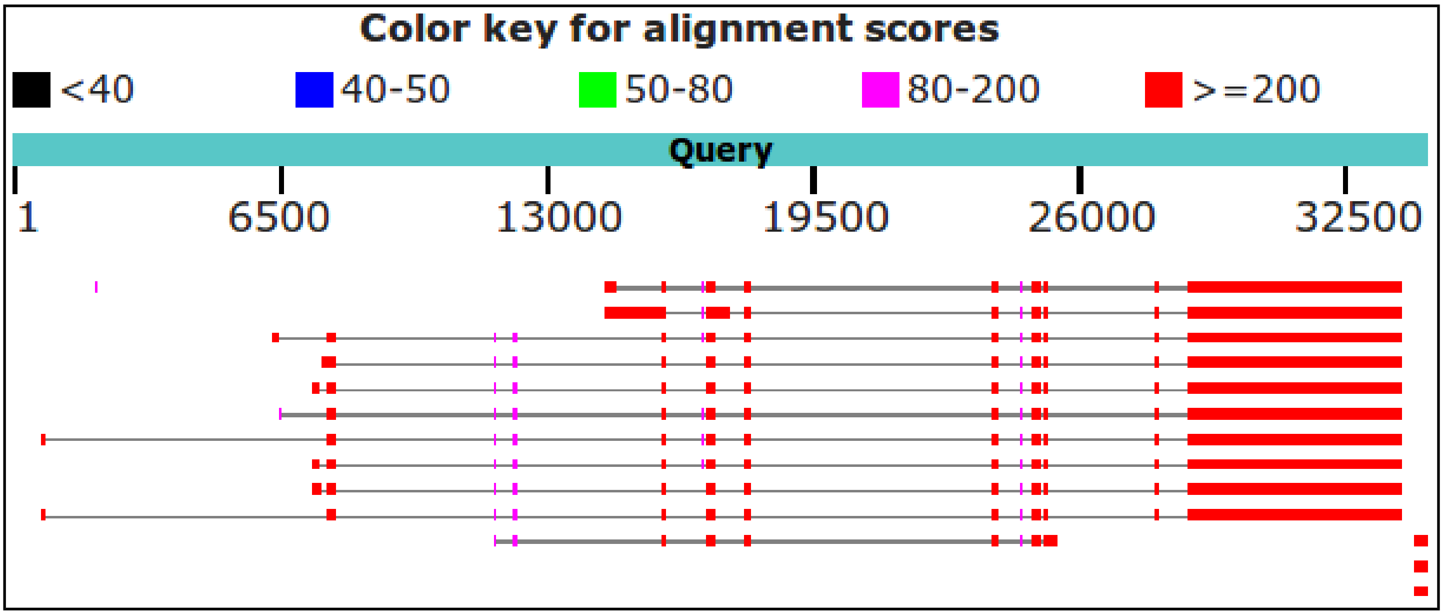
Automatically adjust parameters for short input sequences is independent of **Program selection**, and so remains unaltered.

Which parameters would **blast** need to **automatically adjust** to cater for short input sequences (such as primers being tested for uniqueness), and why?

[2] By default, **megablast** uses **Linear Gap Costs**. That is, it just multiplies the size of the gap with the **Mismatch** penalty. The other two algorithms employ the more common **Affine** strategy, using **Existence** and **Extension** penalties. For more about **Gap Penalties**, go here.

[3] This filter avoids finding "hits" supported only by matches in regions not specific to the query. For example, a polyA tail cannot help to identify a specific mRNA as it is present is all mRNAs. The use of this filter will be evident when we look at the **blast** output.

Finally, ensure all the parameter defaults are back in place[4] and that **megablast** is the **Program Selection**, ask **blast** to **Show results in a new window** and then click on the **BLAST** button. Impressively swiftly, you will have results. At the top of which will be a graphical overview.



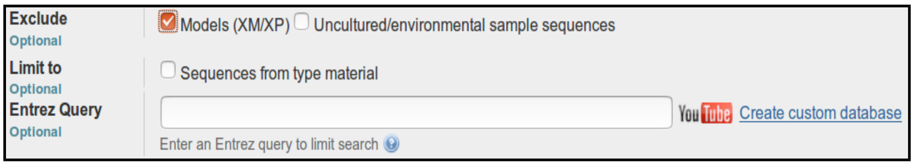
This graphic implies that there are **11** full length matches between the genomic sequence and mRNAs in **RefSeq**. The **RefSeq** entries had to be "gapped" in order to compensate for the introns that are represented in the genomic sequence but not in the mRNA sequences. The **red blocks** therefore represent very closely matching (>=200 brownie points) exons, the lines joining the **red blocks** represent introns that have been spliced out. All **11** full length hits match reasonably uniformly except for the first few exons, implying significant variation in the **5´ UTR**.

Why do you suppose that a few of the exons of the first **11** matches do not achieve the maximum score?

Explain why one exon in the reasonably consistent region, does not appear in all of the transcript matches?

In a previous Practical, you discovered directly that there were **11** high quality "**NM_**" **PAX6** transcripts in **RefSeq**.

Until recently, there was a further **9** "**XM_**" **PREDICTED** transcripts. However, in the last release of **RefSeq**, the **9** less reliable **XM_** transcripts were removed and so were not detected by **blast**. **Ensembl** claimed to have used most, if not all, the high quality **NM_ RefSeq** sequences to aid its transcript predictions. **Ensembl** would have ignored the **XM_ PREDICTED RefSeq** sequences even if they still existed.



Exclude
Optional
☑ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to
Optional
□ Sequences from type material

Entrez Query
Optional
Enter an Entrez query to limit search

blast just sees sequences and, by default, will not be influenced by the quality of the support for their existence. Run as in this exercise, **blast** would always report all **RefSeq PAX6** mRNAs matching the **PAX6** genomic region convincingly, independently of how questionably they are evidenced. However, you could have filtered the target database(s) in various ways, including choosing to **Exclude** all **Modules(XM/XP)** (that is all the more questionable mrna sequences and their amino acid translations). This would not be appropriate here as we wish to mimic the approach of the **NCBI** genome databases which ***DO*** consider **XM/XP** sequences should they exist.

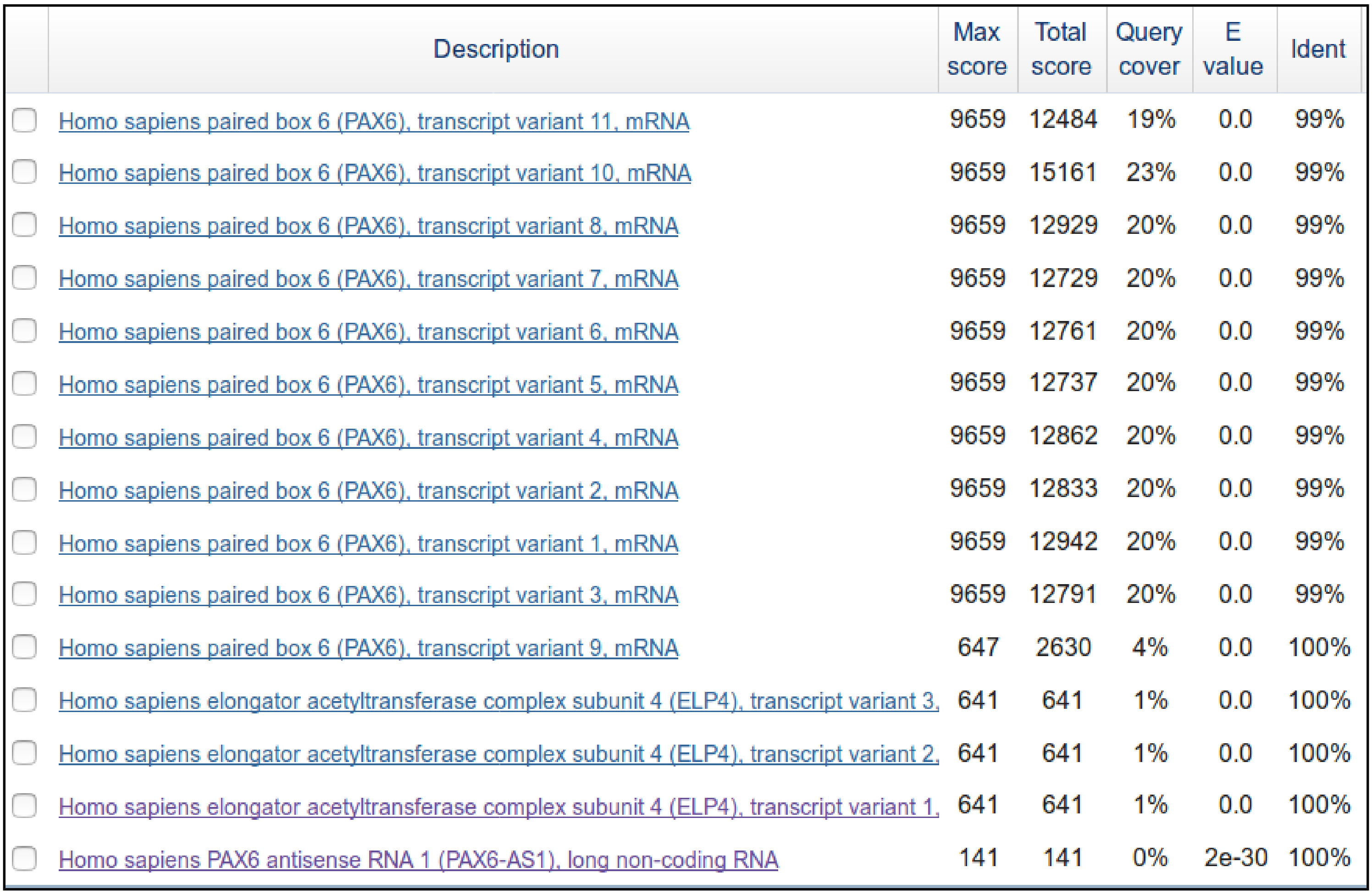
There is a point to pursuing all this detail. You reference a collection of interdependent databases, all of which are updated regularly. More often than not you will notice inconsistencies due to asynchronous updates and differences in database management/interpretation policy. A small price to pay for such a rich source of information, but one of which I suggest it is wise to be aware.

The message of the particular **blast** search here is that it is so easy to predict the same **PAX6** transcripts as you discovered in **MapViewer**, just with a simple **blast** search. That is, you can look things up, or work most of it out for yourself.

[4] If you have any non-default settings, they should be highlighted in yellow.

If you hover over the graphical hits, their origin will be displayed above the graphic[5].

Below the **Graphic Summary** are the **Descriptions**, a simple list of the **15** matches represented in the graphic.



	Description	Max score	Total score	Query cover	E value	Ident
□	Homo sapiens paired box 6 (PAX6), transcript variant 11, mRNA	9659	12484	19%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 10, mRNA	9659	15161	23%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 8, mRNA	9659	12929	20%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 7, mRNA	9659	12729	20%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 6, mRNA	9659	12761	20%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 5, mRNA	9659	12737	20%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 4, mRNA	9659	12862	20%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 2, mRNA	9659	12833	20%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 1, mRNA	9659	12942	20%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 3, mRNA	9659	12791	20%	0.0	99%
□	Homo sapiens paired box 6 (PAX6), transcript variant 9, mRNA	647	2630	4%	0.0	100%
□	Homo sapiens elongator acetyltransferase complex subunit 4 (ELP4), transcript variant 3,	641	641	1%	0.0	100%
□	Homo sapiens elongator acetyltransferase complex subunit 4 (ELP4), transcript variant 2,	641	641	1%	0.0	100%
□	Homo sapiens elongator acetyltransferase complex subunit 4 (ELP4), transcript variant 1,	641	641	1%	0.0	100%
□	Homo sapiens PAX6 antisense RNA 1 (PAX6-AS1), long non-coding RNA	141	141	0%	2e-30	100%

These are such that:

- The top **11** hits, corresponding to the **11** full length hits of the **Graphic Summary,** are the quality (i.e. **NM_** entries with good supporting evidence) **RefSeq** transcripts.

- There follows, corresponding to the **3** small red blobs in the extreme bottom right of the **Graphic Summary**, **3** hits that are the ends of **mRNA**s for the **ELP4** gene. They are exactly where you should expect them to be, assuming you paid full attention to the **ELP4** transcript predictions shown in both the Ensembl and Map Viewer displays of the **Genomic** region around **PAX6**. Reject these contemptuously, they do not pertain to our investigation of **PAX6**.

- The **15th** match, corresponding to the barely visible tiny smudge match to the left of the top **Graphic Summary** hit, is recorded as “**uncharacterized**” and fails to fit in with my story, so I ignore it![6]

So, this **blast** search suggests the existence of **11 PAX6** transcripts supported by **RefSeq** data, as is reported by **Map Viewer**. Also, the results are broadly consistent with the information discovered in **Ensembl**.

Which of the **Refseq PAX6** transcripts corresponds to **isoform 5a**?

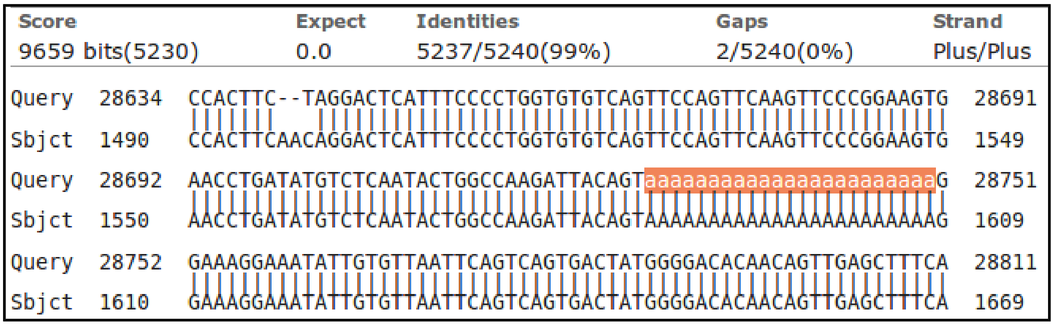
[5] Or you could just read the textual list that follows the graphic if you wish to insist on the simplistic.

[6] Actually, I see now it is a single exon of the **PAX6-AS1** entity pursued so vigorously in the last exercise. Those of you foolish enough to read all the ramble of my answers to questions will recall **PAX6-AS1** with glee! Yep … ignore it.

Moving further down the results you will come to the alignments between the **PAX6** genomic sequence and the matching database entries. All similarity searches use local alignment strategies[7], so you should not be surprised to see a number of alignments for each "hit" in the list. Here we have a genomic query sequence aligned exclusively with mRNA sequences from **RefSeq**. The expectation is therefore to find an alignments corresponding to exons. The alignments are ordered by quality, though you are provided with a **Sort by:** menu to alter the order to taste[8].

Look at the first alignment for the best matching **PAX6** transcript. It is the alignment of the very last exon of a **RefSeq** transcript with the end of the gene you exported from **Ensembl**.

Notice the lower case string of '**a**'s. The case indicates that they were ignored (**filtered**) as a **Low complexity region** whilst **megablast** was looking for identically matching words that might suggest matching regions. By themselves, the '**a**'s are not sufficient evidence that a biological match exists. Only because the surrounding sequence is compellingly similar, can it be assumed that such a match does exist. The '**a**'s are replaced (lower case to indicate they were filtered) when the final alignment is computed. If you look a little further down the same alignment, you will see several other runs of '**a**'s and '**t**'s for which the same explanation applies.



Score	Expect	Identities	Gaps	Strand
9659 bits(5230)	0.0	5237/5240(99%)	2/5240(0%)	Plus/Plus

```
Query  28634  CCACTTC--TAGGACTCATTTCCCCTGGTGTGTCAGTTCCAGTTCAAGTTCCCGGAAGTG  28691
              |||||||   ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1490   CCACTTCAACAGGACTCATTTCCCCTGGTGTGTCAGTTCCAGTTCAAGTTCCCGGAAGTG  1549

Query  28692  AACCTGATATGTCTCAATACTGGCCAAGATTACAGTaaaaaaaaaaaaaaaaaaaaaaaG  28751
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1550   AACCTGATATGTCTCAATACTGGCCAAGATTACAGTAAAAAAAAAAAAAAAAAAAAAAAG  1609

Query  28752  GAAAGGAAATATTGTGTTAATTCAGTCAGTGACTATGGGGACACAACAGTTGAGCTTTCA  28811
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1610   GAAAGGAAATATTGTGTTAATTCAGTCAGTGACTATGGGGACACAACAGTTGAGCTTTCA  1669
```

[7] To use a global approach would be to imply that you were only interested in database entries that matched your query sequence from end to end. Generally, this is not true. You would usually be interested in a database sequence that was similar over any significant region.

[8] Why not try them? End up with the alignments for the top hit in **E value** order.

Comparing **Genomic** sequence against **Protein** sequences to predict **Coding** exons.

Now use a version of **blast** (called **blastx**) to compare your genomic sequence with a protein database. **blastx** will translate a DNA query sequence in all six reading frames and compare each translation with a protein sequence database. Thus, in a similar fashion to that employed by the **Ensembl** pipeline, protein coding regions of the genomic DNA can be identified. For clarity, we will use only the well annotated human proteins of the **SwissProt** section of **Uniprot**. First go to the home of **blast** at:

```
http://blast.ncbi.nlm.nih.gov/Blast.cgi
```

Select

. Use the **Enter Query Sequence Browse** (or **Choose File**) button to **upload file** pax6_genomic.fasta.

In the **Choose Search Set** section, set the **Database** to **UniProtKB/Swiss-prot prot(swissprot)**. Specify the **Organism** as **human (taxid:9606)**.

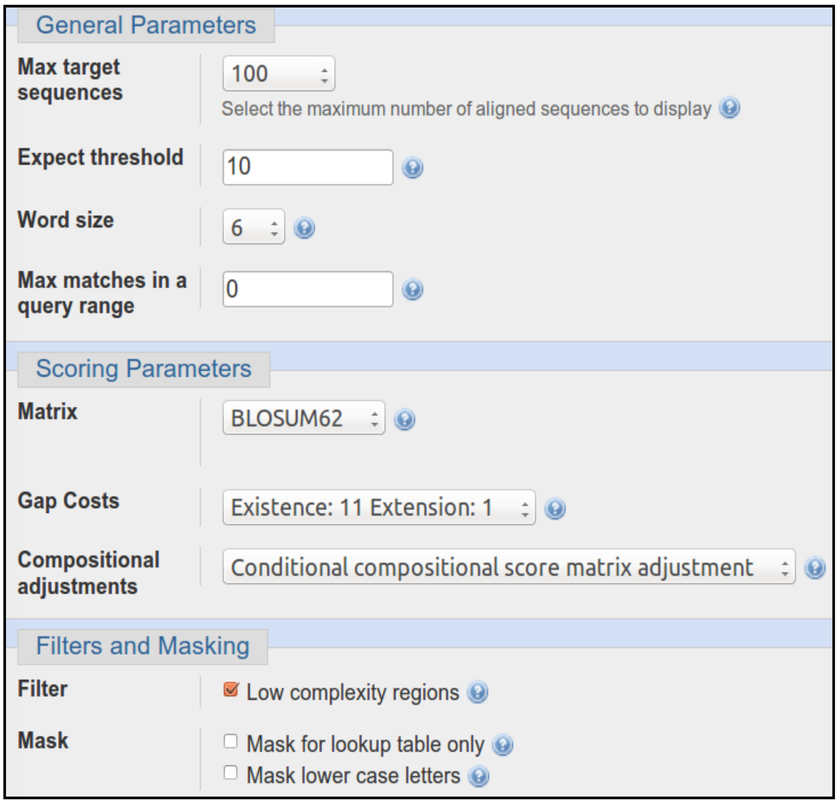
Take a look at the **Algorithm parameters**[9].

The **Word size** choice is **2, 3** or **6** . The default is **6**. We seek very close matches here, so the largest **Word size** would seem appropriate.

The default scoring matrix is **BLOSUM62**, but choices from both the **BLOSUM** and **PAM** families are offered.

The **Compositional adjustments** parameter offers the opportunity to refine the chosen scoring matrix to reflect the residue composition of the sequences being compared in one of a number of ways. Click on the relevant ? button for further enlightenment. I must admit, I was left with questions after reading the **Help**, but some attempt to customise the evaluation of an alignment to reflect sequence composition does seem like an excellent idea.

Low complexity regions will be filtered by default.



General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Expect threshold: 10

Word size: 6

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

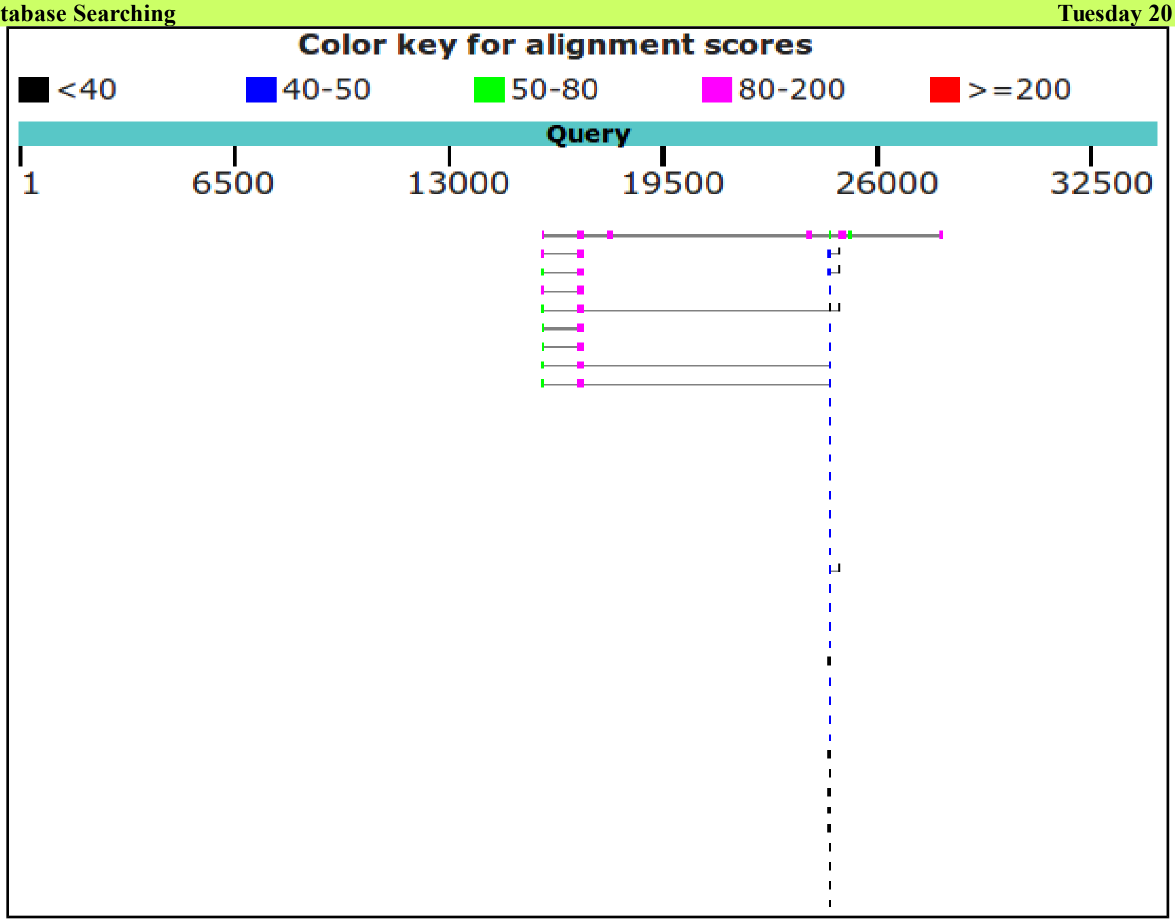
Filter: ☑ Low complexity regions

Mask: □ Mask for lookup table only
□ Mask lower case letters

Change nothing other than to ask **blast** to **Show results in a new window** and click the **BLAST** button.

After minimal thought, **blastx** will thrust its conclusions before you. Hover over the graphical hits for identification.

[9] Here I will assume we have talked about these parameter and you are reasonably well informed of the issues.

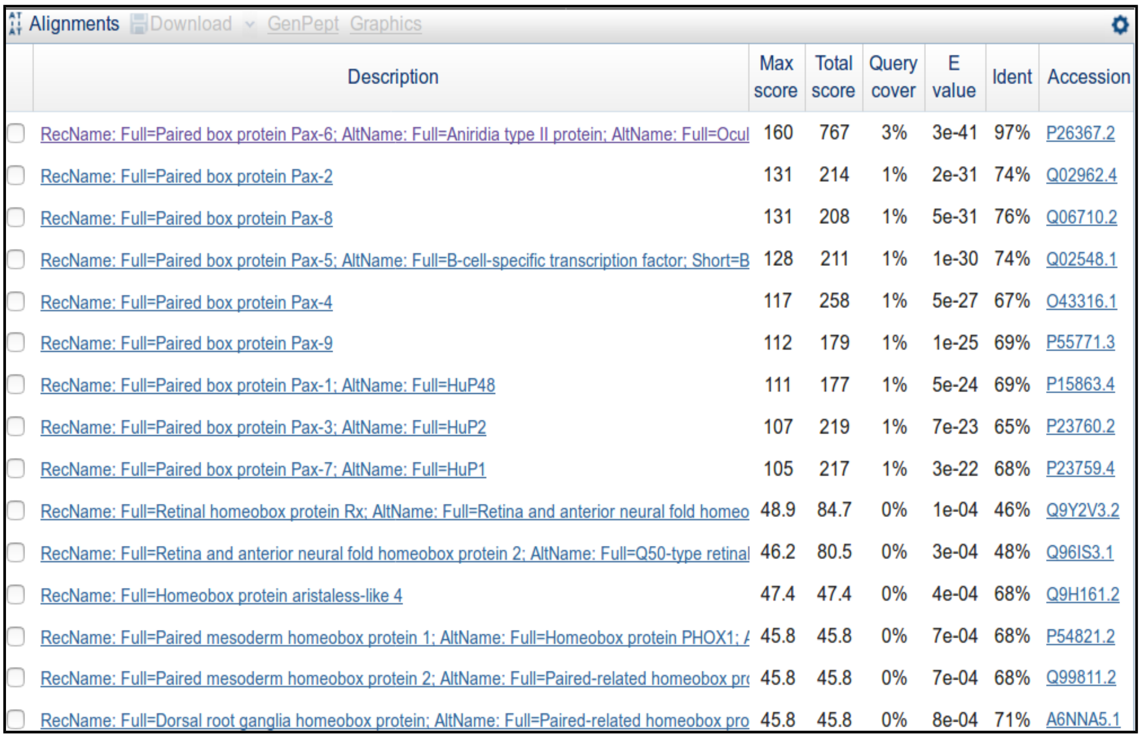


What are the **9** strongest matches around base position **16,750**?

Why would you expect exactly **9** matches around this point?

What do you make of the plethora of matches around **24,000**?

Move down to the textual list of the matches. Hopefully as you fully expected you will find the expected number of **Paired box** matches at the top of the list followed by many many **Homeobox** matches.

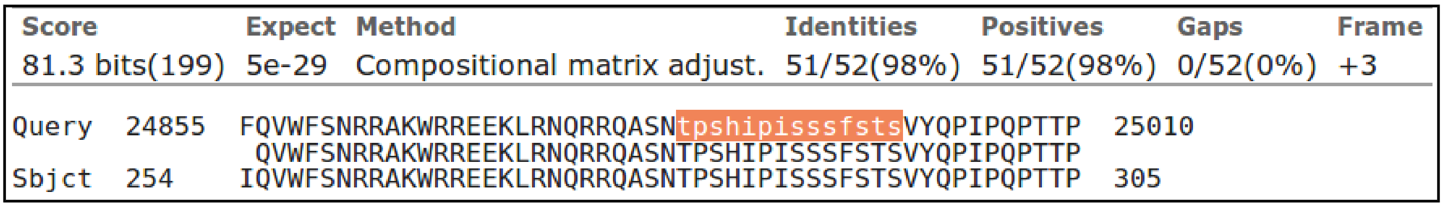


	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	RecName: Full=Paired box protein Pax-6; AltName: Full=Aniridia type II protein; AltName: Full=Ocul	160	767	3%	3e-41	97%	P26367.2
□	RecName: Full=Paired box protein Pax-2	131	214	1%	2e-31	74%	Q02962.4
□	RecName: Full=Paired box protein Pax-8	131	208	1%	5e-31	76%	Q06710.2
□	RecName: Full=Paired box protein Pax-5; AltName: Full=B-cell-specific transcription factor; Short=B	128	211	1%	1e-30	74%	Q02548.1
□	RecName: Full=Paired box protein Pax-4	117	258	1%	5e-27	67%	O43316.1
□	RecName: Full=Paired box protein Pax-9	112	179	1%	1e-25	69%	P55771.3
□	RecName: Full=Paired box protein Pax-1; AltName: Full=HuP48	111	177	1%	5e-24	69%	P15863.4
□	RecName: Full=Paired box protein Pax-3; AltName: Full=HuP2	107	219	1%	7e-23	65%	P23760.2
□	RecName: Full=Paired box protein Pax-7; AltName: Full=HuP1	105	217	1%	3e-22	68%	P23759.4
□	RecName: Full=Retinal homeobox protein Rx; AltName: Full=Retina and anterior neural fold homeo	48.9	84.7	0%	1e-04	46%	Q9Y2V3.2
□	RecName: Full=Retina and anterior neural fold homeobox protein 2; AltName: Full=Q50-type retinal	46.2	80.5	0%	3e-04	48%	Q96IS3.1
□	RecName: Full=Homeobox protein aristaless-like 4	47.4	47.4	0%	4e-04	68%	Q9H161.2
□	RecName: Full=Paired mesoderm homeobox protein 1; AltName: Full=Homeobox protein PHOX1; /	45.8	45.8	0%	7e-04	68%	P54821.2
□	RecName: Full=Paired mesoderm homeobox protein 2; AltName: Full=Paired-related homeobox pr	45.8	45.8	0%	7e-04	68%	Q99811.2
□	RecName: Full=Dorsal root ganglia homeobox protein; AltName: Full=Paired-related homeobox pro	45.8	45.8	0%	8e-04	71%	A6NNA5.1

Why do you suppose the **Paired box** matches precede the **Homeobox** matches?

How do you suppose the **Max matches in a query range** parameter might be of value if this order was reversed?

Take a look at the alignments. You will see many places where regions have been filtered as non-informative. I suggest the one illustrated was filtered because it would match anywhere that was sufficiently **Serine** rich.



Score	Expect	Method	Identities	Positives	Gaps	Frame
81.3 bits(199)	5e-29	Compositional matrix adjust.	51/52(98%)	51/52(98%)	0/52(0%)	+3

```
Query  24855   FQVWFSNRRAKWRREEKLRNQRRQASNtpshipisssfstsVYQPIPQPTTP  25010
                QVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP
Sbjct  254     IQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP  305
```

How does this “non-informative” region match expectations suggested by **SMART** and the **Feature table** of **UniprotKB** for **PAX6_HUMAN**?

Iterative Database Searching to discover and align sequence families (psi-blast & cobalt).

PSI-BLAST is used to find a comprehensive set of relatives of a protein. First, **BLAST** is used to find closely related proteins. From an alignment of these proteins a general "profile" (a **P**osition **S**pecific **S**coring **M**atrix - **PSSM**) is computed. A **PSSM** is very similar in concept and purpose to an **HMM** profile in that it summarises significant features present in the sequences it represents.

A further search of the protein database is then run using the **PSSM** as a query, and a larger more widely associated group of proteins is found. This larger group is aligned and used to construct another **PSSM**, and the process is repeated until no more significantly matching new sequences can be detected, or the user tires of the whole process.

PSI-BLAST is integrated into the **Secondary Structure Prediction** system **Jpred**. Whenever **Jpred** is asked to compute structure form a single protein sequence, it will use **PSI-BLAST** to construct an aligned family of protein sequences to enable an improved prediction. An aligned family of proteins is a much better starting point than any single protein sequence.

Similar ideas are used by the domain database **PFAM** to create large alignments of domain regions. Hopefully there will be time to glance at **PFAM** alignments and **HMM**s.

Here we will use **PSI-BLAST** directly from the **NCBI** on the **Paired DOMAIN** of the **PAX6** protein that you saved in a file earlier. It should be possible to detect a large family of **PAX** domains and to eventually multiply align them generating something like the alignment from the **PFAM** database.

To investigate **PSI-BLAST** go first to the **NCBI** Home page at:

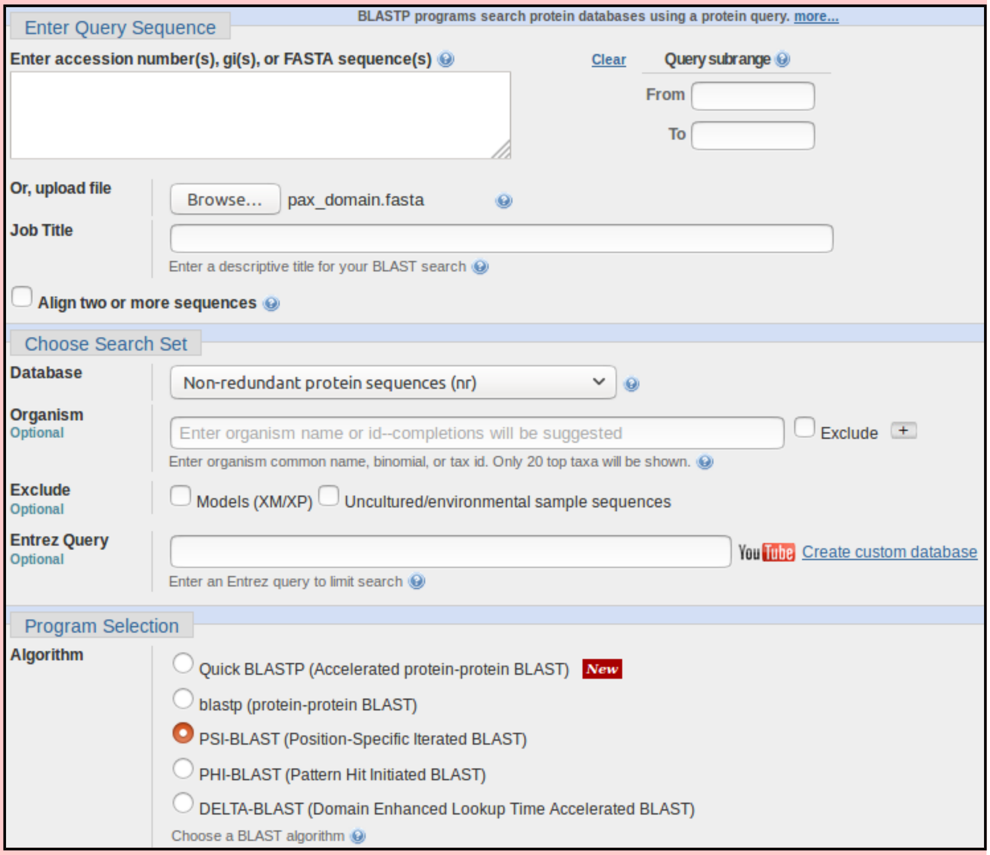
`http://www.ncbi.nlm.nih.gov/`

Click on the **BLAST** option from the **Popular Resources** menu.

Select Protein BLAST from the **Web BLAST** section.

Upload the **PAX6** paired box domain sequence (stored in the file **`pax_domain.fasta`**) using the appropriate **Browse** button.

Select **PSI-BLAST** from the **Program Selection** section. Leave all the others options at their default settings, particularly the option to search all the proteins available.



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange

From

To

Or, upload file: Browse... pax_domain.fasta

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional): YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm:

○ Quick BLASTP (Accelerated protein-protein BLAST) New

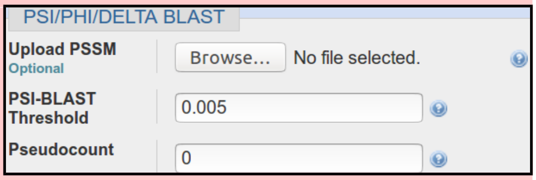
○ blastp (protein-protein BLAST)

● PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm



PSI/PHI/DELTA BLAST

Upload PSSM (Optional): Browse... No file selected.

PSI-BLAST Threshold: 0.005

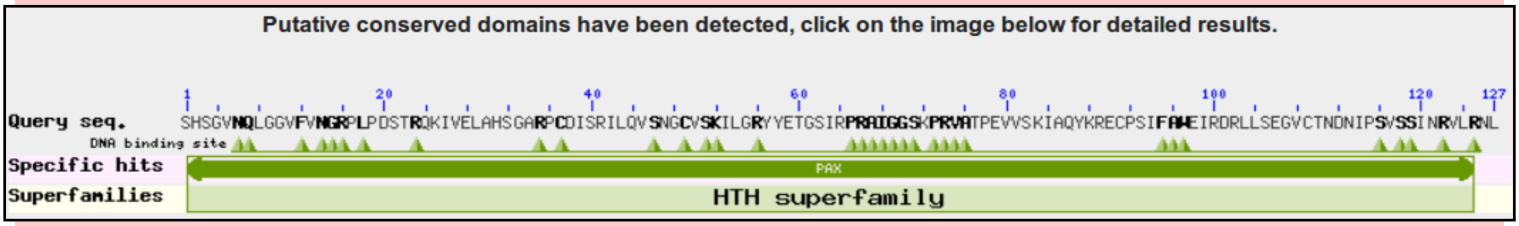
Pseudocount: 0

Before you set **PSI-BLAST** going, click on the **Algorithm parameters** link and take a look at the **PSI/PHI/DELTA BLAST** section. Note the option to use a **PSSM** from a previous run of **PSI-BLAST**, potentially on a different database (but with the same query sequence). Accept the default that database entries scoring better than an **Expect Threshold** of **0.005** be offered for inclusion into the **PSSM** of each successive **PSI-BLAST** iteration. Remember the ? buttons.

What do you suppose the choice of **Pseudocount** might influence?

Elect to **Show results in a new window** and then click on the BLAST button.

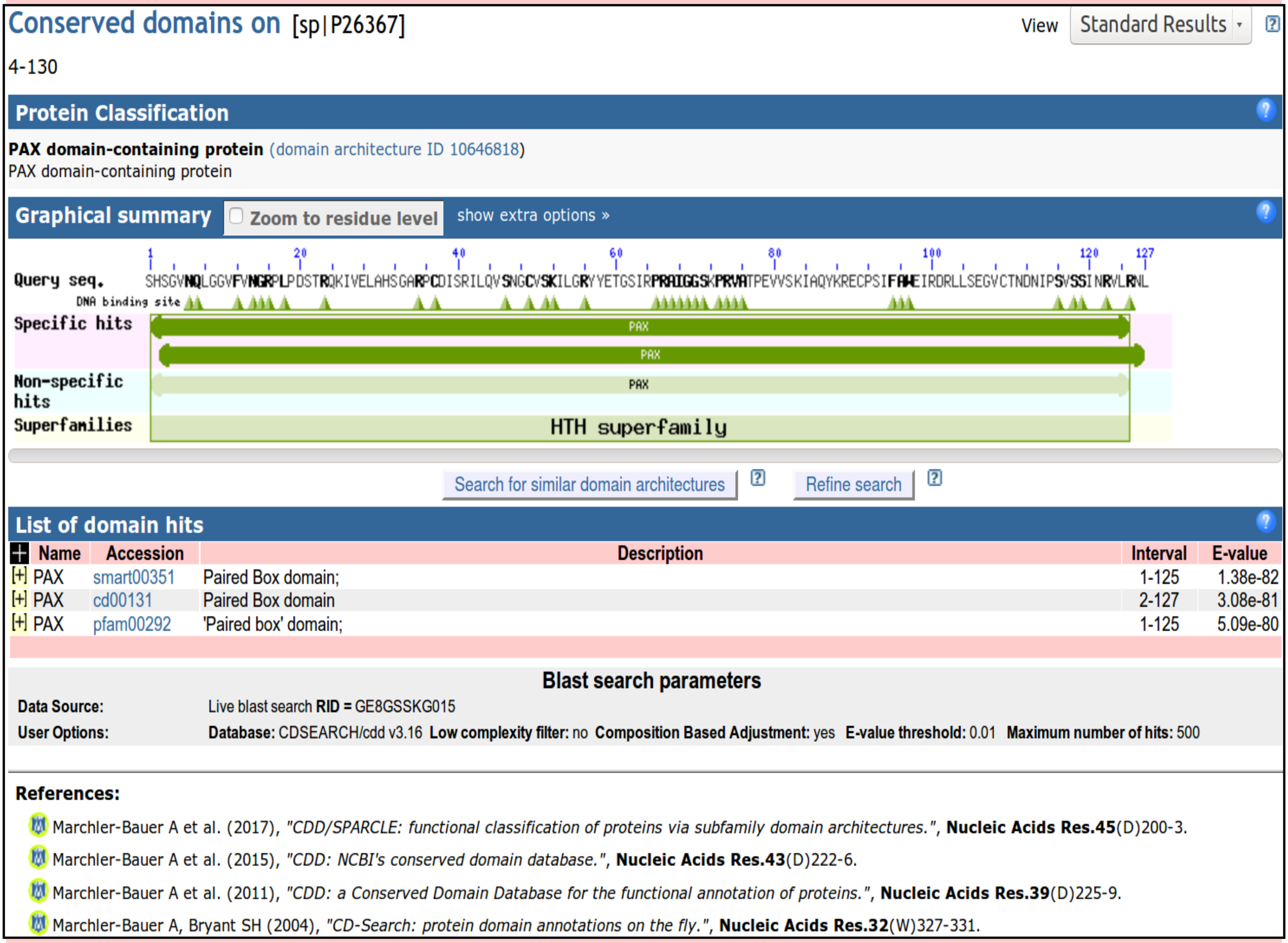
After several moments of deep thought, **PSI-BLAST** will come back with its first set of results, at the top of which is a report that (unsurprisingly) matches have been detected between the query sequence and several domain databases.



Putative conserved domains have been detected, click on the image below for detailed results.

1
20
40
60
80
100
120
127
Query seq.
SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNL
DNA binding site
Specific hits
PAX
Superfamilies
HTH superfamily

For more detail, click on the **Conserved Domains** graphic.



Conserved domains on [sp|P26367]

View Standard Results

4-130

Protein Classification

PAX domain-containing protein (domain architecture ID 10646818)
PAX domain-containing protein

Graphical summary □ Zoom to residue level show extra options »

1
20
40
60
80
100
120
127
Query seq.
SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNL
DNA binding site
Specific hits
PAX
PAX
Non-specific hits
PAX
Superfamilies
HTH superfamily
Search for similar domain architectures
Refine search

List of domain hits

Name	Accession	Description	Interval	E-value
[+] PAX	smart00351	Paired Box domain;	1-125	1.38e-82
[+] PAX	cd00131	Paired Box domain	2-127	3.08e-81
[+] PAX	pfam00292	'Paired box' domain;	1-125	5.09e-80

Blast search parameters

Data Source: Live blast search **RID = GE8GSSKG015**

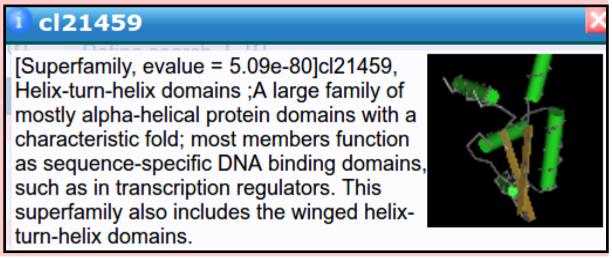
User Options: **Database:** CDSEARCH/cdd v3.16 **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01 **Maximum number of hits:** 500

References:

- Marchler-Bauer A et al. (2017), *"CDD/SPARCLE: functional classification of proteins via subfamily domain architectures."*, **Nucleic Acids Res.45**(D)200-3.
- Marchler-Bauer A et al. (2015), *"CDD: NCBI's conserved domain database."*, **Nucleic Acids Res.43**(D)222-6.
- Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.
- Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.

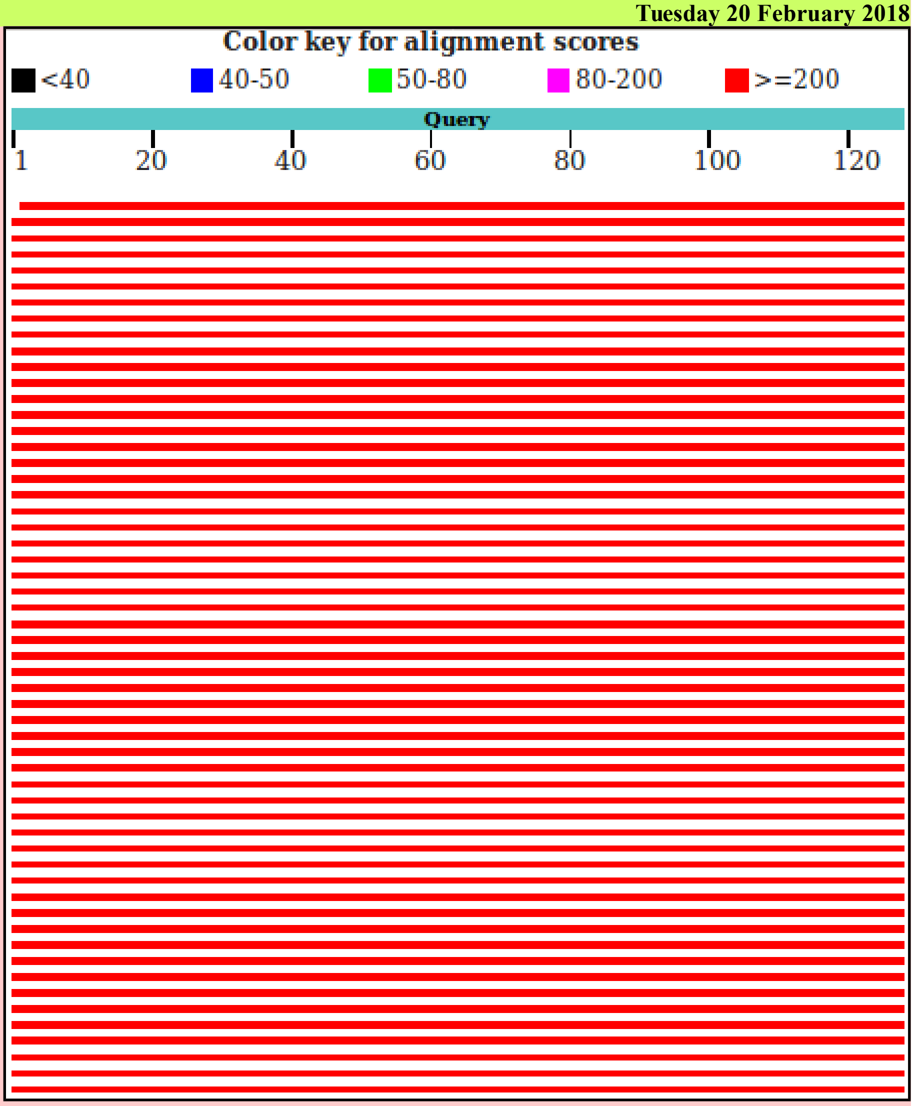
Hover over the **Specific** / **Non-specific hits** and you will see that **SMART**, **Pfam** and the **NCBI Conserved Domains** database matches for a **PAX** domain are all reported. No surprise here.

There is also a **Superfamilies** (derived from **SCOP** as briefly mentioned previously) hit recognising that a **PAX** domain, in common with many other domains, includes **Helix-Turn-Helices**.

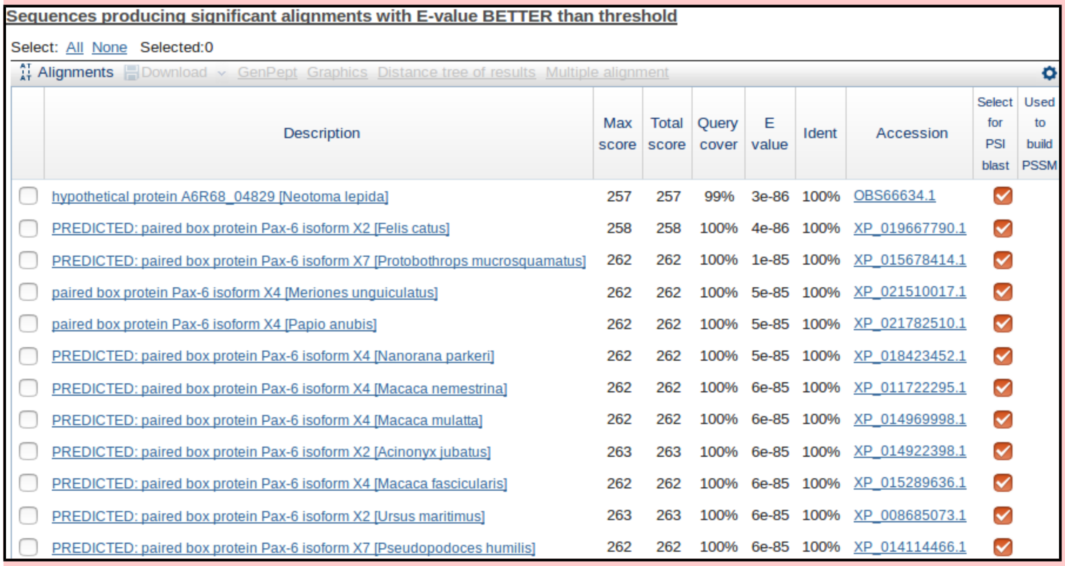


cl21459

[Superfamily, evalue = 5.09e-80]cl21459, Helix-turn-helix domains ;A large family of mostly alpha-helical protein domains with a characteristic fold; most members function as sequence-specific DNA binding domains, such as in transcription regulators. This superfamily also includes the winged helix-turn-helix domains.



Moving back to the main **PSI-BLAST** results, you will see that there are many high quality hits covering the whole length of the query sequence.


Sequences producing significant alignments with E-value BETTER than threshold

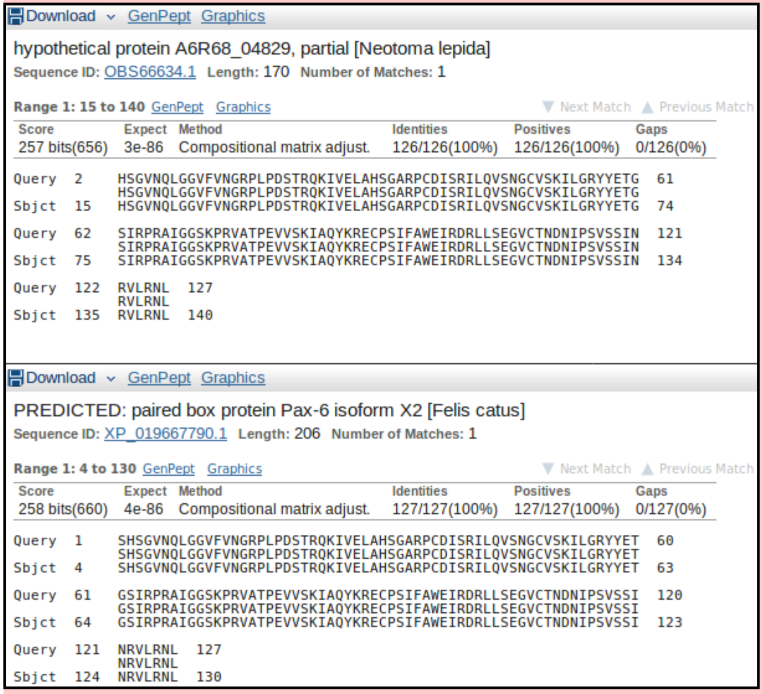
Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
□	hypothetical protein A6R68_04829 [Neotoma lepida]	257	257	99%	3e-86	100%	OBS66634.1	☑	
□	PREDICTED: paired box protein Pax-6 isoform X2 [Felis catus]	258	258	100%	4e-86	100%	XP_019667790.1	☑	
□	PREDICTED: paired box protein Pax-6 isoform X7 [Protobothrops mucrosquamatus]	262	262	100%	1e-85	100%	XP_015678414.1	☑	
□	paired box protein Pax-6 isoform X4 [Meriones unguiculatus]	262	262	100%	5e-85	100%	XP_021510017.1	☑	
□	paired box protein Pax-6 isoform X4 [Papio anubis]	262	262	100%	5e-85	100%	XP_021782510.1	☑	
□	PREDICTED: paired box protein Pax-6 isoform X4 [Nanorana parkeri]	262	262	100%	5e-85	100%	XP_018423452.1	☑	
□	PREDICTED: paired box protein Pax-6 isoform X4 [Macaca nemestrina]	262	262	100%	6e-85	100%	XP_011722295.1	☑	
□	PREDICTED: paired box protein Pax-6 isoform X4 [Macaca mulatta]	262	262	100%	6e-85	100%	XP_014969998.1	☑	
□	PREDICTED: paired box protein Pax-6 isoform X2 [Acinonyx jubatus]	263	263	100%	6e-85	100%	XP_014922398.1	☑	
□	PREDICTED: paired box protein Pax-6 isoform X4 [Macaca fascicularis]	262	262	100%	6e-85	100%	XP_015289636.1	☑	
□	PREDICTED: paired box protein Pax-6 isoform X2 [Ursus maritimus]	263	263	100%	6e-85	100%	XP_008685073.1	☑	
□	PREDICTED: paired box protein Pax-6 isoform X7 [Pseudopodoces humilis]	262	262	100%	6e-85	100%	XP_014114466.1	☑	

The best **500** of these are listed.

All the listed hits are selected for inclusion into the **PSSM** for the next iteration. Unless you feel strongly about any particular entry, leave them all selected.

Note the **Accession Codes** that begin **XP_**. As mention previously, these are less well evidenced protein sequences from the **NCBI** databases.



Download ⌄ GenPept Graphics

hypothetical protein A6R68_04829, partial [Neotoma lepida]
Sequence ID: OBS66634.1 Length: 170 Number of Matches: 1

Range 1: 15 to 140 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
257 bits(656)	3e-86	Compositional matrix adjust.	126/126(100%)	126/126(100%)	0/126(0%)

```
Query  2     HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  61
             HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct  15    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  74

Query  62    SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN  121
             SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN
Sbjct  75    SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN  134

Query  122   RVLRNL  127
             RVLRNL
Sbjct  135   RVLRNL  140
```

Download ⌄ GenPept Graphics

PREDICTED: paired box protein Pax-6 isoform X2 [Felis catus]
Sequence ID: XP_019667790.1 Length: 206 Number of Matches: 1

Range 1: 4 to 130 GenPept Graphics ▼ Next Match ▲ Previous Match

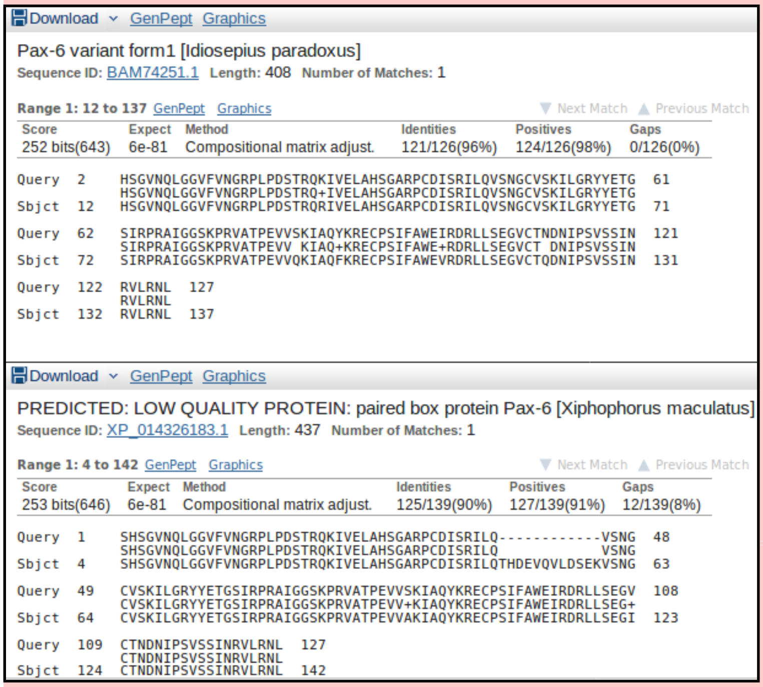
Score	Expect	Method	Identities	Positives	Gaps
258 bits(660)	4e-86	Compositional matrix adjust.	127/127(100%)	127/127(100%)	0/127(0%)

```
Query  1     SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET  60
             SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET
Sbjct  4     SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET  63

Query  61    GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI  120
             GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI
Sbjct  64    GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI  123

Query  121   NRVLRNL  127
             NRVLRNL
Sbjct  124   NRVLRNL  130
```

Move down to the **Alignments** section of the results and you will see that many of the top hits match the query exactly.

Note that many of the top hits come from the **GenPept** database (roughly equivalent to the **TrEMBL** section of **UniProtKB**).

How might the inclusion of poor quality and duplicated sequences have been minimised?



Download ⌄ GenPept Graphics

Pax-6 variant form1 [Idiosepius paradoxus]
Sequence ID: BAM74251.1 Length: 408 Number of Matches: 1

Range 1: 12 to 137 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
252 bits(643)	6e-81	Compositional matrix adjust.	121/126(96%)	124/126(98%)	0/126(0%)

```
Query  2     HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  61
             HSGVNQLGGVFVNGRPLPDSTRQ+IVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct  12    HSGVNQLGGVFVNGRPLPDSTRQRIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  71

Query  62    SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN  121
             SIRPRAIGGSKPRVATPEVV KIAQ+KRECPSIFAWE+RDRLLSEGVCT DNIPSVSSIN
Sbjct  72    SIRPRAIGGSKPRVATPEVVQKIAQFKRECPSIFAWEVRDRLLSEGVCTQDNIPSVSSIN  131

Query  122   RVLRNL  127
             RVLRNL
Sbjct  132   RVLRNL  137
```

Download ⌄ GenPept Graphics

PREDICTED: LOW QUALITY PROTEIN: paired box protein Pax-6 [Xiphophorus maculatus]
Sequence ID: XP_014326183.1 Length: 437 Number of Matches: 1

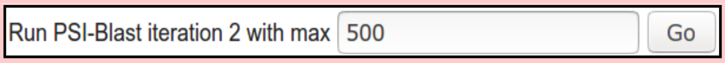
Range 1: 4 to 142 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
253 bits(646)	6e-81	Compositional matrix adjust.	125/139(90%)	127/139(91%)	12/139(8%)

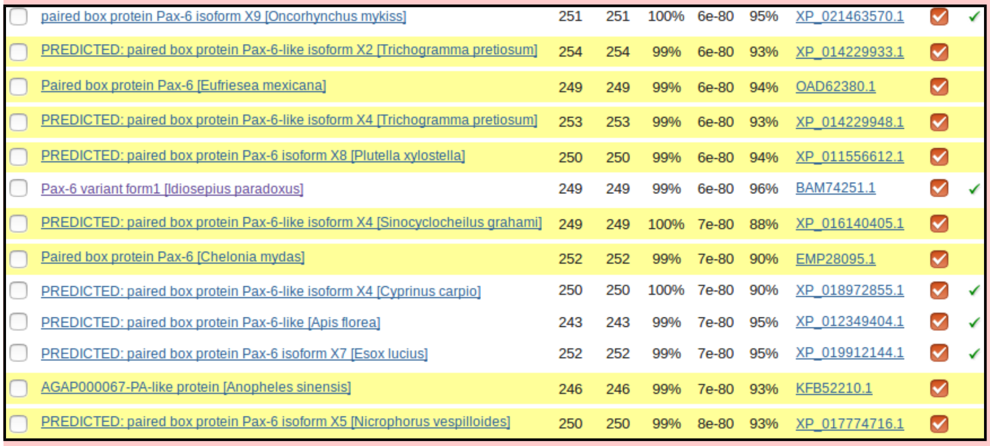
```
Query  1     SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ------------VSNG  48
             SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ            VSNG
Sbjct  4     SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHDEVQVLDSEKVSNG  63

Query  49    CVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGV  108
             CVSKILGRYYETGSIRPRAIGGSKPRVATPEVV+KIAQYKRECPSIFAWEIRDRLLSEG+
Sbjct  64    CVSKILGRYYETGSIRPRAIGGSKPRVATPEVVAKIAQYKRECPSIFAWEIRDRLLSEGI  123

Query  109   CTNDNIPSVSSINRVLRNL  127
             CTNDNIPSVSSINRVLRNL
Sbjct  124   CTNDNIPSVSSINRVLRNL  142
```

Move down far enough and you will see less perfect matches, some of which involve proteins with the extra **14** amino acids of **isoform 5a** of **PAX6_HUMAN**.

Having browsed your results sufficiently, click on the Go button to **Run PSI-Blast iteration 2**. It is at the bottom of the hit list.



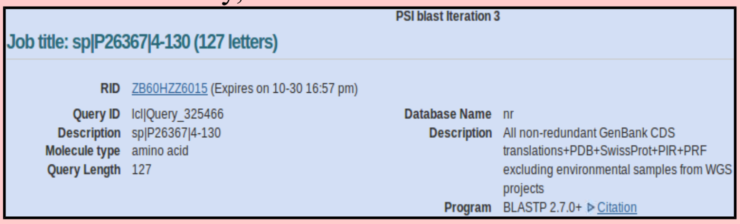
Run PSI-Blast iteration 2 with max 500 Go



☐	paired box protein Pax-6 isoform X9 [Oncorhynchus mykiss]	251	251	100%	6e-80	95%	XP_021463570.1	☑	✓
☐	PREDICTED: paired box protein Pax-6-like isoform X2 [Trichogramma pretiosum]	254	254	99%	6e-80	93%	XP_014229933.1	☑	
☐	Paired box protein Pax-6 [Eufriesea mexicana]	249	249	99%	6e-80	94%	OAD62380.1	☑	
☐	PREDICTED: paired box protein Pax-6-like isoform X4 [Trichogramma pretiosum]	253	253	99%	6e-80	93%	XP_014229948.1	☑	
☐	PREDICTED: paired box protein Pax-6 isoform X8 [Plutella xylostella]	250	250	99%	6e-80	94%	XP_011556612.1	☑	
☐	Pax-6 variant form1 [Idiosepius paradoxus]	249	249	99%	6e-80	96%	BAM74251.1	☑	✓
☐	PREDICTED: paired box protein Pax-6-like isoform X4 [Sinocyclocheilus grahami]	249	249	100%	7e-80	88%	XP_016140405.1	☑	
☐	Paired box protein Pax-6 [Chelonia mydas]	252	252	99%	7e-80	90%	EMP28095.1	☑	
☐	PREDICTED: paired box protein Pax-6-like isoform X4 [Cyprinus carpio]	250	250	100%	7e-80	90%	XP_018972855.1	☑	✓
☐	PREDICTED: paired box protein Pax-6-like [Apis florea]	243	243	99%	7e-80	95%	XP_012349404.1	☑	✓
☐	PREDICTED: paired box protein Pax-6 isoform X7 [Esox lucius]	252	252	99%	7e-80	95%	XP_019912144.1	☑	✓
☐	AGAP000067-PA-like protein [Anopheles sinensis]	246	246	99%	7e-80	93%	KFB52210.1	☑	
☐	PREDICTED: paired box protein Pax-6 isoform X5 [Nicrophorus vespilloides]	250	250	99%	8e-80	93%	XP_017774716.1	☑	

After a few moments, **PSI-BLAST** will return with the results of searching through the database again using the **PSSM** derived from the hits of the first iteration (☑ed). This time the top of the list will be predominantly filled with hits that have already been incorporated into the **PSI-BLAST PSSM**. However, look far enough down the list and you will find some new ones, highlighted yellow.

Once more, click on the Go button to **Run PSI-Blast iteration 3**. That is probably enough! As dear Eddie oft advised, there are typically but three steps to ultimate fulfilment. Previously, I took **8** iterations before there were no more new sequences suggested for inclusion into the **PSMM**. However, I do wonder whether it was worth the effort? Certainly not in the context of this exercise. Trying to continue until no more new sequences can be dangerous, as I discovered the hard way. I once got to iteration **21** before I realised that **PSI-Blast** was playing tricks one me! It was oscillating between two minutely different, perfectly acceptable solutions! Having vented my spleen in shame filled fashion I accepted iteration **21**. I advise that you stop here on "*good enough*" iteration **3**, as I will do this time!

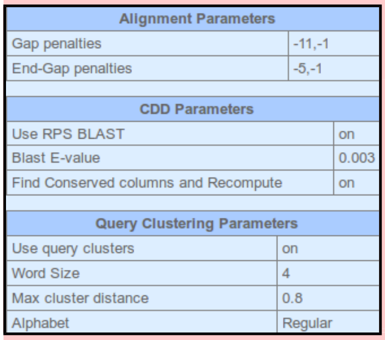


PSI blast Iteration 3

Job title: sp|P26367|4-130 (127 letters)

RID	ZB60HZZ6015 (Expires on 10-30 16:57 pm)		
Query ID	lcl|Query_325466	Database Name	nr
Description	sp|P26367|4-130	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid		
Query Length	127		
		Program	BLASTP 2.7.0+ ▷ Citation

Next, move to the just above the **Graphic Summary** and click on the **Multiple alignment** link. You have elected to use the **NCBI** multiple alignment program **Cobalt** to align the best of the **PAX** domain sequences of your final **PSI-BLAST** iteration (up to **250** sequences that match your query reasonably well, **Expect Score** <= **0.001**, plus the query sequence) .

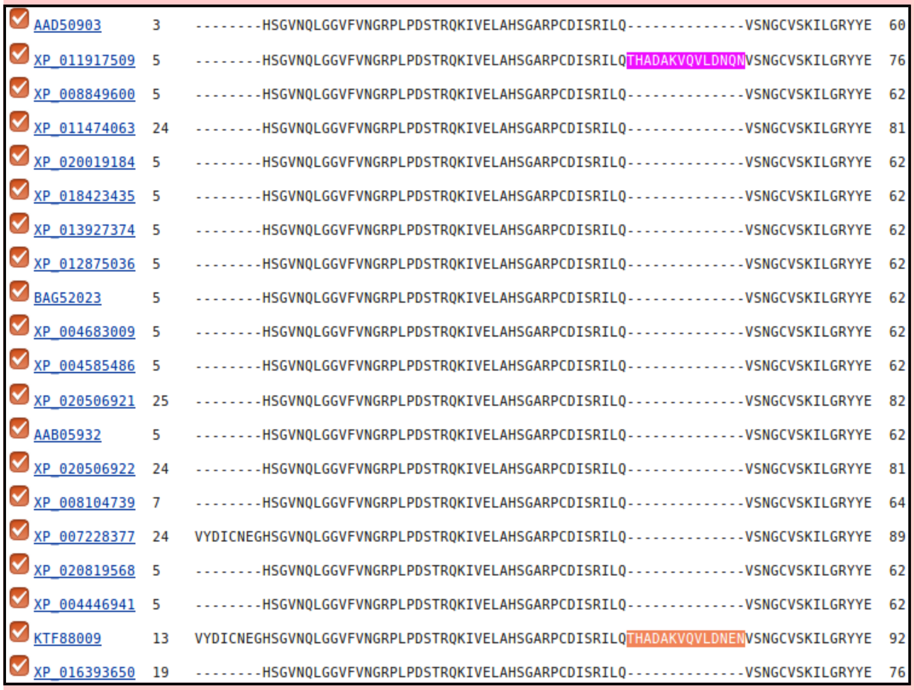


Alignment Parameters	
Gap penalties	-11,-1
End-Gap penalties	-5,-1
CDD Parameters	
Use RPS BLAST	on
Blast E-value	0.003
Find Conserved columns and Recompute	on
Query Clustering Parameters	
Use query clusters	on
Word Size	4
Max cluster distance	0.8
Alphabet	Regular

When it is done, click on the **Alignment parameters** link at the top of the results.

Cobalt reports the parameters it used to make the alignment. It is possible to recompute the alignment with different parameters by using the **Edit and Resubmit** link at the top of the page and then choosing to set **Advanced parameters**. But, maybe not today?

Recording the parameters chosen for any computation is surely extremely important. How else can published computer generated results be reproducible?



```
☑ AAD50903      3   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  60
☑ XP_011917509  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILGRYYE  76
☑ XP_008849600  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ XP_011474063  24  --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  81
☑ XP_020019184  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ XP_018423435  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ XP_013927374  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ XP_012875036  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ BAG52023      5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ XP_004683009  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ XP_004585486  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ XP_020506921  25  --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  82
☑ AAB05932      5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ XP_020506922  24  --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  81
☑ XP_008104739  7   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  64
☑ XP_007228377  24  VYDICNEGHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  89
☑ XP_020819568  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ XP_004446941  5   --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  62
☑ KTF88009      13  VYDICNEGHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENVSNGCVSKILGRYYE  92
☑ XP_016393650  19  --------HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYE  76
```

Move past the list of aligned proteins (why not just hide the **Descriptions** view).

At the top of the actual alignment, set **View Format** to **Plain Text** (…. and then hide the **Descriptions** again??), this being the easiest format to understand in a hurry. The alignment will have very ragged ends, but the important region of **120** or so amino acids representing the **PAX** domain is really quite impressive. In particular, the **isoform 5a** insertion is very convincing.

Cobalt achieves such high quality alignment, partially, by considering the position of matches with **domain** and **motif** databases in addition to sequence composition. Another example of the use of more information leading to improved analysis results.

More on **MSA** later.

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 **Searching for sequence similarities in databases**

When would **Mask lower case letters** be a useful thing to do?

Generally, whenever one might suspect the automatic masking algorithms of **blast** might miss a non informative region in a specific query sequence, obviously.

A specific example might be when a query sequence contained a significant informative region that was known to be common amongst the sequences being searched. If this region was left unmasked, **blast** would pick up so many similar matches to this one region that other interesting similarities might be obscured. By manually masking such a region by changing it to lower case, its matches would not be seen by **blast** and matches with other regions of the query sequence should be more apparent.

Which parameters would **blast** need to **automatically adjust** to cater for short input sequences (such as primers being tested for uniqueness), and why?

The **word size**: Clearly, if you are trying to find matches for a primer (for example) of around **20** base pairs, it would be pretty silly to use a **word size** of **28** (default for **megablast**). A **word** the same size as the primer would find only exact matches. A **word** of about **7** would allow a couple of mismatches and would probably be most generally appropriate.

The **expect score**: As good chance matches between between a short query sequence and a large database will be abundant, it would not be sensible to choose a demanding (i.e. small) **expect score** to represent the limit of significance. In particular, a primer sized query sequence of around **20** base pairs might easily exactly match more than **10** times (generally the default maximum expect score for a significant match) just by chance. After all, there are only **4** bases, a string of **20** is not that long and the databases can be huge! Typically **blast** chooses very high **expect score** cut off for short query sequences, effectively removing the **expect score** filter altogether.

Earlier versions of **blast** did not automatically adjust these parameters. When a short query sequences were selected, suitable adjustment was left to the user. Without sensible parameter adjustment, results could be greatly confusing. For example, a **21** base pair primer could easily match perfectly more than **10** times against a large DNA sequence database. **blast** is set to ignore matches that are expected to occur more than **10** times by chance. Thus even exact matches with such a small sequences would be ignored! Now automatic parameter adjustment is undertaken by **blast**, the user does not really have to think too hard. However, it does seem to be a good idea to know what **blast** is doing and why.

Why do you suppose that a few of the exons of the first **11** matches do not achieve the maximum score?

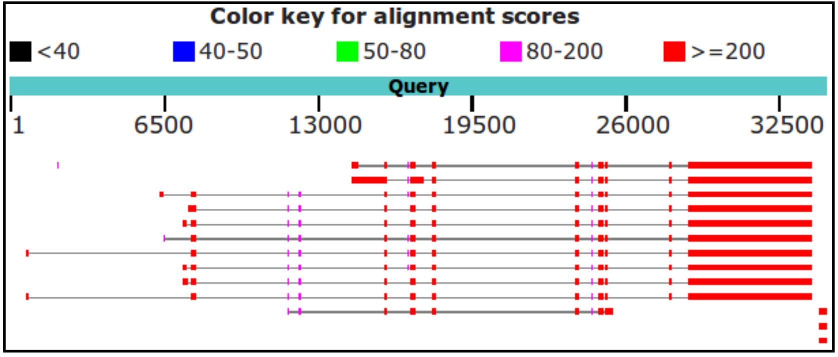
Summary:

Each local region of significant alignment between a database entry and a query sequence is scored independently. The scoring method that governs the alignment score colour in this graphic, reflects both the quality of the match ***and*** its length. Unless a particular region is of sufficient length, it cannot achieve the **200 bit** threshold even if the alignment is perfect. Note that it is the shorter regions that fail to reach the >=200 status. All of the illustrated local alignments associated with **PAX6** transcripts are essentially perfect.

Full Answer:

In common with most database searching programs, **blast** compares query sequences with database entries using a local strategy. The overall evaluation of a particular query sequence is taken to be the highest local score.

Individual local matches are coloured according to individual quality. In this query, all true matches should be perfect, or very nearly so. Scores might therefore be expected to be maximal (>=200). However, they are not? Some only score in the range 80-200.



The score referenced for this purpose is the **bit score**. For a full, no holds barred definition of this score, try here. I prefer this somewhat gentler version:

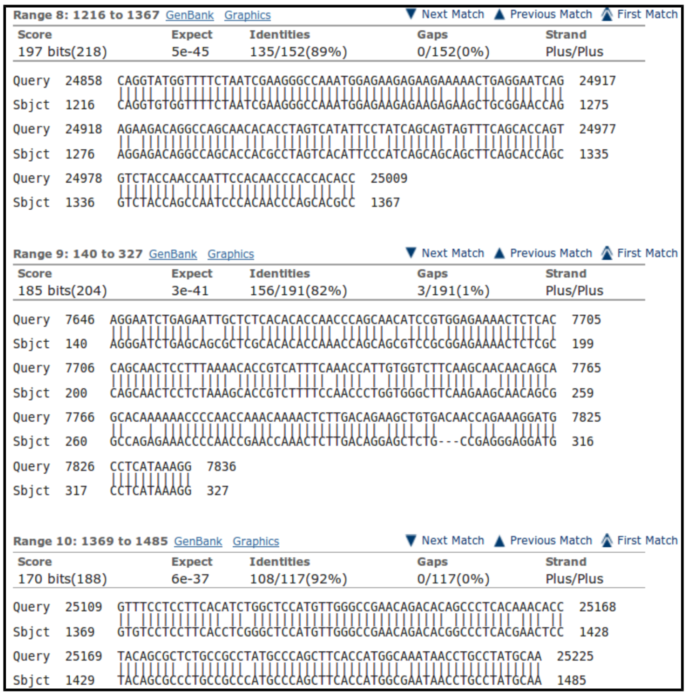
> "The **bit score** gives an indication of how good the alignment is; the higher the score, the better the alignment. In general terms, this score is calculated from a formula that takes into account the alignment of similar or identical residues, as well as any gaps introduced to align the sequences. A key element in this calculation is the "substitution matrix ", which assigns a score for aligning any possible pair of residues. The **BLOSUM62** matrix is the default for most **BLAST** programs, the exceptions being **blastn** and **MegaBLAST** (programs that perform **nucleotide–nucleotide** comparisons and hence do not use protein-specific matrices). Bit scores are normalized, which means that the bit scores from different alignments can be compared, even if different scoring matrices have been used."

Still too scary? The important things to note are that:

- These scores are based on a simple DNA scoring matrix (**1** for a match, **-2** for a mismatch by default for **megablast**), plus penalties for gaps. So scores will be limited by the length of the alignment, ignoring gaps.
- The scores reflect penalties for **indels** (**in**sertions or **del**etions).
- The scores are normalised so that they do not depend on the chosen scoring matrix. This allows bits scores from searches using different scoring matrices to be compared.

Both the scoring matrix dependant **raw scores** and the **bit scores** reflect both the length of an alignment and its quality. **blast** presents the local high scoring regions it discovers ranked by **bit score**. In general, this corresponds to length order. However, a shorter high quality alignment can occasionally outscore a longer less perfect alignment (as illustrated).

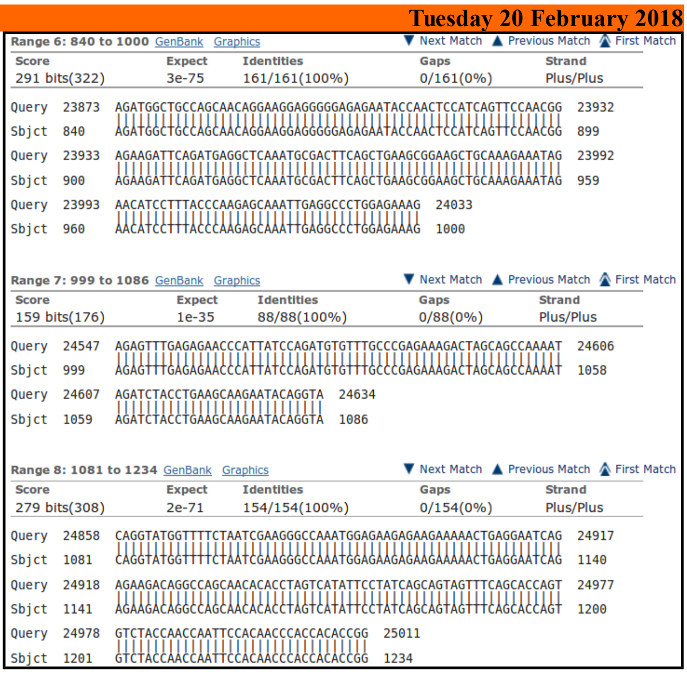
To obtain this illustration I had to use the more sensitive **blastn** algorithm to find more distant alignments (**megablast** is only going to notice really obvious matches) and remove the organism filter to insure that there were less obvious matches to find (all significant matches between any part of the human genome and any human mrna will be too uniformly near exact).



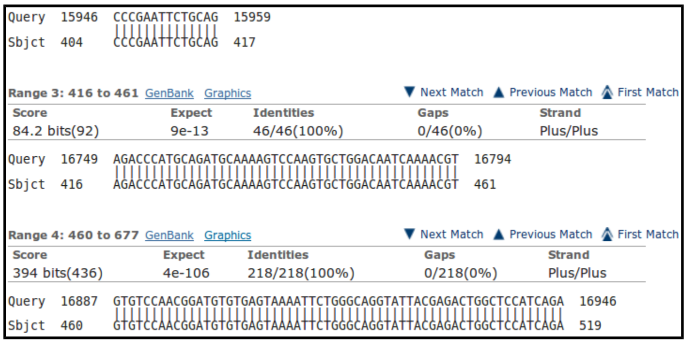
```
Range 8: 1216 to 1367 GenBank Graphics          ▼ Next Match ▲ Previous Match ⩓ First Match
Score            Expect   Identities       Gaps        Strand
197 bits(218)    5e-45    135/152(89%)     0/152(0%)   Plus/Plus

Query  24858  CAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAGAAAAACTGAGGAATCAG  24917
              ||||| |||||||||||||||||||||||||||||||||||||| || ||| |||| |||
Sbjct  1216   CAGGTGTGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAGAGAAGCTGCGGAACCAG  1275

Query  24918  AGAAGACAGGCCAGCAACACACCTAGTCATATTCCTATCAGCAGTAGTTTCAGCACCAGT  24977
              || ||||||||||||| ||| |||||||| ||||| |||||||| || ||||||||||| 
Sbjct  1276   AGGAGACAGGCCAGCACCACGCCTAGTCACATTCCCATCAGCAGCAGCTTCAGCACCAGC  1335

Query  24978  GTCTACCAACCAATTCCACAACCCACCACACC  25009
              |||||||| ||||| |||||||||| ||| ||
Sbjct  1336   GTCTACCAGCCAATCCCACAACCCAGCACGCC  1367


Range 9: 140 to 327 GenBank Graphics            ▼ Next Match ▲ Previous Match ⩓ First Match
Score            Expect   Identities       Gaps        Strand
185 bits(204)    3e-41    156/191(82%)     3/191(1%)   Plus/Plus

Query  7646   AGGAATCTGAGAATTGCTCTCACACACCAACCCAGCAACATCCGTGGAGAAAACTCTCAC  7705
              ||| ||||||| |  |||| |||||||||| |||||| | |||| ||||||||||||| |
Sbjct  140    AGGGATCTGAGCAGCGCTCGCACACACCAAACCAGCAGCGTCCGCGGAGAAAACTCTCGC  199

Query  7706   CAGCAACTCCTTTAAAACACCGTCATTTCAAACCATTGTGGTCTTCAAGCAACAACAGCA  7765
              ||||||||||| |||| ||||||| |||| |||| | |||| ||||||| | ||||||| 
Sbjct  200    CAGCAACTCCTCTAAAGCACCGTCTTTTCCAACCCTGGTGGGCTTCAAGAAGCAACAGCG  259

Query  7766   GCACAAAAAACCCCAACCAAACAAAACTCTTGACAGAAGCTGTGACAACCAGAAAGGATG  7825
              ||   | ||||||||||| ||| ||||||||||||| |||| ||    | ||  ||||||
Sbjct  260    GCCAGAGAAACCCCAACCGAACCAAACTCTTGACAGGAGCTCTG---CCGAGGGAGGATG  316

Query  7826   CCTCATAAAGG  7836
              |||||||||||
Sbjct  317    CCTCATAAAGG  327


Range 10: 1369 to 1485 GenBank Graphics         ▼ Next Match ▲ Previous Match ⩓ First Match
Score            Expect   Identities       Gaps        Strand
170 bits(188)    6e-37    108/117(92%)     0/117(0%)   Plus/Plus

Query  25109  GTTTCCTCCTTCACATCTGGCTCCATGTTGGGCCGAACAGACACAGCCCTCACAAACACC  25168
              || ||||||||||| || |||||||||||||||||||||||||| |||||||| ||| ||
Sbjct  1369   GTGTCCTCCTTCACCTCGGGCTCCATGTTGGGCCGAACAGACACGGCCCTCACGAACTCC  1428

Query  25169  TACAGCGCTCTGCCGCCTATGCCCAGCTTCACCATGGCAAATAACCTGCCTATGCAA  25225
              |||||||| |||||||| |||||||||||||||||||| ||||||||||||||||||
Sbjct  1429   TACAGCGCCCTGCCGCCCATGCCCAGCTTCACCATGGCGAATAACCTGCCTATGCAA  1485
```

You can see evidence of what is occurring in the alignments further down your results. Here is illustrated one of the 80-200 exons that occur in all transcripts at position **24,547**. The match is perfect, but the length of the exon is consistently just to short to get to the heady >=200 level. To make this illustration represent alignments from a particular region, I set **Sort by**: (top of the alignments) to **Query start position**. If you look back at the **blast** graphic, you should be able to easily spot the region of these aligned regions including the one that is 80-200 .

Note how imperfectly **blast** finds exon/intron boundaries. If the start of an intron happens to match the start of the next exon, **blast** will included the bases in two alignments[10]. It is not looking for exons and introns as was **spline**, it just mindlessly seeks matches.



```
Range 6: 840 to 1000 GenBank Graphics          ▼ Next Match ▲ Previous Match ⇞ First Match
Score           Expect   Identities       Gaps         Strand
291 bits(322)   3e-75    161/161(100%)    0/161(0%)    Plus/Plus

Query  23873  AGATGGCTGCCAGCAACAGGAAGGAGGGGAGAGAATACCAACTCCATCAGTTCCAACGG  23932
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  840    AGATGGCTGCCAGCAACAGGAAGGAGGGGAGAGAATACCAACTCCATCAGTTCCAACGG  899
Query  23933  AGAAGATTCAGATGAGGCTCAAATGCGACTTCAGCTGAAGCGGAAGCTGCAAAGAAATAG  23992
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  900    AGAAGATTCAGATGAGGCTCAAATGCGACTTCAGCTGAAGCGGAAGCTGCAAAGAAATAG  959
Query  23993  AACATCCTTTACCCAAGAGCAAATTGAGGCCCTGGAGAAAG  24033
              |||||||||||||||||||||||||||||||||||||||||
Sbjct  960    AACATCCTTTACCCAAGAGCAAATTGAGGCCCTGGAGAAAG  1000


Range 7: 999 to 1086 GenBank Graphics          ▼ Next Match ▲ Previous Match ⇞ First Match
Score           Expect   Identities       Gaps         Strand
159 bits(176)   1e-35    88/88(100%)      0/88(0%)     Plus/Plus

Query  24547  AGAGTTTGAGAGAACCCATTATCCAGATGTGTTTGCCCGAGAAAGACTAGCAGCCAAAAT  24606
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  999    AGAGTTTGAGAGAACCCATTATCCAGATGTGTTTGCCCGAGAAAGACTAGCAGCCAAAAT  1058
Query  24607  AGATCTACCTGAAGCAAGAATACAGGTA  24634
              ||||||||||||||||||||||||||||
Sbjct  1059   AGATCTACCTGAAGCAAGAATACAGGTA  1086


Range 8: 1081 to 1234 GenBank Graphics        ▼ Next Match ▲ Previous Match ⇞ First Match
Score           Expect   Identities       Gaps         Strand
279 bits(308)   2e-71    154/154(100%)    0/154(0%)    Plus/Plus

Query  24858  CAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAGAAAAACTGAGGAATCAG  24917
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081   CAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAGAAGAGAAGAAAAACTGAGGAATCAG  1140
Query  24918  AGAAGACAGGCCAGCAACACACCTAGTCATATTCCTATCAGCAGTAGTTTCAGCACCAGT  24977
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141   AGAAGACAGGCCAGCAACACACCTAGTCATATTCCTATCAGCAGTAGTTTCAGCACCAGT  1200
Query  24978  GTCTACCAACCAATTCCACAACCCACCACACCGG  25011
              ||||||||||||||||||||||||||||||||||
Sbjct  1201   GTCTACCAACCAATTCCACAACCCACCACACCGG  1234
```



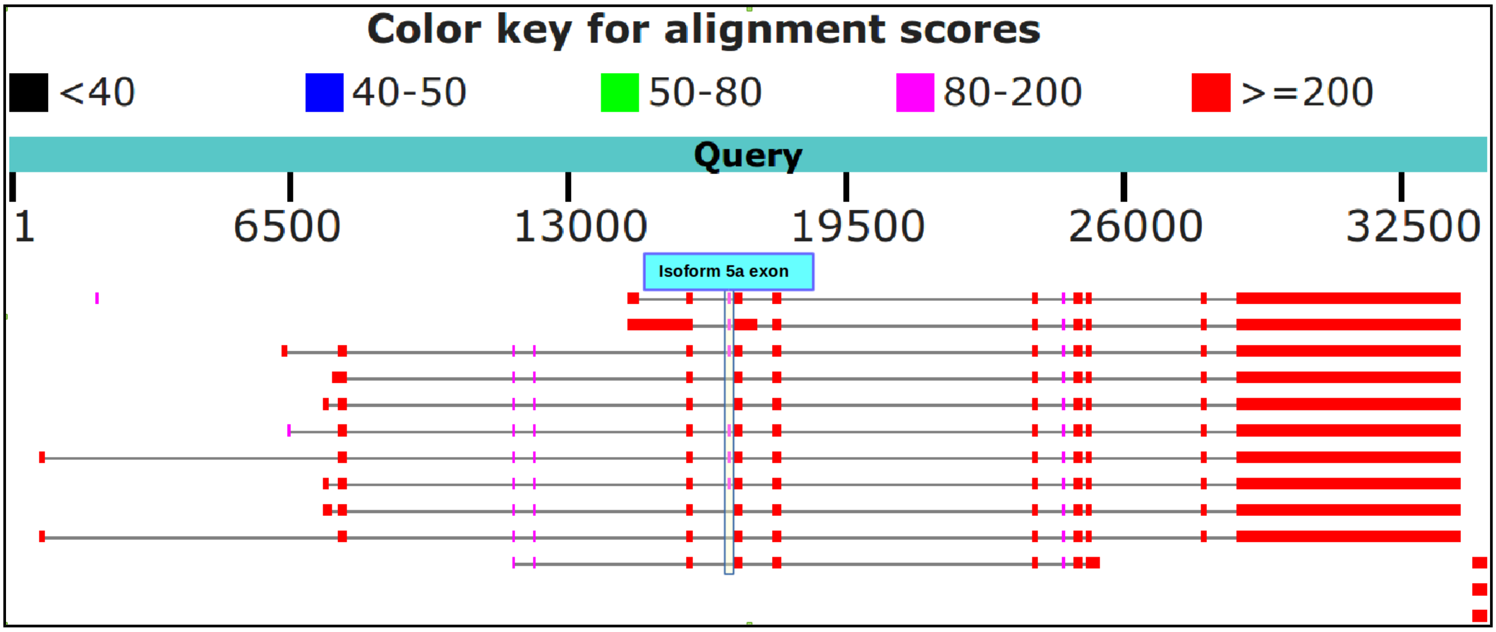
```
Query  15946  CCCGAATTCTGCAG  15959
              ||||||||||||||
Sbjct  404    CCCGAATTCTGCAG  417


Range 3: 416 to 461 GenBank Graphics           ▼ Next Match ▲ Previous Match ⇞ First Match
Score           Expect   Identities       Gaps         Strand
84.2 bits(92)   9e-13    46/46(100%)      0/46(0%)     Plus/Plus

Query  16749  AGACCCATGCAGATGCAAAAGTCCAAGTGCTGGACAATCAAAACGT  16794
              ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  416    AGACCCATGCAGATGCAAAAGTCCAAGTGCTGGACAATCAAAACGT  461


Range 4: 460 to 677 GenBank Graphics           ▼ Next Match ▲ Previous Match ⇞ First Match
Score           Expect   Identities       Gaps         Strand
394 bits(436)   4e-106   218/218(100%)    0/218(0%)    Plus/Plus

Query  16887  GTGTCCAACGGATGTGTGAGTAAAATTCTGGGCAGGTATTACGAGACTGGCTCCATCAGA  16946
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  460    GTGTCCAACGGATGTGTGAGTAAAATTCTGGGCAGGTATTACGAGACTGGCTCCATCAGA  519
```

For a further example, look at the exon that is found only in the **isoform 5a** transcripts. It is tiny (**42** base pairs) and scores well below >=200 even thought it is a perfect match.

Note that the alignment is **46** base pairs long due to **blast** adding on two bases either side that are actually the highly conserved intron start and end base pairs. As you can see, these extra base pairs occur in the preceding and succeeding alignment also.

Explain why one exon in the reasonably consistent region, does not appear in all of the transcript matches?

Well I refer to the **isoform 5a** exon, of course. The tiny inconsistent one about **9** exons in from the right (when it exists). This will, clearly, only occur in **isoform 5a** transcripts.



[10] **2** base pairs (**Sbjct**: **999-1000**, **AG**) occur in both the first two matches illustrated. **6** base pairs are shared between the **2nd** and **3rd** matches (**Sbjct**: **1081-1086**, **CAGGTA**).

Which of the **Refseq PAX6** transcripts corresponds to **isoform 5a**?

Summary:

As I am sure you are tired of noting by now, all the transcripts with the extra tiny exon around position **16,750** in the genomic sequence are **isoform 5a** transcripts. See the illustration for the previous answer.

Full Answer:

The **isoform 5a** transcripts can be spotted most easily from the graphic. They are the ones with the extra small exon slightly to the left of middle (around base position **16,750**). For example, the **first**, **second** and **third blast** matches displayed. If you hover over all the full length matches with your mouse, you will see that they are **transcript variants 11**, **10**, **8**, **7**, **6**, **5**, **4**, **2**, **1**, **3** and **9** (in the vertical order of the graphic).

Stated with the unequalled poetry of **RefSeq Accession Code** and lyrical **Title** Line, the list of those with the extra exon becomes:

TITLE	ACCESSION CODE
Homo sapiens paired box 6 (PAX6), transcript variant 11, mRNA	NM_001310161.1
Homo sapiens paired box 6 (PAX6), transcript variant 10, mRNA	NM_001310160.1
Homo sapiens paired box 6 (PAX6), transcript variant 8, mRNA	NM_001310158.1
Homo sapiens paired box 6 (PAX6), transcript variant 5, mRNA	NM_001258463.1
Homo sapiens paired box 6 (PAX6), transcript variant 4, mRNA	NM_001258462.1
Homo sapiens paired box 6 (PAX6), transcript variant 2, mRNA	NM_001604.5

Yes well, that was fun? The message of the question was to ensure you could see how to spot the **isoform 5a** transcripts (again!), not to list them! But, never mind, doing so was in fine tune with the ennui of the moment.

What are the **9** strongest matches around base position **16,750**?

Summary:

Matches between the regions of the **PAX6** genomic region encoding the **PAX6 Paired Box** domain and **SwissProt** protein sequences representing human proteins including a **Paired Box** domain.

Why would you expect exactly **9** matches around this point?

Summary:

Because that is how many human proteins including a **Paired Box** domain are suggested to exist according to **Interpro** (as shown in a previous Practical). There is **PAX6** plus its **8** paralogues, imaginatively all named:

PAX1, **PAX2**, **PAX3**, **PAX4**, **PAX5**, **PAX6**, **PAX7**, **PAX8** & **PAX9**

What do you make of the plethora of matches around **24,000**?

Summary:

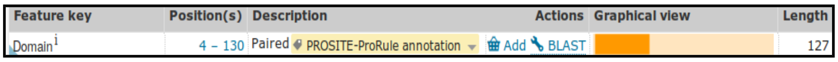
These are matches between the regions of the **PAX6** genomic region encoding the **PAX6 Homeobox** domain and **SwissProt** protein sequences representing human proteins including a **Homeobox** domain. As you discovered earlier from **Interpro**, there are lots of such proteins.

The thin line joining features implies that those features relate to the same database entry. Notice that **4** of the **9** proteins including a **Paired box** domain near the beginning, also include a **Homeobox** domain further along. This is exactly as was suggested by the **SMART** annotation you examined earlier.

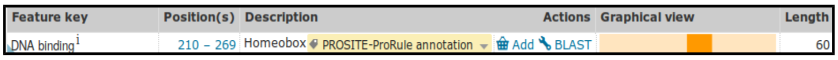
Full Answer:

Well, a couple of graphics to reinforce what has already been claimed and make life more precise and colourful.

First, recall from **UniProtKB** the positions of the two domains in **PAX6**.

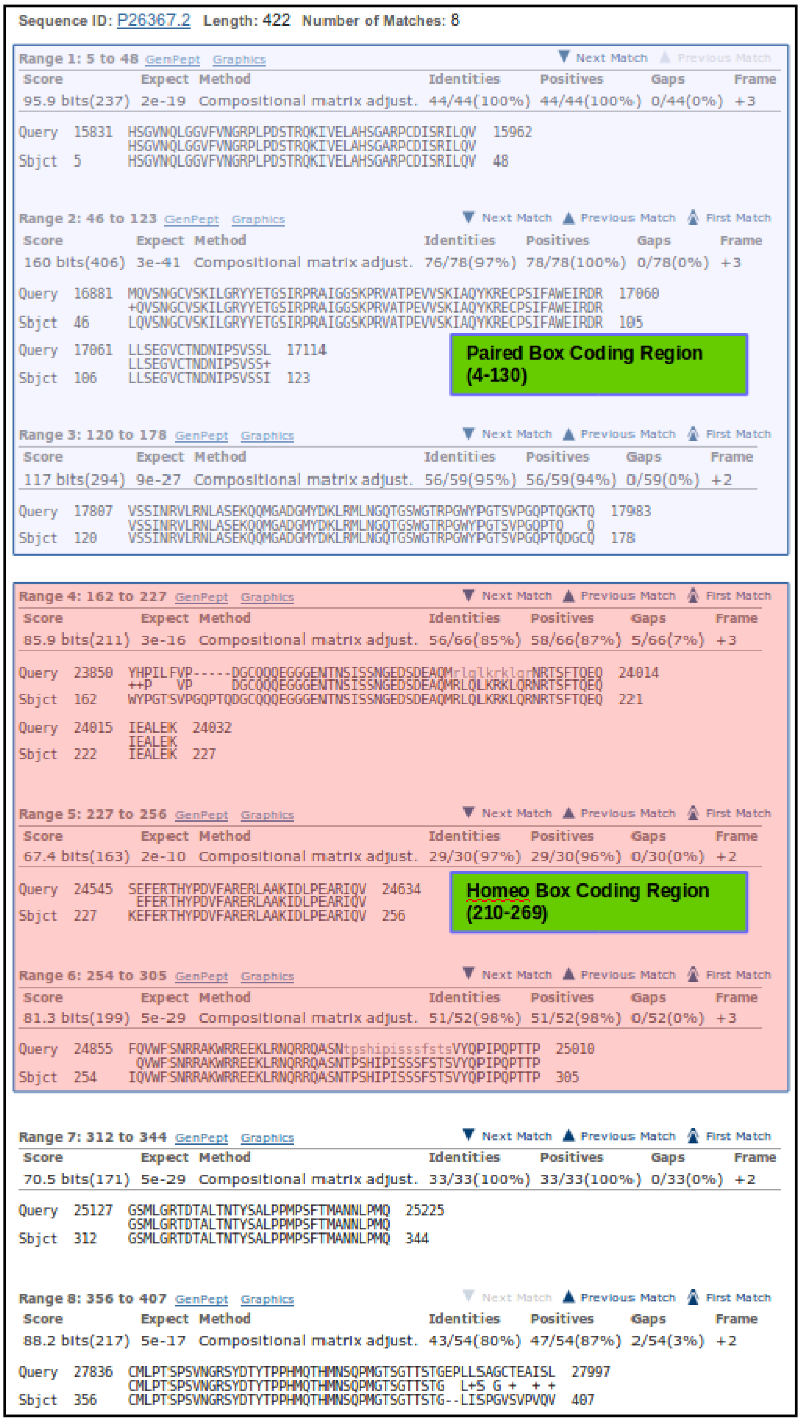


Feature key	Position(s)	Description	Actions	Graphical view	Length
Domain[i]	4 – 130	Paired PROSITE-ProRule annotation	Add BLAST		127



Feature key	Position(s)	Description	Actions	Graphical view	Length
DNA binding[i]	210 – 269	Homeobox PROSITE-ProRule annotation	Add BLAST		60

Next, order the **blastx** alignments by **Subject start position**.



```
Sequence ID: P26367.2  Length: 422  Number of Matches: 8

Range 1: 5 to 48  GenPept  Graphics                              Next Match   Previous Match
 Score          Expect  Method                          Identities    Positives     Gaps      Frame
 95.9 bits(237) 2e-19   Compositional matrix adjust.   44/44(100%)  44/44(100%)  0/44(0%)  +3

Query  15831  HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQV  15962
              HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQV
Sbjct  5      HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQV  48


Range 2: 46 to 123  GenPept  Graphics               Next Match   Previous Match   First Match
 Score          Expect  Method                          Identities   Positives     Gaps      Frame
 160 bits(406)  3e-41   Compositional matrix adjust.   76/78(97%)  78/78(100%)  0/78(0%)  +3

Query  16881  MQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR  17060
              +QVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR
Sbjct  46     LQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR  105

Query  17061  LLSEGVCTNDNIPSVSSL  17114
              LLSEGVCTNDNIPSVSS+
Sbjct  106    LLSEGVCTNDNIPSVSSI  123
```

Paired Box Coding Region (4-130)

```
Range 3: 120 to 178  GenPept  Graphics               Next Match   Previous Match   First Match
 Score          Expect  Method                          Identities   Positives    Gaps      Frame
 117 bits(294)  9e-27   Compositional matrix adjust.   56/59(95%)  56/59(94%)  0/59(0%)  +2

Query  17807  VSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQGKTQ  17983
              VSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQ   Q
Sbjct  120    VSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQ  178


Range 4: 162 to 227  GenPept  Graphics               Next Match   Previous Match   First Match
 Score          Expect  Method                          Identities   Positives    Gaps      Frame
 85.9 bits(211) 3e-16   Compositional matrix adjust.   56/66(85%)  58/66(87%)  5/66(7%)  +3

Query  23850  YHPILFVP-----DGCQQQEGGGENTNSISSNGEDSDEAQMrlqlkrklqrNRTSFTQEQ  24014
              ++P   VP     DGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQ
Sbjct  162    WYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQ  221

Query  24015  IEALEK  24032
              IEALEK
Sbjct  222    IEALEK  227


Range 5: 227 to 256  GenPept  Graphics               Next Match   Previous Match   First Match
 Score          Expect  Method                          Identities   Positives    Gaps      Frame
 67.4 bits(163) 2e-10   Compositional matrix adjust.   29/30(97%)  29/30(96%)  0/30(0%)  +2

Query  24545  SEFERTHYPDVFARERLAAKIDLPEARIQV  24634
               EFERTHYPDVFARERLAAKIDLPEARIQV
Sbjct  227    KEFERTHYPDVFARERLAAKIDLPEARIQV  256
```

Homeo Box Coding Region (210-269)

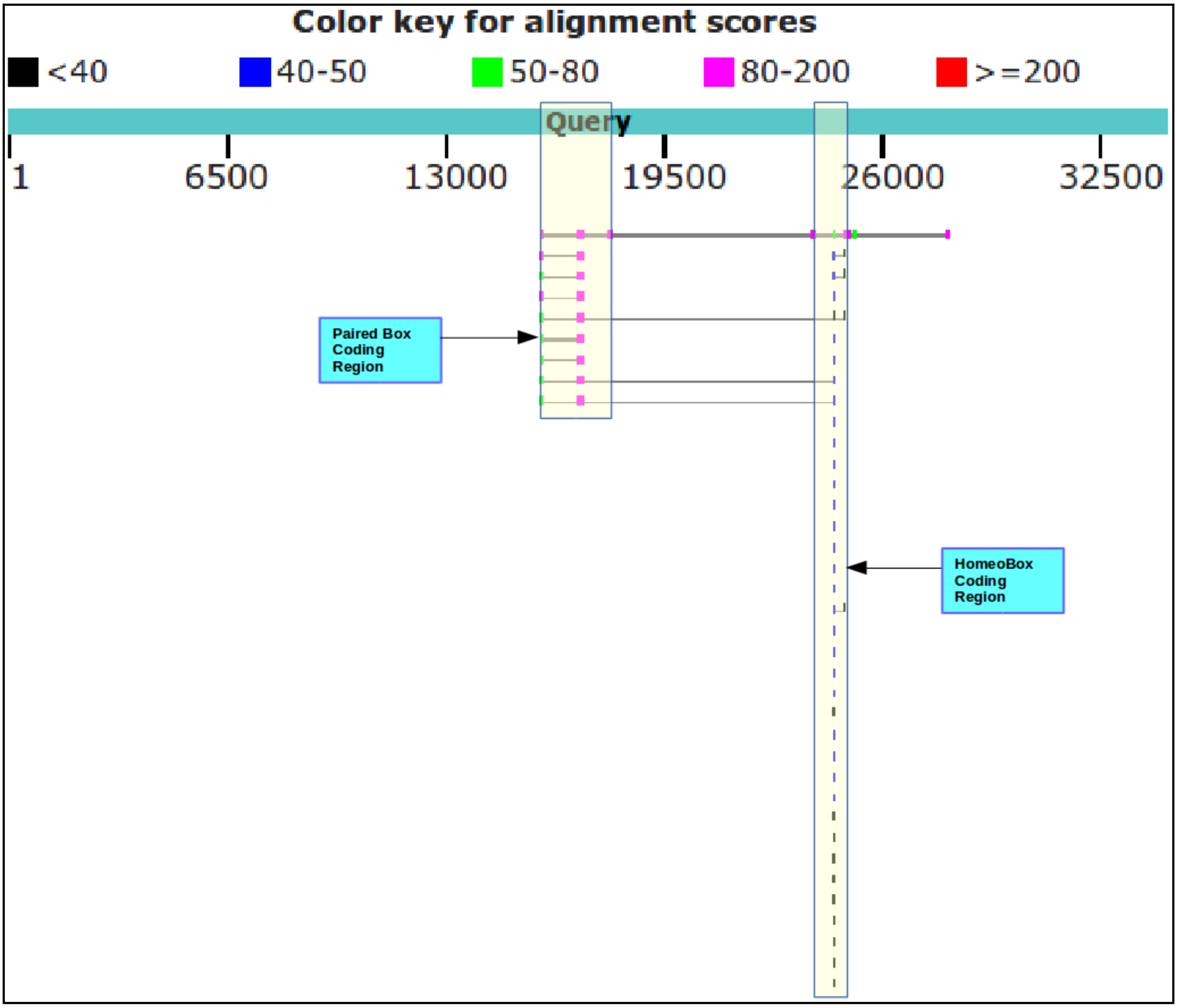
```
Range 6: 254 to 305  GenPept  Graphics               Next Match   Previous Match   First Match
 Score          Expect  Method                          Identities   Positives    Gaps      Frame
 81.3 bits(199) 5e-29   Compositional matrix adjust.   51/52(98%)  51/52(98%)  0/52(0%)  +3

Query  24855  FQVWFSNRRAKWRREEKLRNQRRQASNtpshipisssfstsVYQPIPQPTTP  25010
               QVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP
Sbjct  254    IQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP  305


Range 7: 312 to 344  GenPept  Graphics               Next Match   Previous Match   First Match
 Score          Expect  Method                          Identities    Positives     Gaps      Frame
 70.5 bits(171) 5e-29   Compositional matrix adjust.   33/33(100%)  33/33(100%)  0/33(0%)  +2

Query  25127  GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ  25225
              GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ
Sbjct  312    GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ  344


Range 8: 356 to 407  GenPept  Graphics               Next Match   Previous Match   First Match
 Score          Expect  Method                          Identities   Positives    Gaps      Frame
 88.2 bits(217) 5e-17   Compositional matrix adjust.   43/54(80%)  47/54(87%)  2/54(3%)  +2

Query  27836  CMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGEPLLSAGCTEAISL  27997
              CMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTG  L+S G +  + +
Sbjct  356    CMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTG--LISPGVSVPVQV  407
```

Then see, from the first of the **blastx** alignments, it is the first **2** and a bit aligned regions that correspond to the **Paired Box** coding region.

The next **3** matching sections cover the whole of the **HomeoBox** coding region (with a fair overlap each side).

The final **2** matching sections are not involved in either domain.

With this understanding, one can decorate the **blastx** graphic in a fashion that makes the entirely obvious even ***MORE*** apparent than it was in the first place?



Well, I think it is a nice picture anyway.

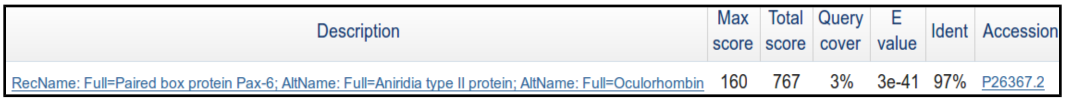
Why do you suppose the **Paired box** matches precede the **Homeobox** matches?

Because they score more highly and so, in the opinion of **blast**, are more worthy. Primarily, they score more highly because they are longer. The list is ranked by **E Value**. Good matches with long sequence are less likely to occur by chance than equally good matches with shorter sequences.

Possibly a more interesting question[11] might have been: "Why are not all the hits which include both domains at the top of the list?". Surely they should be, as they match over a longer proportion of the query sequence and so must, in general at least, be of the greatest significance.

They do not always come at the top of the list because **blast** scores each matching region individually and uses the ranking scores associated with the single region with the highest **E Value** to evaluate the similarity of the entire database entry with the query. This has to be a dubious practice surely? But, it appears to work, so why complain.

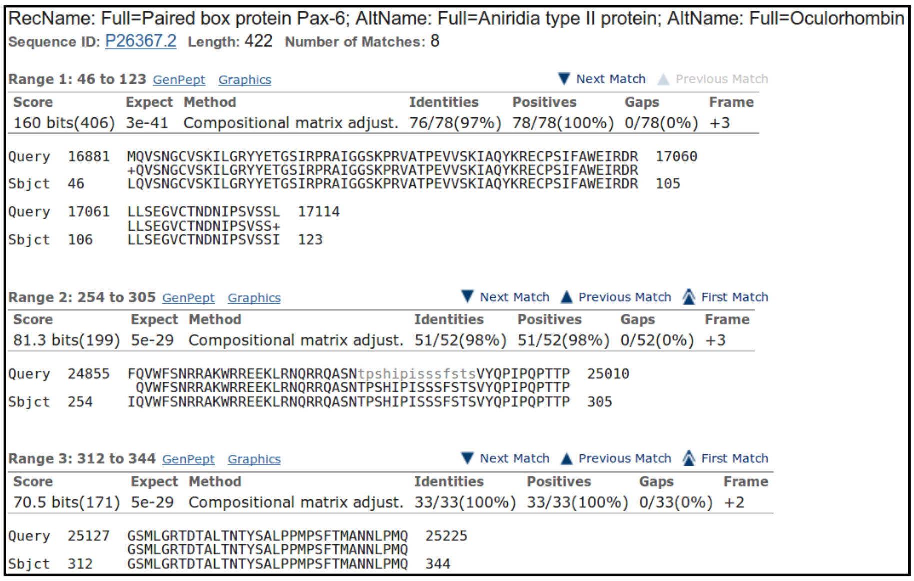
To justify this last assertion, Look at your top hit.



Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Paired box protein Pax-6; AltName: Full=Aniridia type II protein; AltName: Full=Oculorhombin	160	767	3%	3e-41	97%	P26367.2

E Val = **3e-41**, **Max score** = **160**, **Total score 767** associated with the whole of **P26367.2**

Now look at the first few individual regional alignments for this hit.



```
RecName: Full=Paired box protein Pax-6; AltName: Full=Aniridia type II protein; AltName: Full=Oculorhombin
Sequence ID: P26367.2  Length: 422  Number of Matches: 8

Range 1: 46 to 123 GenPept  Graphics                              ▼ Next Match  ▲ Previous Match
 Score          Expect  Method                         Identities    Positives     Gaps       Frame
 160 bits(406)  3e-41   Compositional matrix adjust.   76/78(97%)    78/78(100%)   0/78(0%)   +3

Query  16881  MQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR  17060
              +QVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR
Sbjct  46     LQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR  105

Query  17061  LLSEGVCTNDNIPSVSSL  17114
              LLSEGVCTNDNIPSVSS+
Sbjct  106    LLSEGVCTNDNIPSVSSI  123


Range 2: 254 to 305 GenPept  Graphics                   ▼ Next Match  ▲ Previous Match  ▲ First Match
 Score           Expect  Method                         Identities    Positives     Gaps       Frame
 81.3 bits(199)  5e-29   Compositional matrix adjust.   51/52(98%)    51/52(98%)    0/52(0%)   +3

Query  24855  FQVWFSNRRAKWRREEKLRNQRRQASNtpshipisssfstsVYQPIPQPTTP  25010
               QVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP
Sbjct  254    IQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP  305


Range 3: 312 to 344 GenPept  Graphics                   ▼ Next Match  ▲ Previous Match  ▲ First Match
 Score           Expect  Method                         Identities     Positives      Gaps       Frame
 70.5 bits(171)  5e-29   Compositional matrix adjust.   33/33(100%)    33/33(100%)    0/33(0%)   +2

Query  25127  GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ  25225
              GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ
Sbjct  312    GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ  344
```

As you can see, the **E Value** and **Max score** values used to evaluate the whole protein were computed from just the best (ranked by **E Value**) local alignment! Crude, but never mind.

The **Total score** for the entire protein is the sum (rounded up to the nearest integer) of all the bit scores for all **8** local alignments computed for this protein (I suggest you just trust me on this assertion).

[11] That I did not ask, because I only just thought of it.

How do you suppose the **Max matches in a query range** parameter might be of value if this order was reversed?

If **Paired boxes** had been more prolific, then the number of **Paired box** matches might have filled the **blast** hit list before the highest scoring **Homeo box** hit was registered.

If **Homeo boxes** were longer, and so justified a better **E value**, then the number of **Homeo box** matches might have filled the **blast** hit list before the highest scoring **Paired box** hit was registered.

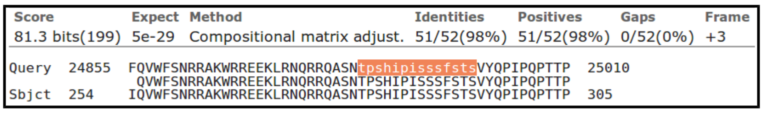
Either of these situations would be very unfortunate, but easily avoided by setting the **Max matches in a query range** parameter to something sensible (**50** say). This would ensure that only the top **50** items in the **blast** hit list would be dominated by the strongest hit.

UNFORTUNATELY … although that is the intention of this parameter, it currently simply will not work, except in very particular circumstances, because of the way it is implemented. This is a great pity, because it is a very good idea, in principle.

I will spare you the details as, despite energetic debate, the **NCBI** people appear to have no intention of changing things, although they do appear to accept my arguments? Or maybe they just humour me?

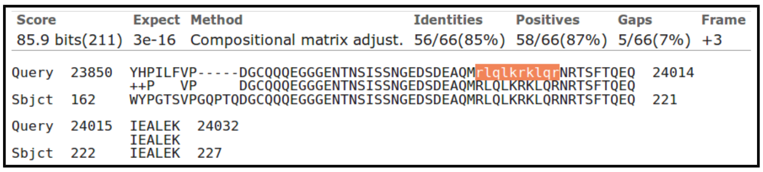
How does this "non-informative" region match expectations suggested by **SMART** and the **Feature table** of **UniprotKB** for **PAX6_HUMAN**?

blast identifies two non-informative regions. I only discussed the prettiest one above. The region discussed is comprised largely of **Serines**, **Prolines**, **Threonines** & **Isoleucines** the **15** residues between **294-308**.



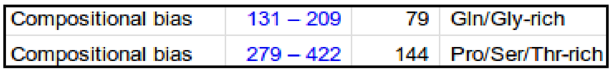
```
Score           Expect  Method                      Identities    Positives     Gaps       Frame
81.3 bits(199)  5e-29   Compositional matrix adjust. 51/52(98%)   51/52(98%)    0/52(0%)   +3

Query  24855  FQVWFSNRRAKWRREEKLRNQRRQASNtpshipisssfstsVYQPIPQPTTP  25010
               QVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP
Sbjct  254    IQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP  305
```

The second (to be found much further down your **blast Alignments** output) is comprised entirely of **Arginines**, **Luecines** and **Lysines** and **Glutamines**, the **10** residues between **203 - 212**.



```
Score           Expect  Method                      Identities    Positives     Gaps       Frame
85.9 bits(211)  3e-16   Compositional matrix adjust. 56/66(85%)   58/66(87%)    5/66(7%)   +3

Query  23850  YHPILFVP-----DGCQQQEGGGENTNSISSNGEDSDEAQMrlqlkrklqrNRTSFTQEQ  24014
              ++P   VP     DGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQ
Sbjct  162    WYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQ  221

Query  24015  IEALEK  24032
              IEALEK
Sbjct  222    IEALEK  227
```

UniprotKB also suggests there are two **compositionally biased regions**.



Compositional bias	131 – 209	79	Gln/Gly-rich
Compositional bias	279 – 422	144	Pro/Ser/Thr-rich

Well, hardly an exact match, but there is approximate agreement? One would certainly suppose that **blast** is only willing to mask fairly severe cases of **compositional bias**. It is also probable that **blast** has a rather more mechanistic (i.e. non-biological) interpretation of what **computational bias** is?

SMART also predicts the more obvious region of **computational bias**, rather more generally:

> "An octapeptide and/or a homeodomain can occur C-terminal to the paired domain, as well as a Pro-Ser-Thr-rich C-terminus"

Not important points in themselves of course, the real message of the exercise is that you can discover so much by either:

Looking things up in databases

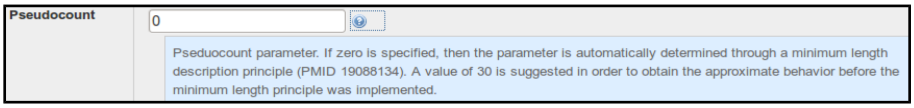
or:

Using the simple analytical software tools yourself.

From your investigations of **PSI-Blast**

What do you suppose the choice of **Pseudocount** might influence?

I clicked with confidence upon the link to the help. It opined as illustrated.



Pseudocount

0

Pseduocount parameter. If zero is specified, then the parameter is automatically determined through a minimum length description principle (PMID 19088134). A value of 30 is suggested in order to obtain the approximate behavior before the minimum length principle was implemented.

I learn that the default choice of **0** does not mean **0**, but instead suggests leaving the value choice to **PSI-Blast**. To discover what a psuedocount might be, I suppose the next step is to read **PMID 19088134**? There is most certainly no elucidation amongst the strangle of words offered here?

The article **Abstract** says:

"**P**osition **s**pecific **s**core **m**atrices (**PSSMs**) are derived from multiple sequence alignments to aid in the recognition of distant protein sequence relationships. The **PSI-BLAST** protein database search program derives the column scores of its **PSSMs** with the aid of **pseudocounts**, added to the observed amino acid counts in a multiple alignment column. In the absence of theory, the number of **pseudocounts** used has been a completely empirical parameter. This article argues that the minimum description length principle can motivate the choice of this parameter. Specifically, for realistic alignments, the principle supports the practice of using a number of **pseudocounts** essentially independent of alignment size. However, it also implies that more highly conserved columns should use fewer **pseudocounts**, increasing the inter-column contrast of the implied **PSSM**s. A new method for calculating **pseudocounts** that significantly improves **PSI-BLAST**'s; retrieval accuracy is now employed by default."

The article itself, continues in like vein ….. how about we close our eyes and accept the defaults? I would just wonder why the whole thing does not commence with, at least an attempt, to answer the question in the forefront of my inquiry, which is .. "**WHAT, in the current context, IS a pseudocount?**". I do not believe it is as tricky as they appear to wish us to believe. I will try again later, when my view of the world is less storm infested.

In the meantime I will take comfort in the claim that:

"A new method for calculating **pseudocounts** that significantly improves **PSI-BLAST**'s; retrieval accuracy is now employed by default."

Jolly good!

2016.12.04: Aha! **Wikipedia** to the rescues once more. Maybe I will donate again? Wonderful service.

One must forgive the **NCBI** people for not explaining what a **pseudocount** is, as they did not, as I first thought, invent the term. It is an idea/strategy of far wider and general application as wikipedia explains.

My interpretation of this article (feel free to disagree/correct) in the current context is:

A **PSSM** is a representation of a **M**ultiple protein **S**equence **A**lignment (**MSA**) based on the amino acid frequencies observed, independently, in each column of that **MSA**. Their purpose is to identify other protein regions of the same size that might be homologous. If a given amino acid is not represented at all in a given column of an **MSA**, the probability of a match for any compared sequence that includes that missing amino acid in that position is implied to be **0** (i.e. impossible!) even if the rest of the region matches extremely well.

Generally speaking, that would be a nonsense! Solution? Add a tiny bit (a **pseudocount** even) to all amino acid counts that come to **0**. Then "*impossible*" becomes "*extremely unlikely*", which makes a bit more sense. A trifle more poetry than science here, but I think I follow the logic.

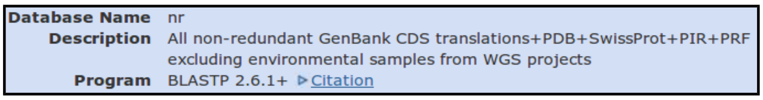
A popular way of implementing **pseudocounts** is due to **Pierre-Simon Laplace**. A French chap who was pretty famous for having good ideas. His strategy, nattily known as **Laplace's Rule of Succession**, was to add a **psuedocount** of **1** to ***ALL*** the real counts and so pervert the message of the data uniformly. Nice one **Pierre**.

I am not entirely sure why, but this all reminds me of one of the many dubious culinary practices of my dear mother (when not in the kitchen, an unsurpassed example of the human female condition!). To-whit, when confronted with a spice or condiment with which she was unfamiliar, she would avoid the unacceptable **zero condition** by adding a swift **pseudocount** (sometimes **two**!) into whatever she was brewing at the time. The principle being that of "*just in case*" and the avoidance of the horror filled possibilities of "*missing an exciting new flavour*".

She would protect the family from any ill effects by assiduously, testing the **psuedocount** side effects upon its most dispensable member … the youngest son, say? If he still frisked after a given period, she would let loose the potion upon the rest of the family. Happily, I survive! But repeated **pseudocount** experimentations may well explain much of the condition of what remains.

How might the inclusion of poor quality and duplicated sequences have been minimised?

At the top of your output is recorded some details of the conditions under which you database search was undertaken. This is a very important step towards making your results reproducible. Not sufficient I would opine. Surely the database versions and a complete record of the parameters used by **blast** are required in order to be able to exactly reproduce a search?



Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.6.1+ ▷Citation

But at least the version of **blast** and the databases that were searched are recorded. The collection of databases searched is rather optimistically called “**nr**”, for non-redundant. A bit of an exaggeration I would think. Surely **PDB** and **SwissProt** overlap a trifle? But let us not be too picky, in fact, surely a noble attempt to remove duplication between these databases has been made, understandably, imperfectly.

The collection of databases that is **nr** includes “*All non-redundant **GenBank CDS** translations*” (aka **GenPept**) which, like it European broad equivalent **TrEMBL**, includes some pretty dubious sequences.

I would think that if one wanted to maximise quality and minimise duplication, it would be best to pick just one good quality database. **SwissProt** is the obvious choice. **blast**, in general, and **PSI-BLAST** in particular, allows such a selection.

However, today the objective is not refinement!!! Bloat is good! More the merrier! Never mind the quality, just admire the volume.

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.

A glance at **PFAM** alignments and **HMM**s.

Actually a very long "glance". Intended to back up a group discussion and/or for people going through these notes by themselves. If you are doing this exercise in a class environment, please just speed read or leave this stuff for later.

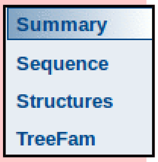
I will provide detailed exercise notes, so you can easily produce similar results yourself, but, a quick browse of the results will be sufficient to back up a class discussion I suggest.

Searching **PFAM**

Go to the home of Pfam at:

http://pfam.xfam.org/

Select the **VIEW A SEQUENCE** option. Enter **pax6_human** (or the corresponding accession code) into the proffered space and press the **Go** button. You will be taken to a **Summary** of the **PFAM** version of what is known about this sequence. Links are provided to several other views of this information, most of which you have already considered. The possibilities include the opportunity to generate easily a phylogenetic tree based upon **PAX6** from the TreeFam database, which is fun if nothing else. We will not be seriously covering phylogeny in the course of these exercises, but why not try it anyway by clicking on the **TreeFam** link.



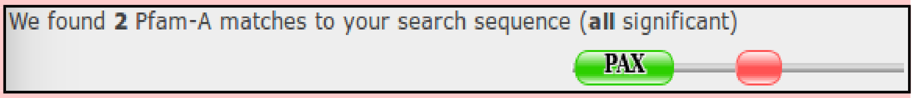
Fine, but you are just looking at what has already been decided. Here we set out to discover, by analysis. How could you use **Pfam** for a sequence that has yet to be annotated.

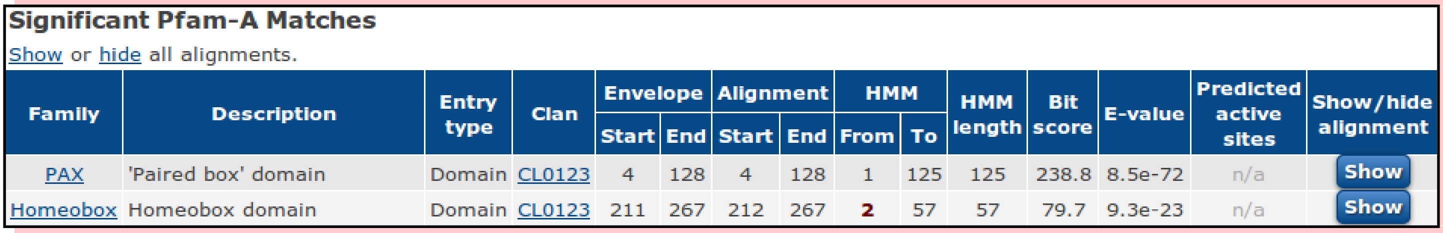
Go back to the home of **Pfam** at:

http://pfam.xfam.org/

This time select the **SEQUENCE SEARCH** option. Copy and paste the sequence of **PAX6_HUMAN** into the appropriate box. Click on the **Go** button.

You should discover nothing you did not expect. This same conclusions, but via direct investigation of the sequence rather than database lookup (or as a component of your **Interpro** analysis).

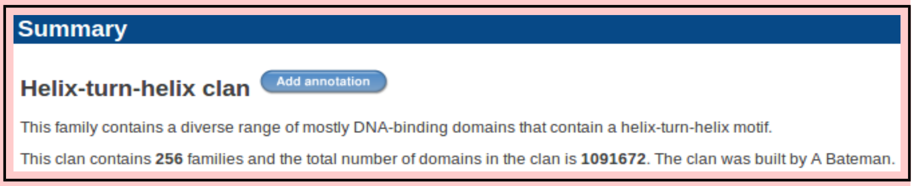

We found **2** Pfam-A matches to your search sequence (**all** significant)


Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
PAX	'Paired box' domain	Domain	CL0123	4	128	4	128	1	125	125	238.8	8.5e-72	n/a	Show
Homeobox	Homeobox domain	Domain	CL0123	211	267	212	267	2	57	57	79.7	9.3e-23	n/a	Show

Have a look around generally, but in the course of your investigations, Click on one of the **CL0123** links. You will see that both the **PAX** and **Homeobox Pfam** families belong to a collection of families (a **Clan**, a similar idea to the **Superfamily** and **Gene3D** domain clusters you met earlier) all of which contain **helix-turn-helix** motifs and are mostly involved in **DNA binding**. Unsurprisingly, the clan in question is the **Helix-turn-helix** clan.


Summary

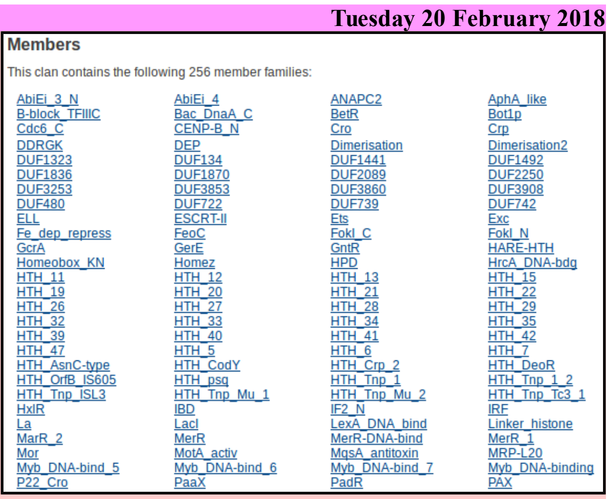
Helix-turn-helix clan Add annotation

This family contains a diverse range of mostly DNA-binding domains that contain a helix-turn-helix motif.

This clan contains **256** families and the total number of domains in the clan is **1091672**. The clan was built by A Bateman.

Notice that **PFAM** reports the matches it finds as being with entries of the **Pfam-A** database (rather than just with **Pfam**). This reflects that, as with a number of the other databases you have considered (including **UniProtKB**, **RefSeq**, **Prosite** …), **PFAM** entries vary considerably in credibility. At one time **PFAM** was offered in two distinct sections, **Pfam-A** and **Pfam-B**. **Pfam-A** was comprised of the more reliable, manually annotated, domain models. **Pfam-B** was entirely computer generated. A few years ago, access to **Pfam-B** was removed from public use as its domain models rarely represented "*meaningful potential new domains*". The **PFAM** team now advise that users regard **Pfam-A** and **PFAM** as effectively synonymous.

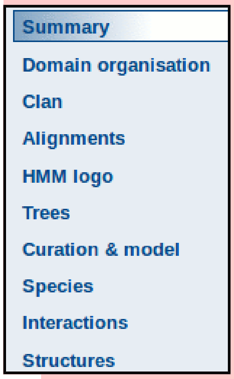
From the **Helix-turn-helix clan** page, select the link to the PFAM PAX family.



Members

This clan contains the following 256 member families:

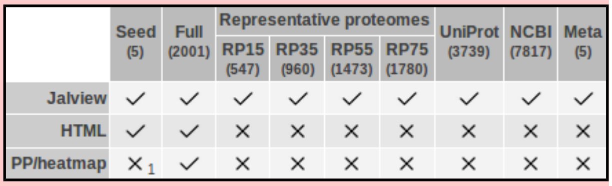
AbiEi_3_N	AbiEi_4	ANAPC2	AphA_like
B-block_TFIIIC	Bac_DnaA_C	BetR	Bot1p
Cdc6_C	CENP-B_N	Cro	Crp
DDRGK	DEP	Dimerisation	Dimerisation2
DUF1323	DUF134	DUF1441	DUF1492
DUF1836	DUF1870	DUF2089	DUF2250
DUF3253	DUF3853	DUF3860	DUF3908
DUF480	DUF722	DUF739	DUF742
ELL	ESCRT-II	Ets	Exc
Fe_dep_repress	FeoC	FokI_C	FokI_N
GcrA	GerE	GntR	HARE-HTH
Homeobox_KN	Homez	HPD	HrcA_DNA-bdg
HTH_11	HTH_12	HTH_13	HTH_15
HTH_19	HTH_20	HTH_21	HTH_22
HTH_26	HTH_27	HTH_28	HTH_29
HTH_32	HTH_33	HTH_34	HTH_35
HTH_39	HTH_40	HTH_41	HTH_42
HTH_47	HTH_5	HTH_6	HTH_7
HTH_AsnC-type	HTH_CodY	HTH_Crp_2	HTH_DeoR
HTH_OrfB_IS605	HTH_psq	HTH_Tnp_1	HTH_Tnp_1_2
HTH_Tnp_ISL3	HTH_Tnp_Mu_1	HTH_Tnp_Mu_2	HTH_Tnp_Tc3_1
HxlR	IBD	IF2_N	IRF
La	LacI	LexA_DNA_bind	Linker_histone
MarR_2	MerR	MerR-DNA-bind	MerR_1
Mor	MotA_activ	MqsA_antitoxin	MRP-L20
Myb_DNA-bind_5	Myb_DNA-bind_6	Myb_DNA-bind_7	Myb_DNA-binding
P22_Cro	PaaX	PadR	PAX



From here, choose Alignments from the menu on the left of the page.

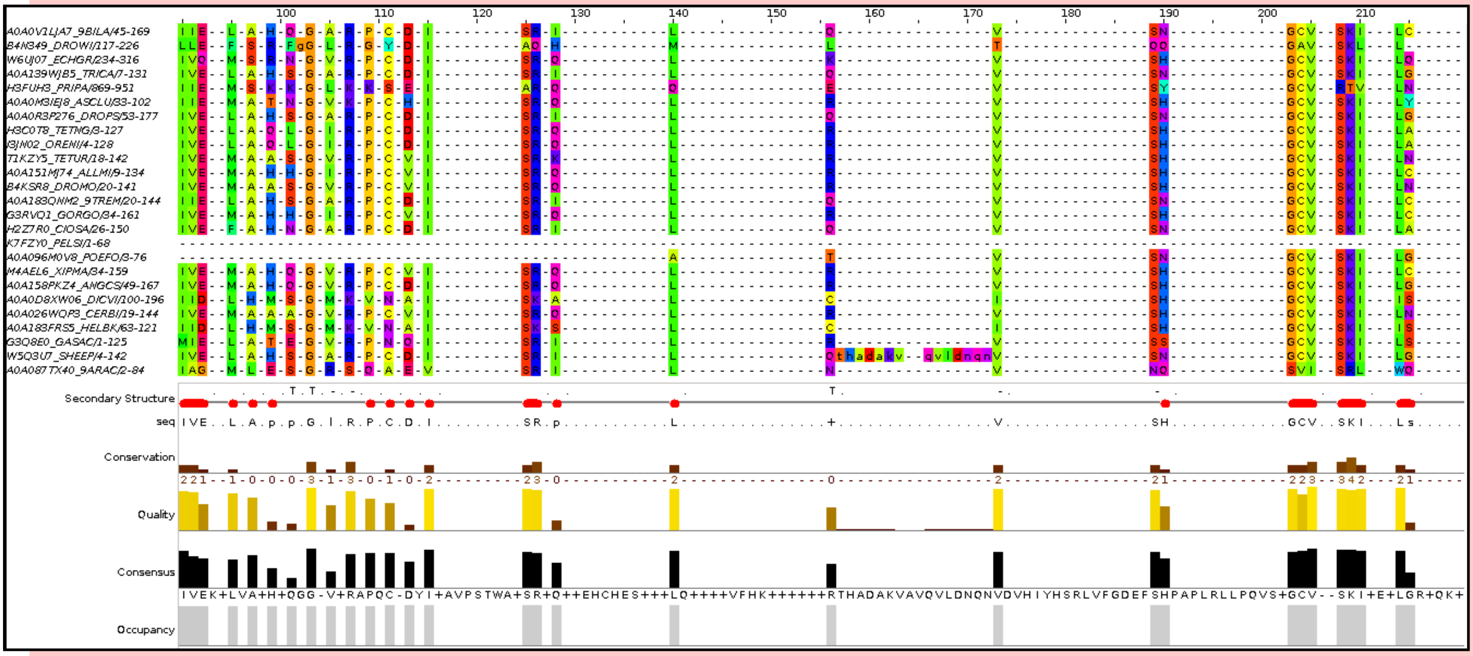
The plan now is to look at two alignments. First an alignment of all the **PAX** domains to which **PFAM** admits the existence (currently **2001**). Then the alignment of the carefully selected representative "**Seed**" sequences (currently just **5**) from which the **PFAM HMM** model for the **PAX** domain is computed.

In the **View options** section, click on the tick in the **Full** column of the Jalview[12] Row. A new window will thrust its way onto your screen offering the requested alignment displayed by **Jalview**.



	Seed (5)	Full (2001)	Representative proteomes RP15 (547)	RP35 (960)	RP55 (1473)	RP75 (1780)	UniProt (3739)	NCBI (7817)	Meta (5)
Jalview	✓	✓	✓	✓	✓	✓	✓	✓	✓
HTML	✓	✓	✗	✗	✗	✗	✗	✗	✗
PP/heatmap	✗ 1	✓	✗	✗	✗	✗	✗	✗	✗

More **Jalview** functionality is claimed when running **Jalview** via **Java Web Start**, so click on the start Jalview via Java Web Start button[13]. In a new window, you should now see the alignment garishly coloured for your delight[14]. The alignment is automatically generated by the program **HMMER3** and, at first glance, is not very impressive! The region illustrated is that around the **isoform 5a 14** amino acid insertion. You should be able to see the gap in that alignment, but … what are all the other gaps?



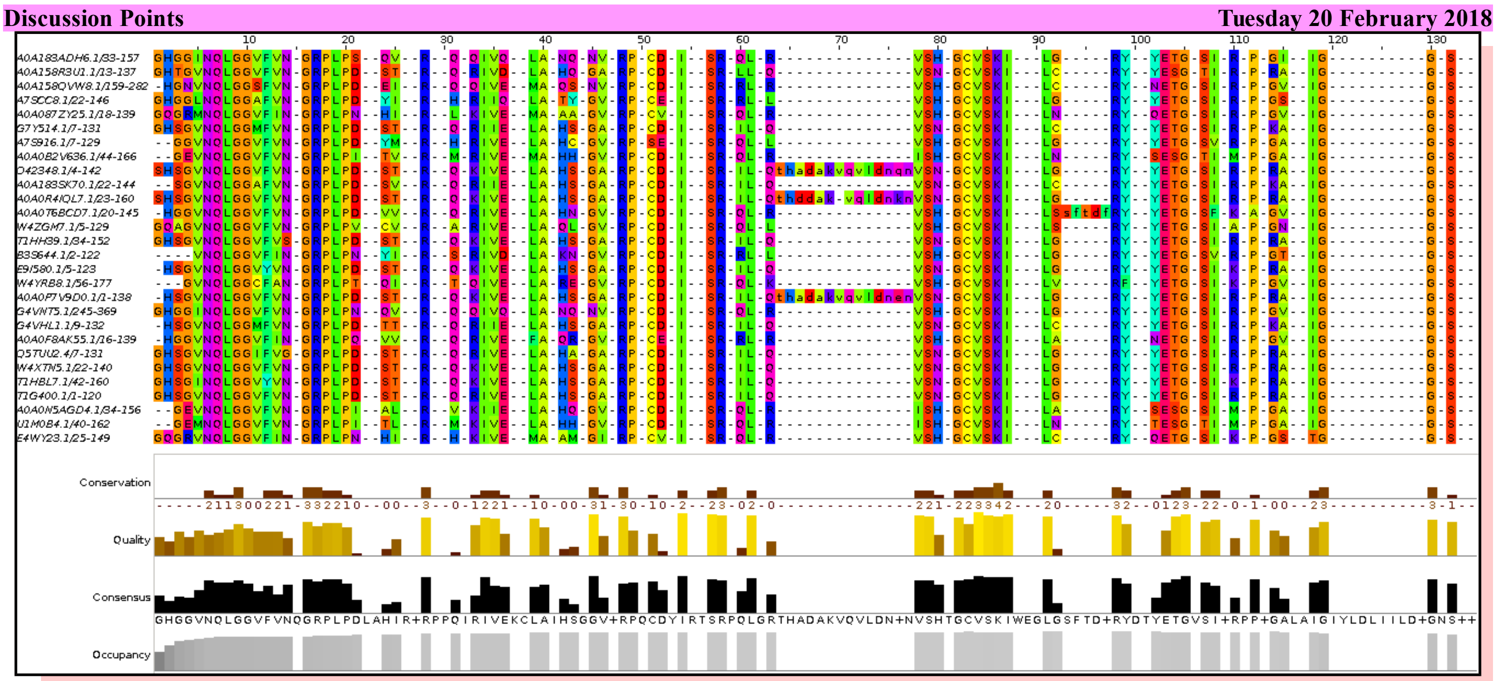
To be fair to **PFAM** (and **HMMER3**), this alignment is generated only for cosmetic purposes. It is the **Seed** alignment that is used to represent a **PAX** domain. Also, a while ago when the were slightly less than **2001** aligned sequences, I discovered that one could massively improve the look of this alignment by removing relatively few (about **10**) outlying sequences (not very good science but very satisfying nonetheless).

Rather than repeat by tedious alignment editing again, I this time elected to look at one of the **Representative proteome** alignments. The illustration here is the same region as above from **RP15**. Much better!

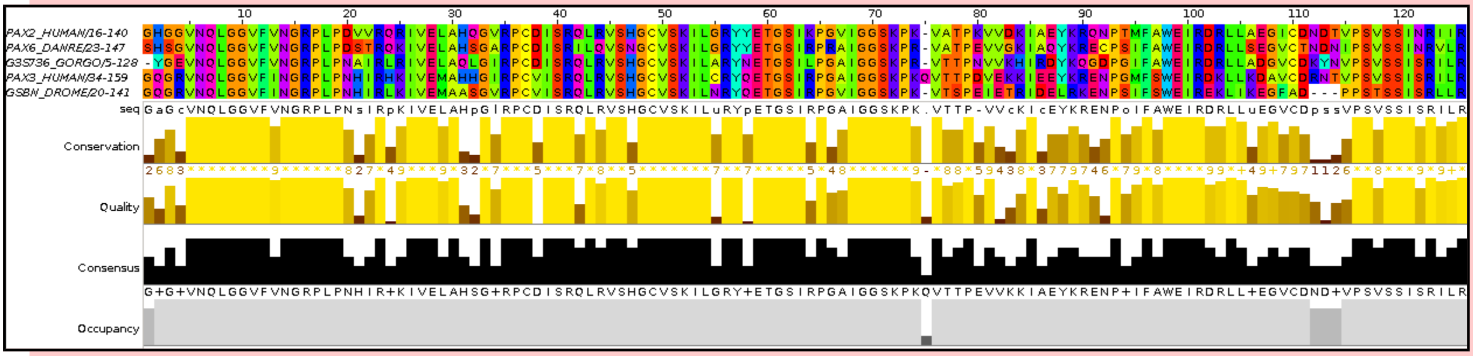
[12] A very nice Java tool for viewing and editing alignments that we will use again.

[13] Exactly what you have to do next should be intuitive (mostly a matter of replying affirmatively to a series of foolish questions), but can vary according to operating system and browser. Whatever is required to display the alignment – **do it**.

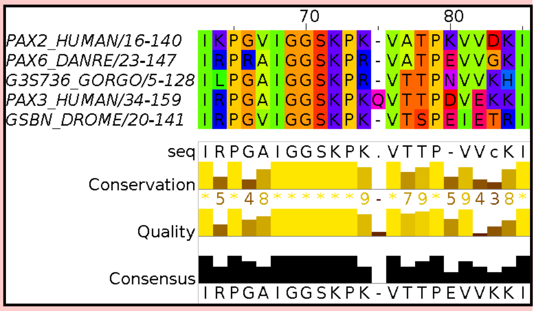
[14] On some systems, there can be problems getting **Java Web Start** to behave properly. Ask if you have any difficulty.



Now to take a look at the Seed alignment. Move back to the Alignments section of the **Pfam PAX** entry page. In the **View options** section, click on the tick in the **Seed** column of the **Jalview** Row. Click on the **start Jalview** via Java Web Start button to start the **Java Web Start** version of **Jalview**.


```
                    10        20        30        40        50        60        70        80        90        100       110       120
PAX2_HUMAN/16-140   GHGGVNQLGGVFVNGRPLPDVVRQRIVELAHQGVRPCDISRQLRVSHGCVSKILGRYYETGSIKPGVIGGSKPK-VATPKVVDKIAEYKRQNPTMFAWEIRDRLLAEGICDNDTVPSVSSINRIIR
PAX6_DANRE/23-147   SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPR-VATPEVVGKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLR
G3S736_GORGO/5-128  -YGEVNQLGGVFVNGRPLPNAIRLRIVELAQLGIRPCDISRQLRVSHGCVSKILARYNETGSILPGAIGGSKPR-VTTPNVVKHIRDYKQGDPGIFAWEIRDRLLADGVCDKYNVPSVSSISRILR
PAX3_HUMAN/34-159   GQGRVNQLGGVFINGRPLPNHIRHKIVEMAHHGIRPCVISRQLRVSHGCVSKILCRYQETGSIRPGAIGGSKPKQVTTPDVEKKIEEYKRENPGMFSWEIRDKLLKDAVCDRNTVPSVSSISRILR
GSBN_DROME/20-141   GQGRVNQLGGVFINGRPLPNHIRLKIVEMAASGVRPCVISRQLRVSHGCVSKILNRYQETGSIRPGVIGGSKPK-VTSPEIETRIDELRKENPSIFSWEIREKLIKEGFAD---PPSTSSISRLLR
seq                 GaGcVNQLGGVFVNGRPLPNsIRpKIVELAHpGIRPCDISRQLRVSHGCVSKILuRYpETGSIRPGAIGGSKPK.VTTP-VVcKIcEYKRENPoIFAWEIRDRLLuEGVCDpssVPSVSSISRILR
Conservation
                    2683*******9******827*49***9*32*7***5***7*8**5*******7**7*****5*48******9-*88*59438*3779746*79*8****99*+49+7971126**8***9*9+*
Quality
Consensus
                    G+G+VNQLGGVFVNGRPLPNHIR+KIVELAHSG+RPCDISRQLRVSHGCVSKILGRY+ETGSIRPGAIGGSKPKQVTTPEVVKKIAEYKRENP+IFAWEIRDRLL+EGVCDND+VPSVSSISRILR
Occupancy
```

Here is the alignment of the **Seed** sequences from which the profile **HMM** for **PAX** is calculated. None of the **5** seed sequences include the **14** extra amino acids noted previously[15]. Human **PAX6** is not a seed sequence.

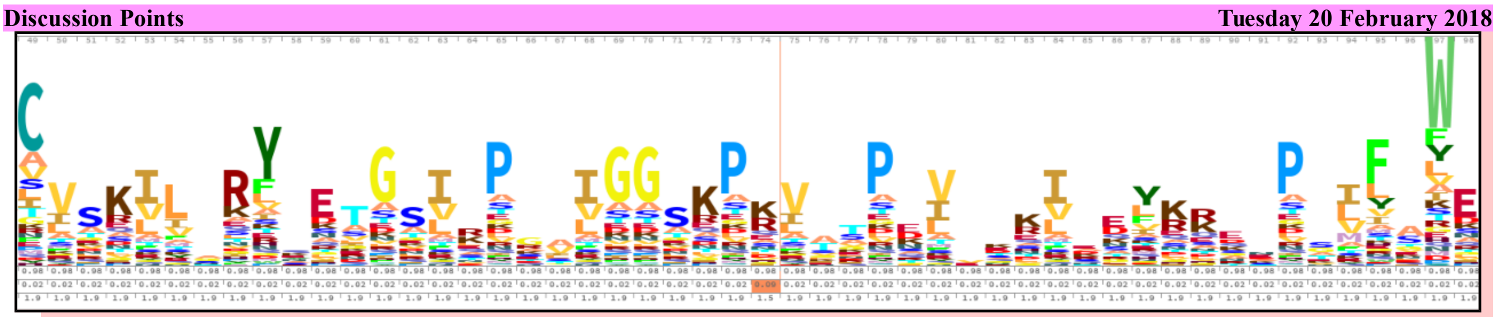

```
                          70        80
PAX2_HUMAN/16-140   IKPGVIGGSKPK-VATPKVVDKI
PAX6_DANRE/23-147   IRPRAIGGSKPR-VATPEVVGKI
G3S736_GORGO/5-128  ILPGAIGGSKPR-VTTPNVVKHI
PAX3_HUMAN/34-159   IRPGAIGGSKPKQVTTPDVEKKI
GSBN_DROME/20-141   IRPGVIGGSKPK-VTSPEIETRI
seq                 IRPGAIGGSKPK.VTTP-VVcKI
Conservation
                    *5*48******9-*79*59438*
Quality
Consensus
                    IRPGAIGGSKPK-VTTPEVVKKI
```

Notice particularly position **75** where **4** of the **5 Seed** sequences are gapped. Only one sequence, **PAX3_HUMAN**, has an amino acid recorded, a **Q** (Glutamine). The **Consensus** character at this point is **"-"**. **Jalview** has it own way to calculate the **Consensus**. Read the documentation for the official explanation. Informally: for positions where there is no dominant amino acid code, + means "more than one possibility", **-** means "predominantly a gap".

Back again to the PFAM PAX family page. Click on the **HMM Logo** link on the left of the page. This is a way of visualising the **HMM** profile computed from the seed sequence alignment you have just been viewing. The logos are indubitably very beautiful. There is a link their documentation just above the picture.

Notice first columns **49** (C), **65**(**P**), **73**(**P**), **92**(**P**) and **97**(**W**). These positions (and several others) represent positions in the **Seed alignment** that are **100%** conserved. Nevertheless, the **Logo** appears to admit the possibility of alternative amino acids in these positions of a real **PAX** domain? This observation illustrates that this **Logo** is not a simplistic representation of an **alignment** (as would be a simple pattern as found in **Prosite**, for example). It is instead, a representation of the profile **HMM** (**pHMM**) derived from the **Seed alignment**. The **pHMM** admits the possibility of a viable **PAX** domain deviating from strict adherence to the pattern suggested by the **Seed alignment**, even where the alignment appears to suggest no variation. These possibilities are computed using such evidences as the scoring matrices discussed earlier.

[15] **Full** alignment columns that are not represented in the seed alignment (and so do not contribute to the calculation of the **HMM**), are shown in lower case. As you can see from the **Full** alignment illustration, including the **14** extra **isoform 5a** positions.



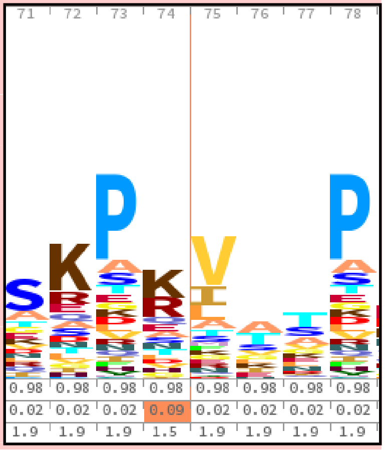
Further evidence of the flexibility of the **pHMM** is the way that **isoform 5a PAX** domains are detected (see **Full alignment**) even though no **isoform 5a** sequences are included in the **Seed** set.

Stated simply, a **pHHM**, of the type used by **PFAM**, is comprised of a number of likelihood scores for each position of the alignment from which it is computed. They are:

- **20** scores representing the likelihood of each amino acid occurring in that position of a "true" domain match
- **1** score representing the likelihood of that position being omitted from a "true" domain match (i.e. a **deletion**)
- **1** score representing the likelihood of the inclusion of an extra amino acid before that position in a "true" domain match (i.e. an **insertion**)
- **20** scores representing the likelihoods of each amino acid being that which is inserted, given an **insertion** event

In the light of that lucid description of a **pHMM**, consider the heavily gapped position of the **Seed alignment** at position **75**. In this position, **4** of the **5** aligned sequences have been gapped, the remaining sequence has a **Q**.

This position does not appear in the **Logo** (although there is a position **75** … which relates to position **76** of the alignment … which seems a bit silly to me!). This implies that the **HMM** represents the data at position **75** thus:



"Generally not present, but a relatively high chance of an insertion which is most likely to be a **Q**"

The alternative, equivalent, representation would be:

"Generally a **Q**, but a relatively high chance of a deletion"

Had the second alternative been selected, the **Logo** would have shown a healthy **Q** at position **75**. The **Logo** is not sufficiently sophisticated to indicate the high deletion likelihood that would be recorded in the **pHMM**.

A thin brownish line is placed in the **Logo** to indicate where position **75** was omitted. The **Logo** is not a precise enough representation to clearly show that the insertion is likely to be a **Q** …. but this will be recorded in the **pHMM**.

DPJ – 2017.12.23